GENERATIVE ADVERSARIAL NETWORKS FOR SYNTHETIC DATA GENERATION: A COMPARATIVE STUDY

Claire Little, Mark Elliot, Richard Allmendinger, Sahel Shariati Samani
Centre for Digital Trust and Society
University of Manchester
INTRODUCTION

- Generative Adversarial Networks (GANs) (Goodfellow et al. 2014) are gaining increasing attention as a means of synthesising data.
- GANs have so far been used predominantly for image generation.
- Less research into structured microdata synthesis.
  - e.g. synthesising census or social survey data.
- We compare two GANs with two statistical methods:
  - generate synthetic census data
  - perform analysis using disclosure risk and utility metrics.
GENERATIVE ADVERSARIAL NETWORKS (GANS)

- Composed of two neural networks
  - Generator, $G$
  - Discriminator, $D$
- Discriminator aims to determine whether a sample of data is from the real distribution or whether it was generated by $G$
- Generator creates data samples in order to fool the discriminator
  - Generator never sees the original data and learns only from error
- Performance improves over time
- Success if the discriminator cannot determine fake from real data

Example of GAN architecture
STUDY DESIGN

- Census data
  - 1991 Individual Sample of Anonymised Records (SAR) for the British Census (ONS 2013), a 2% sample (1,116,181 records) including adults and children
  - We subsetted one geographical region (n=104,267, 9.34% of total)
  - Twelve variables used (11 categorical, 1 numerical)

<table>
<thead>
<tr>
<th>Area</th>
<th>Age</th>
<th>Sex</th>
<th>Marital Status</th>
<th>Economic group</th>
<th>Ethnic group</th>
<th>Birth country</th>
<th>Tenure</th>
<th>Social class</th>
<th>Long term ill</th>
<th>Num quals</th>
<th>Family type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Birmingham</td>
<td>28</td>
<td>F</td>
<td>Single</td>
<td>Employed ft</td>
<td>White</td>
<td>England</td>
<td>Rent LA</td>
<td>Skilled</td>
<td>No</td>
<td>one</td>
<td>Lone no dep. child</td>
</tr>
<tr>
<td>Walsall</td>
<td>10</td>
<td>M</td>
<td>Single</td>
<td>NA</td>
<td>Indian</td>
<td>England</td>
<td>Rent private</td>
<td>NA</td>
<td>No</td>
<td>none</td>
<td>Married dep. child</td>
</tr>
<tr>
<td>Dudley</td>
<td>78</td>
<td>M</td>
<td>Married</td>
<td>Retired</td>
<td>White</td>
<td>Scotland</td>
<td>Own outright</td>
<td>NA</td>
<td>Yes</td>
<td>none</td>
<td>Married no dep child</td>
</tr>
</tbody>
</table>
STUDY DESIGN

• Synthesis Methods
  • Statistical
    • Synthpop (Nowok et al. 2016) – CART based
    • DataSynthesizer (Ping et al. 2017, Zhang et al., 2017) – uses Bayesian networks
  • GAN
    • CTGAN (Xu et al. 2019)
    • TableGAN (Park et al. 2018)

All methods used default parameters and generated synthetic data the same size as original dataset (n=104,267)
STUDY DESIGN

- **Metrics**
  - **Disclosure risk**
    - Measured using the Targeted Correct Attribution Probability (TCAP) (Taub & Elliot, 2019)
    - Provides a score between 0 and 1
      - Higher value implies higher risk
  - **Utility**
    - Propensity mean squared error (pMSE) (Snoke et al. 2018, Woo et al. 2009)
    - Confidence interval overlap (CIO)
    - Ratio of estimates (ROE)
- **Risk-Utility comparison**
  - R-U confidentiality map (developed by Duncan et al. 2004)
    - plots overall utility score against TCAP (risk) score
  - Ideally disclosure risk is minimised and utility is maximised
STUDY DESIGN

• Metrics
  • Disclosure risk
    • Measured using the Targeted Correct Attribution Probability (TCAP) (Taub & Elliot, 2019)
    • Provides a score between 0 and 1
      • Higher value implies higher risk
  • Utility
    • Propensity mean squared error (pMSE) (Snoke et al. 2018, Woo et al. 2009)
    • Confidence interval overlap (CIO)
    • Ratio of estimates (ROE)

• Risk-Utility comparison
  • R-U confidentiality map (developed by Duncan et al. 2004)
    • plots overall utility score against TCAP (risk) score
    • Ideally disclosure risk is minimised and utility is maximised
STUDY DESIGN

• Metrics
  • Disclosure risk
    • Measured using the Targeted Correct Attribution Probability (TCAP) (Taub & Elliot, 2019)
    • Provides a score between 0 and 1
      • Higher value implies higher risk
  • Utility
    • Propensity mean squared error (pMSE) (Snoke et al. 2018, Woo et al. 2009)
    • Confidence interval overlap (CIO)
    • Ratio of estimates (ROE)

