Facilitating access to administrative records with synthetic data

Gillian Raab
Beata Nowok
Dawn Everington
Chris Dibben
Authors and affiliations

Beata Nowok       Chris Dibben      Gillian Raab    Dawn Everington

School of Geosciences, University of Edinburgh
Administrative Data

- Collected by government departments and other organisations
  - registration, transaction and record keeping,
  - delivering a service or for day-to-day operations
  - not research-ready

- Important new resource for social scientists
  - coverage, methodology
  - better understanding of our society
  - better informed government policy
4 Administrative Data Research Centres (ADRCs)
[secure environments, research support, original research, local data negotiations]
- England – led by University of Southampton
- Northern Ireland – led by Queens University Belfast
- Scotland – led by University of Edinburgh
- Wales – led by Swansea University

Administrative Data Service (ADS) – led by UK Data Archive, University of Essex
[network coordination, first point of contact, UK data negotiations]
Safe settings

- On specific locations only, currently
  - England: London, Southampton, Titchfield
  - Northern Ireland: Belfast
  - Scotland: Edinburgh
  - Wales: Swansea, Cardiff

- For more details see adrn.ac.uk
Safe setting in the BioQuarter

Just you and the data!
Drawbacks of safe settings?

- Geography – need to travel
- Restricted work space
- No internet access
- Restrictions on what can be taken out
  - Any written material taken out must be inspected by safe-setting staff
  - Electronic output must pass disclosure rules
  - Safe-setting staff must review all such output
- The administrative data cannot be used for training courses
Synthetic data

What is it?
Data that resembles the original data
No records that correspond to real individuals or other units
But designed to make it give similar analytical results as would be found from the original data (good utility)

History
Originally proposed for disclosure control over 20 years ago
Many theoretical papers from the early 2000’s
Real applications started to appear a few years later
US Bureau of the Census
Others in Canada, New Zealand, Germany

Disclosure risk
Not zero, but evaluations of applications suggest it is low.
Perceived risk may be as important as actual risk
**Observed (input)**

<table>
<thead>
<tr>
<th>Sex</th>
<th>Age</th>
<th>Education</th>
<th>Marital status</th>
<th>Income</th>
<th>Life satisfaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>FEMALE</td>
<td>57</td>
<td>VOCATIONAL/GRAMMAR</td>
<td>MARRIED</td>
<td>800</td>
<td>PLEASED</td>
</tr>
<tr>
<td>MALE</td>
<td>41</td>
<td>SECONDARY</td>
<td>UNMARRIED</td>
<td>1500</td>
<td>MIXED</td>
</tr>
<tr>
<td>FEMALE</td>
<td>18</td>
<td>VOCATIONAL/GRAMMAR</td>
<td>UNMARRIED</td>
<td>NA</td>
<td>PLEASED</td>
</tr>
<tr>
<td>FEMALE</td>
<td>78</td>
<td>PRIMARY/NO EDUCATION</td>
<td>WIDOWED</td>
<td>900</td>
<td>MIXED</td>
</tr>
<tr>
<td>FEMALE</td>
<td>54</td>
<td>VOCATIONAL/GRAMMAR</td>
<td>MARRIED</td>
<td>1500</td>
<td>MOSTLY SATISIFIED</td>
</tr>
<tr>
<td>MALE</td>
<td>20</td>
<td>SECONDARY</td>
<td>UNMARRIED</td>
<td>-8</td>
<td>PLEASED</td>
</tr>
<tr>
<td>FEMALE</td>
<td>39</td>
<td>SECONDARY</td>
<td>MARRIED</td>
<td>2000</td>
<td>MOSTLY SATISIFIED</td>
</tr>
<tr>
<td>MALE</td>
<td>39</td>
<td>SECONDARY</td>
<td>MARRIED</td>
<td>1197</td>
<td>MIXED</td>
</tr>
<tr>
<td>FEMALE</td>
<td>38</td>
<td>VOCATIONAL/GRAMMAR</td>
<td>MARRIED</td>
<td>NA</td>
<td>MOSTLY DISSATISFIED</td>
</tr>
<tr>
<td>FEMALE</td>
<td>73</td>
<td>VOCATIONAL/GRAMMAR</td>
<td>WIDOWED</td>
<td>1700</td>
<td>PLEASED</td>
</tr>
<tr>
<td>FEMALE</td>
<td>54</td>
<td>SECONDARY</td>
<td>WIDOWED</td>
<td>2000</td>
<td>MOSTLY SATISIFIED</td>
</tr>
<tr>
<td>MALE</td>
<td>30</td>
<td>VOCATIONAL/GRAMMAR</td>
<td>UNMARRIED</td>
<td>900</td>
<td>MOSTLY SATISIFIED</td>
</tr>
<tr>
<td>MALE</td>
<td>68</td>
<td>SECONDARY</td>
<td>MARRIED</td>
<td>-8</td>
<td>DELIGHTED</td>
</tr>
<tr>
<td>MALE</td>
<td>61</td>
<td>PRIMARY/NO EDUCATION</td>
<td>MARRIED</td>
<td>-8</td>
<td>MIXED</td>
</tr>
</tbody>
</table>

