A deep generative model for feasible and diverse population synthesis

Kim, Eui-Jin, and Prateek Bansal. "A deep generative model for feasible and diverse population synthesis." Transportation Research Part C: Emerging Technologies 148 (2023): 104053.

B4 門坂佑真

Abstract

- 人口合成(Population synthesis)において，**sampling zero**をカバーしつつ，**structural zero**を最小限に抑えるための2つの損失関数を提案
- 生成モデルの改善指標として，**feasibility**と**diversity**を用いる
- 提案した損失関数により，従来のモデルと比較して精度向上を達成
- ABMの初期段階である合成人口を改善することで，後続のモデリング段階への誤差伝搬を回避する

- Two loss functions to cover **sampling zeros** while minimizing **structural zeros** in population synthesis.
- **Feasibility** and **diversity** are used as improvement metrics for the generative model.
- The proposed loss functions achieve higher accuracy compared to traditional models.
- By improving the synthetic population, which is the first stage of ABM, the propagation of errors to later modeling stages is avoided.

1.Introduction

ABM & Population synthesis

- ABMの入力としては合成人口とその意思決定プロセスが含まれる
- この研究では，個別属性の結合分布を模倣することを目指す
- 旅行需要モデル（**activity-based model**）に焦点を当てる（個別の詳細な属性が必要である）ため，人口合成はthe regional household travel survey（HTS，人口の約1~5%）に依存する

- The inputs for ABM include the synthetic population and its decision-making process.
- This study aims to mimic the joint distribution of individual attributes.
- It focuses on the travel demand model (**activity-based model**), which requires detailed individual attributes, and thus the population synthesis relies on the regional household travel survey (HTS), covering about 1-5% of the population. (Castiglione et al., 2014)

ABM & Population synthesis

- 人口合成には3つの段階がある
    1. **個別属性の現実的で多様な組み合わせを表す合成プールを生成する**
    2. 将来のターゲットに対する代表的な合成人口を構築するために重み計数を推定
    3. 将来の人口合成を生成してABMに割り当てられる
- 従来の方法（re-weighting）ではHTSサンプルには存在しないが，実際の人口には存在する属性の組み合わせを生成できない
- **生成モデル(GM)**が属性の結合確率分布を学習することでこの制約を克服できる

- Population synthesis consists of three stages (Borysov et al., 2019; Rich, 2018):
    1. **Generating synthetic pools representing realistic and diverse combinations of individual attributes.**
    2. Estimating weighting factors to construct a representative synthetic population for future targets.
    3. Generating future synthetic populations and allocating them to the ABM.
- Traditional methods (re-weighting) cannot generate attribute combinations that do not exist in HTS samples but are present in the actual population.
- **Generative models (GMs)** can overcome this limitation by learning the joint probability distribution. (Farooq et al., 2013; Saadi et al., 2016; Sun et al., 2018; Sun and Erath, 2015)

The generated data from GM

- GMにより生成されたデータ(属性の組み合わせ)は4つのグループに分けられる

1. **general sample** : サンプルデータの中かつ生成されたデータの中にも存在する
2. **missing sample** : サンプルデータの中に存在するが生成されたデータの中には存在しない
3. **sampling zero** : 実際の人口には存在するが，小規模サンプルの中には存在しない
4. **structural zero** : 実際には存在しないため本来ないべきであるのに，生成されてしまう

- The data (attribute combinations) generated by GMs can be divided into four groups:

1. **general sample** : Exists both in the sample data and the generated data.
2. **missing sample** : Exists in the sample data but not in the generated data.
3. **sampling zero** : Exists in the actual population but not in the small-scale sample.
4. **structural zero** : Should not exist because it does not exist in the actual population, but is generated.

The generated data from GM

- 理想的な合成人口は，全てのsampling zeroとgeneral sampleを含み，missing sampleがなく，structural zeroを最小限に抑えること
- **sampling zeroを増やすとstructural zeroも増えるというトレードオフ**が存在するため，このような生成はほぼ不可能

- The ideal synthetic population would include all sampling zeros and general samples, with no missing samples, and minimize structural zeros.
- However, there is **a trade-off where increasing sampling zeros also increases structural zeros**, making such generation almost impossible.

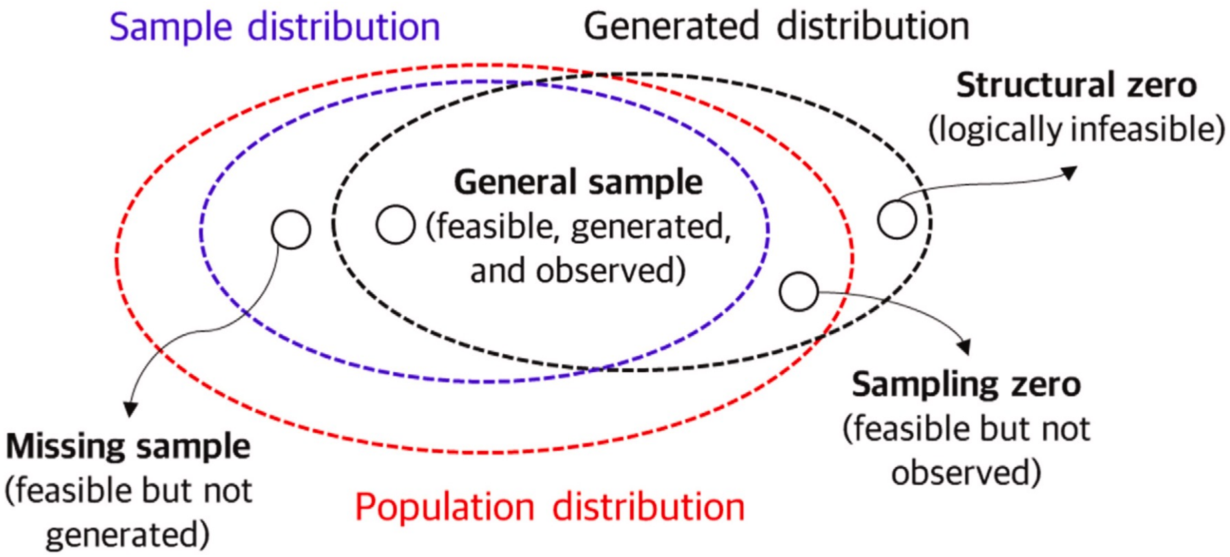


Fig. 1. Conceptual diagram of the general sample, missing sample, sampling zero, and structural zero.

Applying a deep generative model (DGMs)

- 最近ではVAEやGAN等の**DGM**で再現性高く属性を生成可能になったが，**過学習**(overfitting)の恐れ →sampling zeroの生成能力低下
- 性能評価のために2つの指標(**feasibility**と**diversity**)を導入(後述)
- **ミニバッチ内の生成されたデータ分布からトレーニングサンプル分布全体への距離**を測定し，それらを新たな損失関数としてDGMに採用することで前述のトレードオフを制御する

- Recently, deep generative models (**DGMs**) have been able to generate attributes with high reproducibility. However, there is a risk of **overfitting**, which can lead to a decrease in the ability to generate sampling zeros. (Kim et al., 2022)
- To evaluate performance, two metrics(**feasibility&diversity**) are introduced:
- By measuring **the distance from the generated data distribution in a mini-batch to the entire training sample distribution** and incorporating these distances into new loss functions, the trade-off mentioned above can be controlled.

Hypothetical population and sample

- **提案手法の検証には全人口の属性を必要とする**が，今回は100万人以上もの行動データの大規模サンプルを用いることでこの問題を回避
- 全サンプルデータを仮想人口（***h*-population**），その中の一部を仮想サンプル（***h*-sample**）とする
- これまではHTSのサンプル数の制限により，structural zeroの割合とsampling zeroの割合に偏りが生じていた可能性がある

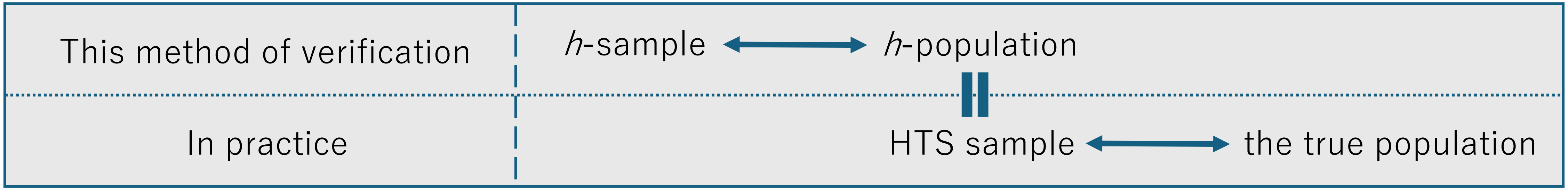
- **The validation of the proposed method typically requires attributes of the entire population**, but in this study, we circumvent this requirement by using a large-scale sample (>1 million) of travel behavior data.
- We treat the entire sample data as a hypothetical population (***h*-population**) and a portion of it as a hypothetical sample (***h*-sample**) for training.
- Previous studies have been limited by the sample size of HTS data, potentially leading to biases in the of structural zeros and sampling zeros rates.

2.Data acquisition

2-1 Data preparation and assumptions

- 2010年、2016年、2021年に韓国で行われたHTSデータを組み合わせて作成された個人属性を持つ大規模なサンプルデータを使用することで、前述の全人口データの入手不可能性の問題を回避
- 全データセットを*h*-populationとし，そのうち5%を*h*-sampleとした

- The issue of unavailability of the entire population data is circumvented by using a large-scale sample data with individual-level attributes created by combining HTS data from South Korea in 2010, 2016, and 2021.
- The entire dataset is treated as the h-population, with only 5% used as the h-sample.



This method of verification

h-sample ⟷ *h*-population

‖

In practice

HTS sample ⟷ the true population

2-1 Data preparation and assumptions

- 移動手段や出発時刻は個人の一般的な移動パターンを表している
- カテゴリ属性のみから構成されるデータは,有限な属性の組み合わせでデータ分布を表現することが可能

- “The major travel mode of regular travel and the major departure time of regular travel represent an individual's general travel patterns.
- The data consisting only of categorical attributes can represent the data distribution with a finite number of attribute combinations.

