The Korean Journal of Applied Statistics
2023, Vol. 36, No. 2, 141–166

DOI: https://doi.org/10.5351/KJAS.2023.36.2.141
Print ISSN1225-066X / Online ISSN 2383-5818

A comparison of synthetic data approaches using utility and disclosure risk measures

Seongbin An[a], Trang Doan[b], Juhee Lee[c], Jiwoo Kim[d], Yong Jae Kim[e], Yunji Kim[a], Changwon Yoon[a], Sungkyu Jung[e], Dongha Kim[d], Sunghoon Kwon[b], Hang J Kim[f], Jeongyoun Ahn[1,a], Cheolwoo Park[2,g]

[a]Department of Industrial & Systems Engineering, KAIST;
[b]Department of Applied Statistics, Konkuk University;
[c]Department of Statistics, Kyungpook National University;
[d]Department of Statistics, Sungshin Women's University;
[e]Department of Statistics, Seoul National University;
[f]Division of Statistics and Data Science, University of Cincinnati;
[g]Department of Mathematical Sciences, KAIST

Abstract

This paper investigates synthetic data generation methods and their evaluation measures. There have been increasing demands for releasing various types of data to the public for different purposes. At the same time, there are also unavoidable concerns about leaking critical or sensitive information. Many synthetic data generation methods have been proposed over the years in order to address these concerns and implemented in some countries, including Korea. The current study aims to introduce and compare three representative synthetic data generation approaches: Sequential regression, nonparametric Bayesian multiple imputations, and deep generative models. Several evaluation metrics that measure the utility and disclosure risk of synthetic data are also reviewed. We provide empirical comparisons of the three synthetic data generation approaches with respect to various evaluation measures. The findings of this work will help practitioners to have a better understanding of the advantages and disadvantages of those synthetic data methods.

Keywords: deep generative model, disclosure risk, nonparametric Bayesian, sequential regression, synthetic data, utility

1. 서론

최근 학술적 연구를 비롯한 다양한 목적으로 공공데이터에 대한 수요가 늘어나면서 통계작성기관들의 데이터 공개에 대한 요구가 늘어나고 있다. 하지만, 개인 혹은 단체와 같이 개별단위에 대한 정보를 담고 있는 마이크로데이터(microdata)는 대부분의 경우 민감한 정보를 포함하고 있기 때문에 공공에 바로 공개하기에는 여러

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2022-0-00937, Solving the problem of increasing the usability and usefulness synthetic data algorithm for statistical data.)

[1] Corresponding author: Department of Industrial & Systems Engineering, KAIST, 291 Daehak-Ro, Yuseong-Gu, Daejeon 34141, Korea. E-mail: jyahn@kaist.ac.kr

[2] Corresponding author: Department of Mathematical Sciences, KAIST, 291 Daehak-Ro, Yuseong-Gu, Daejeon 34141, Korea. E-mail: parkcw2021@kaist.ac.kr

Published 30 April 2023 / journal homepage: http://kjas.or.kr

Table 1: Summary of generative techniques and evaluation measures

구 분	방 법	
재현자료 생성기법	순차적 회귀분석 비모수 베이지안 인공지능 기반 : CTGAN, TVAE	
유용성 지표	대역 유용성	Propensity score, 거리 측도, α-정밀도, β-재현율
	특정 유용성	신뢰구간 중첩
노출 위험도 지표	신원 노출 위험도, 속성 노출 위험도, 독창성 점수	

가지 위험이 따른다. 이에, Rubin (1993)과 Little (1993)은 정보보호 차원에서 다중대체(multiple imputation)를 통해 실제 관측값이 아닌 원본자료를 기반으로 인위적으로 생성된 재현자료(synthetic data)[3]를 생성하여 대신 배포하는 방법을 최초로 제시하였으며, 해당 방법은 미국, 호주, 캐나다 등 많은 국가의 통계작성기관에서 채택되었다 (Drechsler와 Reiter, 2009). 최근 한국도 정보개방의 흐름에 맞춰 국가기관을 중심으로 재현자료의 생성 및 적용에 대한 소개와 다양한 연구가 진행되고 있다 (Kim과 Park, 2019; Park 등, 2020).

재현자료는 민감한 정보만을 재현하고 나머지 정보는 그대로 사용하는 부분 재현자료(partially synthetic data)와 전체 데이터를 재현하는 완전 재현자료(fully synthetic data)로 구분할 수 있다. 각 종류의 재현자료에 대한 연구는 개념적 차이로 인해 독립적으로 이루어졌는데, 부분 재현자료에 대한 초기 연구로는 그 개념을 제시한 Little (1993)과 추론과 생성에 대해 다룬 Reiter (2003, 2005)가 대표적이며, 완전 재현자료에 대한 초기 연구로는 Rubin (1993)을 확장한 Raghunathan 등 (2003)이 있다. 또한, 최근에는 인공신경망 기반의 심층 생성모형(deep generative model)을 바탕으로 재현자료를 생성하는 방법도 연구되고 있는데 (Xu 등, 2019), 이는 완전 재현자료의 생성방법으로 이해할 수 있다.

재현자료를 제대로 활용을 하기위해 생성만큼 중요한 것이 재현자료의 평가이다. 재현자료의 생성 및 배포는 개인이나 기업 정보 침해 위험이 없는 안전한 데이터를 공개하여 활용을 촉진시키는 것을 목적으로 한다. 하지만, 재현자료를 활용한 분석과 원본자료를 활용한 분석이 크게 다를 경우 오히려 연구자 및 사용자가 잘못된 결론을 내리게 할 수 있다. 이에, 재현자료는 원본자료의 변수들간의 관계성을 유지하면서 정보유출의 위험성을 줄이는 것이 중요하며, 이는 각각 유용성(utility)과 노출 위험성(disclosure risk) 측도로 평가할 수 있다. 재현자료의 평가지표에 대한 연구로는 Karr 등 (2006), Drechsler와 Reiter (2009), Snoke 등 (2018) 등이 존재하며 노출 위험성 측도에 대한 연구는 상대적으로 정보 유출 가능성이 큰 부분 재현자료를 중심으로 이루어졌다.

본 논문에서는 최근 제안된 대표적인 재현자료 생성 기법들을 소개하고 그동안 다양한 연구에서 제시되었던 평가 지표들을 이용하여 이들을 비교하고자 한다. 그리고, 이를 통해 현재까지 개발된 재현자료 기법들과 평가 지표들에 대한 이해를 돕는 동시에 앞으로 진행될 재현자료 연구에 토대가 되고자 한다. Table 1은 본 논문에 사용된 재현자료 생성방법들과 평가지표들을 요약해 놓았고, 구체적 내용은 각각 2장과 3장에 설명하였다. 그리고, 재현자료 생성에 사용된 통계청에서 제공하는 전국사업체조사 데이터(SURVEY EST data)는 2.1장에서 설명하였다. 본 연구에서는 노출 위험도가 상대적으로 낮은 완전 재현자료의 생성에 대해서 다루며, Table 1에 있는 재현자료의 유용성과 노출 위험성 측도들을 중심으로 비교한다. 3.1장에서 소개하는 유용성 측도는 크게 대역 유용성 지표(global utility measure)와 특정 유용성 지표(specific utility measure), 이 두 가지로 분류할 수 있다. 대역 유용성 지표는 자료 전체의 분포적인 특성을 얼마나 비슷하게 유지하는지에 대한 측도인 반면, 특정 유용성 지표는 특정 분석이 데이터의 적용될 것을 가정하고, 해당 분석에서 원본자료와 재현자료가 얼마나 유사한 결과를 나타내는지를 기반으로 유용성을 판단한다. 3.1.1장과 3.1.2장에서

[3] Park과 Kim (2016)에서 재현(再現)자료로 번역. 이전 문헌에서는 합성데이터 (Park 등, 2013) 또는 인위자료 (Lee, 2013)로도 번역되었음.

Table 2: SURVEY EST data description

구 분	변 수 명	변 수 설 명
범주형	SEX	대표자 성별 (남/여)
	SUMMAT_CD	매출 금액 (9단계 범주)
연속형	WORKER_T	총 종사자수
	EMP_T	상용근로 종사자수
	BIS_MNTH	영업개월 수

소개하는 propensity score measure, 분포간의 거리 측도는 대표적인 대역 유용성 지표의 예시이다. 특정 유용성 지표의 대표적인 예로는 3.1.3장에서 소개하는 신뢰구간 중첩 지표(confidence interval overlap)이 있으며, 이는 데이터가 선형회귀 분석에 사용되는 경우를 가정하고 유용성을 평가한다. 3.2장에서는 신원 노출 위험도(identity disclosure risk)와 속성 노출 위험도(attribute disclosure risk) 등의 노출 위험성 측도들을 소개한다. 또한, 3.3장에서는 최근 Alaa 등 (2022)에서 제시된 새로운 재현자료의 노출 위험성과 유용성 평가지표를 소개한다. 이어 4장에서는 소개된 재현자료 방법들에 의해 생성된 전국사업체조사 재현데이터들을 다양한 평가지표들을 이용하여 비교하며, 5장에서는 본 논문을 요약하고 앞으로 고려해 볼 점들을 기술하였다.

2. 재현자료 생성 기법

재현자료 생성 기법들을 기술하기 앞서 사용될 표기들을 먼저 소개한다. 재현 대상이 되는 원본자료를 $\mathcal{D}_o$, 재현자료를 $\mathcal{D}_s$, 원본자료의 관측값을 x_{ij} (j 번째 변수의 i 번째 관측치), 재현자료의 관측값은 y_{ij}로 표기한다. 원본자료와 재현자료의 관측값의 개수는 각각 n_o, n_s로 표기하며, 이번 연구에서는 $n_o = n_s$으로 한다. 그리고, 자료 변수의 개수는 p로 표기한다.

이 장에서는 먼저 재현에 사용될 원본자료를 2.1장에서 설명하고, 순차회귀모형(2.2장), 비모수 베이지안(2.3장), 인공 신경망(2.4장) 등 세 가지 재현자료 생성 방법들을 설명한다. 본 논문에서는 원본자료 $\mathcal{D}_o$의 모든 변수의 관측값을 재현하는 완전 재현자료 생성이 목표이다.

2.1. SURVEY EST 데이터셋 설명

본 평가 연구에 사용된 자료는 통계청에서 운영하는 마이크로데이터 통합서비스(microdata integrated service; MDIS)를 통해 다운로드받은 2019년 전국사업체조사 데이터이다. 전국 조사대상 사업체 중 한국표준산업분류 코드 56에 해당하는 음식점 및 주점업 개인사업체 $n_o = 694,741$개를 본 과제의 원본자료로 정의하였다. 분석의 편의를 위해 MDIS에서 제공되는 20여개의 변수 중 Table 2에 설명된 $p = 5$가지 변수를 선택 또는 가공하여 SURVEY EST 데이터셋을 생성하였으며, 앞으로 소개될 각각의 재현자료 방법들을 적용하였다. Table 2에 있는 변수 중 '영업개월 수'는 원본자료에 있는 '창업연도' 및 '창업월' 변수를 조사기준일에 해당하는 2019년 12월을 기준으로 환산하여 생성한 변수이다.

본 연구에 사용된 SURVEY EST 데이터셋은, MDIS에서 제공하는 실제 데이터보다는 변수의 개수가 적지만, 일반적인 사업체 조사가 지니고 있는 특성을 그대로 가지고 있기 때문에 실제적인 예시로서 적합하다. 첫째, SURVEY EST 데이터셋은 전국사업체조사·경제총조사처럼 이항형, 다항형, 연속형 변수가 골고루 포함되어 있다. 또한 SURVEY EST의 '총 종사자수'와 '상용근로 종사자수'간에는 부등식(WORKER_T ≥ EMP_T)이 반드시 만족되어야 하며, 이는 에디팅 룰(edit rule)을 이용하여 정제된 자료를 생성·공표하는 통계청의 사업체 조사들의 특성을 대표한다.

Table 3: 순차회귀모형을 이용한 SURVEY EST 재현의 네 가지 방법

	X_1	X_2	X_3	X_4	X_5
1	SUMMAT_CD	SEX	WORKER_T	EMP_T	BIS_MNTH
	SWR	Reg	CART	CART	CART
2	SEX	WORKER_T	EMP_T	SUMMAT_CD	BIS_MNTH
	SWR	CART	CART	CART	CART
3	SEX	WORKER_T	EMP_T	SUMMAT_CD	BIS_MNTH
	SWR	CART-S	CART-S	Reg	CART-S
4	SEX	WORKER_T	EMP_T	SUMMAT_CD	BIS_MNTH
	SWR	CART-S	CART-S	CART-S	CART-S

각 방법의 첫 줄은 변수를, 두 번째 줄은 조건부분포 추정방법을 나타내며, "SWR"은 (2.1)의 단순복원추출, "CART"는 (2.2), "CART-S"는 (2.3), "Reg"는 다항 로지스틱 회귀모형을 의미한다.

