Population synthesis with geographic coordinates
Supplementary Materials

1 DATA DESCRIPTION

1.1 data_isp

data_isp describes the home characteristics of the properties given as collateral for the mortgages issued by Intesa Sanpaolo between January 2016 and August 2024. The dataset contains 549247 samples and a detailed selection of home characteristics that are further described in Supplementary Table 1.

Table 1: Description of features in the Intesa Sanpaolo dataset.

Feature	**Type**	**Description**
Coordinates (x)	Numeric	Longitude of the property used as collateral.
Coordinates (y)	Numeric	Latitude of the property used as collateral.
Surface	Numeric	Logarithm base 10 of the total surface area of the property, measured in square meters.
Price	Numeric	Logarithm base 10 of selling price of the property (in euros).
Flag garage	Boolean	True if the property includes a garage.
Flag annex	Boolean	True if the property has a basement or rooftop storage space.
Flag A/C	Boolean	True if the property is equipped with air conditioning.
Flag missing A/C	Boolean	True if information about air conditioning is unavailable.
Flag multi floor	Boolean	True if the property spans multiple floors.
Flag missing multi floor	Boolean	True if information about the “Multi floor” flag is missing.
Construction year	Categorical	Groups properties into 5 categories based on construction year: (1) 1500–1965, (2) 1965–1985, (3) 1985–2005, (4) 2005–2024, (5) “Missing” for properties without data.
Energy class	Categorical	Groups properties by energy class (in parenthesis): (1) “High” (A1, A2, A3, A4), (2) “Medium” (B, C, D), (3) “Low” (E, F, G); (4) “Missing” if energy class is not available.
Cadastral code	Categorical	Groups properties based on cadastral classification (cadastral code within parenthesis): (1) “Villas and elegant homes” (A01, A07, A08), (2) “Civil homes” (A02), (3) “Standard homes” (A03), and (4) “Popular homes” (A04, A05).
Floor	Categorical	Groups properties by floor level: (1) “Floor 0” (ground), (2) “Floor 1”, (3) “Floor 2”, (4) “Floor 3”, (5) “Floor 4+” (fourth or higher), and (6) “Missing” (no floor data available).

1.2 `data_airbnb`

Meanwhile, `data_airbnb` is a dataset compiled by Airbnb Inside between 2025-03-01 and 2025-06-30, describing the characteristics of accommodations listed across 15 cities worldwide. More specifically, the dataset contains 184347 samples, split among the the different cities as follows: 9,171 in Austin (USA), 11,377 in Barcelona (Spain), 4,454 in Brisbane (Australia), 17,540 in Cape Town (South Africa), 10,878 in Copenhagen (Denmark), 25,133 in Hawaii (USA), 3,010 in Hong-Kong (Hong-Kong), 4,888 in Lyon (France), 19,718 in Mexico city (Mexico), 7,161 in Montreal (Canada), 7,312 in Naples (Italy), 42,130 in Paris (France), 5,398 in Seattle (USA), 1,339 in Singapore (Singapore) and 4,148 Washington DC (USA). Supplementary Table 2 contains more details about the features available in the dataset.

Table 2: Description of features in the Airbnb dataset.

Feature	Type	Description
Coordinates (x)	Numeric	Longitude of the accommodation listed in Airbnb.
Coordinates (y)	Numeric	Latitude of the accommodation listed in Airbnb.
Bedrooms	Numeric	Number of bedrooms in the home proposed for rent.
Beds	Numeric	Number of beds in the home proposed for rent.
Bathrooms	Numeric	Number of beds in the home proposed for rent.
Accommodates	Numeric	Number of guest allowed in the home proposed for rent.
Price per night	Numeric	Logarithm base 10 of the price per night.
Review score	Numeric	Review score rating on Airbnb. The minimum value is 1 and the maximum is 5.
Reviews per month	Numeric	Average number of reviews per month (computed from the publication date of the announcement).
Flag air conditioning	Boolean	True if the home is equipped with air conditioning.
Flag elevator	Boolean	True if the home has an elevator.
Flag self check-in	Boolean	True if self check-in is allowed to access to the home.
Flag pets	Boolean	True if pets are allowed in the home proposed for rent.
Flag private living room	Boolean	True if the guest has a private living room.
Flag pool	Boolean	True if the home is equipped with a pool.
Flag backyard	Boolean	True if the home has a backyard.
Room type	Categorical	Groups homes into 4 categories: (1) Entire home/apartment; (2) Hotel room; (3) Private room; (4) Shared room.

