Innovative modelling with synthetic data in preparation for the UK 2021 Census

Ioannis Kaloskampis
Senior Data Scientist, UK Office for National Statistics

Abstract

Government organisations, businesses, academia, members of the public and other decision-making bodies require access to a wide variety of administrative and survey data to make informed and accurate decisions. However, the collecting bodies are often unable to share rich micro-data without risking breaking the confidentiality checks required to obtain this data. This can restrain the efforts of the data science community to build robust models to tackle major challenges such as climate change, economic deprivation and global health.

In this work, we propose to address the problem of micro-data sharing by building a synthetic data generation system, using robust statistical, machine learning and state of the art deep learning techniques, such as generative adversarial networks and autoencoders. Our system produces synthetic datasets that are close to the originals for a variety of real-world applications and includes data quality measures and privacy disclosure checks, for which we use the differential privacy framework.

We present a recent application of our system, the generation of multiple synthetic datasets for testing the load balancing and functions, used in the processing pipeline of the UK 2021 Census Rehearsal. Additionally, we present results for several real-world datasets and discuss the quality and data privacy checks, comparing the advantages and disadvantages of different algorithms.

Our work is highly impactful on both industry and research, as it allows datasets such as Census, Labour force survey and other official data to be shared more widely and securely for processing and analysis. It can also supplement existing datasets providing a cost-effective way to gather admin and survey data.

Keyowords: *synthetic data, privacy, machine learning, disclosure control, administrative data*