

Enhanced data fusion and anonymization for administrative and survey data

Abstract

The fusion and anonymization of multiple heterogeneous data sources remain major challenges in applied statistics. In this work, we consider the joint use of demographic and fiscal census data together with several sample surveys. The objective is to integrate these sources in order to obtain a coherent representation of the overall population and to enable the evaluation of policy changes, such as reforms of the fiscal system, while ensuring that the resulting data are fully synthetic and thus completely anonymized.

We present a modeling framework for merging data sets that share the same type of statistical units (e.g., households), and we show how this framework can be enhanced by incorporating information from data sets defined on different units (e.g., individuals). We also address the issue of harmonizing surveys that rely on distinct sampling designs. The proposed approach leads to a fully anonymized synthetic data set that preserves the main statistical properties of the original data and can be directly used for analysis by end users.

Keywords: (Variational) Autoencoders; Generative Adversarial Networks; Missing Data; Sampling Weights; Wasserstein Distance.