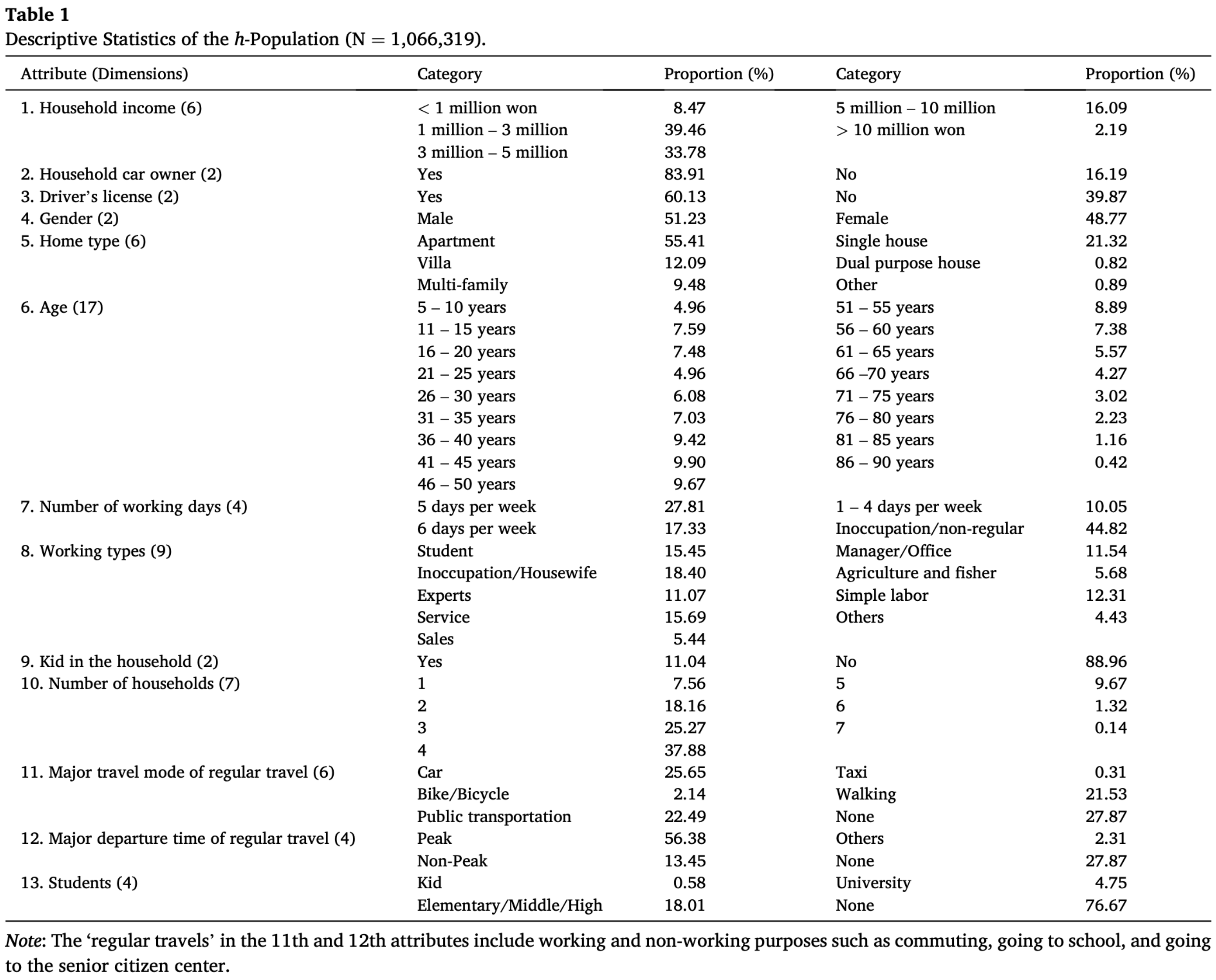


Table 1
Descriptive Statistics of the *h*-Population (N = 1,066,319).

Attribute (Dimensions)	Category	Proportion (%)	Category	Proportion (%)
1. Household income (6)	< 1 million won	8.47	5 million – 10 million	16.09
	1 million – 3 million	39.46	> 10 million won	2.19
	3 million – 5 million	33.78		
2. Household car owner (2)	Yes	83.91	No	16.19
3. Driver’s license (2)	Yes	60.13	No	39.87
4. Gender (2)	Male	51.23	Female	48.77
5. Home type (6)	Apartment	55.41	Single house	21.32
	Villa	12.09	Dual purpose house	0.82
	Multi-family	9.48	Other	0.89
6. Age (17)	5 – 10 years	4.96	51 – 55 years	8.89
	11 – 15 years	7.59	56 – 60 years	7.38
	16 – 20 years	7.48	61 – 65 years	5.57
	21 – 25 years	4.96	66 –70 years	4.27
	26 – 30 years	6.08	71 – 75 years	3.02
	31 – 35 years	7.03	76 – 80 years	2.23
	36 – 40 years	9.42	81 – 85 years	1.16
	41 – 45 years	9.90	86 – 90 years	0.42
	46 – 50 years	9.67		
7. Number of working days (4)	5 days per week	27.81	1 – 4 days per week	10.05
	6 days per week	17.33	Inoccupation/non-regular	44.82
8. Working types (9)	Student	15.45	Manager/Office	11.54
	Inoccupation/Housewife	18.40	Agriculture and fisher	5.68
	Experts	11.07	Simple labor	12.31
	Service	15.69	Others	4.43
	Sales	5.44		
9. Kid in the household (2)	Yes	11.04	No	88.96
10. Number of households (7)	1	7.56	5	9.67
	2	18.16	6	1.32
	3	25.27	7	0.14
	4	37.88		
11. Major travel mode of regular travel (6)	Car	25.65	Taxi	0.31
	Bike/Bicycle	2.14	Walking	21.53
	Public transportation	22.49	None	27.87
12. Major departure time of regular travel (4)	Peak	56.38	Others	2.31
	Non-Peak	13.45	None	27.87
13. Students (4)	Kid	0.58	University	4.75
	Elementary/Middle/High	18.01	None	76.67

Note: The ‘regular travels’ in the 11th and 12th attributes include working and non-working purposes such as commuting, going to school, and going to the senior citizen center.

2-2 Sampling zeros according to a sampling rate

- *h*-populationの属性の組み合わせの種類は264,005であるのに対し，*h*-sampleからは30,837 (11.7%)しか抽出されなかった(図中の黄色い線)
- 赤い線は，属性の組み合わせを観測数に応じて重み付けしたときに，*h*-sampleがどれだけ全体の組み合わせをカバーできているかということを表している

- The h-population has 264,005 unique combinations of attributes, but only 30,837 (11.7%) were extracted from the h-sample (shown by the yellow line in the figure).
- The red line represents how much of the total combinations the h-sample covers when the combinations are weighted according to the number of observations.

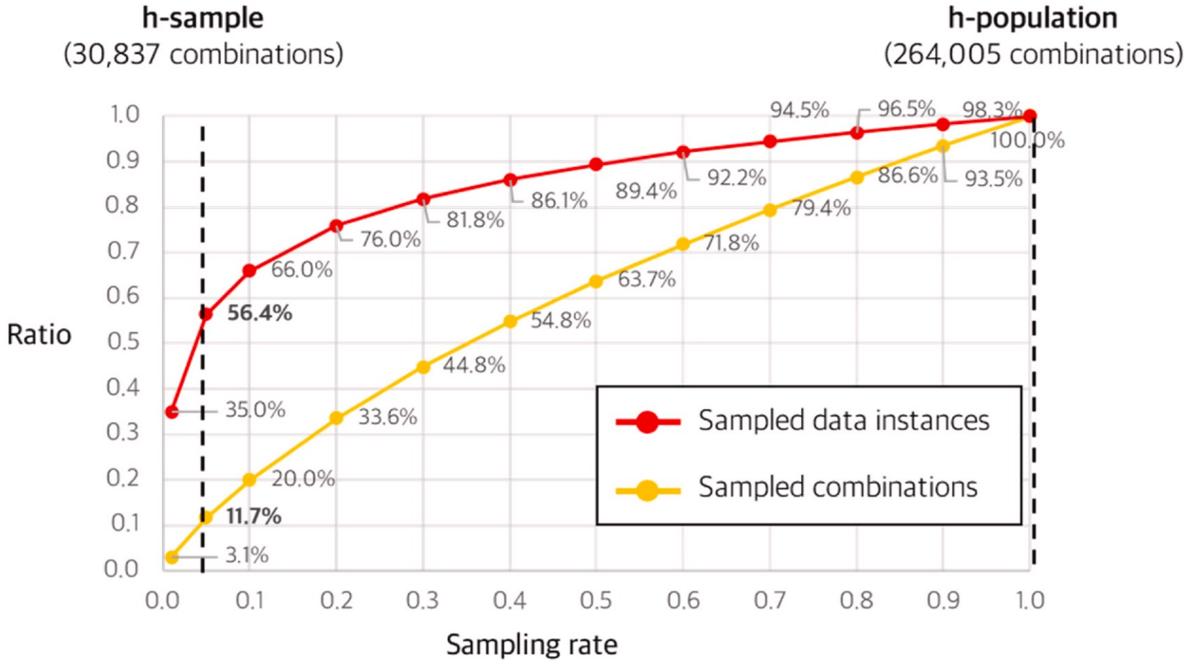


Fig. 2. The degree of sampling zeros according to the sampling rate.

3.Methodologies

3-1 Hypothesis for sampling and structural zero

Hypothesis

> sampling zeroは，structural zeroよりも**全体のトレーニングサンプル分布の境界に近い**場所に位置する可能性が高い
>
> The sampling zeros are more likely to be located nearer the boundary of the entire training sample distribution than the structural zeros.

- sampling zeroとstructual zeroは確率密度が非常に低い(ミニバッチ内のサンプル分布はカバーできていない)疎結合領域に位置するため，全体のトレーニングサンプル分布を使用する必要がある
- 一部のsampling zeroは，structural zeroよりもサンプル境界から遠い場合がある
  - 理由1 : サンプルの確率密度の形状が人口と異なる可能性がある
  - 理由2 : 離散空間内の距離は属性間の関係を考慮していない

- Sampling zeros and structural zeros are located in the loosely connected areas where the probability density is very low (not covered by the sample distribution in the mini-batch), so it is necessary to use the entire training sample distribution.
- Some sampling zeros may be farther from the sample boundary than structural zeros.
  - Reason 1: The shape of the probability density of the sample may differ from that of the population.
  - Reason 2: The distance in the discrete space does not take into account the contextual relationship between attributes.

3-1 Hypothesis for sampling and structural zero

Hypothesis

The sampling zeros are more likely to be located nearer the boundary of the entire training sample distribution than the structural zeros.

