A Unified Framework for De-Duplication and Population Size Estimation
Master Class e-learning by Brunero Liseo

Statistical Area
October 15th 2020
14.30-16.00

Abstract

Record linkage refers to a variety of algorithmic and statistical methods for finding entries related to the same entity in different, usually large, data sets. De-duplication methods generalizes the above task to the case when records belonging to the same entity are present in the same data set. In this talk I discuss a method to perform a de-duplication process via a latent entity model, where the observed records - usually containing information in terms of categorical variables - are perturbed versions of a set of key variables drawn from a finite population of N different entities.

Also, the population size N is considered unknown.
As a result, a salient feature of the proposed method is the capability to account for the de-duplication uncertainty in the population size estimation. As by-products of the approach, I illustrate the relationships between de-duplication problems and capture-recapture models and obtain a more suitable prior distribution on the linkage structure.
On the computational side, a novel algorithm is proposed to sample from the posterior distribution of the matching configuration based on the marginalization of the key variables at a population level.
The performance of the proposed method will be illustrated using two synthetic data sets consisting of German names. Finally a real data application is presented, using records from two lists containing information related to death casualties in the recent Syrian conflict.

Trainer

Brunero Liseo (BSc, Statistics, Sapienza, Roma; PhD, Statistics, Sapienza, Roma) is Professor of Statistics at the School of Economics, Sapienza Università di Roma. Before that, he was a lecturer in Probability at the department of Statistics, Sapienza Università di Roma, from 1992 to 1998. His areas of interest are on foundations of inference, Bayesian modelling, distribution theory and official statistics, in particular data linkage, integration and small area estimation. He wrote more than 60 papers on refereed journals.

He has been a member of the Board of the International Society for Bayesian Analysis. He is now the Director of the Ph. D. of Economics at Sapienza, Rome. He is also a member of the scientific advisory board of Istat and a member of the Italian chapter of WADA (World Anti-Doping Organization).

He is at the moment Principal Investigator on a broad project sponsored by Sapienza University on data integration in Demography.

Homepage: https://sites.google.com/a/uniroma1.it/brulis/home

Info

Per partecipare alla Master class in lingua inglese è necessario inviare una mail entro lunedì **12 ottobre** a tutor@istat.it, specificando nell'oggetto "**MasterClass Liseo**". Agli iscritti sarà inoltrato successivamente il link per partecipare.