2.2. 순차회귀모형을 이용한 재현자료 생성

순차회귀모형(sequential regression modeling)을 이용하여 재현자료를 생성하는 방법은 R 패키지 `synthpop` (Nowok 등, 2016)에 구현되어 있다. 순차회귀 재현자료 생성에서는 원본자료의 확률변수 $X_1, \ldots, X_p$의 결합분포를 순차회귀모형을 이용하여 추정하며, 결합분포에서 재현자료 표집(sampling) 역시 적합된 회귀모형을 이용하여 순차적으로 표집한다.

먼저 결합분포의 추정에 대해 설명한다. 원본자료의 p개 변수의 결합분포를

$$f\left(x_1, \ldots, x_p\right) = f_1\left(x_1\right) f_2\left(x_2 \mid x_1\right) \cdots f_p\left(x_p \mid x_1, x_2, \ldots, x_{p-1}\right)$$

와 같이 분해한 뒤, 각각의 조건부 분포 $f_j(x_j|x_1, \ldots, x_{j-1})$를 회귀모형으로 추정한다. 이 분해에 이용되는 변수의 순서에 따라 결합분포의 추정값이 다르므로, 이 순서는 사전에 이용자가 정해주어야 한다. 첫 번째 변수의 분포 $f_1(\cdot)$는 원본자료 $\mathcal{D}_o$의 첫 번째 변수 관측값들 $\{x_{11}, \ldots, x_{n_o 1}\}$의 point mass 분포로 정한다. 즉, 첫 번째 변수 X_1이 연속형이든 범주형이든

$$\hat{f}_1\left(x\right) = \hat{P}\left(X_1 = x\right) = \frac{1}{n_o} \sum_{i=1}^{n_o} 1\left(x = x_{i1}\right) \tag{2.1}$$

이다. 여기서, $1(\cdot)$은 지시함수이다. 변수가 연속형인 경우에는 커널밀도함수추정 등을 이용한 분포 추정도 가능하지만, 본 논문의 실험에서는 첫 번째를 범주형 변수로 정하였다.

조건부 분포 $f_j(x_j|x_1, \ldots, x_{j-1})$를 추정할 때는 변수 X_j의 종류 (연속형, 이산형, 범주형 등)에 따라 적당한 회귀모형을 정해주어야 한다. 이해를 돕기 위해 개념적인 예로서 X_j가 연속형인 경우를 생각해 보자. 선형회귀모형 $X_j = \beta_0 + \sum_{l=1}^{j-1} \beta_l X_l + \epsilon$을 원본자료 $\mathcal{D}_o$를 이용해 적합하여 $(X_j|x_1, \ldots, x_{j-1})$의 분포를 평균이 $\hat{\beta}_0 + \sum_{l=1}^{j-1} \hat{\beta}_l x_l$인 정규분포로 추정할 수 있다. 물론, 실제 데이터에서 확인하기 어려운 오차항의 정규 분포 가정, 그리고 선형성 가정이 어긋날 수 있으므로, 정규성을 가정한 선형회귀모형보다는 의사결정나무(classification and regression trees; CART) (Breiman 등, 2017) 등의 비모수회귀모형이 더 자주 쓰인다 (Reiter, 2005). 본 논문의 재현자료 생성 실험에서는 의사결정나무와 다중 로지스틱 회귀모형만을 이용하였으나, 순차회귀 재현자료 생성에서는 선형회귀모형을 포함한 다양한 방법이 쓰일 수 있다.

다음으로 실험에 사용한 의사결정나무와 로지스틱 회귀모형을 이용한 조건부 분포의 추정에 대해 설명한다. 의사결정나무를 이용할 때 X_j가 범주형 변수일 때는 분류 의사결정나무를, 수치형일 때는 회귀 의사결정나무를 적합한다. 의사결정나무 적합에는 R의 `rpart` 패키지를 이용하며, 연속형 변수에 대해서는 엔트로피를, 범주형 변수에 대해서는 지니 계수를 불순도로 정하고, 과적합을 방지하기 위해 적합된 의사결정나무의 잎 노드 크기가 5이상(`synthpop` 패키지의 default 옵션)이 되도록 정한다. 여기서 원본자료 $\mathcal{D}_O$의 변수 $X_1, \ldots, X_{j-1}$

을 설명변수로, 변수 X_j를 반응변수로 두고 의사결정나무를 적합한다. 일반적인 의사결정나무의 예측은 각 잎 노드에 해당하는 반응변수의 값들의 평균(회귀나무의 경우) 또는 비율이 가장 높은 최빈값(분류나무의 경우)으로 정해지지만, 재현자료 생성에 사용되는 의사결정나무의 잎 노드에는 해당하는 반응변수 값들의 조건부 point mass 분포를 포함한다. 간단한 예제로 첫 번째 변수가 이항일 때, 예를 들어 $X_1 \in \{$ 남성, 여성$\}$일 때, $(X_2|X_1)$의 조건부분포를 추정하는 의사결정나무는 단 두 개의 잎 노드(남성, 여성)만 있으며, 첫 번째 잎 노드에 해당하는 조건부분포는

$$\hat{f}_2\left(x \mid X_1 = \text{남성}\right) = \frac{1}{n_m} \sum_{i=1}^{n_m} 1\left\{x_{i2} = x \;\&\; x_{i1} = \text{남성}\right\} \tag{2.2}$$

이 된다. 이 때 $n_m = \#\{i = 1, \ldots, n_o : x_{i1} = \text{남성}\}$이다. 변수 X_2가 수치형인 경우에는 이 조건부분포 추정량을 커널밀도함수추정량으로 대체할 수 있다. 이 경우 주어진 너비 $h > 0$와 가우시안 커널 K_h에 대해 다음과 같이 조건부 분포 추정량을 정한다:

$$\hat{f}_2^{(\text{sm})}\left(x \mid X_1 = \text{남성}\right) = \frac{1}{n_m h} \sum_{i:x_{i1}=\text{남성}} K\left(\frac{x - x_{i2}}{h}\right). \tag{2.3}$$

반응변수에 해당하는 X_j가 이항 또는 순서가 없는 범주형일 때는 (다항) 로지스틱 회귀모형을 이용할 수 있다. 이 논문의 실험에서는 설명변수 $X_1, \ldots, X_{j-1}$에 조건부로 다항분포를 따르는 X_j의 각 값의 확률을 다항 선형 로지스틱 회귀모형을 이용하여 적합하였으며, 각 범주(예를 들어, x = 남성 또는 여성)에 따라 조건부 확률 $\hat{P}(X_j = x|x_1, \ldots, x_{j-1}) = \pi_x(x_1, \ldots, x_{j-1})$를 추정한다. 이 때, $\pi_x(x_1, \ldots, x_{j-1})$는 x의 값에 따라 다른 계수를 가지는 로지스틱 함수의 형태다.

위의 회귀모형 적합을 통한 조건부분포 추정을 순차적으로 마치고 나면 재현자료를 생성한다. 이는 순차회귀모형 적합에서 사용한 변수의 순서대로 추정된 조건부분포에서의 단순임의표집을 통해 얻어진다. 첫 번째 변수의 분포를 point mass 분포로 추정했으므로 n_s개의 재현자료 $y_{11}, \ldots, y_{n_s 1}$는 원본자료 $x_{11}, \ldots, x_{n_o 1}$에서 단순복원추출(sampling with replacement)한다. 재현자료의 두 번째 변수값 y_{i2}은 재현된 첫 번째 변수값 y_{i1}에 해당하는 조건부분포 $\hat{f}_2(\cdot|X_1 = y_{i1})$ (식 (2.2) 또는 (2.3))에서 임의표집한다. 수치형 변수를 임의표집한 뒤에는 원본자료의 precision에 맞게 반올림해주는 과정을 거친다. 예를 들어, X_2가 정수형 변수라면 생성된 값이 $y_{i2} = 1.321$일 때, $y_{i2} = 1$로 대체한다. 주어진 변수 순서에 따라 나머지 변수의 재현된 값을 순차적으로 임의표집함으로써 전체 데이터에 대한 재현자료 생성을 완료한다.

만약 한 부 이상의 재현자료를 생성하고자 한다면, 추정된 조건부분포를 이용한 재현자료 생성을 여러 번 반복하거나, 재현자료의 관측값의 개수 n_s를 크게 한 뒤 자료를 분절하는 방법을 사용한다.

2.2.1. 순차회귀모형을 이용한 SURVEY EST 데이터의 재현

순차회귀를 이용한 재현자료 생성은 조건부 분포를 추정할 변수의 순서와 각 조건부 분포의 추정 시 사용되는 회귀모형 적합 방법에 따라 달라진다. 본 연구에서는 네 가지의 변수 순서와 회귀모형 적합 방법의 조합을 시도하여 비교하였으며, Table 3에 이 네 가지 방법을 정리하였다. 모든 방법에서 첫 번째 변수의 분포를 항상 point mass로 추정하기 때문에 수치형 변수는 첫 번째 변수로 고려하지 않고 범주형 변수인 `SEX` 또는 `SUMMAT_CD`로 정하였다.

네 가지 방법 중 첫 두 방법은 수치형 변수값의 재현에서도 원본자료에 존재하는 값들만을 이용하는 소위 "hot deck" 재현방법이며, 나머지 방법들은 커널밀도함수추정량을 이용하여 원본자료에 존재하지 않는 값도 재현할 수 있다는 점에서 차이가 있다. 다만, SURVEY EST 데이터의 수치형 변수들이 정수형이므로, 재현자료를 반올림한 후에는 모두 원본자료에도 존재하는 값이 됨을 확인할 수 있었다.

2.3. 비모수 베이지안 모형을 이용한 재현자료 생성

비모수 베이지안 방법을 이용한 재현자료생성 기법은 미국 인구 통계국이나 미 노동 통계청 등의 해외 국가 통계기관에서 종종 사용되고 있다 (Hu와 Savitsky, 2018; Kim 등, 2021). 본 연구에서는 비모수 베이지안 모형 중, 연속형 변수와 범주형 변수가 혼합되어 있는 자료의 분포를 추정하는데 우수한 것으로 알려진 Murray와 Reiter (2016)의 HCMM-LD (hierarchically coupled mixture model with local dependence) 모형을 이용하여 재현자료를 생성하였다.

HCMM-LD 모형은 다음과 같이 계층적 형태로 자료의 형태를 설명한다.

가정 1. 범주형 변수 $\mathbf{X}_{i,\text{범주}} = (X_{i1}, \ldots, X_{ip_{\text{범주}}})$들이 혼합 다항 분포(mixture multinomial distributions)를 따르는 것으로 가정한다.

가정 2. 연속형 변수 $\mathbf{X}_{i,\text{연속}} = (X_{i,p_{\text{범주}}+1}, \ldots, X_{ip})$들이 범주형 변수를 설명변수로 사용하는 혼합 회귀 모형 (mixture regression model)을 따르는 것으로 가정한다.

가정 3. 혼합 다항 분포와 혼합 회귀 모형에 사용되는 구성요소들(mixture components)이 디리클레 과정 (Dirichlet process; DP)을 따르며, 두 변수 타입간의 의존성을 설명하는 요소 역시 DP를 따른다고 가정한다.

이를 수식으로 표현하면 다음과 같다:

$$\Pr\left(X_{ij} = d \mid s_i, \left\{\boldsymbol{\psi}_{1j}, \ldots, \boldsymbol{\psi}_{Sj}\right\}\right) = \psi_{s_i jd}, \quad i = 1, \ldots, n_o,\ j = 1, \ldots, p_{\text{범주}},\ d \in \left\{1, \ldots, l_j\right\},\ s_i \in \{1, \ldots, S\}; \quad (2.4)$$

$$f\left(\mathbf{x}_{i,\text{연속}} \mid \mathbf{x}_{i,\text{범주}}, r_i, \{B_1, \ldots, B_R\}, \{\Sigma_1, \ldots, \Sigma_R\}\right) \sim N\left(B_{r_i}\mathbf{x}_{i,\text{범주}}, \Sigma_{r_i}\right), \quad r_i \in \{1, \ldots, R\}. \quad (2.5)$$

이 때, l_j는 j 번째 범주형 변수가 가질 수 있는 수준의 총 개수이며, $\boldsymbol{\psi}_{sj} = (\psi_{sj1} \cdots \psi_{sjl_j})$ 로 정의된다. 또한 s_i와 r_i는 소속을 나타내는 지시변수로, DP 모형을 작동시키는 중요 요소이다. 먼저 범주형 변수의 결합분포가 비슷한 개체들이 모여서 군집 s를 형성하고, 그 군집에 해당하는 개체들 $\{i : s_i = s\}$만이 해당 군집의 범주형 결합 분포(joint categorical distribution)의 모수 $\{\boldsymbol{\psi}_{s1}, \ldots, \boldsymbol{\psi}_{sp_x}\}$를 결정한다. 마찬가지로 비슷한 다변량 회귀식으로 연속형 변수들의 결합분포를 설명할 수 있는 개체들이 모여서 군집 r을 형성하고 그 군집에 해당하는 개체들 $\{i : r_i = r\}$만이 회귀계수 B_r과 공분산행렬 Σ_r을 추정하는데 이용된다.

