# Population synthesis with deep generative model: a joint household-individual approach

Abdoul Razac Sané, Rachid Belaroussi, Pierre Hankach, Pierre-Olivier Vandanjon

Sané *et al. Computational Urban Science* (2025) 5:34
https://doi.org/10.1007/s43762-025-00195-9

Computational Urban Science

ORIGINAL PAPER

Open Access

# Population synthesis with deep generative model: a joint household-individual approach

Abdoul Razac Sané[1*], Rachid Belaroussi[2†], Pierre Hankach[3†] and Pierre-Olivier Vandanjon[1†]

**Abstract**

This paper introduces two novel deep generative frameworks for synthetic population generation that jointly model household and individual attributes. In leveraging Variational Autoencoders (VAEs), we propose herein the SVAE-Pop2 method, which employs a single VAE with fixed-size padded inputs, along with the MVAE-Pop2 method, which uses dedicated models for various household sizes. Evaluated on a French household travel survey dataset, our experiments reveal that while both approaches effectively reproduce the actual population's characteristics, MVAE-Pop2 achieves greater fidelity in joint attribute distributions. The proposed methodologies suggest improvements in agent-based simulations and urban modeling by means of generating realistic, multi-layered synthetic populations.

**Keywords** Synthetic population, Machine learning, Deep generative model, Variational autoencoders, Household-individual, Two layers

## 1 Introduction

Population synthesis is the process of generating a representative synthetic population with realistic demographic and behavioral attributes in order to simulate social interactions and urban dynamics. As a prerequisite for agent-based modeling, it plays an important role in both computational and urban sciences and moreover is essential for developing intelligent, resilient cities. Serving as the foundational step in agent-based simulations, population synthesis enables researchers to better understand human behavior and social dynamics, while enabling practitioners to make well-informed decisions. Population synthesis therefore contributes to multiscale urban modeling, which integrates different levels of urban data for comprehensive analysis; in addition, it supports mobility analysis and visualization, thereby facilitating the extraction of meaningful patterns from movement data.

One of the key aspects in the generation of synthetic populations is the joint modeling of individual and household characteristics. Traditional approaches often treat individuals independently, ignoring the household structure. However, in transport and urban modeling, the ability to generate a two-layered population is essential since an individual's decisions and experiences depend on both personal attributes and family context. For example, in mobility studies, commute patterns, car ownership and mode of transport are all strongly influenced by household composition, the employment status of other members, and shared resources (Fabrice Yaméogo et al. 2021). Similarly, in epidemiological studies, the spread of infectious diseases is highly dependent on household structures since transmission dynamics are influenced by interactions among cohabiting individuals (Goeyvaerts et al. 2017). An accurate representation of household-individual interactions significantly enhances the realism

†Rachid Belaroussi, Pierre Hankach and Pierre-Olivier Vandanjon contributed equally to this work.

*Correspondence:
Abdoul Razac Sané
abdoul-razac.sane@univ-eiffel.fr
[1] AME-SPLOTT, University Gustave Eiffel, All. des Ponts et Chaussées, 44340 Bouguenais, France
[2] COSYS-GRETTIA, University Gustave Eiffel, 5 Bd Descartes, 77420 Champs-sur-Marne, France
[3] MAST-LAMES, University Gustave Eiffel, All. des Ponts et Chaussées, 44340 Bouguenais, France

and predictive power of agent-based models across various domains.

Traditional approaches to generating two-layered synthetic populations can be broadly classified into three categories: Synthetic Reconstruction (SR), Combinatorial Optimization (CO), and Statistical Learning (SL) (Sun et al. 2018; Fabrice Yaméogo et al. 2021). Both SR and CO techniques construct synthetic populations by replicating individual records. These methods are predominantly deterministic, ensuring that the same input data always yield the same outcome (although CO methods can incorporate stochastic processes to optimize the assignment of individuals to households). In contrast, SL methods generate synthetic populations by estimating the joint probability distribution of demographic attributes and then sampling from this distribution, thus rendering them inherently stochastic.

Variational Autoencoders, which are a subset of the SL methods, have emerged as a promising machine learning approach for synthetic population generation and in recent years have been increasingly explored and compared to classical methods ever since the foundational work of Borysov et al. (2019a) to the methodological paper of Sané et al. (2025). As deep generative models, VAEs offer several advantages over traditional approaches. Specifically, they can learn the joint distribution of both numerical and categorical attributes, hence enabling a more comprehensive representation of the population. One key advantage of machine learning-based methods is their ability to generate agents that are not present in the sample data, although this ability does come with the risk of producing unrealistic agents that do not exist in the actual population. Another notable benefit of these methods is their capacity to generate diverse synthetic populations even from small sample sizes.

Despite their potential, these models have, to date, been predominantly applied to the generation of synthetic populations at a single layer—either households or individuals—rather than integrating both levels simultaneously.

Very recently, a few studies in synthetic population generation have leveraged deep generative models to jointly model household and individual characteristics (Aemmer and MacKenzie 2022; Qian et al. 2024). As described in the literature review section below, the main limitation of the architecture proposed in Aemmer and MacKenzie (2022) is the possibility of generating inconsistent households, while the limitation of Qian et al. (2024) lies the computation burden, which constrains operational applications.

This paper introduces a deep generative framework that harnesses the full potential of advanced neural architectures in order to generate two-layered synthetic populations. Its contributions are fourfold, as follows:

- SVAE-Pop2: we propose a novel approach that uses a Single Variational Autoencoder (VAE) to jointly generate entire households and their members in a unified, end-to-end process.
- MVAE-Pop2: we introduce an alternative architecture based on a Multi-VAE framework, where a distinct VAE is trained for each household size, improving the model's capacity to capture structure-specific dependencies.
- Empirical evaluation: We conduct a comprehensive evaluation of both approaches using a representative case study, highlighting their ability to generate realistic and coherent synthetic populations.
- Open and reusable codebase: We provide well-documented, publicly available code to ensure full reproducibility and facilitate adaptation to various application contexts.

This paper is organized as follows. First, the generative methods applied to population synthesis will be reviewed, in examining the existing approaches and their limitations. Next, the fundamental principles of the Variational Autoencoder (VAE), which forms the theoretical basis of our work, will be introduced, followed by a detailed description of our methodology. Subsequently, the experimental results, accompanied by a comparative analysis of the approaches' performance, will be presented. The final sections will discuss these results, highlight the contributions and limitations of our approach, and conclude by suggesting directions for future research.

## 2 Literature review

Generative models capture the underlying probability distribution of a dataset, enabling them to generate new data points similar to those in the training set (Duda et al. 2001). Unlike discriminative models, which focus on learning decision boundaries between classes, generative models are aimed at modeling the data distribution itself. They are used in various fields, including: text generation (De Rosa and Papa 2021), image generation (Van den Oord et al. 2016), speech synthesis (Kong et al. 2020), and drug discovery (Zeng et al. 2022).

Hinton et al. (2006) introduced a layer-wise, unsupervised pretraining method for efficiently training deep neural networks; this method would later influence generative models, like Variational Autoencoders (VAEs) and Generative Adversarial Networks Networks (Goodfellow et al. 2014). Their method describes Restricted Boltzmann Machines, which are generative models, and

explains how stacking such machines forms a Deep Belief Network capable of capturing complex data distributions.

Very few studies in the literature however actually use these techniques to generate synthetic populations for microsimulations. Among the first of recent studies, Borysov et al. (2019b) employed Variational Autoencoders (VAE) to generate a synthetic population of travelers. These authors showed that VAEs yield better results on high-dimensional data compared to conventional Gibbs sampling and Bayesian networks. The study by Johnsen et al. (2022) implemented two deep generative models, i.e. Conditional Variational Auto-Encoder (CVAE) and Conditional Generative Adversarial Networks (CGAN), to generate an urban population in relying on aggregated population features and real estate infrastructure information. The results of this study demonstrate that CVAE outperforms both empirical methods and CGAN.

Yameogo et al. (2021) compared four algorithms by generating a two-layered population for the city of Nantes (France). They described Hierarchical Iterative Proportional Fitting, Iterative Proportional Update (IPU), Generalized Ranking and Relative Entropy Minimization, all within a common framework.

