

# Population Synthesis with Geographic Coordinates

Jacopo Lenti  
Sapienza University of Rome  
Rome, Italy & Intesa Sanpaolo  
Turin, Italy  
jcp.lenti@gmail.com

Lorenzo Costantini  
Intesa Sanpaolo  
Turin, Italy  
lorenzo.costantini@intesasanpaolo.com

Ariadna Fosch  
Institute for Biocomputation and  
Physics of Complex Systems (BIFI),  
University of Zaragoza  
Zaragoza, Spain.  
arifosch@gmail.com

Anna Monticelli  
Intesa Sanpaolo Innovation Center  
Turin, Italy  
anna.monticelli@intesasanpaolo.com

David Scala  
Intesa Sanpaolo  
Turin, Italy  
david.scala@intesasanpaolo.com

Marco Pangallo  
CENTAI Institute  
Turin, Italy  
marco.pangallo@gmail.com

## ABSTRACT

It is increasingly important to generate synthetic populations with explicit coordinates rather than coarse geographic areas, yet no established methods exist to achieve this. One reason is that latitude and longitude differ from other continuous variables, exhibiting large empty spaces and highly uneven densities. To address this, we propose a population synthesis algorithm that first maps spatial coordinates into a more regular latent space using Normalizing Flows (NF), and then combines them with other features in a Variational Autoencoder (VAE) to generate synthetic populations. This approach also learns the joint distribution between spatial and non-spatial features, exploiting spatial autocorrelations. We demonstrate the method by generating synthetic homes with the same statistical properties of real homes in 121 datasets, corresponding to diverse geographies. We further propose an evaluation framework that measures both spatial accuracy and practical utility, while ensuring privacy preservation. Our results show that the NF+VAE architecture outperforms popular benchmarks, including copula-based methods and uniform allocation within geographic areas. The ability to generate geolocated synthetic populations at fine spatial resolution opens the door to applications requiring detailed geography, from household responses to floods, to epidemic spread, evacuation planning, and transport modeling.

## KEYWORDS

Synthetic populations; Normalizing Flows; Geolocated data; Variational autoencoders; Agent-Based Models; Synthetic data

### ACM Reference Format:

Jacopo Lenti, Lorenzo Costantini, Ariadna Fosch, Anna Monticelli, David Scala, and Marco Pangallo. 2026. Population Synthesis with Geographic Coordinates. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026)*, Paphos, Cyprus, May 25 – 29, 2026, IFAAMAS, 10 pages. <https://doi.org/10.65109/NCOT6560>



This work is licensed under a Creative Commons Attribution International 4.0 License.

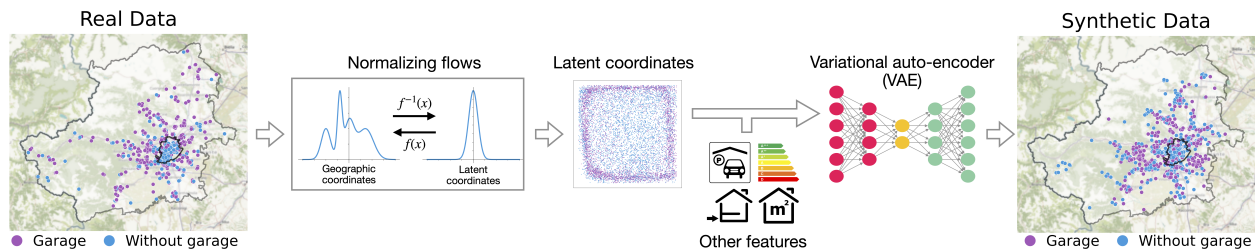
*Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026)*, C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems ([www.ifaamas.org](http://www.ifaamas.org)). <https://doi.org/10.65109/NCOT6560>

## 1 INTRODUCTION

In data-driven Agent-Based Models (ABMs) [46], it is crucial to place agents or elements of the environment at specific geographic coordinates. For instance, in ABMs of household responses to flood risk, homes must be located at coordinates with varying exposure to inundation [26, 30, 45]. Similarly, in epidemiological ABMs, individuals need to be placed at real residences, schools, or workplaces to capture realistic patterns of disease spread [6, 44]. Beyond these cases, geolocation is also critical in applications such as traffic simulation, urban mobility, and evacuation planning, where distance and spatial constraints shape interactions [32]. For all these examples, geographic aggregates such as postcode areas or administrative units are not sufficient.

In an ideal setting, spatial ABMs could be initialized directly from geolocated data. In practice, however, this is rarely feasible, often due to privacy concerns, since geolocation can easily reveal individual identities. This creates the need for methods that construct synthetic populations of individuals or of the places where they live and work, such as homes, schools, or workplaces. In some cases, spatial information may be available from GPS traces, land-use maps, or similar sources, in which case the task reduces to assigning synthetic agents to specific places—a process still typically guided by rules of thumb [17, 58?]. More commonly, only a geolocated sample is available, without land use maps or similar spatial information. The sample may be in a secured server, and it may not be possible to directly use the data for ABM simulation. The challenge then is to generate geolocated synthetic populations that reproduce the observed geographic distribution—for instance, placing residential units in residential areas rather than in water bodies or industrial zones. To the best of our knowledge, no existing method can generate geolocated synthetic populations directly from a geolocated sample, without relying on additional spatial data or heuristic assumptions.

One reason why population synthesis methods have so far struggled to incorporate geographic coordinates is simply that it is difficult. Spatial data involve large empty areas and highly variable densities, and treating coordinates like standard continuous variables that have far more regular distributions is unlikely to yield meaningful results. The common alternative has been to use geographic aggregates (such as census tracts or provinces) as discrete variables [7, 10, 11, 19, 57], but this approach is far from ideal, even



**Figure 1: Overview of the proposed population synthesis generation approach. The real geolocated population is given as input to the generator. Normalizing Flows are trained to map the real geographic coordinates to a simple latent space. Together with all other home features, these latent coordinates are used to train a Variational Autoencoder. Finally the Variational Autoencoder samples synthetic populations that resemble the input data. Left and right panels compare a random sample of 1,000 real and synthetic homes (respectively) in the province of Turin (gray lines). Synthetic data reproduces real patterns, with higher presence of garage in the outskirts of Turin city (black lines in the maps), and lower presence in the city center.**

beyond the issue that it may provide a too coarse grained spatial description. Using categorical variables for space is problematic partly due to the curse of dimensionality, and partly because it ignores spatial correlations. For example, when generating synthetic homes, a model may learn from the sample that in census tracts in the center of a historical city it is unlikely to find homes with a garage, yet this information does not transfer to adjacent tracts in the city, as a tract downtown is treated like a tract in the outskirts.

In this paper, we introduce a method for constructing synthetic populations with geographic coordinates, addressing both challenges outlined above in a unified framework. First, we tackle the non-regularity of spatial data by mapping coordinates into a latent space using Normalizing Flows (NF), a recently developed class of neural networks [47]. Next, we feed these latent coordinates, along with other features, into a Variational Autoencoder (VAE) [36], so that their joint representation in a second latent space captures correlations between spatial and non-spatial variables. As a generative model, this framework allows us to produce arbitrarily many synthetic samples that replicate both spatial distributions and feature relationships. Figure 1 qualitatively shows the accuracy of our approach and provides a schematic representation of the generative model. Examples for other provinces and different attributes are available in Supplementary Materials (SM)<sup>1</sup>.