이 때, 개체 i가 여러 군집에 속할 수 있는 가능성을 고려하고, '적정한' 수의 군집이 모형 적합에 이용되도록 하기 위해서 Murray와 Reiter (2016)는 Ishwaran와 James (2001)의 stick-breaking construction을 이용하여 다음의 DP 사전분포를 가정하였다:

$$\Pr\left(s_i = s \mid k_i = k, \{\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_K\}\right) = \lambda_{ks} = \tilde{\lambda}_{ks}\prod_{s'=1}^{s-1}\left(1 - \tilde{\lambda}_{ks'}\right) \text{ where } \tilde{\lambda}_{ks} \sim \text{Beta}\left(1, \alpha^{(s)}\right) \text{ for } s = 1, \ldots, S-1 \text{ and } \tilde{\lambda}_{kS} = 1;$$

$$\Pr\left(r_i = r \mid k_i = k, \{\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_K\}\right) = \eta_{kr} = \tilde{\eta}_{kr}\prod_{r'=1}^{r-1}\left(1 - \tilde{\eta}_{kr'}\right) \text{ where } \tilde{\eta}_{kr} \sim \text{Beta}\left(1, \alpha^{(r)}\right) \text{ for } r = 1, \ldots, R-1 \text{ and } \tilde{\eta}_{kR} = 1;$$

$$\Pr\left(k_i = k \mid \boldsymbol{\pi}\right) = \pi_k = \tilde{\pi}_k\prod_{k'=1}^{k-1}\left(1 - \tilde{\pi}_{k'}\right) \text{ where } \tilde{\pi}_k \sim \text{Beta}\left(1, \alpha^{(k)}\right) \text{ for } k = 1, \ldots, K-1 \text{ and } \tilde{\pi}_K = 1.$$

이 때, 최상위 수준에 설정되어 있는 군집 지시변수 k_i는 s_i와 r_i의 의존성을 다양한 형태로 설명할 수 있도록 한다. 이에 따라, HCMM-LD 모형은 범주형 변수의 결합 분포와 연속형 변수의 회귀 모형 간의 결합 양식을 다양하게 지시할 수 있다는 장점을 가지며, 모형 mis-specification에 대한 우려를 최소화 할 수 있는 구성을 가지고 있다.

재현자료 데이터 $\mathcal{D}_s = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$의 생성은, 위의 모형을 원본자료 $\mathcal{D}_o = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$에 적합한 후, 다음의 사후 예측분포(posterior predictive distribution)로부터 새로운 값을 랜덤하게 생성하는 과정을 거친다:

$$f\left(\mathbf{y}_i \mid \mathcal{D}_o\right) = \int f\left(\mathbf{y}_i \mid \Theta\right) f\left(\Theta \mid \mathcal{D}_o\right) d\Theta. \quad (2.6)$$

이 때, Θ는 위의 HCMM-LD모형에서 소개된 모든 모수를 의미하며, $f(\Theta|\mathcal{D}_o)$는 원본자료를 이용하여 추정한 Θ의 사후분포를 의미한다. 식 (2.6)에서 재현자료 데이터의 사후 예측분포 계산에 적분이 필요한 것으로 표시되어 있지만, 실제 계산에서는 마르코프 연쇄 몬테카를로(Markov chain Monte Carlo; MCMC)로 Θ의 추정값을 반복적으로 업데이트하면서, 매 반복때마다 주어진 Θ와 $f(\mathbf{y}_i|\Theta)$를 이용하여 $\mathcal{D}_s = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$를 생성한다. 이러한 방식으로 생성된 $\mathcal{D}_s$와 식 (2.6)의 분포에서 생성된 값이 동일한 분포를 가진다는 것은 에르고딕 이론(ergodic theorem)에 의해 증명된다.

여기서 사용된 비모수 베이지안 모형은 MCMC 방법으로 모수를 추정하기 때문에 다른 빈도론 방법보다 계산시간이 오래 걸리는 단점을 가지고 있다. 하지만 사용자가 임의로 설정해야 하는 tuning parameter가 거의 없기 때문에 사용자에 상관없이 일관된 재현자료를 생성할 수 있다는 장점을 가지고 있다. 또한, 모든 확률모형 가정을 상세히 기술하여, 재현자료가 어떻게 원본자료 분포를 있는 그대로 재현해주는지 그 과정 자체를 확률모형으로 설명해 준다. 이는 다른 기계학습 기반 재현자료 생성방법에 비해 재현자료 생성과정의 투명성을 보장한다는 장점을 가지고 있다.

2.3.1. 비모수 베이지안 모형을 이용한 SURVEY EST 데이터의 재현

수식 (2.4)와 (2.5)의 기호를 이용하여 SURVEY EST 데이터의 비모수 베이지안 모형 적합을 설명하면 다음과 같다. 범주형 변수 SEX와 SUMMAT_CD는 $\mathbf{x}_{i,\text{범주}} = (x_{i1}, x_{i2})$가 되며, 각각의 총 수준의 개수는 이항변수인 SEX의 경우 $l_1 = 2$, SUMMAT_CD의 경우 $l_2 = 9$으로 설정된다. 나머지 변수들은 연속형 변수 벡터 $\mathbf{x}_{i,\text{연속}} = (x_{i3}, x_{i4}, x_{i5})$로 설정된다. 범주형 변수간의 결합분포는 혼합 다항분포로, 범주형 변수와 연속형 변수간의 관계는 혼합회귀모형의 회귀계수로서, 그리고 연속형 변수간의 결합분포는 혼합 회귀분석모형의 오차항의 공분산 행렬로서 설명된다.

앞에서 설명하였듯이 비모수 베이지안 모형은 사용자가 특정해야 하는 모형의 구조나 tuning parameters가 거의 없으며, 유일하게 결정해야 하는 부분은 각 혼합 모형에 지시변수가 가질 수 있는 최대값으로, 본 실험에서는 $(R, S, K) = (30, 50, 40)$으로 설정하였다. 각각의 최대값이 충분히 크게 설정되어 있는지 여부는 MCMC에서 생성되는 r_i, s_i, k_i들을 요약하여 확인할 수 있다. 본 실험에서는 평균적으로 (20.0, 25.6, 20.0)개의 군집들이 (r, s, k)들의 최대 개수로 각각 추정되어, 이미 설정하였던 (R, S, K)가 충분히 크다는 점을 확인하였다. 충분히 수렴이 된 Markov chain에서 사후분포 및 재현자료를 생성하기 위해 1,500번의 burn-in iterations을 반복하였고, 각 재현자료를 거의 독립인 상태로 뽑기 위해, burn-in 후 매 500 iterations에서 한 개씩 총 5개의 재현자료를 생성하였다.

2.4. 인공 신경망을 이용한 재현자료 생성

앞의 두 방법론과 함께 심층 생성 모형을 활용한 재현자료 생성 기법에 대해서도 고려한다. 다양한 심층 생성 모형이 존재하지만 본 연구에서는 잠재변수(latent variable)를 이용한 방법론을 이용한다. 원본자료의 확률변수를 $\mathbf{X} \in \mathbb{R}^p$, 잠재변수를 $\mathbf{Z} \in \mathbb{R}^d$라 하면 생성모형의 결합 분포를 다음과 같이 수식화할 수 있다:

$$f_{\mathbf{X},\mathbf{Z}}(\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}) = f_{\mathbf{Z}}(\mathbf{Z} = \mathbf{z})\, f_{\mathbf{X}|\mathbf{Z}}(\mathbf{X} = \mathbf{x} \,|\, \mathbf{Z} = \mathbf{z}).$$

주어진 실수 a가 d번 반복된 벡터를 $\mathbf{a}_d$라 표현하고, 벡터 $\mathbf{v}$에 대해서 diag($\mathbf{v}$)를 대각 원소가 $\mathbf{v}$로 이루어진 대각행렬이라 하자. 본 연구에서는 잠재변수의 주변 분포 $f_{\mathbf{Z}}$는 $\mathcal{N}(\mathbf{0}_d, \text{diag}(\mathbf{1}_d))$를 사용한다. 또한, 조건부 분포 $f_{\mathbf{X}|\mathbf{Z}}$를 수식화할 때 심층 인공 신경망 모형(deep neural network; DNN) $f_{\boldsymbol{\theta}}(\mathbf{z}) : \mathbb{R}^d \to \mathbb{R}^p$을 사용한다. 어떤 학습 방법을 사용하느냐에 따라 $f_{\mathbf{X}|\mathbf{Z}}$와 $f_{\boldsymbol{\theta}}(\mathbf{z})$ 사이의 관계가 달라지므로 이에 대한 자세한 내용은 추후 서술한다.

잠재변수를 이용한 심층 생성 모형의 학습은 크게 두 가지로 나누어 볼 수 있는데, 적대적 학습(adversarial learning) 방법론과 우도 함수 기반 방법론이 그것이다. 첫 번째 방법의 경우 생성 모형 외에 구분자(discrim-

inator) 모형을 정의하는데, 이는 주어진 데이터가 원본자료인지 재현된 자료인지를 분간하는 이진 분류기를 뜻한다. 구분자 모형은 원본자료와 생성 모형이 생성한 재현자료를 최대한 정밀하게 분류하는 방향으로, 생성 모형은 원본자료와 최대한 비슷하게 데이터를 생성하여 구분자 모형이 잘 분간하지 못하는 방향으로, "적대적" 관계에 놓인 두 모형을 동시에 학습하는 최소-최대(mini-max) 학습 방식을 사용한다. 본 연구에서는 적대적 학습 방법론 중에서 테이블 형태의 데이터 재현에 특화된 CTGAN (conditional tabular generative adversarial network) (Xu 등, 2019) 알고리즘을 고려한다.

우도 함수 기반 알고리즘은 말 그대로 재현 대상이 되는 X의 우도 함수값이 최대가 되는 방향으로 학습하는 것을 의미한다. 주어진 $\mathbf{X}$에 대해서 로그 우도 함수값은 아래와 같이 적을 수 있다:

$$\log f_{\mathbf{X}}(\mathbf{X}=\mathbf{x}) = \log \int f_{\mathbf{X},\mathbf{Z}}(\mathbf{X}=\mathbf{x},\mathbf{Z}=\mathbf{z})\, d\mathbf{z} = \log \int f_{\mathbf{X}|\mathbf{Z}}(\mathbf{X}=\mathbf{x} \mid \mathbf{Z}=\mathbf{z})\, p(\mathbf{Z}=\mathbf{z})\, d\mathbf{z}.$$

하지만, 조건부 분포 $f_{\mathbf{X}|\mathbf{Z}}$의 형태가 매우 복잡하기 때문에 위의 적분을 계산하는 것은 거의 불가능에 가깝다. 이를 해결하기 위해 MCMC 혹은 이의 발전된 방법 등을 활용해 위의 식을 정밀하게 근사하는 것에 초점을 두어 다양한 대안이 제안되었으나, 현재까지 가장 널리 사용하는 방법은 변분 모형(variational model)을 도입하여 로그 우도의 하한인 ELBO (evidence lower bound)를 최대화하는 방식이다. 본 연구에서는 ELBO를 최대화하여 생성 모형을 학습하는 대표적인 방법론인 VAE (variational auto-encoders) (Kingma와 Welling, 2013)를 테이블 자료 생성 문제에 응용한 TVAE (tabular VAE) (Xu 등, 2019) 알고리즘을 고려한다.

2.4.1. 데이터 전처리

일반적으로 재현의 목표 대상이 되는 테이블 자료는 연속형 자료와 범주형 자료가 혼재하고 있기 때문에, 전처리 없이 심층 생성 모형을 적용한다면 학습의 실패 또는 불안정성을 유발할 수 있다. 따라서, 성공적인 학습을 위해서는 데이터 전처리가 필수적이며, 본 연구에서는 Xu 등 (2019)에서 사용한 전략을 적용하였다.

