WP. 20
ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)

CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION

STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Tarragona, Spain, 26-28 October 2011)

Topic (iv): Balancing data quality and data confidentiality

Quantitative Methods to Assess Data Confidentiality and Data Utility for Microdata in Japan

Prepared by Ito Shinsuke, Meikai University and
Murata Mariko, Statistical Information Institute for Consulting and Analysis, Japan

Quantitative Methods to Assess Data Confidentiality and Data Utility for Microdata in Japan

Shinsuke Ito[*] and Mariko Murata[**]

[*] Faculty of Economics, Meikai University, 1 Akemi Urayasu Chiba 279-8550 Japan
E-mail: ssitoh@meikai.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 National Statistics Center.

[**] Statistical Information Institute for Consulting and Analysis, 5F Nogaku-Shorin Bldg., 3-6 Kanda-Jinbocho, Chiyoda-ku Tokyo 101-0051 Japan
E-mail: marikomurata@sinfonica.or.jp
Mariko Murata was a research fellow at the National Statistics Center until the end of June 2011.

Abstract: Compared to the United States, Canada, Australia and many European countries, Japan has only recently begun making anonymized microdata available, and there exist only few empirical studies on disclosure limitation methods, disclosure risk and information loss regarding microdata in Japan. Although perturbative methods such as additive noise and swapping including microaggregation are not currently used for official anonymized microdata in Japan, this lack of empirical studies makes it worthwhile to examine the applicability of perturbative methods to official Japanese microdata.

This paper gives an overview of disclosure limitation methods that are currently used for official microdata in Japan, and on this basis describes the potential of pertubative methods such as microaggregation and additive noise as disclosure limitation methods. The paper then explores issues related to data utility and data confidentiality of microdata in order to determine the relevance of perturbative methods as a disclosure avoidance method for official microdata in Japan.

1 Introduction: Current Disclosure Limitation Methods for Official Microdata in Japan

Following the revision of the Statistics Act, anonymized microdata from official statistics have been released in Japan since April 2009. Currently, anonymized official microdata from the ‘National Survey of Family Income and Expenditure’, the ‘Survey on Time Use and Leisure Activities’, the ‘Employment Status Survey’ and the ‘Housing and Land Survey’, all of which are conducted by the Statistical Bureau of Japan, are available. Anonymized official microdata from the ‘Comprehensive Survey of Living Conditions’ conducted by the Ministry of Health, Labour and Welfare are also made available.

Table 1 gives an overview of the disclosure limitation methods applied to anonymized microdata from official statistics based on surveys conducted by the Statistical Bureau of Japan. These include resampling, recoding, top-coding and bott-

		National Survey of Family Income and Expenditure	Survey on Time Use and Leisure Activities	Employment Status Survey	Housing and Land Survey
Comparison					
	Resampling Rate	80%	80%	80%	10%
	Geographical Area	Three major metropolitan areas' or 'Others'	Three major metropolitan areas' or 'Others'	Three major metropolitan areas' or 'Others'	Prefectures
	Age Bracket	Five-year age brackets persons 15 years or older and one-year age brackets for children under 15	Five-year age brackets persons 10 years or older and one-year age brackets for children under 10	Five-year age brackets persons 15 years or older and one-year age brackets for children under 15	Five-year age brackets persons 15 years or older and one-year age brackets for children under 15
	Classification According to Age	Age of persons 85 years and over is top-coded.	Age of persons 85 years and over is top-coded.	Age of persons 85 years and over is top-coded.	Age of persons 85 years and over is top-coded.
	Household Members	Households with eight or more members are deleted.	Households with eight or more members are deleted.	Households with eight or more members are deleted.	Households with eight or more members are deleted.
	Children	Households with three or more members of the same age are deleted.	Households with three or more members of the same age are deleted.	Households with three or more members of the same age are deleted.	Households with three or more members of the same age are deleted.
Specific Characteristics					
	Dwelling Size	Top-coding and/or bottom-coding	-	-	Top-coding and/or bottom-coding
	Yearly Household Income etc.	Top-coding and deletion of further details on items	-	-	-

Note: Anonymized Microdata from the Comprehensive Survey of Living Conditions conducted by the Ministry of Health, Labour and Welfare are also currently released in Japan, and several disclosure limitation methods such as top-coding and/or bottom-coding and recoding are applied for the anonymized microdata from the Comprehensive Survey of Living Conditions.
Source: http://rcisss.ier.hit-u.ac.jp/Japanese/micro/anonym02.html (Japanese only).

