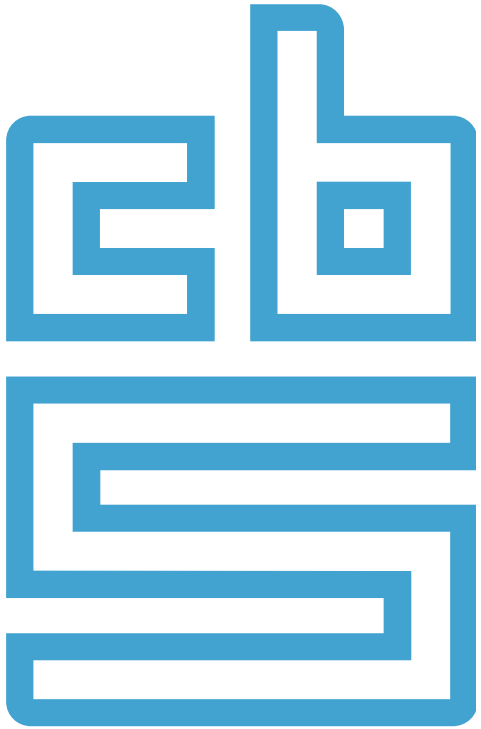


Access to different kinds of Statistics Netherlands' microdata

Presentation at the UNECE Work Session on Statistical Data Confidentiality, December 2021

Eric Schulte Nordholt (e.schultenordholt@cbs.nl)
Statistics Netherlands, Division Socio-economic and spatial statistics

Contents

- Introduction
- The release of public use files
- The release of microdata under contract
- On-site access
- Remote access
- Co-operation projects
- Conclusions
- Discussion

Introduction (1)

Statistical offices have a lot of statistical information

Researchers want to analyse this information more and more at the microdata level

Privacy concerns!

Statistical Disclosure Control in the Netherlands:

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

Introduction (2)

Information from NSIs: tabular data and microdata

Monopoly of NSIs on microdata

Eighties of last century:

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

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

The release of public use files (1)

For everybody, but severe protection

Rules for public use 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 files (2)

Rules for public use 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

The release of public use files (3)

Conclusion

Public use files are useful for

- educational purposes (based on data from the Employee Register, the Labour Force Survey and the Housing Survey) and
- promoting the Population and Housing Census (in particular participating in the IPUMS project)

The release of microdata under contract (1)

Only for bonafide researchers (under contract)

Rules for microdata for researchers:

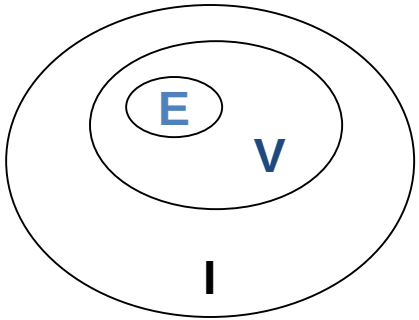
1. No direct identifiers
2. Rule against spontaneous recognition: each combination of an extremely identifying variable, a very identifying variable and an identifying variable should occur at least 100 times in the population
3. Extension of this rule: maximum level of detail of some variables (occupation, 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 under contract (2)

Identifying variables

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

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



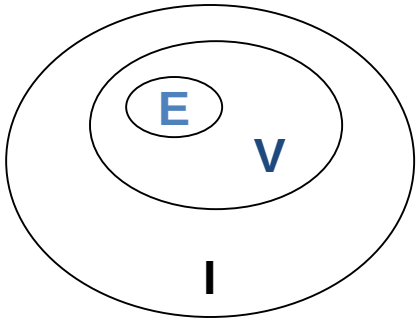
The release of microdata under contract (3)

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

- 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
- Only access to the data needed for the project
- No direct identifiers included in the secure use files

On-site facility temporarily closed due to the Covid crisis

Remote access

- 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 (labour intensive!)

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 European General Data Protection Regulation)

Policy of Statistics Netherlands: “Remote access, unless”

Conclusions

- Legal framework should allow microdata access
- Both public use files and microdata under contract can be produced easily with μ- ARGUS
- Microdata for on-site and remote access may contain all variables except direct identifiers
- Check of output from on-site and remote access analyses is labour intensive
- Co-operation projects have given Statistics Netherlands a stronger and more relevant position
- Building a relation of trust is important (joint responsibility)

Discussion

Are there questions or remarks?