# Population Synthesis using Incomplete Information

Tanay Rastogi, Daniel Jonsson and Anders Karlström

**Abstract**

This paper presents a population synthesis model that utilizes the Wasserstein Generative-Adversarial Network (WGAN) for training on incomplete microsamples. By using a mask matrix to represent missing values, the study proposes a WGAN training algorithm that lets the model learn from a training dataset that has some missing information. The proposed method aims to address the challenge of missing information in microsamples on one or more attributes due to privacy concerns or data collection constraints. The paper contrasts WGAN models trained on incomplete microsamples with those trained on complete microsamples, creating a synthetic population. We conducted a series of evaluations of the proposed method using a Swedish national travel survey. We validate the efficacy of the proposed method by generating synthetic populations from all the models and comparing them to the actual population dataset. The results from the experiments showed that the proposed methodology successfully generates synthetic data that closely resembles a model trained with complete data as well as the actual population. The paper contributes to the field by providing a robust solution for population synthesis with incomplete data, opening avenues for future research, and highlighting the potential of deep generative models in advancing population synthesis capabilities.

**Keywords**: population synthesis, microsample, WGAN

# 1 Introduction

Transportation simulation models, using agent-based models (ABMs), are widely used for various tasks like predicting travel demand, evaluating policy impacts or analyzing travel behavior. These models usually need complete information about individuals' social and household characteristics from an area covering cities, towns, or even countries (Bastarianto et al. (2023)). Ideally, such information could be collected from census data at an individual or household level, and then we could draw a certain number of samples as a synthetic population. Statistical authorities in many countries have also made available a microsamples of individual-level data from the whole population and can be used in place of census data. Travel surveys that capture complete demographic and socioeconomic attributes at a comparable sampling rate can also act as a good replacement. In addition to these microsamples, aggregated marginal information on a regional or zonal level is usually available from the Bureau of Statistics. However, acquiring granular, individual-level data is challenging. Issues such as privacy concerns, as well as the technical and financial constraints of data gathering, often impede accessibility to detailed data.

In order to tackle difficulties in data collection, population synthesis algorithms were created to create synthetic populations. They provide ABM transport models with a reliable alternative to actual populations. Population synthesis techniques generate a comprehensive list of a simulated population, each accompanied by corresponding attribute data. The objective of population synthesis is to optimally utilize the existing microsamples, along with the additional aggregated or marginal information on each attribute of interest, in order to generate agents that closely aligns with the underlying population structure (Sun and Erath (2015)). These simulated agents can thereafter be employed to evaluate the impact of factors such as governmental policies on the region or conduct studies that would be prohibitively costly, ethically questionable, or just unfeasible using actual population data.

As highlighted in the research by Rich (2018); Borysov et al. (2019) the population synthesis methods, typically consist of three steps - 1). Starting solution where a synthetic pool of individuals are generated to represent a diverse combinations of attributes. This is usually done based on the available microsamples. 2). Fitting stage where weighting factors for the synthetic pools are estimated to construct

the representative synthetic population for future targets. 3). Allocation stage where synthetic agents are generated and are assigned to ABM transport models. In this paper, we focus exclusively on the first stage—the generating of appropriate representative samples for a given population without considering how such samples can be aligned with future targets. This step is crucial because any underlying error in this step will propagate in the ABM models and hence affect the forecast results.

In the population synthesis literature, there has been a trend toward simulation based probabilistic models for the first stage i.e. generation of starting solution. Research from Farooq et al. (2013); Sun and Erath (2015); Borysov et al. (2019); Garrido et al. (2020); Kim and Bansal (2023) utilized simulation-based generative models to create synthetic populations for the first stage of population synthesis. One crucial similarity in each of these studies is that the microsample used in the experiments are from single data-source and are complete i.e. there is no missing information for any attributes in the microsample. The quality of generated synthetic populations depends, of course, on the quality and detail of the microsample. While the data quality has been improving, it has not kept pace with the growing interest in microsimulations at the scale of individuals tagged with many associated attributes. Available microsamples are often thin and incomplete. Survey microsample cannot comprehensively cover all the variations of different attributes found in the actual population. Frequently, these microsamples exhibit incomplete data on one or more attributes due to errors in data collection or the respondents intentionally withholding information or not collecting it to ensure privacy. Additionally, to improve the attribute richness of the microsample, they can be combined with supplemental data from other sources. For example, travel survey in a certain region from multiple distinct organizations can be combined to add more data and attributes in the microsample. However, there may be situations in which one or more attributes are absent in either of the surveys. Hence, there is a need for an imputation algorithm can be used to estimate missing values based on data that was observed/measured in microsample. This will ensure that the synthetic population generated by population synthesis models is complete.

In this context we propose a new approach that uses the incomplete microsamples in order to draw the synthetic populations from it. In the context of this study, the *incomplete microsamples* is one which has missing info on one or more attributes of the sample. We use a population synthesis model based on the Wasserstein

Generative Adversarial network (WGAN) suggested by Kim and Bansal (2023). The contribution of this study lies in the proposal of a novel technique in the WGAN training algorithm that enables the model to effectively learn using incomplete training data. The aim of this study is to propose a methodology that can effectively addresses the challenges associated with missing data, while ensuring a degree of accuracy that is at least equivalent to the methods described in the existing literature.

The remainder of the article is structured in the following manner: ***Sec.*** 2 provides a overview of the existing literature on the subject matter and highlights the contribution of this paper within the field. The proposed training methodology for the WGAN model is presented in ***Sec.*** 3. We first briefly present the original WGAN training method and then describe the proposed changes to it. To evaluate and access the performance of the proposed training method, we utilize a microsample from a Swedish national travel survey. The experimental setup, metric evaluation, results, and discussions on these are provided in ***Sec.*** 4. Finally, in ***Sec.*** 5, the article concludes by summarizing the analysis and suggesting potential avenues for future research.

# 2 Literature review

As proposed by Sun et al. (2018); Borysov et al. (2019) the population synthesis methods can be divided into two primary categories: deterministic and simulation methods. Deterministic typically consider the microsample represents the true correlation structure among the attributes. These methodologies tries to expand microsamples by fitting them to a aggregated marginal distribution. Introduced by Deming and Stephan (1940), Iterative Proportional Fitting (IPF) is one of the important key techniques used for population synthesis that combines the microsample data with aggregated marginal. In the review from E. Ramadan and P. Sisiopiku (2020), the authors show the extensive study on IPF methods and it has been continuously developing by adding various extensions to deal with emerging issues. Several studies like Beckman et al. (1996); Zhu and Ferreira

(2014); Rich (2018) have used variation of IPF algorithm to generate synthetic population. Typically, for IPF methods, all the attributes in the microsample have to be discrete and with limited categories. Fitting for large number of individual attributes quickly becomes computationally and memory-wise expensive. Because of the dependency on the original microsample, the IPF methods cannot approximate high-dimensional data. A common issue that comes with high-dimensional data is the problem of zero-cells, which in addition to rendering of sparse sample may also lead to convergence and division by zero problems described by Choupani and Mamdoohi (2016).

In recent years, probabilistic-based simulation methods have gained momentum, offering more robust solutions to the challenges faced by deterministic methods. Research by Borysov et al. (2019) emphasizes that simulation-based methods provide a systematic way of interpolating data. Even if specific agents do not exist in the original data, it may still be possible to sample these specific agents by combining agents in the original data. These methods excel in addressing high-dimensional problems, offering better scaling properties and fulfilling the need for more detailed populations. One of the first model that uses a probabilistic-simulation framework for population synthesis was introduced by Farooq et al. (2013) where they employed a Markov Chain Monte Carlo (MCMC) algorithm based on Gibbs sampling to draw from a partial joint distribution of data, simulating draws from the original distribution. Subsequently, Sun and Erath (2015) utilized a Bayesian network to model the joint distribution function for multiple individual attributes. While these methods generally outperform conventional deterministic models, they may encounter zero-cell problems, especially when tested on larger datasets.

The emergence of Deep Generative Models (DGM) in the machine learning community has introduced new possibilities for population synthesis. Borysov et al. (2019) introduced a Variational Autoencoder (VAE) model to synthesize a population based on Danish Trip Diary. They compared the model against conventional algorithms like IPF and other generative models like Gibbs sampling and Bayesian Network. In their experiments, they found that the VAE model was able to address the problem of sampling zeros by generating agents that are virtually different from those in the original data but have similar statistical properties.

Notably, Generative Adversarial Networks (GANs), introduced by Goodfellow et al. (2014), have been applied to population synthesis. Initially designed for image data, GAN models demonstrate the potential to learn high-dimensional features using neural networks and produce high-quality synthetic data. This has led to the utilization of GAN models in synthesizing not only images but also music, text, and structured tabular data. In recent studies, WGANs based models have been used extensively for generating tabular data. Walia et al. (2020) present the WGAN model using gradient penalty to produce tabular data that is indistinguishable from real data. Baowaly et al. (2019) proposed an two version of WGAN model for generating binary synthetic long-form electronic health related data. The authors claim that the improved models can generate more realistic-looking synthetic data that can be used to train other machine learning models better than previous proposed models. Xu et al. (2019) presents a new model called CTGAN which uses a conditional generator based in WGAN and present itself as the new state-of-art tabular generation model.

In the domain of population synthesis, there are some research that have used WGAN. Garrido et al. (2020) extended the research by Borysov et al. (2019) and applied the WGAN model to Dutch trip diary data, comparing the models against the VAE as well as Bayesian models. In their study, the WGAN model outperformed the VAE model while producing a significantly lower number of structural zeros in the data. Later, Kim and Bansal (2023) introduced two new loss functions to the WGAN models with the aim of ensuring that the trained generator produces fewer structural zeros and more sampling zero data. They tested the models against the naïve WGAN model and VAE, showcasing that the new loss functions indeed help WGAN models produce significantly fewer structural zeros while maintaining a good level of sampling zero data. As highlighted in the previous section, none of the mentioned research is based on the microsamples that are incomplete, which is often the case in real world. Hence, there is a need for a population synthesis method that can impute the missing information in the microsamples.

There are many different methods to impute missing information in tabular data. In the research Emmanuel et al. (2021), authors highlight many classical imputation methods that are based on machine learning techniques like KNNImpute, MICE, MissForest and SMOTE. With the popularity of generative models, many GAN based models have been proposed for data imputation. Yoon et al. (2018) proposed a model called Generative Adversarial Imputation Nets (GAIN) where a they used

a hint vector to train a generator-discriminator model that can imputes the missing components conditioned on what is actually observed, and outputs a completed vector. Neves et al. (2022) improves upon the GAIN model by using WGAN as base and propose three different models called SGAIN, WSGAIN-CP, and WSGAIN-GP.

In this paper, we use the WGAN model from Kim and Bansal (2023) and include the ideas presented by Neves et al. (2022) to train the generator model with a mask vector that indicate the location of missing information in the training data. In this regard, the proposed method is different from Neves et al. (2022) as we do not aim to create a WGAN imputation model, rather just uses the techniques to handle training generator model with missing data.

# 3 Methodology

In mathematical terms, we consider a population of agents $n = 1, 2, \ldots, N$ with each agent defined by vectors of $K$ attributes i.e. variables with agent's individual characteristics and household characteristics. This population of agents is a representation of the actual population of $N$ individuals. The population synthesis problem is concerned with estimation of joint distribution of synthetic population $\hat{P}(X)$ that approximate the true joint distributions of attributes across a real population $P(X)$. The data for creating a model for $P(X)$ come in form of dis-aggregated data collected from survey data in form of microsample $X$, typically with a sample size of $M < N$. To represent the incomplete microsamples, $X_r$, we picked $q$ attributes and replaced $r\%$ of rows (selected at random), in $X$ with NaN values. For this study, the microsamples, $X$ and $X_r$ represents the training data for the propsoed WGAN model.