범주형 값을 갖는 자료의 경우 원-핫 인코딩(one-hot encoding)을 통해 이진 분류 벡터로 변환하였다. 즉, l_j가지의 범주를 갖는 변수 X_j는 각 원소가 0 또는 1만을 가지는 l_j차원 벡터로 전처리하였다. 예를 들어, $X_j = l$이라면 원-핫 인코딩 벡터는 l 번째 원소만 1이고 나머지는 모두 0인 l_j차원 벡터가 된다. 연속형 값을 갖는 변수의 경우 해당 변수의 주변 분포가 혼합 정규 분포를 따른다고 가정하고, 경험 분포에 변분 혼합 정규 모형(variational Gaussian mixture model; VGM)을 적합해서 군집의 수와 각 군집마다 정규 분포의 평균과 분산을 동시에 추정한다. 주어진 관측값이 주어졌을 때 앞서 적합한 VGM 모형을 이용하여 관측값이 어떤 군집에 속하는지에 대한 정보와 군집 정규 분포의 평균과 분산을 통해 정규화한 값의 정보를 이용해 관측값을 변환한다. 여기서, 관측치가 속하는 군집의 정보는 범주형 값을 가지므로 이 또한 앞서 범주형 자료를 처리했던 것처럼 원-핫 인코딩 벡터로 변환해준다.

2.4.2. CTGAN

CTGAN 알고리즘에서 관측 자료는 잠재변수와 DNN $f_{\boldsymbol{\theta}}(\mathbf{z})$을 통한 결정적 관계를 가진다고 모형화한다. 즉, 아래의 수식을 통해 생성 자료의 분포를 나타낼 수 있다:

$$\mathbf{Z} \sim \mathcal{N}(\mathbf{0}_d, \mathrm{diag}(\mathbf{1}_d)), \quad \mathbf{X} \mid (\mathbf{Z}=\mathbf{z}) = f_{\boldsymbol{\theta}}(\mathbf{z}).$$

CTGAN 알고리즘은 수많은 적대적 학습 방법론 중에서 WGAN (Arjovsky 등, 2017)에 기반한 방법론을 사용한다. WGAN 방법론의 목적 함수는 다음과 같다:

$$\min_{\boldsymbol{\theta}} \max_{\boldsymbol{\eta}} \mathbb{E}_{\mathbf{X}\sim P_{\mathbf{X}}}\left[d(\mathbf{X};\boldsymbol{\eta})\right] - \mathbb{E}_{\mathbf{Z}\sim N(\mathbf{0}_d,\mathbf{1}_d)}\left[d\left(f_{\boldsymbol{\theta}}(\mathbf{Z});\boldsymbol{\eta}\right)\right] \text{ subject to } \left\|d(\mathbf{x};\boldsymbol{\eta})\right\|_L \leq 1.$$

여기서, $P_{\mathbf{X}}$는 관측한 자료의 분포를 뜻하고, $d(\mathbf{x};\boldsymbol{\eta})$는 모수 $\boldsymbol{\eta}$를 통해 만들어진 DNN 구분자, $\|d(\mathbf{x};\boldsymbol{\eta})\|_L$는 구분자 함수 $d(\mathbf{x};\boldsymbol{\eta})$의 Lipschitz 상수를 뜻한다. 즉, WGAN은 원본자료와 재현자료를 가장 잘 분간하는 Lipschitz 상수가 1인 구분자를 고려했을 때에도 분간이 어렵도록 하는 심층 생성 모형을 학습하는 것이다. 실제 CTGAN 구현시, WGAN의 안정적인 학습을 위해 gradient penalty를 추가한 WGAN-GP (Gulrajani 등, 2017) 방법론을 사용한다. 이 밖에도 불균형한 범주형 자료의 생성 품질을 높이기 위해 조건부 GAN 모형을 추가적으로 고려하며, 모드 붕괴(mode collapse) 등의 문제를 효과적으로 해결하기 위해 PacGAN (Lin 등, 2018)의 아이디어도 차용한다.

2.4.3. TVAE

TVAE는 변분 분포를 도입하여 관측자료의 로그 우도 함수 대신 하한인 ELBO를 최대화하는 VAE 방법을 통해 심층 생성 모형을 학습한다. TVAE는 데이터의 분포를 다음과 같이 정의한다:

$$\mathbf{Z} \sim \mathcal{N}\left(\mathbf{0}_d, \operatorname{diag}\left(\mathbf{1}_d\right)\right) f_{\mathbf{X}|\mathbf{Z}}\left(\mathbf{X}=\mathbf{x} \mid \mathbf{Z}=\mathbf{z};\boldsymbol{\theta}\right) = \prod_{j=1}^{p} f_j\left(x_j;\mu_j\left(\mathbf{z};\boldsymbol{\theta}\right)\right).$$

각 원소 x_j, $j=1,\ldots,p$에 대해 x_j가 연속형이라면 $f_j(x_j;\mu_i(\mathbf{z};\boldsymbol{\theta}))$는 평균이 $\mu_j(\mathbf{z};\boldsymbol{\theta})$, 분산이 σ^2인 정규 분포를 사용하며, 이진형이라면 평균이 $\mu_j(\mathbf{z};\boldsymbol{\theta})$인 베르누이 분포를 가정한다.

모수 $\boldsymbol{\eta}$를 이용해 정의되는 변분 모형을 $g_{\mathbf{Z}|\mathbf{X}}(\mathbf{Z}=\mathbf{z}|\mathbf{X}=\mathbf{x};\boldsymbol{\eta})$라 하면, 위의 생성 모형으로부터 도출되는 ELBO는 다음과 같이 표현할 수 있다:

$$\log f_{\mathbf{X}}\left(\mathbf{X}=\mathbf{x};\boldsymbol{\theta}\right) \geq \int \log\left(\frac{f_{\mathbf{X}|\mathbf{Z}}\left(\mathbf{X}=\mathbf{x} \mid \mathbf{Z}=\mathbf{z};\boldsymbol{\theta}\right)}{g_{\mathbf{Z}|\mathbf{X}}\left(\mathbf{Z}=\mathbf{z} \mid \mathbf{X}=\mathbf{x};\boldsymbol{\eta}\right)}\right) \cdot g_{\mathbf{X}|\mathbf{Z}}\left(\mathbf{Z}=\mathbf{z} \mid \mathbf{X}=\mathbf{x};\boldsymbol{\eta}\right) d\mathbf{z} =: \operatorname{ELBO}\left(\boldsymbol{\theta},\boldsymbol{\eta}\right).$$

두 함수 $\boldsymbol{\mu}(\mathbf{x};\boldsymbol{\eta}), \boldsymbol{\sigma}^2(\mathbf{x};\boldsymbol{\eta}) : \mathbb{R}^p \to \mathbb{R}^d$리고 하면, TVAE는 변분 모형 $g_{\mathbf{Z}|\mathbf{X}}(\mathbf{Z}=\mathbf{z}|\mathbf{X}=\mathbf{x};\boldsymbol{\eta})$을 평균이 $\boldsymbol{\mu}(\mathbf{x};\boldsymbol{\eta})$, 공분산 행렬이 $\operatorname{diag}(\boldsymbol{\sigma}^2(\mathbf{x};\boldsymbol{\eta}))$인 다변량 정규 분포로 사용한다. TVAE는 위의 ELBO식을 최대화하는 모수 $\boldsymbol{\theta},\boldsymbol{\eta}$를 추정한다.

2.5. 재현자료 생성 기법의 특징과 차이점

지금까지 소개한 순차회귀방법, 비모수 베이지안 방법, 그리고 인공 신경망을 이용한 재현 자료 생성 기법의 특성과 장단점을 간단히 비교한다. 세 방법 모두 독립적인 개인들로 이루어진 테이블 형태의 자료의 재현에 용이하다. 즉, 횡단면(cross-sectional) 데이터의 재현은 가능하지만, 패널데이터(panel data) 또는 종단자료(longitudinal data)로의 확장은 아직 깊이있는 연구가 되어 있지 않다. 시계열자료 역시 순차회귀와 비모수 베이지안 방법을 이용한 재현방법은 아직 연구되지 않았다. 다만 인공신경망을 이용한 시계열자료의 재현은 최근에 많은 연구가 되었다 (Yoon 등, 2019).

순차회귀를 이용한 재현자료 생성은 R 패키지 `synthpop`를 이용하여 쉽게 이용할 수 있으며, 본 연구에 사용된 중간 크기의 자료의 재현에 필요한 계산시간이 매우 짧다는 큰 장점이 있으나, 순차회귀모형을 적합할 변수의 순서와 구체적인 모형의 선택에 따라 그 성능이 달라지는 단점이 있다. 이에 비해 비모수 베이지안 모형은 사용자가 임의로 결정해야 하는 부분이 없어 일관된 재현자료 생성이 가능하지만, 계산시간이 오래 걸리는 단점이 있다. 이 두 방법론 모두 이 논문에서 다루는 완전 재현자료 생성 뿐 아니라 부분 재현자료 생성도 가능하며, 재현 데이터를 생성하는 확률모형이 명확하기 때문에, 재현자료 생성과정의 투명성을 보장한다는 장점이 있다.

인공 신경망을 이용한 재현자료 생성은 위에서 설명한 두 기법과는 확연히 대비되는 특성이 있다. 테이블 데이터의 재현을 염두에 두고 개발된 CTGAN과 TVAE는 Python 라이브러리 `SDV`를 이용하여 쉽게 사용할 수 있으며, 인공 신경망의 훈련에 필요한 시간이 길지 않은 장점이 있지만, 인공 신경망의 성공적인 훈련을 위한

전처리가 필수적이며, 재현자료가 생성된 확률모형의 우도 계산이 명확하지 않은 단점이 있어 우도 기반의 재현자료 검증이 어렵다. 또한, 부분 재현자료 생성 방법에 대해서는 아직 연구된 바가 없으나, CTGAN이 고려한 조건부 생성 모형을 활용한다면 어렵지 않게 확장이 가능할 것으로 생각된다.

본 연구에서는 세 가지 방법론들이 모두 이상적으로 잘 작동할 것으로 예상되는 테이블 형태의 데이터인 SURVEY EST의 재현자료를 생성하여 평가한다.

3. 재현자료의 평가 지표

재현된 자료의 유용성과 노출 위험도를 측정하기 위해 여러가지 평가 지표들이 개발되었는데, 그 중 몇 가지 대표적인 지표들을 검토하고자 한다. 3.1.1장과 3.1.2장은 대역 유용성 지표들인 propensity score와 분포간의 거리들을, 3.1.3장은 특정 유용성 지표인 선형회귀모형분석 하에서의 신뢰구간 중첩 지표를 소개한다. 3.2장에서는 대표적인 노출 위험도인 신원 노출 위험도와 속성 노출 위험도에 대해 설명하며 각 노출 위험도에 대한 지표를 소개한다. 마지막으로, 3.3장은 최근 Alaa 등 (2022)가 제안한 새로운 형태의 유용성과 노출 위험도 지표들을 소개한다.

3.1. 유용성 측도

3.1.1. Propensity score measure

Rosenbaum과 Rubin (1983)이 제시한 propensity score는 원래 처리 효과를 추정하는 기법으로 개발되었으나, Woo 등 (2009)은 이를 재현자료의 유용성 평가 지표로 사용하였고, 지금까지 대표적인 지표 중 하나로 사용되어지고 있다. Propensity score는 공변량 X가 주어졌을 때, 처리 그룹(treatment group)으로 배치될 확률 $\Pr(\text{treatment} = 1|X)$로 정의된다. 본 논문에서는 원본자료와 재현자료 중에서 재현자료로 배치되는 경우가 처리 그룹으로 배치되는 경우이다.

Propensity score를 구하는 과정은 다음과 같다. 먼저, 원본자료 $\mathcal{D}_o$와 재현자료 $\mathcal{D}_s$를 병합한 뒤, 재현자료에서 생성된 관측치에는 1, 원본자료에서 생성된 관측치에는 0을 부여하는 반응변수 T_{Syn}를 생성한다. 그리고 병합된 자료와 T_{Syn}를 각각 설명변수와 반응변수로 하여 학습된 분류모델에서 각 관측치의 재현자료로 판단될 확률 $\Pr(T_{Syn} = 1|X)$을 추정한다. 만약 원본자료와 재현자료의 구조가 완전히 동일하여 구분할 수 없다면 병합된 자료의 각 i 번째 관측치의 propensity score $\hat{p}_i$는 재현자료의 비율 $c = n_s/(n_s + n_o)$를 값으로 가지며, 반대로 완벽하게 구분되는 경우 $\hat{p}_i$는 0 또는 1의 값을 가지게 된다. Woo 등 (2009)은 원본자료 $\mathcal{D}_o$와 재현자료 $\mathcal{D}_s$ 각각의 propensity score인 $\Pr(T_{Syn} = 1|\mathcal{D}_o)$와 $\Pr(T_{Syn} = 1|\mathcal{D}_s)$의 분포가 비슷할수록 재현자료의 유용성이 높다고 해석한다. 본 연구에서는 로지스틱 회귀모형과 의사결정나무를 분류모델로 활용하였다. 의사결정나무의 경우 `Scikit-learn`의 `tree` 라이브러리를 사용하였으며, hyperparameter는 min_samples_split = 20, min_samples_leaf = 5, max_depth = 30으로 설정하였다.

