Statistical Disclosure Control methods for microdata in the Netherlands

Eric Schulte Nordholt
Senior researcher and project leader of the Census
Statistics Netherlands
Division Socio-economic and Spatial Statistics

e.schultenordholt@cbs.nl

4th Baltic Nordic Conference on Survey Statistics in Helsinki
24-28 August 2015

Contents

- Introduction
- The release of public use microdata files
- The release of microdata for researchers
- On-site access
- Remote access
- Co-operation projects
- Harmonisation
- Microdata availability
- Conclusions

Introduction (1)

Statistical offices have a lot of statistical information

Researchers and policy makers want this information

Privacy!

Statistical Disclosure Control in the Netherlands:

publish and release as much detail as possible without disclosing sensitive information that can be attributed to individual respondents

Introduction (2)

Information from NSIs: tabular data and microdata

Monopoly of NSIs

Eighties of last century:

- **end of monopoly**
- **less microdata available (risk awareness)**

How to end the ‘cold war’ between Statistics Netherlands and the academia?

The release of public use microdata files (1)

For everybody, but severe protection

Rules for public use microdata files:

1. **Microdata must be at least one year old**
2. **No direct identifiers or direct regional variables**
3. **Only 1 kind of indirect regional variables. Values of indirect regional variables sufficiently scattered. Each area should contain at least 200,000 persons in the target population and should consist of municipalities from at least six of the twelve provinces. No dominating municipality in any area.**
4. **At most 15 indirect identifiers**
5. **No sensitive variables**

The release of public use microdata files (2)

Rules for public use microdata files (continued):

6. **Sampling weights should not provide additional identifying information**
7. **Rule against spontaneous recognition: at least 200,000 individuals in the population for each category of an identifying variable**
8. **Another rule against spontaneous recognition: at least 1000 individuals in the population for each category in the crossing of two identifying variables**
9. **At least 5 households per combination of categories of household variables**
10. **Records should be in random order**

Conclusion: public use microdata files are useful for educational purposes and promoting the Census

The release of microdata for researchers (1)

Only for bonafide researchers (under contract)

Rules for microdata for researchers:

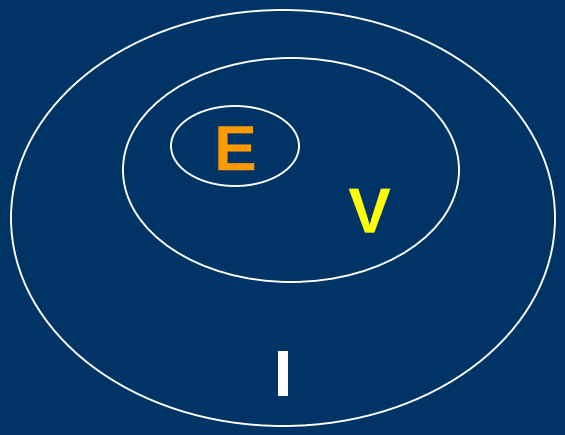
1. No direct identifiers
2. Rule against spontaneous recognition: each combination of an extremely identifying variable, a very identifying variable and an identifying variable should occur at least 100 times in the population
3. Extension of this rule: maximum level of detail of some variables (occupation, level of education, branch of economic activity) is determined by the most detailed direct regional variable
4. Each region that can be distinguished in the microdata should contain at least 10,000 inhabitants
5. No direct regional variables in panel data

The release of microdata for researchers (2)

Identifying variables

- **Direct (formal) identifiers**
  - Name, address, citizen service number, …

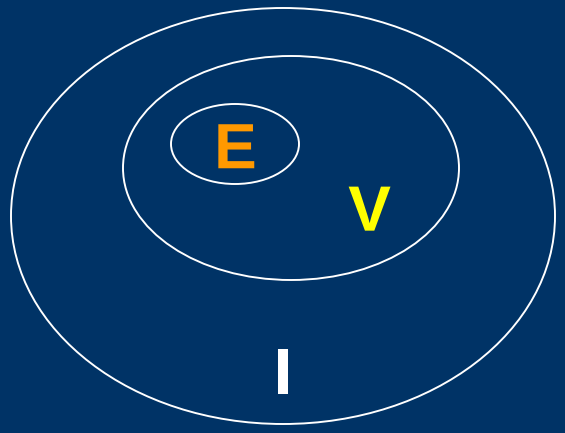
- **Indirect identifiers, differentiated into**
  - **Extremely identifying (E)**
  - **Very identifying (V)**
  - **Identifying (I)**