Because this is the first paper generating a synthetic population with geographic coordinates from a geolocated sample only, there are no off-the-shelf evaluation methods. Thus, a contribution of this paper is to propose an evaluation protocol that can be used by future methods that build on our work. Specifically, our protocol evaluates models along three dimensions: fidelity, utility, and privacy. Fidelity measures how closely the synthetic population resembles the real one. Utility assesses whether the synthetic data can be effectively used for real-world tasks. Privacy ensures that no sensitive information from the original individuals can be inferred or recovered from the synthetic population.

We find that our method (NF+VAE) beats other baselines at the combination of these metrics in the two case studies evaluated: (i) the generation of synthetic homes using a mortgage dataset in Italy [12]; (ii) the generation of homes listed in Airbnb in 15 cities [1, 2]. In both cases, the inclusion of NF is crucial to obtain

an accurate mapping of the geographic coordinates, while the use of the VAE allows us to capture spatial autocorrelation better than other traditional methods.

Summarizing, NF+VAE is an adaptable method that can be used to generate geolocated synthetic populations with a combination of numerical and categorical features. Moreover, due to the loss of information during the VAE encoding, the synthetic population does not replicate individual records from the original dataset, thereby fully preserving privacy. Our approach opens the door to improved sharing of geolocated sensitive data, by producing synthetic datasets that remain faithful to the real ones, while safeguarding user privacy.

**Example.** Our main case study concerns synthetic homes used as mortgage collateral. In collaboration with Intesa Sanpaolo (one of the leading commercial banks in Italy) and Intesa Sanpaolo Innovation Center, we developed an ABM to study the impact of flood risk on housing prices, as households shift their demand from at-risk homes to safer ones. For this application, it was essential that the geographic distribution of homes in the simulation reflected actual flood risk maps: it would make little sense to place most homes in high-risk zones if only a small fraction are located there in reality. At the same time, spatial information is highly sensitive, since precise coordinates could make it possible to identify individual clients. These two conditions illustrate a typical use case for our model. By building synthetic populations of homes, we can recreate geographic distributions present real data, while avoiding privacy leakage of sensible information about the bank’s clients. Finally, the synthetic populations can be used as input of the ABM simulations, thus maintaining the properties of the original dataset and preserving privacy.

**Related work.** Population synthesis typically combines a few principled methods with many problem-specific assumptions [3]. Our contribution falls within the class of sample-based methods [15, 29], where a sample is available from which all relevant statistical distributions can be learned and new instances generated. This differs slightly from the classical population synthesis problem, where one typically has a sample at a higher geographic aggregation and marginal distributions at the desired aggregation, often addressed with Iterative Proportional Fitting [11]. Sample-free methods also exist [10, 19], and active research is comparing them to sample-based approaches [39].

<sup>1</sup>SM is available at <https://github.com/jacleti/NFVAE-population-synthesis/blob/main/SM-Population-synthesis-with-geographic-coordinates.pdf>

While several directions have been explored within population synthesis (e.g., matching individuals to households [7, 56]), two are particularly relevant for our work. One line of research concerns the use of generative modeling, and VAEs in particular, for population synthesis [4, 13, 15, 29, 50]. A key advantage is that, being probabilistic rather than deterministic, VAEs never return exact copies of individuals from the training sample, thereby preventing identification. Probabilistic approaches were already available through Bayesian Networks [54, 58], Gibbs Sampling [25], and related methods. However, deep learning approaches offer at least two advantages in population synthesis. First, it helps overcome the curse of dimensionality, since traditional methods scale poorly with the number of features, and categorical variables treated with one-hot encoding further increase the dimensionality. Second, it addresses zero-cell problem, exploring novel plausible combinations of features.

Another line of research concerns geolocated synthetic populations, moving beyond geographic aggregates. This area has received much less attention. A common approach is to distribute units randomly within polygons such as neighborhoods or along lines such as roads, as in the SPEW package [28]. When map information is available, for example land use maps or GIS features [17, 58? ], it can be used to place agents more precisely according to heuristic rules, such as avoiding industrial areas or water bodies for households. Our work is complementary to this research: it enables the assignment of geographic coordinates to agents in a principled manner even when map information is not available, provided that a sample with localization data exists.

In the wide context of synthetic data, the evaluation of the generated samples is an active field of research. Synthetic data are used for different purposes, such as data sharing [8], improvement of downstream models [20], and fairness [55]. As synthetic data are required to be similar to real data, univariate and multivariate distance measures are employed [17, 23, 33]. Along with fidelity, privacy is a major concern, as many measures can be defined with different assumptions on the data or model availability to the attacker [16, 18, 43]. As in our work, utility is a common choice for assessing the quality of the synthetic data [31, 33, 52]. Additionally, synthetic data are evaluated in terms of expressivity, efficiency, diversity, and generalization [5, 18].

## 2 PROBLEM STATEMENT

We consider a dataset  $D$  of size  $N^R$ , consisting of variables  $x_1, \dots, x_n$ . Among them,  $x_1$  and  $x_2$  correspond to latitude and longitude, while the remaining features may be numeric, integer, categorical, or boolean. Our goal is to develop a generative model  $\mathcal{G}$  that produces a synthetic dataset  $\tilde{D} := \mathcal{G}(D, N^S)$  of size  $N^S$  with the same variables of  $D$ . Specifically,  $\tilde{D}$  should satisfy the following properties:

- **Fidelity.** Synthetic data should be similar to the original data. For a distance function  $d$ , we require  $d(D, \tilde{D})$  to be small. The distance function can be defined in multiple ways, depending on whether the emphasis is on spatial coordinates, other features, or their combination.
- **Utility.** Synthetic data should be useful to draw realistic conclusions about the original data. If  $m_D$  denotes a model trained on  $D$ , and  $m_{\tilde{D}}$  the same model trained on  $\tilde{D}$ , then their output on

a common input should be similar. Formally, for a distance  $d$  defined on the output space of  $m$ , we require  $d(m_D(D), m_{\tilde{D}}(D))$  to be small.

- **Privacy.** Synthetic data should not reveal sensitive information about individuals in  $D$ . We partition  $D$  in  $D^1$  and  $D^2$ .  $D^1$  is used to train the generative model, i.e.  $\tilde{D}^1 := \mathcal{G}(D^1, N^S)$ . Given a distance function returning the minimal distance between a sample  $p$  and the elements of a population, privacy requires that a classifier cannot determine whether  $p$  was used in the training of the generative model. In other terms, considering  $p_1 \in D^1$  and  $p_2 \in D^2$ , the distances  $d(p_1, \tilde{D}^1)$  and  $d(p_2, \tilde{D}^1)$  do not help a classifier detecting if  $p_1 \in D^1$ . This approach is in line with the Membership Inference Attacks (MIA) [18, 51, 53] and evaluates if the generative model unintentionally exposes sensitive information contained in the real data.

To address this problem, we first formalize the design of  $\mathcal{G}$ . Afterwards, we propose meaningful distance measures and predictive models that operationalize fidelity, utility, and privacy. Finally, we compare  $\mathcal{G}$  against competitive generative approaches across these evaluation criteria.

## 3 METHODS

We develop a model that takes as input a dataset of units, each associated with spatial coordinates. These units may represent homes, as in our case studies, or any type of geolocated entity. Each unit can be described by categorical, real, or discrete features. The output of the model is a synthetic dataset of variable size that reproduces the joint distributions of the features with realistic spatial pattern, without replicating the original data.