Sun et al. (2018), have proposed synthetic population generation methods to capture the hierarchical structure that exists between individuals in a household. Casati et al. (2015) introduced a method referred to as"hierarchical MCMC"(or hMCMC), that follows a multi-step process. First, individuals in the same household are ordered according to their role (i.e. owner, spouse, children, others). Second, using MCMC (Farooq et al. 2013), the characteristics of both households and owners are generated. Third, spouses and children are generated conditionally based on the joint probabilities of the household and owner data produced during the second step. Finally, other household members are generated conditionally based on the joint probabilities derived from the second and third steps.

The combination of generative models with two-level frameworks is a relatively new area of research that has remained largely unexplored. The main challenge lies in the fact that the input size of a Variational Autoencoder (VAE) must be fixed, whereas households consist of a variable number of individuals. Consequently, representing both a household and its members in a unified format is non-trivial, as the dimensionality of the input vector naturally depends on household size.

The work by Aemmer et al. (2022) was among the first to bridge these fields and overcome this difficulty. Their method generates a synthetic population of households and individuals using a VAE/CVAE framework in two steps: first, a synthetic population of households is generated; then, synthetic individuals are generated one by one, conditioned on the household data. However, one major limitation of this approach is its inability to capture the hierarchical structure within households, nor does it model the interdependencies between individuals. As a result, the individuals generated are independent of one another and may not adhere to the constraints inherent in a realistic household structure (e.g. a household composed of a couple with two children). This limitation can have significant implications for downstream applications, such as activity-based transport modelling, where interactions within households can greatly impact individual behaviours. For instance, creating inconsistent households — such as a household which includes only two children — can result in inaccurate simulations of daily mobility patterns or resource allocation. Similarly, the timing, destination and mode of transport chosen by one individual often depend on the activities and constraints of other household members, such as coordinating school drop-offs, shared commutes or joint use of a family vehicle. Capturing these interdependencies is essential in order to produce synthetic populations that are both realistic and policy-relevant.

The very recent study by Qian et al. (2024) sought to address this limitation by proposing a novel approach that combines VAE with fine-tuning. Their method transforms both the household and its individuals into a single vector and then employs a VAE to generate the synthetic population. To account for the hierarchical structure between household members, the authors introduced a new loss function, called *Decoupled Binary Cross-entropy (D-BCE)*, that quantifies the relationships between each attribute of the household and its individuals. Although this approach has shown some promising results, it has not been validated at a large scale. In fact, the D-BCE loss function is computationally expensive, with a complexity of $O(n^2)$ compared to the standard Binary Cross-Entropy (BCE), whose complexity is $O(n)$. Moreover, this function requires a large amount of data in order to be effective. This computational burden poses a significant challenge to the practical implementation of large-scale urban simulations or planning tools, where the rapid generation of diverse and realistic populations is crucial. Without scalable solutions, it remains difficult to apply such methods in operational contexts, which limits their impact on real-world decision-making.

We are proposing herein two methods to overcome these limitations:

- the first one is very close to the architecture proposed by Qian et al. (2024) yet with a light cost function;
- the second is original and consists of building one VAE per size of household.

## 3 Deep generative modeling of a synthetic population

### 3.1 The variational autoencoder principle

Variational Autoencoder (VAE) is a powerful yet unsupervised generative model designed to learn structured representations of complex data. This is achieved by mapping latent variables to observable data through a carefully designed nonlinear transformation. Such an approach allows VAEs to generate new data samples by sampling from a learned latent space and then transforming these samples back into the data domain using a decoder function. The structure of a Variational Autoencoder (VAE) is illustrated in Fig. 1a. According to this architecture, the variables $\mu$, $\sigma$, and $\varepsilon$ all share the same dimension $m$ as the latent space $Z$.

During training, the encoder processes an input data sample $X$ and outputs two vectors, $\mu$ (mean) and $\sigma$ (standard deviation), both of size $m$. These vectors define the parameters of a multivariate normal distribution, from which the latent variable $Z$ is drawn using the formula: $Z = \mu + \sigma \cdot \varepsilon$. Here, the symbol $\cdot$ denotes element-wise multiplication, and $\varepsilon$ is a random variable sampled from a standard multivariate normal distribution: $\varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

A key technique employed in this process is the *reparameterization trick*. Instead of directly sampling $Z$ from a stochastic distribution, which would make gradient-based optimization difficult, the model expresses $Z$ as a deterministic transformation of $\varepsilon$. This adjustment allows gradients to be backpropagated smoothly through the network during training.

Once $Z$ has been obtained, the decoder maps it back into the original data space in order to reconstruct $X$, in producing an output $\hat{X}$. The model then evaluates how closely $\hat{X}$ matches the original $X$, in guiding the optimization process. Training involves minimizing a loss function that balances reconstruction accuracy and the regularization of $Z$, through use of using backpropagation to update the encoder and decoder parameters.

The loss function of our VAE architecture comprises two main components, as expressed in Eq. 1:

$$L(\phi, \theta) = \underbrace{\sum_{a=1}^{N} \sum_{k=1}^{K} \sum_{d=1}^{D_k} x_a^{kd} \log \hat{x}_a^{kd}}_{\text{Reconstruction cost}} + \underbrace{\beta \times \mathbb{KL}(\mathcal{N}(\mu, \sigma) \| \mathcal{N}(\mathbf{0}, \mathbf{I}))}_{\text{KL cost}} \tag{1}$$

whre $N$ is the number of agents (individuals or households), $K$ is the number of attributes, and $D_k$ is the number of modalities for attributes $k$. $\mu$ (mean) and $\sigma$ (standard deviation), both of size $m$ constitutes the encoder outputs.

The first term accounts for the *reconstruction error*, which measures how well the VAE reproduces the input data $X$ in its output $\hat{X} = P_\phi(Q_\theta(X))$. The second term, a *regularization term* represented by the *Kullback-Leibler (KL) divergence*, quantifies how much the learned latent space distribution deviates from the standard Gaussian prior $\mathcal{N}(\mathbf{0}, \mathbf{I})$. By enforcing this constraint, the VAE ensures that latent representations are structured in a way that facilitates both smooth sampling and interpolation.

The hyperparameter $\beta \in \mathbb{R}_+$ serves as weighting coefficients, intended to balance the trade-off between reconstruction accuracy and regularization (Higgins et al. 2017).

Once trained, the VAE can generate new synthetic samples by drawing a latent variable $Z$ from the standard Gaussian prior $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and then passing it through the decoder. This process maps the sampled points back to the original data space, in creating new data instances. Figure 1b illustrates this process.

### 3.2 Variational autoencoder for single-layered synthetic population generation

The application of Variational Autoencoders (VAEs) to a population synthesis follows a straightforward process. A practitioner starts with a sample of the population, typically obtained from a census or household travel survey. The goal here is to generate a synthetic

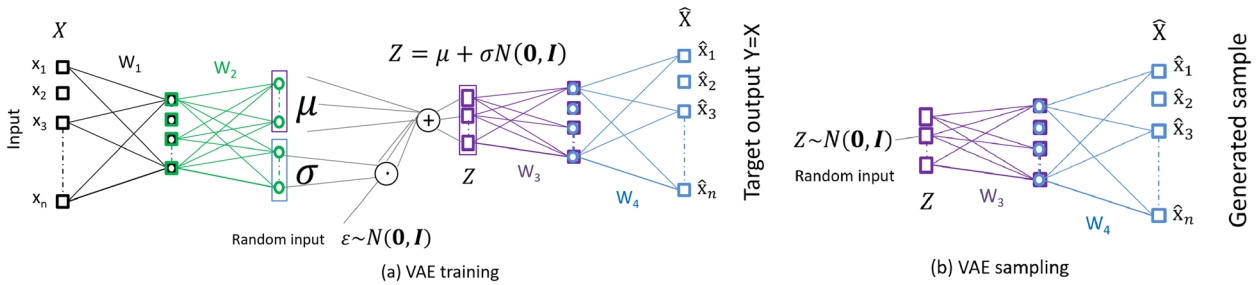


**Fig. 1** Variational Autoencoder architecture : **a** During training, the random variable $\varepsilon$ is injected into the latent space $Z$ as an external input; **b** After training, input a $Z$ following a normal random law is input in order to generate a new sample

population whose size matches that of the actual population, in using this sample as a basis.

