A Study of the Impact of Synthetic Data Generation Techniques on Data Utility using the 1991 UK Samples of Anonymised Records

Jennifer Taub
Mark Elliot
Joseph Sakshaug

Motivations for this Study

- To provide a grounded measure in real analysis for synthetic data
- To combine objective and subjective measures of utility for synthetic data

Outline for this talk

- Background information on synthetic data
- Background information on data utility
- Description of Methods
- Findings
- Concluding Remarks

Background on Synthetic Data

What is Synthetic Data?

- Synthetic data is a way of protecting data privacy
- Synthetic data creates a brand new dataset based on a model of the original dataset
- Ideally, synthetic data contains none of the original respondents, yet yields valid statistical analyses

History of Synthetic Data

- In 1993 Rubin first introduced synthetic data
- Rubin's (1993) proposal entailed treating all data as if it were missing values and imputing the data from the conditional.
- Little (1993) introduced a method that would only replace the sensitive units referred to as partially synthetic data.

Non-Parametric Ways to Generate Synthetic Data

- CART (Reiter, 2005)
- Random Forests (Cailo and Reiter, 2010)
- Bagging (Drechsler and Reiter, 2011)
- Support Vector Machines (Drechsler, 2010)
- Genetic Algorithms (Chen et al, 2016)

- Drechsler and Reiter (2011)— CART yielded the highest data utility compared to other machine learning methods.
- Nowok (2015)—CART performed better than other tree-learning methods, but not as well as the parametric synthetic data.

Background on Data Utility

Previous Utility Tests

Drechsler and Reiter (2009) classify two existing types of utility measures for synthetic data:

- Comparisons of broad differences between the original and released data (**Broad Measures**)
- Comparisons of differences in specific models between original and released data (**Narrow Measures**)

Broad Measures

Broad measures quantify some kind of statistical distance between the original and released data:

- Kullback-Leibler
- Hellinger distance.
- Propensity Score Measure mean-squared error (pMSE) (Woo et al, 2009)
  - Adapted by Snoke et al (2016)to fit synthetic data
    - the pMSE ratio
    - the standardized pMSE

Narrow Measures

- Narrow measures tend to be tailored to a certain kind of analysis and may have low utility for other kinds of analyses. (Karr et al, 2006)
- Traditionally, in synthetic data, narrow measures have been used to measure data utility. CIO used in:
  - Drechsler et al (2008)
  - Drechsler and Reiter (2009)

Purdam and Elliot (2007)

- Tested the data utility of local suppression and PRAM , using the 1991 SAR
- They went through 10 published papers and re-ran the analyses using the SDC perturbed data.
- They found that these SDC measures had a significant impact on the usability of the data and the accuracy of the analyses.

Purdam-Elliot Methodology

- **Systematically sampled narrow measure** vs **traditional ad hoc narrow measure**.
- We re-ran the same analyses from the same papers used by Purdam and Elliot (2007).
  - The papers that they selected show-cased a variety of different analytic techniques
  - Directly comparable to Purdam and Elliot's allowing comparison between CART synthetic data, suppression and PRAM

Description of Replicated Papers

Paper	Description
Ballard (1996)	Bar charts exploring different demographics of Pakistanis in the UK
Champion (1996)	Frequency tables looking at migration in the UK
Drinkwater and O'Leary (1997)	Probit regression of Welsh citizens on whether employed with focus on Welsh fluency
Eade et al (1996)	Cross-tab ethnicity by social class and gender
Gardiner and Hill (1996)	Logistic regression of car ownership
Gardiner and Hill (1997)	Frequency tables exploring cycling by area and other variables
Gould and Jones (2000)	Multilevel logistic model for long-term limiting illness, level 2: area, by age, gender, ethnicity, and interaction terms
Green (1997)	Frequency tables looking at participation in growing industries and occupations
Leventhal (1994)	Bar chart comparing prevalence of target group by different demographics

Methods

- We used two datasets:
  - The original 1991 SAR without any perturbation as a control
  - A CART generated synthetic version of the 1991 SAR, which was generated using *synthpop* (Nowok et al, 2016).

Confidence Interval Overlap (Drechsler et al , 2008)

$$J_k = \frac{1}{2}\left(\frac{U_{,k} - L_{,k}}{U_{org,k} - L_{org,k}} + \frac{U_{,k} - L_{,k}}{U_{syn,k} - L_{syn,k}}\right)$$

where *U,k* and *L,k*, denote the upper and lower bound of the intersection of the confidence intervals from both the original and synthetic data for estimate *k*, and *Uorg,k* and *Lorg,k* represent the upper and lower bound of the original data, and *Usyn,k* and *Lsyn,k* the synthetic data.

Ratio of Counts

If $y^1_{orig} > y^1_{synth}$ then:

$$\frac{y^1_{synth}}{y^1_{orig}}$$

If $y^1_{orig} < y^1_{synth}$ then:

$$\frac{y^1_{orig}}{y^1_{synth}}$$

If $y^1_{orig} = y^1_{synth}$ then:

The ratio of counts is 1

Key for paper effect severity (From Purdam and Elliot, 2007, p. 1109)

Severity	Description
Severe	The results were sufficiently different from the original, leading the analyst to come to a different conclusion
Moderate	Change in emphasis rather than a completely different finding
No effect	Indicates that the figures may have been slightly different but the overall pattern was not, indicating the same conclusion could be consistently drawn

Findings

Description of the Effect of the Synthetic data

Paper	CIO	Ratio of Counts	Severity Rating
Ballard (1996)	0.568	0.675	No effect
Champion (1996)	0.531	0.836	No effect
Drinkwater and O'Leary (1997)	0.475	0.543	Severe
Eade et al (1996)	0.293	0.799	Moderate
Gardiner and Hill (1996)	0.414	0.632	Severe
Gardiner and Hill (1997)	0.546	0.561	Moderate
Gould and Jones (2000)	0.583	0.626	Moderate
Green (1997)	0.370	0.768	No effect
Leventhal (1994)	N/A	0.840	No effect

Summary of the Effect of Perturbations

(Information on PRAM and Suppressions from Purdam and Elliot, 2007, p. 1110)

	No Effect	Moderate	Severe
CART	4	3	2
Suppressions	5	5	0
Post-randomisation	2	7	1
Both Suppressions and PRAM	1	5	4

Relationship between the Severity Rating and Utility Metrics

- ANOVA
  - Ratio of Counts:
    $F(2,6) = 3.167$, $p = 0.115$
  - Confidence Interval Overlap:
    $F(2,5) = 0.083$, $p = 0.921$
- Pearson's Correlation
  - Between RoC and CIO:
    $r = -0.423$

- Problems with sample sizes concerning CIO
- Frequency tables and regression models two different trends

Which analyses are CART synthetic data good/bad for?

- Bad Statistics
  - Ex. Gardiner and Hill- 2 black cyclists
- Small Subsets
- Frequency tables vs regression models
- Model Complexity

Concluding Remarks

- Mixed set of results
- When compared to local suppression and PRAM, CART synthetic data performed better than both combined, and was on par with PRAM.
- Nowok (2016) similar utility tests using synthpop CART synthetic data – logistic regression CIO 0.50 vs Gardiner and Hill (1996) logistic regression CIO 0.41
- Which analyses are CART synthetic data good for?
- Advantages to the Purdam- Elliot methodology
  - Systematically sampled narrow measure
  - Data used in unexpected ways

Further Research

- Failure to cross- validate different types of utility measures
- Further Research
  - Using the Purdam-Elliot methodology on other types of synthetic data
  - Will further information provide a better basis for cross-validating utility measures?