### 3.1 Preliminaries

Our proposal NF+VAE is a generative model that combines Normalizing Flows (NF) and Variational Autoencoders (VAE). In this section, we introduce these two frameworks. We point to [47] and [21, 36] for more extensive reviews in the topics.

**Normalizing Flows.** Normalizing Flows (NF) are a class of models that transform a simple distribution  $Z$  into a more complex distribution  $X$  through a sequence of  $K$  invertible mappings  $f_i$ , each parameterized by  $\theta$ . Their invertible structure enables efficient mapping in both directions, from  $Z$  to  $X$  and vice versa. The overall transformation can be written as  $x = f_\theta(z) = f_K \circ \dots \circ f_2 \circ f_1$  [47], where  $z \sim Z$ . Since  $f_i$  is bijective, it is possible to use the change of variable formula,

$$p_\theta(x) = p_\theta(z) \prod_{i=1}^K \left| \det \left( \frac{\partial f_i^{-1}}{\partial z_i} \right) \right| = p_\theta(z) \left| \det \left( \frac{\partial f^{-1}}{\partial x} \right) \right|, \quad (1)$$

where  $z_i = f_i(z_{i-1})$ . In practice,  $f_i$  are chosen so that the Jacobian determinants in Equation (1) are efficient to compute, ensuring tractable log-likelihoods. Training proceeds by minimizing the KL divergence between the data distribution and the flow-based model. In such a way, the model learns  $\theta$  and, consequently, the bijective mapping between the base and data distributions. The forward pass maps data into the simple base space, while the backward pass maps samples from the base back to the real data distribution. NF are widely adopted with different goals. In generative modeling,

they allow to generate new samples by drawing  $z \sim Z$  and applying  $f_\theta(z)$ , while in variational inference they provide expressive approximate posteriors when the true posterior is unknown. In our settings, Normalizing Flows represent flexible frameworks for capturing complex geographic patterns in a simple latent space. A wide array of flow architectures have been proposed, including Planar and Radial Flows, Autoregressive Flows, Piecewise Linear and Piecewise Quadratic Flows [37]. In this work, we adopt Neural Spline Flows [22], which are invertible transformations based on monotonic rational-quadratic splines, offering both flexibility and computational tractability.

**Variational Autoencoders.** Variational Autoencoders (VAE) are another class of generative models widely adopted in machine learning [36]. VAE rely on two components, an encoder  $\mathcal{E}$  and a decoder  $\mathcal{D}$ . The encoder  $\mathcal{E}$  maps the original data distribution  $X$  into a lower-dimensional latent space with a simple prior distribution (typically Gaussian),  $Z = \mathcal{E}(X)$ . The decoder  $\mathcal{D}$  reconstructs the samples from the latent space back into the data space,  $\tilde{X} = \mathcal{D}(Z)$ . Both  $\mathcal{E}$  and  $\mathcal{D}$  are neural networks, whose parameters are learned during a training phase. These neural networks are trained by minimizing a loss function that is composed of the reconstruction loss  $\mathcal{L}_R$  and the KL-divergence  $\mathcal{L}_{KL}$ . On the one hand,  $\mathcal{L}_R$  measures the discrepancy between  $X$  and  $\tilde{X}$  to ensure the generated samples resemble the original. On the other hand,  $\mathcal{L}_{KL}$  regularizes the latent representation by aligning  $Z$  with the chosen prior distribution. Since the latent space has dimension that is lower than the data space, the encoding of  $\mathcal{E}(X)$  implies a discard of information. As a result, the decoder learns to generate new samples that are similar to  $X$ , but not identical.

### 3.2 Generative Model

Our proposed model, NF+VAE, generates synthetic geolocated data points by combining NF and VAE. Let  $D$  denote the target  $n$ -dimensional dataset, and  $D_{(x,y)} \in \mathbb{R}^2$  the subset of  $D$  containing only the geographic coordinates. We first train a NF that maps  $D_{(x,y)}$  into a simpler latent distribution,  $Q^{NF} \in \mathbb{R}^2$ . This step is crucial, because geographic coordinates usually encode highly complex spatial patterns associated with natural and artificial constraints, such as mountains, lakes, urban barriers, and areas with varying population densities. Thus, the NF transforms  $D_{(x,y)}$  into latent coordinates  $Z_{(x,y)}$ .

We then construct  $D^{NF}$ , which is a copy of  $D$ , where  $D_{(x,y)}$  is replaced by  $Z_{(x,y)}$ . Next, we train a VAE to generate synthetic data resembling  $D^{NF}$ . The encoder maps  $D^{NF}$  into a  $k$ -dimensional Gaussian latent space, with  $k < n$ , while the decoder reconstructs the original space. We define the loss of the VAE as  $\mathcal{L}_{VAE} = \alpha_{GEO} \mathcal{L}_{GEO} + \alpha_R \mathcal{L}_R + \alpha_{KL} \mathcal{L}_{KL}$ , where (i)  $\mathcal{L}_{GEO}$  is the Euclidean distance between the geographic coordinates of the original and the synthetic data, (ii)  $\mathcal{L}_R$  is the Euclidean distance between all other features, and (iii)  $\mathcal{L}_{KL}$  is the KL-divergence between the encoded data and the Gaussian distribution. Since small errors of the geographic coordinates in the synthetic data lead to the generation of implausible samples, we decided to weight geographic features more than the others features, by setting  $\alpha_{GEO} > \alpha_R$ . Given the granularity of our datasets, we are not interested in generating

samples that differ too much from the available ones. For this reason, we prioritize realistic reconstruction over excessive variation,  $\alpha_R > \alpha_{KL}$ . Once  $\alpha = (\alpha_{GEO}, \alpha_R, \alpha_{KL})$  and the hyperparameters of the neural networks are chosen, we can train the model.

In order to generate  $N^S$  synthetic observations, we draw  $N^S$   $k$ -dimensional Gaussian samples and feed them to the decoder. The trained decoder maps these samples to the output  $n$ -dimensional space, generating  $\tilde{D}^{NF}$ . Therefore, we apply the trained NF to map  $\tilde{D}_{(x,y)}^{NF}$ , which is distributed as  $Q^{NF}$ , back to the realistic geographic coordinates. This yields to the final synthetic dataset  $\tilde{D}$ , containing  $N^S$  observations and the same variables of  $D$ . Thanks to the design of the loss, the generated samples are jointly distributed similar to the original data. However, given the compression of information induced by the encoder of the VAE, the generated samples are not identical to the original ones.

