Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 00 (2026) 000–000

Transportation Research Procedia

www.elsevier.com/locate/procedia

World Conference on Transport Research - WCTR 2026 Toulouse 6-10 July 2026

THE FRAGILITY OF SYNTHETIC POPULATIONS OVER TIME: AN EVALUATION OF FORECASTING STRATEGIES FOR AGENT-BASED MODELS

Vianey Darsel[a,*], Etienne Côme[a], Jeppe Rich[b], Francisco Camara Pereira[b], Latifa Oukhellou[a]

[a]*COSYS-GRETTIA, Université Gustave Eiffel, Champs-sur-Marne, France*
[b]*Danmarks Tekniske Universitet, Kongens Lyngby, Denmark*

Abstract

Agent-based simulation models are widely used to assess the future mobility projects. However, their effectiveness hinges on the accuracy of synthetic populations, which are often generated using data from a single reference year. This study explores the temporal validity of these populations by evaluating their performance over multi-year datasets in a simulation-like environment. We focus on multi-dimensional agents, ensuring that all attributes critical to the simulation are comprehensively represented. To assess the quality of synthetic populations, we compare three generation methods: direct use of original training data, Bayesian Network sampling, and diffusion model. We then evaluate four agent-forecasting approaches: static projection (both idealistic and realistic), resampling, and dynamic projection. While dynamic projection holds potential, it requires complex longitudinal data to model the evolution of agent attributes. To address this, we introduce a heuristic model based on aggregated statistics, though its limitation lies in ignoring attribute correlations, which reduces its accuracy. Our results demonstrate that synthetic populations lose relevance over time, emphasizing the necessity of effective agent-forecasting. Among the methods tested, idealistic static projection yields the best performance but requires data that is difficult to obtain. The other methods show limited improvements, highlighting the ongoing challenges in accurately forecasting agent attributes.

Keywords: Agent-forecasting; Population Synthesis; Agent-based Simulation; Deep Generative Models

* Corresponding author. Tel.: +33.1.81.66.86.79.
E-mail address: vianey.darsel@univ-eiffel.fr

Nomenclature

${}^{y}\mathbf{X}$	Multi-dimensional distribution of the population for year y
${}^{y}X_k$	Distribution of the k-th attribute of ${}^{y}\mathbf{X}$
${}^{y}\mathbf{S}^n$	n realisations of a random variable following the distribution of ${}^{y}\mathbf{X}$
${}^{y}\mathbf{s}^i$	The i-th realisation of a random variable following the distribution of ${}^{y}\mathbf{X}$
${}^{y}\tilde{\mathbf{S}}$	A set of samples approximating a set of realisations from ${}^{y}\mathbf{X}$
N_y	Target size for ${}^{y}\tilde{\mathbf{S}}$
Ω_k	Sample space for ${}^{y}X_k$. It corresponds to all modalities that ${}^{y}X_k$ could take over all years.
ω_k	A value from Ω_k.
$f_k(\mathbf{S}, \omega_k)$	The observed frequency of the modality ω_k in the set $\mathbf{S}$.
${}^{y}u_j$	An updating rule in the dynamic projection model.
$g(\mathbf{S})$	Function associated to a generative model to generate a synthetic population with similar characteristic than the input population $\mathbf{S}$.

1. Introduction

Agent-based simulation models are powerful tools for estimating population flows within a geographic area. These models rely on a synthetic population that represents the total population, generated from limited datasets. While population synthesis has been extensively studied for reconstructing populations at a fixed point in time (typically the year of the original data), this article addresses a less explored challenge: agent-forecasting, or generating populations for future scenarios.

Agent-based simulations are particularly effective for analyzing mobility scenarios, whether based on current conditions or hypothetical settings. They play a critical role in the early stages of transportation projects, such as new transit lines, infrastructure development, or policy interventions, by enabling impact assessment and prediction. However, such projects often span decades: for example, Copenhagen's Inderhavnsbroen bridge took five years to complete[1], while the Grand Paris Express metro lines are expected to require 13 to 20 years[2]. This long-term nature underscores the need for robust agent-forecasting methods.

Agent-forecasting shares conceptual links with demographic forecasting, which projects future population trends at an aggregate level (Booth, 2006). While demographic forecasts provide macro-level insights (e.g., population size, age distribution, and geographic spread), agent-forecasting operates at a microscopic level, generating individual-level attributes. Although national-level forecasting methods are well-documented, small-area projections remain underdeveloped (Wilson et al., 2022). Agent-forecasting can bridge this gap by deriving aggregated statistics from synthetic populations, offering granularity for both public and private sectors to anticipate consumption patterns and policy needs (Mazzuco and Keilman, 2020).

Unlike population synthesis, agent-forecasting has received limited attention in the literature. There are three primary methodological approaches. The first two approaches (static projection and resampling) replicate samples from the initial dataset to satisfy statistical constraints for the target year, either on the marginals or on the distribution of selected attributes. These methods depend on the quality of the initial set, which becomes less relevant as time goes by. In this work, we investigate the impact of the interval between the training and testing data on the quality of the population, its representativeness, and its realism. The last method, dynamic projection, involves updating each individual's profile annually to reflect ageing and other changes. While longitudinal data at the individual-level would be ideal for capturing individual-level dynamics, such data is rarely available. Consequently, dynamic projection methods often rely on longitudinal data for one or a limited number of attributes to infer transition laws. In contrast, simulation mod-

[1] Inderhavnsbroen - Wikipedia
[2] Grand Paris Express - Wikipedia

els typically represent individuals through multiple interdependent attributes—for example, eight in Hörl and Balac (2021a). However, transition data for these attributes are generally scarce, which constrains dynamic projections to a reduced set of attributes—for instance, three in Ballas et al. (2005a), Namazi-Rad et al. (2014), and Kukic et al. (2023). To address this limitation, we propose a dynamic projection approach based on aggregated data, which is typically more readily available. This approach is admittedly simplistic, as it does not capture complex interdependencies and is subject to stochastic variation. It is therefore not expected to achieve high efficiency. Its primary purpose is to enable a comprehensive comparison with existing methods and to assess the extent of performance achievable under such simplified conditions.

The quality of agent-forecasting also depends on the initial synthetic population. In population synthesis, several methods have been proposed to generate multi-dimensional individuals, using probabilistic models (Farooq et al., 2013; Sun and Erath, 2015), or Deep Generative Models (DGMs) (Borysov et al., 2019; Garrido et al., 2020; Darsel et al., 2025a). An asset of these methods is their ability to create samples out of the training data, but also to generate an infinite pool of individuals, which can be useful for methods that copy individuals from the initial dataset. In this work, we investigate the impact of using a population generated by a Bayesian Network, a probabilistic model, and by a diffusion model, a DGM, on the forecasted population.

Furthermore, most existing work on agent-forecasting represent individuals using only a limited set of attributes, despite the fact that simulations are capable of handling higher-dimensional representations, which can improve the simulation's quality. In this work, we examine how the different forecasting methods perform when the number of attributes increases. Moreover, previous research has typically evaluated generated populations on data from only one or two years, neglecting to investigate the evolution of the quality of the population over time. In this paper, we explore the performance of the different methods over multiple years.

This article addresses the key gaps identified in the existing literature by:

- Assessing the reliance of synthetic populations from various sources (training data or generated by generative models) and evaluating their quality over multiple years.
- Proposing a method for dynamic projection when the transitions between states are not available, based on aggregated statistics.
- Evaluating the different methods (static projection, resampling, and dynamic projection) for agent-forecasting on multi-dimensional individuals over years.
- Comparing generative methods (Bayesian network, and diffusion model) for initializing synthetic populations and generating diverse sample pools in agent-forecasting.

In Section 2, we present a literature review on agent-forecasting. The different projection methods are then derived in Section 3, with a closer attention to dynamic projection, for which we derive a new model. Then, in Section 4 we describe our experiments framework, including the data used for the experiments, and the evaluation methodology. We present the results on evaluating population synthesis solutions over time and on comparing agent-forecasting methods. Some discussions about the results are given in Section 5.

2. Literature Review

Agent-forecasting methods can be categorized into three distinct approaches, discussed below. These approaches can be combined into hybrid models, which we address in the last subsection.

2.1. Static projection

Static models rely on a microdata sample of the population and adjust the weights of these samples to align with specific macroscopic parameters or dynamics for the targeted population. They generate populations directly from the initial dataset, without simulating intermediate steps or evolutionary processes. Ballas et al. (2005b) explain that an advantage of this method is that it can allow to simulate "what-if" scenarios, thanks to the ability of static methods by imposing specific characteristics on the generated population.

Ballas et al. (2006) first generate a population with a population synthesis re-weighting technique, for the area of Leeds. Then, different scenarios on closing a factory in the area are tested, and modifications in the household attributes are directly applied. The impact on simulation is then analyzed with this modified population. Re-weighting is also performed by Harding et al. (2011), who use the population projections in age, sex, and employment status together to re-weight the initial microsamples. Harding et al. (2011) and Tanton and Edwards (2012) explain that re-weighting methods are only suitable for short term forecasts due to lack of long-term data that are mandatory to get accurate forecasts on the macro indicators. However, Lomax et al. (2022) explain how a re-weighting method can be efficient to test multiple long-term scenarios. Lomax et al. (2022) adjust the weights based on the sex, age, and ethnicity of the population.

