Population synthesis using Variational Auto-Encoders (VAE)

Abdoul Razac SANÉ, PhD Student
Laboratory : Splott

Thesis advisor : Pierre-Olivier VANDANJON, Splott
Supervisors : Rachid Belaroussi, Grettia
Pierre Hankach, Mast

Université Gustave Eiffel

September 19, 2024

Definition : Synthetic population is a generic representation of a group of individuals or households, that mimics the characteristics and behaviors of the actual population.

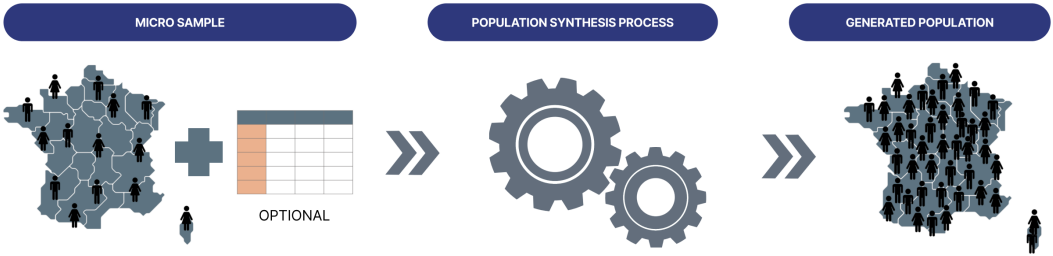
⇒ **Objectifs**

- Reproduce realistic populations from aggregate data (e.g., census data).
- Provide data for simulations (e.g., urban planning, transportation, public health).
- Enable analysis without compromising personal data privacy.

Definition : Synthetic population is a generic representation of a group of individuals or households, that mimics the characteristics and behaviors of the actual population.

⇒ **Objectifs**

- Reproduce realistic populations from aggregate data (e.g., census data).
- Provide data for simulations (e.g., urban planning, transportation, public health).
- Enable analysis without compromising personal data privacy.



Plan

1. Population synthesis methods
2. Variational autoencoder
3. Case study
4. Results
5. Conclusion and perspectives

Figure: Classification of methods, inspired by Yameogo et al. [2021]

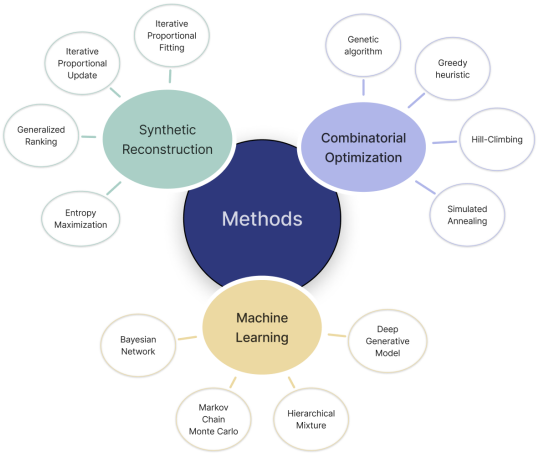


Figure: Classification of methods, inspired by Yameogo et al. [2021]

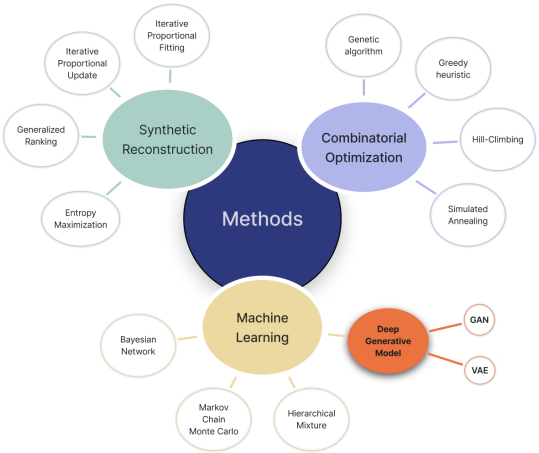


Figure: The autoencoder is an unsupervised model trained to copy its inputs. It compresses the input vector X with dimensions n into the latent representation Z with fewer dimensions and then reconstructs it back into the data space