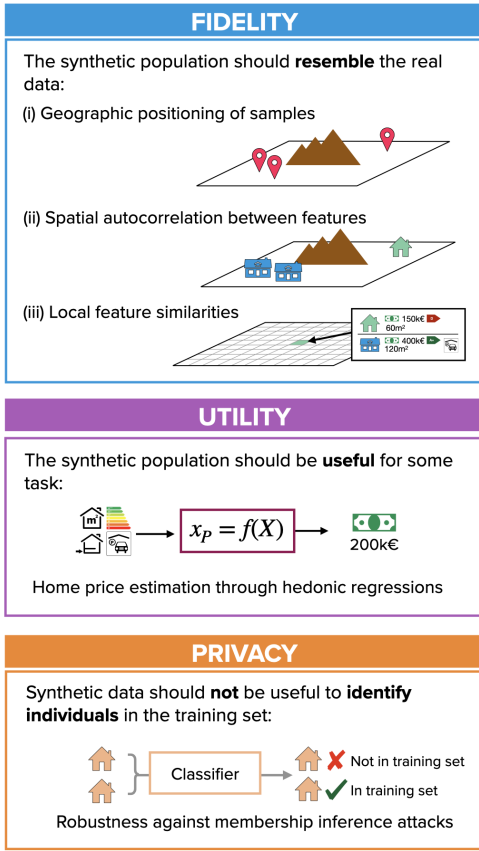
### 3.3 Evaluation setup

We design an evaluation pipeline (see Figure 2) to compare the synthetic populations generated by our proposed method against a set of benchmark models. Following prior works on synthetic data [18], we evaluate the quality of synthetic data along three dimensions, which are *fidelity*, *utility*, and *privacy*. First, the synthetic data must closely resemble the original data, reproducing both distributional patterns and correlations among variables. Given the inherent differences between geographic coordinates and the other features, we adopt three complementary measures of fidelity based on (i) geographic coordinates, (ii) spatial autocorrelation, and (iii) local features. In SM, we also assess the fidelity of correlations between non-geographic features. Second, the synthetic data must be useful, which means that the analyses performed on them must yield insights comparable to those derived from the original data. In practice, this implies that models trained on synthetic data should provide results consistent with models trained on real data. Third, synthetic data must preserve privacy. Consequently, the synthetic data should not permit the recovery of individual information from the original dataset.

**Fidelity - Geographic coordinates.** To assess the similarity of the distributions of geographic coordinates, we adopt the sliced-Wasserstein distance (*SW*) [42]. The Wasserstein distance, also known as Earth mover’s distance, builds on Optimal Transport theory, and it quantifies the minimal cost of transforming one distribution into another. Given two probability measures  $\mu$  and  $\nu$  on a metric space  $(X, d)$ , the  $p$ -Wasserstein distance is the minimal expected cost of transporting mass to transforms  $\mu$  into  $\nu$ , where the cost is measured as the  $p$ -th power distance between points. However, for more than one dimension, computing the Wasserstein distance becomes computationally expensive. The sliced-Wasserstein distance mitigates the issue, by maintaining the properties of the classical Wasserstein distance, while significantly reducing the computational cost [38]. The sliced-Wasserstein distance, *SW*, projects the high-dimensional distribution onto many one-dimensional subspaces and averages the Wasserstein distance between such projections [14]. Relying on previous work, we set  $p = 2$  [27].

The distance used to measure the fidelity of the geographic coordinates is

$$d_G^F = SW \left( D_{(x,y)}, \tilde{D}_{(x,y)} \right). \quad (2)$$



**Figure 2: Description of the evaluation framework, based on fidelity, utility, and privacy. Fidelity measures (i) the similarity of the distribution of geographic coordinates, (ii) the similarity of the spatial autocorrelations, and (iii) the similarity of the houses generated in each grid cells. Utility assesses the quality of a model trained on synthetic data in predicting the house prices in real data. Privacy measures the robustness against membership inference attacks.**

**Fidelity - Spatial autocorrelation.** Beyond geographic coordinates, synthetic data should preserve feature correlations, particularly spatial autocorrelations. We measure this using Moran’s Index [40], which measures the correlation between the spatial proximity and the similarity of a variable of interest  $x$  for each pairs of data points. Spatial proximity is measured through a weight matrix  $W = (w_{ij})_{i,j=1}^N$ , which assigns higher weights to pairs of observations that are closer with each other. In our study, we adopt a distance-based weight  $w_{i,j} = \mathbb{1}_{d(x,y) < m}$ , where  $d(i,j)$  is the Euclidean distance between the geographic coordinates and  $m$  the first percentile of all pairwise distances  $d(x,y)$ . In SM, we report a robustness analysis with an alternative weighting function.

The computation of spatial autocorrelation for high-dimensional data is an open problem, and we define a measure based on Principal Components (PCs). From the real data, we remove the geographic coordinates and we extract the first PCs,  $y_1, \dots, y_l$ , explaining 95%

of variance, then project the synthetic data onto the same subspace. For each  $y_j$  we compute Moran’s Index for the real data,  $I_j$ , and for the synthetic data,  $\tilde{I}_j$ . We then compute Moran’s Index of real data as  $I = \sum_{j=1}^l \lambda_j I_j$ , where  $\lambda_j$  is the eigenvalue associated with the  $j$ -th PC. Analogously,  $\tilde{I} = \sum_{j=1}^l \lambda_j \tilde{I}_j$  is the Moran’s Index of the synthetic data. The distance measure used to quantify the fidelity of the spatial autocorrelation is

$$d_S^F = |I - \tilde{I}|. \tag{3}$$

This definition of  $d_S^F$  allows to reduce the number of features, while assigning more weight to the most explanatory components.

**Fidelity - Local features.** Subsequently, we evaluate the fidelity of the features at a local scale. Similar to previous work on synthetic tabular data [34], we lay on Principal Component Analysis (PCA) to evaluate the fidelity of the generative model. As done for computing  $d_S^F$ , we remove latitude and longitude, we extract the PCs  $y_1, \dots, y_l$  from the real data explaining 95% of variance, and then we project the synthetic data onto the same subspace. Following methods from climate science and geography [9, 35], we partition the study region into a uniform grid of step size  $0.01^\circ$  (approximately 1km). For each cell  $c_h$  in the grid, we compute  $y_j^h$  and  $\tilde{y}_j^h$ , which are the projections on  $y_j$  of the real and synthetic datasets, after filtering only the observations falling within  $c_h$ . We compute  $d_{L,h}^F = \sum_{j=1}^l \lambda_j (y_j^h - \tilde{y}_j^h)^2$ , where  $\lambda_j$  is the eigenvalue associated with the  $j$ -th PC. Finally, we average  $d_{L,h}^F$  across the set of grid cells  $C'$ , which are the cells containing more than zero units both in real and synthetic data.

$$d_L^F = \frac{1}{|C'|} \sum_{h: c_h \in C'} d_{L,h}^F \tag{4}$$

In this way, we measure the fidelity of the synthetic data as the average distance between the average real observation in a grid cell and the average synthetic observation in the same cell. By computing  $d_L^F$  only on  $C'$ , we exclude all empty cells in any of the two datasets. Our motivations for following this choice are that (i) cells that are not in  $C'$  are not useful to evaluate the feature similarity between the two datasets, and (ii) spatial differences across the dataset are already measured in  $d_G^F$ .

**Utility.** In order to assess the utility of the synthetic populations, we want to quantify the discrepancy between the output of a model trained on real data and the output of the same model trained on synthetic data. This approach is also known as Train on Synthetic Test on Real (TSTR) [24].

Relying on a typical real estate context, we develop a hedonic regression model that estimates the home price from the other features [49]. In the hedonic regression, we include the spatial fixed effects, which quantify the effect of being in a specific subregion on the home price. To account for this fixed effect, we include the categorical variable  $x_s$  encoding the subregion associated with each observations, among the  $K$  available subregion.

Let  $x_p$  be the home price, then we write the hedonic regression task for home  $i$  as

$$\log x_p^i = \sum_{j \neq p} \omega_j x_j^i + \sum_r \mu_r \mathbb{1}_{x_s^i = r} + \epsilon^i, \tag{5}$$

where  $\mu_r$  is the spatial fixed effect related to subregion  $r$ ,  $\epsilon^i$  is the error term, and  $\omega_j$  the regression coefficient for variable  $j$ . We call  $m_D$  the model (5) trained on the real data  $D$ , while  $m_{\tilde{D}}$  denotes the model trained on the synthetic data  $\tilde{D}$ .