Propensity score mean squared error

Snoke 등 (2018)에서는 다음과 같이 propensity score 기반의 유용성 평가 지표인 propensity mean squared error (pMSE)를 사용하여 재현자료의 유용성을 측정하는 것을 제시하였다.

$$\text{pMSE} = \frac{1}{n_s + n_o} \sum_{i=1}^{n_s+n_o} (\hat{p}_i - c)^2 . \tag{3.1}$$

원본자료와 재현자료의 분류가 힘들수록, 즉 재현자료의 유용성이 높을수록 pMSE는 0에 가까운 값이 된다. 하지만, 재현자료의 비율과 학습된 모델 등의 요소에 따라 pMSE 값에 대한 판단 기준이 달라질 수 있으므로

Snoke 등 (2018)에서는 pMSE의 귀무 분포(null distribution)를 활용한 유용성 평가 지표인 pMSE-ratio와 standardized pMSE를 제시한다.

pMSE 귀무 분포

재현자료가 원본자료를 잘 재현하였다고 가정한 후, Snoke 등 (2018)에서 제시하는 pMSE의 귀무 분포를 구하는 방법은 크게 이론적 방법과 재표집(resampling) 방법 두 가지로 나뉜다. 먼저, 이론적 방법은 propensity score를 구하기 위해서 로지스틱 회귀모형을 사용한 경우 pMSE는 카이제곱분포를 따르는 확률변수의 상수배를 한 분포를 따르게 된다:

$$\text{pMSE} \sim \frac{(1-c)^2 c}{n_s + n_o} \chi^2_{p'-1}.$$

이 때, p'은 로지스틱 회귀모형의 학습에 사용된 설명변수의 개수이며 카이제곱분포의 자유도는 $p'-1$이 된다. 따라서, pMSE는 다음과 같은 평균과 표준편차를 가진다:

$$\mathbb{E}(\text{pMSE}) = \frac{(1-c)^2 c}{n_s + n_o}(p'-1), \quad sd(\text{pMSE}) = \frac{(1-c)^2 c}{n_s + n_o}\sqrt{2(p'-1)}.$$

재표집을 통한 재현자료의 귀무 분포 생성 방법은 쌍별(pairwise)방법과 재배열(permutation) 방법 두 가지가 Snoke 등 (2018)에 의해 제시되었다. 쌍별 방법은 여러 셋의 재현자료들 중 한 쌍을 선정하여 임의로 섞은 후 반복적으로 pMSE를 계산하여 귀무 분포를 구하는 방식이고, 재배열 방식은 하나의 재현자료를 사용하되 반응변수 T를 재배열해서 계산한 pMSE들을 활용하여 귀무 분포를 구하는 방법이다.

귀무 분포의 평균과 표준편차를 각각 μ_{null}과 sd_{null}이라고 한다면 pMSE-ratio와 standardized pMSE를 식 (3.2)처럼 정의하며,

$$\text{pMSE-ratio} = \frac{\text{pMSE}}{\mu_{\text{null}}}, \quad \text{standardized pMSE} = \frac{\text{pMSE} - \mu_{\text{null}}}{sd_{\text{null}}}, \tag{3.2}$$

pMSE-ratio 는 1에 가까울수록, standardized pMSE는 0에 가까울수록 재현자료의 유용성이 높음을 나타낸다.

3.1.2. 분포간의 거리 측도

두 확률분포간의 거리를 재는데 흔히 사용되는 측도들 중 Kullback–Leibler (KL) 괴리도 (Kullback과 Leibler, 1951)와 Wasser-Stein 거리 (Villani, 2008)를 이용하여 유용성을 평가한다. 다변량 자료들의 분포를 정밀하게 추정하는 것은 어려운 일이므로, 각 변수별로 일차원 분포간 거리를 구한 뒤 이를 합하는 과정을 통한다.

두 개의 일차원 확률 분포 $f(x)$와 $g(x)$ 사이의 KL 괴리도는

$$D(f\|g) = \int_{-\infty}^{\infty} f(x) \log \frac{f(x)}{g(x)} dx,$$

로 정의되고, r-Wasserstein 거리는 f와 g의 누적분포함수(cumulative distribution function) F_f과 F_g를 이용하여

$$W_r(f,g) = \left(\int_0^1 \left| F_f^{-1}(t) - F_g^{-1}(t) \right|^r \right)^{1/r} dt,$$

로 정의된다. 본 연구에서는 2-Wasserstein 거리를 이용하였다.

Table 4: Notations for §3.2.1

Notation	설명
n	원본(재현)자료의 관측치 개수
f_i	원본자료의 i 번째 관측치에 대해 준식별자 값이 같은 관측치 개수
X_i	원본자료의 i 번째 관측치의 민감 변수 값
p_i	원본자료에서 X_i와 같은 값을 갖는 관측치의 비율
d_i	$1 - p_i$
Y_t	원본자료의 i 번째 관측치와 연결된 재현자료의 민감 변수 값
d'_i	원본자료에서 X_i가 속한 군집에 있는 관측치의 비율

3.1.3. 신뢰구간 중첩 지표

Snoke 등 (2018)에서 정의한 특정 유용성 평가란 원본자료와 재현자료에 대해 특정 분석을 시행한 후 얻은 결과 간의 유사도를 비교하는 것이다. 가장 일반적이고 이해하기 쉬운 특정 유용성 평가 지표로는 선형회귀모형 적합을 통해 얻은 계수들을 비교하는 방법이 있다. 그 중 원본자료와 재현자료를 이용해 구한 각 회귀계수의 신뢰구간 중첩 정도를 계산하는 방법이 많이 사용된다 (Karr 등, 2006; Drechsler와 Reiter, 2009; Snoke 등, 2018).

Karr 등 (2006)에 따르면 j 번째 설명변수에 대응하는 모집단 회귀계수 β_j에 대한 95% 신뢰구간 중첩 정도를 다음과 같이 정의하며,

$$IO_j = 0.5\left[\frac{\min\left(u_{o,j}, u_{s,j}\right) - \max\left(l_{o,j}, l_{s,j}\right)}{u_{o,j} - l_{o,j}} + \frac{\min\left(u_{o,j}, u_{s,j}\right) - \max\left(l_{o,j}, l_{s,j}\right)}{u_{s,j} - l_{s,j}}\right] \text{ for } j = 0, 1, 2, \ldots, p, \tag{3.3}$$

신뢰구간 중첩 지표인 IO는 다음과 같이 정의한다.

$$\text{IO} = \frac{1}{p+1}\sum_{j=0}^{p} IO_j,$$

여기서 $(l_{o,j}, u_{o,j})$와 $(l_{s,j}, u_{s,j})$는 각각 원본자료로부터 추정한 계수 β_j에 대한 95% 신뢰구간과 재현자료로부터 추정한 계수 β_j에 대한 95% 신뢰구간을 뜻한다. IO 값과 재현데이터의 유용성은 비례하는데 IO의 최댓값은 1이다. 또한 원본자료와 재현자료에서 각각 구한 신뢰구간들 사이에 중첩되는 부분이 없을 때 IO는 음수 값을 갖게 되는데, 이 신뢰구간들이 더 많이 떨어질수록 IO 값도 감소한다.

3.2. 재현자료의 노출 위험도 평가 지표

3.2.1. 신원 노출 위험도

EI Emam 등 (2020)에 따르면 신원 노출 위험도란 재현자료의 관측치와 원본자료의 관측치를 옳게 연결할 위험이다. 지금까지 신원 노출 위험도를 측정하는 연구는 부분재현자료라는 전제 하에만 진행되었지만, 완전재현자료일지라도 생성 모델이 과적합되었을 때 재현자료로부터 원본자료의 관측치를 연결하고 그것으로부터 새로운 것을 배우는 것이 가능하다. 변수들은 준식별자(quasi-identifier)와 민감변수(sensitive variable)로 나뉠 수 있는데, 재현자료의 관측치의 신원을 확인할 때는 준식별자가 사용된다. 준식별자는 비교적 쉽게 찾아낼 수 있는 정보이기 때문에 공격자(attacker)가 이미 알고 있다고 가정한다. 준식별자를 제외한 나머지 변수들을 민감변수라고 하는데, 만약 공격자가 알게 되면 어느 정도의 피해가 발생한다. 원본자료와 재현자료의 준식별자가 유일하게 동일하면 신원이 확인될 위험이 매우 높아지지만 민감변수의 값이 비슷하지 않다면 해당 관측치로부터 새로운 것을 배울 가능성이 적다.

EI Emam 등 (2020)에서 정의한 신원 노출 위험도는 Table 4에 정리된 표기법을 사용하여 다음과 같이 계산된다. 우선 신원 자체가 식별되는 지를 나타내는 지시함수 I_i 를 정의하는데, 원본자료의 i 번째 관측치에 대응하는 재현자료 관측치가 있다면 1값을, 그렇지 않으면 0값을 갖게 된다. 다음으로 공격자가 $I_i = 1$ 인 관측치에 한해 민감변수에 대한 새로운 정보를 취득하게 되었는지를 판단하게 되는데 이는 원본자료의 값과 재현된 값 사이의 유사한 정도로 판단한다. 민감변수가 명목형이라면 다음의 부등식을 만족하는 경우를 찾고

$$d_i \times I(X_i = Y_t) \; > \; \sqrt{p_i(1-p_i)}, \quad i = 1, 2, \ldots, n,$$

연속형이라면 k-means를 이용해 값을 군집화 한 후, 다음의 부등식을 만족하는지 알아본다.

$$d_i' \times |X_i - Y_t| \; < \; 1.48 \times \text{MAD}, \quad i = 1, 2, \ldots, n.$$

여기서 MAD는 중위절대편차이다. 원본자료의 i 번째 관측치에 대해 위 부등식을 만족하는 민감변수의 비율이 5% 이상일 때 1을 갖고 그렇지 않으면 0을 갖는 지시함수 R_i를 정의한다. 마지막으로 신원 노출 위험도는 다음과 같이 정의된다:

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{f_i} \times I_i \times R_i\right).$$

이 값이 작을수록 공격자가 신원을 정확하게 추측할 가능성이 작아진다.

3.2.2. 속성 노출 위험도

속성 노출은 공격자가 개인의 신원을 식별할 수는 없지만 특정 민감한 변수의 속성을 추론할 수 있을때 발생한다 (Stan 등, 2015). 완전 재현자료의 각 관측치는 인위적으로 만들어졌으므로 원본자료의 각 관측치와 직접적으로 연관되어있지 않을 수 있으나 속성 노출 위험도는 항상 존재한다. Markus 등 (2020)에서는 속성 노출 위험도를 공격자의 분류문제(attacker's classification problem)로 해석하고, 그에 따른 다양한 지표들을 설명 및 제시하였다. 본 논문에서는 Markus 등 (2020)에서 설명한 지표들의 기초가 되는 correct attribution probability (CAP)에 대해서 설명한다.

CAP은 공격자가 원본자료의 일부 변수를 가지고 있으며, 하나의 특정 변수의 값에 대하여 알고자하는 상황에서 계산된다 (Elliot, 2015). 공격자가 가지고 있는 변수들을 K(key variables), 알고자하는 변수를 T(target variable)로 각각 정의한다. 이 때, 공격자가 원본자료의 key variables K_o와 재현자료를 통해 원본자료의 target variable T_o의 예측값을 얻고자 하는 상황이 앞서 말한 공격자의 분류문제로 정의된다 (Markus 등, 2020). CAP은 K와 T 전부 범주형 변수인 경우 계산이 가능한 지표로 해당 논문 역시 동일한 가정을 기반으로 CAP에 대하여 설명한다.

CAP 값은 원본자료의 각 관측치에 대해 계산된다. 재현자료의 key variables과 target variable을 각각 K_s와 T_s로 정의하면, 원본자료의 i 번째 관측치에 대한 CAP 값 $\text{CAP}_{s,i}$ 는 다음과 같이 재현자료를 이용하여 구할 수 있다:

$$\text{CAP}_{s,i} = \frac{\sum_{i'=1}^{n_s} I\left((T_{s,i'} = T_{o,i}) \bigcap (K_{s,i'} = K_{o,i})\right)}{\sum_{i'=1}^{n_s} I(K_{s,i'} = K_{o,i})}.$$

즉, $\text{CAP}_{s,i}$ 는 원본자료의 i 번째 관측치의 key value $K_{o,i}$ 와 동일한 K를 가지는 재현자료 관측치들 중에 T 역시 원본자료의 i 번째 관측치의 target value $T_{o,i}$와 같은 관측치의 비율이다. 각 $\text{CAP}_{s,i}$ 값은 $K_{o,i}$ 와 동일한 K를 가지는 재현자료의 T값들의 분산이 클수록 작으며, 이는 원본자료에 비해 재현자료에서 K변수들로부터 T 값을 예측하기 어려울수록 $\text{CAP}_{s,i}$ 값들은 작은 값을 가짐을 뜻한다. 이 때, 재현자료 내에 원본자료의 i 번째

관측치의 key value $K_{o,i}$와 동일한 K를 가지는 관측치가 존재하지 않을 수 있다. Elliot (2015)에서는 해당 값을 0으로 지정하거나 무시한 뒤 $\text{CAP}_{s,i}$들의 평균 값을 속성 노출 위험도의 지표로 사용한다.

