

Challenge Examples: Public Institution

Dr. Kiyomi SHIRAKAWA,
National Statistics Center /
Hitotsubashi University



HITOTSUBASHI
UNIVERSITY

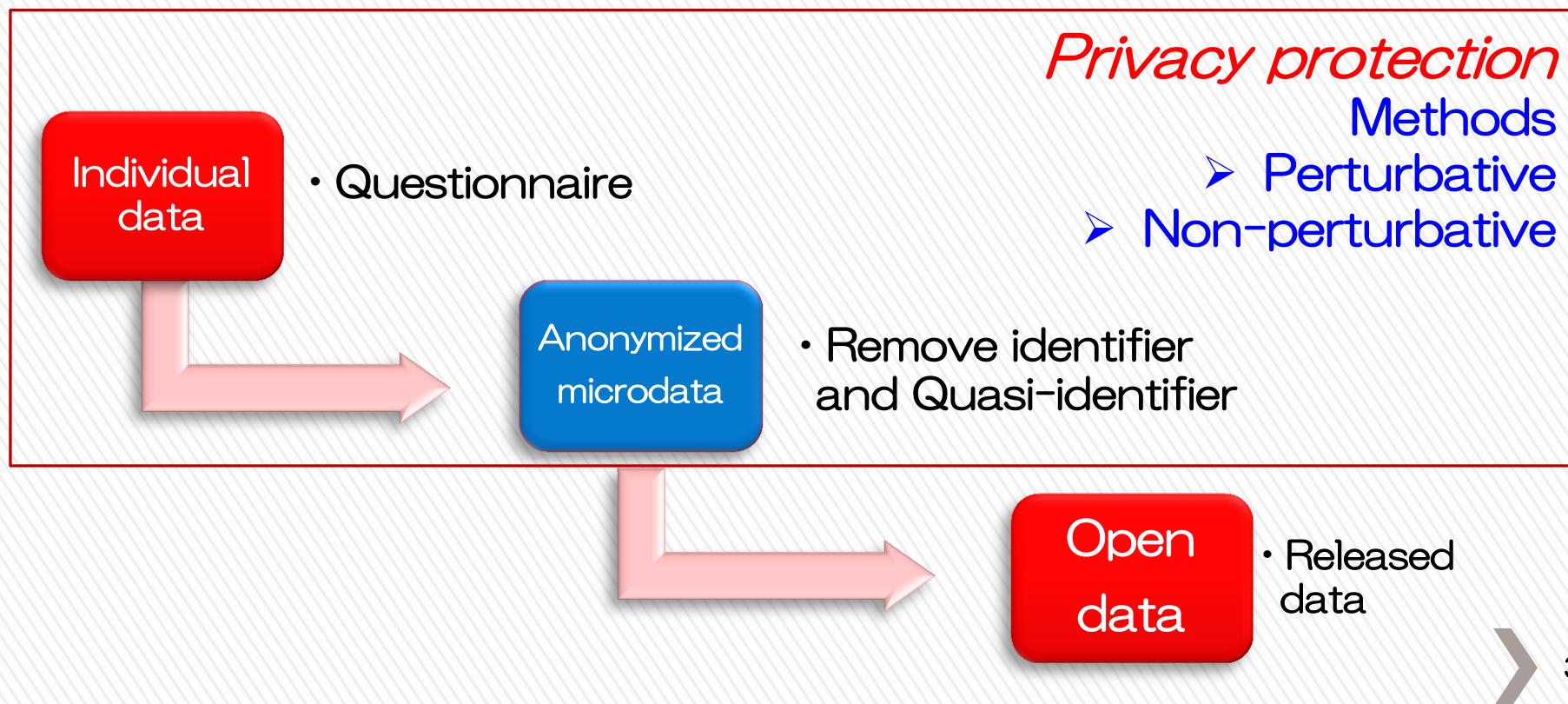


Table of Contents

1. Anonymized microdata
 1. Microdata protection methods
 2. Available anonymized data
 3. Usage performance
2. Anonymized microdata of Population Census
 1. Characteristics of Anonymized Census microdata
 2. Anonymized methods for Population census
 3. Empirical studies
3. Conclusion

1. Anonymized microdata

- » Data containing only anonymized records
- » A record from which direct identifiers have been removed.