2 IMPLEMENTATION DETAILS

Several modeling choices were required along the analyses. They are described below.

Before training NF, we apply Min-Max normalization to bound the geographic coordinates within $[0, 1]$. We adopted a Uniform distribution as the base distribution Q^{NF}, maintaining the same support of the pre-processed input coordinates. For the flow transformations, we use Neural Spline Flows, i.e., invertible transformations based on monotonic rational-quadratic splines [2]. Our architecture consists of 48 layers with 48 neurons each. Training is performed with the Adam optimizer for 20,000 epochs, with learning rate of 10^{-5}. Hyperparameters are tuned to minimize the Energy distance [3] between the distribution of real data and Gaussian samples transformed through the NF on a subset of our datasets.

Before training VAE, (i) categorical variables are one-hot encoded, and (ii) all features are standardized to their z-scores. In `data_isp`, VAE compresses the data from 29 dimensions to 20, and from 5 to 10 in `data_airbnb`. The encoder is a neural network with layers sizes $[64, 32, 32, 20]$, while the decoder mirrors this structure in the reverse order. The activation function of each neuron in the VAE is ReLU [1]. We train the model with the Adam optimizer for 400 epochs, using a learning rate equal to 10^{-3} . These weights are first tuned to minimize a linear combination of the Energy distance between real and synthetic data and the L^2 distance between their correlation matrices. The remaining hyperparameters are then tuned to minimize the loss function, $\mathcal{L}_{VAE}$ (Section 3.3 SM). After training, synthetic samples are post-processed as follows: (i) integer variables are rounded, (ii) z-scores are inverted, and (iii) categorical variables are recovered from one-hot encoding using maximum activation.

3 ADDITIONAL RESULTS FOR DATA_ISP

Here we provide some additional figures and robustness analyses that complement the results described on the main text.

3.1 Other example maps

Supplementary Figure 1 provides some examples of the performance of NF+VAE for four provinces not shown in the main text. In all cases, NF+VAE captures the geographic patterns in the original data and it reproduces accurately the spatial distribution of the attribute shown, regardless of the type of variable (categorical or continuous).