2.2. *Resampling*

Static projection can be computationally expensive, as it requires generating a new population for each year. To address this, Prédhumeau and Manley (2023) propose to adapt it to a resampling method, which introduces some dynamic elements into the process. For the initial year, the population is reconstructed using the available microsample data. Subsequently, the population is updated annually by duplicating the population from the preceding year and then selectively copying or removing individuals to satisfy specific constraints. Unlike static projection, this method focuses on a subset of attributes rather than requiring all marginals. In Prédhumeau and Manley (2023), the joint distribution of age, location and gender is used.

There are two main motivations for this method. Firstly, it avoids the computational cost of static projection that requires a model to be learnt for each year. Secondly, it can handle a larger number of attributes, as the population is updated based on a selection of attributes only – 7 attributes in Prédhumeau and Manley (2023). However, it assumes that the correlations between individual attributes do not change much over time, relying on the initial pool of individuals for reconstruction. Thus, Prédhumeau and Manley (2023) recommend this method only for short- and medium-term forecasts. By only copying individuals from the current population, there is a risk to decrease the diversity in the population, as some combinations of instances may disappear over time.

2.3. *Dynamic projection*

Dynamic models start by sampling a population with a traditional population synthesis model. Then, each agent, either the household or the individual, is updated yearly by considering life events (birth, death, migration...). This ageing process can be interpreted as an adaptation of the component-cohort method for the microscopic scale (van Imhoff and Post, 1998). Indeed, the component-cohort method models the evolution of a population at a macroscopic scale by considering demographic indicators (such as mortality or migration rates) for different sexes and age groups to update the total population at each time step. In dynamic agent-forecasting projection, this evolution is applied stochastically at the individual level, turning the rates into a probability. On average, the demographic indicators are respected.

Various demographic and socio-economic rules can be incorporated to update synthetic populations. For instance, van Imhoff and Post (1998) account for mortality, fertility, and changes in marital status (union formation, separation, marriage, and divorce), using forecasts from Statistics Netherlands, although their model is restricted to a relatively small population of around 10,000 individuals. In contrast, Vencatasawmy et al. (1999) developed SVERIGE, a comprehensive model for the entire Swedish population, structured into ten modules that simulate diverse life events, including migration, education, and employment. However, SVERIGE lacks a systematic evaluation of its outputs. Similarly, Ballas et al. (2005a) constructed a model for Ireland that integrates mortality, fertility, and internal migration. By explicitly modelling internal migration, their framework allows for the simulation of the entire national population, although it omits external migration flows. In this approach, internal migration is determined by the relative attractiveness of each county, which is assumed to be proportional to its population size. More recent contributions, such as Namazi-Rad et al. (2014) and Fatmi and Habib (2017), adopt a two-step strategy. They first reconstruct the full population using population synthesis and re-weighting methods, and subsequently simulate its temporal evolution. Importantly, these studies model dynamics at both the household and individual levels, thereby offering a more detailed representation of population change.

2.4. Hybrid approaches

Although the resampling may look attractive, it requires knowledge of the joint distribution, which may not be available annually. Therefore, Kukic et al. (2023) propose a hybrid approach, between resampling and dynamic projection. The method begins with an initial population, which is then updated using resampling when joint distribution data is available and dynamic projection otherwise. Unlike Prédhumeau and Manley (2023), this approach first ages the population by incrementing each individual's age by one year, before applying resampling.

When an additional microsample becomes available, the resampling technique can be adapted by incorporating samples from the new microsample instead of the initial dataset (Kukic and Bierlaire, 2025). In their work, Kukic and Bierlaire (2025) introduced several refinements to the resampling process. They employed a Monte Carlo Markov Chain sampler as proposed by Kukic et al. (2024), and they used the household size as a criterion for resampling. Additionally, they generated new individuals from the new microsample to match the desired marginal distributions. Excess samples are then removed to maintain the population size while preserving the required proportions. Since resampling allows to handle a higher number of attributes, Kukic and Bierlaire (2025) increase the number of attributes simulated to 8. To the best of our knowledge, in the literature, Kukic and Bierlaire (2025) is the only model that deals with microsamples from several years–two years are used in the training and one year for the evaluation, and it is also the only model with eight attributes which corresponds to the number of attributes generated by eqasim (Hörl and Balac, 2021b).

3. Methodology

Building on the work of Kukic and Bierlaire (2025), we aim to compare the ability of various methods to generate populations ready for simulation, i.e., with a sufficient number of attributes. Existing literature on this topic typically relies on datasets with limited temporal coverage, often using only a single year of data for evaluation. In this paper, we assess the ability of different projection methods to forecast population over multiple years. Our approach uses microsample from only one year, combined with aggregated statistics from subsequent years.

In this section, we describe the derivation of the different models for agent-forecasting. To that extent, we first introduce some notations that are similar for all models. Then, we present the attributes that we want to simulate in the models, as some of them appear in the derivation of the methods.

3.1. Notations

To derive mathematically the different models and the evaluation metrics, we introduce a couple of notations. Let ${}^{y}\mathbf{X} = ({}^{y}X_1, {}^{y}X_2, \ldots, {}^{y}X_d)$ denote the d-dimensional distribution of the population for year y, which we aim to approximate. A set of n realizations of ${}^{y}\mathbf{X}$ is represented by ${}^{y}\mathbf{S}^n$, where each sample ${}^{y}\mathbf{s}^i = \left({}^{y}s^i{}_1, {}^{y}s^i{}_2, \ldots, {}^{y}s^i{}_d\right)$ (for $i \in [1, n]$) corresponds to an individual realization. We denote ${}^{y}\tilde{\mathbf{S}}$ as a set of samples approximating a set of realisations from ${}^{y}\mathbf{X}$. In static projection and resampling, the size of the the generated set is controllable, and we note $N_y = |{}^{y}\tilde{\mathbf{S}}|$ the target size of this set. For $k \in [1, d]$, Ω_k corresponds to all modalities that the attribute ${}^{y}X_k$ could take over all possible years. This set is independent from the year. For a given modality $\omega_k \in \Omega_k$, $f_k(\mathbf{S}, \omega_k)$ denotes the frequency of ω_k in the set $\mathbf{S}$. For clarity, we re-index the years such that the initial year, i.e. the year where data is available, is set to $y = 0$.

3.2. Attributes

In our experiments, we consider eight attributes to describe an individual (three numerical and five categorical attributes). These attributes are presented in Table 1, and are in line with simulation (Hörl and Balac, 2021b). This set is bigger than most of the literature on agent-forecasting, but it is consistent with common practice in population synthesis, where handling eight or more attributes is standard.

Name	Type	Number of instances	min	max
Age	integer	120	0	120
Sex	binary	2		
Last diploma	category	5		
Number of persons in the household	integer	25	0	64
Type of professional activity	category	9		
Married	boolean	2		
Department	category	8		
Number of cars	integer	4	0	3

Table 1. Parameters for the dynamic projection for each attribute

3.3. Derivation of static projection (idealistic)

Static projection requires access to an initial set of realizations from the first year, denoted by ${}^0\mathbf{S}^n$, and some margins for the desired year:

$$\forall k \in [1, d], \forall \omega_k \in \Omega_k, \mathbb{P}({}^yX_k = \omega_k)$$

The generated population consists of samples drawn from the microsample, with each sample's occurrence calculated to respect the specified marginal distributions. Mathematically, for a given year y, the generated population can be expressed as:

$$\begin{aligned} {}^y\tilde{\mathbf{S}} &= \bigcup_{i=1}^{n} \left({}^0\mathbf{s}^i \text{ repeated } {}^yw^i \text{ times} \right) \\ \text{such that } &\begin{cases} \forall k \in [1, d], \forall \omega_k \in \Omega_k, f_k\left({}^y\tilde{\mathbf{S}}, \omega_k\right) = \mathbb{P}\left({}^yX_k = \omega_k\right) \\ |{}^y\tilde{\mathbf{S}}| = \sum_{i=1}^{n} {}^yw^i = N_y \end{cases} \end{aligned} \tag{1}$$

where ${}^yw^i \in \mathbb{N}$ corresponds to the occurrence of ${}^0\mathbf{s}^i$ in ${}^y\tilde{\mathbf{S}}$. The constraints specifies the margins and the size of the desired population.

To integrate the margins to a microsample, Iterative Proportional Fitting (IPF) is a common method in population synthesis (Beckman et al., 1996; Ballas et al., 2005a; Ye et al., 2009). IPF approximates the joint distribution from a microsample and the marginal distributions, making it directly applicable to the current problem. The aim is to estimate the true proportion of each sample in the microsample. Starting from the initial proportions, the weights are iteratively adjusted to align with the marginal constraints. This operation is performed until an error termination criterion is reached.

To improve the quality of the microsample and ensure precision, additional bins may be introduced for specific attributes during the IPF procedure. This adjustment mitigates potential bias caused by an insufficient number of samples in some instances. For example, in our experiments, we create age bins because the microsample contains very few individuals over 100 years old. Without these bins, some age groups might be underrepresented or missing from the microsample, even though they are relevant for certain years. To address this, we used 15-year age intervals.

To obtain the population with the desired marginals, the individuals are picked to fit the estimated proportion. When the desired count c is not an integer, stochastic rounding is applied (Gupta et al., 2015).