A deep generative model represents a category of machine learning models designed to generate novel data samples resembling a given dataset. The core objective is to comprehend and model the inherent patterns, structures, and statistical features embedded in the training data. The fact that neural networks can approximate any function makes them a natural choice for the population synthesis problem, that is, the approximation of the function $P(X)$. This paper primarily focuses on

the WGAN with gradient penalty (WGAN-GP), a generative model based on the framework introduced by Goodfellow et al. (2014) and subsequently enhanced by Gulrajani et al. (2017). The WGAN-GP model generates a synthetic population by transforming a random generated numbers from $K$-dimensional standard normal latent variable, $Z$. The WGAN-GP aims to generate output such that $G(Z) \rightarrow \hat{P}(X)$ such that they are consistent with actual population, $P(X)$.

For training, the generator network, $G(Z)$, is initiated with a draw from a latent variable $Z$. The draw is then transformed in such a way that the output has the same dimensions and shape as the real data. The second network, the discriminator network $D(X)$, receives an observation. This can either be from real data or from the generator $G(Z)$. The objective of the discriminator is to tell whether the information it receives comes from the real data or not. The training process continues until the $D$ network is no longer able to distinguish between generated and synthetic data. The pseudo-code of training of the WGAN-GP is presented in ***Algorithm*** 1.

The learning process in the model is based on $G(Z)$ and $D(X)$ playing this adversarial game based on the loss function given by,

$$\mathscr{L} = \mathscr{L}_D + \mathscr{L}_G + \lambda_{gp} * \mathscr{L}_{GP} + \lambda_{bd} * R_{BD} + \lambda_{ad} * R_{AD} \tag{1}$$

where $\mathscr{L}_D$ is the discriminator loss, $\mathscr{L}_G$ is generator loss and $\mathscr{L}_{GP}$ is the loss from regularization by gradient penalty on the discriminator. $R_{BD}$ and $R_{AD}$ are the regularization terms to control the generation out-of-training samples. The term $\lambda_{gp}$, $\lambda_{bd}$ and $\lambda_{ad}$ are the model hyper-parameter which are manually selected to control the effect gradient penalty and two regularization terms, respectively.

For training data of batch-size $M$, the loss functions can be defined as,

$$\mathscr{L}_D = \frac{1}{M} \sum_{i=1}^{M} -D(X_i) + D(G(Z_i)) \tag{2}$$

$$\mathscr{L}_G = \frac{1}{M} \sum_{i=1}^{M} -D(G(Z_i)) \tag{3}$$

$$\mathscr{L}_{GP} = \frac{1}{M}\sum_{i=1}^{M}(\left\|\nabla_{\tilde{X}_i} D(\tilde{X}_i)\right\|_2 - 1)^2$$
$$\tilde{X}_i = \alpha\hat{X}_i + (1-\alpha)\hat{X}_i \tag{4}$$

where, $\|.\|_2$ is the euclidean norm, $X_i$ and $\hat{X}_i$ are the real and generated data, respectively. $\alpha$ is a random number from a uniform distribution of $\alpha \in U[0,1]$.

The two regularization in the equation are boundary distance regularization $R_{BD}$ and average distance regularization $R_{AD}$ are introduced by Kim and Bansal (2023) in order to promote sampling zero samples and restrict structural zero generation samples, respectively. These functions are expressed as,

$$R_{BD} = \frac{1}{M}\sum_{i=1}^{m} \min_{j\in\{1:N\}, i\in\{1:M\}} (DIST(\widehat{X}_i, X_j)) \tag{5}$$

$$R_{AD} = -\frac{1}{NM}\sum_{i=1}^{M}\sum_{j=1}^{N} (DIST(\widehat{X}_i, X_j)) \tag{6}$$

$$DIST(X_i, X_j) = \sqrt{(X_i - X_j)^2} \tag{7}$$

where, $\widehat{X}_i$ are the generated data of size $M$, $X_j$ is the entire training data of size $N$. The $R_{BD}$ calculates the nearest distance from each generated data in the batch to entire $N$ training data and average them for $M$ generated batch, where as $R_{AD}$ computes the average for average distance to the entire training sample distribution of $M$ generated data.

**Algorithmus 1** WGAN-GP with missing data

**Require:** Generator ($G$), Discriminator ($D$), Latent Variable ($Z$), Training Data ($X$), Epochs ($E$), Batch Size ($M$), $D$ updates per epoch ($n_d$), Gradient Penalty ($\lambda_{gp}$), Boundary distance ($\lambda_{bd}$), Average distance ($\lambda_{ad}$), mask ($Y$).

**Ensure:** Trained Generator $G$, Trained Discriminator $D$

Initialize $G$ and $D$
**for** Epoch $e = 1$ to $E$ **do**
  **for** Batch $M$ in Data Loader **do**
    **for** Update $d = 1$ to $n_d$ **do**
      $Z \leftarrow \mathcal{N}(0, 1)$ of size $M$
      $\acute{X} \leftarrow G(Z)$
      **Multiply with mask,** $\widehat{X} \leftarrow \acute{X} * Y$
      $D_{real} \leftarrow D(X_m)$
      $D_{fake} \leftarrow D(\widehat{X})$
      Wasserstein loss for the $D$, $\mathscr{L}_D \leftarrow -D_{real} + D_{fake}$
      Gradient Penalty $\mathscr{L}_{GP} \leftarrow \lambda_{gp} * \mathbb{E}[(\left\|\nabla_{\tilde{X}_i} D(\tilde{X}_i)\right\|_2 - 1^2)]$
      Update $D$ parameters using $\mathscr{L} \leftarrow \mathscr{L}_D + \mathscr{L}_{GP}$
    **end for**
    Generator Loss $\mathscr{L}_G \leftarrow -D_{fake}$
    Regularization $R_{BD} \leftarrow \min(DIST(\widehat{X}, X^S))$
    Regularization $R_{AD} \leftarrow DIST(\widehat{X}, X^S)$
    Update $G$ parameters using $\mathscr{L} \leftarrow \mathscr{L}_G + \lambda_{bd} R_{BD} + \lambda_{ad} R_{AD}$
  **end for**
**end for**

In order to handle missing data, we used the masking approach presented by Neves et al. (2022). During the training of the WGAN-GP, we introduce a new matrix called mask matrix as input. The mask, $Y$ as being the mask of training data, $X$, where a missing value in $X$ is represented by a zero and any non-missing value is represented by a one. Thus the mask is a binary matrix of same size as $X$. The biggest difference from the original WGAN training algorithm is in line 7 of the ***Algorithm*** 1, where the matrix generated by the generator, $G(Z)$ is multiplied by mask $Y$, before giving as input to discriminator $D$ to get score for fake samples. ***Figure*** 1 shows an example for the sample and its corresponding mask used during the training process.


**Figure 1: Illustration showing an example of sample data with corresponding training data and mask. The missing values are represented as NA in the sample, which are replaced by 0 in training data and mask.**

# 4 Case Study

## 4.1 Travel Survey

The data for this study were obtained from the national travel behavior survey conducted by the Swedish Institute for Transport and Communications Analysis (SIKA), known as Riks-RVU 2005–2006 (Abramowski and Holmström (2007)). This survey was carried out over a one-year period, from October 2005 to September 2006. It includes information on 41,000 individuals aged 6 to 84, selected through a stratified sampling process from the Swedish total population register. For the purposes of this analysis, we focus exclusively on the UPBD dataset within Riks-RVU, which contains detailed individual and household attributes. Each row in the survey represents a weighted sample that has been post-stratified based on strata defined by year, region, age, and sex. The regional classification is primarily at the county level, except for Stockholm County, where it is defined at the municipal level. Age groups are categorized as 6–14, 15–24, 25–44, 45–64, and 65–84 years.

The data were processed through the following steps to prepare them for model training: 1). All attributes were converted into categorical variables. 2). Missing values were imputed using domain knowledge to assign appropriate categories. 3). A full population dataset was generated by replicating each survey row according to its corresponding weight, thereby creating a synthetic population that represents the entire Swedish population in 2005. This replicated dataset comprises information on 8,227,341 individuals.

To assess the accuracy of the generated full population in representing the actual Swedish population, we compared its marginal distributions with publicly available population statistics from Statistics Sweden (SCB) for the period January to December 2006. A heatmap visualization (see ***Figure*** 2) illustrates the count variations of females and males across different counties, comparing SCB data with the generated population. Visual inspection indicates that the synthetic population closely follows the trends observed in the SCB marginals for most regions and age groups. The most significant discrepancies appear in Götaland County and within the 6–14 age group. The elevated errors in Götaland County are likely due to its

relatively smaller population size compared to the entire country. Differences in the 6–14 age group arise from divergent age group definitions: SCB reports for ages 5–14, whereas the generated population uses 6–14. Additional discrepancies may result from the slight misalignment in temporal coverage, as SCB data cover January to December 2006, while the Riks-RVU survey spans October 2005 to September 2006. Despite these minor inconsistencies, the generated full population dataset closely approximates the SCB marginals, supporting the conclusion that it provides a reliable representation of the actual population. Therefore, it can be confidently used as ground truth for subsequent analyses.

### 4.1.1 Population Data

In order to compare if the proposed methodology was able to successfully train WGAN models with missing information, we need a ground truth population data that does not contain any missing information. In that sense, we dropped all rows containing any missing information from the generated full population dataset and retained only 17 distinct attributes pertaining to both individual and household characteristics. ***Table*** 1 displays the list of attributes utilized in the project from the travel survey dataset, with 13 associated with individuals and 4 associated with households. This dataset is designated as *h-population* serving as the reference dataset or "ground truth" for subsequent analysis.

### 4.1.2 Training Data

For training the proposed WGAN models, we create two different type of training dataset. First, we create a training set that is complete. We sample 10% of the *h-population* and deliberately remove certain unique category combinations from it. This refined dataset is labeled as *h-nomis*. By eliminating certain combinations in *h-nomis*, we address potential biases in structural zero and sampling zero rates arising from differences in the distributions of*h-population* and *h-nomis*.

Subsequently, we create multiple incomplete datasets. These datasets are derived from *h-nomis* but include some attributes with missing information. These datasets simulates scenarios where information is either absent in microsamples or when multiple microsamples are merged, resulting in missing information on one or more attributes. The creation of incomplete datasets involves two steps: 1). randomly

