Synthetic Data Challenge

Authors

- Jennifer Taub (UoM)
- Mark Elliot(UoM)
- Gillian Raab (Edinburgh)
- Anne-Sophie Charest (Laval)
- Cong Chen (Public Health England)
- Christine M. O'Keefe (CSIRO)
- Michelle Pistner Nixon(Penn State)
- Joshua Snoke (RAND)
- Aleksandra Slavkovi'c (Penn State)

Motivations for this Study

As an outcome of the Isaac Newton Institute programme on Data Linkage and Anonymisation in 2016, a group formed to run a challenge amongst themselves:

- To test different methods of synthetic data generation against one another
- To test measures for data utility and disclosure risk

Outline of this Talk

- Background information on synthetic data
- Description of Measures used for data utility and disclosure risk
- Findings
- Concluding Remarks

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.

Project Design and Methods

The Dataset

The Scottish historical census of 1901- a sample of 82,851 households in the Edinburgh region. The dataset contains 24 variables: 20 observed, 3 derived, and a unique identifier

Synthetic Datasets

Synthetic Dataset	Method of Synthesis
Raab	CART- *synthpop*
Snoke et al	CART-*synthpop*
Pistner et al	Quantile Regression
Charest	Random Sampling
Chen 1	Simulacrum process- *matlab*
Chen 2	Simulacrum process- *matlab*
Chen 3	Simulacrum process- *matlab*
Chen 4	Simulacrum process- *matlab*

Measuring Data Utility- Narrow Measures

- Frequency Tables and Cross-Tabulations
  - Ratio of Counts (ROC)

$$\frac{\min(y^1_{orig}, y^1_{synth})}{\max(y^1_{orig}, y^1_{synth})}$$

  - Confidence Interval Overlap (CIO)

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

- Regression Models –OLS and logistic regression compared using CIO

Data Utility- Broad Measures

- Multiple Correspondence Analysis (MCA)- The two maps (synthetic and orginal) were compared using Euclidean Distance

$$d = \sqrt{(x_o - x_s)^2 + (y_o - y_s)^2}$$

- Propensity Score- the original and synthetic datasets are combined together into a logistic regression model wherein the following equation calculates the likelihood of the synthetic records being identified.

$$pMSE = \frac{1}{N}\sum_{i=1}^{N}[\hat{p}_i - c]^2$$

Measuring Disclosure Risk using Targeted Correct Attribution Probability (TCAP)

- Intruder scenario
  - The intruder has some information K_i on an individual that they know to be on the original dataset. They want to learn the value of some variable T_i.
  - They have access to the synthetic dataset.
  - They identify all the records in synthetic dataset that match K_i. If the proportion of records with the largest equivalence class on $T_s|K_s$ meets some threshold then they infer that value for T_i. If not then they give up.
  - TCAP captures the proportion of records for the key K that have the same target value on it’s original equivalent.

Measuring Disclosure Risk using Target Correct Attribution Probability (TCAP)

- We use the following Key combinations:
  - Key 3 = sex, marital status, parish
  - Key 4 = Key3 + age group
  - Key 5 = Key4 + presence child
  - Key 6 = Key5 + country of birth,

Results

Utility Results

	Raab	Snoke et al	Pistner et al	Charest	Chen 1	Chen 2	Chen 3	Chen 4
Freq. Tabs	0.863	0.884	0.528	0.913	0.899	0.882	0.931	0.892
Cross-tabs	0.78	0.621	0.347	0.758	0.799	0.676	0.748	0.73
CIO of Means	0.88	0.505	0.145	0.676	0.632	0.575	0.706	0.62
Reg models	0.480	0.429	0.0303	0	0.180	0.221	0.2467	0.255
1-4* PMSE	0.999	0.879	0.831	0.999	0.863	0.878	0.879	0.879
MCA	0.816	0.407	0.5425	0.197	0.611	0.4685	0.4155	0.856
Mean	0.803	0.621	0.404	0.591	0.664	0.617	0.654	0.705