The described approach is idealized, as it assumes full knowledge of all marginal distributions—an impractical scenario in real-world applications. In practice, projections for these margins are often inaccurate and unavailable for many attributes. Nevertheless, we remain interested in assessing the potential performance of a method operating under such ideal conditions.

3.4. Derivation of static projection (realistic)

Since the method described in Section 3.3 lacks realism, we propose a more practical alternative based on the same underlying principle. Introducing artificial degradation to true values would make the results excessively sensitive to experimental conditions, so we avoid altering prediction accuracy. Instead, we only adjust the number of attributes considered.

In this approach, we fit the IPF using all attributes for which projections are available from the French National Institute of Statistics and Economic Studies (INSEE): sex, age, and department [3]. While this framework remains optimistic, by assuming access to true values, it is designed for practical implementation.

3.5. Derivation of the resampling approach

The resampling approach requires access to an initial solution ${}^0\mathbf{S}^n$, and some yearly constraints. In this paper, we will consider the same constraints as in Prédhumeau and Manley (2023), by resampling with the true joint distribution for the age, the sex and the location (${}^yX_{age,sex,loc}$). But unlike Prédhumeau and Manley (2023), we do not copy samples from the current population, but from an initial wide pool of individuals. This allows to solve the potential risk of vanishing some profiles, and keep diversity over time.

The initial solution can be obtained with any algorithm from population synthesis ${}^0\tilde{\mathbf{S}} = g\left({}^0\mathbf{S}^n\right)$, where g is the function associated to the algorithm, which enables to construct the full population from a microsample. Then for a new year y, we copy the population from the previous year: ${}^y\bar{\mathbf{S}} = {}^{y-1}\tilde{\mathbf{S}}$. For each combination of age, sex and location, the desired number of samples in the population is computed, and compared with the current count of samples with that combination. If the current count exceeds the desired number, some samples are randomly removed from the population. Otherwise, if the number of samples is below the desired number, samples are randomly drawn from the initial microsample.

We note that the processes are independent for each group of age, sex and location. Therefore, if we want to sample to one specific year, we can directly sample to the target year from year 0, with the same process.

As for static projection, this refitting algorithm could require using bins. We use the same bins as for the static projection.

While this method remains idealistic in assuming access to true values, it is realistic to assume access to INSEE's projections of the joint distribution for age, sex, and department[4]. These projections are the same data used in our realistic static projection method, but here we use the joint distribution for resampling—unlike the static method, which only considers the marginal distributions.

3.6. Derivation of dynamic projection

In the general case, dynamic projection requires access to an initial solution ${}^0\mathbf{S}^n$, and some updating rules ${}^yu_1, ...{}^yu_r$ for each year. As for the resampling, a population is reconstructed for the initial year ${}^0\tilde{\mathbf{S}} = g\left({}^0\mathbf{S}^n\right)$. From this initial population, individuals are aged each year. To construct the population for the year after ${}^{y+1}\mathbf{S}^i$, each updating rule is applied one after the other on each individual to get the individual for the next year:

$$ {}^{y+1}\mathbf{s}^i = {}^yu_r \circ ... \circ {}^yu_1({}^y\mathbf{s}^i). $$

The individual is kept in the population, based on mortality and migration rules. In our model, we omit the latter, because of the lack of data. We only consider internal migration.

[3] Age Pyramid - INSEE
[4] Age Pyramid - INSEE

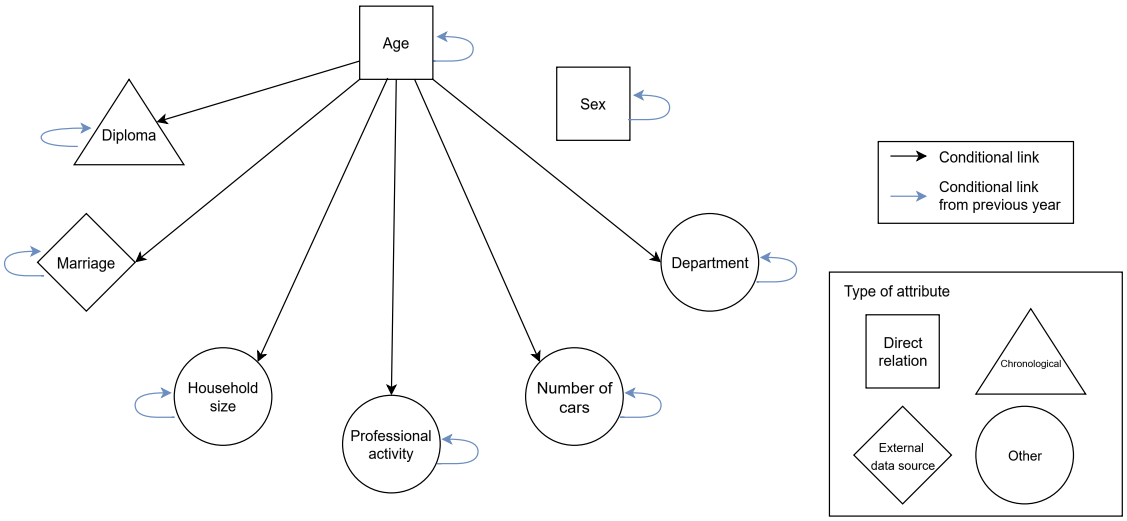

Fig. 1. Attributes used in our experiments with their corresponding type, and the conditional links used for the derivation

In our experiments, we do not benefit from longitudinal data for the full joint distribution, but we have a representation of the population for each year. Therefore, we construct heuristic rules in order to fit aggregated statistics. We distinguish four types of attributes:

- attributes with a direct relation with the year and the previous situation, e.g. age. This first class of attributes does not require any detailed explanations, as the derivation of the updating rule is straightforward.
- attributes that can be computed from external data sources, e.g. marriage rate. The rates are often given for each age. Therefore, we derive the updating rules conditionally to the age.
- attributes constrained by a chronological order between the instances, e.g. highest diploma.
- attributes with no or few constraints, e.g. professional activity or the department.

The latter two require more attention because of the lack of longitudinal or additional data. Figure 1 presents the categorization of attributes used in our experiments, along with the conditional dependencies applied. In our model, each attribute depends on its previous state. Additionally, all attributes except the sex are conditioned by age, reflecting its central role in the individual development.

3.6.1. Updating attributes with ordered modalities

If there is an order between the modalities, no return to a previous state is possible. Therefore, the transition matrix is triangular, reducing the number of unknown quantities.

Among ordered attributes, some have sequential modalities, meaning an individual can only progress to the next modality in a fixed order. Age is an example of attribute that follows this rule: its modalities are strictly ordered, and an individual can only transition from one modality to the immediately subsequent one (or remain in the same modality). For such attributes, the transition matrix has non-null values only on the main diagonal and on the first diagonal upper the main diagonal. In the case of the age, there is no need to compute the transition matrix, as our temporal step is one year, so the transitions are straightforward.

In our experiments, we only have one other attribute that have ordered modalities: the highest degree. For this attribute, each modality can be reached only from the previous modality, with an exception. The highest degree can be obtained from two possible states: holding a high school diploma or a professional diploma. Therefore, in the transition matrix, we have one element non-null outside the two principal diagonals of the triangular matrix.

The different transition matrix corresponding to the different cases mentioned above are given here:

Transition matrix for an ordered attribute:

$$\begin{pmatrix} m_{1,1} & \cdots & \cdots & m_{1,n} \\ 0 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & m_{n,n} \end{pmatrix}$$

Transition matrix for an ordered attribute, when the modalities are mandatorily successive:

$$\begin{pmatrix} m_{1,1} & m_{1,2} & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & m_{n-1,n-1} & m_{n-1,n} \\ 0 & \cdots & \cdots & 0 & m_{n,n} \end{pmatrix}$$

Transition matrix for our degree attribute (successive modalities with one exception):

$$\begin{pmatrix} m_{1,1} & m_{1,2} & 0 & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & m_{i,j} & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & 0 \\ \vdots & & & \ddots & m_{n-1,n-1} & m_{n-1,n} \\ 0 & \cdots & \cdots & \cdots & 0 & m_{n,n} \end{pmatrix}$$

To compute the transition probabilities for the highest diploma, we use the observed frequencies of age and diploma combinations for two consecutive years. We can deduce the transition coefficients to fit the margins. However, we have one degree of freedom in our equation systems due to the extra coefficient. We assume that individuals with a high school diploma are prioritized for admission to higher education. If the target margin for higher education diplomas is not met by individuals who already hold such a diploma or those with a high school diploma, the remaining share is fulfilled by individuals holding a professional diploma.

3.6.2. Updating attributes with unordered modalities

For the last category of attributes, we also use the joint distribution between the attribute of interest and the age for two consecutive years. This time, there are too many transition coefficients to compute. The equation system has no unique solution. So we derive an heuristic method based on two main motivations: respecting the distribution, and preserving a momentum for individual states. The momentum represents that individuals tend to remain in the same state over years, so they are likely to remain with the same characteristics.

