Synthetic Population Generation with MCMC Algorithm

Viktória Várkonyi, Zsolt Vizi
National Laboratory for Health Security, University of Szeged, Szeged, Hungary

In the context of modern data protection the generation of synthetic populations has received special attention. The challenges in data protection have led to an increasing demand for methods that enable the protection of sensitive data. Synthetic population generation is a method that helps create populations modeled on real data. This method allows the creation of data that represent the characteristics of a real population but does not include data that identify real individuals. It can be particularly useful in areas where large amounts of sensitive data need to be managed, such as in the healthcare or financial sectors. Several algorithms have been developed to create synthetic populations. These methods cover different areas of mathematics, such as machine learning, statistical learning and combinatorial optimisation.

In this presentation, we present and compare two synthetic population generation algorithms based on evaluation metrics: IPF algorithm and Gibbs-sampling. It is important to note that these methods also have their limitations. The goal is to model and reproduce real data as accurately as possible. Based on the [1] and the [2] articles, we implemented the two algorithms and created synthetic populations for census data available on the Statistics Canada website (https://www150.statcan.gc.ca/n1/en/catalogue/98M0001X) Statistics Canada website. We then compared the synthetic data generated by the two methods with the original data using the metrics presented in the articles. The two algorithms, the data processing and the evaluations were coded in Python.

The link: https://github.com/Varviki2002/synthetic_pop_gibbs.git.

[1] Farooq, B., Bierlaire, M., Hurtubia, R., Flötteröd, G., Simulation based population synthesis. *Transportation Research Part B: Methodological*, **58** (2013), 243–263.

[2] Prédhumeau, M., Manley, E., A synthetic population for agent-based modelling in Canada. *Scientific Data*, **10(1)** (2023), 148.