Synthetic population data

Creation and use

Antonia Gieschen
Lecturer in Predictive Analytics
University of Edinburgh

Introduction

Antonia Gieschen

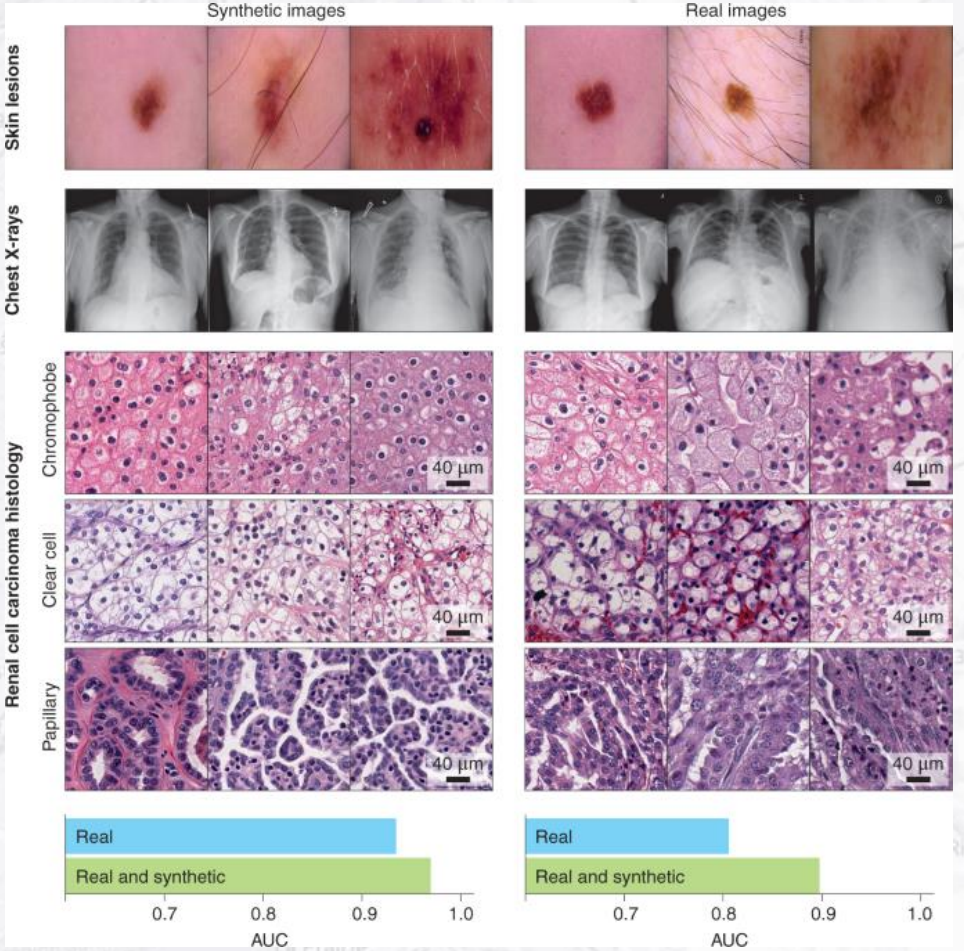
Lecturer in Predictive Analytics

<u>Research interests:</u>

- Financial wellbeing and its intersection to health
- Consumer behaviour and segmentation
- Population health and inequality
- Spatial analysis and clustering

Synthetic data

- Why synthetic data?
  - Wide ranged uses: in medicine, astrophysics etc. to create more samples from existing data where information is limited
  - But also in the social sciences and population studies where we want to protect individuals' privacy
- How is it created?
  - Usually ML based e.g. through neural network approaches
  - In population study often based on census data -> **synthetic populations**