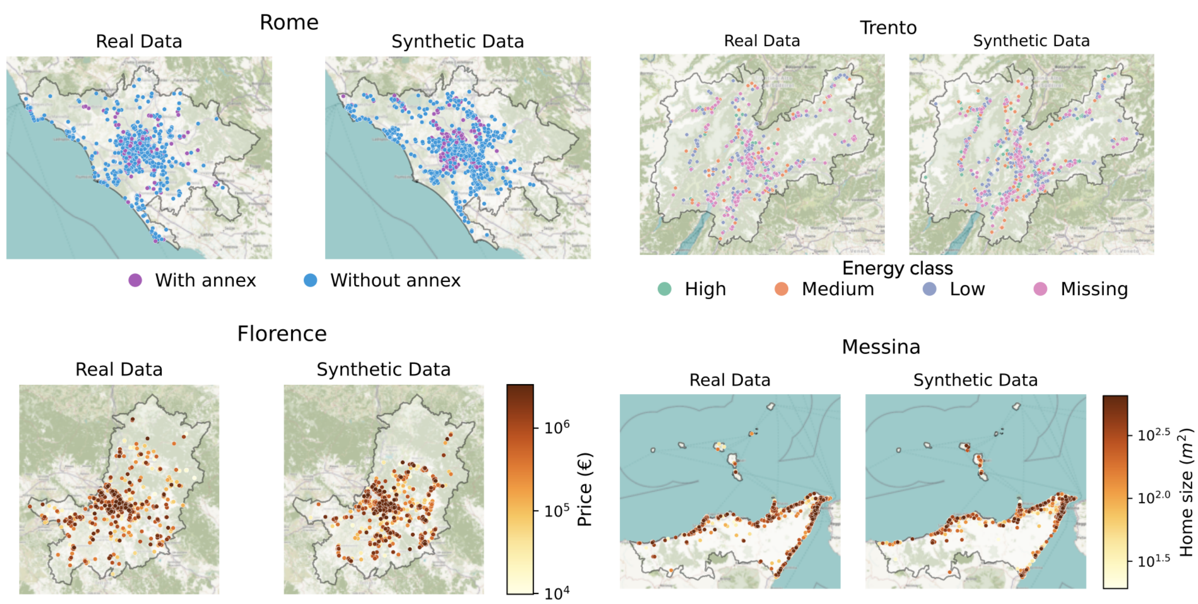

Figure 1: Comparison of the real and synthetic populations for four different provinces with different attributes: binary (i.e., presence of an annex, top-left panels), categorical (i.e., energy class, top-right panels), and continuous –price (bottom-left) and surface (bottom-right). These examples evidence the accuracy of NF+VAE in generating synthetic populations that preserve both the geographic distribution of homes and spatial correlations of their attributes in different contexts.

3.2 Statistics to evaluate boxplots

Supplementary Table 3 reflects the summary statistics of the results shown in Fig. 4 of the main text. We report the extreme values of the distribution across provinces and the 1st, 2nd and 3rd quartiles.

	Geographic Coordinates					
	NF + VAE	VAE	NF + Copula	Copula	Global shuffle	Local shuffle
min-max	[0.007 , 0.109]	[0.024 , 0.582]	[0.002 , 0.090]	[0.009 , 0.141]	[0.018 , 0.325]	[0.003 , 0.059]
Q1(25%)	0.016	0.057	0.006	0.028	0.076	0.017
Median (50%)	0.022	0.095	0.009	0.037	0.112	0.024
Q3 (75%)	0.032	0.160	0.016	0.055	0.141	0.033

	Spatial Autocorrelation					
	NF + VAE	VAE	NF + Copula	Copula	Global shuffle	Local shuffle
min-max	[0.000 , 0.258]	[0.001 , 0.255]	[0.027 , 0.196]	[0.014 , 0.193]	[0.000 , 0.688]	[0.001 , 0.159]
Q1(25%)	0.013	0.017	0.062	0.061	0.026	0.023
Median (50%)	0.028	0.038	0.080	0.081	0.058	0.043
Q3 (75%)	0.062	0.070	0.105	0.102	0.088	0.070

	Local Feature					
	NF + VAE	VAE	NF + Copula	Copula	Global shuffle	Local shuffle
min-max	[0.255 , 0.516]	[0.264 , 0.633]	[0.230 , 0.604]	[0.230 , 0.531]	[0.311 , 0.621]	[0.224 , 0.521]
Q1(25%)	0.366	0.457	0.379	0.365	0.470	0.372
Median (50%)	0.391	0.492	0.409	0.403	0.501	0.410
Q3 (75%)	0.413	0.531	0.443	0.450	0.544	0.446

	Utility					
	NF + VAE	VAE	NF + Copula	Copula	Global shuffle	Local shuffle
min-max	[0.054 , 0.624]	[0.086 , 3.712]	[0.101 , 10.428]	[0.046 , 0.451]	[0.040 , 0.922]	[0.001 , 0.074]
Q1(25%)	0.156	0.317	0.189	0.117	0.137	0.006
Median (50%)	0.196	0.561	0.283	0.163	0.206	0.009
Q3 (75%)	0.271	0.812	0.403	0.235	0.322	0.015

	Privacy					
	NF + VAE	VAE	NF + Copula	Copula	Global shuffle	Local shuffle
min-max	[-0.215 , 0.277]	[-0.233 , 0.187]	[-0.207 , 0.166]	[-0.277 , 0.153]	[-0.176 , 0.337]	[0.064 , 0.440]
Q1(25%)	-0.031	-0.030	-0.037	-0.031	-0.006	0.150
Median (50%)	0.003	0.011	0.002	0.002	0.035	0.187
Q3 (75%)	0.051	0.039	0.035	0.030	0.072	0.225

Table 3: Statistics of the distribution of evaluation metrics in 106 datasets from `data_isp` (extension of Fig. 4 of the main text).