Table 1: List of Disclosure Limitation Methods Applied to Anonymized Microdata from Official Statistics Based on Surveys Conducted by the Statistical Bureau of Japan

om-coding as well as deletion of direct identifiers such as individual names or addresses. The ‘National Survey of Family Income and Expenditure’, the ‘Survey on Time Use and Leisure Activities’ and the ‘Employment Status Survey’ each use a resample rate of 80 percent, while the resample rate for the ‘Housing and Land Survey’ is 10 percent due to the larger sample size.

Information on the geographical area included in the ‘National Survey of Family Income and Expenditure’, the ‘Employment Status Survey’ and the ‘Survey on Time Use and Leisure Activities’ is broken down into ‘three major metropolitan areas’ (comprising the Tokyo area, the Nagoya area and the Osaka area) and ‘Others’ (covering all other areas of Japan). As a result, all three surveys lack detailed information on geographic areas outside the major metropolitan areas.

Individuals’ age is recoded, and therefore age is available only in five-year brackets. In addition, the age of persons 85 years and older is top-coded, resulting in a loss of detail. On the other hand, for anonymized microdata of the Survey on Time Use and Leisure Activities the age of children under 10 years is available in one-year age brackets.

Households who have eight or more members in total and households who have three or more members in the same age are deleted in all four surveys. Top-coding and/or bottom-coding are also applied towards quantitative attributes such as dwelling size, yearly household income, household savings and household liabilities, and further details are often not included.

Two types of anonymized official microdata for the ‘Comprehensive Survey of Living Conditions’ are currently released. These differ in the number of records and the amount of detail on survey items contained in the microdata file. A number of disclosure limitation methods such as top-coding and/or bottom-coding and recoding are applied to both.

The United States, Canada, Australia and many European countries have a longer history of making anonymized microdata available than Japan, and generally release several kinds of anonymized microdata from official statistics such as Population Census and Labour Force Survey. In the United States, Public Use Microdata Sample (PUMS) from the Census of Population and Housing have been publicly released since 1963, and for the 2000 Census, 1% and 5% PUMS files that contain different geographical information have been released.

In the United Kingdom, Samples of Anonymised Records (SARs) from both the 1991 and 2001 Population Census have been released. The 1991 SARs contain Household SAR and Individual SAR. Whereas household SAR are compiled by selecting 1% of records on the level of household unit and are hierarchically structured, individual SAR are compiled by selecting 1% of records on the level of individual persons and contain more detailed geographical information than household SAR. In addition, Small Area Microdata (SAM) have been released for the first time following the 2001 UK Population Census. SAM contain more detailed information on geography than individual SAR and therefore allow for the comparative analysis of smaller geographic areas.

In Japan, a release of more and different types of official microdata might take place in the future. This development would likely require extensive research into the quantitative assessment of disclosure risks and information loss for microdata. However, at present only few empirical studies on disclosure limitation methods, disclosure risk and information loss exist in Japan. The limited research on disclosure limitation methods for individual data might be a contributing factor that has prevented the release of a wider variety of microdata from Japanese official statistics.

This paper aims to propose quantitative methods for assessing data confidentiality and data utility for microdata, and to examine the applicability of these methods for original official microdata.

2 Disclosure Limitation Methods in Japan: An Illustration of Microaggregation

In Europe and North America, a number of perturbative methods are applied to official microdata to achieve disclosure limitation. In the United States, perturbative methods such as noise addition and data swapping are used in the creation of Public Use Microdata Sample (PUMS) of the 2000 Census (Zayatz (2007)). In the United Kingdom, PRAM (Post-Randomisation Method) is applied to Samples of Anonymised Records (SARs) from the 2001 Population Census (De Kort and Wathan (2009)).

While perturbative methods such as additive noise and swapping including microaggregation are not currently adopted for official anonymized microdata in Japan, it is worth examining their applicability for official microdata in Japan in order to enlarge the number of available disclosure limitation methods and potentially improve the usability of anonymized official microdata in case the methods are adopted.