To that extent, we initialize the transition for each attribute ${}^{y}X_k$ to remain in state ω_k using a momentum rate $m_k \in [0, 1]$. The remaining probability mass is then distributed among all possible states (including the previous one) to align with the marginal distributions. In practice, we adapt this momentum rate, for each state, in order to avoid exceeding the margin:

$$M_{\omega_k} = \min\left(m_k, \frac{\mathbb{P}({}^{y+1}X_k = \omega_k)}{\mathbb{P}({}^{y}X_k = \omega_k)}\right). \tag{2}$$

Thus, the transition probability is given by:

$$\mathbb{P}\left({}^{y+1}X_k = \omega'_k | {}^{y}X_k = \omega_k\right) = M_{\omega_k}\mathbb{1}_{\omega'_k=\omega_k} + (1 - M_{\omega_k})\frac{\mathbb{P}({}^{y+1}X_k = \omega'_k) - M_{\omega'_k} \times \mathbb{P}({}^{y}X_k = \omega'_k)}{\sum_{\omega\in\Omega_k} \mathbb{P}({}^{y+1}X_k = \omega) - M_{\omega} \times \mathbb{P}({}^{y}X_k = \omega)} \tag{3}$$

In Equation 3, the first term corresponds to the momentum term and the second term to the random assignment of the remaining weights, so the marginal distributions are respected.

In our experiments, as shown in Figure 1, we condition Equation 3 on the individual's age, as an individual's characteristics—-such as student status—-are highly age-dependent. To better capture the evolving dynamics of each individual and keep realist individuals, the marginal distributions are therefore conditioned on age.

Name	Momentum
Number of persons in the household	0.99
Type of professional activity	0.95
Department	0.98
Number of cars	0.9

Table 2. Momentum values chosen for the attributes with no order between the instances.

3.6.3. Parameters for not ordered attributes

Unlike other the attribute types, each non-ordered attribute requires a hyperparameter to define its momentum value. We have chosen the values based on the expected stability of the instances. The chosen momentum values for each of the 4 non-ordered attributes are provided in Table 2.

By setting the momentum parameter to values below 1, we introduce flexibility that allows for exploration of alternative values and spontaneous transitions between states. The momentum rates are intentionally set high to maintain strong continuity between successive states. Additionally, in cases where there is a significant change in the margin of a given instance, the momentum coefficient is capped by the margin for the following year in Equation 2. This ensures that redistribution remains possible when the share of an instance decreases.

Finally, we can derive an updating function for any of the four types of attributes. Since the last two categories of attributes are largely random and no cross-correlation—except with age—is considered, this method should effectively capture the dynamics of each attribute individually, but may struggle to represent their combined interactions.

3.7. Generation of the initial solution

All different models require access to a microsample ${}^0\mathbf{S}^n$. In static projection and resampling, the samples are replicated to generate the new populations, considering some constraints. However, the number of samples in the initial microsample is limited. Therefore, the generated populations could lack diversity, as it is impossible to produce samples outside the scope of the training data. This problem is known as the zero-cell problem in the population synthesis literature (Guo and Bhat, 2007).

The extensive literature on population synthesis shows that there are reliable methods that enable a a good population reconstruction, thus preventing a lack of diversity. In this work, we propose changing the initial microsample ${}^0\mathbf{S}^n$ to another set drawn from ${}^y\mathbf{X}$ using state-of-the-art population synthesis algorithms. Based on Darsel et al. (2025b), we propose to use two different models–Bayesian Network and a Diffusion Model–to generate the initial microsample ${}^0\mathbf{S}^n$. These generative models generate an unlimited number of individuals. We leverage this ability to adapt the size of the initial microsample, but also to create a large pool of synthetic individuals, thereby enhancing the diversity and range of choices available in static projection and resampling models. This corresponds to generate ${}^0\mathbf{S}^N$, where $N >> N_y$, with N_y target size of the population for year y.

In static projection, changing the initial microsample directly impacts the results, as the initial solution in the IPF algorithm is derived from the proportion observed in the microsample.

In our experiments, we use both the samples generated by generative models, and samples from the training data as the starting solution and the pool of individuals. This allows us to assess the impact of each method.

Details on the parameters of the generative models are given in Appendix A.

4. Results

4.1. Data

In our experiments, we use open-source French census data, which is published every year. This disaggregated data comprises the entire population. The attributes describing the individuals are given in Table 1. They correspond to the more relevant attributes available for an agent-based simulation model, supported by Hörl and Balac (2021a). In

our experiments, we use the data from 2013 to 2021 provided by INSEE[5]. We concentrate on the Île-de-France area, which is the most densely populated region in France.

When the population is projected, only a microsample from 2013 is used for training generative models, or directly as initial solution. Data from 2013 to 2021 is used for evaluation, but also to train reference models to ensure they continue to accurately represent the true population over time. Finally, data from 2014 to 2021 is used to compute the aggregated statistics required for implementing the different projection models and for evaluation purposes. We consider an optimistic scenario in which all these statistics are available and accurate, which is not necessarily the case in practice.

To fit with the literature on population synthesis, we restrict the available microsamples to 1% of the total population. The 1% dataset is constructed with stratified sampling, based on the age, the sex and the department, allowing to keep the true joint distribution for these attributes. In experiments involving a generative model, as presented in Section 3.7, we are free in the size of the population we generate, as the generation of an individual is independent from the others. For the initial solution, we generate a population representing 5% of the total population, to achieve a better and more diverse representation of the population. A larger population would increase the evaluation time.

In addition to the initial solution for 2013, we use generative models to create a pool of potential individuals that can be selected using methods that replicate individuals. We set this pool at 50% of the total population size to ensure variety in our simulations. Although we ultimately generate a forecasted population of 5%, maintaining a larger pool helps to prevent the same individuals from being selected multiple times.

Finally, for the dynamic projection method, some external data sources are used to simulate the birth rates[6], the mortality rates [7], the marriage rates [8] and the divorce rates [9]. All these external data sources are also provided by INSEE. We note that the divorce rates data is no longer published yearly, therefore we consider the figures from 2013 for every year of the simulation. Using the dynamic of one year of data over multiple years is commonly done in dynamic projection (Ballas et al., 2005a; Kukic and Bierlaire, 2025).

4.2. Evaluation

To evaluate a generated population from a given year, we use a test set comprising 90We follow the evaluation criteria from Darsel et al. (2025a). Privacy is not a concern in our case, as comparing with training data from another year is meaningless in forecasting. However, we consider the other two criteria: distribution and realism.

To evaluate the distribution, we consider the Standardized Root Mean Squared Error ($SRMSE$) on the marginals, the bi-variate and the tri-variate distributions. We consider the same metric as Lederrey et al. (2022) and Darsel et al. (2025a) by taking the average score on all combinations of attributes for each size of the $SRMSE$.

First, we introduce the $SRMSE$ for a group of one, two or three variables X_k, X_l, and X_m between the generated data $\mathbf{X}_{gen}$ and the testing data $\mathbf{X}_{test}$, with:

$$
\begin{aligned}
SRMSE_k(\mathbf{X}_{gen}, \mathbf{X}_{test}) &= \sqrt{\sum_{\omega_k \in \Omega_k} \left(f_{gen}(\omega_k) - f_{test}(\omega_k)\right)^2 \times \mid \Omega_k \mid}, \\
SRMSE_{kl}(\mathbf{X}_{gen}, \mathbf{X}_{test}) &= \sqrt{\sum_{\omega_{kl} \in \Omega_{kl}} \left(f_{gen}(\omega_{kl}) - f_{test}(\omega_{kl})\right)^2 \times \mid \Omega_k \mid \times \mid \Omega_l \mid}, \\
SRMSE_{klm}(\mathbf{X}_{gen}, \mathbf{X}_{test}) &= \sqrt{\sum_{\omega_{klm} \in \Omega_{klm}} \left(f_{gen}(\omega_{klm}) - f_{test}(\omega_{klm})\right)^2 \times \mid \Omega_k \mid \times \mid \Omega_l \mid \times \mid \Omega_m \mid}.
\end{aligned}
\tag{4}
$$

[5] The 2013 dataset: https://www.insee.fr/fr/statistiques/2409376. The other years are available on the website with the same name expect the year.

[6] https://www.insee.fr/fr/statistiques/7673408?sommaire=7673431

[7] https://www.insee.fr/fr/statistiques/series/103039135

[8] https://www.insee.fr/fr/statistiques/8570890?sommaire=8571148

[9] Data from 2013 only: https://www.insee.fr/fr/statistiques/fichier/7624542/fm_t28.xlsx

Thus, we can define the average score:

$$\begin{aligned}
\overline{SRMSE}_1(\mathbf{X}_{gen}, \mathbf{X}_{test}) &= \frac{1}{d}\sum_{k=1}^{d} SRMSE_k(\mathbf{X}_{gen}, \mathbf{X}_{test}),\\
\overline{SRMSE}_2(\mathbf{X}_{gen}, \mathbf{X}_{test}) &= \frac{1}{\binom{d}{2}}\sum_{(k,l)\in\binom{\{1,\dots,d\}}{2}} SRMSE_{kl}(\mathbf{X}_{gen}, \mathbf{X}_{test}),\\
\overline{SRMSE}_3(\mathbf{X}_{gen}, \mathbf{X}_{test}) &= \frac{1}{\binom{d}{3}}\sum_{(k,l,m)\in\binom{\{1,\dots,d\}}{3}} SRMSE_{klm}(\mathbf{X}_{gen}, \mathbf{X}_{test}).
\end{aligned} \tag{5}$$

For realism, we consider the metric from Garrido et al. (2020) by computing the rate of structural zeros generated, but we consider the detection method of a structural zero given by Darsel et al. (2025a). In this detection method, a sample is considered as a structural zero, if there is at least one combination of two of its modalities that does not exist in the data. For example, a 6-year-old individual marked as married would be considered an undesired structural zero. This metric is named the Share of Samples with a Couple of Instances Out of Data (SSCIOD). A more detailed derivation of this metric is given by Darsel et al. (2025a).