To assess the utility of a synthetic population, we use  $m_D$  and  $m_{\tilde{D}}$  to predict the home prices on  $D$ , and then we compare the performances of the two models. To this end, we denote with  $m_D(D)$  the home prices predicted by  $m_D$  and  $m_{\tilde{D}}(D)$  the home prices predicted by  $m_{\tilde{D}}$ . Subsequently, we compute  $R^2(m_D) = R^2(m_D(D), x^p)$  and  $R^2(m_{\tilde{D}}) = R^2(m_{\tilde{D}}(D), x^p)$ , where  $R^2$  is the coefficient of determination. Finally, the distance measure used to quantify the utility of the synthetic data is

$$d^U = |R^2(m_D) - R^2(m_{\tilde{D}})|. \quad (6)$$

If  $d^U$  is close to zero, this means that the synthetic data are as useful as real data to accomplish the task of predicting home prices with hedonic regression with spatial fixed effects.

**Privacy.** Many different measures have been proposed to quantitatively assess the privacy of a synthetic dataset [18]. The general goal is to ensure that synthetic data cannot be used to recover sensitive information about individuals in the original dataset. In this work we focus on the risk of membership inference [51, 53], where an adversary attempts to determine whether a particular record was used to train the generative model. Such attacks are concerning because, in some cases, the presence of a record in the training set may itself reveal sensitive information. For example, in our settings, confirming that a home appears in the training implies that the bank issued a mortgage to the homeowner.

To quantify this risk, we design the following procedure. First, we define  $D^1$  as a random sample of 95% of  $D$ . Second, we generate a synthetic population  $\tilde{D}^1$  by training the generative model on  $D^1$ . For any sample  $p \in D$ , we define  $d(p, \tilde{D}^1)$  as the minimum Euclidean distance between  $p$  and the closest  $y \in \tilde{D}^1$ . The distance considers all spatial and non-spatial features, after standard one-hot encoding and rescaling within  $[0, 1]$ . Third, we construct two disjoint subsets,  $Z^{TRAIN}$  and  $Z^{TEST}$ , by splitting  $D$  with a ratio 80-20. Using  $Z^{TRAIN}$ , we train a classifier  $C$  that predicts whether a record  $p$  belongs to  $D^1$  based only on  $d(p, \tilde{D}^1)$ . Fourth, we evaluate  $C$  on  $Z^{TEST}$  using Area Under ROC Curve (AUC-ROC). The AUC-ROC is commonly used in imbalanced datasets, and it measures the probability that a classifier assigns to a random positive a score higher than a random negative. Thus, the privacy preservation of the synthetic population is measured as

$$\rho^P = \text{AUC-ROC}(C(Z^{TEST})) - 0.5. \quad (7)$$

In this way, if  $\rho^P > 0$ ,  $C$  may infer whether a record was part of the training dataset better than random classifier. Since this classification is done by simply matching the record with the most similar observation in the synthetic dataset, this definition strongly relates to the notion of privacy.

We considered a logistic regression as the classifier  $C$  in the main text. In SM, we show that the results are robust considering other classifiers.

## 4 EXPERIMENTAL SETTINGS

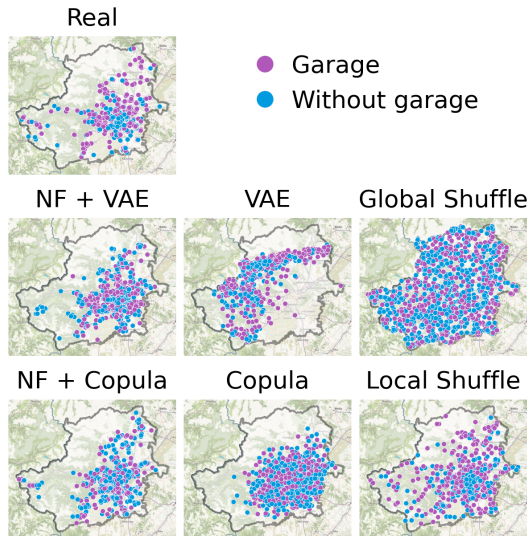
**Data description.** We evaluated the generative model using two buckets of dataset, `data_isp` and `data_airbnb`.

`data_isp` is a unique dataset owned by Intesa Sanpaolo (ISP), one of the leading commercial banks in Italy. Specifically, we got access to the data about the mortgages issued by Intesa Sanpaolo between January 2016 and August 2024 (all data were treated with a GDPR-compliant and privacy preserving protocol). For each mortgage, we have the information about the home given as collateral that is the input of the generative model. The dataset contains 14 features, 4 features are numeric, 6 are boolean, and the remaining 4 are categorical. Numerical features include latitude, longitude, surface, and sale price, while energy class, cadastral code, floor, and construction year are categorical features. Boolean features include presence of air conditioning, annex (i.e., basement or rooftop storage room), and garage among others. The dataset contains 549247 homes across 106 Italian provinces. We built one dataset for each province containing the homes in the province at hand. 7 provinces have  $< 1,000$  homes, 74 provinces have between 1,000 and 5,000 homes, 11 provinces range between 5,000 and 10,000, and 14 provinces have  $> 10,000$  homes. In SM, we provide additional details of the features in `data_isp`.

Secondly, we evaluate the method on `data_airbnb`, which lay on a public dataset of Airbnb provided by Inside Airbnb [2]. Airbnb is a widely used online marketplace for short-term rentals, where hosts describe their properties through a variety of attributes such as location, number of rooms, available amenities, and price per night. We selected a sample of 15 cities across different countries and, for each city, we created a dataset comprising 1 categorical, 7 binary, 4 discrete, and 5 continuous features. A detailed description of `data_airbnb` is provided in SM.

**Benchmarks.** We compare our proposed method, **NF+VAE** (see SM for implementation details), with a set of benchmarks of particular interest.

- **VAE.** As an ablation study, we generated synthetic data with VAE, without transforming the geographic coordinates. In this case, we use the same VAE architecture of NF+VAE.
- **Copula.** We generate synthetic populations with Gaussian copula, relying on the implementation of `sdv` [48]. This method fits a multivariate Gaussian distribution in a latent space, then uses marginal inverse-CDF transformations to map latent samples into the original data domain. In such a way, it maintains marginal distributions while approximating the correlations through the Gaussian copula [41].
- **NF+copula.** We transform the geographic coordinates with the same NF used in NF+VAE, but replacing VAE with Gaussian copula. In such a way, we evaluate also the contribution of NF applied on a different generative model.
- **Global shuffle.** This method samples the original observations with replacement and then assigns geographic coordinates uniformly random within the region covered by the dataset. In our case, the regions of the synthetic populations are the Italian provinces (`data_isp`) and the cities (`data_airbnb`).
- **Local shuffle.** Similar to the previous method but it samples with replacement the observations in the original dataset, while



**Figure 3: Real and synthetic homes generated by the benchmark generators in the province of Turin. In this plot, for each map we show a random sample of 1,000 homes, colored by the presence of garage.**

assigning random coordinates within the boundaries of the same granular subregion of the sampled observation. The subregions in `data_isp` are the postcode areas, while the subregions of `data_airbnb` are the neighborhoods.

By choosing these benchmarks, we compare NF+VAE with a natural competitor (NF+copula), two ablated models (VAE and copula), and two non-trivial null models (Global shuffle and Local shuffle). For all benchmarks, we filter out the synthetic data points that fall outside the borders of the region by matching them against the region’s shapefile. In Figure 3, we show a visual representation of a synthetic population for each method.