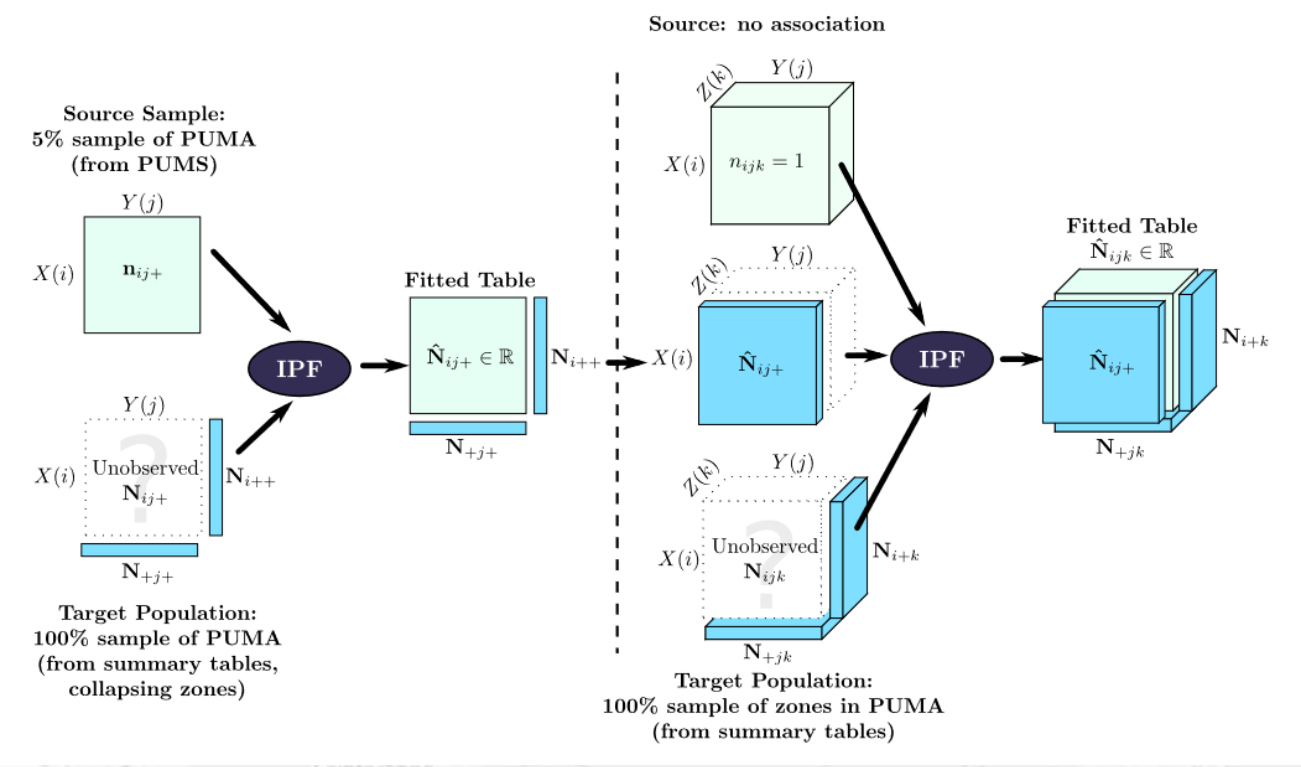
Chen, R.J., Lu, M.Y., Chen, T.Y. *et al.* Synthetic data in machine learning for medicine and healthcare. *Nat Biomed Eng* **5**, 493–497 (2021). https://doi.org/10.1038/s41551-021-00751-8

Synthetic populations

- What are synthetic populations in the context of synthetic data?
- Census data is usually published in two formats:
  - Aggregated data (summary tables) for the whole population by geographic region (Summary Files in USA or Basic Summary Tabulations in Canada)
  - Sample of household or individual data (PUMS in USA or PUMF in Canada)
- If goal is to estimate a whole population, these two have to be connected
- The resulting population is called **synthetic** because its properties are estimated using known totals (from summary tables) and the actual records are sampled individuals/households (from PUMS/PUMF)

Synthetic populations

- Example method: IPF for synthetic population creation



IPF procedure to derive a population table based on sampled household data and known summary table of target population

(Figure from: Pritchard, D. R., & Miller, E. J. (2012). Advances in population synthesis: fitting many attributes per agent and fitting to household and person margins simultaneously. *Transportation*, *39*(3), 685-704.)