In Japan, only a small number of empirical studies on the effectiveness of disclosure limitation methods including perturbative methods have been conducted. Ito *et al*. (2008), Ito (2009) and Ito and Takano (2011) have conducted empirical studies on the effectiveness of microaggregation as a disclosure limitation method. These papers have examined the characteristics of microaggregation, evaluated the effectiveness of microaggregation for individual data from Japanese official statistics, and were among the first in Japan to advocate methods for creating micro-aggregated data that closely resembles individual data using multi-dimensional tabulation. The proposed method of microaggregation in the above papers involves the creation of records with common values for all types of qualitative attributes based on multi-dimensional tabulation. In a next step, records with common values for qualitative attributes are sorted and divided into groups larger than a specific minimum size, and the value of each quantitative attribute is replaced with a measure of central tendency (ex. average value etc.) within each group based on research by Defays and Anwar (1998) and Domingo-Ferrer and Mateo-Sanz (2002).

These papers have also created micro-aggregated data based on individual data from the 'National Survey of Family Income and Expenditure' using techniques such as the individual ranking method, and verified the degree of similarity between micro-aggregated data and original data by measuring information loss. For quantitative attributes, these paper have assessed the information loss of masked data compared to original data using measures such as mean square error, mean absolute error, mean variation of attributes' values, variance-covariance matrices, and correlation matrices based on research by Domingo-Ferrer and Torra (2001a).

Ito (2010) and Ito and Takano (2011) have proposed an appropriate method for assessing data confidentiality of microdata in Japan based on a review of disclosure risk assessment methods for microdata used in Europe and North America, and examined the applicability of this method for micro-aggregated data generated from the Japanese 'National Survey of Family Income and Expenditure'. To assess the data confidentiality of microdata for quantitative attributes, these papers have used methods for measuring the relative risk of various kinds of masked data compared to original data based on record linkage techniques such as developed by Domingo-Ferrer and Torra (2001b), and assessed the degree of "true match" of original data using deterministic record linkage and distance-based record linkage based on research by Domingo-Ferrer and Torra (2001b), Winglee *et al.* (2002) and Herzog *et al.*(2007).

3 The Effectiveness of Perturbative Methods for Microdata in Japan Based on Microdata from the Family Income and Expenditure Survey

This research focuses on the comparative analysis of data confidentiality and data utility of masked data created by applying several perturbative methods to original official microdata. Data utility and data confidentiality of several kinds of masked data are then assessed through quantitative methods. The survey data used for this research is original microdata from the January 2009 'Family Income and Expenditure Survey' which includes 4,220 households where the head of household is currently employed. The disclosure limitation methods applied to this microdata are microaggregation, noise addition, categorization of quantitative attributes, and combined use of the above methods.

In this research, six quantitative attributes of wages/salaries and consumption expenditure in the month when the survey was conducted as well as yearly household income, household savings, household liabilities and dwelling size are perturbed. Before applying perturbation to these quantitative attributes, records are clustered within each category of type of tenure of dwelling (the above categories are recoded into five categories).

The content of perturbative methods used in this research is as follows: For microaggregation, individual ranking method (‘MicroIR’) and sum of Z-scores method (‘MicroZscore’) are used to generate the masked data. For noise addition, Gaussian noise is added for each value of quantitative attributes (‘AddNoise’). If the standard deviation of quantitative attributes of original data is *s*, noise with *N(0, ps)* is generated (where *p* is a parameter) based on research by Domingo-Ferrer and Torra (2001b). Values of *p* used in this research range from 0.01 to 0.5 (ex. ‘AddNoise0.01’). For categorization of quantitative attributes, 10-quantile and 20-quantile are used to generate the masked data (‘CTG10’ or ‘CTG20’). Values of quantitative attributes to which categorization is applied are replaced with averages of quantitative attributes within each category.

On the other hand, for the combined use of disclosure limitation methods, (1) the combined use of individual ranking method and sum of Z-scores method (‘MicroIRZscore’), (2) the combined use of microaggregation (individual ranking method or sum of Z-scores method) and categorization (10-quantile) (‘MicroIRCTG10’ or ‘MicroZscoreCTG10’) and (3) the combined use of noise addition (*p*=0.10,0.16,0.30,0.50) and categorization (ex. ‘AddNoise0.10CTG10’) are applied. For (1), the individual ranking method is applied to wages/salaries and consumption expenditure, and sum of Z-scores method is applied to the other four quantitative attributes such as yearly household income. For (2), wages/salaries and consumption expenditure are micro-aggregated, and the other four attributes are categorized based on 10-quantile. For (3), Gaussian noise is added to wages/salaries and consumption expenditure, and categorization is applied to the other four attributes based on 10-quantile.