1.1 Microdata protection methods



HITOTSUBASHI
UNIVERSITY

» Perturbative methods vs. data types

- Allows for the release of the entire microdata set, although perturbed values rather than exact values are released.

Method	Continuous data	Categorical data
Noise addition	X	
Micro aggregation	X	(X)
Rank swapping	X	X
Rounding	X	
Resampling		
PRAM		X
MASSC		X

“X” denotes applicable and “(X)” denotes applicable with some applicable.



4

1.1 Microdata protection methods



HITOTSUBASHI
UNIVERSITY

» Non-perturbative methods vs. data types

- Does not rely on distortion of the original data but on partial suppressions or reductions of detail

Method	Continuous data	Categorical data
Sampling		X
Global recoding	X	X
Top and bottom coding	X	X
Local suppression		X



5

1.2 Available anonymized data



HITOTSUBASHI
UNIVERSITY

(Last modified: 2014/11/28)

Survey		Distribution data	Distribution schedule in FY2015
MIC	Population Census	2000, 2005	—
	Labour Force Survey	1989 to 2010	2011
	Housing and Land Survey	1993, 1998, 2003	—
	National Survey of Family Income and Expenditure	1989, 1994, 1999, 2004	—
	Employment Status Survey	1992, 1997, 2002	—
	Survey on Time Use and Leisure Activities	Questionnaire A (Time Use/Leisure Activities) 1991, 1996, 2001, 2006	Questionnaire B (Time Use) 2001, 2006

MIC: Ministry of Internal Affairs and Communications



6

<http://www.nstac.go.jp/services/archives.html>



1.3 Usage performance

Survey (Ministry of Internal Affairs and Communications)	Distribution number *(Files)							
	FY2010		FY2011		FY2012		FY2013	
National Survey of Family Income and Expenditure	17	(68)	12	(35)	13	(52)	8	(40)
Employment Status Survey	9	(38)	16	(45)	11	(62)	15	(46)
Survey on Time Use and Leisure Activities	10	(40)	7	(15)	5	(15)	10	(44)
Housing and Land Survey	6	(12)	1	(3)	1	(3)	3	(5)
Labour Force Survey	—	—	0	(0)	0	(0)	2	(24)
Population Census	—	—	—	—	—	—	1	(1)
計	38	(158)	31	(98)	27	(132)	33	(160)

*"Sum of Distribution number" does not match the calculated value.

2. Anonymized microdata of Population Census

Population Census:

The population census in Japan have been conducted almost every five years since 1920, the 2005 Population Census being the eighteenth one.

- The 2005 Population Census covered 17 topics
12 topics for household members such as “Sex”
and “Year and month of birth”
5 topics for households such as “Type of household”
- The 2000 Population Census covered 22 topics
16 topics for household members
6 topics for households

2.1 Characteristics of Anonymized Census microdata



- Regional segments: prefectures and population 500,000 or municipal district
- Amount of data is 1% of the population: Extracted with household unit
- Provide year: 2000 and 2005
- Items: the number of household members, relationship, age, type of housing, etc.

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

2.2 Anonymized methods for Population census



Anonymized methods

- Suppression
- Recoding
- Sampling
- Cell Suppression (1 or 2)
- Data Swapping

Difference in method



HITOTSUBASHI
UNIVERSITY

Methods for Population Census	Methods for other survey
Suppression Recoding Sampling <u>Cell Suppression(1 or 2)</u> <u>Data Swapping</u>	Suppression Recoding Sampling

(1) Suppression

Ratio of data to be deleted is less than 0.5% with respect to population.