The synaptic weights and hyperparameters of the VAE are learned from the sample, thereby modeling the joint probability distribution of the input attributes. Population generation is then performed by sampling from this learned probability distribution within the latent space, with the number of samples equal to the size of the actual population. From a practical standpoint, the choice of input and output variables determines the nature of the generated entities. If these variables represent individual attributes, then the VAE will generate individuals. Conversely, if they correspond to household characteristics, the VAE will produce households.

In the field of synthetic population generation, this distinction is crucial. Consequently, VAEs are typically employed to generate either households or individuals, and rarely individuals within households, given that the latter approach introduces additional complexities.

### 3.3 Two approaches for modeling two-level synthetic population generation using VAEs

One of the major challenges encountered when generating a synthetic population that integrates both households and their individuals lies in data modeling. In fact, the number of individuals varies from one household to another, while VAE architectures require fixed-sized inputs. To overcome this constraint, we propose two distinct approaches (see Fig. 2):

- A single-model approach, in which each household and its individuals are represented as an extended vector.
- A household size-specific approach, whereby the representation of each household and its individuals is also provided as an extended vector, with a dedicated model for each household size.

#### *3.3.1 Single VAE for two layered population generation (SVAE-Pop2)*

This approach, which relies on a single VAE, requires that the input be of fixed size. To achieve satisfy this requirement, we have defined a vector with a maximum size corresponding to a household composed of $S$ members (in the present example, $S = 5$). When actual households contain fewer than $S$ members, the corresponding vectors are padded with fictitious individuals, each characterized by specific attributes.

$$X = \left( \underbrace{x_1, x_2, ..., x_{k_h}}_{houshold}, \underbrace{x_1^1, x_2^1, ..., x_{k_i}^1}_{individual_1}, \underbrace{x_1^2, x_2^2, ..., x_{k_i}^2}_{individual_2}, \cdots, \underbrace{x_1^S, x_2^S ... x_{k_i}^S}_{individual_S} \right) \quad (2)$$

where $k_h$ is the number of attributes describing the household, and $k_i$ the number of attributes for each individual. $x_k$ is the $k^{th}$ household attribute, while $x_j^i$ is the $j^{th}$ individual attribute of individual $i$, as described in Table 1.

For example, for $S = 5$, a household with three individuals is padded with two fictitious individuals, who in turn are assigned additional modalities for each attribute. As an illustration, for the gender attribute (called *sex*), which originally comprised two modalities, a third modality, referred to as a *fictional gender*, is introduced to represent the attributes of these fictitious individuals.

#### *3.3.2 Multi VAE for two layered population generation (modeling-based multiple VAE) (MVAE-Pop2)*

The second approach consists of developing a specific model for each household size. Thus, when considering a maximum household size of $S = 5$, a total of five specialized VAEs is obtained. This method effectively overcomes the challenges associated with variations in household size. In the following discussion, we will present the inputs of the VAEs specialized in generating households of size 1 (Eq. 3) and of size $s$ (Eq. 4). The number of household attributes equals $k_h - 1$ because household size does not need to be taken into account.

For one individual:

$$X = [\underbrace{x_1, x_2, ..., x_{k_h - 1}}_{houshold}, \underbrace{x_1^1, x_2^1, ..., x_{k_i}^1}_{individual_1}] \quad (3)$$

For a household with $s$ individuals:

$$X = \left( \underbrace{x_1, x_2, ..., x_{k_h - 1}}_{houshold}, \underbrace{x_1^1, x_2^1, ..., x_{k_i}^1}_{individual_1}, \cdots, \underbrace{x_1^s, x_2^s ... x_{k_i}^s}_{individual_s} \right) \quad (4)$$

## 4 Experiments

### 4.1 Case study dataset

The training data for this study stem from the household travel survey conducted in France's Loire-Atlantique Département (a French administrative division similar to a county) (Département-Loire-Atlantique 2015). These data, which are freely accessible, include both individual characteristics and household attributes, in acknowledging that a single household can comprise multiple individuals. This type of study is often carried out at the national scale in order to analyze transportation demand,

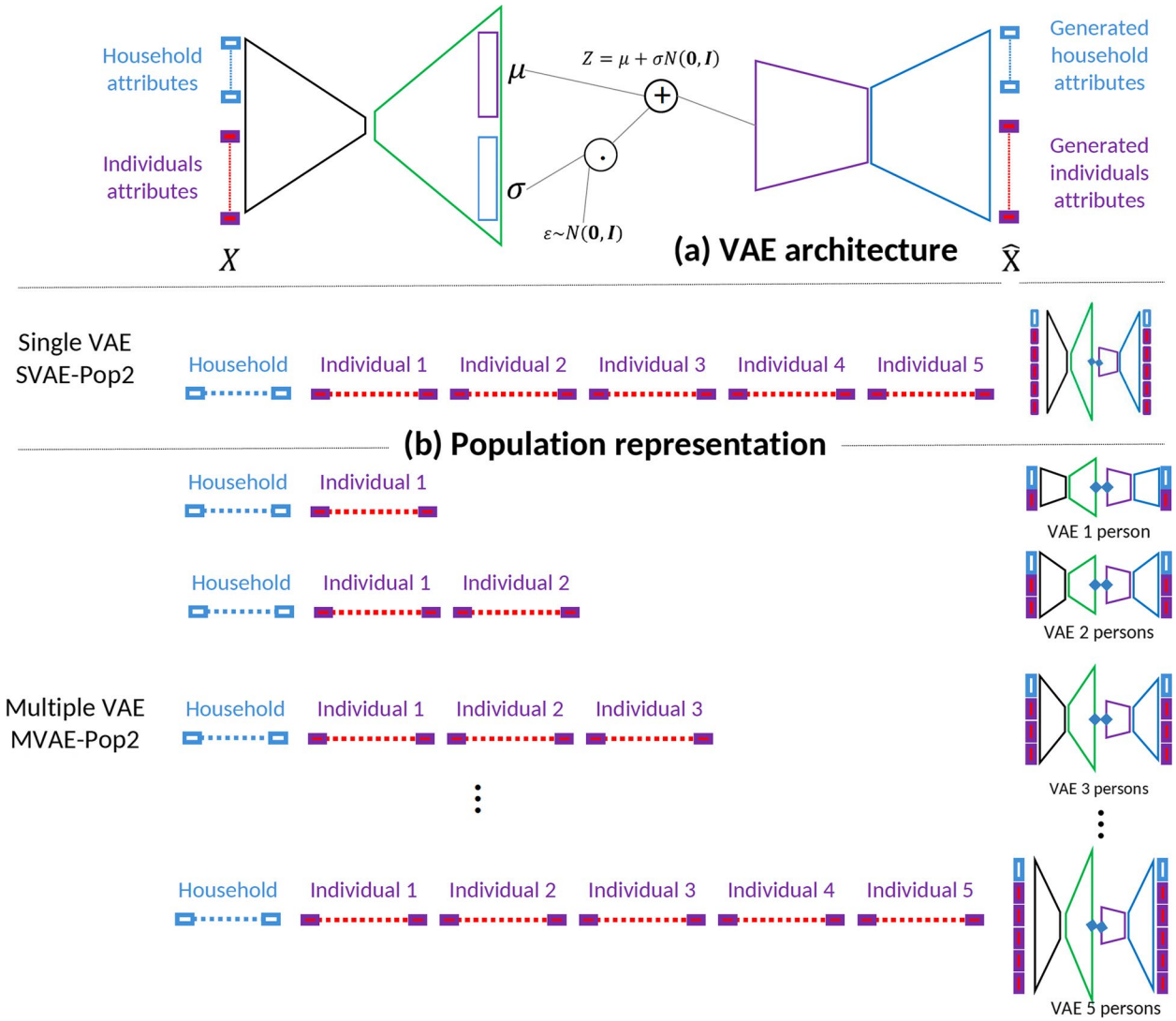


**Fig. 2** **a** General architecture of a VAE for population modeling with two layers (Household, Individuals) fo features; **b** Inputs for SVAE and MVAE representations

hence making these data representative of transportation demand datasets.

