The use of saturated count models for synthesis of large confidential administrative databases

James Jackson1, Robin Mitra2, Brian Francis1 and Iain Dove3
1Department of Mathematics and Statistics, Lancaster University
2Department of Statistical Science, UCL 3Office for National Statistics

The General Data Protection Regulation (GDPR), implemented by the UK in May 2018, requires that businesses and organisations adhere to certain standards when processing personal data, that is, information that relates to identifiable individuals. This is coupled by a drive - by the UK Government, for example - to fully utilize administrative data. Such databases, collected by organisations such as government departments for administrative purposes, can sometimes hold records for an entire population, so are useful from a statistical perspective. My thesis explored the use of synthetic data methods to balance this dichotomy between protecting privacy and maximizing data utility.

Synthetic data sets are generated by fitting and simulating from a model (synthesis model) fit to the original data. The idea is that they can preserve the statistical properties of the original data, hence providing analysts with valid inferences (albeit with more uncertainty associated with estimates). While, as values are not real, disclosure risk should be much lower than in the original data.

Over the past three decades, synthetic data methods for statistical disclosure control have continually developed; methods have adapted to account for different data types, but mainly within the domain of survey data sets.

Administrative databases tend to comprise of mainly categorical variables. Categorical data sets can expressed in a tabular format, where a structured set of cell counts give the frequencies with which each combination of categories is observed (see below).

<table>
<thead>
<tr>
<th>Age</th>
<th>0-18</th>
<th>&gt;18</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geography</td>
<td>England</td>
<td>16</td>
<td>44</td>
</tr>
<tr>
<td></td>
<td>Wales</td>
<td>4</td>
<td>36</td>
</tr>
<tr>
<td>Total</td>
<td>20</td>
<td>80</td>
<td>100</td>
</tr>
</tbody>
</table>

These original counts can be replaced with synthetic counts. We used saturated models to achieve this, along with overdispersed count distributions such as the negative binomial (NB1) model.

The upshot is that there are tuning parameters that are set by the synthesizer (he or she responsible for generating the synthetic data), that arise from using a two- or three-parameter count distribution.

Saturated models guarantee the preservation of relationships between variables in the synthetic data, as no assumptions are made as to which interactions exist.

The computational time is essentially null, as no model fitting is required - the model’s fitted counts are just the observed counts. Thus the method scales equally well to large data sets.

As the fitted counts are just equal to the observed counts, it allows expected properties of the synthetic data to be determined a priori (that is, prior to synthesis). Uncertainty can then be added where it is most needed to add noise to sensitive cells in the original data.

Over the past three decades, synthetic data methods for statistical disclosure control have continually developed; methods have adapted to account for different data types, but mainly within the domain of survey data sets.

My thesis, through the fitting of saturated models, presents a way in which administrative databases can not only be synthesized quickly, but which also allows risk and utility to be formalized in a manner inherently unfeasible in other techniques.

The demand - and need - for near-real-time data was highlighted during the Covid-19 pandemic, and synthetic data sets are a way to facilitate this.

As an article for statistics magazine Significance, Claire McKay Bowen showed how data privacy can be visualised in art-form. Following a similar idea, let the following image of shoppers in a Liverpool street represent a data set with sensitive information.

In an article for statistics magazine Significance, Claire McKay Bowen showed how data privacy can be visualised in art-form. Following a similar idea, let the following image of shoppers in a Liverpool street represent a data set with sensitive information.

The idea is that they can preserve the statistical properties of the original data, hence providing analysts with valid inferences (albeit with more uncertainty associated with estimates). While, as values are not real, disclosure risk should be much lower than in the original data.

As the fitted counts are just equal to the observed counts, it allows expected properties of the synthetic data to be determined a priori (that is, prior to synthesis). Uncertainty can then be added where it is most needed to add noise to sensitive cells in the original data.