Introduce of the embeddings to transform high-dimensional discrete data into a lower-dimensional continuous vector

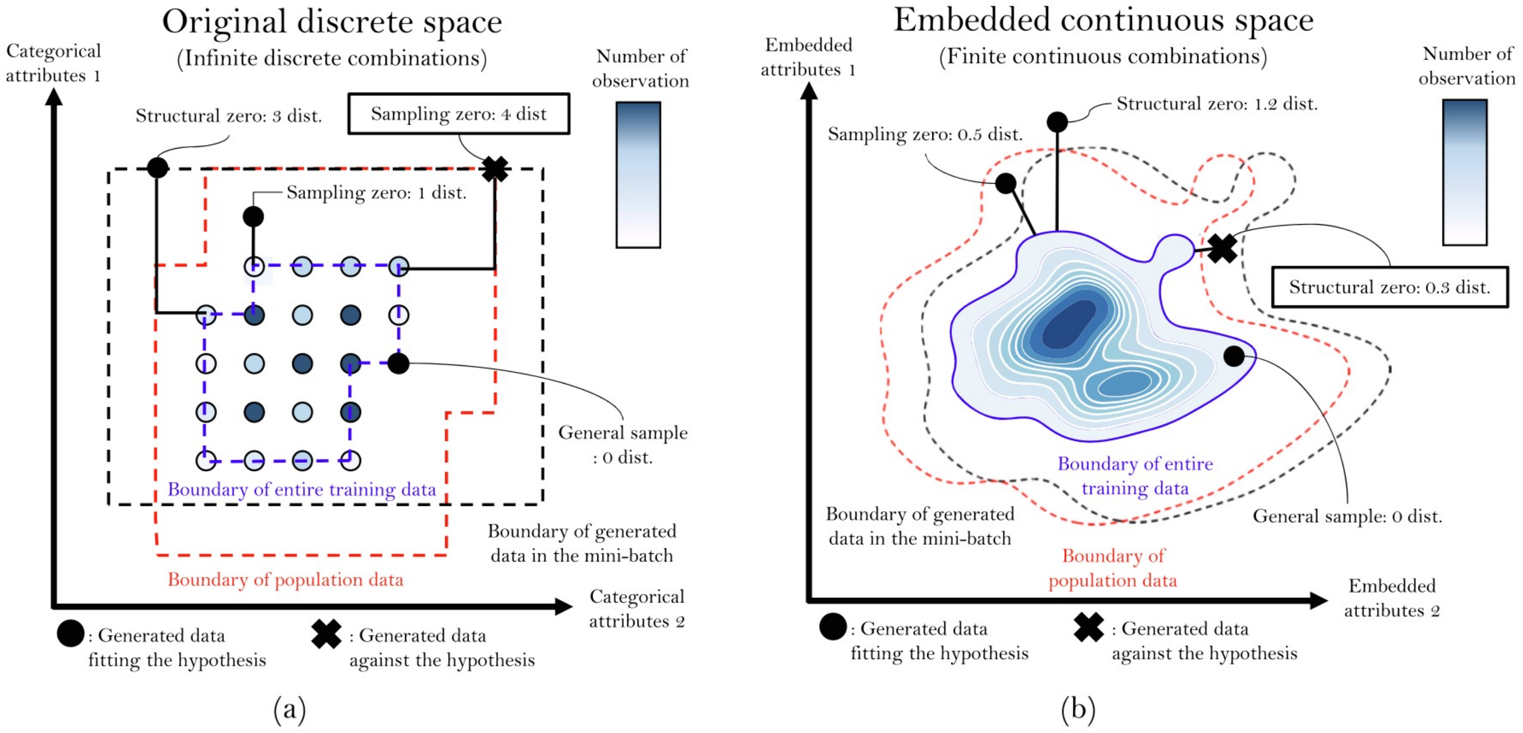

Original discrete space
(Infinite discrete combinations)
Categorical attributes 1
Number of observation
Structural zero: 3 dist.
Sampling zero: 4 dist
Sampling zero: 1 dist.
General sample : 0 dist.
Boundary of entire training data
Boundary of generated data in the mini-batch
Boundary of population data
Categorical attributes 2
● : Generated data fitting the hypothesis
✖ : Generated data against the hypothesis
(a)
Embedded continuous space
(Finite continuous combinations)
Embedded attributes 1
Number of observation
Structural zero: 1.2 dist.
Sampling zero: 0.5 dist.
Structural zero: 0.3 dist.
Boundary of entire training data
Boundary of generated data in the mini-batch
General sample: 0 dist.
Boundary of population data
Embedded attributes 2
● : Generated data fitting the hypothesis
✖ : Generated data against the hypothesis
(b)

Fig. 3. Hypothesis for sampling and structural zero in the (a) original discrete space and (b) embedded continuous space.

3-1 Hypothesis for sampling and structural zero

- 自己教師あり埋め込みネットワークを使用して，マルチカテゴリ属性を数値属性に変換
- 埋め込みネットワークは，不完全な属性セットが与えられたときに，完全な個人属性セットを出力するように訓練
- マスクされた属性を補完する際に，カテゴリ属性間の文脈的関係が学習される

- We use self-supervised embedding networks to transform multi-categorical attributes into numerical attributes.
- The embedding network is trained to output a complete set of individual attributes given an incomplete set of attributes.
- In filling the masked attributes, the contextual relationships among categorical attributes are learned

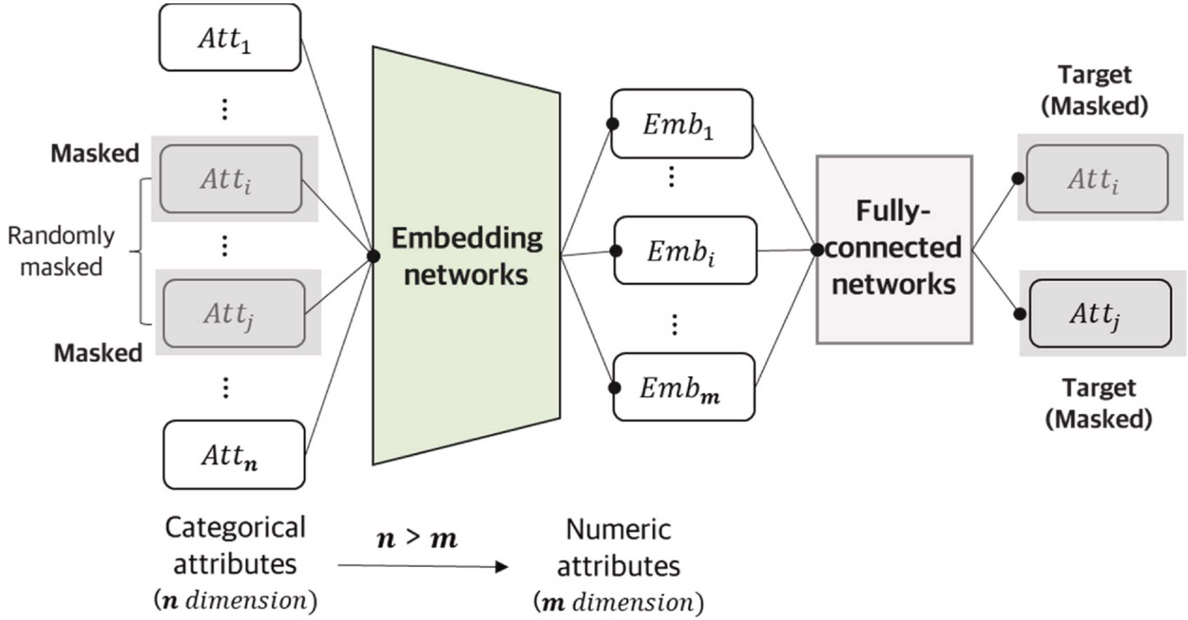


Fig. 4. Model structure of self-supervised embedding networks.

3-1 Hypothesis for sampling and structural zero

- 2つのデータベクトル (X_i, X_j)間の距離は、下の式のように2つのベクトル間のユークリッド距離によって測定される
- 埋め込み空間では2つのデータポイントは連続ベクトルとして表現されるため直接距離の測定可能

- The distance between the two data vectors (X_i, X_j) is measured by the Euclidean distance between the two vectors as shown in the equation below.
- In the embedding space, the two data points are represented as continuous vectors, making it possible to directly measure the Euclidean distance.

the Euclidean distance between two vectors

$$\boxed{\mathrm{Dist}(X_i, X_j) = \sqrt{(X_i - X_j)^2}}$$

3-2 Deep Generative models (DGMs)

- DGMは、サンプルの属性 X^Sを使用して、母集団の属性 $P(X)$の結合確率分布を近似
- 母集団合成に対するGANとVAEの優れたパフォーマンスを考慮し、これらを2つの主要なモデルとして採用

- The DGM approximates the joint probability distribution of the population’s attributes, $P(X)$, using the sample’s attributes, X^S.
- Considering the superior performance of a GAN and a VAE for population synthesis, we adopt them as the two main models.

3-2 Deep Generative models (DGMs)

Generative adversarial network (GAN)

- The generator, $G(z;\ \phi_g)$, parameterized by ϕ_g, estimates the $P_\phi(\hat{X})$, with a z sampled from a prior random distribution, $P(z)$.
- The parameterized discriminator, $D(X; \phi_d)$, outputs whether the attribute combination is the real data (X) or the generated data $(\hat{X} = G(z))$

The value function to estimate $\boldsymbol{\phi_g}$ and $\boldsymbol{\phi_d}$

$$\boxed{\min_{\phi_g}\max_{\phi_d} E_{X\ \ P(X)}[log D(X)] + E_{z\ \ P(z)}[\log(1 - D(G(z)))].}$$

- The D is trained to maximize the value function by imposing the higher $D(X)$ and lower $D(G(z))$.
- The G is trained to minimize $\log(1 - D(G(z)))$ by deceiving D with realistic data.

3-2 Deep Generative models (DGMs)

Generative adversarial network (GAN)

- The generated distribution $P_\phi(\hat{X})$: **discrete** on a K-dimensional set of simplex
- The real distribution $P(X)$: **continuous** over this simplex set

→ instability and saturation of equilibrium

→ employ a Wasserstein GAN (WGAN) with a gradient penalty (GP)

The loss function

$\mathcal{L}_{\mathrm{d}} = \frac{1}{m}\sum_{i=1}^{m} - D(X_i) + D(G(z_i)).$ → maximizing the value function

$\mathcal{L}_{g} = \frac{1}{m}\sum_{i=1}^{m} - D(G(z_i)).$ → minimizing $\log(1 - D(G(z)))$ to train the G

$\mathcal{L}_{GP} = \frac{1}{m}\sum_{i=1}^{m} \lambda\left(\|\nabla_{\widetilde{X}_i} D\left(\widetilde{X}_i\right)\|_2 - 1\right)^2.$ → regularizes it based on D's gradients

$\widetilde{X}_i = \alpha\widehat{X}_i + (1-\alpha)X_i,\ \alpha\ \mathrm{Uniform}[0,1].$ → a weighted average of the generated and real data

3-2 Deep Generative models（DGMs）

Variational autoencoder (VAE)

- The encoder $Q(X;\ \theta_e)$ maps the input data (X) into the parameters of latent prior distribution $P(z)$.
- The decoder $R(z, \theta_d)$ generates the data $(\hat{X})$ mimicking X from the z sampled from $P(z)$.
- The decoder parameter θ_d is estimated together with the encoder parameter θ_e based on a reparameterization trick.