3.3 Tuning VAE loss

The weights of the loss functions are $\alpha = (\alpha_{GEO}, \alpha_R, \alpha_{KL}) = (3, 1, 0.01)$. We chose α_{GEO}, α_R, and α_{KL} through a sensitivity analyses that involved setting $\alpha_R = 1$ and varying α_{GEO} and α_{KL} in the set {1, 3, 9, 27} and $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$, respectively. Then, we evaluated: (i) the energy distance between real and synthetic data, and (ii) the fraction of non-existing synthetic homes with respect to real data across the postcodes in each province. The smaller the energy distance, the higher the similarity between the locations of synthetic and real homes. Studying the fraction of non-existing homes allowed us to further measure the synthetic-real discrepancy considering both the homes' features and locations. To this aim, we transformed continuous home features into categorical ones as follows: a) coordinates were aggregated at postcode and b) homes' surface and price were divided in 5 classes based on the quintiles computed on the real homes. The fraction of non-existing homes is the fraction of synthetic homes with a combination of boolean and categorical (including also the ZIP code and binned surface and price) features that does not exist in the real dataset. The higher this fraction, the worse are the performance of the model in generating realistic homes. We got α_{GEO}, α_R, and α_{KL} studying three Italian provinces –Rome (big province), Bologna (medium province), and Vibo Valentia (small province)– and then applied the resulting hyperparameters to generate the synthetic data for all provinces. Supplementary Figure 2 shows the energy distance and the fraction of non-existing homes as the α_{GEO} and α_{KL} vary in their ranges. The selected hyperparameters ($\alpha_{GEO} = 3$ and $\alpha_{KL} = 10^{-2}$) are among the combinations for which we got the minimum energy distance (i.e., best home locations) and proportion of non-existing (i.e., best joint agreement among homes' locations and features).

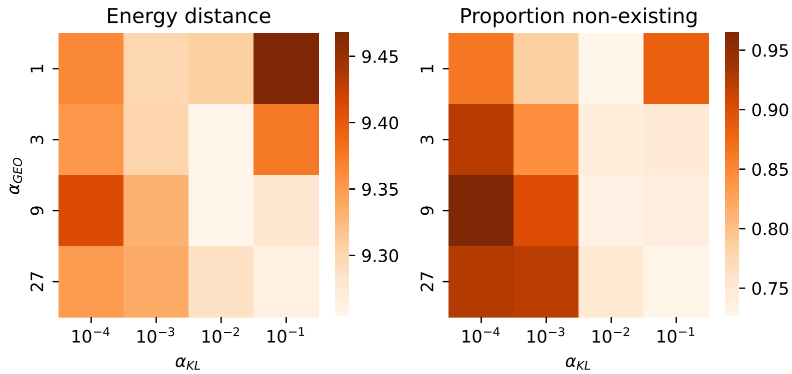

Figure 2: Performance of NF + VAE with all combinations of $\alpha_{GEO} \in \{1, 3, 9, 27\}$, $\alpha_R = 1$, and $\alpha_{KL} \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$. Lighter tiles represent higher similarity measures.

3.4 Robustness Check - Fidelity - Spatial Autocorrelation

As a robustness check we also considered an alternative definition of the distance matrix used to estimate the Moran's I. This alternative configuration considers that $w_{i,j} = \mathbb{1}_{x_j \in \mathcal{N}_i^{20}}$, where $\mathcal{N}_i^{20}$ is the set of 20 closest homes to home i.

Supplementary Figure 3 shows that this alternative distance definition does not result in major changes in the d_S^F for all baselines.

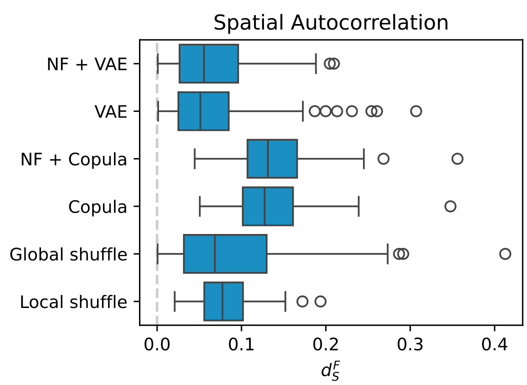

Figure 3: Robustness check of Similarity - Spatial Autocorrelation. We repeat the same analysis presented in Fig. 4b with a different weighting function within the computation of Moran's I.

3.5 Robustness Check Fidelity - Correlation features

The three fidelity evaluation metrics used in the validation pipeline assess whether NF+VAE correctly captures spatial correlations between homes' locations and their attributes. As a robustness check, we additionally introduce a new similarity metric to evaluate whether NF+VAE also preserves non-spatial correlations between home features.