**Synthetic (output)**

<table>
<thead>
<tr>
<th>Sex</th>
<th>Age</th>
<th>Education</th>
<th>Marital status</th>
<th>Income</th>
<th>Life satisfaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>MALE</td>
<td>81</td>
<td>PRIMARY/NO EDUCATION</td>
<td>MARRIED</td>
<td>2100</td>
<td>PLEASED</td>
</tr>
<tr>
<td>MALE</td>
<td>54</td>
<td>VOCATIONAL/GRAMMAR</td>
<td>MARRIED</td>
<td>1700</td>
<td>PLEASED</td>
</tr>
<tr>
<td>FEMALE</td>
<td>32</td>
<td>VOCATIONAL/GRAMMAR</td>
<td>DIVORCED</td>
<td>870</td>
<td>MIXED</td>
</tr>
<tr>
<td>FEMALE</td>
<td>98</td>
<td>PRIMARY/NO EDUCATION</td>
<td>MARRIED</td>
<td>800</td>
<td>MOSTLY DISSATISFIED</td>
</tr>
<tr>
<td>FEMALE</td>
<td>50</td>
<td>PRIMARY/NO EDUCATION</td>
<td>MARRIED</td>
<td>NA</td>
<td>MOSTLY SATISIFIED</td>
</tr>
<tr>
<td>FEMALE</td>
<td>37</td>
<td>VOCATIONAL/GRAMMAR</td>
<td>MARRIED</td>
<td>158</td>
<td>PLEASED</td>
</tr>
<tr>
<td>MALE</td>
<td>28</td>
<td>VOCATIONAL/GRAMMAR</td>
<td>NA</td>
<td>1500</td>
<td>MOSTLY SATISIFIED</td>
</tr>
<tr>
<td>FEMALE</td>
<td>62</td>
<td>PRIMARY/NO EDUCATION</td>
<td>MARRIED</td>
<td>830</td>
<td>MOSTLY SATISIFIED</td>
</tr>
<tr>
<td>MALE</td>
<td>78</td>
<td>PRIMARY/NO EDUCATION</td>
<td>MARRIED</td>
<td>NA</td>
<td>PLEASED</td>
</tr>
<tr>
<td>FEMALE</td>
<td>29</td>
<td>SECONDARY</td>
<td>MARRIED</td>
<td>580</td>
<td>MOSTLY SATISIFIED</td>
</tr>
<tr>
<td>MALE</td>
<td>59</td>
<td>PRIMARY/NO EDUCATION</td>
<td>MARRIED</td>
<td>1300</td>
<td>MOSTLY SATISIFIED</td>
</tr>
<tr>
<td>MALE</td>
<td>41</td>
<td>SECONDARY</td>
<td>UNMARRIED</td>
<td>1500</td>
<td>MIXED</td>
</tr>
<tr>
<td>MALE</td>
<td>18</td>
<td>SECONDARY</td>
<td>UNMARRIED</td>
<td>-8</td>
<td>PLEASED</td>
</tr>
<tr>
<td>FEMALE</td>
<td>73</td>
<td>PRIMARY/NO EDUCATION</td>
<td>WIDOWED</td>
<td>1350</td>
<td>MOSTLY SATISIFIED</td>
</tr>
</tbody>
</table>

Data that look (structurally) like original data but contain artificial units only.
Synthetic microdata questions

How to create it?
The distribution of the data is modelled and synthetic data generated from the models
Detailed specification of models required
Data needs to look plausible
And needs to reproduce the relationships of interest to researchers

How can it be used?
Originally proposal was to use it in place of real data.
But this is now thought to be a step too far
It can be made available to researchers
  • To understand and explore the data structure
  • To develop code to carry out their analyses
  • Final analyses are carried out on the original data
  • This closes the loop and helps to develop better methodology
To produce data for training courses

How freely can it be made available?
This is determined by the data holder
Data holders are concerned with perceived risk
Thus in most cases even synthetic data are restricted to specific researchers
SYLLS project – from 2013

- To develop tools that can be used by staff of the 3 UK longitudinal Studies with access to the original data to produce synthetic data extracts that can be made available more freely than the original data.
- Researchers can explore the synthetic data on their own computers and develop analysis code
- Teaching data sets are another use
- Originally we worked for the staff at the Scottish Longitudinal Study – and we still do
- But we now have a wider remit within ADRC-S to work with all staff making administrative data available
A software tool for producing synthetic versions of sensitive microdata