The loss function

$\mathscr{L}_R = \frac{1}{m}\sum_{i=1}^{m} X_i log\widehat{X}_i.$ → minimizing the discrepancy between the input and generated data

$\mathscr{L}_{KL} = -\beta D_{KL}[Q(\boldsymbol{X})||P(\boldsymbol{z})].$ → minimaizing the Kullback-Leibler (KL) divergence between the prior $P(z)$ and estimated latent distribution that is the output of the encoder

3-2 Deep Generative models（DGMs）

A summay of GAN and VAE

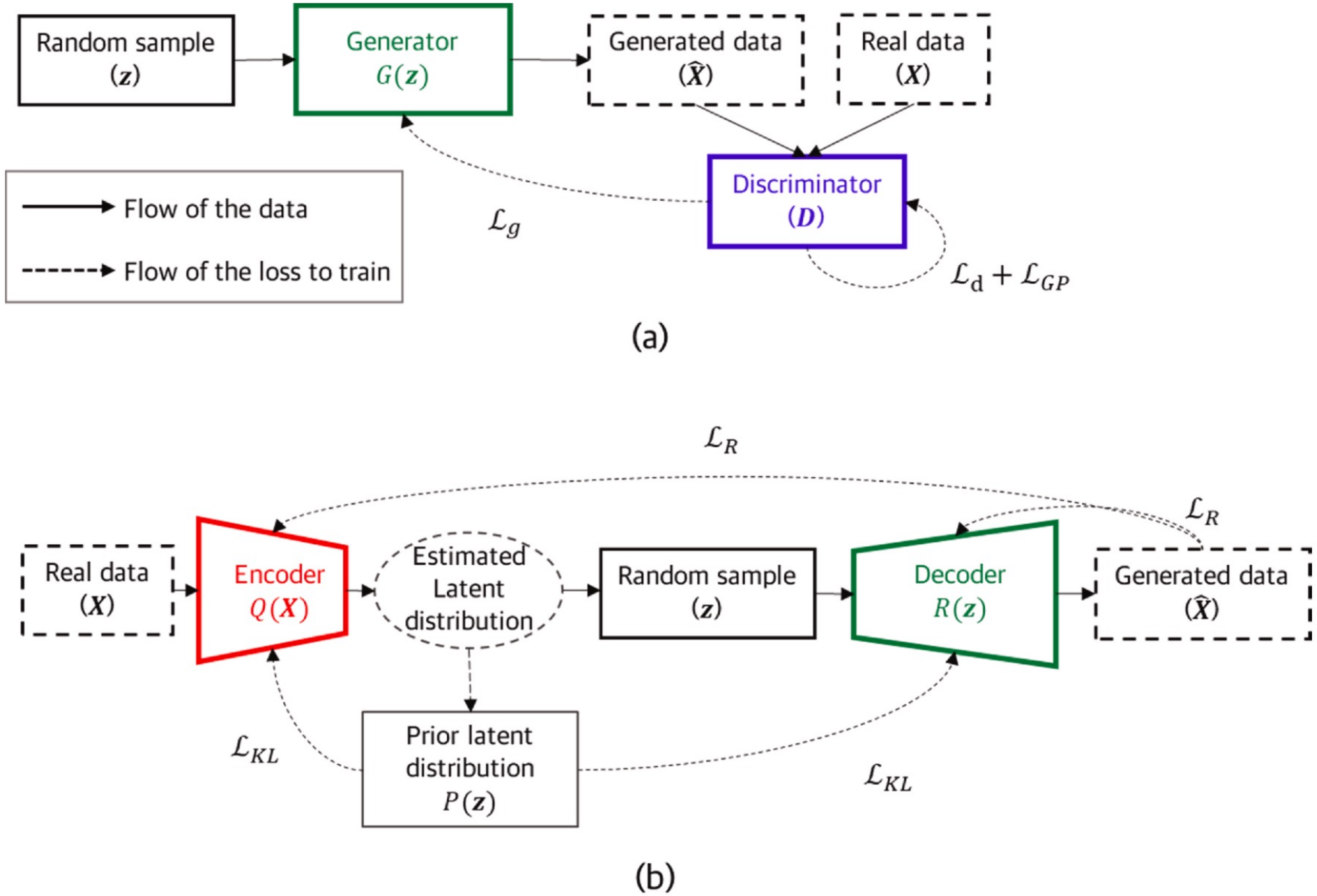


Fig. 5. The training procedure of (a) GAN and (b) VAE.

3-3 Novel loss functions for regularization

Distance to the boundary of the sample distribution

- 全体のトレーニングサンプル分布の境界までの距離が，離散空間でも埋め込み空間でもsampling zeroとstructural zeroを区別するための良い指標であることを示している．
- WGANによって生成されたsampling zeroとstructural zeroは，VAEによって生成されたものより明確に区別される傾向がある．

- The result shows that the distance to the boundary of the entire training sample distribution is a good indicator to distinguish the sampling and structural zeros in discrete and embedding spaces.
- The sampling and structural zeros generated by WGAN tend to be more clearly distinct than those generated by VAE.

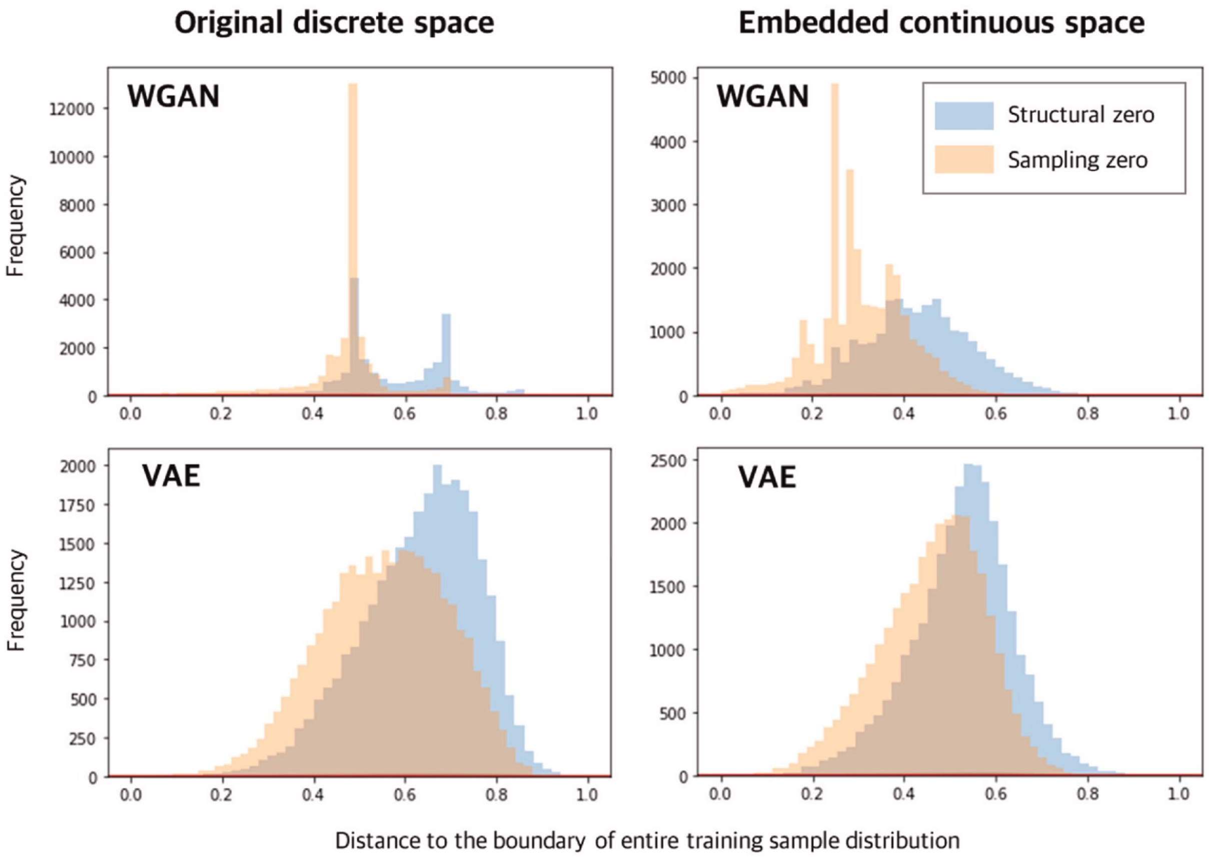


Fig. 6. Histograms of distance from sampling zeros and structural zeros to the boundary of the entire training sample distribution.

3-3 Novel loss functions for regularization

Regularizations based on the distance to the sample distribution

The ‘boundary distance regularization’ ($\boldsymbol{R_{BD}}$)

$$R_{BD}\left(\widehat{\boldsymbol{X}},\boldsymbol{X}^{S}\right)=\frac{1}{M}\sum_{j=1}^{M}\min_{i\in\{1:N\},j\in\{1:M\}}\left(Dist\left(\widehat{X}_j,X_i^{S}\right)\right)$$

→ discouraging the generation of structural zero far from the sample boundary

- R_{BD}はミニバッチ内の各生成データからN個の全トレーニングデータまでの**最短距離**を計算し，M個の生成データに対して平均を取ったもの．
- 各ミニバッチの訓練において全体のトレーニングサンプル分布を考慮して，生成データが訓練データの分布内に留まるようにしている．

- The R_{BD} calculates **the nearest distance** from each generated data in the mini-batch to N entire training data and averages them for M generated data.
- Considering the entire training sample distribution in each mini-batch training

3-3 Novel loss functions for regularization

Regularizations based on the distance to the sample distribution

The ‘average distance regularization’ ($\boldsymbol{R_{AD}}$)

$$R_{AD} = -\frac{1}{NM}\sum_{j=1}^{M}\sum_{i=1}^{N} Dist\left(\widehat{X}_j, X_i^S\right)$$

→ encouraging the generation of missing samples and sampling zero

- R_{AD}はM個の生成データの「全体のトレーニングサンプル分布への**平均距離**」の平均を計算して，生成データが訓練データ全体の分布に対して適切に散らばるようにしている．
- missing sampleとsampling zeroは全体のトレーニングサンプル分布に対する平均距離が長い傾向

- The R_{AD} computes the average for the ‘**average distance** to the entire training sample distribution’ of the M generated data.
- The missing sample and sampling zero tend to have a long average distance to the entire training sample distribution.

3-3 Novel loss functions for regularization

Regularizations based on the distance to the sample distribution

- R_{BD}は基本的にstructural zeroを減らすが，sampling zeroも誤って除去する可能性がある.
- 実際，missing sampleとsampling zeroは全体のトレーニングサンプル分布に対する平均距離が長い

- The R_{BD} is expected to reduce the structural zero but can also falsely remove the sampling zero.
- The missing sample and sampling zero tend to have a long average distance to the entire training sample distribution.