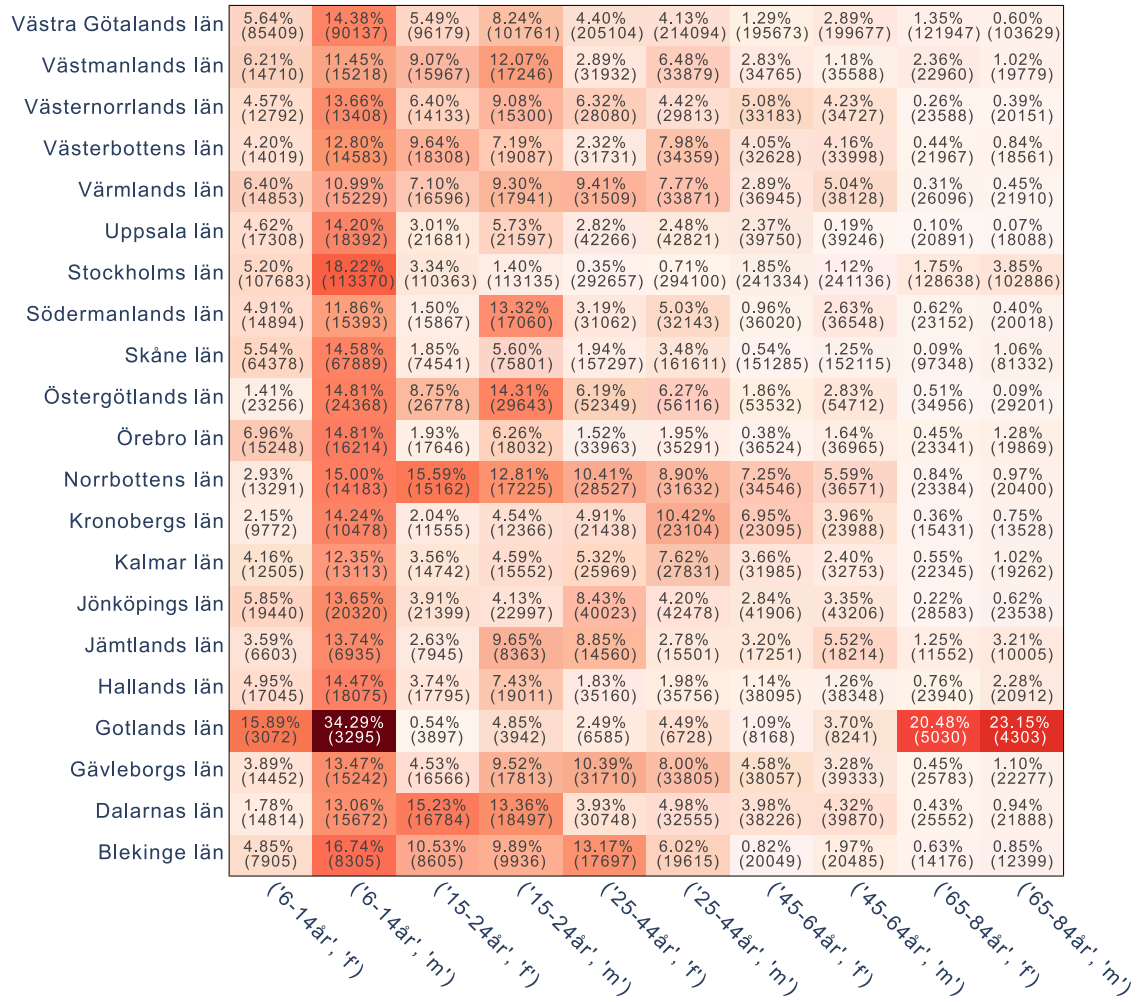



**Figure 2: Heatmap illustrating the disparity between the SCB and generated full population data for the marginal distribution of the population across all 21 counties in Sweden. Each cell displays the % error and corresponding count from SCB 2006 (enclosed in brackets).**

**Table 1: List of Attributes with Category proportions.**

| SN | Attributes(dim) | Definition | Category | Propotion % |
|---|---|---|---|---|
| 1 | AGE(5) | Age | 15-24 år | 8,59 |
| | | | 25-44 år | 34,83 |
| | | | 45-64 år | 31,78 |
| | | | 65-84 år | 11,39 |
| | | | 6-14 år | 13,4 |
| 2 | SEX(2) | Sex | f | 50,96 |
| | | | m | 49,04 |
| 3 | INKUP(5) | Individual income | 100 000 - < 200 000 | 24,01 |
| | | | 200 000 - < 250 000 | 20,28 |
| | | | Under 100 000 | 24,92 |
| | | | 250 000 - < 300 000 | 14,46 |
| | | | 300 000 eller mer | 16,33 |
| 4 | UP_FORV(3) | Main occupation | fulltime | 52,18 |
| | | | notworking | 38,63 |
| | | | parttime | 9,19 |
| 5 | LIVSKAT(8) | Life category | Ungdom, ej hemma | 4,75 |
| | | | Yngre | 13,6 |
| | | | Förälder | 29,85 |
| | | | Uppgift saknas | 3,18 |
| | | | Äldre barnlös | 20,75 |
| | | | Pensionär | 11,35 |
| | | | Barn | 13,1 |
| | | | Ungdom, hemma | 3,43 |
| 6 | SUN2KN(7) | Education level | gymnasial | 38,83 |
| | | | förgymnasial9_ | 8,46 |
| | | | eftergymnasial_2 | 5,79 |
| | | | eftergymnasial2_ | 24,55 |
| | | | förgymnasial_9 | 6,81 |
| | | | i skolan | 14,65 |
| | | | forskarutbildning | 0,91 |
| 7 | BOST_LAN(21) | Residence County | Stockholms län | 20,68 |
| | | | Skåne län | 12,63 |
| | | | Västernorrlands län | 2,77 |
| | | | Örebro län | 3 |
| | | | Östergötlands län | 4,71 |
| | | | Gävleborgs län | 3,04 |
| | | | Västra Götalands län | 17,15 |
| | | | Dalarnas län | 3,1 |
| | | | Norrbottens län | 2,98 |
| | | | Uppsala län | 3,56 |
| | | | Jönköpings län | 3,64 |
| | | | Södermanlands län | 2,87 |
| | | | Värmlands län | 3,07 |
| | | | Västmanlands län | 2,87 |
| | | | Västerbottens län | 2,99 |
| | | | Kalmar län | 2,4 |
| | | | Blekinge län | 1,67 |
| | | | Jämtlands län | 1,33 |
| | | | Kronobergs län | 1,89 |
| | | | Hallands län | 3,03 |
| | | | Gotlands län | 0,61 |
| 8 | KORKORT(2) | Driving license possession | ja | 75,02 |
| | | | nej | 24,98 |

| SN | Attributes(dim) | Definition | Category | Propotion % |
|---|---|---|---|---|
| 9 | SUN2KI(10) | Educational focus | humaniora_och_kons | 3,34 |
| | | | allmän_utbildning | 21,97 |
| | | | pedagogik_och_läraru | 6,43 |
| | | | samhällsvetenskap_ju | 14,24 |
| | | | teknik_och_tillverknir | 18,6 |
| | | | lant_och_skogsbruk_s | 1,19 |
| | | | naturvetenskap_mate | 2,29 |
| | | | tjänster | 4,94 |
| | | | hälso_och_sjukvård_s | 10,98 |
| | | | okänd | 16,03 |
| 10 | SKOL(2) | Student? | nej | 77,13 |
| | | | ja | 22,87 |
| 11 | APLATS(2) | Employed? | ja | 61,01 |
| | | | nej | 38,99 |
| 12 | BILANT(8) | Number of cars | 1 bilar | 52,46 |
| | | | 0 bilar | 14,53 |
| | | | 2 bilar | 28,33 |
| | | | 3 bilar | 3,67 |
| | | | 4 bilar | 0,57 |
| | | | 7 eller fler bilar | 0,11 |
| | | | 5 bilar | 0,24 |
| | | | 6 bilar | 0,09 |
| 13 | KKORT_HH(5) | N driving licenses in the household | inget körkort | 29,74 |
| | | | 1 körkort | 54,38 |
| | | | 2 körkort | 14,74 |
| | | | 3 körkort | 1,08 |
| | | | 4 körkort | 0,07 |
| 14 | HHINK(5) | Household income | 100 000 - < 200 000 | 11,57 |
| | | | 300 000 eller mer | 67,18 |
| | | | Under 100 000 | 3,04 |
| | | | 250 000 - < 300 000 | 9,05 |
| | | | 200 000 - < 250 000 | 9,15 |
| 15 | HHSTORL(6) | Household size | 1 person | 20 |
| | | | 2 personer | 32,72 |
| | | | 4 personer | 21,42 |
| | | | 3 personer | 14,92 |
| | | | 6- personer | 2,54 |
| | | | 5 personer | 8,4 |
| 16 | SNIKOD(13) | Industry | Transport | 4,5 |
| | | | Services | 13,02 |
| | | | Hospitality | 1,69 |
| | | | Agriculture | 31,62 |
| | | | Healthcare | 11,06 |
| | | | Education | 8,14 |
| | | | Retail | 7,83 |
| | | | Government | 4,74 |
| | | | Manufacturing | 11,18 |
| | | | WaterWaste | 0,76 |
| | | | Construction | 3,44 |
| | | | Recreation | 1,86 |
| | | | Extraction | 0,15 |
| 17 | UP_ANST(3) | Type of employment | permanent | 53,24 |
| | | | temporary | 7,78 |
| | | | inte anställd | 38,99 |

selecting a $q$ representing the number of attributes, and 2). introducing NaN values randomly to $r\%$ of each of the $q$ attributes in the *h-nomis* dataset. These datasets are denoted as *h-miss-q-r*, indicating the number of attributes and the proportion of rows with NaN values.

In the study, WGAN model trained with *h-nomis* act as the "benchmark" model against which the other WGAN models trained on *h-miss-q-r* will be analyzed and examined. In the end we generate synthetic poulation from each of the trained models and test them against *h-population* for evaluating the performance of these models.

***Table*** 2 presents a comprehensive summary of the training data employed for the analysis, accompanied by their corresponding attributes.

**Table 2: Statistics on the datasets used for training and analysis.**

| SN | Dataset | N Rows | N Attributes | N Categories | N Unique Combinations | Missing Attributes | Missing Value (%) | N Missing Rows (% of Total) |
|---|---|---|---|---|---|---|---|---|
| 1 | h-population | 5156896 | 17 | 107 | 14811 | - | - | - |
| 2 | h-nomis | 477489 | 17 | 107 | 12811 | - | - | - |
| 3 | h-miss-2-10 | 477489 | 17 | 107 | 12811 | UP_FORV, SUN2KN | 10 | 90576 (18.97%) |
| 4 | h-miss-2-20 | 477489 | 17 | 107 | 12808 | UP_FORV, SUN2KN | 20 | 171707 (35.96%) |
| 5 | h-miss-2-30 | 477489 | 17 | 107 | 12803 | UP_FORV, SUN2KN | 30 | 243406 (50.98%) |
| 6 | h-miss-2-40 | 477489 | 17 | 107 | 12779 | UP_FORV, SUN2KN | 40 | 305761 (64.04%) |
| 7 | h-miss-3-10 | 477489 | 17 | 107 | 12808 | UP_FORV, SUN2KN, SUN2KI | 10 | 129287 (27.08%) |
| 8 | h-miss-3-40 | 477489 | 17 | 107 | 12613 | UP_FORV, SUN2KN, SUN2KI | 40 | 374565 (78.44%) |
| 9 | h-miss-4-10 | 477489 | 17 | 107 | 12807 | UP_FORV, SUN2KN, SUN2KI, HHSTORL | 10 | 164094 (34.37%) |
| 10 | h-miss-4-40 | 477489 | 17 | 107 | 12154 | UP_FORV, SUN2KN, SUN2KI, HHSTORL | 40 | 415914 (87.10%) |
| 11 | h-miss-5-10 | 477489 | 17 | 107 | 12805 | UP_FORV, SUN2KN, SUN2KI, HHSTORL, SNIKOD | 10 | 195286 (40.90%) |
| 12 | h-miss-5-40 | 477489 | 17 | 107 | 11070 | UP_FORV, SUN2KN, SUN2KI, HHSTORL, SNIKOD | 40 | 440523 (92.26%) |

### 4.1.3 Sampling zero and Structural Zero

As a WGAN based models can lean joint probability distributions of the attributes and hence can produce a variety of category combinations. In this section, we outline the different kinds of distribution of the category combination and data samples that can be produced from WGAN models. This concept is visually illustrated in the ***Figure*** 3.