Information loss of masked data compared to original data is assessed using measures of information loss including mean square error and mean variation of correlation matrices. For the individual ranking method, information loss is measured using the average of weights generated after each attribute is micro-aggregated. Weights on original data are also used for the combined use of individual ranking method, sum of Z-scores method and the combined use of individual ranking method and categorization.

To assess data confidentiality for quantitative attributes of microdata, this research measures the degree of “one-to-one true match” of original data using distance-based record linkage based on empirical research by Ito (2010) and Ito and Takano (2011). “One-to-one true match” refers to the relationship between the matched records with identical household numbers in the original data and masked data. Distance-based record linkage is conducted using standardized Euclid distance based on research by Torra *et al*. (2006). Key variables used for distance-based record linkage are type of tenure of dwelling, number of household members, number of working members, and age of the head of household (five-year age brackets) as well as six quantitative variables such as wages/salaries. For age of the head of household, the median within each bracket is used.

Table 2 presents the result of the quantitative assessment of data utility for several masked data using perturbative methods based on original microdata from the ‘Family Income and Expenditure Survey’. The table details the information loss of masked data compared to original data based on mean square error and mean variation of correlation matrices. This result shows that micro-aggregated data created using the individual ranking method is considerably closer to the original data than micro-aggregated data created using sum of Z-scores method. With regards to noise addition, information loss becomes more extensive as the value of *p* increases. Masked data created using 20-quantile is closer to the original data than masked data created using 10-quantile. Also, the mean square error of masked data created using 10-quantile compared to original data is almost the same as of data created using noise addition in the case of p=0.50. For the combined use of disclosure limitation methods, masked data created by applying categorization to the four attributes such as yearly household income is closer than masked data created by applying sum of Z-scores method to the above attributes.

Table 3 also presents the result of a quantitative assessment of data confidentiality for several masked data using perturbative methods. The table shows the number and percentage of records which result in “one-to-one true match”, the number of records which result in “false match[1]” and the number of records which correspond to “n:m match[2]”. In Table 3 the percentage of records that result in a “true match” is lower for the sum of Z-scores method than the percentage for any other method. For noise addition, the percentage of records which result in a “true match” decreases as value of *p* increases, and the percentage of records which result in a “false match” and that of records which result in “n:m match” also increases. In addition, when applying categorization to original data in order to create masked data the percentage of records which result in “true match” is high.

4 Conclusion

This paper proposes methods for quantitatively assessing data utility and degree of confidentiality for several types of masked data created through perturbative methods are applied, and conducts a comparative analysis of information loss and deg-

[1] In the case of “false match”, the relationship between the matched records excluding matched records which have identical household numbers in the original data and masked data is a one-to-one match.

[2] "n:m match" refers to the relationship between the matched records in the original data and masked data, but excluding "one-to-one true match" and "false match" and including 1:n match or n:1 match.

	Information Loss	
	Mean Square Error	Mean Variation
MicroIR	0.000039	0.020757
MicroZscore	0.025357	0.736120
AddNoise0.01	0.000000	0.000708
AddNoise0.05	0.000002	0.003517
AddNoise0.10	0.000012	0.009386
AddNoise0.16	0.000053	0.020264
AddNoise0.20	0.000110	0.029471
AddNoise0.30	0.000432	0.058772
AddNoise0.50	0.002264	0.131800
CTG10	0.002139	0.107590
CTG20	0.001198	0.079517
MicroIRZscore	0.013403	0.543650
MicroIRCTG10	0.000078	0.039800
MiceoZscoreCTG10	0.007535	0.124640
AddNoise0.10CTG10	0.000078	0.034988
AddNoise0.16CTG10	0.000088	0.035931
AddNoise0.30CTG10	0.000186	0.040967
AddNoise0.50CTG10	0.000690	0.066631