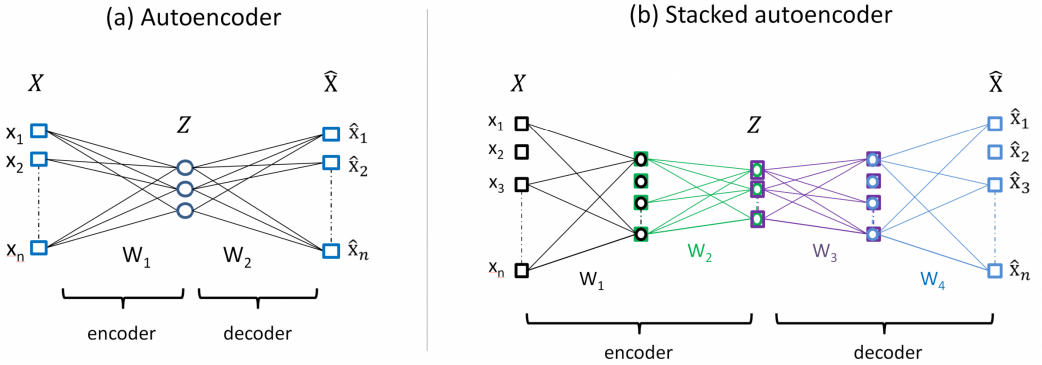
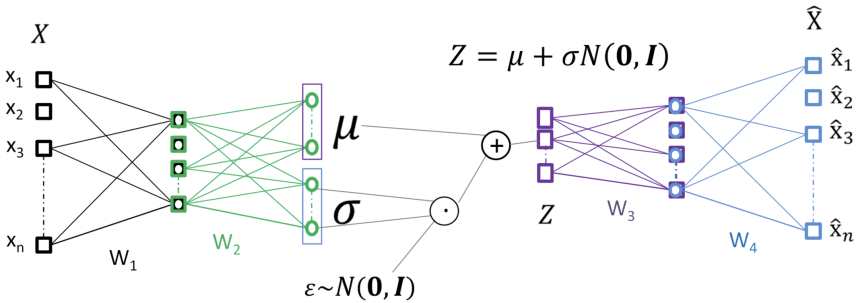


Figure: VAE illustration and loss function



$$L(\phi,\theta) = \sum_{i=1}^{N} \underbrace{c_1 \times \sum_{k\in\{num\}} (\hat{x}_i^k - x_i^k)^2}_{\text{Numerical cost}} + \underbrace{c_2 \times \sum_{k\in\{cat\}} \sum_{d=1}^{D_i} x_i^{kd} \log \hat{x}_i^{kd}}_{\text{Categorical cost}} + \underbrace{\beta \times \mathbb{KL}(\mathcal{N}(\mu^k, \sigma^k) \| \mathcal{N}(\mathbf{0}, \mathbf{I}))}_{\text{KL cost}}$$