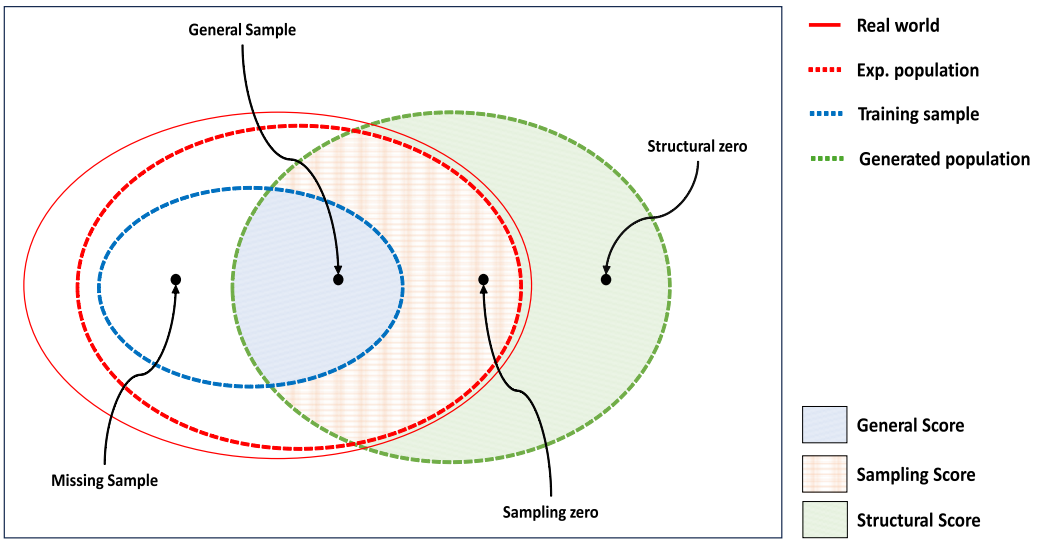



**Figure 3: Conceptual diagram showing the distribution of combination of categories and data types for our study.**

In the illustration, the black box represents the entirety of all possible combinations of the 107 categories in our study. Within this scope, the thin red circle outlines the distribution of category combinations observed in the actual population, highlighted as "Real world" in the illustration. As we are using the population that is derived from a travel survey, it may contain fewer category combinations when compared to actual population. The bold dashed RED circle illustrates the distribution of category combinations in the *h-population* dataset, labeled as "Exp. population" in the illustration and lies within the Real world distribution. The training datasets, *h-nomis* and *h-miss-q-r*, represent a fraction of the *h-population*, with certain combinations deliberately omitted from these samples. Consequently, this creates a distribution of category combinations smaller than the Exp. population, depicted as a bold dashed BLUE circle and referred to as the "Training Sample"

in the illustration. Lastly, the distribution of category combinations from the synthetic population generated by trained WGAN models is depicted as a bold dashed GREEN circle, labeled as "Generated population" in the illustration. This distribution of the may intersect with the other distributions at various levels.

Following the classifications defined by Kim and Bansal (2023), the synthetic population generated by WGAN models can be classified into four categories: general sample, sampling zero, structural zero, and missing sample. If a category combination is present in all distributions, it is termed a "general sample". The amount of general samples produced by the WGAN model is determined by the intersection of all distributions, depicted as the BLUE shaded region in the illustration. Categories that are absent in the Training Sample but feasible in the Exp. population and generated by the WGAN model are referred to as "sampling zero". The number of sampling zeros generated by the WGAN model is computed as the intersection of the Exp. population and Generated population, excluding the distribution of the Training Sample, and is illustrated as the RED shaded region. Next, categories generated by the WGAN model that are not part of either the Exp. population or Training Sample are categorized as "structural zero", shown as the GREEN shaded region. Lastly, samples present in the Training Sample but not generated by the WGAN model are termed "missing samples".

The objective of the population synthesis WGAN model is to achieve the highest level of intersection with both the Exp. population and Training Sample, thereby generating the maximum amount of general samples and sampling zeros while minimizing the number of structural zeros. As depicted in the illustration, it is desirable to have a large area of BLUE and RED shaded regions while reducing the area of the GREEN shaded region.

## 4.2 Model Evaluation and Discussion

To validate the effectiveness of our proposed approach for training WGAN with incomplete data, we initially utilized the *h-nomis* dataset to optimize the hyperparameters of the WGAN model. Subsequently, we applied the same optimized parameters to train additional models using various *h-miss-q-r* datasets. The hyperparameters specific to our scenario include number of layers and neurons for both the discriminator and generator, the dimension of the latent space vector, the learning rate, the regularization values for $\lambda_{bd}$, $\lambda_{ad}$ and gradient penalty $\lambda_{gp}$. The models underwent training on a GPU cluster comprising four NVIDIA GeForce RTX 3080 units, each with 10GB of memory, for a total of 1000 epochs. The optimized model parameters are detailed in ***Table*** 3.

Ultimately, the best-performing models were evaluated against the *h-population* dataset, which serves as a ground truth for the real population. To assess the models, we generated synthetic populations named $G_{nomis}$ using the model trained on the *h-nomis* dataset and use it as a benchmark to access other models. Another set of synthetic population, named $G_{miss-q-r}$, are generated using the model trained on the *h-miss-q-r* dataset. All these synthetically generated populations consist of 5,156,896 data points.

**Table 3: Parameters for the trained WGAN-GP models with regularization.**

| Parameter | Value |
|---|---|
| Discriminator - N Layers | 2 |
| Discriminator - N Neurons | 128 |
| Generator - N Layers | 2 |
| Generator - N Neurons | 128 |
| Latent Vector: N Neurons | 128 |
| Learning Rate | 0.01 |
| $\lambda_{gp}$ | 0.025 |
| $\lambda_{bd}$ | 10 |
| $\lambda_{ad}$ | 1 |

### 4.2.1 Attribute-level Evaluation

We first perform column level check on the WGAN generated synthetic populations to make sure that each attributes individually is able to follows statistical distribution

of ground-truth data, *h-population*. The attribute level check is done using three metrics - category coverage, TV complement and category adherence provided by Datacebo (2024) library.

Category coverage measures whether a attributes column in synthetic attribute covers all the possible categories that are present in ground-truth attribute. This metric first computes the number of unique categories, $C$, that are present in the ground-truth column $r$. Then it computes the number of those categories present in the synthetic attribute, $s$. It returns the proportion ground-truth categories that are in the synthetic data and is defined as,

$$score\_cc = \frac{C_s}{C_r} \tag{8}$$

Total variation (TV) complement metric computes the similarity of a ground-truth attribute vs. a synthetic attribute in terms of the column shapes i.e. the marginal distribution or 1D histogram of the column. This test computes the Total Variation Distance (TVD) between the ground-truth and synthetic attributes. To do this, it first computes the frequency of each category value and expresses it as a probability. The TVD statistic compares the differences in probabilities, as shown in 9.

$$\delta(R, S) = \frac{1}{2} \sum_{\omega \in \eta} |R_\omega - S_\omega| \tag{9}$$

Here, $\omega$ describes all the possible categories in a attribute, $\eta$. Meanwhile, R and S refer to the ground-truth and synthetic frequencies for those categories. The TV complement returns 1-TVD so that a higher score means higher quality and is give by,

$$score\_tv = 1 - \delta(R, S) \tag{10}$$

Category adherence metrics measures whether a synthetic attribute adheres to the same category values as the ground-truth data i.e. the synthetic population should not be inventing new category values that are not originally present in the ground-truth population. This metric extracts the set of unique categories, that are present in the ground-truth attribute, $Cr$. Then it finds the of data points of the

synthetic data, $s$, that are found in the set $C$. The score is the proportion of these data points as compared to all the synthetic data points and is given by,

$$score\_ca = \frac{|s, s \in C_r|}{|s|} \tag{11}$$

***Table*** 4 shows the metric scores for all attributes in the synthetic data generated by different WGAN models, tested against *h-population* dataset. The metric score presented here demonstrate that the proposed WGAN training method successfully trains a model with incomplete data that closely approximates the performance of benchmark model. The visual inspection on the bar graphs, presented in ***Appendix*** A for all 17 attributes for each of the dataset, further proves the analysis results.

Upon deeper evaluation of ***Table*** 4, it becomes apparent that certain $G_{miss-q-r}$ population exhibit slightly superior performance compared to benchmark $G_{nomis}$ population. Additionally, $G_{miss-q-r}$ population perform well for attributes with missing data - UP_FORV, SUN2KN, SUN2KI, HHSTORL, SNIKOD and have similar metrics to $G_{nomis}$. A closer examination reveals that all populations (excluding $G_{miss-4-10}$ and$G_{miss-5-40}$) exhibit lower score_cc for the "KKORT_HH" attribute. This stems from the inability of these WGAN models to generate any sample with the "4-körkort" category. Similarly, the same trend is observed for the "BILANT" attribute, where WGAN models struggle to generate sample with the "6 billar" category. However, such performance levels are deemed acceptable for these trained WGAN models, considering the negligible share of "4-körkort" and "6 billar" categories in the actual population, as indicated in ***Table*** 1.

### 4.2.2 Higher Dimension Evaluation

Following Kim and Bansal (2023); Garrido et al. (2020); Borysov et al. (2019); we compare the categorical partial joint distribution of synthetic population generated by WGAN models. We used a standardized root mean square error (SRMSE) and coefficient of determination $R^2$ as metric for evaluation multi-dimensional distributions. The SRMSE is given by,

**Table 4: Attribute level evaluation of all WGAN models against *h-population***

| SN | attributes | G_nomis | | | G_miss-2-10 | | | G_miss-2-20 | | | G_miss-2-30 | | | G_miss-2-40 | | | G_miss-3-10 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | score_cc | score_tv | score_ca | score_cc | score_tv | score_ca | score_cc | score_tv | score_ca | score_cc | score_tv | score_ca | score_cc | score_tv | score_ca | score_cc | score_tv | score_ca |
| 1 | AGE | 1,00 | 0,99 | 1,00 | 1,00 | 0,99 | 1,00 | 1,00 | 0,99 | 1,00 | 1,00 | 0,98 | 1,00 | 1,00 | 0,99 | 1,00 | 1,00 | 0,99 | 1,00 |
| 2 | SEX | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 0,99 | 1,00 | 1,00 | 0,99 | 1,00 |
| 3 | INKUP | 1,00 | 0,99 | 1,00 | 1,00 | 0,99 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 0,99 | 1,00 | 1,00 | 0,99 | 1,00 | 1,00 | 0,98 | 1,00 |
| 4 | UP_FORV | 1,00 | 0,99 | 1,00 | 1,00 | 0,99 | 1,00 | 1,00 | 0,98 | 1,00 | 1,00 | 0,99 | 1,00 | 1,00 | 0,99 | 1,00 | 1,00 | 0,99 | 1,00 |
| 5 | LIVSKAT | 1,00 | 0,99 | 1,00 | 1,00 | 0,98 | 1,00 | 1,00 | 0,99 | 1,00 | 1,00 | 0,98 | 1,00 | 1,00 | 0,99 | 1,00 | 1,00 | 0,99 | 1,00 |
| 6 | KORKORT | 1,00 | 0,99 | 1,00 | 1,00 | 0,99 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 0,99 | 1,00 | 1,00 | 0,99 | 1,00 |
| 7 | UP_ANST | 1,00 | 1,00 | 1,00 | 1,00 | 0,99 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 0,99 | 1,00 | 1,00 | 0,99 | 1,00 | 1,00 | 1,00 | 1,00 |
| 8 | SUN2KN | 1,00 | 0,98 | 1,00 | 1,00 | 0,99 | 1,00 | 1,00 | 0,99 | 1,00 | 1,00 | 0,99 | 1,00 | 1,00 | 0,98 | 1,00 | 1,00 | 0,97 | 1,00 |
| 9 | SUN2KI | 1,00 | 0,99 | 1,00 | 1,00 | 0,98 | 1,00 | 1,00 | 0,98 | 1,00 | 1,00 | 0,98 | 1,00 | 1,00 | 0,98 | 1,00 | 1,00 | 0,99 | 1,00 |
| 10 | SKOL | 1,00 | 1,00 | 1,00 | 1,00 | 0,98 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 0,99 | 1,00 | 1,00 | 0,99 | 1,00 | 1,00 | 1,00 | 1,00 |
| 11 | APLATS | 1,00 | 1,00 | 1,00 | 1,00 | 0,99 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 0,99 | 1,00 | 1,00 | 0,99 | 1,00 | 1,00 | 0,99 | 1,00 |
| 12 | BOST_LAN | 1,00 | 0,97 | 1,00 | 1,00 | 0,95 | 1,00 | 1,00 | 0,95 | 1,00 | 1,00 | 0,94 | 1,00 | 1,00 | 0,96 | 1,00 | 1,00 | 0,96 | 1,00 |
| 13 | BILANT | 0,88 | 0,98 | 1,00 | 1,00 | 0,98 | 1,00 | 1,00 | 0,98 | 1,00 | 0,88 | 0,99 | 1,00 | 1,00 | 0,98 | 1,00 | 0,88 | 0,98 | 1,00 |
| 14 | KKORT_HH | 0,80 | 0,99 | 1,00 | 0,80 | 0,98 | 1,00 | 0,80 | 0,98 | 1,00 | 0,80 | 0,98 | 1,00 | 0,80 | 0,99 | 1,00 | 0,80 | 0,98 | 1,00 |
| 15 | HHINK | 1,00 | 0,98 | 1,00 | 1,00 | 0,98 | 1,00 | 1,00 | 0,98 | 1,00 | 1,00 | 0,99 | 1,00 | 1,00 | 0,99 | 1,00 | 1,00 | 0,99 | 1,00 |
| 16 | HHSTORL | 1,00 | 0,98 | 1,00 | 1,00 | 0,99 | 1,00 | 1,00 | 0,98 | 1,00 | 1,00 | 0,98 | 1,00 | 1,00 | 0,99 | 1,00 | 1,00 | 0,99 | 1,00 |
| 17 | SNIKOD | 0,92 | 0,97 | 1,00 | 0,92 | 0,97 | 1,00 | 1,00 | 0,98 | 1,00 | 0,92 | 0,98 | 1,00 | 1,00 | 0,97 | 1,00 | 0,92 | 0,97 | 1,00 |
| | **Average** | 0,9764 | 0,9862 | 1,0000 | 0,9837 | 0,9847 | 1,0000 | 0,9882 | 0,9862 | 1,0000 | 0,9764 | 0,9851 | 1,0000 | 0,9882 | 0,9854 | 1,0000 | 0,9764 | 0,9856 | 1,0000 |

