Assessing the Disclosure Risk of Perturbed Enterprise Data

Natalie Shlomo
Southampton Statistical Sciences Research Institute
University of Southampton
Highfield, Southampton
SO17 1BJ, United Kingdom
E-mail: N.Shlomo@soton.ac.uk

One of the aspects of the EU 7th framework Blue-ETS project deals with ways and means to access and release enterprise microdata, for example by generating synthetic datasets for public-use files. Techniques for producing synthetic data are related to methods of multiple imputation where the data is generated from a statistical model based on the sufficient statistics of the original data (Rubin, 1993, Reiter, 2005). In practice, partially synthetic data can be generated where some variables may be preserved or categorized. We assume that one dataset is to be released.

With such highly perturbed microdata, the question is how to assess the disclosure risk. In the case of enterprise microdata, disclosure risk arises from both identity disclosure and attribute disclosure. Identity disclosure is the risk that an identification can be made on the basis of identifying key variables and is related to notions of population uniqueness when dealing with samples. Attribute disclosure is the risk that information can be learnt from sensitive variables which for enterprise microdata, are typically continuous. When attribute disclosure is a concern, the sample is treated as the population.

The risk of disclosure for perturbed enterprise microdata is typically assessed by probabilistic record linkage where the perturbed dataset is matched to the original dataset and the risk measure takes the form of the probability of a correct match (Yancy, et al., 2002, Hawala, et al. 2005). This represents a worst-case scenario since it does not take into account the protection afforded by the sampling and it assumes that an intruder would have access to the original dataset. However, for attribute disclosure risk in enterprise microdata, we assume that the sample is a population and that it is likely that many continuous variables are publically available on external datasets.

Shlomo and Skinner (2010) consider disclosure risk assessment for identity disclosure of misclassified or perturbed sample microdata. In addition, they demonstrate that the probabilistic record linkage framework fits into the probabilistic modelling framework for identity disclosure. The results showed that the disclosure risk as measured by the probability of a correct match depends on the probability of perturbation.

In the probabilistic record linkage framework for assessing disclosure risk, all possible pairs are produced between the perturbed microdata and the original microdata. If variables are not perturbed, they should be used as blocking variables

to reduce the number of pairs. The probability of perturbation is typically measured through a string comparator which calculates a distance between the perturbed and original value of a variable, i.e. the added noise. If a variable is categorized in the dataset, the perturbed value can be estimated as the average of the boundaries of the category.

We consider two options for calculating string comparators:

1. Standardize the noise by subtracting its expectation and dividing by its standard deviation. Calculate the CDF under the normal distribution. With this set-up, if the perturbed value is equal to the original value then the value of the string comparator is 0.5.
2. Consider the following distance metric: $\exp(-|noise| / median(noise))$

   If the perturbed value is equal to the original value, then the value of the string comparator is 1.

We calculate a string comparator for each of the matching variables. Under the conditional independence assumption of probabilistic record linkage, the final matching probability is the weighted average of the string comparators for each of the variables. The weights can be obtained by normalizing the odds of a correct match for each variable as calculated from a logistic regression model where the response variable is the true match status and the explanatory variables are the string comparators. The weights represent the distinguishability of the variable and their contribution to determining a correct match. Based on the final matching probability, we determine the links through a pre-determined Type I Error threshold and calculate the following risk measures: proportion of correct matches out of the total links and the number of correct matches to the number of false matches for the linked pairs.

We demonstrate the technique on a dataset containing Sugar Farms from a 1982 survey of the sugar cane industry in Queensland, Australia which has been perturbed through different rates of correlated noise which preserve the sufficient statistics.