For this case study, the maximum household size is set at $S = 5$, which covers 98.5% of all households, corresponding to 12,575 households and 29,496 individuals.

The variables selected for this analysis, as presented in Table 1 below, pertain to both individuals and their households.

Some originally numerical attributes were recoded as categorical, resulting in a dataset consisting entirely of categorical attributes. This decision was inspired by the findings (Sané et al. 2025; Garrido et al. 2020; Johnsen et al. 2022), which show that Variational Autoencoders (VAEs) perform better on categorical attributes than numerical ones. The rationale behind this transformation is justified in more detail in the section discussion.

Moreover, 70% of the data are used to train the models, while the remaining 30% serve as the test set. The test set, or reference population, will be referred to throughout the remainder of this paper.

### 4.2 Model building

#### *4.2.1 Identification of hyperparameters*

To determine the optimal architecture for SVAE-Pop2, a range of hyperparameters has been explored. Table 2 provides an overview of the hyperparameter space examined to identify the most effective models. This table lists value ranges for several key hyperparameters, including

**Table 1** Overview of household and individual features

| Groups | Attributes | Type[a] | Description |
|---|---|---|---|
| Household (HH) | | | |
| | house_type | cat(5) | House type |
| | house_occupation_type | cat(4) | House occupation type |
| | household_size | cat(5) | Number of persons in the HH |
| | number_of_vehicles | cat(5) | Number of vehicles in the HH |
| | number_of_bikes | cat(4) | Number of bicycles in the HH |
| | has_internet | cat(2) | Has Internet access? |
| Individuals | | | |
| | link_ref_person | cat(5) | Link with the HH reference person |
| | age | cat(6) | Age |
| | sex | cat(2) | Gender |
| | is_adolescent | cat(2) | Is an adolescent? |
| | school_level | cat(5) | Educational level achieved |
| | employed | cat(2) | Is employed? |
| | studies | cat(2) | Is a student? |
| | socioprofessional_class | cat(9) | Socioprofessional status |
| | has_license | cat(2) | Holds a driver's license? |

[a] All attributes are categorical. The number in parentheses reflects the number of possible values

**Table 2** Models' Hyper-Parameters Overview

| Layers | Nodes | Latent dim | Epochs | Batch size | $\beta$ | Activation |
|---|---|---|---|---|---|---|
| 1–3 | 50–100 | 10–40 | 200 | 256 | 0.0005–0.01 | sigmoid, tanh |

the number of layers, the number of neurons per layer, the latent dimension, the number of epochs, the batch size, the $\beta$ value, and the activation function. In all, 210 models were implemented for the SVAE-Pop2 architecture.

This same hyperparameter optimization procedure was then applied to each VAE within the MVAE-Pop2 model.

#### *4.2.2 Comparison metrics*

To evaluate model performance, we have relied on three metrics to measure the distance between two agents: $\pi$, the modality vector of the true agent, and $\hat{\pi}$, the vector predicted by the model. These metrics are: the mean absolute error (*MAE*), the correlation coefficient (*corr*), and the standardized root mean square error (*SRMSE*).

The mean absolute error (*MAE*) (Eq. 5) quantifies the average magnitude of errors in a set of predictions, without considering their direction; it is calculated as the average of the absolute differences between predicted and actual values. The MAE is expressed in the same units as the original data, making it easy to interpret.

$$MAE = \frac{1}{N_{all}} \sum_{l} |\pi_l - \hat{\pi}_l| \tag{5}$$

The correlation coefficient (Eq. 6) measures the degree of fit between the generated data ($\hat{\pi}$) and the test data ($\pi$) by assessing the correlation between their cross-modal distributions.

$$corr = \frac{cov(\pi, \hat{\pi})}{sd(\pi) \times sd(\hat{\pi})} \tag{6}$$

The third metric, the standardized root mean square error (SRMSE) (Eq. 7), quantifies the root mean square error (RMSE) between the generated and real cross-modal frequencies, as normalized by the mean of the real frequencies. This metric is particularly useful when errors are weighted exponentially based on their magnitude. A lower SRMSE indicates a more accurate model.

$$SRMSE = \frac{\sqrt{\frac{1}{N_{all}} \sum_l (\pi_l - \hat{\pi}_l)^2}}{\frac{1}{N_{all}} \sum \pi_l} \tag{7}$$

The SRMSE was used to compare architectures because it is the most robust metric and thus the most commonly employed for this purpose (Sané et al. 2025). Other metrics were also used to interpret the results of the methodologies applied to the representative case study.

After optimizing the hyperparameters, the best-performing architectures based on SRMSE for both modeling approaches (SVAE-Pop2 and MVAE-Pop2) are presented in Table 3. The optimal SVAE-Pop2 model follows an architecture of $245 \times 80 \times 40 \times 80 \times 245$, with a $\beta$ value of 0.005 and a *sigmoid* activation function. This model consists of an encoder and a decoder, each composed of three layers. The encoder includes an input layer of size 245, a single hidden layer of size 80, and a latent space layer of size 40. The decoder mirrors the encoder's structure, with three layers of sizes $40 \times 80 \times 245$. For the MVAE-Pop2 modeling approach, which employs a specialized model for each household size, the optimal architecture varies accordingly. For instance, the best-performing model for households of size 4 adopts an architecture of $176 \times 100 \times 50 \times 10 \times 50 \times 100 \times 176$ with a $\beta$ value of 0.005 and a *tanh* activation function.

#### *4.2.3 Technologies introduced*

The data processing performed for the purpose of model evaluation was carried out using Python. The primary libraries used include: Pandas 1.5, Numpy 1.23, and TensorFlow 2.10. Graphs were created with the Plotnine 0.10 library. The source code and data link are both available in the appendix section titled "Data Availability and Computer Code".

## 5 Results

Once the optimal architectures have been determined at the end of the training phase, a synthetic population is generated for each approach (SVAE-Pop2 and MVAE-Pop2). The total number of generated entities amounts to approximately 1.4 M individuals distributed among 650 k households, which corresponds to the size of the population in the study area. These synthetic populations served as the basis for evaluating and comparing the ability of each model to faithfully reproduce the characteristics of the actual population, with these characteristics assumed to be the same as those of the reference population described in the previous section.

This section provides a detailed analysis of the results obtained, in highlighting the respective performance of each approach, their differences and their practical implications. This evaluation focuses on: individual characteristics, household-specific attributes, and the overall coherence between these two levels of information.

### 5.1 Performance on the individual attributes

The evaluation of how well both approaches perform on individual attributes was carried out in two steps. We first analyzed the marginal distributions of each attribute, then the joint distribution of these attributes was examined.

Figure 3 presents the marginal distributions of the individual attributes. The orange and blue bars correspond to the synthetic populations generated by the SVAE-Pop2 and MVAE-Pop2 approaches, respectively, while the green bars represent the test data used as a reference. Overall, the marginal distributions of the synthetic populations align well with those of the test data, although some discrepancies do remain. The mean absolute error is 3.5 percentage points for SVAE-Pop2 and 2.9 percentage points for MVAE-Pop2, with standard deviations of 2.9 and 2.7, respectively. Moreover, 75% of the modalities exhibit an absolute error below 6 percentage points for SVAE-Pop2 and below 4.9 percentage points for MVAE-Pop2. The maximum errors observed reach 8.6 percentage points for SVAE-Pop2 (for the attribute *studies*) and 8.8 percentage points for MVAE-Pop2 (for the attribute *has_license*).