| SN | attributes | G_miss-3-40 | | | G_miss-4-10 | | | G_miss-4-40 | | | G_miss-5-10 | | | G_miss-5-40 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | score_cc | score_tv | score_ca | score_cc | score_tv | score_ca | score_cc | score_tv | score_ca | score_cc | score_tv | score_ca | score_cc | score_tv | score_ca |
| 1 | AGE | 1,00 | 0,98 | 1,00 | 1,00 | 0,99 | 1,00 | 1,00 | 0,99 | 1,00 | 1,00 | 0,98 | 1,00 | 1,00 | 0,99 | 1,00 |
| 2 | SEX | 1,00 | 1,00 | 1,00 | 1,00 | 0,98 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 |
| 3 | INKUP | 1,00 | 1,00 | 1,00 | 1,00 | 0,99 | 1,00 | 1,00 | 0,99 | 1,00 | 1,00 | 0,99 | 1,00 | 1,00 | 0,99 | 1,00 |
| 4 | UP_FORV | 1,00 | 0,99 | 1,00 | 1,00 | 0,99 | 1,00 | 1,00 | 0,99 | 1,00 | 1,00 | 0,99 | 1,00 | 1,00 | 0,98 | 1,00 |
| 5 | LIVSKAT | 1,00 | 0,98 | 1,00 | 1,00 | 0,98 | 1,00 | 1,00 | 0,98 | 1,00 | 1,00 | 0,98 | 1,00 | 1,00 | 0,98 | 1,00 |
| 6 | KORKORT | 1,00 | 0,99 | 1,00 | 1,00 | 0,99 | 1,00 | 1,00 | 0,99 | 1,00 | 1,00 | 0,99 | 1,00 | 1,00 | 0,99 | 1,00 |
| 7 | UP_ANST | 1,00 | 1,00 | 1,00 | 1,00 | 0,99 | 1,00 | 1,00 | 0,99 | 1,00 | 1,00 | 0,99 | 1,00 | 1,00 | 0,99 | 1,00 |
| 8 | SUN2KN | 1,00 | 0,94 | 1,00 | 1,00 | 0,97 | 1,00 | 1,00 | 0,98 | 1,00 | 1,00 | 0,97 | 1,00 | 1,00 | 0,98 | 1,00 |
| 9 | SUN2KI | 1,00 | 0,95 | 1,00 | 1,00 | 0,98 | 1,00 | 1,00 | 0,95 | 1,00 | 1,00 | 0,98 | 1,00 | 1,00 | 0,89 | 1,00 |
| 10 | SKOL | 1,00 | 0,99 | 1,00 | 1,00 | 0,99 | 1,00 | 1,00 | 0,99 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 0,99 | 1,00 |
| 11 | APLATS | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 0,99 | 1,00 | 1,00 | 0,99 | 1,00 | 1,00 | 0,98 | 1,00 |
| 12 | BOST_LAN | 1,00 | 0,95 | 1,00 | 1,00 | 0,95 | 1,00 | 1,00 | 0,94 | 1,00 | 1,00 | 0,96 | 1,00 | 1,00 | 0,95 | 1,00 |
| 13 | BILANT | 0,75 | 0,98 | 1,00 | 0,88 | 0,98 | 1,00 | 1,00 | 0,98 | 1,00 | 0,88 | 0,98 | 1,00 | 1,00 | 0,99 | 1,00 |
| 14 | KKORT_HH | 0,80 | 0,98 | 1,00 | 1,00 | 0,98 | 1,00 | 0,80 | 0,98 | 1,00 | 0,80 | 0,98 | 1,00 | 1,00 | 0,99 | 1,00 |
| 15 | HHINK | 1,00 | 0,99 | 1,00 | 1,00 | 0,99 | 1,00 | 1,00 | 0,99 | 1,00 | 1,00 | 0,98 | 1,00 | 1,00 | 0,99 | 1,00 |
| 16 | HHSTORL | 1,00 | 0,98 | 1,00 | 1,00 | 0,98 | 1,00 | 1,00 | 0,97 | 1,00 | 1,00 | 0,98 | 1,00 | 1,00 | 0,96 | 1,00 |
| 17 | SNIKOD | 0,92 | 0,98 | 1,00 | 0,92 | 0,98 | 1,00 | 1,00 | 0,98 | 1,00 | 1,00 | 0,98 | 1,00 | 0,92 | 0,96 | 1,00 |
| | **Average** | 0,9690 | 0,9817 | 1,0000 | 0,9881 | 0,9839 | 1,0000 | 0,9882 | 0,9826 | 1,0000 | 0,9809 | 0,9842 | 1,0000 | 0,9955 | 0,9764 | 1,0000 |

$$SRMSE(\pi, \hat{\pi}|k) = \frac{RMSE}{\overline{\pi}} = \frac{\sqrt{\sum_{(i,j)}(\pi_{(i,j)} - \hat{\pi}_{(i,j)})^2/N_b}}{\sum_{(i,j)} \pi_{(i,j)}/N_b} \tag{12}$$

where $\pi$ and $\hat{\pi}$ are k-joint categorical distribution of ground-truth and synthetic population, respectively. $N_b$ is the total number of possible category combinations and $k$ is the number of attributes in the joint. The $R^2$ score are computed using the categorical distribution for both the ground-truth and synthetic population, and is given by,

$$R^2 = 1 - \frac{\sum_{(i,j)}(\pi_{(i,j)} - \hat{\pi}_{(i,j)})^2}{\sum_{(i,j)}(\pi_{(i,j)} - \bar{\pi}_{(i,j)})^2} \tag{13}$$

where $\bar{\pi}$ is the mean of the k-joint categorical distribution of ground-truth population.

We employ distinct subsets of attributes to encompass diverse features of the joint distribution. The evaluation is done for the following k-joint of attributes with total categories in brackets:

- 210-dimensional joint of AGE(5), SEX(2) and BOST_LAN(21).
- 16k-dimensional joint of UP_FORV(3), SUN2KN(7), SUN2KI(10), HHSTORL(6) and SNIKOD(13).
- 7M-dimensional joint of AGE(5), SEX(2), BOST_LAN(21), SUN2KI(10), SNIKOD(13), LIVSKAT(8), SUN2KN(7), and INKUP(5).

***Table*** 5 displays the outcomes derived from the SRMSE and $R^2$ metrics for all WGAN model across various k-joint levels. We also present the counts for the count of unique category combinations that can be generated, are in the *h-population* and are in each synthetic population, for all k-level joints. Comprehensive 45-degree charts for all models are provided in ***Appendix*** B for visual and qualitative evaluation.

The findings indicate that for all $G_{miss-q-r}$ population, their performance falls short of the benchmark $G_{nomis}$ population in terms of both SRMSE and $R^2$ values, across all k-dimensional joints. Consequently, as the amount of missing information in the

training data increases, the SRMSE values (and corresponding $R^2$ values) tend to rise. This trend is particularly evident in the case of the 16k-dimensional joint, where attributes contain missing information in the dimensional joint. As the number of attributes with missing information grows, so does the SRMSE value, with $G_{miss-5-40}$ population exhibiting the highest error. In general, the $G_{miss-q-r}$ population only exhibits SRMSE and $R^2$ very close to the benchmark population $G_{nomis}$. Hence, it can be concluded that the proposed WGAN training method successfully trains a model with incomplete data.

It is noteworthy that with an increase in the number of joint combinations, the SRMSE value rises (while $R^2$ decreases) across all models. This phenomenon occurs because the WGAN models generate a substantially larger number of category combinations compared to the ground-truth population. Some of these combinations are present within the ground-truth population, while others are not. We analyze this behavior by examining the count of sampling and structural zero data points produced by the models and evaluate the models' performance concerning the category combinations.

**Table 5: Higher dimension evaluation for k-joint level distribution of attributes for all models.**

| | | 210-dimensional (Total:210, Real:210) | | | 16k-dimensional (Total:16380, Real:2485) | | | 7M-dimensional (Total:7644000, Real:10395) | | |
|---|---|---|---|---|---|---|---|---|---|---|
| SN | Model | N Comb. Generated | mSRMSE | R2 | N Comb. Generated | mSRMSE | R2 | N Comb. Generated | mSRMSE | R2 |
| 1 | G_nomis | 210 | 0,19 | 0,9813 | 11897 | 1,75 | 0,9781 | 365530 | 25,79 | 0,9127 |
| 2 | G_miss-2-10 | 210 | 0,27 | 0,9625 | 11247 | 1,67 | 0,9801 | 343796 | 27,56 | 0,9003 |
| 3 | G_miss-2-20 | 210 | 0,24 | 0,9689 | 12060 | 1,78 | 0,9775 | 358073 | 26,36 | 0,9088 |
| 4 | G_miss-2-30 | 210 | 0,25 | 0,9668 | 11372 | 1,97 | 0,9725 | 349566 | 26,82 | 0,9056 |
| 5 | G_miss-2-40 | 210 | 0,24 | 0,9710 | 11260 | 1,89 | 0,9746 | 352989 | 27,82 | 0,8984 |
| 6 | G_miss-3-10 | 210 | 0,24 | 0,9704 | 10643 | 2,08 | 0,9693 | 343689 | 28,99 | 0,8896 |
| 7 | G_miss-3-40 | 210 | 0,21 | 0,9760 | 12022 | 2,26 | 0,9636 | 414874 | 26,61 | 0,9070 |
| 8 | G_miss-4-10 | 210 | 0,23 | 0,9722 | 11316 | 2,16 | 0,9670 | 350875 | 26,77 | 0,9059 |
| 9 | G_miss-4-40 | 210 | 0,27 | 0,9610 | 12638 | 2,22 | 0,9651 | 394555 | 26,32 | 0,9091 |
| 10 | G_miss-5-10 | 210 | 0,25 | 0,9669 | 12010 | 2,24 | 0,9642 | 348536 | 29,40 | 0,8865 |
| 11 | G_miss-5-40 | 210 | 0,24 | 0,9690 | 12480 | 2,61 | 0,9515 | 389556 | 28,60 | 0,8926 |