Disclosure Risk

Target	Key	Raab	Snoke et al	Pistner et al	Charest	Chen 1	Chen 2	Chen 3	Chen 4
Employ.	6	0.734	0.747	0.667	0.673	0.727	0.701	0.745	0.738
	5	0.728	0.750	0.674	0.668	0.717	0.701	0.709	0.704
	4	0.731	0.724	0.648	0.635	0.734	0.727	0.681	0.729
	3	0.782	0.775	0.554	0.597	0.550	0.847	0.667	0.791
Occup.	6	0.196	0.246	0.121	0.127	0.214	0.233	0.260	0.276
	5	0.207	0.256	0.154	0.117	0.174	0.254	0.269	0.265
	4	0.196	0.266	0.153	0.088	0.195	0.277	0.284	0.238
	3	0.038	0.400	0.074	0.179	0.250	0.519	0.321	0.372
HH Size	6	0.284	0.257	0.173	0.214	0.233	0.247	0.256	0.228
	5	0.278	0.276	0.143	0.251	0.218	0.226	0.273	0.251
	4	0.272	0.231	0.073	0.161	0.176	0.168	0.219	0.188
	3	0.3	0.186	0	0.091	0.200	0.077	0.234	0.175
Mean		0.396	0.426	0.286	0.317	0.366	0.415	0.410	0.413

Four ways of calculating Risk-Utility Score

- Subtract the risk score from the Utility
- Take the minimum of the Utility score and the Inverse of the Risk score
- Multiply together the utility score and the inverse risk.
- Take the geometric mean of the utility score and the inverse risk score.