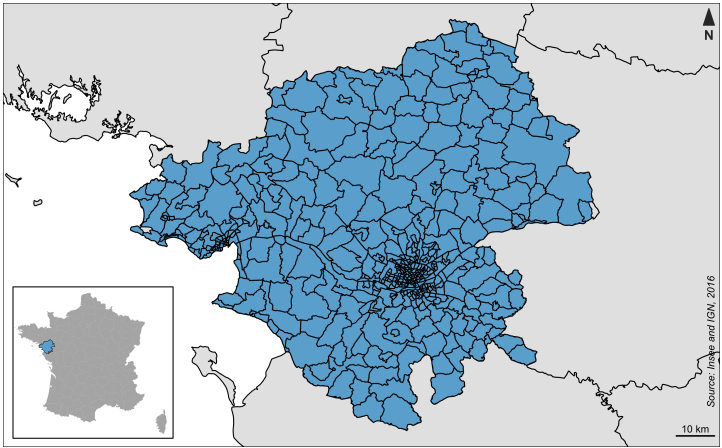
Case study

→ **To yield Loire-Atlantique pop.**

→ **Around 1.4M**

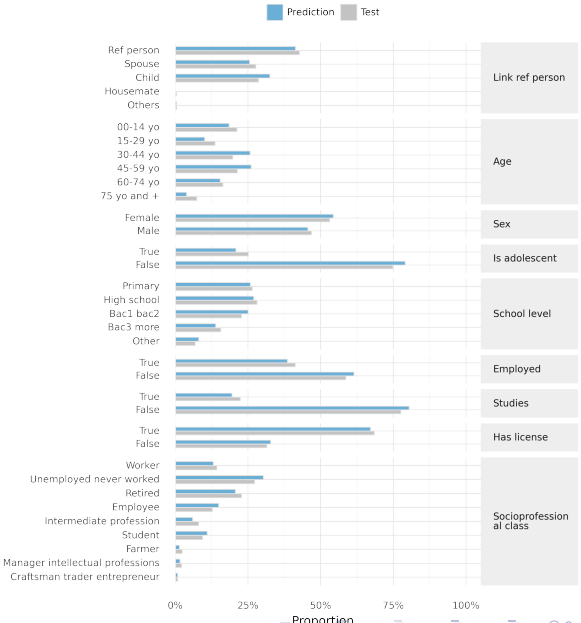
Nantes HTS data

- 12,700 households
- 29,500 people aged 5 and older



Results : Performances on the individual attributes

An error of less than 3% for all attributes except Age. The age categories of age between 15-29 years old and the 75+ years old are underestimated with 5% less than the true population, whereas the 30-44 and 45-59 age categories are overestimated by 5%compared to the test set



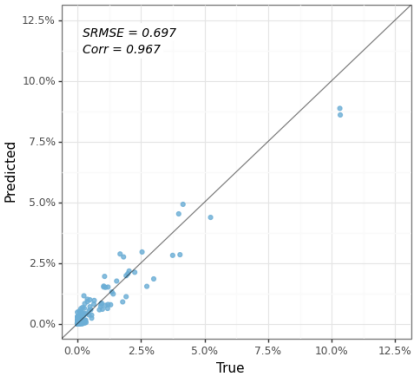


Figure: Household attributes : type of housing, type of house occupancy, number of vehicles and Internet availability

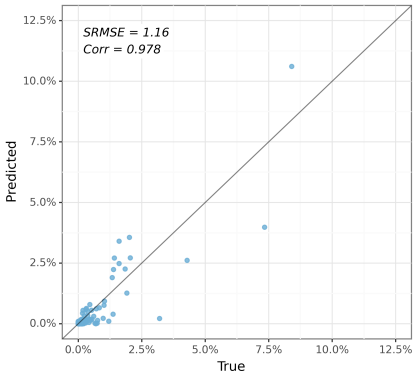
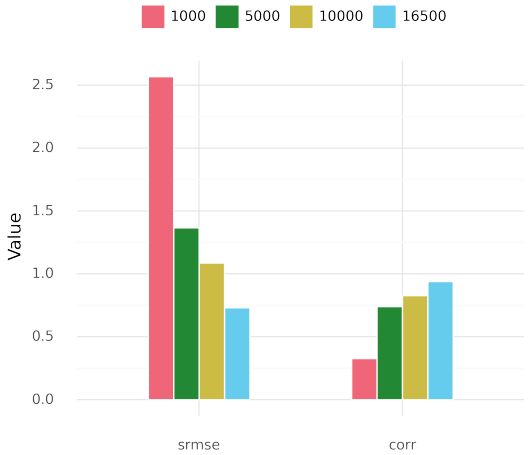


Figure: Individuals attributes : Gender, Educational level, Socio-professional categories and Link with the reference person

Results : Training size influence

Figure: Values of different metrics on the 41.6 k-dimensional joint representation with a growing training set size



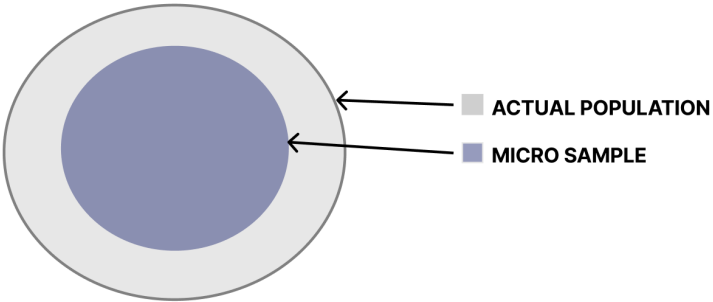


Figure: Zero-cells, Kim and Bansal [2022]