To this end, we removed the geographic coordinates from the real and synthetic data and estimated the Pearson correlation between all features in each dataset. Supplementary Figure 4 shows an example of the resulting correlation matrices for the province of Bari (Apuglia), both the real (m^R) and synthetic data (m^S). The visual similarity between the two matrices already suggests that NF+VAE faithfully reproduces the correlations between non-spatial features. To confirm this observation, we estimate d_C^F, defined as the L^2 distance between m^R and m^S after setting all diagonal elements to zero. Supplementary Figure 5 shows the results of d_C^F for all baselines. The results indicate that the use of VAE is crucial for accurately capturing the correlations among home attributes.

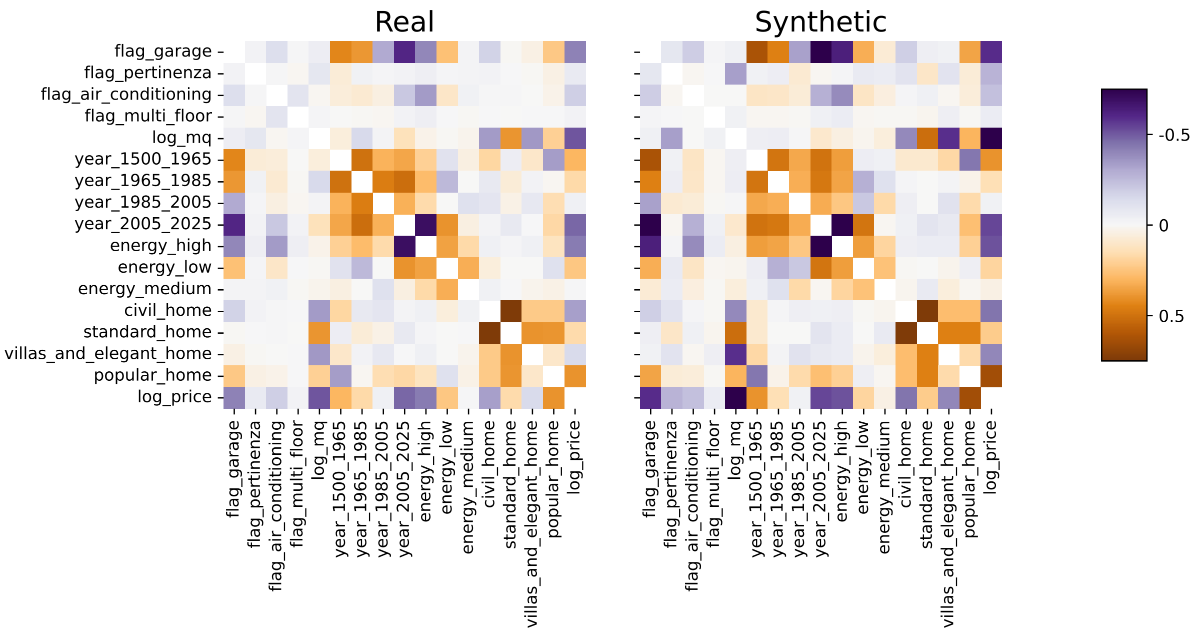

Figure 4: Correlation matrices of real and synthetic populations of Bari province. For the sake of conciseness, we remove the variables related to floors, missing data, and geographic coordinates. We can observe that the synthetic data maintains also the linear correlations among the non-geographic variables.

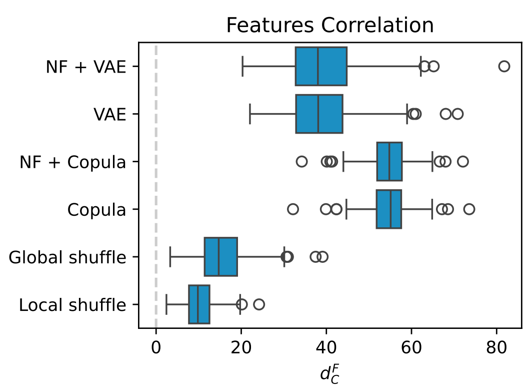

Figure 5: The boxplot shows the distribution of the L^2 distances between real and synthetic homes generated using different approaches. d_C^F describes the L^2 distance between m^R and m^s after setting all diagonal elements to 0 for the homes in the 106 Italian provinces in `data_isp`.