Enhanced synthetic populations

We first create data in which each synthetic agent in our population is defined through census data and its variables.

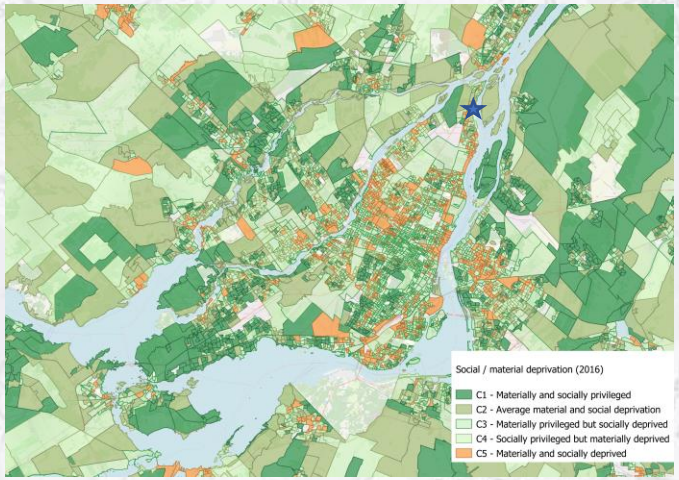
As a non-exhaustive example:

	Age	Gender	Annual household income before taxes	Housing situation	Education	Location
Individual A “Anna”	30-35	Woman	82,589	Renting 2 bedroom apartment	Master’s degree	H1A

Enhanced synthetic populations

We then add further variables to those records.

	Age	Gender	Annual household income before taxes	Housing situation	Education	Location	Subjective financial anxiety level	Socio-economic deprivation in area
Individual A “Anna”	30-35	Woman	82,589	Renting 2 bedroom apartment	Master’s degree	H1A	Low / no problems keeping up	C2: Average level of social and material deprivation



References

Relevant publications:

- Belkhiria, F., Nie, J.Y., Paquet, C., Sengupta, R., Gieschen, A., Talukder, B., Brown, S. and Dubé, L., 2022, December. Using Big Data and Machine Learning for Multilayered Surveillance for Healthy Food Environment and Diet. In 2022 IEEE International Conference on Big Data (Big Data) (pp. 4055-4064). IEEE. https://doi.org/10.1109/BigData55660.2022.10020762
- Gieschen, A., Paquet, C., Sengupta, R., Aunio, A.L., Belkhiria, F., Brown, S. and Dube, L., 2023. SynthEco-A multi-layered digital ecosystem for analysing complex human behaviour in context. International Journal of Population Data Science, 8(3), p.2285. https://doi.org/10.23889/ijpds.v8i3.2285
- Pritchard, D. R., & Miller, E. J. (2012). Advances in population synthesis: fitting many attributes per agent and fitting to household and person margins simultaneously. *Transportation*, *39*(3), 685-704. https://doi.org/10.1007/s11116-011-9367-4

Social deprivation data (visualised on slide 7):

- Institut national de santé publique du Québec (INSPQ). Index of material and social deprivation compiled by the Bureau d'information et d'études en santé des populations (BIESP) from 1991, 1996, 2001, 2006, 2011, 2016 and 2021 Canadian Census data. [https://www.inspq.qc.ca/en/deprivation/material-and-social-deprivation-index]

Images used:

- Synthetic data image on slide 3: Chen, R.J., Lu, M.Y., Chen, T.Y. *et al.* Synthetic data in machine learning for medicine and healthcare. *Nat Biomed Eng* **5**, 493–497 (2021). https://doi.org/10.1038/s41551-021-00751-8
- Background image, base maps and data from OpenStreetMap and OpenStreetMap Foundation (CC-BY-SA). © https://www.openstreetmap.org and contributors.