Table 2: Assessing the Degree of Data Utility for Microdata from the ‘Family Income and Expenditure Survey’

	one-to-one true match		false match	n:m match
MicroIR	4,203	99.60%	0	17
MicroZscore	860	20.38%	611	2,749
AddNoise0.01	4,218	99.95%	0	2
AddNoise0.05	4,214	99.86%	0	6
AddNoise0.10	4,165	98.70%	1	54
AddNoise0.16	3,980	94.31%	15	225
AddNoise0.20	3,748	88.82%	39	433
AddNoise0.30	3,076	72.89%	199	945
AddNoise0.50	1,838	43.55%	556	1,826
CTG10	3,558	84.31%	6	656
CTG20	3,934	93.22%	1	285
MicroIRZscore	1,633	38.70%	435	2,152
MicroIRCTG10	3,800	90.05%	2	418
MiceoZscoreCTG10	2,968	70.33%	68	1,184
AddNoise0.10CTG10	3,780	89.57%	4	436
AddNoise0.16CTG10	3,695	87.56%	6	519
AddNoise0.30CTG10	3,302	78.25%	56	862
AddNoise0.50CTG10	2,657	62.96%	206	1,357

Table 3: Assessing the Degree of Data Confidentiality for Microdata from the ‘Family Income and Expenditure Survey’

ree of confidentiality of this masked data to examine the effectiveness of perturbative methods as a disclosure limitation method for official microdata in Japan.

The result of this empirical research shows that the information loss of masked data tends to be larger for higher levels of noise addition. The result also shows that based on the combined use of disclosure limitation methods such as noise addition and microaggragation, the percentage of records which result in “true match” is potentially small.

These methods allow the relative measurement of information loss and degree of confidentiality for masked data created through perturbation. This research aims to contribute to the examination of perturbative methods towards a potential future adoption of these methods for the creation of anonymized official microdata in Japan.

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

Defays, D. and Anwar, M.N. (1998) “Masking Microdata Using Micro-Aggregation”, *Journal of Official Statistics*, Vol.14, No.4, pp.449-461.

De Kort, S. and Wathan, J. (2009) “Guide to Imputation and Perturbation in the Samples of Anonymised Records”

http://www.ccsr.ac.uk/sars/resources/imputation.doc

Domingo-Ferrer, J. and Torra, V. (2001a) “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.

Domingo-Ferrer, J. and Torra, V. (2001b) “A Quantitative Comparison of Disclosure Control Methods for Microdata”, Doyle *et al*.(eds.) *Confidentiality, Disclosure, and Data Access: Theory and Practical Application for Statistical Agencies*, Elsevier Science, Amsterdam, pp.111-133.

Domingo-Ferrer, J. and Mateo-Sanz, J. M.(2002) “Practical Data-oriented Microaggregation for Statistical Disclosure Control”, *IEEE Transactions on Knowledge and Data Engineering*, vol.14, no.1, pp.189-201.

Herzog, T. N., Scheuren, F. J., Winkler, W. E.(2007) *Data Quality and Record Linkage Techniques*, Springer, New York.

Ito, S., Isobe, S., Akiyama, H.(2008) “A Study on Effectiveness of Microaggregation as Disclosure limitation methods: Based on National Survey of Family Income and Expenditure”, *NSTAC Working Paper*, No.10, pp.33-66 (in Japanese).

Ito, S.(2009) “On Microaggregation as Disclosure limitation methods”, *Journal of Economics, Kumamoto Gakuen University*, Vol.15, No.3・4, pp.197-232 (in Japanese).

Ito, S.(2010) “A Method to Quantitatively Assess the Confidentiality of Official Microdata”, *Meikai Economic Review*, Vol.22, No.2, pp.1-17 (in Japanese).

Ito, S. and Takano, M. (2011) “A Method to Quantitatively Assess Confidentiality and Potential Usage of Official Microdata in Japan”, Paper presented at the 58th World Statistics Congress of the International Statistical Institute at the Convention Centre Dublin.

Torra, V., Abowd, J. and Domingo-Ferrer, J (2006) “Using Mahalanobis Distance-Based Record Linkage for Disclosure Risk Assessment”, Domingo-Ferrer, J. and Franconi, L.(eds.) *Privacy in Statistical Databases: CENEX-SDC Project International Conference, PSD 2006 Rome, Italy, December 13-15, 2006 : Proceedings*, Springer, Berlin, pp.233–242.

Winglee, M., Valliant, R., Clark, J., Lim, Y., Weber, M., Strudler, M. (2002) “Assessing Disclosure Protection for the SOI Public Use File”, Paper Presented at Proceedings of the Annual Meeting of the American Statistical Association.

http://www.amstat.org/sections/SRMS/Proceedings/

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.