A Method to Quantitatively Assess Confidentiality and Potential Usage of Official Microdata in Japan

Shinsuke Ito[1]
Meikai University, Faculty of Economics
1 Akemi Urayasu
Chiba 279-8550 Japan
E-mail: ssitoh@meikai.ac.jp

Masahiro Takano[2]
Economic and Social Research Institute
3-1-1 Kasumigaseki Chiyoda-ku
Tokyo 100-8970 Japan
E-mail: masahiro.takano@cao.go.jp

1. Introduction

Following the revision of the Statistics Act, anonymized microdata from official statistics have been released in Japan since April 2009. Whereas the United States, Canada, Australia and many European countries generally release several different types of anonymized microdata, official releases of microdata in Japan have been limited to one type of anonymized microdata.

In the United States, Public Use Microdata Sample (PUMS) from the Census of Population and Housing have been publicly released since 1963. 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. 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 as part of data releases 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 many of the above countries, disclosure limitation methods are used to protect confidential information contained in the data, and extensive research exists on how to quantitatively assess disclosure risks and information loss for microdata. In Japan, there are still few empirical studies on disclosure limitation methods, disclosure risk and information loss. The small number of empirical studies on the effectiveness of various disclosure limitation methods for individual data might be a factor that prevents the release of a wider variety of microdata from Japanese official statistics.

This paper gives an overview of disclosure avoidance methods that are currently used for official microdata in Japan. Furthermore, it describes microaggregation as a methodology and discusses the effectiveness of microaggregation as a disclosure limitation method for individual data from Japanese official statistics. Based on these findings, this paper then proposes an appropriate method for assessing the confidentiality and potential usages of microdata and examines the applicability of this method for original official microdata.

---

[1] Shinsuke Ito is a part-time researcher at the National Statistics Center and conducts research on disclosure limitation methods for microdata in co-ordination with officials at the National Statistics Center.

[2] Masahiro Takano was an official at the National Statics Center until the end of March 2011, and is currently a member of the Economic and Social Research Institute (ESRI, since April 2011).

2. Current Disclosure Limitation Methods for Official Microdata in Japan

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" are publicly available in Japan. Table 1 gives an overview of the disclosure limitation methods used in each survey. These include resampling, recoding, top-coding, bottom-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 of 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). Therefore, for all three surveys there is a lack of detailed information on geographic areas outside the major metropolitan areas. Individuals" age is recoded, and as a result age is available only in five-year brackets. In addition, the age of persons 85 years and over is top-coded and therefore the level of detail is severely limited. On the other hand, the age of children under10 years is available in one-year brackets. Furthermore, 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 applied towards quantitative attributes such as dwelling size. Other important attributes such as yearly household income, household savings and household liabilities are top-coded, and further details on each item are not included.

3. Microaggregation as a Disclosure Limitation Method in Japan

The National Statistics Center has conducted empirical studies about the effectiveness of disclosure limitation methods, and has acknowledged microaggregation as one disclosure limitation method. Ito *et al.* (2008) and Ito (2009) examine the characteristics of microaggregation, evaluate the effectiveness of microaggregation for individual data from Japanese official statistics, and are the first to advocate methods for creating micro-aggregated data that closely resembles individual data using multi-dimensional tabulation in Japan. The proposed method of microaggregation 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 other qualitative attributes are sorted and divided into groups larger than a specific minimum size, and the value of each quantitative attribute for records 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). Second, these papers create micro-aggregated data based on individual data from the „National Survey of Family Income and Expenditure" (original data) using the technique of microaggregation such as individual ranking method, and verify the degree of similarity between micro-aggregated data and original data. This empirical research shows that micro-aggregated data created using the individual ranking method is considerably closer to the original data than micro-aggregated data created using microaggregation without sorting (i.e. dividing records into groups according to the arrangement of records in the original data and replacing value of each quantitative attribute for records with average values within each group).

Ito (2010) proposes an appropriate method for assessing the confidentiality of microdata in Japan based on a review of disclosure risk assessment methods for microdata used in Europe and North America, and examines the applicability of this method for micro-aggregated data generated from the Japanese „National Survey of Family Income and Expenditure". Ito (2010) also compares the degree of confidentiality based on record linkage technique, and assesses the degree of "true match" of original data and micro-aggregated data using deterministic record linkage and distance-based record linkage based on research by Domingo-Ferrer and Torra (2001b) and Winglee *et al*. (2002). The results show that for all kinds of micro-aggregated data, the percentage of records that result in a "true match" is higher for distance-based record linkage than for deter-

Table 1: List of Disclosure Limitation Methods Applied to Anonymized Official Microdata Currently Released in Japan

		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 brackest for children under 15	Five-year age brackets persons 10 years or older and one-year brackest for children under 10	Five-year age brackets persons 15 years or older and one-year brackest for children under 15	Five-year age brackets persons 15 years or older and one-year brackest for children under 15
	Classification According to Age	Age of persons 85 years and over is topcoded.	Age of persons 85 years and over is topcoded.	Age of persons 85 years and over is topcoded.	Age of persons 85 years and over is topcoded.
	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	-	-	-

Source: http://rcisss.ier.hit-u.ac.jp/Japanese/micro/anonym02.html (Japanese only).

ministic record linkage.