The release of microdata for researchers (3)

Examples of identifying variables

- **Extremely identifying:**
  - Regional variables (residence, work, …)
- **Very identifying:**
  - Sex, nationality
  - \+ Extremely identifying variables
- **Identifying:**
  - Age, occupation, education
  - \+ Very identifying variables



On-site access (1)

- Researchers work in a secure area of the statistical institute
- Researchers can apply the standard statistical software packages and also bring their own programmes
- Researchers and their superiors have to sign that they will not disclose the individual information of respondents

On-site access (2)

The On-site room

- Standard PCs in a Windows network
- No rights to access the rest of Statistics Netherlands' network
- No internet, printer, CD-ROM or USB
- MS-Office suite, SPSS, SAS
- Own software on request (own license)
- Fee for the use of the data (preparation) and the use of the PCs

Remote access (1)

- Combination of advantages of on desk (no commuting to the NSI) and on-site (large detail in microdata)
- Security risks are high, especially with remote execution (no intermediary between the researcher and the statistical institute)
- Remote access has become very popular in several countries

Both on-site access and remote access require output checking

Remote access (2)

Remote access pilot in 2005 at Statistics Netherlands

Advantages of remote access

- at own institute
- 24/7 availability
- ability to play around with the data, without confidentiality checks until final output
- controlled safe settings

Disadvantages of on-site

- only at premises SN
- only working hours
- no direct contact with colleagues
- special offices needed

Remote access (3)

Remote access (4)

Only authorised users from selected research institutes allowed (under contract)

On-site:

- Users cannot enter or leave Statistics Netherlands unaccompanied

Remote access:

- Biometric identification

- Public Key Infrastructure (PKI) Certificates
- Username and password

Remote access (5)

Detailed microdata stay at Statistics Netherlands

On-site:

- Network separate from production
- No internet
- No printer
- No CD-ROM or USB
- Desired output checked by Statistics Netherlands staff

Remote access:

- Network separate from production
- Citrix connection (on special PCs at the institute of the researcher)
- Desired output checked by Statistics Netherlands staff

Co-operation projects

Special contracts with research institutes

Three different situations:

- Only Statistics Netherlands is publishing output
- Also partners publish output but the protection rules of Statistics Netherlands apply
- Other partners also publish and have their own rules (respecting the national privacy act)

Harmonisation (1)

More information about the Dutch traditional Censuses (including those of 1960 and 1971):

http://www.volkstellingen.nl/en/

For 1960 and 1971 the same variables as for 2001

- if not available: constructed based on existing variables in Census data

Variables not internationally harmonised (e.g. sex, age, marital status, household position, country of birth, economic status, household size and country of citizenship)

- same classification and priority rules as for 2001

Harmonisation (2)

Household size and country of citizenship:
- missing for 1960

Religious denomination (philosophy of life):
- only for 1960 and 1971

Place of residence one year prior to the census:
- only for 2001

International classifications
- Branch of current economic activity: ISIC / NACE
- Occupation: ISCO
- Level of educational attainment: ISCED

Harmonisation (3)	1960	1971	2001
Sex	X	X	X
Age	X	X	X
Country of citizenship		X	X
Marital status	X	X	X
Household position	X	X	X
Religious denomination	X	X	
Country of birth	X	X	X
Household size		X	X
Place of residence one year prior to the census			X
Economic status	X	X	X
Level of educational attainment	X	X	X
Occupation	X	X	X
Branch of current economic activity	X	X	X

Microdata availability

One percent samples for three years (1960, 1971 and 2001)

IPUMS (Integrated Public Use Microdata Series):
http://www.ipums.org/international/index.html

Weighting to population totals

Protecting according to rules for public use files

Microdata sets for all three years available for research!
DANS (Data Archiving and Networked Services):
http://www.dans.knaw.nl/en/

Conclusions

- Both public use microdata files and microdata for researchers can be produced easily with μ-ARGUS
- Lots of protection techniques are available, most popular strategy is to first recode the identifying variables, and then to suppress the remaining unsafe combinations
- Microdata for on-site and remote access may contain all variables except direct identifiers
- Co-operation projects have given Statistics Netherlands a stronger and more relevant position

Thank you for your attention!

Time for questions and discussion