### 4.2.3 Sampling and Structural Zero Evaluation

Based on the definitions provided in ***Section*** 4.1.3, we extracted general sample, sampling zero, and structural zero data for each generated population. This was done for k-joint attribute combination described in ***Section*** 4.2.2 and at varied sampling levels. We computed the ratio of both general sample and sampling zero in the generated population in comparison to the combinations present in the ground-truth population (*h-population*). We extracts the set of general samples ($GS$) or sampling zero ($SZ$) that exists in *h-population*. Then do the same for all WGAN generated population. The score is defined as,

$$score\_gs = \frac{GS_{generated}}{GS_{ground-truth}} \tag{14}$$

$$score\_sz = \frac{SZ_{generated}}{SZ_{ground-truth}} \tag{15}$$

The ratio of structural zero is determined by calculating the total number of structural zero ($STZ$) instances generated by the WGAN models against all the unique combinations, $C$ produced by the WGAN model itself and is defined as,

$$score\_stz = \frac{STZ_{generated}}{C_{generated}} \tag{16}$$

In alignment with Kim and Bansal (2023), we also implemented precision and recall to compare the models. Precision check weather the synthetic data generated new attributes combinations that still resembles the actual population. Recall measures the extent of over-fitting to the training sample. The value of precision and recall is given by,

$$Precison = \frac{1}{C}\sum_{j=1}^{C} 1_{\hat{\pi}_j \in \pi} \tag{17}$$

$$Recall = \frac{1}{C}\sum_{j=1}^{C} 1_{\pi_j \in \hat{\pi}} \tag{18}$$

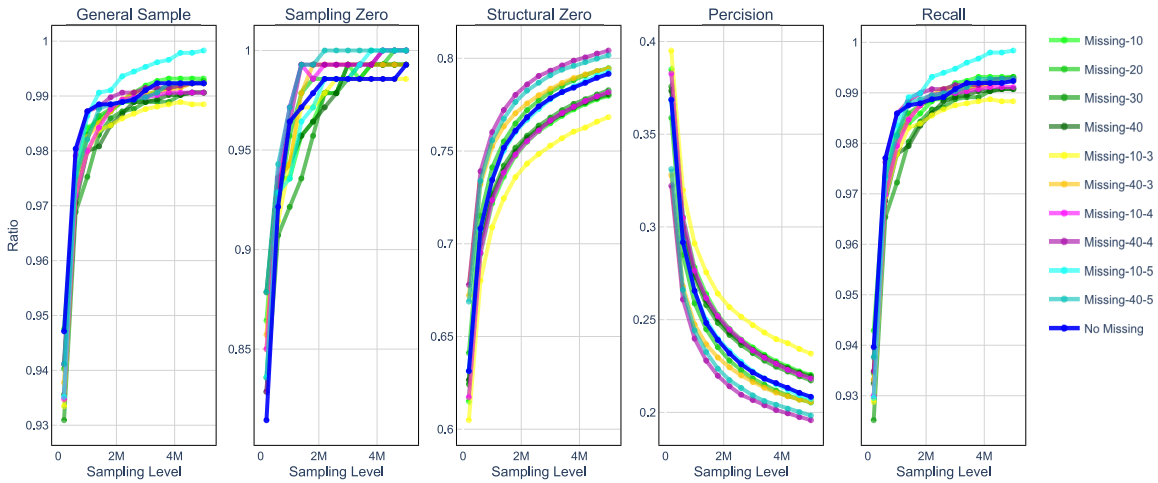
***Figure*** 4 displays the evaluation plots for all models, obtained from synthetic data generated from for 16k-dimensional and 7M-dimensional categorical joint, mentioned in ***Section*** 4.2.2.

In both presented plots, the metrics from all models exhibit similar or superior scores compared to the benchmark $G_{nomis}$ population. The metric score presented here demonstrate that the proposed WGAN training method successfully trains a model with incomplete data that closely approximates the performance of benchmark model.

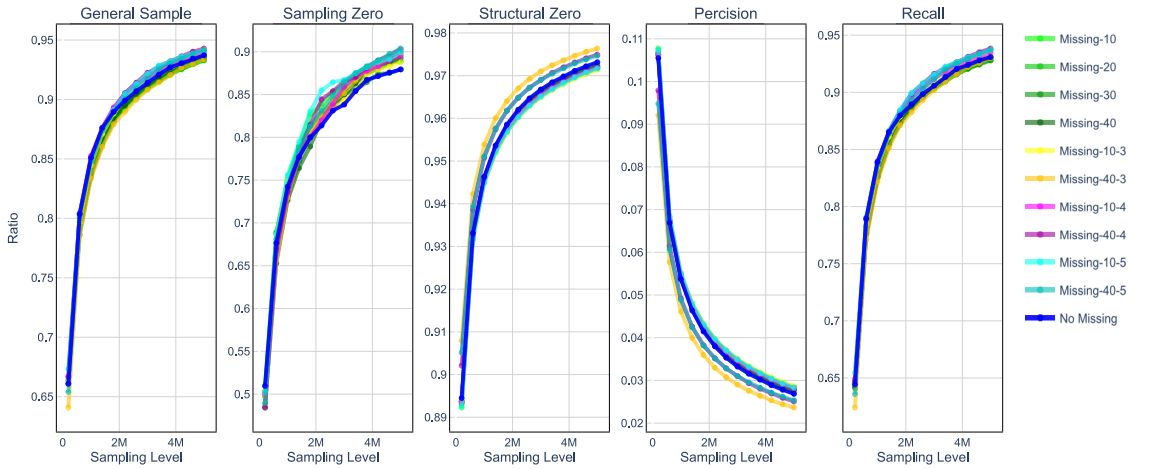
With respect to the metric analysis, all models have successfully generated nearly all general samples and sampling zeros present in the ground-truth population. Generally, the number of sampling zeros and general samples generated increases as more data is sampled but decreases with the expansion of the k-dimensional space. For instance, when sampling from a 7M-dimensional distribution at a sampling rate of 5M data points, approximately 93.69% $\pm$ 0.36% of general samples and 89.42% $\pm$ 0.8% of sampling zeros are captured. At the same time, the number of structural zero also increases with the sampling rate as well as k-dimensional space, across all models. This is due to the fact, as the number of attribute dimensions is increasing, more and more combinations are possible within the data. The issue is that the total number of combinations that are present in the ground-truth population is very small compared to all possible unique combinations that are possible in an actual real world population. In this study, the *h-population* only contains 10 395 unique combinations against the 7M possible combinations in the data which is only 0.13% of all possible combinations. WGAN model struggles to restrict generation of structural zeros for such a small set of unique combinations in ground-truth data, especially for a high

dimensional cases. This in fact is evident from the very low values for precision for all trained models, including the the base model $G_{nomis}$.

In comparison to the results reported by Kim and Bansal (2023), the precision values in our study are significantly lower. Possible explanation for this difference could be the unequal distribution of unique category combinations in the ground-truth data. Regarding Kim and Bansal (2023), the dataset consists of 264,005 distinct combinations, while the *h-population* used in the study only includes 14,811 distinct combinations. This dataset is 17 times smaller than the dataset mentioned in Kim and Bansal (2023). Consequently, the precision and recall metrics rely on the quantity of distinct category combinations found in the actual data. Models trained on datasets with a greater number of category combinations will exhibit superior performance in terms of precision and recall metrics. This can be attributed to the capacity of WGAN models to generate a vast array of category combinations.




**(a) 16k-dim joint**




**(b) 7M-dim joint**

**Figure 4: Plots with the ratio of general sample, sampling zero, structural zero, precision and recall for 16k and 7M dimensional joint data at different sampling levels.**

# 5 Conclusion

The paper presents a novel method for population synthesis using WGAN models, which effectively train on incomplete data to ensure the synthetic population generated by the models is complete. This property of the model is especially useful when publicly accessible microsamples have missing information on one or more attributes; due to errors in data collection, privacy concerns resulting in data being withheld, or when information is missing during the merging of multiple microsamples. The proposed methodology utilizes a mask matrix to depict missing values in training data that allows the WGAN model to train on datasets that contain missing attributes.

The training method was validated using data from the Swedish national travel survey. We conducted a comparison between the benchmark model that was trained using complete data against models that were trained using data with different levels of missing information. The population generated from all the trained models was evaluated at the attribute-level and higher k-dimensional level to assess model's capability in generating sampling and structural zeros. For all the evaluation metrics, the results obtained from trained WGAN on incomplete data exhibit a high degree of similarity with the benchmark WGAN model trained on complete data. The validation results affirm the efficacy of the suggested training technique employing mask matrix in proficiently managing incomplete data, leading to synthetic populations that closely resemble the real population.

Upon closer examination of the evaluation results, it became evident that all trained WGAN models exhibited suboptimal performance on metrics relying on precision or SRMSE calculations, particularly in high-dimensional scenarios. This is attributed to the presence of numerous structural zeros generated by WGAN models, stemming from the limited number of unique category combinations in the travel survey data used for training in the study compared to the vast range of potential category combinations in real world. It is concluded that improved population data with a higher number of category combinations would enhance the metric results. Despite the significant number of challenges posed by structural zeros, all models excel in generating general samples and accurately sampling zeros, across various combinations of attributes and sampling levels.

The paper makes a substantial contribution to the field by providing a strong solution for population synthesis using incomplete data. This discovery presents new opportunities for future investigation, emphasizing the capacity of deep generative models to enhance the abilities of population synthesis, which is essential for agent-based models (ABMs) employed in transportation simulations and other fields.

# 6 Future works

Subsequently, the forthcoming course of action entails synthesizing the future population by employing the trained WGAN models. The conventional approach for this task involves employing Iterative Proportional Fitting (IPF) or Combinatorial Optimization (CO) techniques on a representative sample of the current year's population. This allows for the generation of a simulated population for a given future scenario that closely matches the desired distribution characteristics. An intriguing avenue for investigation would involve examining the feasibility of accomplishing this solely using an alternative deep learning model. Conditional Tabular GANs (CT-GAN) is a promising starting point for synthesizing data based on specified conditions. However, the current design of CT-GAN does not allow for conditioning on marginals. Further research is required to investigate the application of CT-GAN in generating synthetic populations for future scenarios, while considering marginal conditions.

# 7 Acknowledgment

The computations and data handling was enabled by the supercomputing resource Berzelius provided by National Supercomputer Centre at Linköping University and the Knut and Alice Wallenberg foundation.

# References


Abramowski, L. and Holmström, A. (2007). Den nationella resvaneundersökningen 2005-2006 (RES). Technical report, Swedish Institute for Transport and Communications Analysis (SIKA), Stockholm.

Baowaly, M. K., Lin, C.-C., Liu, C.-L., and Chen, K.-T. (2019). Synthesizing electronic health records using improved generative adversarial networks. *Journal of the American Medical Informatics Association : JAMIA*, 26(3):228–241.

Bastarianto, F. F., Hancock, T. O., Choudhury, C. F., and Manley, E. (2023). Agent-based models in urban transportation: review, challenges, and opportunities. *European Transport Research Review*, 15(1):19.

Beckman, R. J., Baggerly, K. A., and McKay, M. D. (1996). Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice*, 30(6):415–429.

Borysov, S. S., Rich, J., and 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.

Choupani, A. A. and Mamdoohi, A. R. (2016). Population Synthesis Using Iterative Proportional Fitting (IPF): A Review and Future Research. In *Transportation Research Procedia*, volume 17, pages 223–233. Elsevier B.V.

Datacebo (2024). SD Metrics.

Deming, W. E. and Stephan, F. F. (1940). On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known. *Source: The Annals of Mathematical Statistics*, 11(4):427–444.

E. Ramadan, O. and P. Sisiopiku, V. (2020). A Critical Review on Population Synthesis for Activity- and Agent-Based Transportation Models. In *Transportation Systems Analysis and Assessment*. IntechOpen.

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., and Tabona,

O. (2021). A survey on missing data in machine learning. *Journal of Big Data*, 8(1):140.

Farooq, B., Bierlaire, M., Hurtubia, R., and Flötteröd, G. (2013). Simulation based population synthesis. *Transportation Research Part B: Methodological*, 58:243–263.