4.3. Comparing the initial solution over years

In our first experiments, we would like to evaluate the error due to omitting the evolving population. To that extent, we consider the three methods for generating the initial solution derived in Section 3.7 – direct use of the training data, data generated by a Bayesian Network (BN), and data generated by a diffusion model. On the one hand, we generate the population based on the data from 2013 with the different methods, and we evaluate the generated population on the test set for each year between 2013 and 2021. On the other hand, we compare the scores with populations generated with the same method, but using the training data of the same year as the evaluation data.

In Figure 2, we observe the evolution of the different scores over years for the different configurations.

The results for 2013 state that the training data performs better than the population generated with the Bayesian Network, and the diffusion model. These scores are most of the time flat, when the models are trained with yearly data, specially for training data and Bayesian Network.

However, for the three $\overline{SRMSE}$ metrics, we observe a rise of the scores for the data generated from 2013 data over the years, whatever the generative process is. As this rise occurs only on model based on 2013 data, we deduce that it is due to a poorer representation of the population for years between 2014 and 2021, and not to the ability of the models to catch the current population. For a given $\overline{SRMSE}$, the lower the initial score is, the higher the increasing is. Indeed, we observe that the scores for the training data, which has the best initial scores, rise more than the scores for diffusion models. For instance, when dealing with the marginals, the score for the training set is multiplied by 10.2 between 2013 and 2021, but only by 1.19 for diffusion. For the tri-variate distribution, the increase for the training set is $\times 3.07$ against $\times 1.15$ for diffusion.

A last observation regarding the models using 2013 data is that Bayesian Network performs very well when evaluating the distribution of the marginal distribution, comparing with the other generative model: diffusion. But, we observe that the gap with the diffusion model gets smaller over time for the tri-variate distribution. Diffusion is even better for some specific years. This makes the diffusion model interesting as it seems to be good at capturing the correlations between attributes.

The realism–measured by the $SSCIOD$–remains relatively stable for the 2013 population. This stability is expected, as a realistic individual profile should maintain its realism over time. However, we note a variation in the $SSCIOD$ for diffusion in models trained annually. These fluctuations stem from differences in the quality of the learning process and the training data. Notably, the 2013 model performs significantly worse than models from other years.

From these first experiments, we highlight that using one year of data results in a loss of the relevance of the generated population over years. Therefore, projection methods for population synthesis are of interest, as data becomes

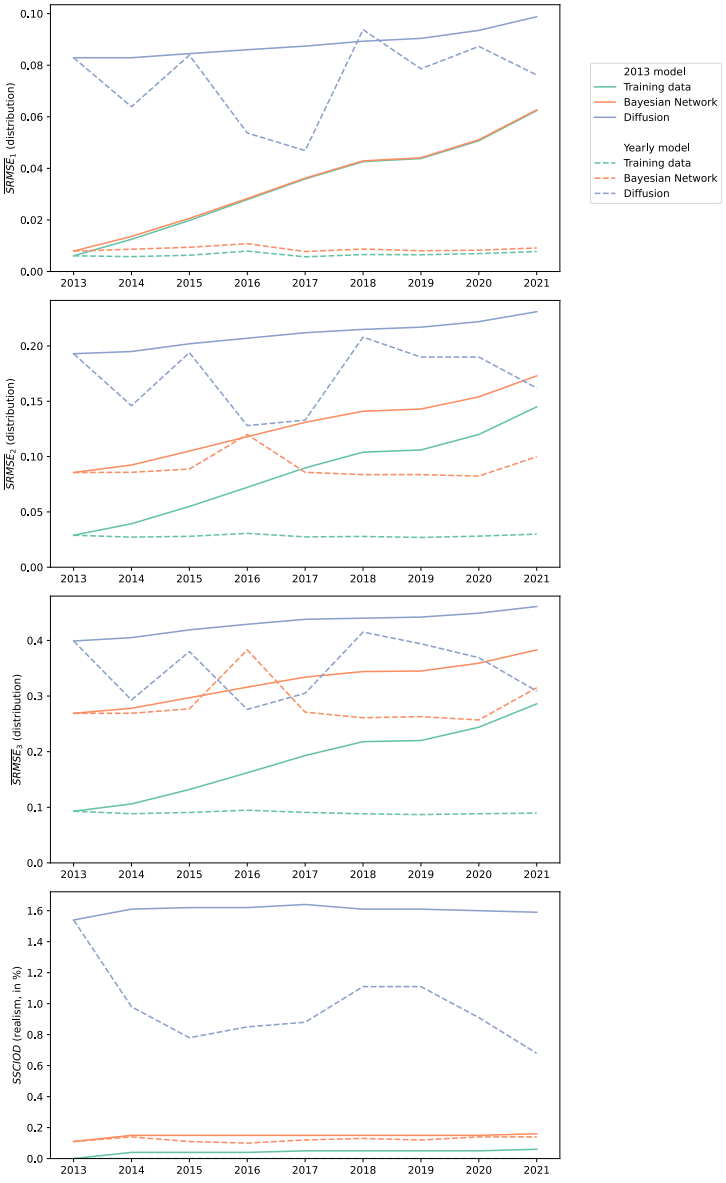

Fig. 2. Comparison of the models trained with 2013 data and with yearly data over years

less accurate over time. As this synthetic population is used as input for agent-based simulations, the same conclusion can be drawn about the simulations losing accuracy over time.

4.4. Comparing projection methods over years

As shown in Section 4.3, a synthetic population tends to be outdated after a couple of years. We aim now to evaluate the capacity of the different projection approaches–static projection (idealistic and realistic), resampling and dynamic projection–to forecast a population to overcome this issue.

Figure 3 compares the different scores for the different projection methods with the different initial solutions.

Overall, the idealistic static projection method outperforms the others, regardless of the initial population, as we expected. Its performance deteriorates only slightly over time, though this decline is more pronounced when training

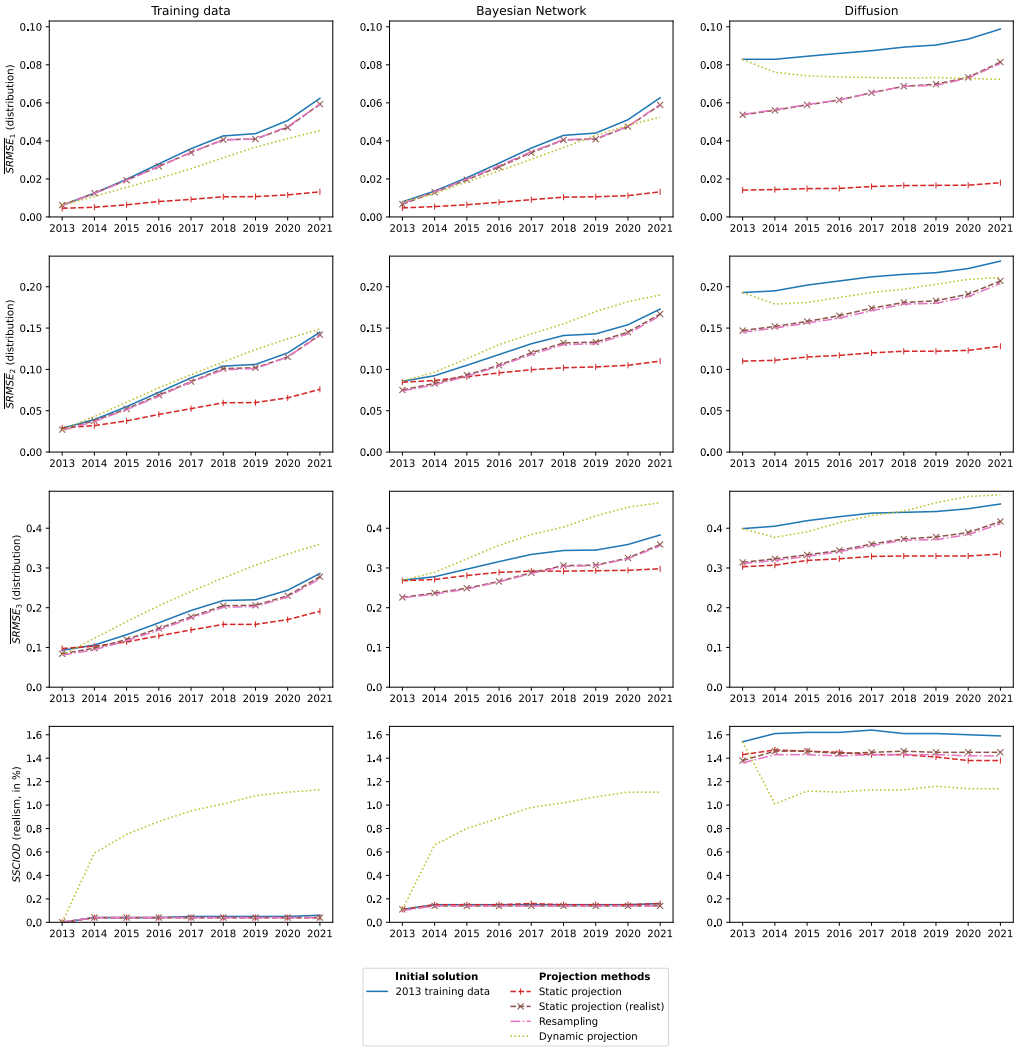

Fig. 3. Comparison of the projection methods for different initial solutions

data are used as the initial population, particularly for bi- and tri-variate statistics. The benefits of static projection are especially evident for diffusion models, which rely on the IPF steps to achieve strong scores for marginal distributions. This improvement also positively impacts bi- and tri-variate distributions. In this context, the refitting step in the static projection method significantly enhances population quality, even for the initial 2013 solution. While diffusion models initially underperform compared to other models, their results become comparable to those of the Bayesian Network across all years after refitting, and even align with training data in later experimental years.

