5 Safes of secure access to confidential data

Felix Ritchie
Bristol Economic Analysis
University of the West of England

Safe data: ideas and tools

Administrative Data Research Network

UK Data Service

HM Revenue & Customs

Justice

Office for National Statistics

The Health Foundation

Stages of microdata assessment

1. Identify the need for confidentiality protection
2. Analyse data characteristics and use
   - release mechanism
   - user needs (essential/irrelevant variables)
3. Define disclosure scenario and assess risk
4. Identify the most appropriate method(s)
5. Implement

Is there a need for protection?

- Can you conceive of confidentiality risks in
  - Facebook data?
  - The Police National Computer records?
  - Speed camera data?

Data characteristics

- Direct identifiers rarely a problem
  - little research value
- Indirect identifiers more problematic
  - need to understand how variables can be combined

User characteristics

- Why does it matter, how someone may use the protected data?

- Protecting data is not cost-free
  - can we be more cost-effective?
- Protecting data affects the analysis
  - how do you know if you are protecting it sensibly and usefully?
- Microdata protection characteristics reported in terms of univariate characteristics
  - is this relevant?

Risk scenarios

- Stage 1: what **could** go wrong?
- Stage 2: what **is likely** to go wrong?
- Stage 3: (non-data protection measures)
- Stage 4: manage the residual risk in the data

Data protection options

- Broad option 1: reduce detail
  - Recoding
  - Top/bottom coding
  - Rounding
- Broad option 2: change values
  - Microaggregation
  - Swapping values of identifying variables (PRAM)
  - Rank swapping
  - Adding noise
- Local suppression
- Synthetic data

- how will users react to these?

Evaluating alternatives

- Most have been around a long time
  - well-studied
  - advantages and drawbacks understood
- but all struggle to balance protection and data damage
  - confidentiality protection can be quantified
  - ‘damage’ cannot – it depends upon likely use
  - very little knowledge about multivariate consequences

  $\Rightarrow$ not a fair fight

- Remember the introduction:
  - where you start from affects where you end up
- Judgment is important!

Tools

- mu-Argus
  - standalone software package recommended by Eurostat for government statisticians
  - software and manual:
    - http://neon.vb.cbs.nl/casc/mu.htm
- SDCmicro
  - R package used by others
  - software and manual:
    - http://www.inside-r.org/packages/cran/sdcMicro/docs/sdcMicro
  - new documentation being developed UKDA

Evaluation tools: be careful!

- Good for providing comparison between SDC methods (*relative* risk of methods)
  - quick and easy to explore what changes have biggest effect
- More problematic when trying to define *absolute* risk
  - the numbers **have no meaning**

More information

- UK Anonymisation network
  - http://ukanon.net/

- Hundepool A., Domingo-Ferrer J., Franconi L., Giessing S., Lenz R., Naylor J., Schulte Nordholt E., Seri G., de Wolf P-P. (2010) *Handbook on Statistical Disclosure Control*
  - http://www.cros-portal.eu/sites/default/files//SDC_Handbook.pdf

- Hafner H.-P., Ritchie F. and Lenz R. (2015) "User-centred threat identification for anonymized microdata“
  - http://www2.uwe.ac.uk/faculties/BBS/BUS/Research/Economics%20Papers%202015/1503.pdf