CAP 지표는 다음과 같이 원본자료 내에서도 구할 수 있다:

$$\text{CAP}_{o,i} = \frac{\sum_{i'=1}^{n_o} I\left(\left(T_{o,i'} = T_{o,j=i}\right) \cap \left(K_{o,i'} = K_{o,i}\right)\right)}{\sum_{i'=1}^{n_o} I\left(K_{o,i'} = K_{o,i}\right)}.$$

여러가지 재현자료들에서 구한 CAP 값들의 평균 $\mathbb{E}(\text{CAP}_s)$들과 원본자료에서 구한 CAP 값의 평균 $\mathbb{E}(\text{CAP}_o)$ 값을 비교하여 각 재현자료들에 대한 속성 노출 위험도에 대하여 판단할 수 있게 된다. 물론 CAP 지표 역시 자료에서 구한 하나의 통계량이기 때문에 원본자료와 차이가 큰 재현자료의 경우 유용성 측면에서 좋지 않은 재현자료일 수 있다. 하지만, 만약 동일한 유용성을 가지는 재현자료들이 있다면 $\mathbb{E}(\text{CAP}_s)$의 비교를 통해 그 중에서 속성 노출 위험도가 가장 낮은 재현자료에 대해 판단할 수 있다.

4장에서는 CAP을 활용하여 각 재현자료에 대한 속성 노출 위험도를 제시한다. CAP 지표는 범주형 자료에 대해서만 정의되었기에, 재현자료 내의 각 연속형 변수들에 대하여 일변량 k-means를 실시하여 연속형 변수들을 범주형 변수로 전환한 뒤 속성 노출 위험도를 측정한다. 또한, T와 K 역시 설정할 필요가 있다. 우선, 변수 중 가장 민감한 정보로 볼 수 있는 SUMMAT_CD(매출 금액)을 T로 설정하고 다른 변수 전부를 K로 설정한다. 이렇게 구한 CAP 은 실제 연속변수를 그대로 사용하거나 특정 변수들만 부분적으로 K로 설정하는 경우에 비해 보수적으로 구하게 되어 실제 위험도 보다는 높은 값이 나오게 됨을 짐작할 수 있다.

3.3. α-정밀도, β-재현율, 독창성 점수

Alaa 등 (2022)은 재현자료를 평가하기 위한 세 가지 지표로서 α-정밀도(precision), β-재현율(recall), 그리고 독창성 점수(authenticity score)를 제시한다.

1. α-정밀도는 재현자료가 원본자료를 얼마나 충실하게 재현하는가에 대한 평가 지표로서, 예를 들어 α-정밀도가 높은 재현자료는 현실성이 높은 관측치들을 포함한다.
2. β-재현율은 재현자료가 원본자료의 다양성을 충분히 반영하는가에 대한 평가 지표로서, 예를 들어 β-재현율이 낮은 재현자료는 원본자료의 일부만을 반복적으로 재현한 것으로 이해할 수 있다.
3. 독창성 점수는 재현자료가 얼마나 원본자료에 존재하지 않는 새로운 관측치들을 만들어 내는가에 대한 평가 지표이다. 이는 재현자료가 원본자료를 과적합하여 관측치들을 그대로 사용하고 있는가를 평가하기 위해 제시되었다.

여기서, α-정밀도와 β-재현율은 재현자료의 유용성을 측정하는 지표로, 독창성 점수는 정보노출의 위험성을 측정하는 지표로 이해할 수 있다.

3.3.1. α-정밀도와 β-재현율

원본 데이터 $\mathcal{D}_0$의 확률 분포의 서포트 안에서 α 만큼의 확률을 가지는 가장 작은 부분집합(α-support)을 $\mathcal{S}_o^\alpha$라고 한다면, α-정밀도 P_α는

$$P_\alpha := \Pr\left(x_s \in \mathcal{S}_o^\alpha\right), \text{ for } \alpha \in [0, 1]$$

로 정의되어, 재현자료의 관측치(x_s)가 원본자료 분포의 α-support에 포함될 확률을 의미한다. 정의상 α-support는 해당 분포의 최빈값(mode) 근처에서 형성되기 때문에, 재현자료가 원본자료의 분포에서 나타날 가능성을 계산하는 것으로 이해할 수 있다.

유사하게 β-재현율 R_β는 P_α 와는 대칭적으로

$$R_\beta := \Pr\left(x_o \in \mathcal{S}_s^\beta\right), \text{ for } \beta \in [0,1]$$

로 정의되어, 원본자료의 관측치(x_o)가 재현자료 분포의 β-support ($\mathcal{S}_s^\beta$)에 포함될 확률을 의미한다. 이는 재현자료의 분포가 원본자료를 얼마나 포함하는지에 대한 것으로 이해할 수 있다.

정의에서 알 수 있듯이 위 두 측도는 α와 β의 값에 따라 α-정밀도 곡선과 β-재현율 곡선을 형성한다. 이때, 성공적인 재현이 이루어졌다면 두 개 곡선 모두 45도선을 형성해야한다는 정리를 이용하여, P_α와 R_β의 평균절대편차(mean absolute deviation)를

$$\triangle P_\alpha := \int_0^1 |P_\alpha - \alpha|\, d\alpha, \quad \triangle R_\beta := \int_0^1 \left|R_\beta - \beta\right| d\beta,$$

로 각각 정의할 수 있다. 해당 값들은 0에서 1/2 사이의 값을 가지며 0에 가까울 수록 재현이 올바르게 되었다는 것을 의미한다. 최종적으로, 용이한 해석을 위해 통합 α-정밀도와 통합 β-재현율을 각각 $IP_\alpha = 1 - 2\triangle P_\alpha$, $IR_\beta = 1 - 2\triangle R_\beta$ 로 정의하여 0과 1사이 값을 가지며 1에 가까울수록 유용한 재현을 의미하는 통합 지표로 사용한다.

하지만, 원본자료와 재현자료에서 직접 토대를 정확히 추정하는 것은 매우 어렵다. 따라서, Alaa 등 (2022)에서는 One-class SVM (Schölkopf 등, 2001) 기법에 착안하여 원본자료와 재현자료를 고차원의 구체(hypersphere) 형태의 서포트를 가지도록 embedding하여 잠재공간에서 α-정밀도와 β-재현율을 계산한다.

3.3.2. 독창성 점수

독창성 점수 $A \in [0,1]$는 재현자료의 확률 분포($\mathbb{P}_s$)를 원본자료와 상이한 값들의 분포와 원본자료와 거의 흡사한 값들의 분포가 혼합된 형태로 보는 관점에서 제시되었다. 즉

$$\mathbb{P}_s = A \cdot \mathbb{P}'_s + (1 - A) \cdot \delta_{o,\epsilon}, \tag{3.4}$$

이며, 여기서 $\mathbb{P}'_s$는 원본자료와 다른 재현자료의 관측치만으로 구성된 분포를 의미하고, $\delta_{o,\epsilon}$은 원본자료의 경험 분포(empirical distribution)에 가우시안 잡음(Gaussian noise) $\mathcal{N}(0, \epsilon^2)$을 더한 분포를 의미한다. 즉, 재현자료의 분포($\mathbb{P}_s$)를 완전히 새로운 분포와 원본자료의 분포의 혼합 분포로 생각하고 그 계수를 A라고 생각하는 것으로 이해할 수 있다.

하지만, 실제 데이터에서 식 (3.4)의 A를 추정하는 것은 불가능에 가깝기에, Alaa 등 (2022)에서는 거리 기반의 통계량 a_j를 이용한다:

$$a_j = 1\left\{d_{s,j} \le d_{o,i^*}\right\},$$

여기서 $d_{s,j}$는 j 번째 재현자료 관측치(y_j)와 가장 가까운 원본자료의 관측치(x_{i^*})간의 거리를, d_{o,i^*}는 x_{i^*}와 원본자료에서 자기 자신을 제외한 가장 가까운 관측치간의 거리를 의미한다. 즉, a_j는 j 번째 재현자료 관측치가 특정 원본자료에 다른 원본자료 관측치보다도 더 가깝게 존재하는지를 나타내는 지시변수로, 이를 활용한 우도비율검정(likelihood ratio test)을 통해 독창성 점수를 추정한다. 자세한 추정 내용은 Alaa 등 (2022)에서 확인할 수 있다.

3.4. 평가 지표들의 특징과 차이점

지금까지 소개한 다양한 평가 지표들의 특징과 장단점을 간단히 정리한다.

pMSE는 대표적인 대역 유용성 지표로 원본자료와 재현자료를 분포적으로 구분할 수 있는가를 수치화한 지표이다. 모든 변수들의 분포를 개별적으로 비교하지 않고 변수들간의 관계성도 고려하여 유용성을 평가할

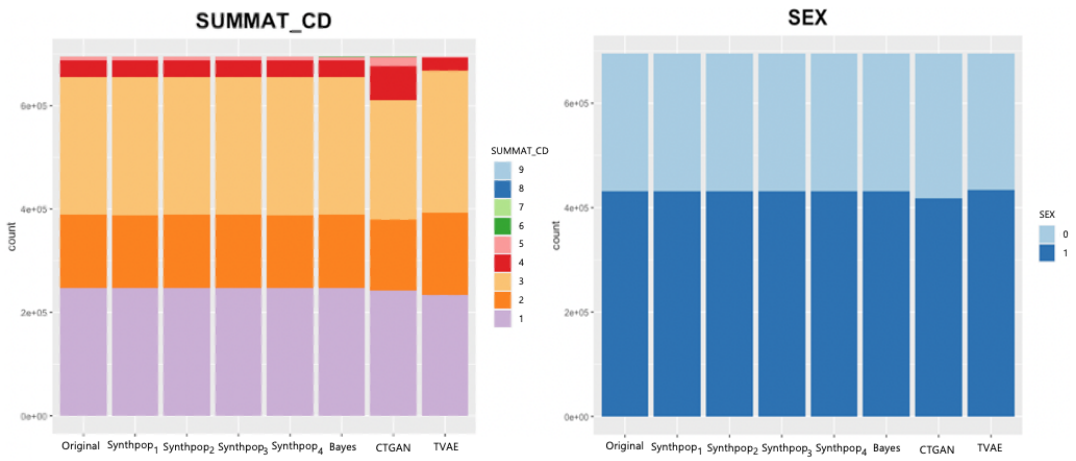

Figure 1: *Bar plots for variables SUMMAT_CD and SEX variables.*

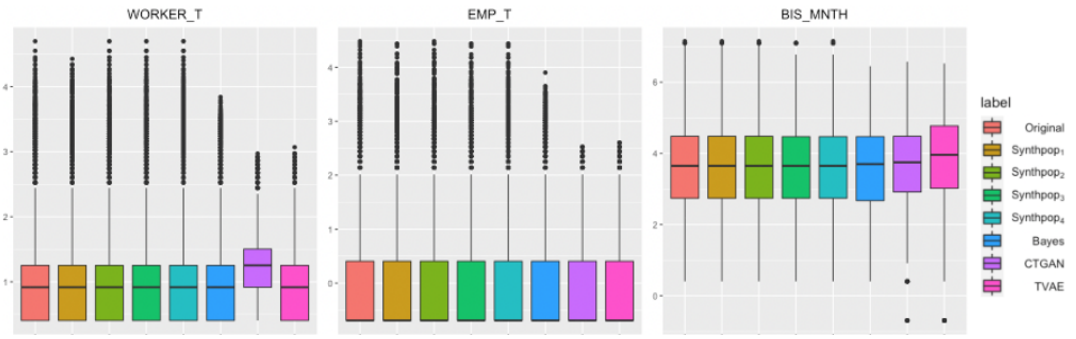

Figure 2: *Boxplots for variables WORKER_T, EMP_T, and BIS_MNTH on log scale.*

수 있다는 장점이 있다. 하지만, 재현자료 관측치의 크기, 재현자료와 원본자료를 구분하기 위해 사용되는 분류 모델에 따라 pMSE 값의 기준이 달라진다는 단점이 있기에, pMSE의 귀무 분포를 고려하여 유용성에 대한 판단을 할 필요가 있다. 또한, 재현자료의 비율에 따라 pMSE 값의 범위가 달라지므로, 재현자료 생성 방법을 비교할 경우 동일한 크기의 재현자료들에 대하여 비교하여야 한다.