However, these strong results are largely attributable to the richness of the projection data, as all attributes are incorporated during the fitting stage. In realistic settings, the static projection still outperforms the initial solution, but the improvement is far less substantial—particularly when using training data, where gains are minimal.

The results closely mirror those of the resampling approach, as anticipated, since both methods use the same projection data. Resampling performs slightly better due to its access to additional information through the joint distribution.

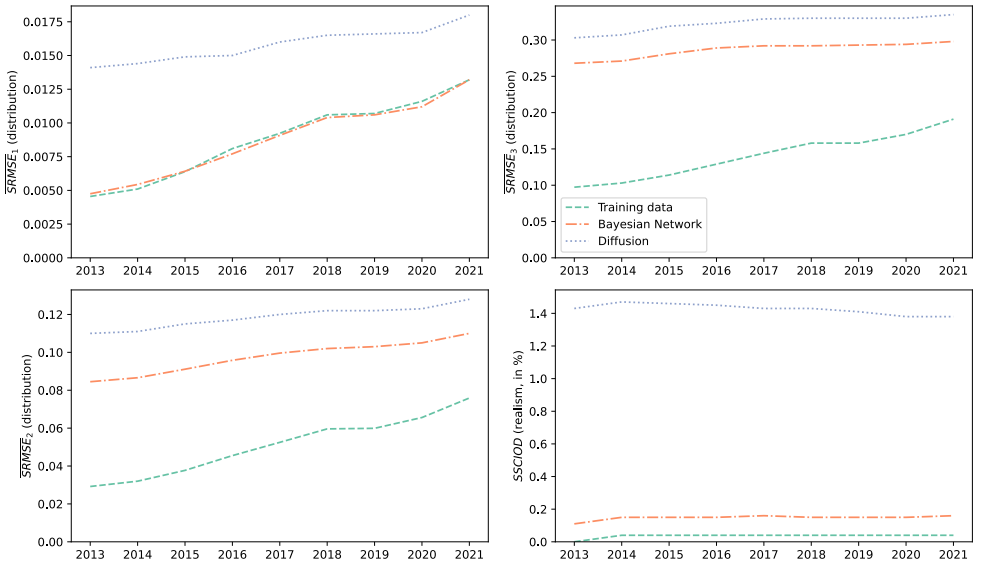

Fig. 4. Comparison of the scores for static projection

The minor differences between the two methods suggest that the IPF effectively estimates the joint distribution. So in this case, the joint distribution is useless and the marginal distributions are enough. It is more valuable to have access to the marginal distributions for more attributes rather than getting the joint distribution. While the improvements from realistic static projection and resampling may seem modest, they reflect the inherent stability of the attributes considered (sex, age, and location). Since these attributes evolve slowly over time, the disparity between the initial population and updated samples remains limited.

The dynamic projection approach yields slight improvements in marginal distributions (i.e., the the $\overline{SRMSE}_1$ metric), across all three settings. However, it worsens results for other metrics compared to the baseline (the initial 2013 solution), as anticipated. Its implementation seems to disrupt the capture of more complex correlations, rendering it unsuitable in its current form. Combining the dynamic projection method with the diffusion model notably improves the realism of the generated population in the first year of the experiment.

To examine the static projection method in more detail, we compare the populations fitted with this model for the various initial solutions shown in Figure 4. The results show that using training data consistently yields the highest scores across all metrics, with the possible exception of $\overline{SRMSE}_1$. However, the performance gap between the training data and the diffusion model gradually diminishes over time.

As expected, the realism metric ($SSCIOD$) yields constant scores for methods that draw samples from a fixed pool—such as static projection and resampling—since the realism of the samples closely reflects the realism of the sampling method itself, which corresponds to the realism observed in the first year of the experiment.

5. Discussions

All our projection methods assume access to the accurate aggregated information for the target years—an assumption rarely met in practice. The dynamic projection method, for instance, requires knowledge of age-conditional margins for attributes that cannot be directly computed or derived from longitudinal data. In our experiments, the idealistic static projection method relies on 168 marginal coefficients annually, while the dynamic projection demands 3,038 coefficients. Resampling, though limited to a subset of attributes, still requires 112 coefficients yearly to capture the joint distribution. For dynamic projection, one could consider reusing the same transition probabilities across years, as proposed by Namazi-Rad et al. (2014) and Kukic et al. (2023). However, this approach is unsuitable for static projection and resampling, as it would generate identical populations year after year.

Regarding population forecasts, Wilson et al. (2022) provides an extensive review of small-area forecasting methods by age and sex. Policymakers frequently conduct such forecasts—for example, in France Age Pyramid - INSEE) and Denmark (Statbank) for their projections. Refitting with age, sex, and location aligns with our resampling model and is more feasible in practice. Thus, the realistic static projection and resampling methods are better suited for real-world applications.

For the dynamic projection method, we could have anticipated stronger performance on the $\overline{SRMSE}_1$ metric, given its reliance on marginal frequencies and age-conditional evolution laws. However, achieving the expected margins requires an accurate age distribution fit. Our model does not account for external migrations, which can disrupt age balance over time due to age-correlated migration flows. This omission may partly explain the method's underwhelming results.

Similarly, the idealistic static projection method does not achieve a perfect $\overline{SRMSE}_1$ score, despite optimization via the IPF algorithm. This is not due to overly lenient stopping criteria but rather to the binning process in IPF. Some attributes, such as individuals aged 110, exist in the true population but are absent from the training data, necessitating binning. This discrepancy in granularity introduces additional evaluation errors.

With the idealistic static projection, the relative ranking of the models remains unchanged, but the gap between the diffusion model and the Bayesian Network has narrowed—particularly for multivariate errors. This highlights the power of using margins to refine the synthetic population generated from the diffusion model. Diffusion does not generate a well balanced set, but rather a good starting solution to pick samples in, for population synthesis or agent-forecasting. Given the low error in the training data compared to the diffusion model's initial error, the training data margins can serve as a reliable proxy for the true margins when the latter are unavailable. This trick could be extend to any population synthesis algorithm that struggles in respecting the margins.

All three forecasting methods improve the marginal distribution for the diffusion model. This improvement is intuitive for static projection and resampling, as both methods use constraints on the aggregated statistics. For dynamic projection, however, the improvement is less immediately apparent. The enhancement emerges starting in 2014 and becomes more pronounced each subsequent year. While the marginals inform the derivation of the updating rules, they do not impose hard constraints. Nevertheless, we can mathematically justify why this method helps refitting the marginals. When marginal distributions remain relatively stable over time, i.e. when Equation 2 yields m_k, the error is reduced by a factor proportional to the momentum. In Appendix B, we provide a formal proof for cases where the momentum is uniform across all instances of a given modality.

The realism score for the diffusion model could eventually appear weak in both Figure 2 and 3, as it performs the worst by far. However, the share of 1% of unrealistic samples is low compared to most generative models, what have been explored in Darsel et al. (2025a). Therefore, the realism figures are acceptable in our different experiences.

Another concern regarding using the microsample as the initial solution is about privacy. As the data is directly copied for two methods, the generated population would contain directly the information about true individuals at the initial year.

6. Conclusion & Perspectives

This study shows that synthetic populations inevitably become outdated over time, regardless of the generative model used. This highlights the importance of using agent-forecasting techniques to keep populations up to date and ensure they are reliable for simulation purposes.

We evaluated four forecasting methods—idealistic static projection, realistic static projection, resampling, and dynamic projection—applied to a population spanning 2013 to 2021. The idealistic static projection represents an unattainable benchmark in practice but illustrates the method's maximum potential. For dynamic projection, longitudinal data for all attributes was unavailable, as the inclusion of additional attributes (intended to better fit simulation scenarios) made such data harder to obtain. To address this, we introduced a heuristic approach based on evolving margins to generate transitions for attributes without explicit evolution laws. However, this method does not account for cross-modal interactions.

As expected, our experiments confirmed that idealistic static projection delivers the best performance, while dynamic projection fails to improve population quality when the initial solution is already robust. Resampling and

realistic static projection yield similar results, suggesting that access to the joint distribution (resampling) is not essential—marginal distributions (realistic static projection) suffice to achieve comparable performance.

When comparing different models for generating the initial population—training data, Bayesian Network samples, and diffusion model samples—training data consistently produced the best results across all forecasting methods. However, the diffusion model underscored the value of a refitting step (static projection applied to the initial 2013 population), which significantly improved the quality of the generated population. Since Bayesian Networks inherently fit the training data margins, this step is unnecessary for them. We recommend performing a refitting step whenever the $\overline{SRMSE}_1$ of the generated population is substantially worse than that of the training data.