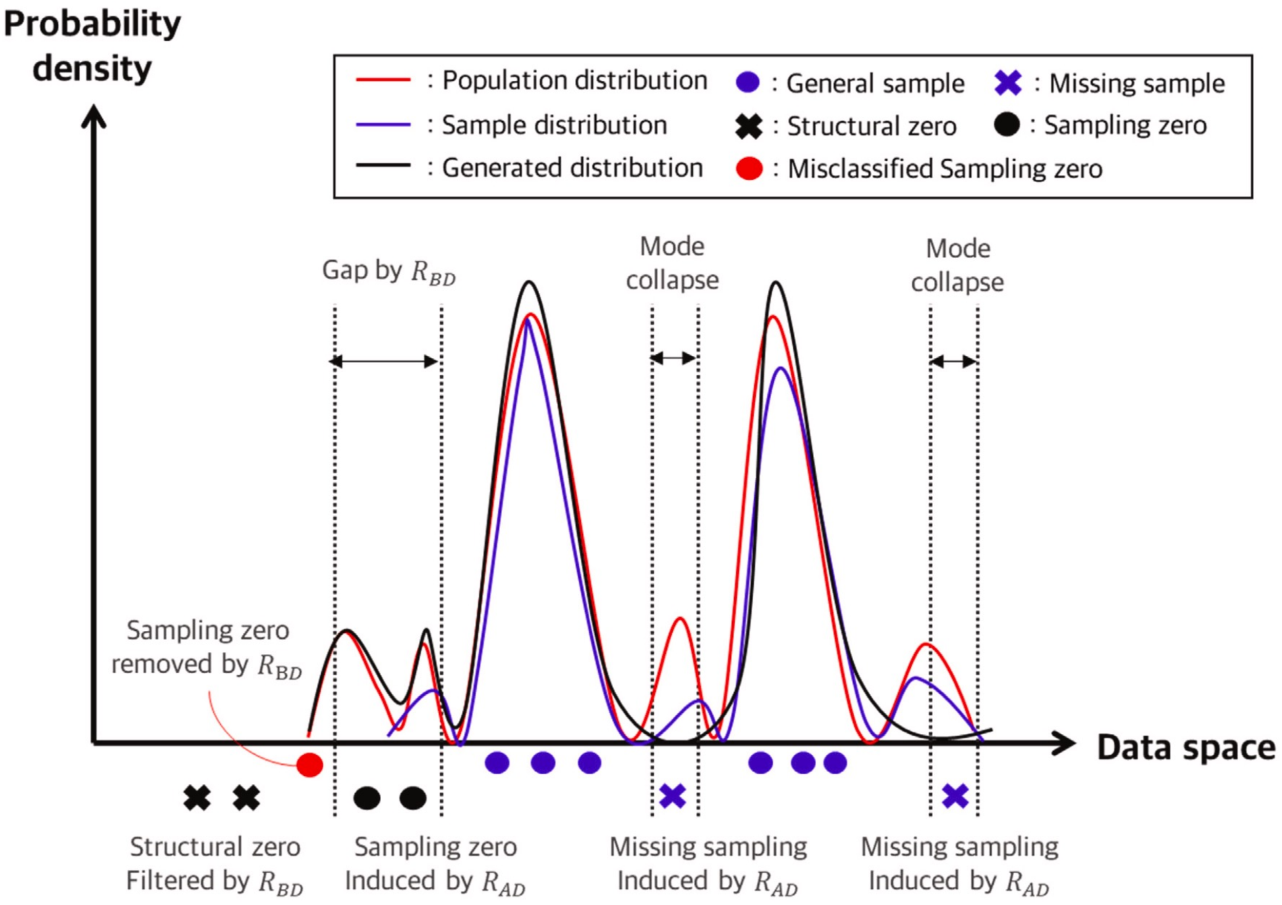

Probability density
: Population distribution
: Sample distribution
: Generated distribution
● : General sample
✖ : Structural zero
● : Misclassified Sampling zero
✖ : Missing sample
● : Sampling zero
Gap by R_{BD}
Mode collapse
Mode collapse
Sampling zero removed by R_{BD}
Data space
Structural zero Filtered by R_{BD}
Sampling zero Induced by R_{AD}
Missing sampling Induced by R_{AD}
Missing sampling Induced by R_{AD}

Fig. 7. The generated data considered by R_{BD} and R_{AD} on the simplified data space.

3-3 Novel loss functions for regularization

Variants of the DGM

The final loss functions of WGAN and VAE adopting $\boldsymbol{R_{BD}}$ and $\boldsymbol{R_{AD}}$

$$\mathscr{L}_{\mathrm{WGAN}} = \mathscr{L}_{\mathrm{d}} + \mathscr{L}_{\mathrm{g}} + \mathscr{L}_{\mathrm{GP}} + \gamma_{BD}^{GAN} R_{BD} + \gamma_{AD}^{GAN} R_{AD}$$

$$\mathscr{L}_{\mathrm{VAE}} = \mathscr{L}_{R} + \mathscr{L}_{KL} + \gamma_{BD}^{VAE} R_{BD} + \gamma_{AD}^{VAE} R_{AD}$$

$\gamma_{BD}^{GAN}, \gamma_{AD}^{GAN}, \gamma_{BD}^{VAE}, \gamma_{AD}^{VAE}$: the regularization weights

- R_{BD} : discouraging the generation of structural zero far from the sample boundary
- R_{AD} : encouraging the generation of missing samples and sampling zero

3-4 Evaluation metrics

Distributional similarity

Standardized root mean square error (SRMSE)

$$\mathrm{SRMSE}(\boldsymbol{\pi},\widehat{\boldsymbol{\pi}})=\frac{\mathrm{RMSE}(\boldsymbol{\pi},\widehat{\boldsymbol{\pi}})}{\overline{\boldsymbol{\pi}}}=\frac{\sqrt{\sum_{(k,k')}\left(\boldsymbol{\pi}_{(k,k')}-\widehat{\boldsymbol{\pi}}_{(k,k')}\right)^2\Big/N_b}}{\sum_{(k,k')}\boldsymbol{\pi}_{(k,k')}\Big/N_b},$$

→ evaluating the distributional similarity of marginal and bivariate distribution

$\boldsymbol{\pi}$: categorical distributions of *the h*-population
$\widehat{\boldsymbol{\pi}}$: categorical distributions of the generated data
N_b : the total number of category combinations

- 過学習の評価は難しいが，低次元分布を比較することで計算を簡単にしている．

- Evaluating overfitting is challenging.
- Comparing low-dimensional distributions helps simplifying the computation.

3-4 Evaluation metrics

Feasibility and diversity

Feasibility : 生成されたデータが元のデータとどれほど似ているか，元のデータをどれだけ再現できているかを表す指標． / Indicating how well the generated data resembles the population data.

The metrics to evaluate the feasibility

$$\text{Precision} = \frac{1}{M}\sum_{j=1}^{M} 1_{\hat{x}_j \in X}$$ → measuring the ratio of generated data included in the feasible *h*-population, that is one minus structural zero rates

Diversity : 生成データが元データのバリエーションをどれだけ捉えているかを表す指標．過学習の程度を評価． / Refering to the degree to which the generated data captures the population variations

The metrics to evaluate the diversity

$$\text{Recall} = \frac{1}{N}\sum_{i=1}^{N} 1_{x_i \in \hat{X}}$$ → measuring the ratio of *h*-population’s combinations included in the generated data proportional to the sampling zero

3-4 Evaluation metrics

Feasibility and diversity

Trade-off relationship : sampling zeroを増やすとstructural zeroも増える(precisionとrecallもトレードオフ)
/ Precision and recall have a trade-off relationship.

The metrics indicating the overall quality

$$\boxed{\text{F1 score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}}$$

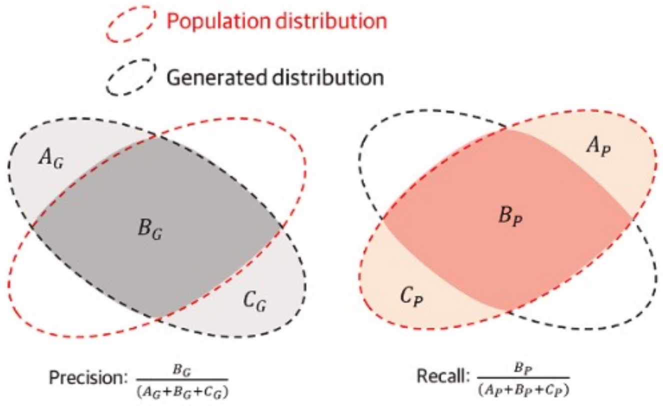


Fig. 8. Conceptual diagram of precision and recall for evaluating generative model.