The adequacy of the marginal distributions is further confirmed by Fig. 4, which compares the marginal distributions of the test data (on the x-axis) with those of the synthetic populations (on the y-axis). The points are generally aligned along the line with a slope of 1, thus indicating a good level of correspondence between the distributions. However, the SVAE-Pop2 approach (in orange) exhibits a slightly higher dispersion compared to MVAE-Pop2 (in blue), which is corroborated by the SRMSE values: 0.17 for SVAE-Pop2 vs. 0.15 for

**Table 3** Overview of the best-performing models

| Model | Input size | Architecture | $\beta$ | Activation |
|---|---|---|---|---|
| SVAE-Pop2 | 245 | $245 \times 80 \times 40 \times 80 \times 245$ | 0.005 | sigmoid |
| MVAE-Pop2 | | | | |
| $size = 1$ | 59 | $59 \times 100 \times 40 \times 100 \times 59$ | 0.0005 | sigmoid |
| $size = 2$ | 98 | $98 \times 50 \times 20 \times 50 \times 98$ | 0.001 | sigmoid |
| $size = 3$ | 127 | $127 \times 80 \times 20 \times 80 \times 127$ | 0.005 | tanh |
| $size = 4$ | 176 | $176 \times 100 \times 50 \times 10 \times 50 \times 100 \times 176$ | 0.005 | tanh |
| $size = 5$ | 215 | $215 \times 100 \times 20 \times 100 \times 215$ | 0.0005 | tanh |

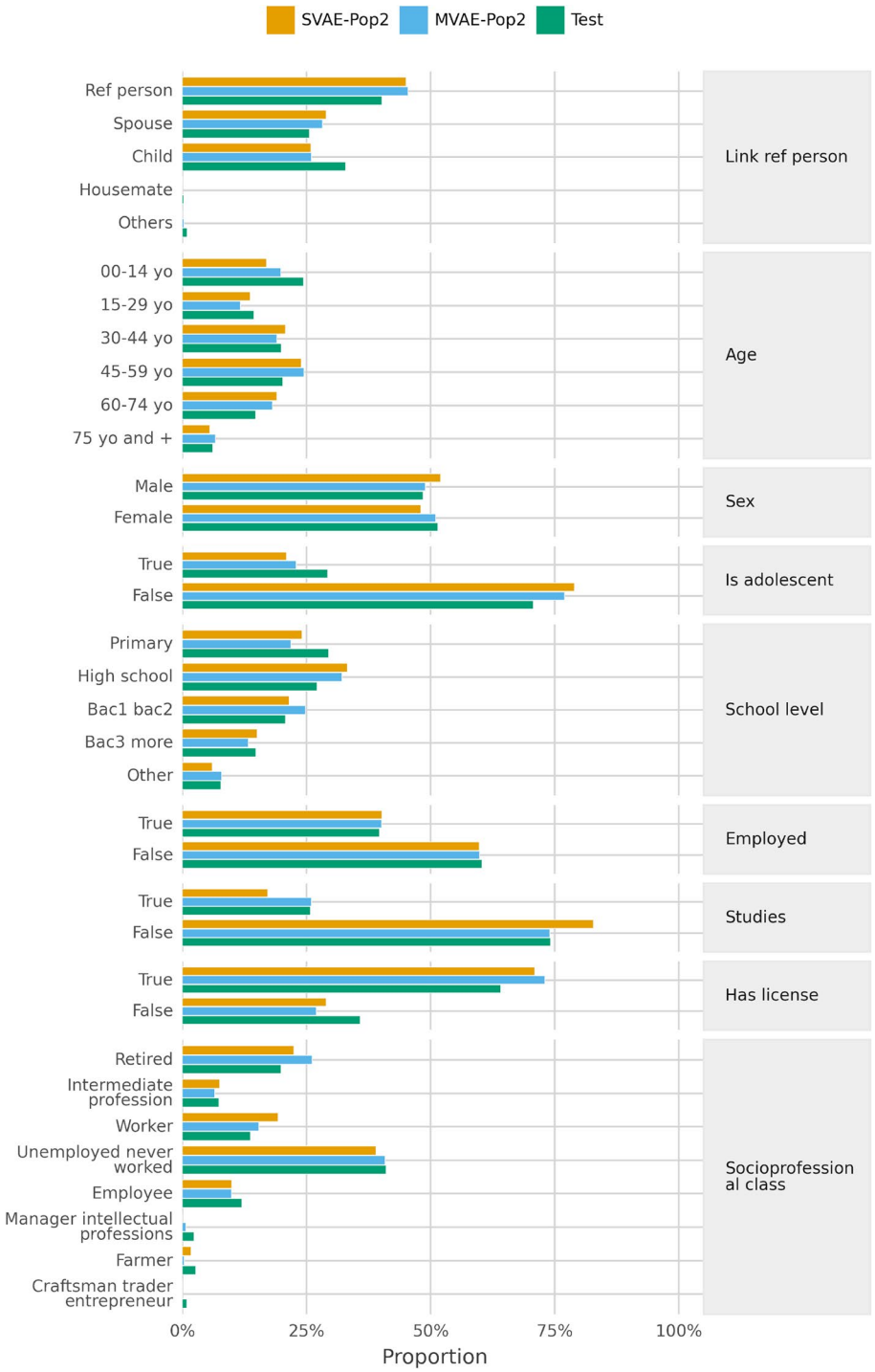


**Fig. 3** Marginal distributions for the true population (green) and samples generated using VAE (orange and blue) for individual attributes

MVAE-Pop2. The linear correlation coefficient is identical for both approaches, at 0.98.

The second evaluation focuses on comparing the joint distributions of individual attributes. For this purpose, we selected five attributes frequently used in mobility studies: *link_ref_person, age, sex, has_license, socioprofessional_class* (Bouzouina et al. 2021; Plaut 2006; Sun and Zacharias 2021; Brückmann et al. 2021; Wang and Cao 2017; Wang et al. 2022). Figure 5 illustrates this comparison; it can be observed that the MVAE-Pop2 approach

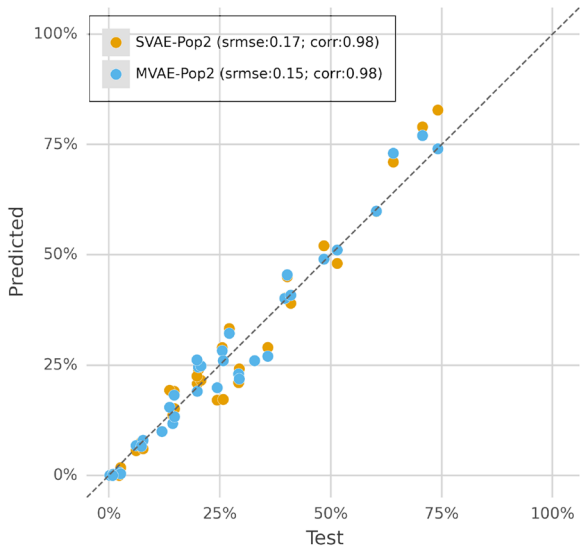


**Fig. 4** Goodness of fit of the marginal distributions of generated individual data to the test values

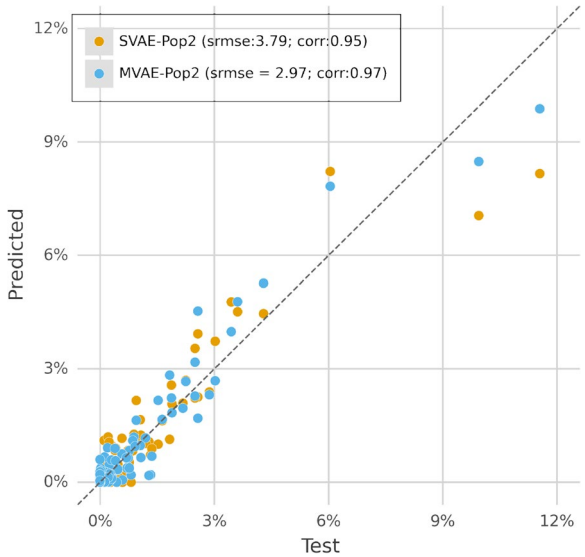


**Fig. 5** Multivariate comparison between predicted and test data for individual attributes (link_ref_person, age, sex, has_license, socioprofessional_class)

(in blue) exhibits a lower dispersion compared to SVAE-Pop2 (in orange), with an SRMSE of 3.79 for SVAE-Pop2 versus 2.97 for MVAE-Pop2. Furthermore, the linear correlation coefficient equals 0.95 for SVAE-Pop2 and 0.97 for MVAE-Pop2, thus confirming a higher fidelity of the joint distributions with the latter approach.