Garrido, S., Borysov, S. S., Pereira, F. C., and Rich, J. (2020). Prediction of rare feature combinations in population synthesis: Application of deep generative modelling. *Transportation Research Part C: Emerging Technologies*, 120.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative Adversarial Nets. In Z. Ghahramani and M. Welling and C. Cortes and N. Lawrence and K.Q. Weinberger, editor, *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. (2017). Improved Training of Wasserstein GANs. In I. Guyon and U. Von Luxburg and S. Bengio and H. Wallach and R. Fergus and S. Vishwanathan and R. Garnett, editor, *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Kim, E. J. and Bansal, P. (2023). A deep generative model for feasible and diverse population synthesis. *Transportation Research Part C: Emerging Technologies*, 148.

Neves, D. T., Alves, J., Naik, M. G., Proença, A. J., and Prasser, F. (2022). From Missing Data Imputation to Data Generation. *Journal of Computational Science*, 61.

Rich, J. (2018). Large-scale spatial population synthesis for Denmark. *European Transport Research Review*, 10(2).

Sun, L. and Erath, A. (2015). A Bayesian network approach for population synthesis. *Transportation Research Part C: Emerging Technologies*, 61:49–62.

Sun, L., Erath, A., and Cai, M. (2018). A hierarchical mixture modeling framework for population synthesis. *Transportation Research Part B: Methodological*, 114:199–212.

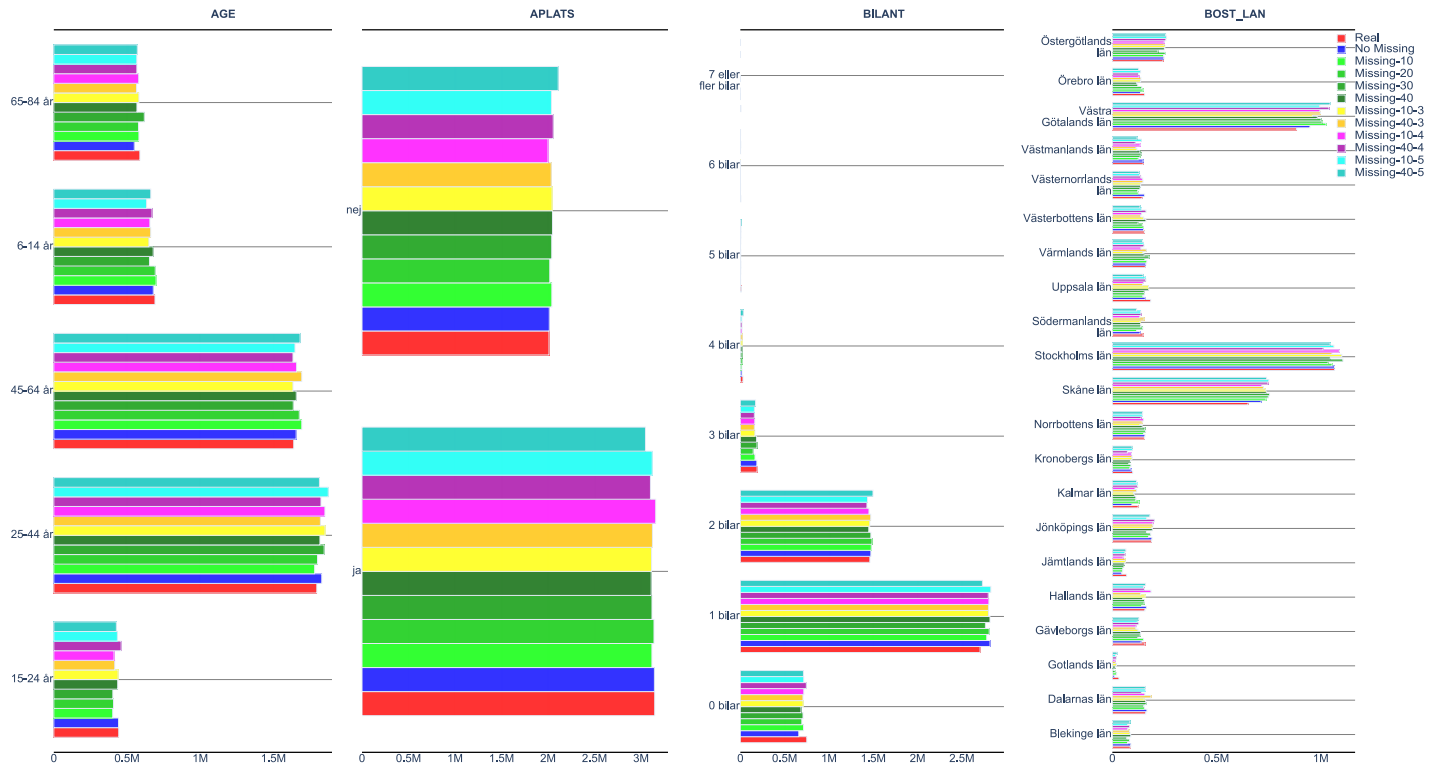
Walia, M., Tierney, B., and Mckeever, S. (2020). Synthesising Tabular Data using Wasserstein Conditional GANs with Gradient Penalty (WCGAN-GP). In *AICS 2020: 28th Irish Conference on Artificial Intelligence and Cognitive Science*, Dublin.

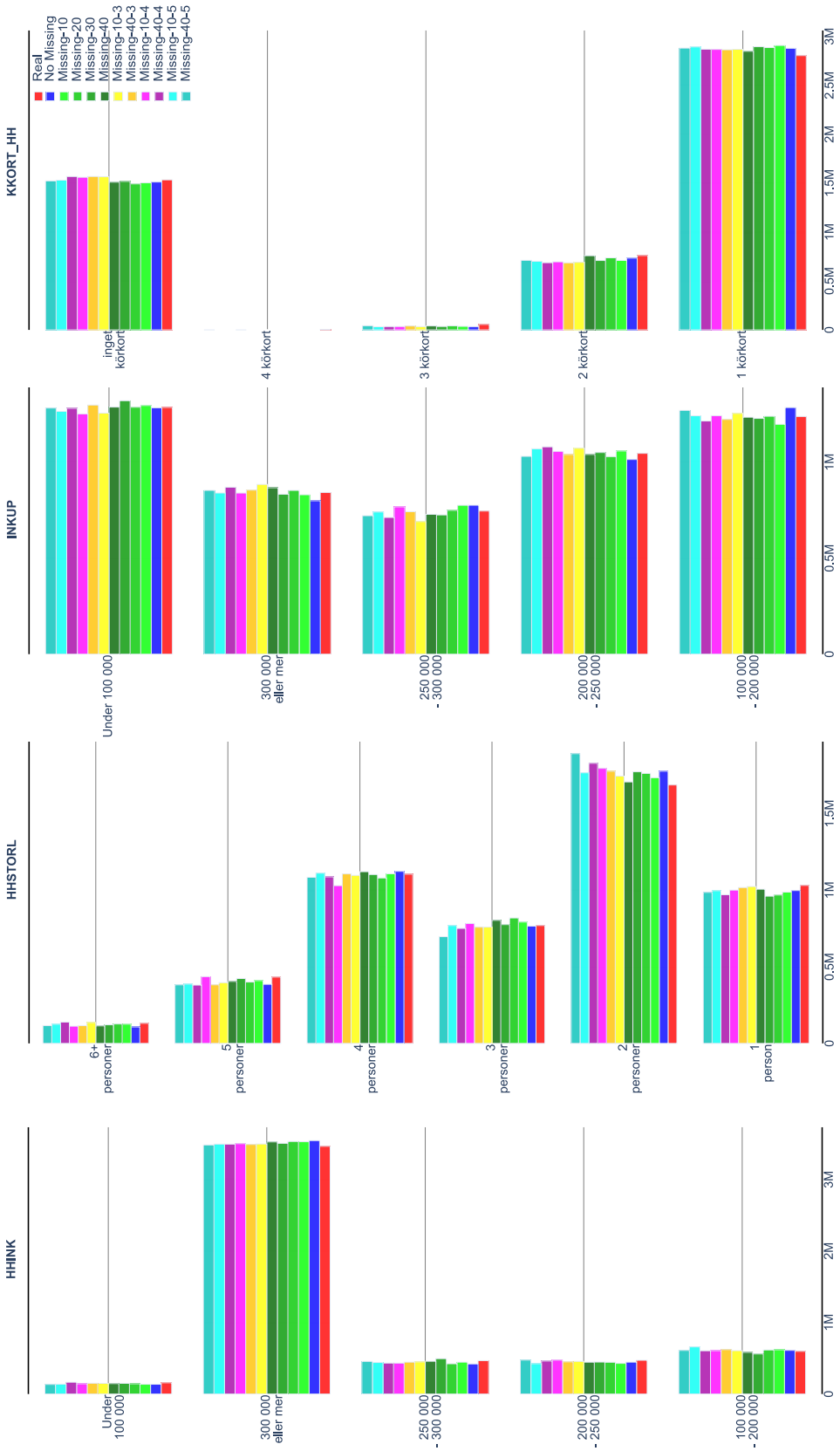
Xu, L., Skoularidou, M., Cuesta-Infante, A., and Veeramachaneni, K. (2019). Modeling Tabular data using Conditional GAN. *Advances in Neural Information Processing Systems (NeurIPS)*, 32:7335–7345.

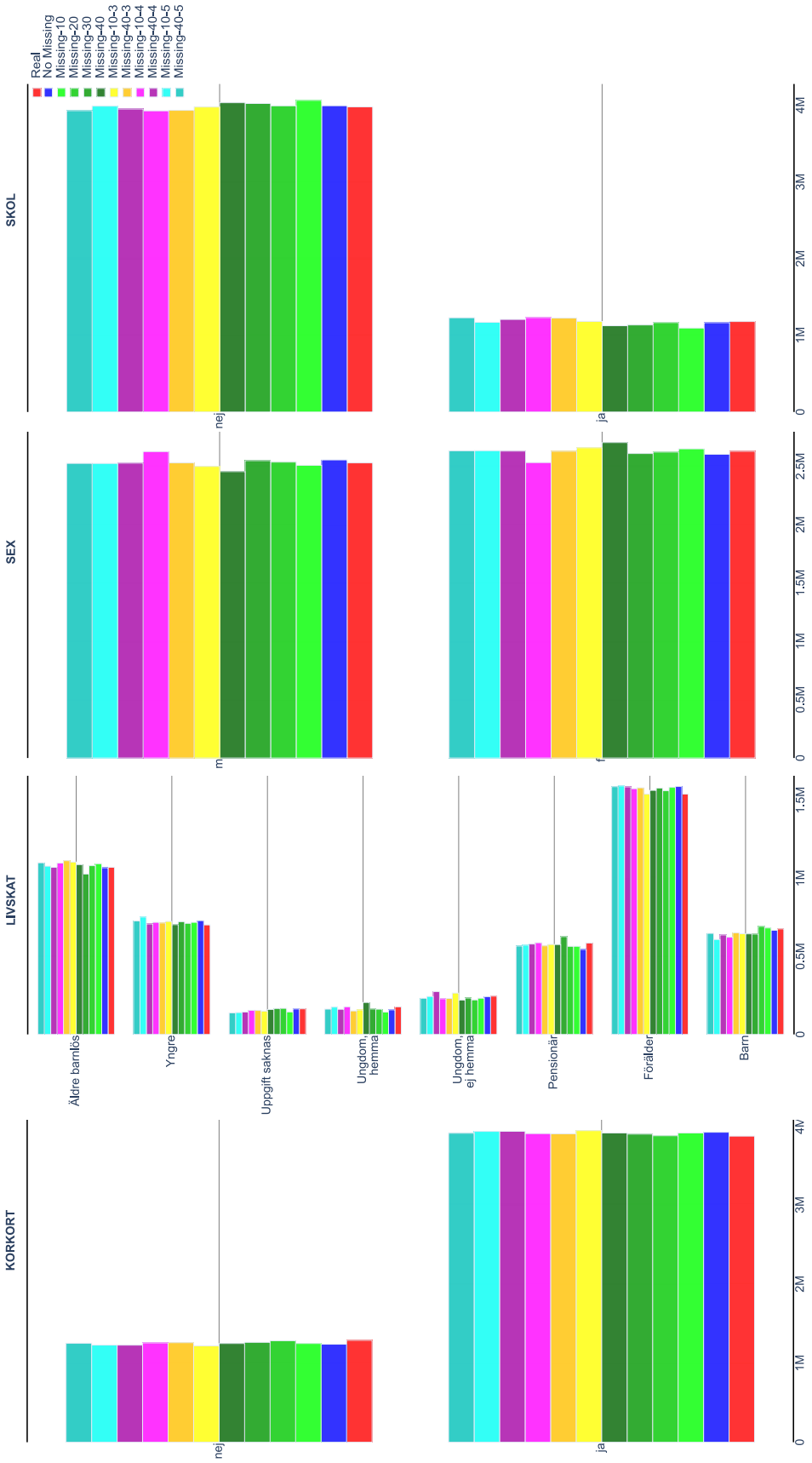
Yoon, J., Jordon, J., and van der Schaar, M. (2018). GAIN: Missing Data Imputation using Generative Adversarial Nets. *Proceedings of the 35th International Conference on Machine Learning*, pages 5689–5698.