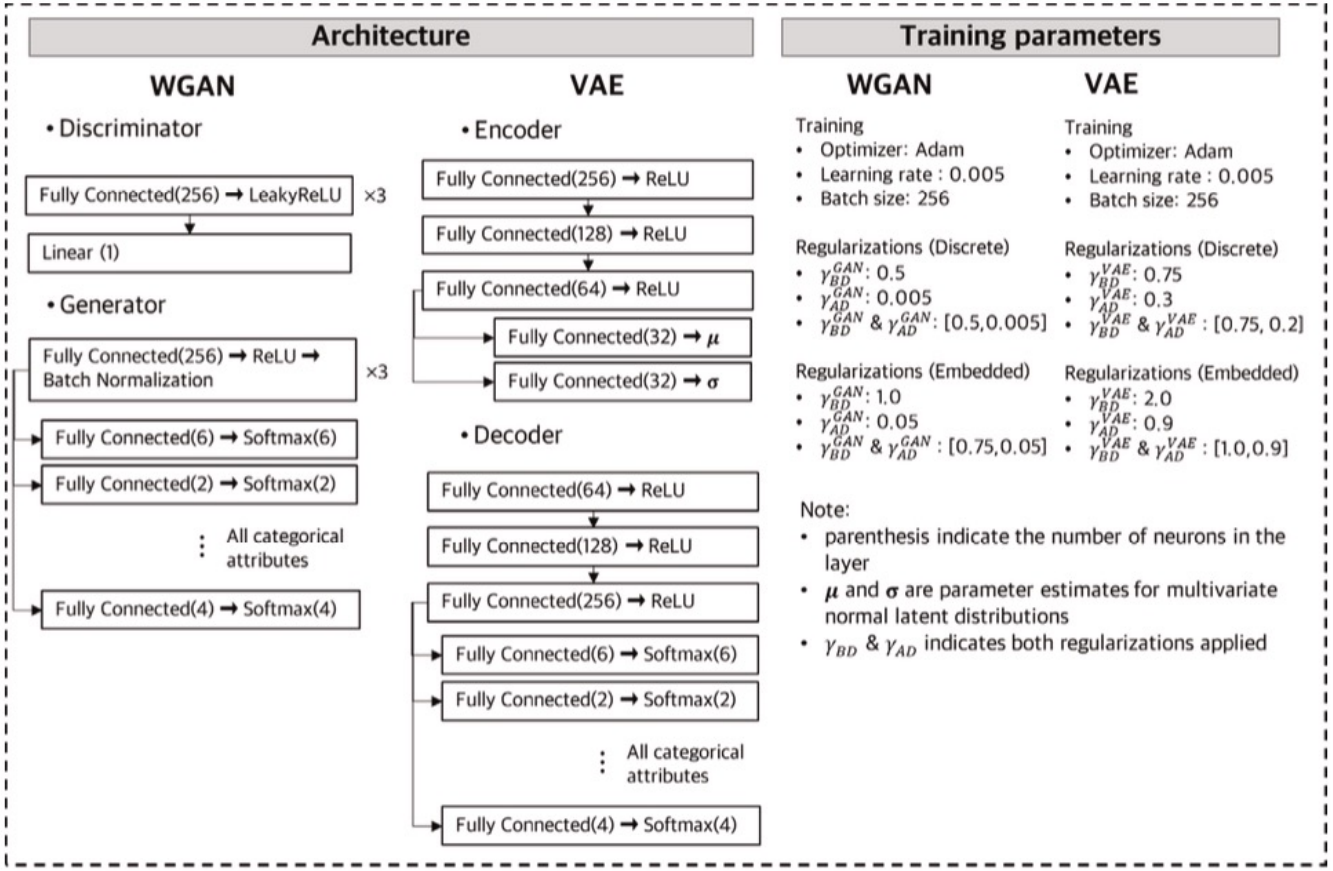
4.Empirical application

4-1 Model evaluation results

Evaluated models

- The prototypical agent approach : re-weighting
- BN : a GM decomposing the joint distribution into a set of partial conditional distributions to learn the data-generating process efficiently
- The vanilla DGM
- The proposed DGM : incorporating R_{BD} and R_{AD} into the loss function

DGM generates the data with the h-population's size.



Architecture

WGAN

- Discriminator
  - Fully Connected(256) → LeakyReLU ×3
  - Linear (1)
- Generator
  - Fully Connected(256) → ReLU → Batch Normalization ×3
  - Fully Connected(6) → Softmax(6)
  - Fully Connected(2) → Softmax(2)
  - ⋮ All categorical attributes
  - Fully Connected(4) → Softmax(4)

VAE

- Encoder
  - Fully Connected(256) → ReLU
  - Fully Connected(128) → ReLU
  - Fully Connected(64) → ReLU
  - Fully Connected(32) → $\boldsymbol{\mu}$
  - Fully Connected(32) → $\boldsymbol{\sigma}$
- Decoder
  - Fully Connected(64) → ReLU
  - Fully Connected(128) → ReLU
  - Fully Connected(256) → ReLU
  - Fully Connected(6) → Softmax(6)
  - Fully Connected(2) → Softmax(2)
  - ⋮ All categorical attributes
  - Fully Connected(4) → Softmax(4)

Training parameters

WGAN

Training
- Optimizer: Adam
- Learning rate : 0.005
- Batch size: 256

Regularizations (Discrete)
- γ_{BD}^{GAN}: 0.5
- γ_{AD}^{GAN}: 0.005
- γ_{BD}^{GAN} & γ_{AD}^{GAN}: [0.5,0.005]

Regularizations (Embedded)
- γ_{BD}^{GAN}: 1.0
- γ_{AD}^{GAN}: 0.05
- γ_{BD}^{GAN} & γ_{AD}^{GAN} : [0.75,0.05]

VAE

Training
- Optimizer: Adam
- Learning rate : 0.005
- Batch size: 256

Regularizations (Discrete)
- γ_{BD}^{VAE}: 0.75
- γ_{AD}^{VAE}: 0.3
- γ_{BD}^{VAE} & γ_{AD}^{VAE} : [0.75, 0.2]

Regularizations (Embedded)
- γ_{BD}^{VAE}: 2.0
- γ_{AD}^{VAE}: 0.9
- γ_{BD}^{VAE} & γ_{AD}^{VAE} : [1.0,0.9]

Note:
- parenthesis indicate the number of neurons in the layer
- $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are parameter estimates for multivariate normal latent distributions
- γ_{BD} & γ_{AD} indicates both regularizations applied

Fig. 9. Calibrated hyperparameters of WGAN and VAE.

4-1 Model evaluation results

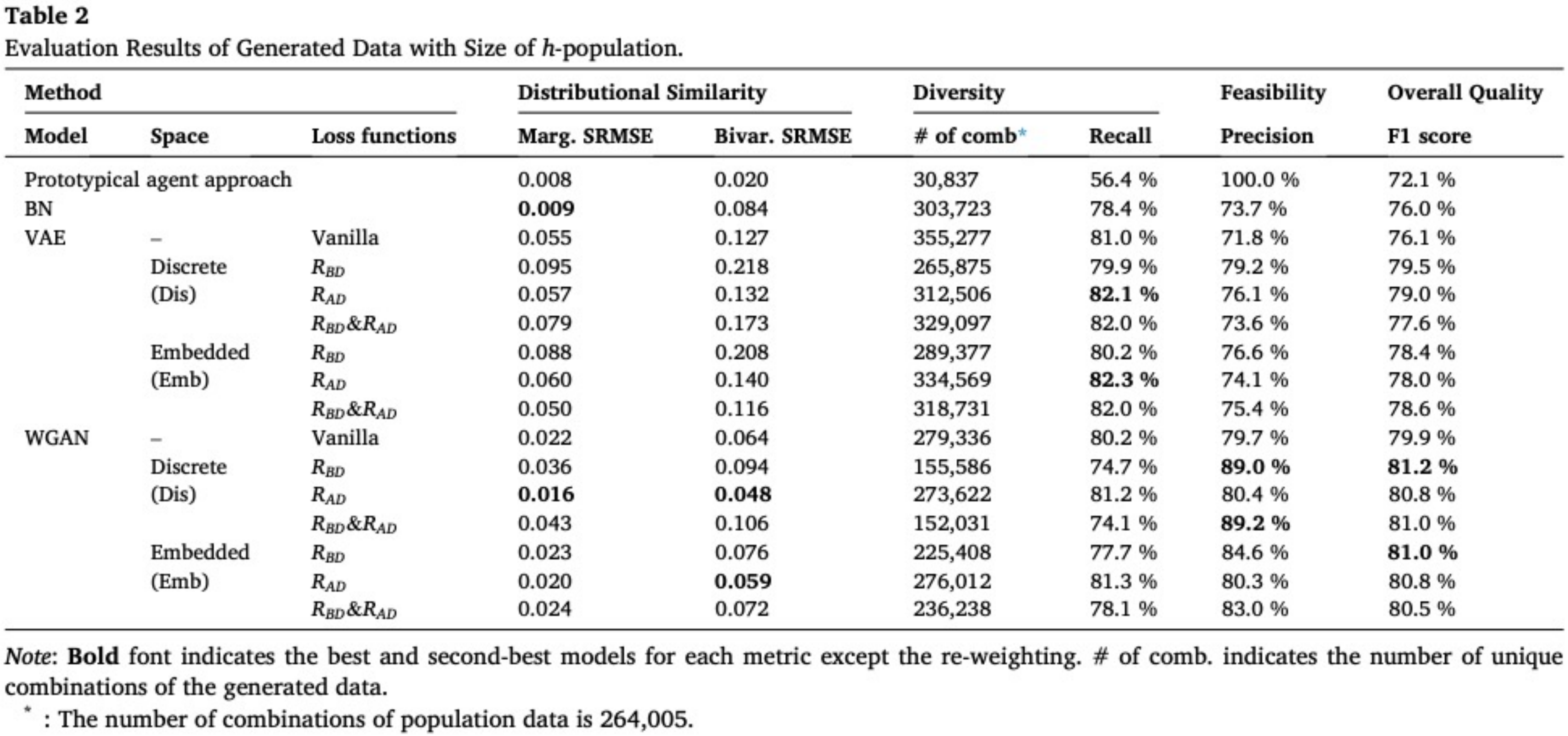


Table 2
Evaluation Results of Generated Data with Size of *h*-population.

Method			Distributional Similarity		Diversity		Feasibility	Overall Quality
Model	Space	Loss functions	Marg. SRMSE	Bivar. SRMSE	# of comb*	Recall	Precision	F1 score
Prototypical agent approach			0.008	0.020	30,837	56.4 %	100.0 %	72.1 %
BN			**0.009**	0.084	303,723	78.4 %	73.7 %	76.0 %
VAE	–	Vanilla	0.055	0.127	355,277	81.0 %	71.8 %	76.1 %
	Discrete (Dis)	R_{BD}	0.095	0.218	265,875	79.9 %	79.2 %	79.5 %
		R_{AD}	0.057	0.132	312,506	**82.1 %**	76.1 %	79.0 %
		R_{BD}&R_{AD}	0.079	0.173	329,097	82.0 %	73.6 %	77.6 %
	Embedded (Emb)	R_{BD}	0.088	0.208	289,377	80.2 %	76.6 %	78.4 %
		R_{AD}	0.060	0.140	334,569	**82.3 %**	74.1 %	78.0 %
		R_{BD}&R_{AD}	0.050	0.116	318,731	82.0 %	75.4 %	78.6 %
WGAN	–	Vanilla	0.022	0.064	279,336	80.2 %	79.7 %	79.9 %
	Discrete (Dis)	R_{BD}	0.036	0.094	155,586	74.7 %	**89.0 %**	**81.2 %**
		R_{AD}	**0.016**	**0.048**	273,622	81.2 %	80.4 %	80.8 %
		R_{BD}&R_{AD}	0.043	0.106	152,031	74.1 %	**89.2 %**	81.0 %
	Embedded (Emb)	R_{BD}	0.023	0.076	225,408	77.7 %	84.6 %	**81.0 %**
		R_{AD}	0.020	**0.059**	276,012	81.3 %	80.3 %	80.8 %
		R_{BD}&R_{AD}	0.024	0.072	236,238	78.1 %	83.0 %	80.5 %