※0.5% ... Criteria stipulated in anonymous guidelines

Area A	Household members					
	1 person	...	7 people	8 people	9 people	10 people or more
	25.4		1.8	0.53	0.12	0.04

↓ ↓
Suppression

(2) Recoding



Detection of the segment of less than 0.5%
→ Integration of similar segment

Nationality					
Japanese	Korea	Chinese	...	Brazilian	Peruvian
98.78	0.37	0.27	...	0.17	0.03



Recoding

Nationality	
Japanese	Foreigner
98.78	1.22

(2) Recoding



Detection of the segment of less than 0.5%
→ Integration of segment

Age group					
80 - 84 years	85 - 90	90 - 94	95 - 99	100 - 104	105 - 109
2.2	1.04	0.37	0.08	0.01	0

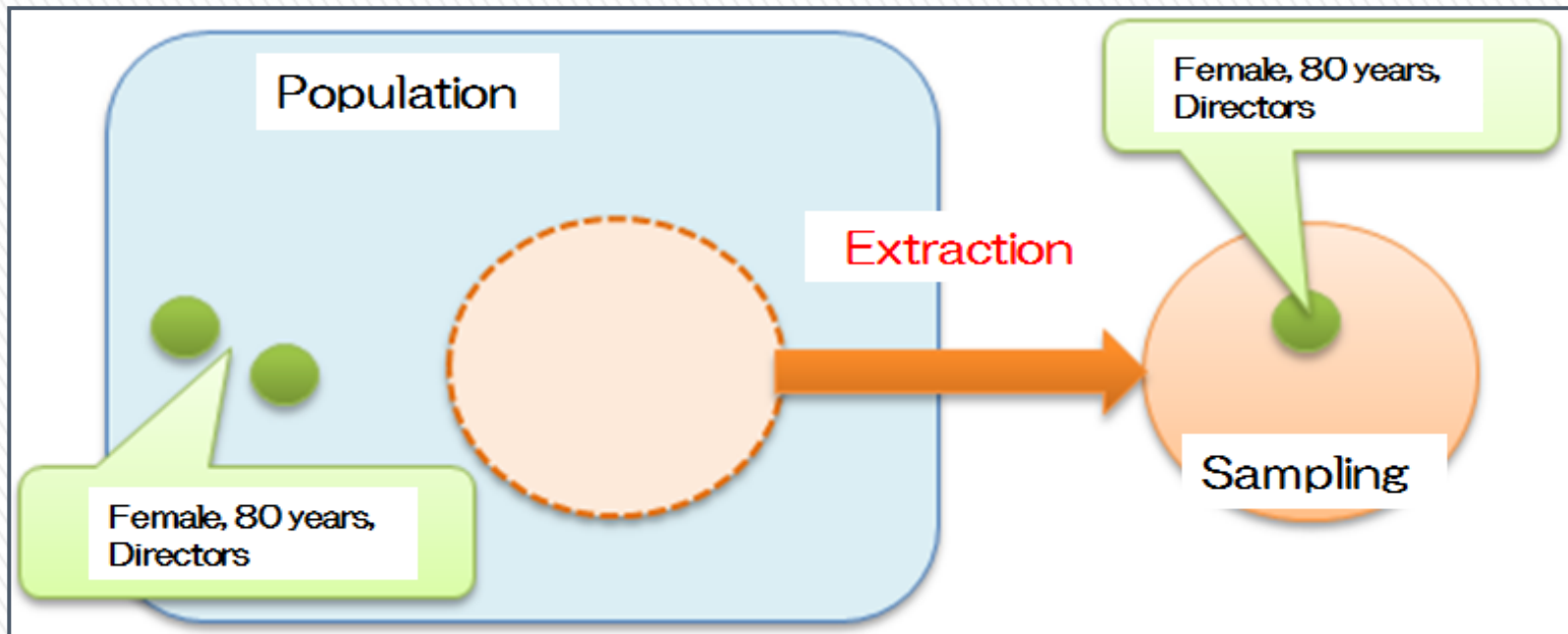
A horizontal bracket is drawn below the table, spanning from the '85 - 90' column to the '105 - 109' column. A vertical line descends from the center of this bracket to the 'Top coding' label below.

Top coding

Age group	
80~84 years	85 years or more
2.2	1.47

(3) Sampling

Extract object data from the population



Non unique population:

The risk in the sample unique is low.

(3) Sampling

Sampling by systematic sampling

1. Determination of the starting point
2. The starting point is the second household
3. The extracted every 100 households

Number of households

001	...		
002	...	→	Extraction
003	...		
004	...		
	:		
100	...		
101	...		
102	...	→	Extraction
	:		

Sampling interval:
100

(4) Cell Suppression



- Published result table is identification information.
 - Remove less than 2 households in the result table

Published Result Table
Case study:
Sex, Nationality

	Male	Female
Japanese	5	4
Foreigner	3	1

Anonymized microdata

Number of households	Sex	Nationality
001	1	1
001	2	1
002	1	1
002	2	2
002	1	1
003	1	1
:	:	:

Unit of remove:
Households

Population unique

(5) Data Swapping



The combination is not in Published result table:
New cell suppression appearance.