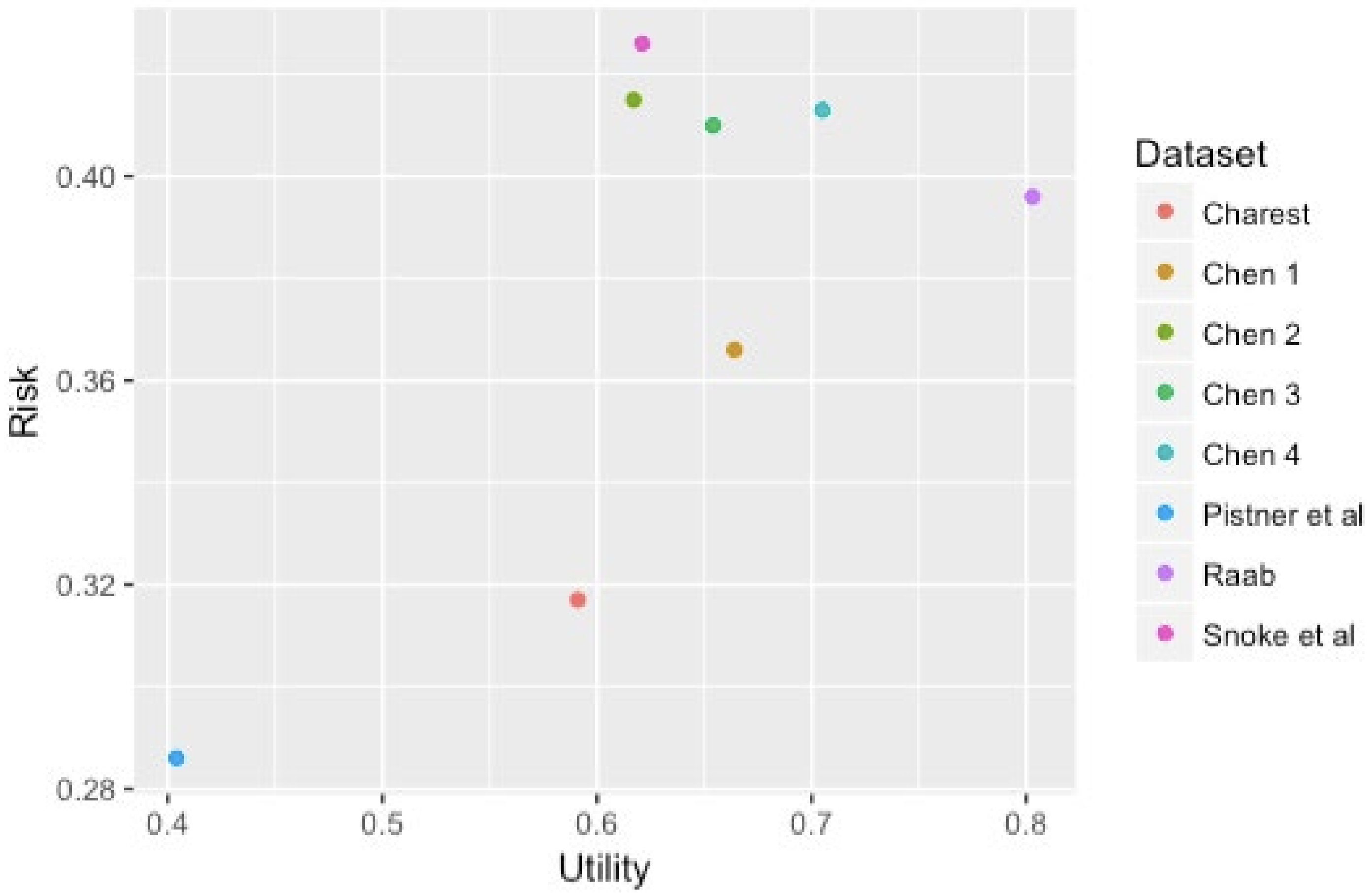
Total Risk Utility Score

	Raab	Snoke et al	Pistner et al	Charest	Chen 1	Chen 2	Chen 3	Chen 4
Utility score	0.803	0.621	0.404	0.591	0.664	0.617	0.654	0.705
Risk score	0.396	0.426	0.286	0.317	0.366	0.415	0.41	0.413
Utility - Risk	0.407	0.195	0.118	0.274	0.298	0.202	0.244	0.292
Min(Utility, Inverse Risk)	0.604	0.574	0.404	0.591	0.634	0.585	0.59	0.587
Utility * Inverse Risk	0.485	0.356	0.288	0.402	0.421	0.361	0.386	0.414
Geometric Mean (Utility, Inverse Risk)	0.696	0.597	0.537	0.634	0.649	0.601	0.621	0.643

Difficulties Calculating the Risk-Utility Score

- While both risk and utility were both scaled from 0 to 1, they are not measuring the same thing and therefore are not on the same scale and not directly comparable.
- alternative determine what is a sufficient utility score or a sufficient risk score and then merely optimising the other score (as suggested by Duncan and Stokes, 2004).
  - However, what is an acceptable risk or utility score is to a certain extent a matter of judgement about the types of analyses that will be performed on the dataset and on the sensitivity of a given dataset.

Risk Utility Map



Reasons for Different Risk-Utility Scores

- **Pre-processing** - Raab and Snoke both used CART, however they used different stratification of variables.
- **Synthesizing Order** - Raab and Snoke same method but different synthesizing order. While Pistener and Snoke had similar synthesizing order but different methods.
- **Random Variance** - All 4 Chen datasets used the same method, but produced different risk utility profiles
- **Synthesis method** - As previously studied different synthesizers produce different risk utility profiles. For examples Pistener's quantile regression lowered the risk scores as promised while sacrificing utility, while CART datasets had high utility but high risk.

Discussion and Conclusion

- This paper has presented a case study comparison of different synthetic datasets.
  - More trials would be need to draw conclusive results on the best way to generate synthetic data.
    - Other methods
    - Other data
    - Honing the utility battery
- However, this type of comparative approach will allow us to focus on the most effectively methods.