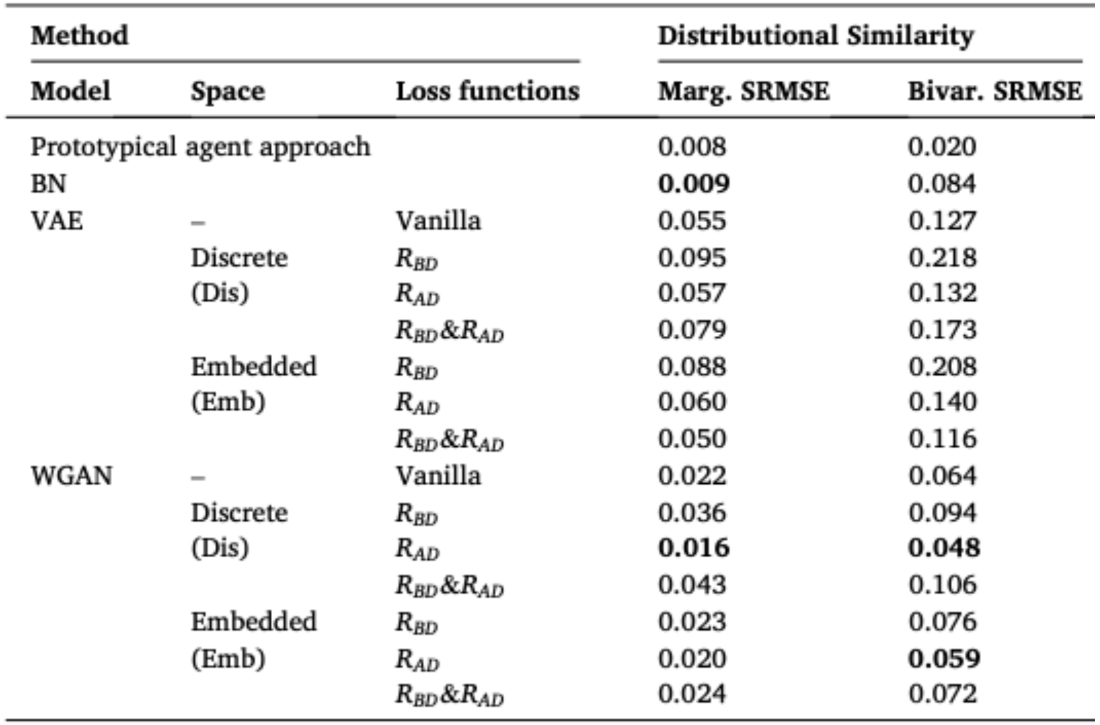
Note: **Bold** font indicates the best and second-best models for each metric except the re-weighting. # of comb. indicates the number of unique combinations of the generated data.

* : The number of combinations of population data is 264,005.

4-1 Model evaluation results

Distributional similarity

- The BN shows the best performance for marginal SRMSE among GMs (even close to re-weighting), while the ‘WGAN-Dis-R_{AD}’ is the best for bivariate distribution.
- All the WGAN variants outperform their VAE counterparts.
- The R_{AD} loss function in discrete and embedded spaces increases the distributional similarity of vanilla WGAN but decreases those of VAE.
- The R_{BD} loss function increases the SRMSE for all cases, indicating that the SRMSE cannot reflect aspects related to the structural zero.

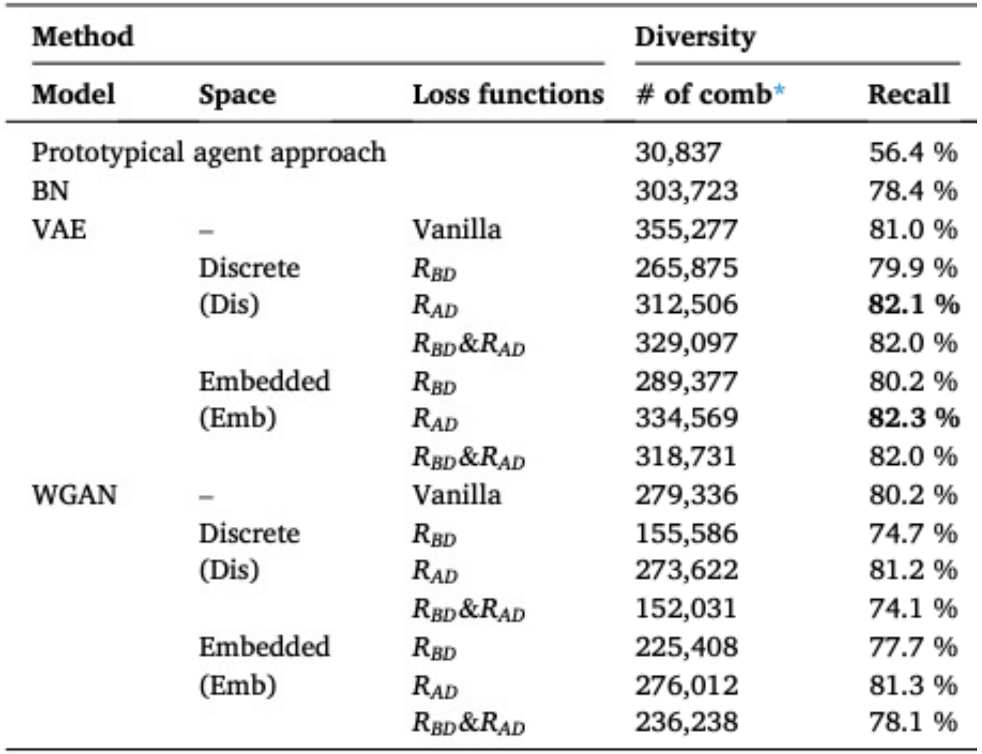


Method			**Distributional Similarity**	
Model	**Space**	**Loss functions**	**Marg. SRMSE**	**Bivar. SRMSE**
Prototypical agent approach			0.008	0.020
BN			**0.009**	0.084
VAE	–	Vanilla	0.055	0.127
	Discrete (Dis)	R_{BD}	0.095	0.218
		R_{AD}	0.057	0.132
		R_{BD}&R_{AD}	0.079	0.173
	Embedded (Emb)	R_{BD}	0.088	0.208
		R_{AD}	0.060	0.140
		R_{BD}&R_{AD}	0.050	0.116
WGAN	–	Vanilla	0.022	0.064
	Discrete (Dis)	R_{BD}	0.036	0.094
		R_{AD}	**0.016**	**0.048**
		R_{BD}&R_{AD}	0.043	0.106
	Embedded (Emb)	R_{BD}	0.023	0.076
		R_{AD}	0.020	**0.059**
		R_{BD}&R_{AD}	0.024	0.072

4-1 Model evaluation results

Diversity measured by recall

- As expected, the re-weighting exhibits poor diversity, but the proposed DGM significantly enhances the diversity.
- The VAE variants marginally outperform the WGAN variants in terms of diversity.
- The R_{AD} loss function consistently improves the diversity of WGAN and VAE in both embedding and discrete spaces, but the improvements are marginal.
- The improvements from embedding and discrete spaces are almost identical.
- The number of generated attribute combinations is partially related to the recall but is not proportional.



Method			**Diversity**	
Model	**Space**	**Loss functions**	**# of comb***	**Recall**
Prototypical agent approach			30,837	56.4 %
BN			303,723	78.4 %
VAE	–	Vanilla	355,277	81.0 %
	Discrete (Dis)	R_{BD}	265,875	79.9 %
		R_{AD}	312,506	**82.1 %**
		R_{BD}&R_{AD}	329,097	82.0 %
	Embedded (Emb)	R_{BD}	289,377	80.2 %
		R_{AD}	334,569	**82.3 %**
		R_{BD}&R_{AD}	318,731	82.0 %
WGAN	–	Vanilla	279,336	80.2 %
	Discrete (Dis)	R_{BD}	155,586	74.7 %
		R_{AD}	273,622	81.2 %
		R_{BD}&R_{AD}	152,031	74.1 %
	Embedded (Emb)	R_{BD}	225,408	77.7 %
		R_{AD}	276,012	81.3 %
		R_{BD}&R_{AD}	236,238	78.1 %

4-1 Model evaluation results

Diversity measured by recall

- By definition, the recall increases in proportion to the number of generated data points.
- The superior diversity of VAE over WGAN becomes more prominent for small sampling rates.

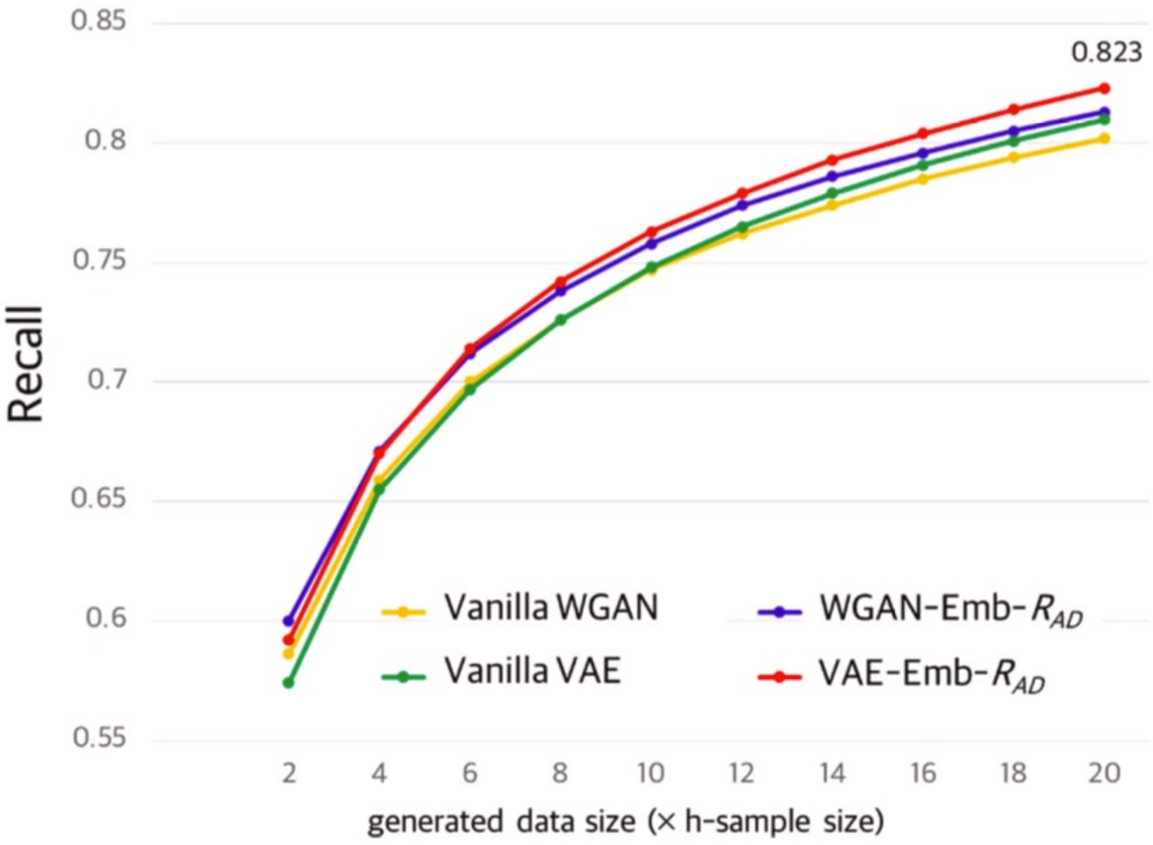
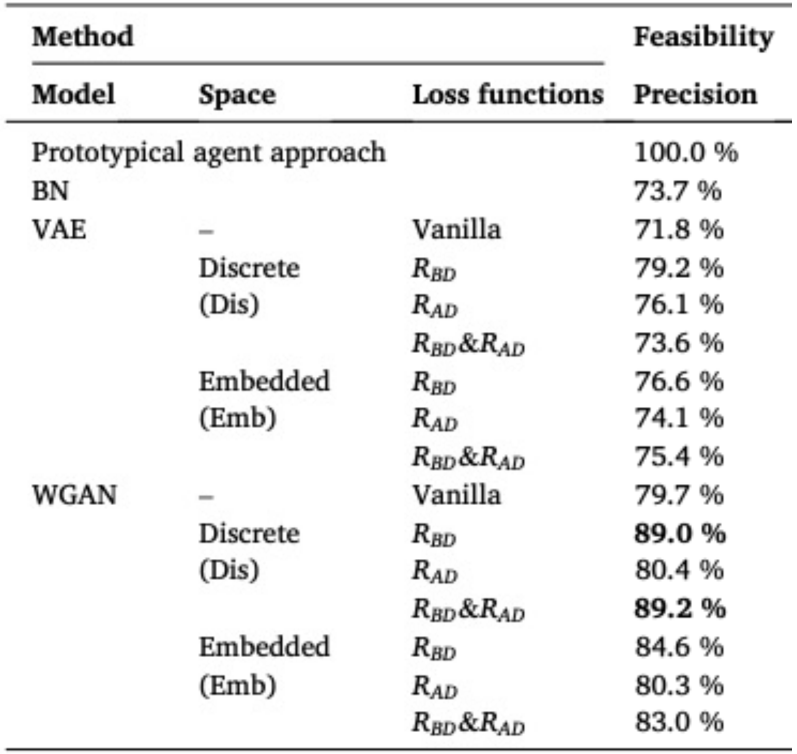