The basic idea behind the method:
To transform a database by exchanging values of
confidential variables among individual records.

Records are exchanged in such a way that low-order
frequency counts or marginals are maintained.



Data Swapping image



HITOTSUBASHI
UNIVERSITY

Area A

Area	Number of households	Item 1	Item 2	Item3
A	001	01	1	5
A	001	02	3	2
A	002	01	2	1
A	002	03	2	3
A	003	01	1	4
A	003	02	2	4
A	003	03	2	4
:	:	:	:	:

Area B

Area	Number of households	Item 1	Item 2	Item3
B	001	01	1	2
B	002	01	1	2
B	002	02	2	2
B	002	03	2	2
B	003	01	1	2
B	003	02	2	1
B	004	01	2	1
:	:	:	:	:



Data
Swapping

Regional and item 3 is replaced with other households in other regions.

2.2 Empirical studies



Anonymized methods

- Suppression
- Recoding
- Sampling
- Cell Suppression (1 or 2)
- Data Swapping

(1) Sampling

Since 1% extraction,

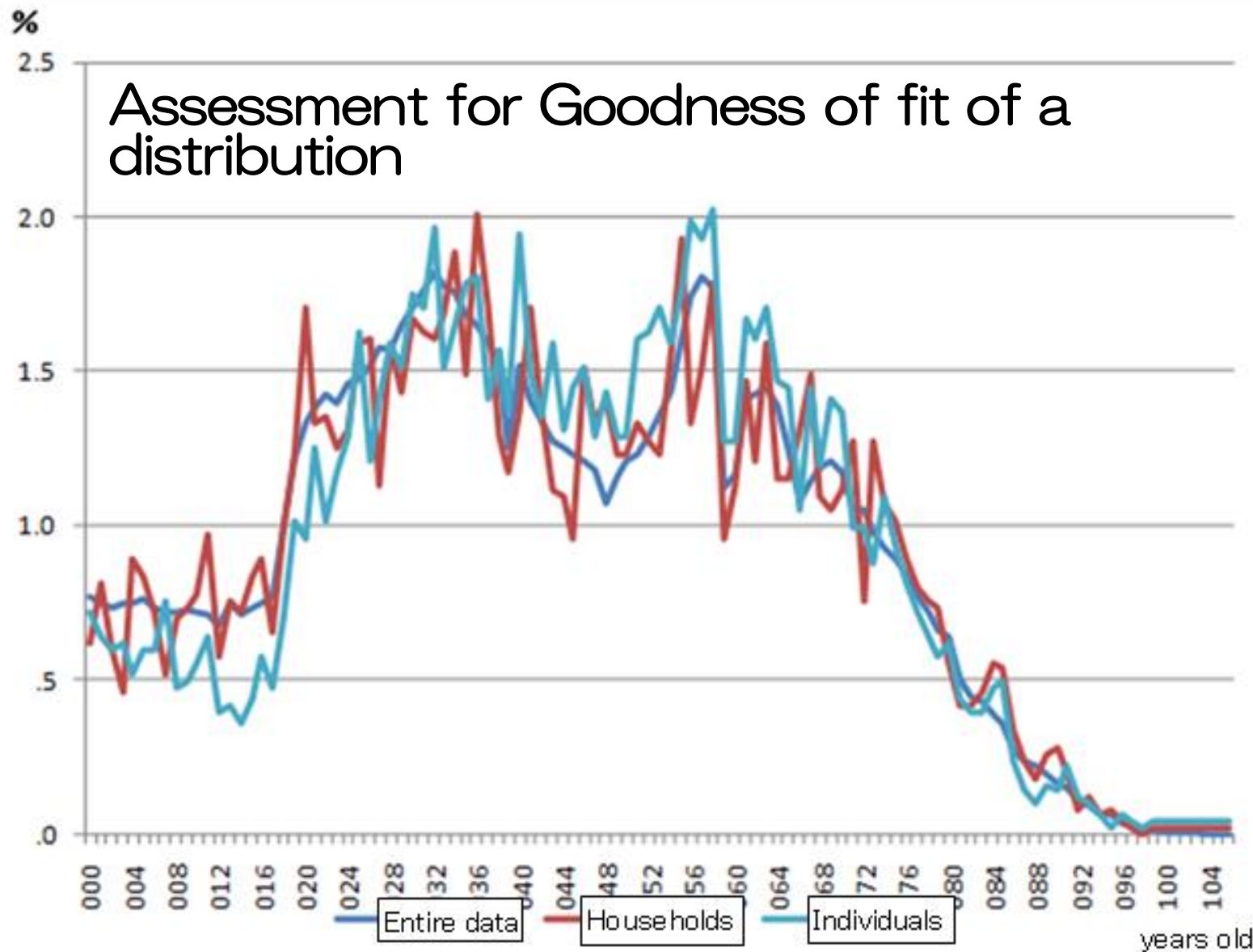
- Is Error increased?
- By extraction with households, has influence on the individual distribution?