3.6 Robustness Check - Utility

Supplementary **??** evaluates the robustness of the utility results by adopting an alternative notion of utility. While Fig. 4d focus on the regression task, Supplementary **??** assesses utility in a classification setting. Specifically, we train a classifier on synthetic data and evaluate it on real data for a set of binary variables. Following the same protocol used in the privacy evaluation, we randomly split the dataset into a training set (95%) and a held-out test set. We use a Random Forest classifier to predict each binary target. Let m_D denote the classifier trained on real data, $m_{\tilde{D}}$ the classifier trained on synthetic data, and $AUC(m)$ the Area Under the ROC Curve of model m evaluated on the real test set. We measure utility as $d_u = \left|AUC(m_{\tilde{D}}) - AUC(m_D)\right|$. Thus, $d_u \approx 0$ indicates that the synthetic dataset is as informative as the real one dataset for predicting the selected binary variables. We assess the utility of all generators on three representative binary features: 'Flag Garage', 'Flag A/C', and 'Flag Annex'.

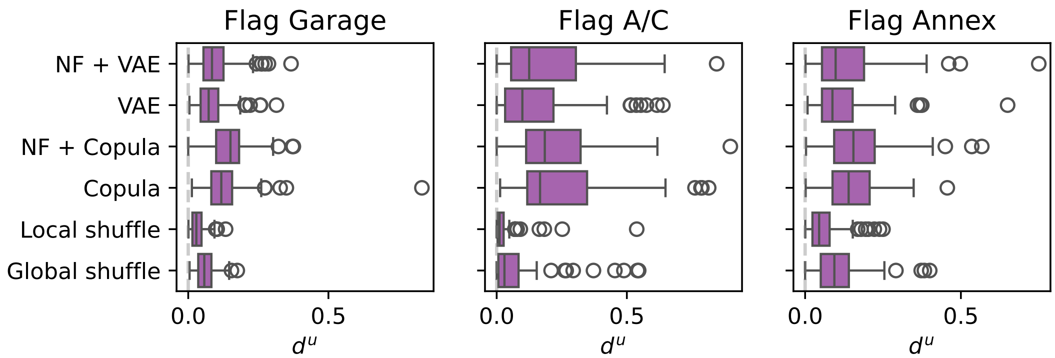

Figure 6: Robustness analysis of Utility. We compare the utility metric across diverse measures of d^U that are based on different task. In particular, d^U measures the capability of classifying the observations in the test set along the features of 'Flag Garage', 'Flag A/C', and 'Flag Annex'.

3.7 Robustness Check - Privacy

Supplementary Figure 7 presents the same results as in Fig. 4e considering different classifiers. As one can see, all synthetic generators do not present privacy leakage since the median difference between 0.5 and the AUC-ROC (ρ^P) is close to 0. Instead, the local shuffle approach has median values of ρ^P higher than 0 for several classifier other than the Logistic Regression. For example, Gaussian Naive Bayes, KNeighbors, Random Forest, and Multi-Layer Perceptron (MLP).

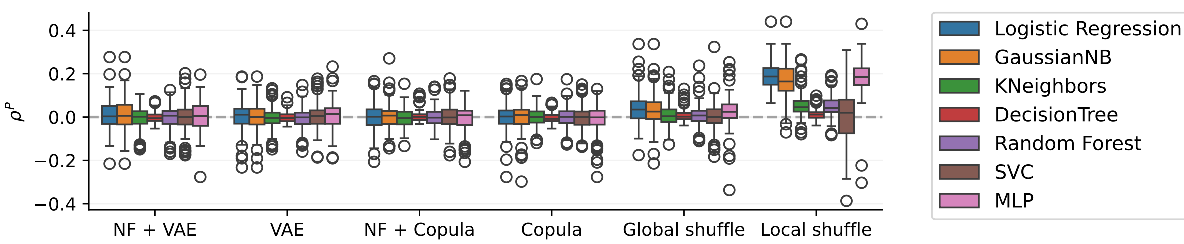

Figure 7: Robustness analysis of Privacy. We compare the measures of privacy as described in Fig. 4e with different classifiers. This figure extends ??c, which presents only Logistic Regression.

3.8 Zero-cell Problem

We analyze the well-known zero-cell problem in synthetic data. To check how the methods tackle this problem, we match the geographic coordinates with the corresponding subregion, and we group the other continuous variables in 5 bins corresponding to the quintiles of the real dataset. In this way, we transformed continuous variables into categorical variables, allowing to compare the combinations of features in the data. In doing so, we note that median proportion of samples generated by NF+VAE that are not present in the original (binned) data is 0.65. While for the other methods this median is 0.83 for VAE, 0.93 for copula, 0.93 for NF+copula, 0.86 for Global shuffle, and 0 for Local shuffle.

