

Enhanced Data Fusion, Machine Learning, and Anonymization for Tax Systems

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 tax reforms, while ensuring that the resulting data are fully synthetic and therefore completely anonymized.

We present a modeling framework for merging data sets that share the same type of statistical units (e.g., households), and show how this framework can be extended by incorporating information from data sets defined on different statistical units (e.g., individuals). We also address the harmonization of surveys relying on different sampling designs. The proposed methodology produces a fully synthetic and anonymized data set that preserves the main statistical properties of the original data while remaining directly usable for statistical analyses and policy evaluation.

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