분포간의 거리를 활용한 대역 유용성 지표는 pMSE와 다르게 원본자료와 재현자료에서 각 변수의 분포를 각각 직접적으로 쉽게 비교하여 유용성을 판단할 수 있다는 장점이 있다. 그러나, 변수간의 상관성을 고려하지 못한다는 단점이 있다.

Karr 등 (2006)에서 제시한 신뢰구간 중첩 지표는 특정 유용성 지표의 하나로 회귀분석 결과의 유사도를 평가하는 것이다. 이는 회귀분석이 재현자료를 얻는 목표 중 하나일 때 쉽게 비교가능하다는 장점이 있으나, 반응 변수 설정에 따라 결과가 달라지는 단점이 있다.

EI Emam 등 (2020)에서 제시한 신원 노출 위험도의 특징은 원본자료와 재현자료에서 준식별자와 민감변수의 값이 서로 같은 관측치들을 연결한다는 점에서 이해하기 쉬운 지표이다. 하지만 구현 시간이 오래 걸리고 변수를 준식별자와 민감변수로 구분할 때 이를 나누는 명확한 기준이 없어 사용자의 판단에 의존한다는 단점이 있다. 속성 노출 위험도를 측정하기 위해 제시된 CAP (Elliot, 2015) 역시 직관적으로 쉽게 이해가능한 노출

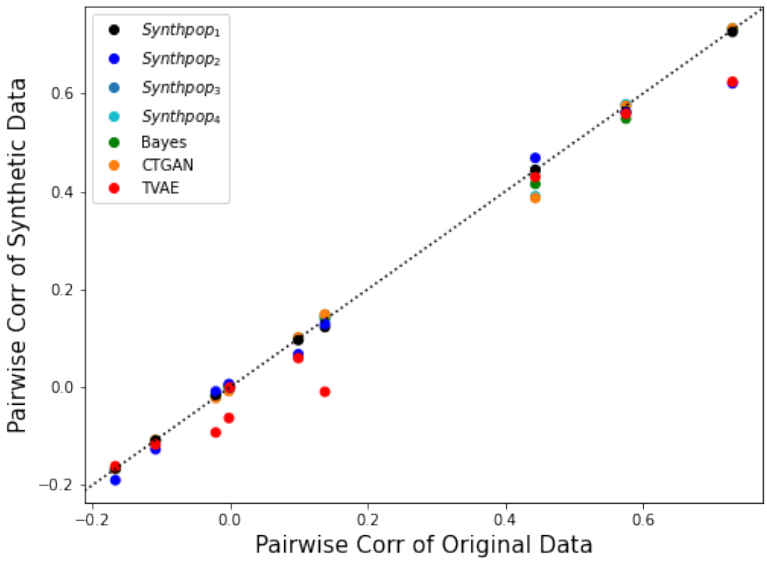

Figure 3: *Pairwise correlation comparison between the original and synthetic data sets. Points near the reference line indicates higher similarity to the original data.*

위험도 지표이다. 하지만, 하나의 변수가 노출되는 정도를 특정 유용성에 대한 지표로도 해석가능 하기에 데이터 사용 목적에 맞게 적절히 해석하고 사용할 필요가 있다. 또한, 데이터의 특성과 target variable의 설정에 따라 값이 상이할 수 있기에 항상 원본자료에서 구한 CAP 값을 함께 고려하여 재현자료의 속성 노출 위험도를 판단하여야한다. 이 지표는 현재 범주형 변수에 한정하여 사용할 수 있으므로 연속형 변수로의 적용 시 군집분석에 의한 범주화를 거쳐야 하는 단점이 있다. 따라서, 앞으로 연속형 변수에 대한 속성 노출 위험도에 대한 다양한 연구가 필요할 것으로 보인다.

Alaa 등 (2022)에서 제시한 α-정밀도, β-재현율, 그리고 독창성 점수의 가장 큰 특징은 잠재공간에서 원본자료와 재현자료의 토대를 추정하고 자료 사이의 거리를 구한다는 점에서 앞서 소개한 다른 평가 지표들과 확연히 차이가 난다. 해당 평가 지표들은 테이블 형태뿐만 아니라, 이미지 등 다양한 형태의 데이터의 재현에 대한 평가 지표로서 고안되었다는 점에서 기인한 특징으로 이해할 수 있다. 다만, 이로 인해 잠재공간으로 임베딩할 때의 hyperparameter 설정을 어떻게 하는가에 따라 결과가 다르게 나올 수 있다는 점과 계산량이 많다는 단점이 존재한다. 또한, 세 가지 지표 모두 데이터 전체가 아닌 각 관측치에 대한 평가 지표로서 고안되었다는 특징도 존재한다. 각 평가 지표의 특징을 살펴보자면, 대부분의 유용성 측도들은 원본자료와 재현자료의 분포가 얼마나 유사한가 혹은 분석결과가 얼마나 유사한가와 같이 하나의 기준만을 제시하는 반면 Alaa 등 (2022)은 유용성 측도를 충실도를 측정하는 α-정밀도와 다양성을 측정하는 β-재현율 두 개의 지표로 분리해서 제시한다. 따라서, 어느 차원에서 재현자료의 유용성이 부족한지 더 심도있는 분석이 가능하다. 마지막으로, 독창성 점수는 신원 노출 위험도와 속성 노출 위험도와 다르게 특정 변수의 설정을 고려하지 않고, 관측치간 거리를 기반으로 노출 위험도를 측정한다는 특징이 존재한다. 이는 재현자료 생성모형이 원본자료에 얼마나 과적합되었는가를 거리를 통해 측정하는 것으로, Hilprecht 등 (2019)에서 제시한 Monte Carlo 방법을 이용한 회원 추론 공격(membership inference attack)의 개념과 유사한 것으로 이해할 수 있다.

Table 5: Evaluation results of synthetic data sets based on utility and disclosure risk measures

Measure \ Data		$\mathcal{D}_O$	Synthpop_1	Synthpop_2	Synthpop_3	Synthpop_4	Bayes	CTGAN	TVAE
pMSE	Value (LR)	2e-6 (9.9e-7)	4.3e-7 (1.1e-7)	9.8e-7 (7e-7)	1.2e-6 (3.4e-7)	9.1e-7 (2.9e-7)	6.2e-6 (1.3e- 6)	3.1e-3 (3.7e-5)	2.1e-3 (3.9e-5)
	Ratio (LR)	1.08 (0.676)	0.513 (0.135)	1.194 (0.806)	1.442 (0.236)	1.169 (0.397)	7.39 (2.194)	3726.999 (150.54)	2628,764 (266.18)
	Standardized (LR)	0.115 (0.958)	-0.719 (0.2)	0.257 (1.161)	0.737 (0.465)	0.245 (0.599)	8.966 (3.497)	6164.843 (526.293)	4034.097 (363.172)
	Value (CART)	0.007 (9.9e-4)	0.002 (5.6e-4)	0.004 (9.4e-5)	0.005 (1e-4)	0.004 (1.4e-4)	0.018 (7e-5)	0.03 (1.2e-4)	0.024 (1.5e-4)
	Ratio (CART)	0.998 (0.14)	0.537 (0.118)	0.935 (0.015)	1.065 (0.035)	0.924 (0.038)	3.622 (0.062)	4.996 (0.6)	4.914 (0.099)
	Standardized (CART)	0.032 (0.868)	-2.819 (0.625)	-0.427 (0.119)	0.365 (0.176)	-0.457 (0.253)	13.902 (1.817)	22.198 (4.1)	25.061 (4.151)
거리기반	KLD	0.5e-3 (5.3e-5)	9.3e-5 (3.7e-6)	0.1e-3 (1.9e-5)	0.1e-3 (1.1e-5)	0.1e-3 (1.9e-5)	0.005 (0.1e-3)	0.097 (0.8e-3)	0.076 (0.8e-3)
	WD	1.474 (0.293)	0.970 (0.282)	0.915 (0.172)	0.984 (0.150)	1.051 (0.201)	3.537 (0.045)	5.270 (0.072)	18.001 (0.152)
신뢰구간 중첩 지표		0.720 (0.097)	0.785 (0.098)	0.303 (0.085)	0.556 (0.074)	0.364 (0.100)	0.387 (0.075)	-2.178 (0.065)	-12.098 (0.100)
노출 위험도	신원 노출 위험도	0.051 (2e-4)	0.056 (3e-4)	0.055 (3e-4)	0.055 (6e-4)	0.055 (6e-4)	0.049 (3e-4)	0.046 (3e-4)	0.050 (8e-5)
	속성 노출 위험도	0.446 (2e-4)	0.446 (1.6e-4)	0.442 (2.9e-4)	0.437 (3e-4)	0.442 (3.1e-4)	0.442 (1.7e-4)	0.419 (1.4e-4)	0.439 (2e-4)
Alaa 등 (2022)	α-정밀도	0.988 (0.3e-3)	0.988 (0.001)	0.978 (0.001)	0.987 (0.4e-3)	0.978 (0.001)	0.986 (0.4e-3)	0.953 (0.001)	0.974 (0.001)
	β-재현율	0.987 (0.4e-3)	0.987 (0.001)	0.983 (0.001)	0.986 (0.001)	0.983 (0.4e-3)	0.986 (0.4e-3)	0.978 (0.3e-3)	0.947 (0.002)
	독창성	0.354 (0.4e-3)	0.352 (0.001)	0.353 (0.001)	0.352 (0.001)	0.354 (0.001)	0.367 (0.001)	0.393 (0.001)	0.372 (0.001)

4. 재현자료 기법들 비교 분석

이 장에서는 SURVEY EST 데이터를 이용, 2장에서 소개된 순차회귀모형(Synthpop), 비모수 베이지안(Bayes), 인공신경망(CTGAN과 TVAE) 재현자료 기법들로 생성한 재현데이터들을 비교 분석한다. 순차회귀모형의 경우 2.2장에서 설명한 것처럼 네 가지 방법을 이용하여 Synthpop_1, Synthpop_2, Synthpop_3, Synthpop_4 재현자료를 생성하였다. 앞서 밝혔듯이 원본자료와 재현자료의 개수는 동일하며($n_o = n_s$) 각각의 방법당 5부의 재현자료를 생성하였다.

3장에서 소개된 평가 지표를 이용하여 비교하기 앞서 Figures 1–3에서는 각 재현자료에 대한 탐색적 자료분석을 실시하고 원본자료와 비교하였다. 각 방법내에서 생성된 재현자료들 5부끼리는 결과가 유사하였으며 여기서는 각 방법 당 하나의 재현데이터셋의 결과만 보고한다. Figure 1은 범주형 변수인 SUMMAT_CD와 SEX에 대한 성분막대도표를 나타낸다. CTGAN을 제외한 모든 방법들이 원본자료와 비슷한 비율구성을 보임을 알 수 있다. 도표에서 확인할 수는 없으나 TVAE에서는 SUMMAT_CD의 7, 8, 9의 값이, CTGAN은 9의 값이 생성되지 않았다. Figure 2에서는 연속형 변수인 WORKER_T, EMP_T, BIS_MNTH 의 상자그림을 나타내고 있다. 각 변수는 0을 포함하며 매우 비대칭한 분포를 가지기에 모든 관측치에 0.5를 더한 뒤 log를

취하였다. 그림에서 알 수 있듯이 순차회귀모형으로 생성한 자료들이 비모수 베이지안이나 인공 신경망 방법으로 생성된 자료들보다 원본자료와 좀 더 비슷한 양상을 보였다. 특히 상자 그림의 3분위수를 벗어난 값들의 분포에서 많은 차이를 보였으며, 이를 통해 순차회귀모형이 원본자료의 큰 값들 생성에 강점을 보이는 것을 확인할 수 있었다.

Figure 3에서 쌍별 상관계수(pairwise correlation coefficient) 계산은 Khamis (2008)를 참고하여 변수 쌍의 종류에 따라 다른 상관계수 방법을 사용했다. 연속형 변수들 쌍에 대해서는 피어슨의 상관계수(Pearson's coefficient of correlation), 명목형과 연속형 변수 쌍에 대해서는 점이연상관계수(point-biserial correlation coefficient), 순서형과 연속형 변수 쌍에 대해서는 스피어만의 순위상관계수(Spearman's rank correlation coefficient), 그리고 명목형과 순서형 변수 쌍에 대해서는 순위이연상관계수(rank-biserial correlation coefficient)를 사용하였다. 원본자료에서 구한 쌍별 상관계수와 각 재현자료에서 구한 쌍별 상관계수를 기준선인 45도 선에 대해 비교한 결과이다. Synthpop_1과 비모수 베이지안 방법이 기준선과 가장 가까웠으며, 그 다음으로는 Synthpop_2, Synthpop_3, Synthpop_4, 그리고 CTGAN, TVAE 순으로 기준선과 가까운 양상을 보였다.