4. Usage Potential and Degree of Confidentiality of microdata

Several methods of assessing the usage potential of microdata exist. For example, numerous papers compare original data and masked data based on information loss (Mateo-Sanz *et al*. (2005) etc.). Woo *et al*. (2009) develop global measures of data utility for masked data based on propensity score, cluster analysis, or cumulative distribution function.

When measuring information loss for microdata, it is important to use a method of assessing usage potential that fits the characteristics of attributes. For quantitative attributes, this paper aims to assess the information loss of masked data compared to original data using measures of information loss 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). For qualitative attributes, a method of measuring information loss based on entropy-based measures has been developed by Kooiman *et al*.(1998), Domingo-Ferrer and Torra (2001a).

To assess the disclosure risk of microdata for quantitative attributes, methods for measuring the relative risk of various kinds of masked data compared to original data based on record linkage techniques have been developed, for example by Domingo-Ferrer and Torra (2001b). This paper presents a comparative analysis of the degree of confidentiality based on record linkage techniques, and assesses the degree of “true match” of original data using deterministic record linkage, distance-based record linkage, and probabilistic record

linkage (Torra and Domingo-Ferrer (2003)), based on the empirical research by Ito (2010). Distance-based record linkage is conducted using both Euclid distance and Mahalanobis distance based on research by Torra *et al*. (2006).

Domingo-Ferrer and Torra (2001b) also proposes a method of assessing disclosure risk based on probabilistic record linkage for qualitative attributes. This paper suggests the creation of contingency tables for both original data and masked data and compares the percentage of unique cells in each contingency table of original data with that of masked data in order to measure of the degree of confidentiality quantitatively.

This paper also includes an empirical study that assesses the usage potential and the degree of confidentiality of masked data based on original microdata from the 2004 „National Survey of Family Income and Expenditure". For this, several kinds of masked test data are created. Five quantitative attributes including yearly household income and household savings are microaggregated using techniques such as the individual ranking method. The qualitative attributes of occupation of the head of households and type of tenure of dwelling are recoded, and the age of the head of households is recoded in five-year age brackets and/or topcoded for persons 85 years or older. Using entropy-based measures of information loss for qualitative attributes, the usage potential of the microdata is assessed quantitatively, and a way to measure information loss for both qualitative attributes and quantitative attributes is demonstrated. This paper also measures the degree of confidentiality of masked data to original data using contingency tables for qualitative attributes, and examines the potential of assessing the degree of confidentiality of masked data for qualitative attributes, including the possibility of using record linkage techniques to measure the relative degree of confidentiality of masked data for quantitative attributes. Lastly, this research includes a comparative analysis of the usage potential and degree of confidentiality of several masked test data using the R-U map by Duncan *et al*. (2001), and identifies the applicability of the R-U map for comparative analysis of the usage potential and degree of confidentiality based on official microdata in Japan.

5. Conclusion

This paper outlines disclosure avoidance methods that are currently used for official microdata in Japan, and examines microaggregation as a disclosure avoidance method for official microdata in Japan. 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 the applicability of perturbative methods to official microdata in Japan.

Second, this paper proposes methods of quantitatively assessing the usage potential and degree of confidentiality for official microdata in Japan, and conducts a comparative analysis of information loss and degree of confidentiality of masked data using the R-U map. This method enables a relative measurement of information loss and degree of confidentiality of masked data using perturbation for Japanese microdata.

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.

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.

Duncan, G. T., Keller-McNulty, S. and Stokes, S. L.(2001) "Disclosure Risk vs. Data Utility: the R-U Confidentiality Map" *Technical Report 121*, US National Institute of Statistical Sciences, Durham, North Carolina.

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 Avoidance 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 Avoidance 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).

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

Mateo-Sanz, J. M., Domingo-Ferrer, J. and Sebé, F.(2005) "Probabilistic Information Loss Measures in Confidentiality Protection of Continuous Microdata" *Data Mining and Knowledge Discovery*, vol.11, pp.181-193.

Torra, V. and Domingo-Ferrer, J. (2003) "Record Linkage Methods for Multidatabase Data Mining", Torra, V. (ed.) *Information Fusion in Data Mining, Springer*, Berlin, pp.101-132.

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/

Woo, M., Reiter, J. P., Oganian, A., Karr, A. F.(2009) "Global Measures of Data Utility for Microdata Masked for Disclosure Limitation", *The Journal of Privacy and Confidentiality*, Vol.1, No.1, pp.111-124.