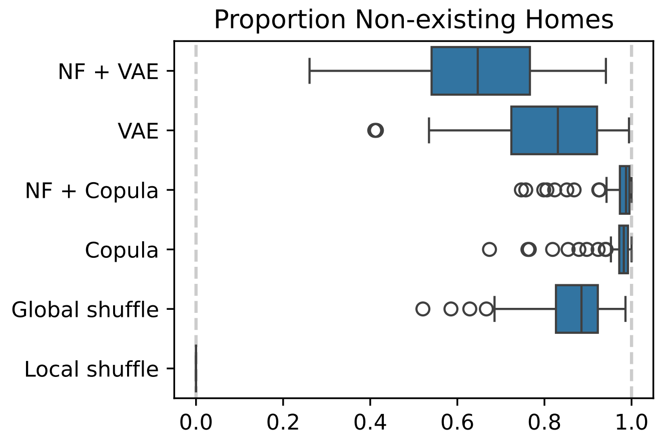

Figure 8: Proportion of homes described by combinations of features that are not present in the original dataset.

4 AIRBNB RESULTS: BOXPLOTS AND STATISTICS TABLE

We evaluated the performance of NF+VAE in generating synthetic populations of `data_airbnb` through the same pipeline described in Sec. 3.3 of the main text. The results across the 15 cities studied are shown in Supplementary Figure 9, while Supplementary Table 4 shows some more detailed statistics.

Fidelity. As in the `data_isp` evaluation, the fidelity of geographic coordinates shows that methods using NF achieve performances comparable to Local shuffle, while VAE alone does not capture correctly geographic distributions. However, in this case, VAE-based methods and copula-based ones achieve similar performances when evaluating spatial autocorrelation, where NF+VAE slightly presents the best spatial autocorrelations. Both approaches also perform similarly for the local feature fidelity evaluation, achieving performances similar to the null models except for the VAE ablation case. We hypothesize that the increased performance of Copula-based methods in `data_airbnb` derives form the lower spatial heterogeneity of the city setting, when compared to the province scale.

Utility. We evaluated the utility with a regression setting where we predict the price per night based on the accommodation features. In this case, we considered spatial fixed effects defined at the neighborhood scale, retrieved from a spatial matching algorithm starting from the accommodation coordinates. This approach mirrors the hedonic regressions used in the main text to estimate the sell price of a home based on its features. Supplementary Figure 9d presents our results. NF+VAE, NF+Copula, and Copula approaches have similar performances, slightly worse than the global shuffle. Instead, VAE shows the worst performances.

Privacy. We checked whether the methods present privacy leakages on `data_airbnb`. Supplementary Figure 9e shows that all the generative approaches (except the local shuffle) preserve the privacy of the input data. In fact, the median difference between 0.5 and the AUC-ROC is close to 0. The Local shuffle model is still significantly larger than 0, which indicates that shuffling inside the neighborhood is still not sufficient to preserve the privacy in the Airbnb dataset. This result is in line with what shown in the main text on the `data_isp`, highlighting that Local shuffle is not a privacy-preserving approach to generate geo-located synthetic samples.

Summary. We emphasize that `data_airbnb` exhibits different properties from `data_isp`, with fewer features and less clear patterns. In particular, several attributes, such as bedrooms, self check-in, and pets allowance, show low spatial autocorrelation. Similarly, hedonic regressions achieve lower accuracy. Despite these challenges, the overall evaluation of NF+VAE on `data_airbnb` yields results that are consistent with those observed in `data_isp`. Specifically, NF+VAE attains fidelity measures comparable to Local shuffle, utility is similar to copula-based methods, while privacy is preserved by all non-shuffle methods.