Our analysis assumes an optimistic scenario where 1% of the total population is available—a proportion that may not reflect real-world conditions. In practice, training data are often smaller and less representative, potentially limiting the generalizability of our findings. Future research could explore the impact of smaller, less representative training datasets. Additionally, we used a broad geographical attribute (department), which lacks discriminative power. Investigating finer geographical granularity could increase forecasting complexity but may also enhance simulation accuracy.

We also operated under the idealized assumption that all aggregated statistics required for projection methods are perfectly accurate. In reality, such data are often subject to measurement errors, which could weaken or even preclude the implementation of these methods. For instance, the idealistic static projection assumes access to precise marginal distributions for all attributes at the target year—a requirement that is rarely feasible in practice.

Despite its superior performance, idealistic static projection still shows a gradual decline in multivariate distributions over time, leaving room for improvement. This method replicates samples from the initial population, effectively constraining the final distribution to remain close to the original. Consequently, attribute correlations in the final population closely mirror those in the initial population. Capturing evolving inter-attribute correlations—by modifying individuals from the initial population—represents a promising yet unexplored avenue for enhancing current methods.

Unlike Kukic and Bierlaire (2025), who worked at the household level, our experiments were conducted at the individual level. A household-level approach could be particularly relevant for dynamic projection, as some attributes are shared among household members or derived directly from household structure (e.g., household size). While constructing updating rules at this level requires extensive data, exploring the differences in results between individual- and household-level frameworks would be valuable.

Appendix A. Population synthesis models implementation

In this appendix, we present the parameters of the generative models used for the generation of the initial solution.

A.1. Bayesian Network implementation

A Probabilistic Graphical Model needs to be learned first. To that extent, we use a hill-climbing approach with the BIC criterion (Koller and Friedman, 2009). Then, the probabilities are computed based on the observed frequencies. In our experiments, we use the Python package called pgmpy (Ankan and Textor, 2024).

A.2. Diffusion implementation

Our implementation of diffusion is the same as Darsel et al. (2025a), and is an adaptation from TabSyn model (Zhang et al., 2023). This model is depicted by Figure A.5. First, a network embeds the mixed-type data into continuous data. The embedding network is performed by a Transformer trained with a VAE structure. Then, a diffusion model is trained on the continuous embedded data. For more detailed information on the model, we recommand to refer to the previously mentioned contributions.

The architectures for our models are given by Figure A.6 and A.7. The hyperparameters are the same used in Darsel et al. (2025a).

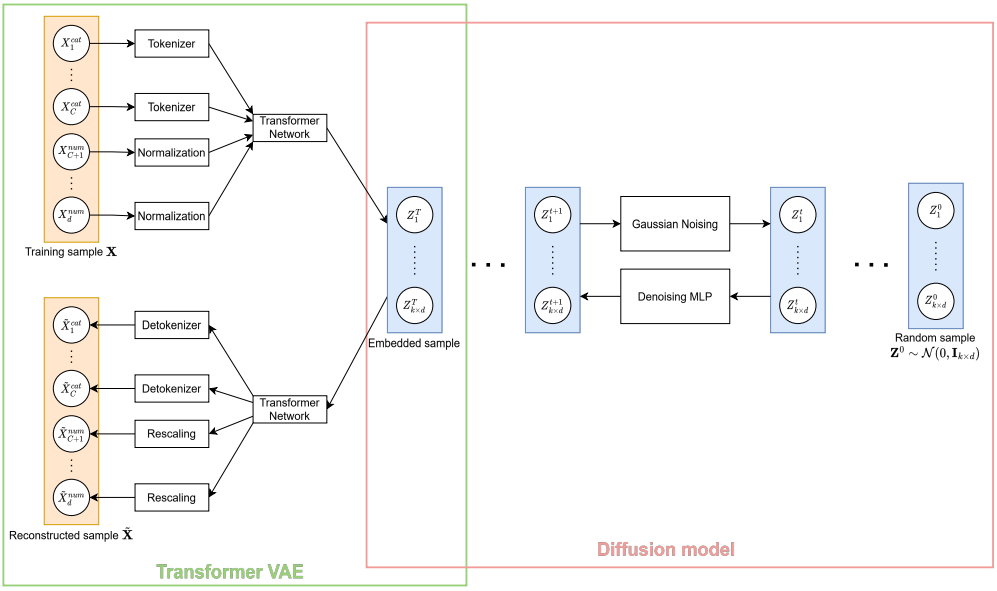

Fig. A.5. Architecture for the Transformer VAE network used for data embedding. d is the dimension of the data. The process is slightly different for numerical and categorical variables.

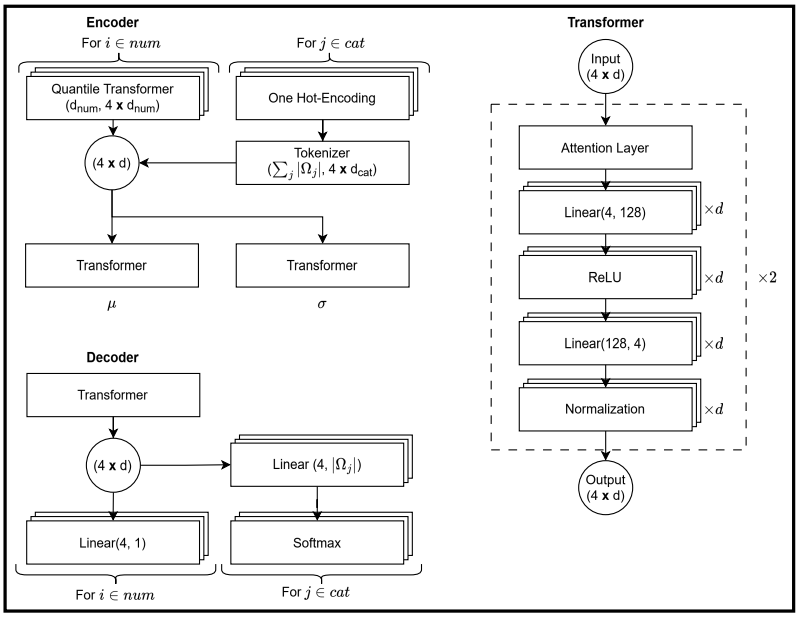

Fig. A.6. Architecture for the Transformer VAE network used for data embedding. d is the dimension of the data. The process is slightly different for numerical and categorical variables.

Appendix B. Error vanishing with momentum in dynamic projection

Here, we would like to demonstrate that an error on the margins can be reduced thanks to the momentum term. Using the notations from Section 3, we consider a modality ω_k with an error ϵ (i.e. ${}^{y}f_k(\omega_k) = \mathbb{P}({}^{y}X_k = \omega_k) + \epsilon$). We compute the expectation for the frequency for the next year. Between two years, we assume that the changes in the margins are small enough for all modalities, so $\forall\omega \in \Omega_k, M_\omega = m_k \in]0, 1[$. When deriving $\mathbb{E}[{}^{y+1}f_k(\omega_k)]$, we get:

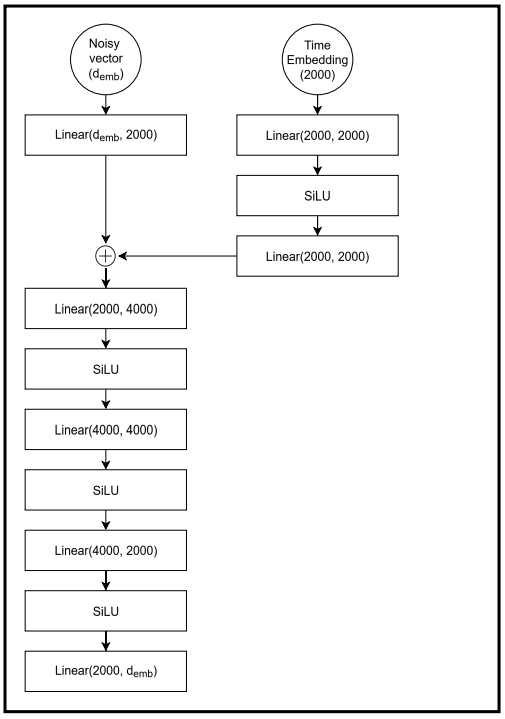

Fig. A.7. Architecture for the diffusion network used for data embedding. d_{emb} is the dimension of the embedded data.

$$
\begin{aligned}
\mathbb{E}[{}^{y+1}f_k(\omega_k)] &= {}^{y}f_k(\omega_k) \times M_{\omega_k} + (1 - M_{\omega_k}) \times \frac{\mathbb{P}({}^{y+1}X_k = \omega_k) - M_{\omega_k}\mathbb{P}({}^{y}X_k = \omega_k)}{\sum_{\omega \in \Omega_k} \mathbb{P}({}^{y+1}X_k = \omega) - M_{\omega}\mathbb{P}({}^{y+1}X_k = \omega)} \\
&= {}^{y}f_k(\omega_k) \times m_k + (1 - m_k) \times \frac{\mathbb{P}({}^{y+1}X_k = \omega_k) - m_k\mathbb{P}({}^{y}X_k = \omega_k)}{\sum_{\omega \in \Omega_k} \mathbb{P}({}^{y+1}X_k = \omega) - m_k\mathbb{P}({}^{y+1}X_k = \omega)} \\
&= (\mathbb{P}({}^{y}X_k = \omega_k) + \epsilon) \times m_k + (1 - m_k) \times \frac{\mathbb{P}({}^{y+1}X_k = \omega_k) - m_k\mathbb{P}({}^{y}X_k = \omega_k)}{1 - m_k} \\
&= \mathbb{P}({}^{y+1}X_k = \omega_k) + \epsilon \times m_k
\end{aligned}
\tag{B.1}
$$

The expected error is now $\epsilon \times m_k$.