• Risk-Utility comparison
  • R-U confidentiality map (developed by Duncan et al. 2004)
    • plots overall utility score against TCAP (risk) score
  • Ideally disclosure risk is minimised and utility is maximised
RESULTS

Histograms comparing original data with synthetic data for age

Synthpop closely matched the age distribution whilst TableGAN struggled
Data produced by Synthpop and DataSynthesizer had similar counts to the original data. TableGAN did not manage to identify all categories.
# RESULTS

The basket of utility metrics

<table>
<thead>
<tr>
<th>Metric</th>
<th>Synthpop</th>
<th>DataSynthesizer</th>
<th>CTGAN</th>
<th>TableGAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>pMSE</td>
<td><strong>0.00015</strong></td>
<td>0.01438</td>
<td>0.03162</td>
<td>0.17529</td>
</tr>
<tr>
<td>$1 - (4 \times \text{pMSE})$</td>
<td><strong>0.9994</strong></td>
<td>0.9425</td>
<td>0.8735</td>
<td>0.2988</td>
</tr>
<tr>
<td>ROE univariate (mean)</td>
<td><strong>0.981</strong></td>
<td>0.821</td>
<td>0.743</td>
<td>0.499</td>
</tr>
<tr>
<td>ROE bivariate (mean)</td>
<td><strong>0.847</strong></td>
<td>0.616</td>
<td>0.587</td>
<td>0.255</td>
</tr>
<tr>
<td>CI Overlap (mean)</td>
<td><strong>0.506</strong></td>
<td>0.365</td>
<td>0.410</td>
<td>-</td>
</tr>
<tr>
<td>Overall utility</td>
<td><strong>0.833</strong></td>
<td>0.686</td>
<td>0.653</td>
<td>0.351</td>
</tr>
</tbody>
</table>

Synthpop had optimal results for all metrics.
## Results

TCAP scores for the synthetic methods, four key sizes

<table>
<thead>
<tr>
<th>Target</th>
<th>Key</th>
<th>Synthpop</th>
<th>DataSynthesizer</th>
<th>CTGAN</th>
<th>TableGAN</th>
<th>Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>LTILL</td>
<td>6</td>
<td>0.935</td>
<td>0.929</td>
<td>0.912</td>
<td>0.911</td>
<td>0.774</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.897</td>
<td>0.898</td>
<td>0.891</td>
<td>0.907</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.894</td>
<td>0.899</td>
<td>0.889</td>
<td>0.907</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.936</td>
<td>0.951</td>
<td>0.931</td>
<td>0.901</td>
<td></td>
</tr>
<tr>
<td>FAMTYPE</td>
<td>6</td>
<td>0.709</td>
<td>0.623</td>
<td>0.598</td>
<td>0.301</td>
<td>0.223</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.725</td>
<td>0.658</td>
<td>0.639</td>
<td>0.384</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.736</td>
<td>0.654</td>
<td>0.651</td>
<td>0.416</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.809</td>
<td>0.608</td>
<td>0.648</td>
<td>0.420</td>
<td></td>
</tr>
<tr>
<td>TENURE</td>
<td>6</td>
<td>0.596</td>
<td>0.429</td>
<td>0.490</td>
<td>0.217</td>
<td>0.329</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.504</td>
<td>0.376</td>
<td>0.453</td>
<td>0.336</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.500</td>
<td>0.350</td>
<td>0.447</td>
<td>0.341</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.496</td>
<td>0.353</td>
<td>0.482</td>
<td>0.279</td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>0.728</td>
<td>0.644</td>
<td>0.669</td>
<td>0.527</td>
<td>0.442</td>
</tr>
</tbody>
</table>

Synthpop had highest disclosure risk, TableGAN had the lowest.
### Results

**RU Confidentiality map and table of results**

<table>
<thead>
<tr>
<th>Method</th>
<th>Utility (overall)</th>
<th>Risk (TCAP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synthpop</td>
<td>0.833</td>
<td>0.728</td>
</tr>
<tr>
<td>DataSynthesizer</td>
<td>0.686</td>
<td>0.644</td>
</tr>
<tr>
<td>CTGAN</td>
<td>0.653</td>
<td>0.669</td>
</tr>
<tr>
<td>TableGAN</td>
<td>0.351</td>
<td>0.527</td>
</tr>
</tbody>
</table>

Risk-Utility relationship appears to approximately follow a straight line – excluding the original data.
CONCLUSIONS

- Trade-off between utility and disclosure risk appears to fall on a relatively straight line
- Synthpop showed both highest utility and disclosure risk
- TableGAN had lowest disclosure risk but with unacceptably low data quality

- Methods only tested on a single dataset
- Methods only tested on a subset of records
- Bucket of analyses for the utility tests needs expanding
- Default parameters used for each method
FUTURE WORK

- Much wider range of tests examining effects of parameter changes on the RU map
- Investigating other GAN architectures
- Investigating whether any method can effectively optimise both risk and utility
- Testing on larger datasets (number of variables and cases) and determining scalability of methods
THANK YOU FOR LISTENING!