## 5 RESULTS

We evaluate model performance following the validation pipeline described in Section 3.3, along the dimensions of fidelity, utility, and privacy. Our main results show the performance of the model when applied to `data_isp`, while we include an equivalent validation analysis in the SM using `data_airbnb`.

**Fidelity.** First, we focus on the fidelity of geographic coordinates, by computing the sliced-Wasserstein distance with 1000 projections. As shown in Figure 4a, the methods that use NF achieve performances comparable to Local shuffle (median  $d_G^F$ : 0.022 for NF+VAE, 0.009 for NF+copula, and 0.024 for Local shuffle). By contrast, VAE fails to adequately capture the geographic distribution (median  $d_G^F$ : 0.095). This highlights the importance of transforming geographic coordinates through a trained NF, before using them in the generative model. Second, when evaluating the spatial autocorrelations (Figure 4b), we observe that the VAE-based methods capture spatial dependencies better than copula-based methods (median  $d_S^F$ : 0.028 for NF+VAE, 0.080 for NF+copula, and 0.043 for Local shuffle). In SM we show that the results of Figure 4b are consistent by choosing a different weighting function in the computation of

Moran’s I. Third, in the local features fidelity (Figure 4c) NF+VAE and NF+copula outperform the other methods (average distances: 0.391 for NF+VAE, 0.409 for NF+copula, and 0.410 for Local shuffle). These results indicate that both methods generate plausible homes within individual grid cells.

**Utility.** We use hedonic regressions, where the spatial fixed effects are determined by the postcode areas for `data_isp`, and the neighborhoods for `data_airbnb`. Since Local shuffle replicates the exact same features of the original data, it achieves the highest utility. However, when comparing the remaining methods, we observe that NF+VAE and copula outperform Global shuffle (median  $d^u$ : 0.196 for NF+VAE, 0.283 for NF+copula, 0.163 for copula, 0.206 for Global shuffle). Since Global shuffle merely replicates homes with random locations, the high utility achieved by NF+VAE (and copula) supports the importance of correctly replicating the spatial distributions of samples in building reliable model. In the SM we further validated the utility of the synthetic data considering a classification task.

**Privacy.** Finally, we compare the privacy robustness against a Logistic Regression classifier for the analyzed benchmarks, Figure 4e. We note that Local shuffle is significantly larger than 0, showing weak robustness against membership inference attacks. This means that it is possible to classify a data point as part of the training set based on the closest synthetic data point. All other methods are robust against the classifier, as the privacy measure is not significantly different from zero. In SM we provide an extension of Figure 4e, showing that these privacy robustness is consistent across a set of standard machine learning classifiers.

**Summary.** By looking at the overall results, we can detect the following insights.

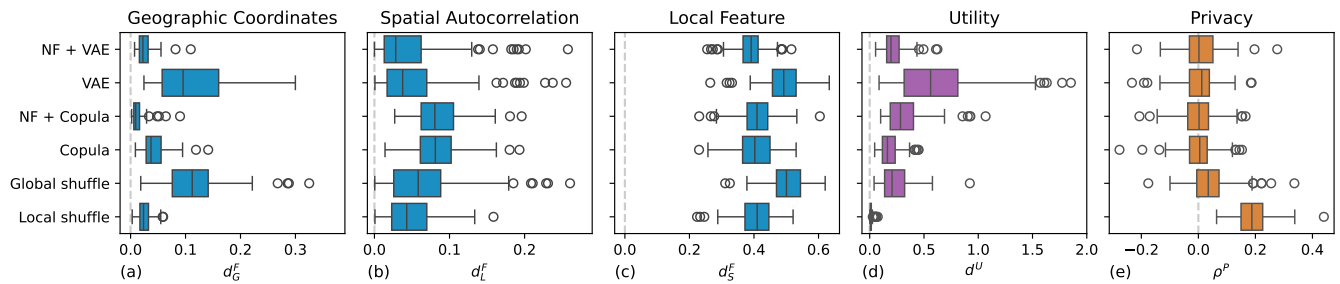
- The study of the ablated models reveal the importance of NF in learning the spatial distributions.
- VAE-based models capture the spatial autocorrelations, thus favoring NF+VAE over NF+copula.
- NF+VAE also produces synthetic populations that are useful for downstream models.
- All non-shuffle models are robust privacy attacks, thus excluding Local shuffle.

Overall, our experimental analyses show the superiority of NF+VAE in the combination of fidelity, utility, and privacy. In SM we show that similar conclusions hold for `data_airbnb`. However, in `data_airbnb` we also notice that NF+copula achieves performances that are comparable to NF+VAE.

## 6 DISCUSSION

We proposed and evaluated a method for generating geolocated synthetic populations that combines Normalizing Flows (NF) and Variational Autoencoders (VAE). Our evaluation pipeline is based on fidelity, utility, and privacy. Using this pipeline, we assess the quality of our model, NF+VAE, across an extensive set of datasets (106 in `data_isp` + 15 in `data_airbnb`). Overall, our approach improves other competitors in terms of fidelity and utility, while maintaining the privacy in the generation of geolocated units.

**Flexibility.** Our method requires minimal customization, beyond standard preprocessing and the definition of network architectures.



**Figure 4: Distributions of evaluation metrics in data\_isp.** (a) *Fidelity - Geographic coordinates*, i.e., sliced-Wasserstein distance geographic coordinates, (b) *Fidelity - Spatial autocorrelation*, distance between spatial autocorrelations in the PCs, (c) *Fidelity - Local features*, distance between average home per spatial grid cell, (d) *Utility*, distance between  $R^2$  in predicting real log-price with a model trained with real and synthetic data, (e) *Privacy*, difference between AUC-ROC of a classifier trained to infer the membership in the original dataset. Best performances are close to 0 in all methods. Detailed statistics in SM.

NF+VAE easily handles variables of different types and dimensions. Additionally, it requires only the region shapefile, without the need of additional geographic information, such as subregion boundaries, streets, or points of interest.

*Zero-cell problem.* While traditional sampling-based methods struggle to reproduce combinations of features that are not present in the real data, NF+VAE generates novel but realistic combinations. As shown in SM, Local shuffle only replicates existing samples, thus failing to address zero-cell problem- Contrarily, NF+VAE produces a median of 0.65 homes that are not present in the real data. This strength characterizes this class of generative models: just as in computer vision VAE produces realistic images not present in the training data, NF+VAE can synthesize homes that are plausible but not present in the original data.

*Generalization.* Although this work is motivated by housing market analysis, the method naturally generalizes to many other domains. In ABMs, geolocated populations can represent diverse spatial entities, such as households at their residences, workers at their workplaces, or students at their schools. Since privacy concerns usually restrict the use of collected data, making synthetic data represents a compelling alternative.

*Towards standard evaluation.* To our knowledge, this is the first systematic comparison of synthetic population generators for geolocated data. In accordance with the usual practices in machine learning, we underline the importance of rigorous evaluation of the proposed method, with a systematic benchmark against representative baselines. We hope that this work fosters further research on the evaluation of geographic synthetic data.

**Limitations.** Despite these strengths, several limitations remain.

*Model implementation.* Since NF+VAE combines two deep learning approaches, it reflects a set of modeling and optimization choices. Although we demonstrated that our design meets the initial requirements, refined implementations of NF and VAE could further improve the quality of the generated data.

