Office for National Statistics

From ‘Reproducible’ to ‘Reproduced’

Andrew Engeli

Head of Policy, Secure Research Service, ONS (GBR)

The nature of confidential data

- Difficult to acquire
- Difficult to provision
- Difficult to access
- High risk / disclosive
- ***High value / high impact***

Unique characteristics of confidential data

- Timebound
  - ✓ Survey data rarely change, administrative data often change
  - ✓ Not always flat files (e.g. Census)
  - ✓ Version control critical, but out of hands of researcher
- Geographically bound
  - ✓ Data restricted to legal jurisdictions
- Data architecture challenges
  - ✓ Synthetic datasets not acceptable
  - ✓ Data linkage – probabilistic linking?
- Disclosure risks
  - ✓ Intermediate stage in research process – disclosure control
  - ✓ What methodology? Differential privacy? Perturbation

Reproducible – a necessary condition

Reproducible = the condition of being able to be reproduced

Q: *Is it sufficient that projects are reproducible?*

A: ***No***

- Reproducibility is a necessary condition, but alone it is not sufficient
- <u>The sufficient condition is ‘reproduced’</u>

Reproducible -> Reproduced

- Reproducible -> onus is on the individual researcher/team
- Reproduced -> onus is on the research community

Why should reproducible research be reproduced?

- ✓ Trust
- ✓ Transparency
- ✓ Third party verification

Challenges of reproduction with confidential data

Who are the stakeholders?

- ✓ Journal editors
- ✓ Research organisations
- ✓ Data processors (providers)
- ✓ Data owners
- ✓ End users (of the research)
- ✓ ***Data subjects - the public!***

What are the barriers?

- ✓ Resources
- ✓ Technical skills
- ✓ ***Data access***

A Call to Arms

“Data providers have a practical and ethical duty to ensure that research conducted through their services is reproducible….”

…and reproduced!

The vision for the UK

- Statutory instrument that facilitates reproducibility (Digital Economy Act 2017, Research Code of Practice)
- Research and data infrastructure that supports reproducibility (ADR UK, UKDS and ONS)
- Ability to spread reproducibility across the UK statistical system (ONS Best Practice and Impact division, Government Statistical System, Champion networks)
- Regulatory oversight (Office of Statistics Regulation)

The wider vision

1. Reproduction of research with confidential data will almost always take place in a national context (exception of IO)

2. UK can work with TTPA such as Cascad to leverage established infrastructure

3. Common international standard that encapsulates journey from reproducible -> reproduced (ISO?)