### 5.2 Performance on the household attributes

The evaluation of how well both approaches perform on household attributes relied on a methodology similar to that used for individual attributes. Figures 6 and 7 respectively present the marginal distributions and the univariate comparison between the test data and the synthetic populations generated by the two approaches.

The performance metrics indicate a mean absolute error of 4.2 for SVAE-Pop2 and 3.4 for MVAE-Pop2, with standard deviations of 3.3 and 3.7, respectively. The third quartile of absolute errors is 5.5 for SVAE-Pop2 and 4.1 for MVAE-Pop2, meaning that 75% of the absolute errors lie below these values. The maximum errors observed are 12 for SVAE-Pop2 and 12.9 for MVAE-Pop2, reaching their peaks on the attribute *number_of_vehicles* in both cases.

Figure 7 illustrates this comparison in the form of a scatter plot. A nearly similar dispersion is observed for both approaches, although MVAE-Pop2 (in blue) appears to exhibit a slightly lower dispersion than SVAE-Pop2 (in orange). The SRMSE values are 0.22 for SVAE-Pop2 and 0.21 for MVAE-Pop2, while the linear correlation coefficients are 0.98 and 0.99, respectively.

Regarding the joint household attributes, Fig. 8 compares the distributions of the attributes *number_of_vehicles* and *has Internet*. For this analysis, we selected the variables *number of vehicles* and *has Internet*, which are recognized in the literature as household factors influencing mobility (Bouzouina et al. 2021; Plaut 2006; Sun and Zacharias 2021; Brückmann et al. 2021; Wang and Cao 2017; Wang et al. 2022). In this case, MVAE-Pop2 (in blue) shows a higher dispersion than SVAE-Pop2 (in orange), with an SRMSE of 0.52 compared to 0.45. However, the linear correlation coefficient is identical for both approaches, at 0.97.

In order to analyze this relatively lower performance of the MVAE, we also evaluated the performance of each sub-model within the MVAE framework on household-level attributes. Table 4 presents these results along with the training sample size for each household size. The findings indicate that sub-models trained on smaller samples perform less accurately (compare the sub-models for Household size 3,5 to the sub-models for Household size 1,2,4). In other words, the performance of each sub-model tends to improve as the size of its training dataset increases. This explains that the MVAE performs slightly worse than the SVAE on the Household attributes.

### 5.3 Performance on the joint household-individual cross-modalities

The final stage of the evaluation involves comparing the joint distributions that integrate both household and individual attributes. For this purpose, we selected the attributes *household size, age and sex*, which are common to both levels of information. The cross-analysis of these attributes allows us to assess

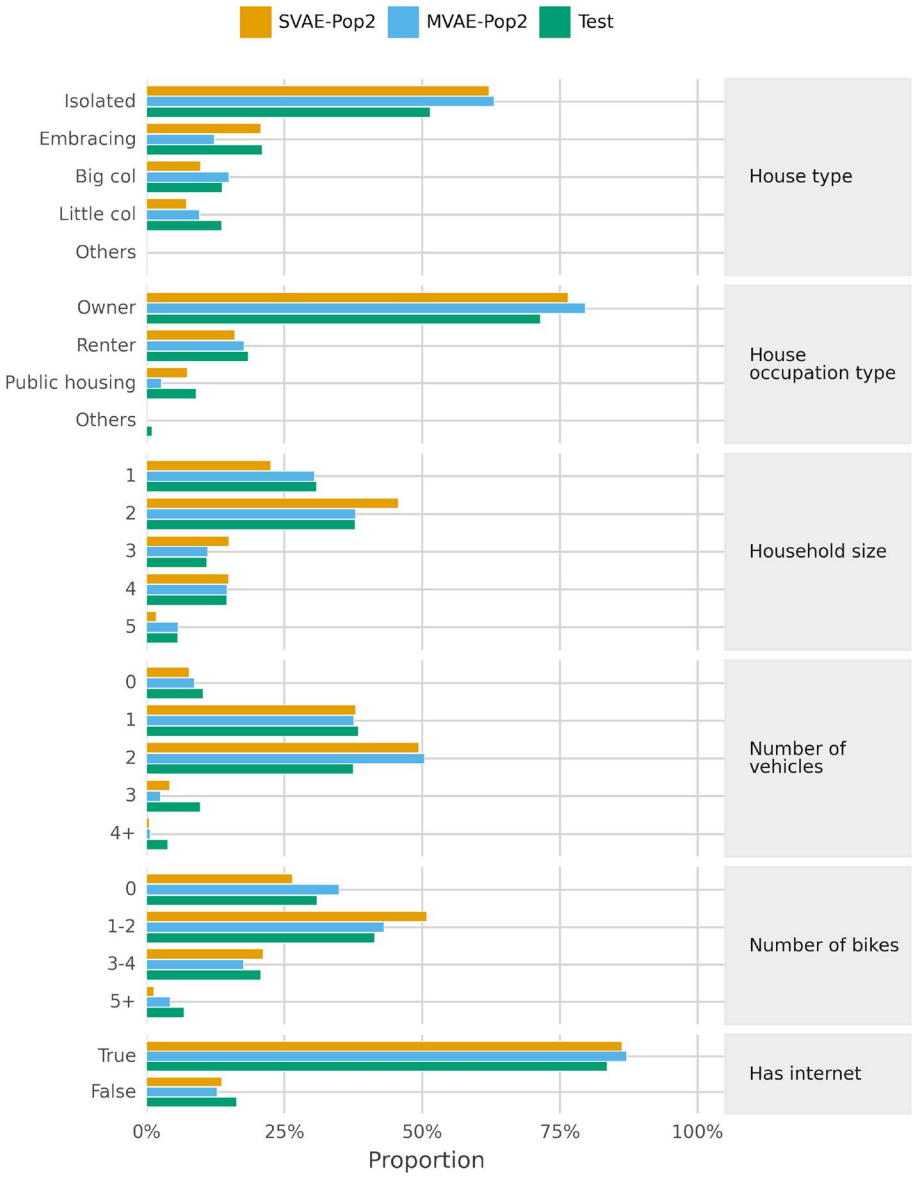


**Fig. 6** Marginal distributions for the true population (green) and samples generated using VAE (orange and blue) for household attributes

the coherence between household characteristics and individual features by examining the gender distribution across different age groups and household sizes.

Figure 9 illustrates this comparison. It can be observed that the MVAE-Pop2 approach (in blue) exhibits a lower dispersion than SVAE-Pop2 (in orange), thus indicating a better fidelity of the joint distributions. The SRMSE values are 0.49 for SVAE-Pop2 and 0.36 for MVAE-Pop2, while the linear correlation coefficients equal 0.93 and 0.96, respectively.