R package

synthpop

version 1.4-0

http://cran.r-project.org/package=synthpop
https://github.com/bnowok/synthpop
https://www.jstatsoft.org/article/view/v074i11
Overview of **synthpop** functions

**descriptive models**
- `compare.synds()`
- `summary.synds()`
- `compare.fit.synds()`
- `glm.synds()` `summary.fit.synds()`
- `utility.tab()`
- `utility.gen()`

**data**
- `read.obs()`
- `syn()`
- `sdc()`
- `write.syn()`
The Scottish Longitudinal Study

-One of three UK studies
  - ONS-LS (England and Wales)
  - SLS (Scotland)
  - NILS (Northern Ireland)

-What data are held in the SLS?
  - Censuses data from 1991-2001 2011 linked over time (5% of Scottish population)
  - Linked births, deaths, marriages. Migration records from GP registrations
  - Other administrative data sources e.g. education, health and others

-How can researchers obtain SLS data
  - Apply to do a project
  - Have it approved by the research board
  - Have safe-researcher accreditation
  - Each user gets a customised extract of linked data prepared for them to use with the variables they ask for
  - Visit the SLS safe setting to carry out analyses
Scottish Longitudinal Study (SLS) safe setting
Example SLS project

- Collaborative project with Scottish Government Social Research
- Young people not in education or training (NEETS)
- ADRC-S staff are running a training course next week to teach researchers methods of handling administrative data
- Synthetic data sets have been prepared for the training course using some of the data from this project

Consequences, risk factors, and geography of young people not in education, employment or training (NEET)

This paper summarises key findings from a study into the consequences, risk factors, and geographies of young people not in education, employment or training (NEET) over the past two decades. The study uses the Scottish Longitudinal Study (SLS) which links anonymised individual records from the 1991, 2001, 2011 Censuses and a wide range of data from a variety of sources. Scotland’s censuses are also used in the analysis of the geographies of NEET.