Fig. 11. The changes in the diversity of DGMs according to the number of generated data.

4-1 Model evaluation results

Feasibility measured by precision

- The WGAN with strength in producing realistic data shows higher precision than other models.
- The proposed R_{BD} loss function enhances the precision of WGAN and VAE by 9.3 % and 7.4 % compared to their vanilla counterparts.
- The precision enhancement from R_{BD} loss function is more prominent in the discrete space than in the embedding space.



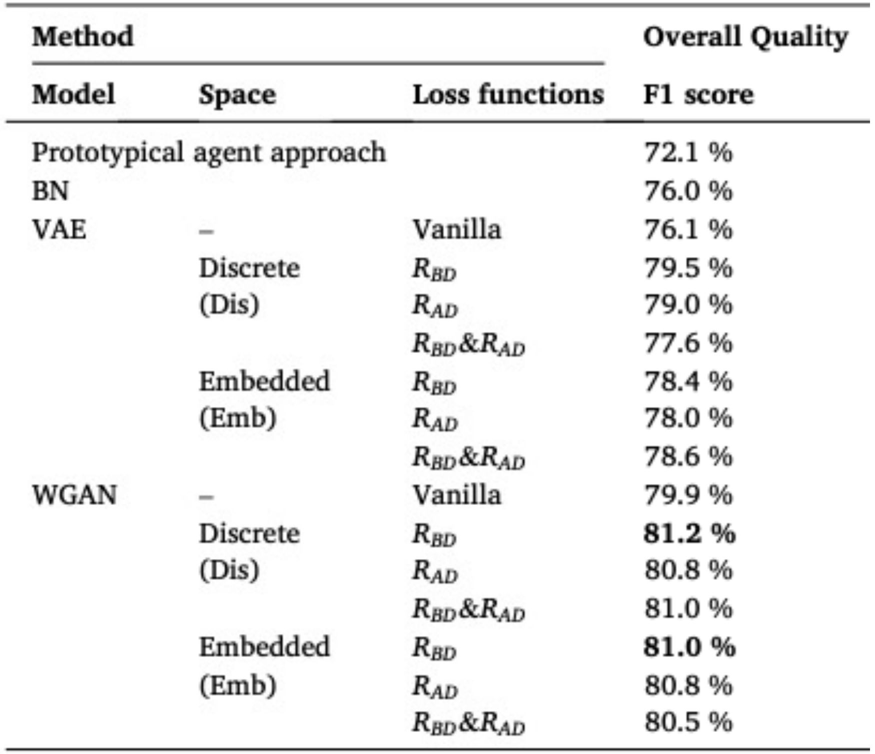
Method			**Feasibility**
Model	**Space**	**Loss functions**	**Precision**
Prototypical agent approach			100.0 %
BN			73.7 %
VAE	–	Vanilla	71.8 %
	Discrete (Dis)	R_{BD}	79.2 %
		R_{AD}	76.1 %
		R_{BD}&R_{AD}	73.6 %
	Embedded (Emb)	R_{BD}	76.6 %
		R_{AD}	74.1 %
		R_{BD}&R_{AD}	75.4 %
WGAN	–	Vanilla	79.7 %
	Discrete (Dis)	R_{BD}	**89.0 %**
		R_{AD}	80.4 %
		R_{BD}&R_{AD}	**89.2 %**
	Embedded (Emb)	R_{BD}	84.6 %
		R_{AD}	80.3 %
		R_{BD}&R_{AD}	83.0 %

4-1 Model evaluation results

Overall quality

- In terms of overall quality, WGAN outperforms VAE. But the authors suggest that the choice between them depends on the ultimate goal due to their different strengths.
- When both R_{BD} and R_{AD} are incorporated into the loss functions, the performance is inferior to other patterns.

Based on the trade-off between feasibility and diversity, it is better to find the optimal weight for each loss function.



Method			**Overall Quality**
Model	**Space**	**Loss functions**	**F1 score**
Prototypical agent approach			72.1 %
BN			76.0 %
VAE	–	Vanilla	76.1 %
	Discrete (Dis)	R_{BD}	79.5 %
		R_{AD}	79.0 %
		R_{BD}&R_{AD}	77.6 %
	Embedded (Emb)	R_{BD}	78.4 %
		R_{AD}	78.0 %
		R_{BD}&R_{AD}	78.6 %
WGAN	–	Vanilla	79.9 %
	Discrete (Dis)	R_{BD}	**81.2 %**
		R_{AD}	80.8 %
		R_{BD}&R_{AD}	81.0 %
	Embedded (Emb)	R_{BD}	**81.0 %**
		R_{AD}	80.8 %
		R_{BD}&R_{AD}	80.5 %

4-2 Sensitivity analysis of loss functions for regularization

- The higher the γ_{BD}, the higher the precision, and it monotonically increases while sacrificing the recall.
- If γ_{AD} is too large, the generated data will deviate beyond the scope of improving diversity, leading to decreases in both recall and precision.

The users can fine-tune the VAE or WGAN using the proposed two loss functions for regularization, according to their objectives.

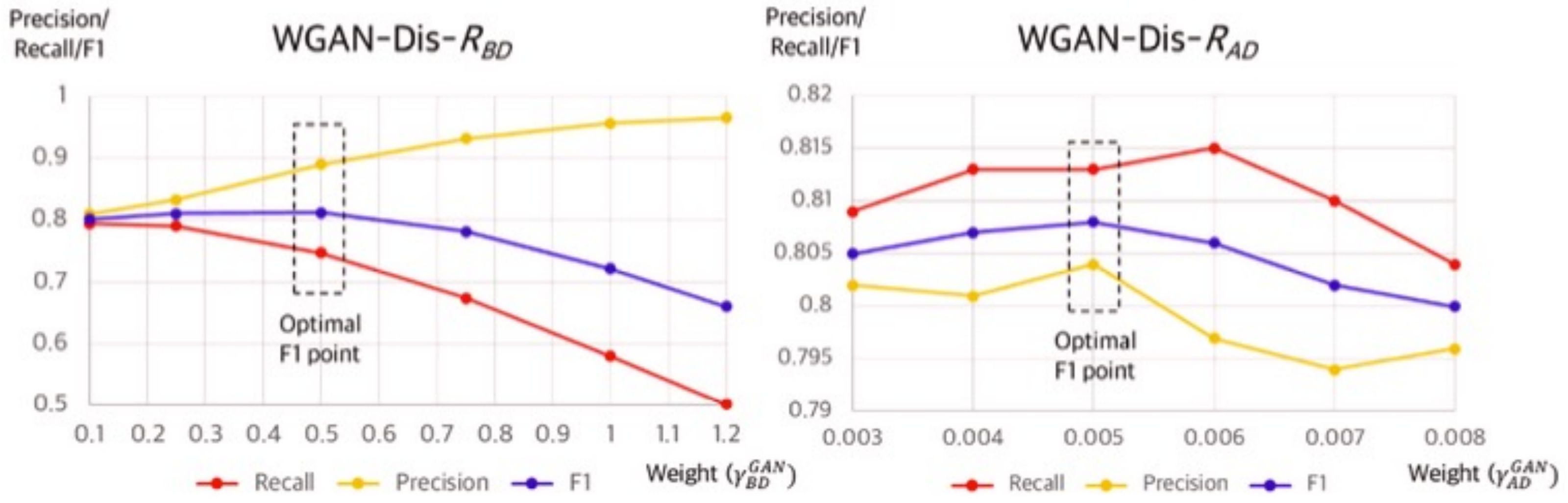


Fig. 12. Sensitivity analysis of the weights of loss functions on WGAN performance.

5.Conclusion

5 Conclusion

Objective

DGMのトレーニングにおいてsampling zeroの生成を促しつつ，strucral zeroを最小限に抑える新しい損失関数を提案 / This study proposes novel loss functions for regularization in the training of DGMs to encourage generating sampling zeros while maintaining the structural zeros at a minimum.

Novel insigths and achievement

- 分布類似性だけでDGMを評価することは十分ではない / Evaluating DGMs based solely on distributional similarity is not sufficient.
- DGMはre-weightingよりoverall qualityが高い / DGMs achieve higher overall quality compared to re-weighting methods.
- VAEはdiversityを向上させ，WGANはfeasibilityを向上させる / VAEs improve diversity, while WGANs enhance feasibility.
- 新しい損失関数はfeasibilityとdiversityのトレードオフを制御することでDGMのパフォーマンスを向上させる / New loss functions enhance DGM performance by controlling the trade-off between feasibility and diversity.

5 Conclusion

Future work

- 合成人口よりも複雑なアクティビティパターンの生成や人口合成に地理的属性を組み込むことに，効果的な正則化の開発
- 土地や人口の変化に対してオンライン学習をすることで，DGMが将来の合成人口におけるdiversityを高めることができる

所感

- 合成人口という語も知らなかったので最初はなかなか掴みどころがなく1周目はかなり苦労した
- 機械学習（特にDGM）の良い勉強になった
- ただ機械学習を実際に触ったことがないので，なかなか式が直感的にイメージしにくい
- 距離の導入というシンプルな方法でけっこう効果があるのは面白い