*Evaluation metrics.* Our evaluation framework involves a set of arbitrary choices to measure fidelity, utility, and privacy. The adopted definitions and metrics are not exhaustive, and they can vary with the application. For example, we evaluated utility based on the regression of house prices, while alternative measures might be more appropriate in other domains. In the case of ABMs, the utility could

compare the ABM outputs when initialized with the real and synthetic populations. Similarly, we focused on membership inference as a privacy attack, but alternatives could provide equally relevant insights. For instance, attribute inference attack aims to recover sensitive individual attributes from the synthetic data and partial knowledge of the real data. Given the absence of a unique superior standard for privacy in synthetic data, we recognize that evaluation must be tailored to the intended use case. Another possibility to address privacy is to use generative models relying on differential privacy. Thus, the level of privacy can be defined in advance.

*Evaluation desiderata.* In addition to fidelity, utility, and privacy, the evaluation could include other features, such as efficiency and expressivity [18]. Regarding efficiency, we acknowledge that training NF+VAE requires substantially more computational resources than simpler approaches, such as copula-based or shuffle-based generators. Our results, however, show that replacing the VAE with copula maintains comparable performances and significantly reduces the training cost. Still, NF remains crucial for transforming geographic coordinates in a convenient representation. As for expressivity, given the strong capabilities demonstrated by VAE in other domains, such as image generation, we expect NF+VAE to exhibit a comparable expressive power in population synthesis.

*Benchmarks.* The inclusions of additional baselines, such as Combinatorial Optimization, Generative Adversarial Networks (GANs), or Bayesian Networks, would provide a more comprehensive assessment. Our benchmarks comprises (i) a simple and valid alternative (NF+copula), (ii) the ablated models to isolate the effect of NF, and (iii) the shuffle-based models, acting as null models.

*Data availability.* Finally, the effectiveness of NF+VAE depends on data availability. Both VAE and NF perform best with large, high-dimensional datasets, and may struggle with overfitting or weak pattern learning when data are limited. However, given the increasing availability of granular real-world datasets, we believe our approach is applicable in a wide range of context.

**Reproducibility.** All code used for model training and evaluation is available at <https://github.com/jacienti/NFVAE-population-synthesis>. Due to strict privacy regulations governing financial data, we cannot release data\_isp either at the individual or aggregate level. However, all analyses conducted on the public available data\_airbnb are fully reproducible.