Main findings
Consequences of NEET status
- Young people, who were NEET, remained disadvantaged in their level of educational attainment 10 and 20 years later. More than one in five of NEET young people in 2001 had no qualifications in 2011, compared with only one in twenty five of non-NEETS.
- There is a ‘scarring effect’ on economic activity. In comparison with their non-NEET peers, NEET young people in 2001 were 2.5 times as likely to be unemployed or economically inactive 10 years later.
- The scarring effect is also evident in the occupational positions that NEET young people take up, if they entered employment. For example, NEET young people in 2001 were 2.5 times as likely as their non-NEET peers to work in a low status occupation in 2011, if they found work.
- NEET experiences are associated with a higher risk of poor physical health after 10 and 20 years. The risk for the NEET group was 1.6-2.5 times that for the non-NEET group, varying with different health outcomes.
- NEET experiences are associated with a higher risk of poor mental health after 10 and 20 years. The risk of depression and anxiety prescription for the NEET group is over 50% higher than that for the non-NEET group.
- Young people who were NEET in 1991 and remained economically inactive in 2001 consistently demonstrated significantly poorer outcomes in 2011 than those who were non-NEET in 1991 and economically active in 2001 and
<table>
<thead>
<tr>
<th>flag</th>
<th>synid</th>
<th>startdate</th>
<th>finishdate</th>
<th>incidenttype</th>
<th>intaltprov</th>
<th>noprovdays</th>
</tr>
</thead>
<tbody>
<tr>
<td>FALSE_DATA</td>
<td>4</td>
<td>10/03/2008</td>
<td>10/03/2008</td>
<td>Fighting</td>
<td>No provision</td>
<td>1</td>
</tr>
<tr>
<td>FALSE_DATA</td>
<td>7</td>
<td>19/01/2009</td>
<td>26/01/2009</td>
<td>Physical assault with no weapon against pupil</td>
<td>No provision</td>
<td>16</td>
</tr>
<tr>
<td>FALSE_DATA</td>
<td>7</td>
<td>04/04/2010</td>
<td>NA</td>
<td>Verbal abuse of staff</td>
<td>No provision</td>
<td>2</td>
</tr>
<tr>
<td>FALSE_DATA</td>
<td>7</td>
<td>04/04/2010</td>
<td>NA</td>
<td>Threat of physical violence, no weapon, against staff</td>
<td>No provision</td>
<td>2</td>
</tr>
<tr>
<td>FALSE_DATA</td>
<td>7</td>
<td>04/04/2010</td>
<td>NA</td>
<td>General or persistent disobedience</td>
<td>No provision</td>
<td>2</td>
</tr>
<tr>
<td>FALSE_DATA</td>
<td>17</td>
<td>20/05/2009</td>
<td>21/05/2009</td>
<td>Substance misuse â€“ alcohol</td>
<td>No provision</td>
<td>2</td>
</tr>
<tr>
<td>FALSE_DATA</td>
<td>20</td>
<td>07/03/2009</td>
<td>07/03/2009</td>
<td>Verbal abuse of staff</td>
<td>No provision</td>
<td>2</td>
</tr>
<tr>
<td>FALSE_DATA</td>
<td>20</td>
<td>16/03/2009</td>
<td>20/03/2009</td>
<td>Physical assault with no weapon against pupil</td>
<td>No provision</td>
<td>2</td>
</tr>
<tr>
<td>FALSE_DATA</td>
<td>20</td>
<td>27/04/2009</td>
<td>04/05/2009</td>
<td>Physical assault using improvised weapon against pupil</td>
<td>No provision</td>
<td>13</td>
</tr>
<tr>
<td>FALSE_DATA</td>
<td>20</td>
<td>03/09/2009</td>
<td>08/09/2009</td>
<td>Refusal to attend class</td>
<td>No provision</td>
<td>5</td>
</tr>
<tr>
<td>FALSE_DATA</td>
<td>20</td>
<td>03/09/2009</td>
<td>08/09/2009</td>
<td>General or persistent disobedience</td>
<td>No provision</td>
<td>5</td>
</tr>
<tr>
<td>FALSE_DATA</td>
<td>20</td>
<td>03/09/2009</td>
<td>08/09/2009</td>
<td>Physical assault using weapon against pupil</td>
<td>No provision</td>
<td>5</td>
</tr>
<tr>
<td>FALSE_DATA</td>
<td>22</td>
<td>08/05/2008</td>
<td>12/05/2008</td>
<td>Physical assault with no weapon against pupil</td>
<td>No provision</td>
<td>6</td>
</tr>
<tr>
<td>FALSE_DATA</td>
<td>36</td>
<td>01/11/2007</td>
<td>05/11/2007</td>
<td>General or persistent disobedience</td>
<td>Other</td>
<td>0</td>
</tr>
<tr>
<td>FALSE_DATA</td>
<td>55</td>
<td>28/10/2008</td>
<td>30/10/2008</td>
<td>General or persistent disobedience</td>
<td>No provision</td>
<td>4</td>
</tr>
<tr>
<td>FALSE_DATA</td>
<td>55</td>
<td>09/11/2008</td>
<td>10/11/2008</td>
<td>Verbal abuse of staff</td>
<td>No provision</td>
<td>2</td>
</tr>
<tr>
<td>FALSE_DATA</td>
<td>58</td>
<td>11/02/2010</td>
<td>12/02/2010</td>
<td>Fighting</td>
<td>No provision</td>
<td>2</td>
</tr>
<tr>
<td>FALSE_DATA</td>
<td>65</td>
<td>06/09/2007</td>
<td>11/09/2007</td>
<td>Threat of physical violence using weapon or improvised</td>
<td>No provision</td>
<td>6</td>
</tr>
<tr>
<td>FALSE_DATA</td>
<td>65</td>
<td>19/10/2008</td>
<td>26/10/2008</td>
<td>General or persistent disobedience</td>
<td>No provision</td>
<td>10</td>
</tr>
<tr>
<td>FALSE_DATA</td>
<td>65</td>
<td>29/01/2009</td>
<td>30/01/2009</td>
<td>Threat of physical violence, no weapon, against pupil</td>
<td>No provision</td>
<td>0</td>
</tr>
</tbody>
</table>
Comparing real and synthetic data

Source: Scottish Longitudinal Study
Comparing real and synthetic data

Source: Scottish Longitudinal Study
NEET 2011 by household type 2001

Source: Scottish Longitudinal Study
NEET 2011 by exclusions 2006-2010

Source: Scottish Longitudinal Study
Utility

- These examples worked OK
- Synthetic data are only as good as the models that created them.
- Creating synthetic data, even with synthpop is not easy
- Detailed specification is needed to provide useful and plausible data
Disclosure risk

- Perceived risk may be important. After synthesis, we carry disclosure control.
  - Including:
    - Labelling the data as FALSE DATA
    - Removing any sample uniques that appear in the synthesised data
    - Top and bottom coding
  - Making synthetic data available only to trained and approved researchers
  - Ensuring that training data is only used for the course
Acknowledgements

The help provided by staff of the Longitudinal Studies Centre Scotland is acknowledged. The LSCS is supported by the ESRC/JISC, the Scottish Funding Council, the Scientists Office and the Scottish Government. The authors alone are responsible for the interpretation of the data. Census output is Crown copyright and is reproduced with the permission of the Controller of HMSO and the Queen’s Printer for Scotland.