- Entire data
- 1 %Sampling (Households)
- 1 %Sampling (Individuals)

Comparison of the
distribution in the entire
data and sampling data

(1) Sampling



(2) Suppression



Detection of the segment of less than 0.5%
→ A large number of remove

The identifier of seven items in high risk
(Sex, Type of household, Type and tenure of dwelling, Type of building and number of stories, Area of floor space of dwelling)



Decide of deletion the criteria by a distribution

(2) Suppression



Case Study:

Age of head of household of male, Age of eldest son

- If the age difference is large or small, population unique many appearance

Cell Suppression	Ratio of population
Remove by 0.5% criteria	13.0%
Optimal removed criteria	0.3%

(3) Cell Suppression

Published result table is identification information.

Matching:

*Matching of Published result table and
anonymized data*



Remove less than 2 households in the result table

(3) Cell Suppression

Create test data based on the identifier of seven items in high risk

(Sex, Type of household, Type and tenure of dwelling, Type of building and number of stories, Area of floor space of dwelling)



Create a cross-table of 7 items

→ Frequency of 2 or less:
Remove household

(3) Cell Suppression

Region analysis:

Cell suppression rate

→ The highest area is 0.147%.

	Ratio in a range
Prefectures	0.001% to 0.666%
Population 500,000 or more municipal district	0.003% to 0.147%

(3) Cell Suppression

Verification in a cross-table of the original data and anonymized data

Case Study: head of household of sex,
householder age 5-year-old class

Female 50-54 years of age

Original data
4,017

Anonymized data
3,999

The difference in the 18 records,
The difference between the ratio is
0.45%

The maximum difference is 0.45%.
→ No problem in usefulness.

(4)Data Swapping



Changing the Data Swapping rate:
How much is the difference between the
original data?



Create some Swapped data

Create a cross table for inspection
→ Verify the difference between the
original data

Different swapping rates



HITOTSUBASHI
UNIVERSITY

This trial is a basic data for determining the Data Swapping rate.

Swapping rate 0.01%

	Male	Female
single	0	-1
married	0	0
bereavement	0	0
divorce	0	1
unknown	0	0

Swapping rate 0.1%

	Male	Female
single	1	2
married	-1	-3
bereavement	0	-1
divorce	1	5
unknown	-1	-3

Swapping rate 0.5%

	Male	Female
single	-2	7
married	2	-3
bereavement	-1	2
divorce	8	-2
unknown	-7	-4

Swapping rate 1%

	Male	Female
single	3	5
married	0	-2
bereavement	7	26
divorce	6	-23
unknown	-16	-6



3. Conclusion



- (1) Study of the method for increasing the confidentiality without deleting the population unique
- (2) Studies on a creating method corresponding to the needs



Examination of confidentiality and usability with a focus on Data Swapping

References

- » Official Statistics-based Services
<http://www.nstac.go.jp/services/archives.html>
- » Technical Research
<http://www.nstac.go.jp/services/reference.html>
- » Privacy in Statistical Databases 2014
<http://unescoprivacychair.urv.cat/psd2014/>
- » Work session on statistical data confidentiality
<http://www.unece.org/index.php?id=31958#/>