## 6 Discussion

The main contribution of this work lies in its ability to fill a gap in the literature regarding the generation of two-level synthetic populations. By proposing two innovative approaches, SVAE-Pop2 and MVAE-Pop2, we have demonstrated that these methods are capable of generating synthetic populations that faithfully replicate the test data at both the individual and household levels. The experimental results indicate that, although both approaches deliver a convincing performance, the

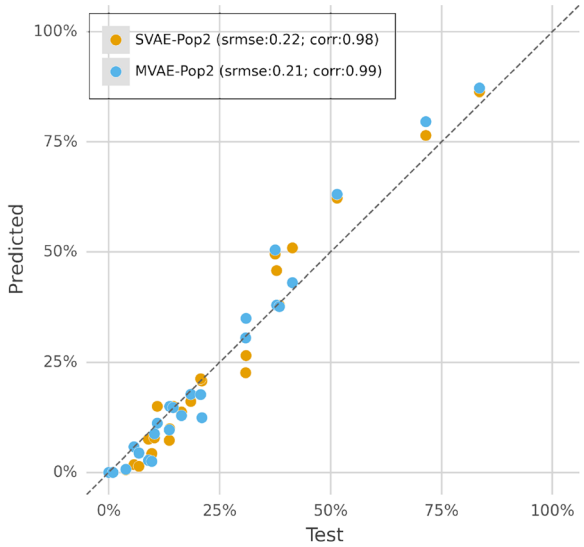


**Fig. 7** Goodness of fit of the marginal distributions of generated household data to the test values

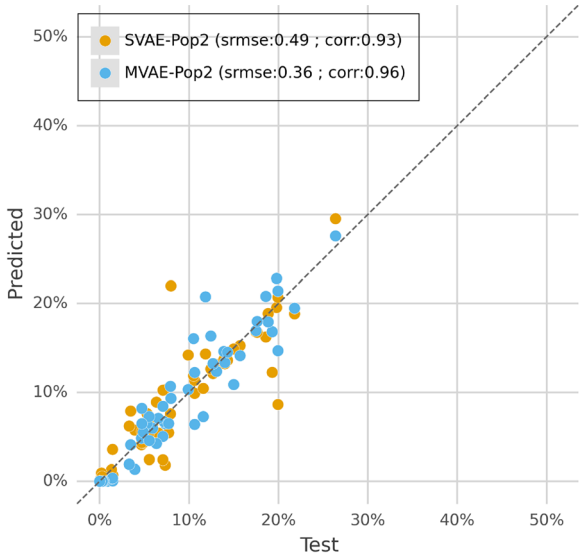


**Fig. 9** Combined multivariate comparison between predicted and test data (household_size, age, sex)

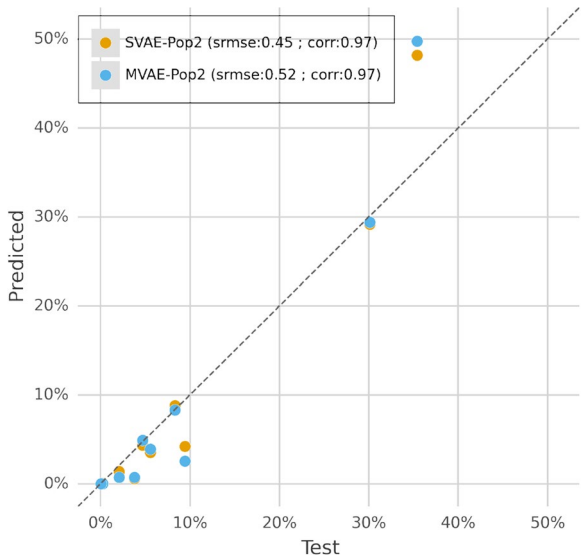


**Fig. 8** Multivariate comparison between predicted and test data for household attributes (number of vehicles, has Internet)

**Table 4** Overview of how the MVAE sub-models performed on household attributes

| Household size | N[a] | SRMSE | Corr |
|---|---|---|---|
| 1 | 3058 | 0.330 | 0.985 |
| 2 | 3822 | 0.623 | 0.979 |
| 3 | 1122 | 1.218 | 0.911 |
| 4 | 1478 | 0.666 | 0.983 |
| 5 | 580 | 1.564 | 0.900 |

[a] Size of the training sample

MVAE-Pop2 model yields slightly superior results, particularly in terms of the accuracy of the joint attribute distributions.

However, these approaches also present certain limitations. The MVAE-Pop2 model, which relies on a dedicated model for each household size, requires a sufficiently large sample for each of these sizes. This requirement can be problematic for underrepresented household sizes, which may explain the less satisfactory performance for certain household attributes compared to SVAE-Pop2. Moreover, the MVAE-Pop2 approach necessitates additional information, such as household size, for its proper application.

Conversely, the SVAE-Pop2 approach offers the advantage of employing a single model, thus leveraging the entire dataset during training regardless of household size. However, to ensure a fixed-size input, this method resorts to padding incomplete households with fictitious individuals. While this strategy is practical, it may introduce biases, particularly when the sample size is limited, thereby affecting the model's robustness and generalization capacity. This explains the less satisfactory performance for individual attributes compared to MVAE-Pop2.

The choice between these two approaches should therefore be guided by data availability. If the sample is sufficiently rich for each household size, then the MVAE-Pop2 approach appears to be more suitable, especially when a large number of attributes must be generated, i.e. necessitating a substantial number of individuals per household size. On the other hand, in

contexts where data are limited by household size, the SVAE-Pop2 approach provides a more relevant alternative since it allows inputting the entire dataset without the need for partitioning by household size.

Although both approaches show promising results, their application will ultimately depend on the specific context and availability of data. Future research could explore adaptive strategies or hybrid combinations of these methods to further optimize the quality of the generated synthetic populations while mitigating the limitations identified in this study.

Regarding the categorization of originally numerical attributes, this choice was motivated not only by the findings of previous studies (Sané et al. 2025; Aemmer and MacKenzie 2022; Johnsen et al. 2022; Garrido et al. 2020), which consistently show that Variational Autoencoders (VAEs) tend to perform better on categorical data in synthetic population generation tasks, but also by broader technical considerations. Specifically, handling mixed data types with VAEs requires a heterogeneous reconstruction loss—typically combining binary cross-entropy for categorical variables, mean squared error for numerical variables, and the Kullback-Leibler divergence for latent space regularization. These loss components operate on different scales, which can complicate optimization, destabilize training, and hinder convergence. By using a fully categorical representation, we ensure a homogeneous loss formulation that simplifies optimization and enhances model robustness, particularly important in our two-level (household–individual) architecture. Moreover, in the context of transportation modeling, discretized variables (such as age groups or income brackets) are common practice and align with the level of granularity typically required for applications such as travel demand estimation or scenario-based planning. While this approach involves a trade-off between precision and training stability, it does not compromise the realism or policy relevance of the generated synthetic populations. We acknowledge this simplification and consider future extensions to mixed-type VAEs as a promising direction.

The architectures presented enable extending a sample to cover the entire population. However, users often also have access to aggregated data. We therefore recommend generating the synthetic population using the proposed architectures and then applying an algorithm such as Iterative Proportional Updating (IPU) to adjust the population. An even better solution would be to integrate the aggregated data directly into the synthetic population generation process, so that the VAE can take this information into account.

## 7 Conclusion and future work

This paper has presented two methods for generating individuals within households using VAEs: the SVAE-Pop2 method, based on a single VAE, and the MVAE-Pop2 method, which employs a dedicated VAE for each household size. In a representative case study, both architectures produced a synthetic population consistent with the actual population, with a slight advantage observed for the MVAE-Pop2 approach.

SVAE-Pop2 is a unified generic solution when data are limited or unevenly distributed across household sizes. In contrast, MVAE-Pop2 proves more effective in settings with large, well-balanced samples, especially when modeling complex intra-household dependencies. These differentiated strengths make each approach suitable for distinct application scenarios.

The proposed methods are particularly well-suited for real-world applications, especially in domains such as urban traffic modeling and mobility analysis. By explicitly modeling the interdependencies between individuals within the same household, our methods allow for a more holistic representation of household behavior. This is essential in transportation modeling, where the activities of one household member often influence those of others (e.g., joint trips, school runs, coordinated commuting or car-sharing within the household). Furthermore, these methods are highly relevant in contexts where only limited data are available, such as household travel surveys. They enable the generation of large, realistic populations from relatively small samples, thus providing a powerful tool for scaling up microdata to the population level while preserving critical structural properties.

To promote reproducibility and practical use, the source code is publicly available and well-documented, facilitating adaptation to a variety of datasets and use cases.

The use of VAEs in synthetic population generation has thus reached a new milestone, evolving from the generation of individuals or households to the generation of individuals within households. This constitutes a critical step toward producing synthetic populations that more closely mirror real-world social structures and mobility patterns. The next step will consist of incorporating aggregated data into the VAE generation process, thereby further enhancing the fidelity and robustness of the synthetic populations.