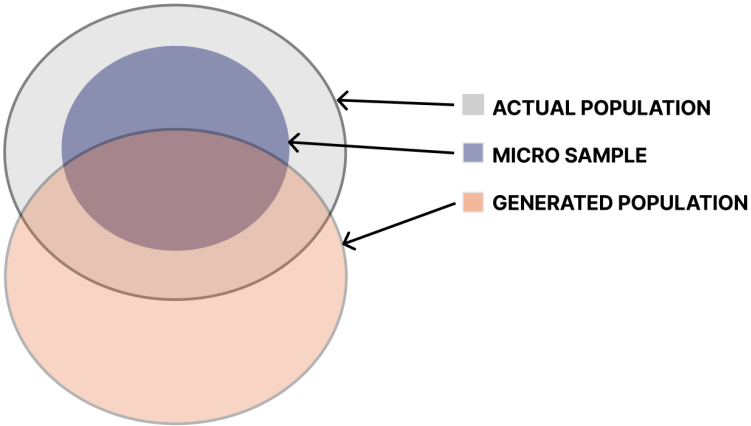


Figure: Zero-cells, Kim and Bansal [2022]

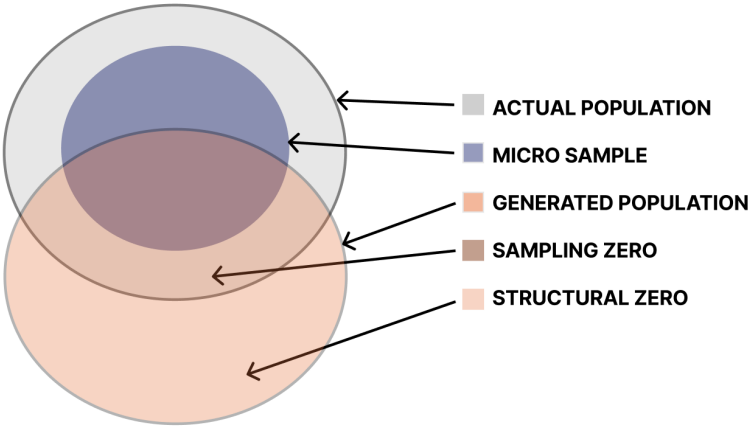
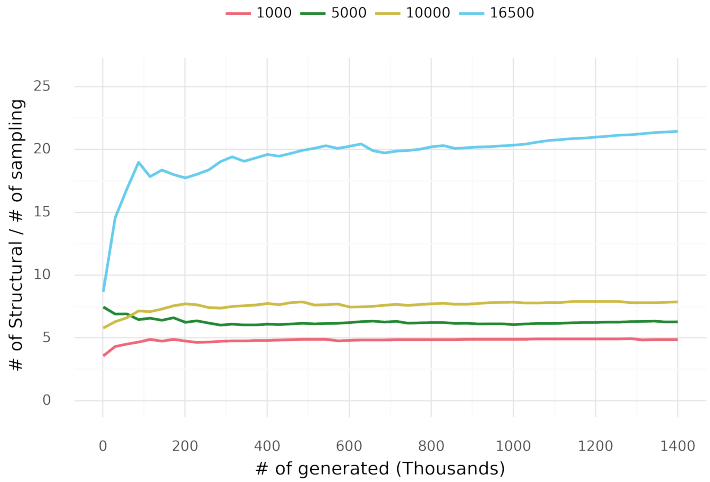


Figure: Zero-cells, Kim and Bansal [2022]

Figure: Ratio between the number of structural zeros to sampling zeros versus the size of the generated population



Presenting a VAE model for generating a synthetic population of agents using mixed data and their subsequent evaluation.

☺ Remarkable performance of VAEs

- The average errors for attribute marginal data were less than 3 percentage points
- Ability to generate new individuals not present in the training sample
- Take into account mixed data, but categorical attributes becomes more robust.
- Capable of processing high-dimensional data while maintaining good performance with a reasonable sample size.

☹ Main limitation

- Structural zeros : the use of VAE requires post-processing to eliminate them.

Perspectives ...

- Taking marginal data into account to improve model performance
- To generate individuals in households,
- Integrating sequential data: activity plans including consumer-side logistics

Thank you for your attention ... ☺

Boyam Fabrice Yameogo, Pierre-Olivier Vandanjon, Pascal Gastineau, and Pierre Hankach. Generating a Two-Layered Synthetic Population for French Municipalities: Results and Evaluation of Four Synthetic Reconstruction Methods. *Journal of Artificial Societies and Social Simulation*, 24(2):5, 2021. ISSN 1460-7425. doi: 10.18564/jasss.4482. URL http://jasss.soc.surrey.ac.uk/24/2/5.html.

Eui-Jin Kim and Prateek Bansal. A Deep Generative Model for Feasible and Diverse Population Synthesis, August 2022. URL http://arxiv.org/abs/2208.01403. arXiv:2208.01403 [cs, stat].