Guidelines for output checking

Results of the ESSNet-SDC project

Jan Mol and Anco Hundepool

Overview

ESSNet SDC

Difficulty of output checking

Two models

- Principles based
- Rule of thumb

Examples

ESSNEt-SDC Results

Task 1. Dissemination of CENEX results
(new handbook/courses)
Task 2. Making SDC-tools better usable by NSIs
Task 3. Output checking
Task 4. Communication: WEB, FAQ
Task 5. Improvement of software for microdata
Task 6. Improvement of software for tabular data
Task 7. Synthetic data
Task 8. Analysis of problems on linked tables

ESSNet SDC

Task 3 Output checking

3.a. Guidelines

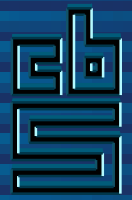
(UK, NL, DE, IT)

- Questionnaire send out
- Many discussions followed
- Final report available

3.b. Automatic output checking

Very difficult, but jacknife (DE)

→Report available, Prototype (R) available, Mainly to safeguard this research

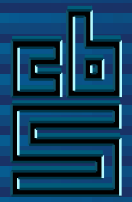


Guidelines group

Special task in the ESSNet-SDC

Partners:

- Maurice Brandt
- Luisa Franconi
- Christopher Guerke
- Anco Hundepool
- Maurizio Lucarelli
- Jan Mol
- Felix Richie
- Giovanni Seri
- Richard Welpton



What makes discosure control in an RDC difficult?

Researchers work with very sensitive datasets, but do very valuable work

Researchers can do a lot of different types of analysis (SPSS, SAS, Stata, own software…)

Infinity number of different outputs

What makes discosure control in an RDC difficult?

Unpredictible outputs

Outputs hard to understand

Need for flexibility

Two types of error to prevent

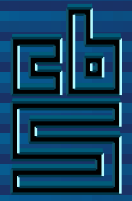
- Confidentiality errors
- Inefficiency errors

Two models

- **Principles based model**
  - Preventing confidentiality AND inefficiency errors
  - Relatively complex to use
- **Rules of the thumb model**
  - Preventing confidentiality errors
  - Relatively easy to use

NOT OPPOSITES BUT EXTENSIONS

Valid for both Onsite/RDC and Remote Access



Principles based model

NSI-staff need to be trained in flexible models

- Nothing is ruled out

Researchers need to be trained as well. Then they will understand us/our problems; in their own benifit

- They should provide us with the information on what they have done

Principles based model

Researchers must make it clear to us what they have done

- Which new variables have been derived
- Which selections/subsets have been made

Classification of outputs into safe/unsafe

- Based upon functional form, not the data

Safe versus unsafe

Safe:

- e.g. regression coefficients
- will normally be released
- NSI takes active decision not to release

Unsafe:

- e.g. tables
- will not be released unless…
- researcher demonstrates to NSI why the output is safe

No unconditional yes/no

Safe versus unsafe

- In the guidelines a whole list of models etc.
- Often behind a model is a freq.-table; could help
- Ideally 2 checkers (subjet matter and RDC)
- Heavy hurden how to control (price of success)
- Remote Access could reduce this: intermediate results stay at RA

Potential problems

Training of NSIs and researchers

- Positive engagement essential and benificial

Devolution of responsabilities

- Responsability lies with the checker and the researchers

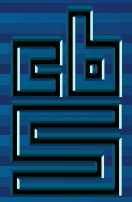
'Rule of the thumb' model

Ignore inefficiency errors

Set of rules with high thresholds

Can be applied
- rather automatically
- by staff with less experience

Not risk-free
- no blind application of rules



Principles of Rule of Thumb

1. > 10 unweighted contributors to a cell
2. > 10 degrees of freedom
3. Group disclosure (90%)
4. Largest contributor < 50% cell total

Use of 'rule of thumb'

Users

- Naive researchers
- Inexperienced NSIs
- Rather automatic SDC mechanisms

Starting point for each output

- Even when using the full principles based model
- Quickly directing attention to the tricky parts

Examples

	Rule of Thumb	Principles based
Regression coefficients	Release	Release
Tables	Release if > 10 unweighted units in each cell	Release if researcher demonstrates safety
Mininma & maxima	Don't release	Release if non-discosive

Other issues

Guidelines on minimum/best practice for management

- Contracts
- Acceptable outputs
- Size of output/costs
- Statistical versus content analysis

Annex on classification of all outputs

- Living document – web based?
- All help is welcome to extend!

Conclusions

- Final ESSNet version of guidelines is avalable at http://neon.vb.cbs.nl/casc
- More work needed. Still very complex
- Automation is still far away
- Starting point for new RDCs and RAs
- Harminisation is a prerequisite for cross-border access and multi-country research

Thanks