**Acknowledgements**
Not applicable.

**Data availability and computer code**
The data used in this study have been drawn from the open-source Household Travel Survey (HTS) conducted in Nantes, France. The data and Python code developed for this work are available at https://osf.io/p3gt9/?view_only=8d42294e9dc34ea4b8ef5c6bdceba9b4. Special attention has been given to the modularity of the code and the clarity of the comments, hence ensuring that users can easily adapt the code to suit their needs.

**Authors' contributions**
Conceptualization, Abdoul Razac Sané; Methodology, Abdoul Razac Sané, Rachid Belaroussi, Pierre-Olivier Vandanjon, Pierre Hankach; Software, Abdoul Razac Sané; Validation, Abdoul Razac Sané, Rachid Belaroussi, Pierre-Olivier Vandanjon, Pierre Hankach; Formal analysis, Abdoul Razac Sané; Investigation, Abdoul Razac Sané; Writing—original draft preparation, Abdoul Razac Sané, Rachid Belaroussi, Pierre-Olivier Vandanjon, Pierre Hankach; Writing—review and editing, Abdoul Razac Sané, Rachid Belaroussi, Pierre-Olivier Vandanjon, Pierre Hankach.

**Funding**
Not applicable.

**Data availability**
The data used in this study have been drawn from the open-source Household Travel Survey (HTS) conducted in Nantes, France (https://data.loire-atlantique.fr/explore/dataset/224400028-enquete-deplacements-en-loire-atlantique/information). The data and Python code written for this work are available at https://osf.io/p3gt9/?view_only=8d42294e9dc34ea4b8ef5c6bdceba9b4.

## Declarations

**Ethics approval and consent to participate**
Not applicable.

**Competing interests**
The authors hereby declare no competing interests.

Received: 28 March 2025 Revised: 13 June 2025 Accepted: 29 June 2025
Published online: 08 July 2025

**References**

Aemmer, Z., & MacKenzie, D. (2022). Generative population synthesis for joint household and individual characteristics. *Computers, Environment and Urban Systems, 96*, 101852. https://doi.org/10.1016/j.compenvurbsys.2022.101852

Borysov, S. S., Rich, J., & Pereira, F. C. (2019). How to generate micro-agents? A deep generative modeling approach to population synthesis. *Transportation Research Part C: Emerging Technologies, 106*, 73–97. https://doi.org/10.1016/j.trc.2019.07.006

Borysov, S. S., Rich, J., & Pereira, F. C. (2019b). Scalable Population Synthesis with Deep Generative Modeling. *Transportation Research Part C: Emerging Technologies, 106*, 73–97. arXiv:1808.06910

Bouzouina, L., Baraklianos, I., Bonnel, P., & Aissaoui, H. (2021). Renters vs owners: The impact of accessibility on residential location choice. Evidence from lyon urban area, france (1999–2013). *Transport Policy, 109*, 72–84. https://doi.org/10.1016/j.tranpol.2021.05.022

Brückmann, G., Willibald, F., & Blanco, V. (2021). Battery electric vehicle adoption in regions without strong policies. *Transportation Research Part D: Transport and Environment, 90*, 102615.

Casati, D., Müller, K., Fourie, P. J., Erath, A., & Axhausen, K. W. (2015). Synthetic Population Generation by Combining a Hierarchical, Simulation-Based Approach with Reweighting by Generalized Raking. *Transportation Research Record: Journal of the Transportation Research Board, 2493*(1), 107–116. https://doi.org/10.3141/2493-12

De Rosa, G. H., & Papa, J. P. (2021). A survey on text generation using generative adversarial networks. *Pattern Recognition, 119*, 108098.

Département-Loire-Atlantique (2015). Enquête déplacements (2015) en Loire-Atlantique. https://data.loire-atlantique.fr/explore/dataset/224400028_enquete-deplacements-en-loire-atlantique/information/ Accessed 06 June 2023

Duda, R. O., Stork, D. G., & Hart, P. E. (2001). *Pattern classification* (2nd ed.). Wiley.

Fabrice Yaméogo, B., Gastineau, P., Hankach, P., & Vandanjon, P.-O. (2021). Comparing methods for generating a two-layered synthetic population. *Transportation research record, 2675*(1), 136–147.

Farooq, B., Bierlaire, M., Hurtubia, R., & Flötteröd, G. (2013). Simulation based population synthesis. *Transportation Research Part B: Methodological, 58*, 243–263. https://doi.org/10.1016/j.trb.2013.09.012

Garrido, S., Borysov, S. S., Pereira, F. C., & Rich, J. (2020). Prediction of rare feature combinations in population synthesis: Application of deep generative modelling. *Transportation Research Part C: Emerging Technologies, 120*, 102787. https://doi.org/10.1016/j.trc.2020.102787

Goeyvaerts, N., Santermans, E., Potter, G. E., Torneri, A., Kerckhove, K. V., Willem, L., Aerts, M., Beutels, P., & Hens, N. (2017). Household members do not contact each other at random: implications for infectious disease modelling. *Proceedings of the Royal Society B: Biological Sciences, 285*. https://doi.org/10.1101/220202

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Networks. arXiv:1406.2661. Publisher: arXiv Version Number: 1.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., & Lerchner, A. (2017). beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Sy2fzU9gl Accessed 09 May 2023

Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural computation, 18*(7), 1527–1554.

Johnsen, M., Brandt, O., Garrido, S., & Pereira, F. (2022). Population synthesis for urban resident modeling using deep generative models. *Neural Computing and Applications, 34*(6), 4677–4692. https://doi.org/10.1007/s00521-021-06622-2

Kong, J., Kim, J., & Bae, J. (2020). Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems, 33*, 17022–17033.

Plaut, P. O. (2006). The intra-household choices regarding commuting and housing. *Transportation Research Part A: Policy and Practice, 40*(7), 561–571. https://doi.org/10.1016/j.tra.2005.10.001

Qian, X., Gangwal, U., Dong, S., & Davidson, R. (2024). A Deep Generative Framework for Joint Households and Individuals Population Synthesis. arXiv:2407.01643

Sané, A. R., Vandanjon, P.-O., Belaroussi, R., & Hankach, P. (2025). A comprehensive investigation of variational auto-encoders for population synthesis. *Journal of Computational Social Science, 8*(1), 13. https://doi.org/10.1007/s42001-024-00332-0

Sun, L., Erath, A., & Cai, M. (2018). A hierarchical mixture modeling framework for population synthesis. *Transportation Research Part B: Methodological, 114*, 199–212. https://doi.org/10.1016/j.trb.2018.06.002

Sun, Z., & Zacharias, J. (2021). Do housing tenure and public transport provision matter in automobile use in bedroom suburban communities? evidence from beijing. *Journal of Housing and the Built Environment, 36*(1), 241–262.

Van den Oord, A., Kalchbrenner, N., Espeholt, L., Vinyals, O., Graves, A., & et al. (2016). Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems, 29*. https://proceedings.neurips.cc/paper_files/paper/2016/file/b1301141feffabac455e1f90a7de2054-Paper.pdf. Accessed 04 June 2024

Wang, D., & Cao, X. (2017). Impacts of the built environment on activity-travel behavior: Are there differences between public and private housing residents in hong kong? *Transportation Research Part A: Policy and Practice, 103*, 25–35. https://doi.org/10.1016/j.tra.2017.05.018

Wang, X., Yan, X., Zhao, X., & Cao, Z. (2022). Identifying latent shared mobility preference segments in low-income communities: Ride-hailing, fixed-route bus, and mobility-on-demand transit. *Travel Behaviour and Society, 26*, 134–142. https://doi.org/10.1016/j.tbs.2021.09.011

Yameogo, B. F., Vandanjon, P.-O., Gastineau, P., & Hankach, P. (2021). 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. https://doi.org/10.18564/jasss.4482

Zeng, X., Wang, F., Luo, Y., gu Kang, S., Tang, J., Lightstone, F. C., Fang, E. F., Cornell, W., Nussinov, R., & Cheng, F. (2022). Deep generative molecular design reshapes drug discovery. *Cell Reports Medicine, 3*(12), 100794. https://doi.org/10.1016/j.xcrm.2022.100794

## Publisher's Note