	Geographic Coordinates					
	NF + VAE	VAE	NF + Copula	Copula	Global shuffle	Local shuffle
min-max	[0.001 , 0.209]	[0.005 , 0.690]	[0.001 , 0.414]	[0.002 , 0.572]	[0.009 , 0.905]	[0.001 , 0.092]
Q1(25%)	0.003	0.010	0.002	0.004	0.016	0.004
Median (50%)	0.004	0.049	0.004	0.011	0.064	0.006
Q3 (75%)	0.015	0.106	0.006	0.023	0.118	0.013

	Spatial Autocorrelation					
	NF + VAE	VAE	NF + Copula	Copula	Global shuffle	Local shuffle
min-max	[0.010 , 0.292]	[0.012 , 0.156]	[0.018 , 0.341]	[0.017 , 0.347]	[0.021 , 0.346]	[0.012 , 0.216]
Q1(25%)	0.029	0.042	0.047	0.043	0.049	0.015
Median (50%)	0.055	0.063	0.065	0.064	0.063	0.033
Q3 (75%)	0.105	0.111	0.121	0.117	0.122	0.063

			Local Feature			
	NF + VAE	VAE	NF + Copula	Copula	Global shuffle	Local shuffle
min-max	[0.513 , 2.631]	[0.575 , 3.855]	[0.371 , 2.588]	[0.378 , 2.537]	[0.365 , 2.744]	[0.302 , 2.520]
Q1(25%)	0.883	1.433	0.860	0.853	0.765	0.636
Median (50%)	1.247	1.864	1.181	1.115	1.018	0.962
Q3 (75%)	1.615	2.563	1.574	1.419	1.571	1.393

			Utility			
	NF + VAE	VAE	NF + Copula	Copula	Global shuffle	Local shuffle
min-max	[0.082 , 0.772]	[0.188 , 4.433]	[0.095 , 0.783]	[0.100 , 0.585]	[0.031 , 0.277]	[0.001 , 0.013]
Q1(25%)	0.151	0.337	0.162	0.122	0.052	0.003
Median (50%)	0.215	0.618	0.183	0.176	0.069	0.004
Q3 (75%)	0.312	1.084	0.307	0.247	0.101	0.005

			Privacy			
	NF + VAE	VAE	NF + Copula	Copula	Global shuffle	Local shuffle
min-max	[-0.062 , 0.054]	[-0.070 , 0.138]	[-0.138 , 0.038]	[-0.122 , 0.051]	[-0.030 , 0.132]	[0.069 , 0.304]
Q1(25%)	-0.026	-0.040	-0.039	-0.018	0.004	0.144
Median (50%)	-0.003	0.012	-0.007	-0.012	0.027	0.205
Q3 (75%)	0.010	0.023	0.011	0.014	0.061	0.252

Table 4: Statistics of the distribution of evaluation metrics in 15 datasets from `data_airbnb` (extension from Supplementary fig. 9).

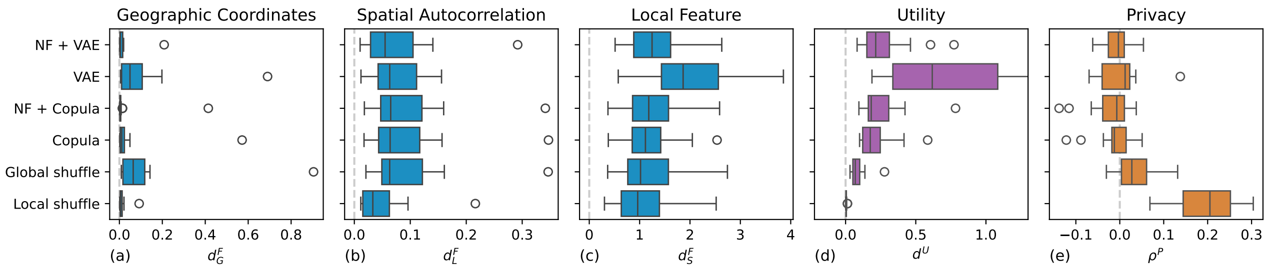

Figure 9: Distributions of evaluation metrics in 15 datasets from Airbnb. (a) *Fidelity - Geographic coordinates*, i.e., sliced-Wasserstein distance geographic coordinates, (b) *Similarity - Spatial autocorrelation*, Euclidean distance between Moran's Index in the first PCs of real and synthetic homes, (c) *Fidelity - Local features*, distance between average home per spatial grid cell, (d) *Utility*, average absolute difference 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. Detailed statistics are available in Supplementary Table 4.

REFERENCES

[1] Abien Fred Agarap. 2018. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375* (2018).

[2] Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. 2019. Neural spline flows. *Advances in neural information processing systems* 32 (2019).

[3] Gábor J Székely and Maria L Rizzo. 2013. Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference* 143, 8 (2013), 1249–1272.