## REFERENCES

- [1] AirBnB. <https://www.airbnb.com>.
- [2] InsideAirBnB. <https://www.insideairbnb.com>.
- [3] Abhijin Adiga, Aditya Agashe, Shaikh Arifuzzaman, Christopher L. Barrett, Richard J. Beckman, Keith R. Bisset, Jiangzhuo Chen, Youngyun Chungbaek, Stephen G. Eubank, Sandeep Gupta, et al. 2015. Generating a synthetic population of the United States. *Network Dynamics and Simulation Science Laboratory, Tech. Rep. NDSSL* (2015), 15–009.
- [4] Zack Aemmer and Don MacKenzie. 2022. Generative population synthesis for joint household and individual characteristics. *Computers, Environment and Urban Systems* 96 (2022), 101852.
- [5] Ahmed Alaa, Boris Van Breugel, Evgeny S. Saveliev, and Mihaela Van Der Schaar. 2022. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In *International conference on machine learning*. PMLR, 290–306.
- [6] Alberto Aleta, David Martín-Corral, Michiel A. Bakker, Ana Pastore y Piontti, Marco Ajelli, Maria Litvinova, Matteo Chinazzi, Natalie E. Dean, M. Elizabeth Halloran, Ira M. Longini Jr, et al. 2022. Quantifying the importance and location of SARS-CoV-2 transmission events in large metropolitan areas. *Proceedings of the National Academy of Sciences* 119, 26 (2022), e2112182119.
- [7] Theo Arentze, Harry Timmermans, and Frank Hofman. 2007. Creating synthetic household populations: Problems and approach. *Transportation Research Record* 2014, 1 (2007), 85–91.
- [8] Samuel A. Assefa, Danial Dervovic, Mahmoud Mahfouz, Robert E. Tillman, Prashant Reddy, and Manuela Veloso. 2020. Generating synthetic data in finance: opportunities, challenges and pitfalls. In *Proceedings of the First ACM International Conference on AI in Finance*. 1–8.
- [9] Maximilian Auffhammer and Wolfram Schlenker. 2014. Empirical studies on agricultural impacts and adaptation. *Energy Economics* 46 (2014), 555–561.
- [10] Johan Barthélemy and Philippe L. Toint. 2013. Synthetic population generation without a sample. *Transportation Science* 47, 2 (2013), 266–279.
- [11] Richard J. Beckman, Keith A. Baggerly, and Michael D. McKay. 1996. Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice* 30, 6 (1996), 415–429.
- [12] Anna Bellaver, Lorenzo Costantini, Ariadna Fosch, Anna Monticelli, David Scala, and Marco Pangallo. 2025. Floods do not sink prices, historical memory does: How flood risk impacts the Italian housing market. *arXiv preprint arXiv:2502.12116* (2025).
- [13] Nathan Blackthorn, Andrew Arash Mahyari, and Ashok Srinivasan. 2024. Training Variational Autoencoders for Population Synthesis in Public Health with Missing Data. In *2024 IEEE International Conference on Big Data (BigData)*. IEEE, 4969–4973.
- [14] Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. 2015. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision* 51, 1 (2015), 22–45.
- [15] Stanislav S. Borysov, Jeppe Rich, and Francisco C. Pereira. 2019. How to generate micro-agents? A deep generative modeling approach to population synthesis. *Transportation Research Part C: Emerging Technologies* 106 (2019), 73–97.
- [16] Tânia Carvalho, Nuno Moniz, Pedro Faria, and Luis Antunes. 2022. Survey on privacy-preserving techniques for data publishing. *arXiv preprint arXiv:2201.08120* (2022).
- [17] Kevin Chapuis, Patrick Taillandier, Misslin Renaud, and Alexis Drogoul. 2018. Gen\*: a generic toolkit to generate spatially explicit synthetic populations. *International Journal of Geographical Information Science* 32, 6 (2018), 1194–1210.
- [18] Graham Cormode, Samuel Maddock, Enayat Ullah, and Shripad Gade. 2025. Synthetic Tabular Data: Methods, Attacks and Defenses. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*. 5989–5998.
- [19] Jan de Mooij, Tabea Sonnenschein, Marco Pellegrino, Mehdi Dastani, Dick Ettema, Brian Logan, and Judith A. Versteegen. 2024. GenSynthPop: generating a spatially explicit synthetic population of individuals and households from aggregated data. *Autonomous Agents and Multi-Agent Systems* 38, 2 (2024), 48.
- [20] Ayesha Siddiqua Dina, AB Siddique, and D. Manivannan. 2022. Effect of balancing data using synthetic data on the performance of machine learning classifiers for intrusion detection in computer networks. *Ieee Access* 10 (2022), 96731–96747.
- [21] Carl Doersch. 2016. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908* (2016).
- [22] Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. 2019. Neural spline flows. *Advances in neural information processing systems* 32 (2019).
- [23] Erica Espinosa and Alvaro Figueira. 2023. On the quality of synthetic generated tabular data. *Mathematics* 11, 15 (2023), 3278.
- [24] Cristóbal Esteban, Stephanie L. Hyland, and Gunnar Rätsch. 2017. Real-valued (medical) time series generation with recurrent conditional gans. *arXiv preprint arXiv:1706.02633* (2017).
- [25] Bilal Farooq, Michel Bierlaire, Ricardo Hurtubia, and Gunnar Flötteröd. 2013. Simulation based population synthesis. *Transportation Research Part B: Methodological* 58 (2013), 243–263.
- [26] Tatiana Filatova. 2015. Empirical agent-based land market: Integrating adaptive economic behavior in urban land-use models. *Computers, Environment and Urban Systems* 54 (2015), 397–413.
- [27] Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. 2021. POT: Python Optimal Transport. *Journal of Machine Learning Research* 22, 78 (2021), 1–8. <http://jmlr.org/papers/v22/20-451.html>
- [28] Shannon Gallagher, Lee F. Richardson, Samuel L. Ventura, and William F. Eddy. 2018. SPEW: synthetic populations and ecosystems of the world. *Journal of Computational and Graphical Statistics* 27, 4 (2018), 773–784.
- [29] Sergio Garrido, Stanislav S. Borysov, Francisco C. Pereira, and Jeppe Rich. 2020. Prediction of rare feature combinations in population synthesis: Application of deep generative modelling. *Transportation Research Part C: Emerging Technologies* 120 (2020), 102787.
- [30] Toon Haer, WJ. Wouter Botzen, Hans de Moel, and Jeroen CJH. Aerts. 2017. Integrating household risk mitigation behavior in flood risk analysis: an agent-based model approach. *Risk Analysis* 37, 10 (2017), 1977–1992.
- [31] Lasse Hansen, Nabeel Seadat, Mihaela van der Schaar, and Andrija Petrovic. 2023. Reimagining synthetic tabular data generation through data-centric AI: A comprehensive benchmark. *Advances in neural information processing systems* 36 (2023), 33781–33823.
- [32] Alison J. Heppenstall, Andrew T. Crooks, Linda M. See, and Michael Batty. 2011. *Agent-based models of geographical systems*. Springer Science & Business Media.
- [33] Mikel Hernandez, Gorka Epelde, Ane Alberdi, Rodrigo Cilla, and Debbie Rankin. 2022. Synthetic data generation for tabular health records: A systematic review. *Neurocomputing* 493 (2022), 28–45.
- [34] Dayananda Herurkar, Ahmad Ali, and Andreas Dengel. 2025. Evaluating Generative Models for Tabular Data: Novel Metrics and Benchmarking. *arXiv preprint arXiv:2504.20900* (2025).
- [35] Jon A. Kimerling, Kevin Sahr, Denis White, and Lian Song. 1999. Comparing geometrical properties of global grids. *Cartography and Geographic Information Science* 26, 4 (1999), 271–288.
- [36] Diederik P. Kingma, Max Welling, et al. 2019. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning* 12, 4 (2019), 307–392.
- [37] Ivan Kobyzev, Simon JD. Prince, and Marcus A. Brubaker. 2020. Normalizing flows: An introduction and review of current methods. *IEEE transactions on pattern analysis and machine intelligence* 43, 11 (2020), 3964–3979.
- [38] Soheil Kolouri, Kimia Nadjahi, Umot Simssekli, Roland Badeau, and Gustavo Rohde. 2019. Generalized sliced wasserstein distances. *Advances in neural information processing systems* 32 (2019).
- [39] Maxime Lenormand and Guillaume Defuant. 2013. Generating a Synthetic Population of Individuals in Households: Sample-Free Vs Sample-Based Methods. *Journal of Artificial Societies and Social Simulation* 16, 4 (2013), 12.
- [40] Patrick AP. Moran. 1948. The interpretation of statistical maps. *Journal of the Royal Statistical Society. Series B (Methodological)* 10, 2 (1948), 243–251.
- [41] Roger B. Nelsen. 2006. *An introduction to copulas*. Springer.
- [42] Sloan Nietert, Ziv Goldfeld, Ritwik Sadhu, and Kengo Kato. 2022. Statistical, robustness, and computational guarantees for sliced Wasserstein distances. *Advances in Neural Information Processing Systems* 35 (2022), 28179–28193.
- [43] Pablo A. Osorio-Marulanda, Gorka Epelde, Mikel Hernandez, Imanol Isasa, Nicolas Moreno Reyes, and Andoni Beristain Iraola. 2024. Privacy mechanisms and evaluation metrics for synthetic data generation: A systematic review. *IEEE Access* 12 (2024), 88048–88074.
- [44] Marco Pangallo, Alberto Aleta, R. Maria del Rio-Chanona, Anton Pichler, David Martín-Corral, Matteo Chinazzi, François Lafond, Marco Ajelli, Esteban Moro, Yami Moreno, Alessandro Vespignani, and J. Doyne Farmer. 2023. The unequal effects of the health–economy trade-off during the COVID-19 pandemic. *Nature Human Behaviour* (2023), 1–12.
- [45] Marco Pangallo, Matteo Coronese, Francesco Lamperti, Guido Cervone, and Francesca Chiaromonte. 2024. Climate change attitudes in a data-driven agent-based model of the housing market. (2024). In preparation.
- [46] Marco Pangallo and R. Maria del Rio-Chanona. 2025. Data-driven economic agent-based models. In *The economy as an evolving complex system IV*. SFI Press, Santa Fe, N.M., ?
- [47] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. 2021. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research* 22, 57 (2021), 1–64.
- [48] Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. 2016. The synthetic data vault. In *2016 IEEE international conference on data science and advanced analytics (DSAA)*. IEEE, 399–410.
- [49] Sherwin Rosen. 1974. Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of political economy* 82, 1 (1974), 34–55.
- [50] Abdoul Razac Sané, Pierre-Olivier Vandanjon, Rachid Belaroussi, and Pierre Hankach. 2025. A comprehensive investigation of variational auto-encoders for population synthesis. *Journal of Computational Social Science* 8, 1 (2025), 13.

- [51] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 3–18.
- [52] Joshua Snoke, Gillian M Raab, Beata Nowok, Chris Dibben, and Aleksandra Slavkovic. 2018. General and specific utility measures for synthetic data. *Journal of the Royal Statistical Society Series A: Statistics in Society* 181, 3 (2018), 663–688.
- [53] Amy Steier, Lipika Ramaswamy, Andre Manoel, and Alexa Haushalter. 2025. Synthetic data privacy metrics. *arXiv preprint arXiv:2501.03941* (2025).
- [54] Lijun Sun and Alexander Erath. 2015. A Bayesian network approach for population synthesis. *Transportation Research Part C: Emerging Technologies* 61 (2015), 49–62.
- [55] Boris Van Breugel, Trent Kyono, Jeroen Berrevoets, and Mihaela Van der Schaar. 2021. Decaf: Generating fair synthetic data using causally-aware generative networks. *Advances in Neural Information Processing Systems* 34 (2021), 22221–22233.
- [56] Peijun Ye, Bin Tian, Yisheng Lv, Qijie Li, and Fei-Yue Wang. 2020. On iterative proportional updating: Limitations and improvements for general population synthesis. *IEEE Transactions on Cybernetics* 52, 3 (2020), 1726–1735.
- [57] Xin Ye, Karthik Konduri, Ram M Pendyala, Bhargava Sana, and Paul Waddell. 2009. A methodology to match distributions of both household and person attributes in the generation of synthetic populations. In *88th Annual Meeting of the transportation research Board, Washington, DC*, Vol. 36.
- [58] Meng Zhou, Jason Li, Rounaq Basu, and Joseph Ferreira. 2022. Creating spatially-detailed heterogeneous synthetic populations for agent-based microsimulation. *Computers, Environment and Urban Systems* 91 (2022), 101717.