References

Ankan, A., Textor, J., 2024. pgmpy: A Python Toolkit for Bayesian Networks. Journal of Machine Learning Research 25, 1–8. URL: http://jmlr.org/papers/v25/23-0487.html.

Ballas, D., Clarke, G., Dewhurst, J., 2006. Modelling the Socio-economic Impacts of Major Job Loss or Gain at the Local Level: a Spatial Microsimulation Framework. Spatial Economic Analysis 1, 127–146. URL: https://doi.org/10.1080/17421770600697729, doi:10.1080/17421770600697729. publisher: RSA Website _eprint: https://doi.org/10.1080/17421770600697729.

Ballas, D., Clarke, G.P., Wiemers, E., 2005a. Building a dynamic spatial microsimulation model for Ireland. Population, Space and Place 11, 157–172. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/psp.359, doi:10.1002/psp.359. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/psp.359.

Ballas, D., Rossiter, D., Thomas, B., Clarke, G., Dorling, D., 2005b. Geography matters: Simulating the local impacts of national social policies. Geography matters: Simulating the local impacts of national social policies URL: https://ora.ox.ac.uk/objects/uuid:30b0d85f-9174-4aab-a06a-541576831c3c. iSBN: 9781859352656 Publisher: Joseph Rowntree Foundation.

Beckman, R.J., Baggerly, K.A., McKay, M.D., 1996. Creating synthetic baseline populations. Transportation Research Part A: Policy and Practice 30, 415–429. URL: https://www.sciencedirect.com/science/article/pii/0965856496000043, doi:10.1016/0965-8564(96)00004-3.

Booth, H., 2006. Demographic forecasting: 1980 to 2005 in review. International Journal of Forecasting 22, 547–581. URL: https://www.sciencedirect.com/science/article/pii/S016920700600046X, doi:10.1016/j.ijforecast.2006.04.001.

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. URL: https://www.sciencedirect.com/science/article/pii/S0968090X1831180X, doi:10.1016/j.trc.2019.07.006.

Darsel, V., Come, E., Oukhellou, L., 2025a. Robust and Reproducible Evaluation Framework for Population Synthesis Models — Application to Probabilistic and Deep Generative Models. URL: https://papers.ssrn.com/abstract=5295092.

Darsel, V., Côme, E., Oukhellou, L., 2025b. Population Synthesis with Deep Generative Models -is it worth it? Exploring new models and metrics, in: 13th Symposium of the European Association for Research in Transportation, Technische Universität München, Munich (Germany), Germany. URL: https://hal.science/hal-05130079.

Farooq, B., Bierlaire, M., Hurtubia, R., Flötteröd, G., 2013. Simulation based population synthesis. Transportation Research Part B: Methodological 58, 243–263. URL: https://www.sciencedirect.com/science/article/pii/S0191261513001720, doi:10.1016/j.trb.2013.09.012.

Fatmi, M.R., Habib, M.A., 2017. Baseline Synthesis and Microsimulation of Life-stage Transitions within an Agent-based Integrated Urban Model. Procedia Computer Science 109, 608–615. URL: https://www.sciencedirect.com/science/article/pii/S1877050917310359, doi:10.1016/j.procs.2017.05.366.

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. URL: https://www.sciencedirect.com/science/article/pii/S0968090X20306975, doi:10.1016/j.trc.2020.102787.

Guo, J.Y., Bhat, C.R., 2007. Population Synthesis for Microsimulating Travel Behavior. Transportation Research Record 2014, 92–101. URL: https://doi.org/10.3141/2014-12, doi:10.3141/2014-12. publisher: SAGE Publications Inc.

Gupta, S., Agrawal, A., Gopalakrishnan, K., Narayanan, P., 2015. Deep Learning with Limited Numerical Precision, in: Proceedings of the 32nd International Conference on Machine Learning, PMLR. pp. 1737–1746. URL: https://proceedings.mlr.press/v37/gupta15.html. iSSN: 1938-7228.

Harding, A., Vidyattama, Y., Tanton, R., 2011. Demographic change and the needs-based planning of government services: projecting small area populations using spatial microsimulation. Journal of Population Research 28, 203–224. URL: https://doi.org/10.1007/s12546-011-9061-6, doi:10.1007/s12546-011-9061-6.

Hörl, S., Balac, M., 2021a. Open synthetic travel demand for Paris and Île-de-France: Inputs and output data. Data in Brief 39, 107622. URL: https://www.sciencedirect.com/science/article/pii/S2352340921008970, doi:10.1016/j.dib.2021.107622.

Hörl, S., Balac, M., 2021b. Synthetic population and travel demand for Paris and Île-de-France based on open and publicly available data. Transportation Research Part C: Emerging Technologies 130, 103291. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21003016, doi:10.1016/j.trc.2021.103291.

van Imhoff, E., Post, W., 1998. Microsimulation Methods for Population Projection. Population: An English Selection 10, 97–138. URL: https://www.jstor.org/stable/2998681. publisher: Institut National d'Etudes Demographiques.

Koller, D., Friedman, N., 2009. Probabilistic Graphical Models: Principles and Techniques. MIT Press. Google-Books-ID: 7dzpHCHzNQ4C.

Kukic, M., Benchelabi, S., Bierlaire, M., 2023. Hybrid Simulator for Capturing Dynamics of Synthetic Populations, in: 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC), pp. 2646–2651. URL: https://ieeexplore.ieee.org/abstract/document/10422198, doi:10.1109/ITSC57777.2023.10422198. iSSN: 2153-0017.

Kukic, M., Bierlaire, M., 2025. Adaptive synthetic generation using one-step Gibbs Sampler.

Kukic, M., Li, X., Michel Bierlaire, 2024. One-step Gibbs sampling for the generation of synthetic households. Transportation Research Part C: Emerging Technologies 166, 104770. URL: https://www.sciencedirect.com/science/article/pii/S0968090X24002912, doi:10.1016/j.trc.2024.104770.

Lederrey, G., Hillel, T., Bierlaire, M., 2022. DATGAN: Integrating expert knowledge into deep learning for synthetic tabular data. URL: http://arxiv.org/abs/2203.03489, doi:10.48550/arXiv.2203.03489. arXiv:2203.03489 [cs].

Lomax, N., Smith, A.P., Archer, L., Ford, A., Virgo, J., 2022. An Open-Source Model for Projecting Small Area Demographic and Land-Use Change. Geographical Analysis 54, 599–622. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/gean.12320, doi:10.1111/gean.12320. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/gean.12320.

Mazzuco, S., Keilman, N. (Eds.), 2020. Developments in Demographic Forecasting. Springer Nature. URL: https://library.oapen.org/handle/20.500.12657/42565, doi:10.1007/978-3-030-42472-5. accepted: 2020-10-13T12:29:52Z.

Namazi-Rad, M.R., Mokhtarian, P., Perez, P., 2014. Generating a Dynamic Synthetic Population – Using an Age-Structured Two-Sex Model for Household Dynamics. PLOS ONE 9, e94761. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0094761, doi:10.1371/journal.pone.0094761. publisher: Public Library of Science.

Prédhumeau, M., Manley, E., 2023. A synthetic population for agent-based modelling in Canada. Scientific Data 10, 148. URL: https://www.nature.com/articles/s41597-023-02030-4, doi:10.1038/s41597-023-02030-4. publisher: Nature Publishing Group.

Sun, L., Erath, A., 2015. A Bayesian network approach for population synthesis. Transportation Research Part C: Emerging Technologies 61, 49–62. URL: https://www.sciencedirect.com/science/article/pii/S0968090X15003599, doi:10.1016/j.trc.2015.10.010.

Tanton, R., Edwards, K., 2012. Spatial Microsimulation: A Reference Guide for Users. Springer Science & Business Media. Google-Books-ID: aNOkfcgsBZAC.

Vencatasawmy, C.P., Holm, E., Rephann, T., Esko, J., Swan, N., Öhman, M., Åström, M., Alfredsson, E., Holme, K., 1999. Building a spatial

microsimulation model .

Wilson, T., Grossman, I., Alexander, M., Rees, P., Temple, J., 2022. Methods for Small Area Population Forecasts: State-of-the-Art and Research Needs. Population Research and Policy Review 41, 865–898. URL: https://doi.org/10.1007/s11113-021-09671-6, doi:10.1007/s11113-021-09671-6.

Ye, X., Konduri, K., Pendyala, R., Sana, B., Waddell, P., 2009. Methodology to match distributions of both household and person attributes in generation of synthetic populations .

Zhang, H., Zhang, J., Shen, Z., Srinivasan, B., Qin, X., Faloutsos, C., Rangwala, H., Karypis, G., 2023. Mixed-Type Tabular Data Synthesis with Score-based Diffusion in Latent Space. URL: https://openreview.net/forum?id=4Ay23yeuz0.