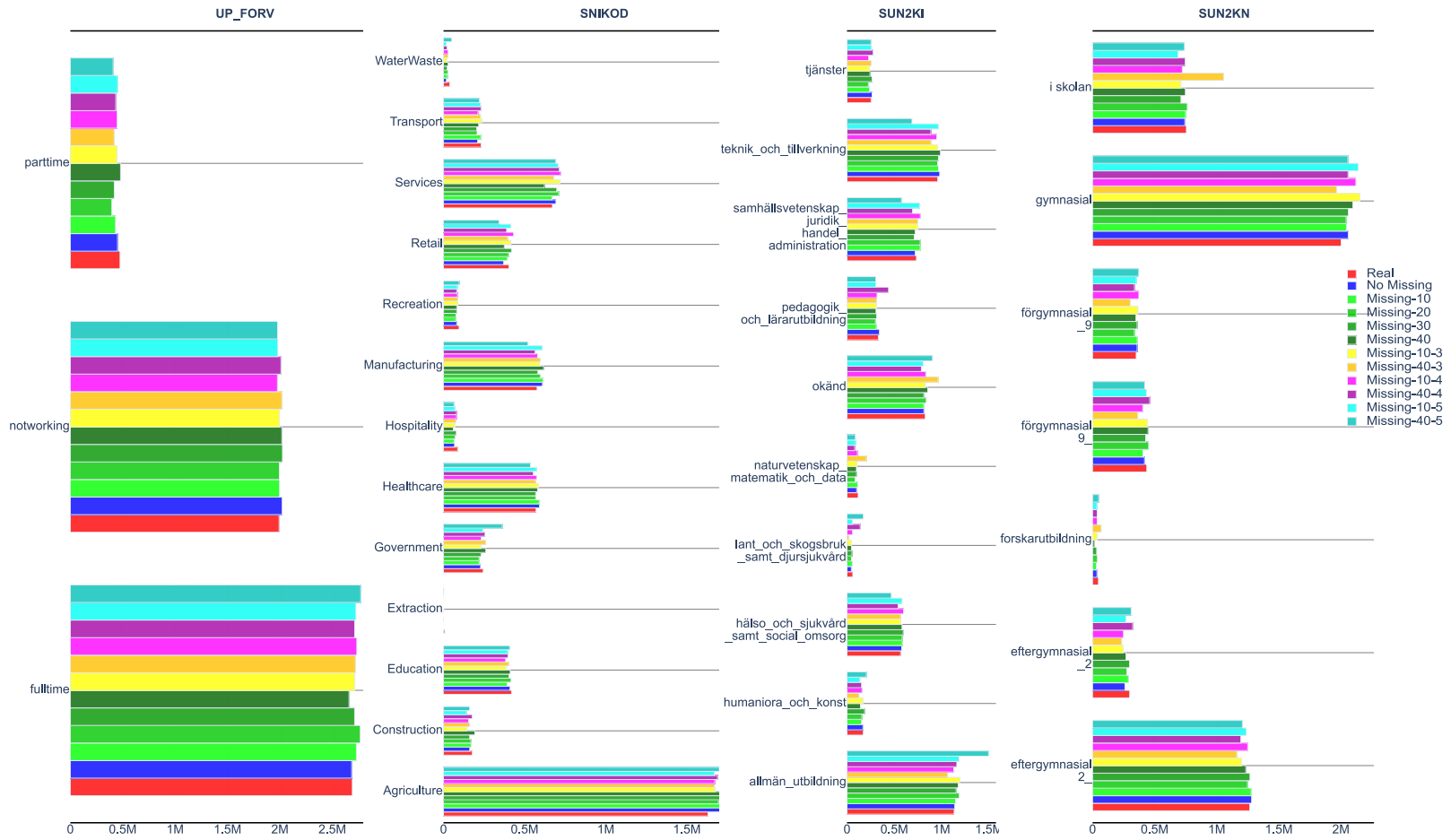
Zhu, Y. and Ferreira, J. (2014). Synthetic Population Generation at Disaggregated Spatial Scales for Land Use and Transportation Microsimulation. *Transportation Research Record: Journal of the Transportation Research Board*, 2429(1):168–177.

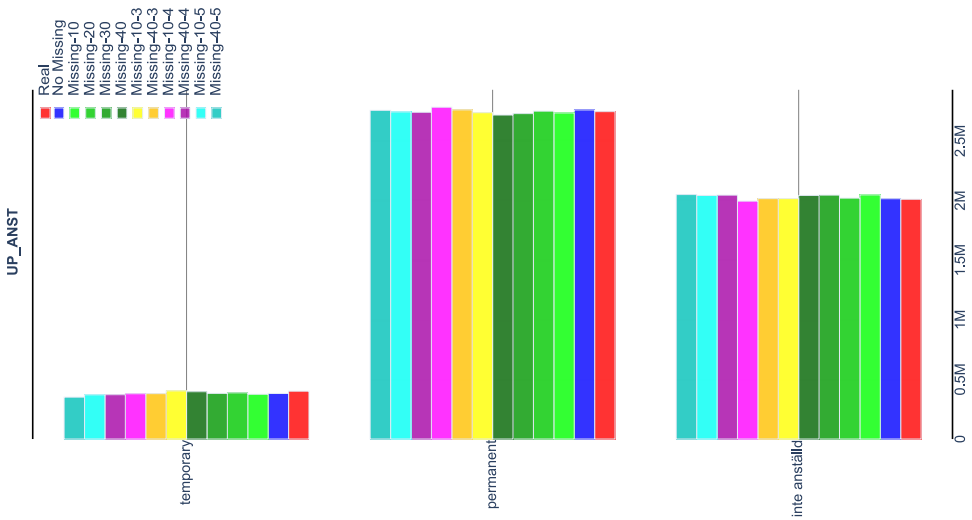
# A Bar graphs for 17 attributes for different types of data



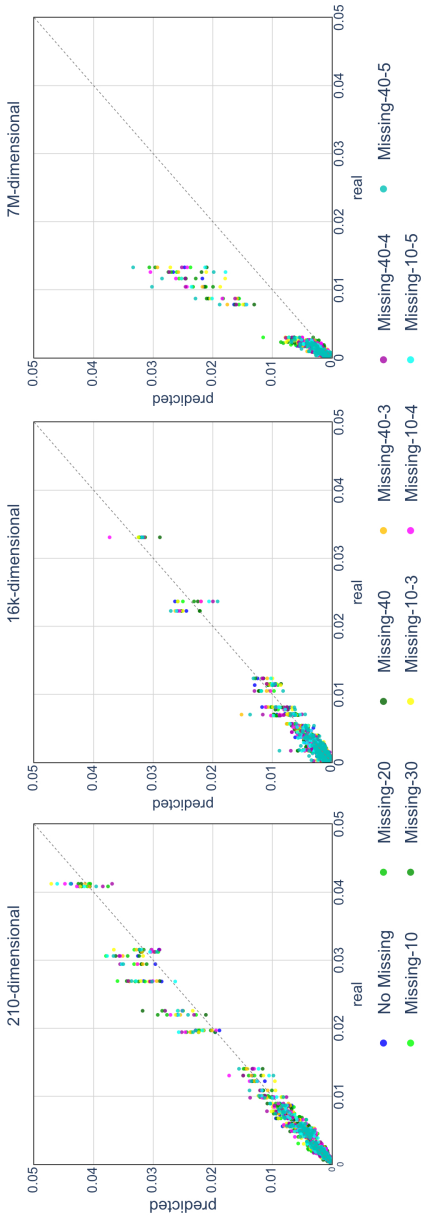

KKORT_HH
INKUP
HHSTORL
HHINK
Real
No Missing
Missing-10
Missing-20
Missing-30
Missing-40
Missing-10-3
Missing-40-3
Missing-10-4
Missing-40-4
Missing-10-5
Missing-40-5
inget körkort
4 körkort
3 körkort
2 körkort
1 körkort
Under 100 000
300 000 eller mer
250 000 - 300 000
200 000 - 250 000
100 000 - 200 000
6+ personer
5 personer
4 personer
3 personer
2 personer
1 person
0
0,5M
1M
1,5M
2M
2,5M
3M


Real
No Missing
Missing-10
Missing-20
Missing-30
Missing-40
Missing-10-3
Missing-40-3
Missing-10-4
Missing-40-4
Missing-10-5
Missing-40-5
SKOL
nej
ja
SEX
m
f
LIVSKAT
Äldre barnlös
Yngre
Uppgift saknas
Ungdom, hemma
Ungdom, ej hemma
Pensionär
Förälder
Barn
KORKORT
nej
ja
0
0.5M
1M
1.5M
2M
2.5M
3M
4M


UP_FORV
parttime
notworking
fulltime
0
0.5M
1M
1.5M
2M
2.5M
SNIKOD
WaterWaste
Transport
Services
Retail
Recreation
Manufacturing
Hospitality
Healthcare
Government
Extraction
Education
Construction
Agriculture
0
0.5M
1M
1.5M
SUN2KI
tjänster
teknik_och_tillverkning
samhällsvetenskap_
juridik_
handel_
administration
pedagogik_
och_lärarutbildning
okänd
naturvetenskap_
matematik_och_data
lant_och_skogsbruk
_samt_djursjukvård
hälso_och_sjukvård
_samt_social_omsorg
humaniora_och_konst
allmän_utbildning
0
0.5M
1M
1.5M
SUN2KN
i skolan
gymnasial
förgymnasial
_9
förgymnasial
9_
forskarutbildning
eftergymnasial
_2
eftergymnasial
2_
0
0.5M
1M
1.5M
2M
Real
No Missing
Missing-10
Missing-20
Missing-30
Missing-40
Missing-10-3
Missing-40-3
Missing-10-4
Missing-40-4
Missing-10-5
Missing-40-5


UP_ANST
Real
No Missing
Missing-10
Missing-20
Missing-30
Missing-40
Missing-10-3
Missing-40-3
Missing-10-4
Missing-40-4
Missing-10-5
Missing-40-5
temporary
permanent
inte anställd
0
0.5M
1M
1.5M
2M
2.5M

# B 45-degree charts for higher dimensional qualitative assessment


210-dimensional
16k-dimensional
7M-dimensional
predicted
real
0
0.01
0.02
0.03
0.04
0.05
No Missing
Missing-10
Missing-20
Missing-30
Missing-40
Missing-10-3
Missing-40-3
Missing-10-4
Missing-40-4
Missing-10-5
Missing-40-5