Table 5는 각 재현방법을 이용해 생성된 5부 재현자료셋에서 얻어진 평가지표의 평균과 표준편차(괄호 안의 값)를 보여주고 있다. 여기서, $\mathcal{D}_0$는 원본자료를 단순임의표집을 통해 두 개의 데이터셋으로 나눈 후, 원본자료와 재현자료의 역할을 수행했을 때의 평가 결과로, 재현이 성공적으로 이루어졌을 경우에 대한 기준점으로 제시되었다. 이 또한 5번 반복되어 평균과 표준편차를 구하였다.

3.1.1장의 식 (3.1)에 따르면 원본자료와 재현자료의 비율이 동일한 상황에서 pMSE의 최대값은 0.25이므로 pMSE 값(value)만으로 비교했을 때에는 모든 방법의 pMSE 값들이 매우 작은 차이를 가진다. 하지만 식 (3.2)에 정의된 pMSE-ratio 와 standardized pMSE의 관점에서는 순차적 회귀모형을 이용한 재현방법들이 원본자료를 통한 기준값과 가장 가까운 양상을 보였다. 순차적 회귀모형을 이용한 재현방법들 다음으로는 비모수 베이지안, TVAE, CTGAN을 이용한 재현방법 순으로 유용성이 높다고 평가할 수 있다. 이러한 경향은 로지스틱 회귀모형(LR)과 의사결정나무(CART) 두 경우 모두에서 동일하다. 3.1.1장에서 설명한 pMSE의 귀무 분포의 예시는 Appendix에 실려있다.

3.1.2장에서 다룬 KL 괴리도(KLD)와 Wasserstein 거리(WD)에 따른 평가의 경우 두 개의 거리 지표가 거의 일관된 평가 결과를 보임을 알 수 있다. 순차적 회귀모형을 이용한 재현방법이 기준값에 가장 가까우며, 그 뒤로는 비모수 베이지안과 인공 신경망을 이용한 재현방법 순의 결과를 보여주었다.

3.1.3장에서 다룬 신뢰구간 중첩 지표의 경우 반응변수를 영업개월 수를 뜻하는 'BIS_MNTH'로 설정하고 나머지 변수들을 설명변수로 정한 후 분석했다. Table 5에서 신뢰구간 중첩 지표 부분의 결과를 기준점 값과 비교해보면 Synthpop_1 방법이 제시된 재현방법들 중 가장 효과적임을 알 수 있다. 그 뒤 차례로 Synthpop_3, 비모수 베이지안 방법, Synthpop_4, Synthpop_2, 그리고 인공 신경망을 이용한 방법인 CTGAN, TVAE 순으로 성능이 좋음을 알 수 있다.

3.2.1장에서 다룬 신원 노출 위험도의 경우 매출 금액을 뜻하는 'SUMMAT_CD'를 민감변수로 설정하고 나머지 변수들을 준식별자로 정한 후 평가했다. EI Emam 등 (2020)에 따르면 민감변수의 유형이 연속형일 때 k-means를 이용해 범주화 시켜야 하지만 민감변수로 설정한 'SUMMAT_CD'의 유형이 범주형이기 때문에 특별한 처리 없이 그대로 사용하였다. 그리고, 준식별자의 유형이 연속형일 때 고유한 값(unique value)의 종류가 많아지기 때문에, 임의의 원본자료의 관측치에 대해 준식별자의 값이 동일한 재현자료의 관측치를 찾는데 계산 시간이 오래 걸린다. 따라서, 60만개가 넘는 관측치를 가진 SURVEY EST 데이터 전체를 평가하는 데에 어려움이 있어 원본자료와 재현자료에서 각각 50,000개의 관측치를 임의로 추출하고 평가했다. 다음으로, 3.2.2장에서 다룬 CAP 지표를 구하기 위해 우선 연속형 변수들을 각각 범주화하였다. 원본자료의 각 연속형 변수들에 k-means++ (Arthur와 Vassilvitskii, 2007)를 실시하였고, elbow method를 바탕으로 군집의 수를 결정하였다. 그리고, 이를 기준으로 각 재현자료의 연속형 변수들 역시 범주화하였다. CAP 값을 구하기 위해서 앞서 설명했듯이 가장 민감한 정보로 볼 수 있는 SUMMAT_CD(매출 금액)을 target variable로 설정하였으며,

나머지 변수들을 key variables로 설정하였다. $\mathcal{D}_o$에서 제시된 CAP 값은 다른 지표들과 동일하게 원본 자료를 무작위로 분할하여 임의의 원본자료와 재현자료로 설정한 뒤 구한 값이다. 동일한 K를 가지지 않는 관측치를 제외하고 각 관측치에서의 CAP 값의 평균을 Table 5에 제시하였다.

노출 위험도의 결과를 살펴보면 여러 재현 방법들 중 인공 신경망을 이용한 CTGAN과 TVAE 방법에서 신원 노출 위험도와 속성 노출 위험도 모두 낮은 값을 가지고, 특히 CTGAN 방법을 사용했을 때 두 위험도에서 모두 가장 낮은 값을 가지는 것을 확인할 수 있다. 반면 순차적 회귀모형을 사용한 방법 중 하나인 Synthpop_1 방법을 사용했을 때 두 위험도에서 모두 가장 높은 값을 가지는 것을 확인할 수 있다. 3.1장에서 다룬 유용성 지표의 결과가 Synthpop_1 방법을 사용했을 때 성능이 가장 좋고 인공 신경망 방법을 사용했을 때 성능이 좋지 않았던 것을 고려한다면 노출 위험도와 유용성이 반비례한다는 사실을 확인할 수 있다.

3.3장에서 다룬 α-정밀도, β-재현율, 그리고 독창성 점수의 경우, 원본자료와 재현자료 사이의 거리를 구하는데 필요한 시간복잡도가 $O(n_0 \cdot n_s)$이기에 관측치가 60만개가 넘는 자료 전체를 평가하는 것은 매우 어렵다. 이에, 원본자료와 재현자료에서 각각 5,000개의 관측치를 임의로 추출하고 평가하는 과정을 100번 반복하여 그 평균으로 값을 추정하였다. 또한, support를 추정하기 위한 embedding 과정에서의 hyperparameter는 Alaa 등 (2022)에서 주어진 코드의 기본값으로 모두 고정하였다. α-정밀도와 β-재현율은 $\mathcal{D}_0$에서 0.98의 값을 보이며 가장 이상적인 재현의 경우인 1에 매우 가까운 값을 보였으며, 순차적 회귀모형과 비모수 베이지안 방법을 이용한 재현자료 모두 이에 근접한 수치를 보여주었다. 하지만, 인공 신경망을 이용한 CTGAN과 TVAE 재현자료의 경우 위 재현방법들에 비해 비교적 좋지 않은 수치를 보여주었다. 독창성 점수의 경우 순차적 화귀모형과 비모수 베이지안 방법은 기준값에 근접한 반면, 인공신경망을 이용한 방법의 경우 오히려 기준값을 상회하는 값을 보여주었다. Appendix에는 α-정밀도 (P_α)와 β-재현율 (R_β) 곡선의 예시가 실려있다.

5. 결론

본 논문에서는 순차적 회귀모형, 비모수 베이지안, 인공 신경망 재현 방법들을 여러 가지 유용성과 노출 위험성 지표들을 이용하여 비교하였다. 4장에서 볼 수 있듯이 분석한 SURVEY EST 데이터에 대해서는 전체적으로 순차적 회귀모형, 비모수 베이지안, 인공 신경망 재현 방법 순으로 유용성 지표들의 값이 제시된 기준값에 가까웠다. 그러나, 노출 위험도에서는 반대로 인공신경망 재현 방법들이 위험도가 낮고 순차적 회귀모형이 높은 경향을 보였다. 이런 관점에서 보면 비모수 베이지안 방법이 두 지표의 균형을 이룬 재현자료를 생성한다고 볼 수 있다. 하지만, 본 연구에서 구한 결과는 사례연구에 해당되므로 좀 더 일반적인 비교 결과를 얻기위해 다양한 형태와 크기의 데이터들에 적용해 볼 필요가 있다. 또한, 각 재현 방법들이 여러 옵션을 달리 함에 따라 재현 결과가 달라질 수 있기 때문에 이에 대한 세심한 연구가 필요하다. 인공 신경망 방법의 경우, diffusion model이나 score-based generative model처럼 성능이 향상된 방법들이 계속적으로 개발되고 있어 앞으로 이들을 이용한 재현자료 생성을 고려할 필요가 있다 (Dhariwal과 Nichol, 2021; Song과 Ermon, 2019; Song 등, 2020).

Appendix

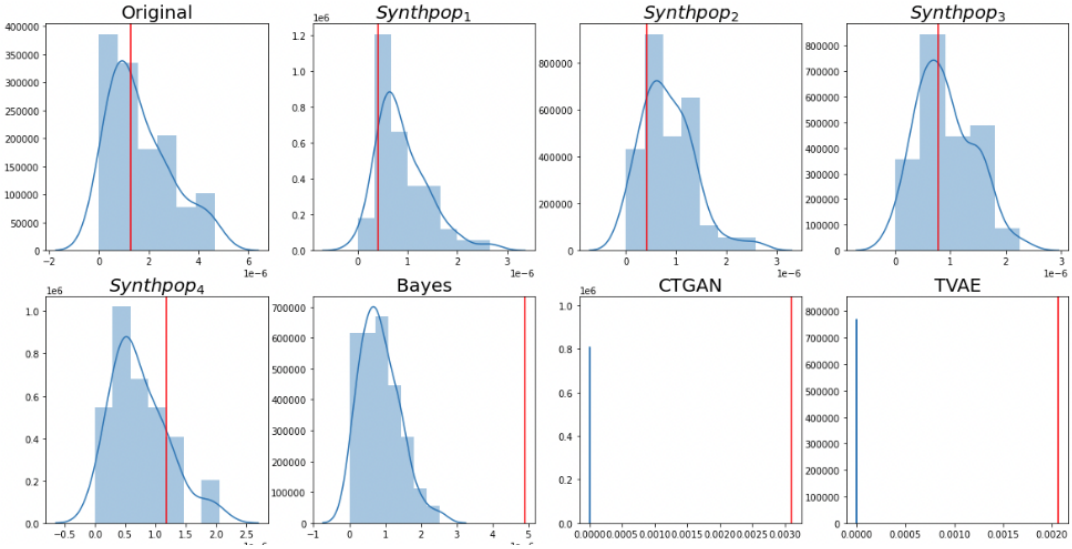

Figure 4: *Null distribution of pMSE based on logistic regression with observed pMSE (red line). They are drawn using the first synthetic datasets.*

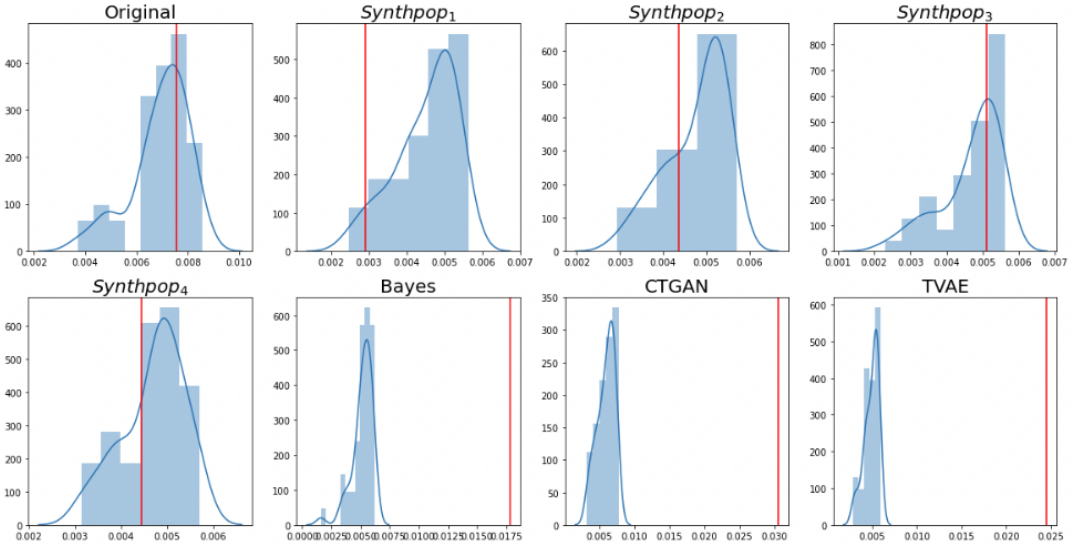

Figure 5: *Null distribution of pMSE based on CART with observed pMSE (red line). They are drawn using the first synthetic datasets.*

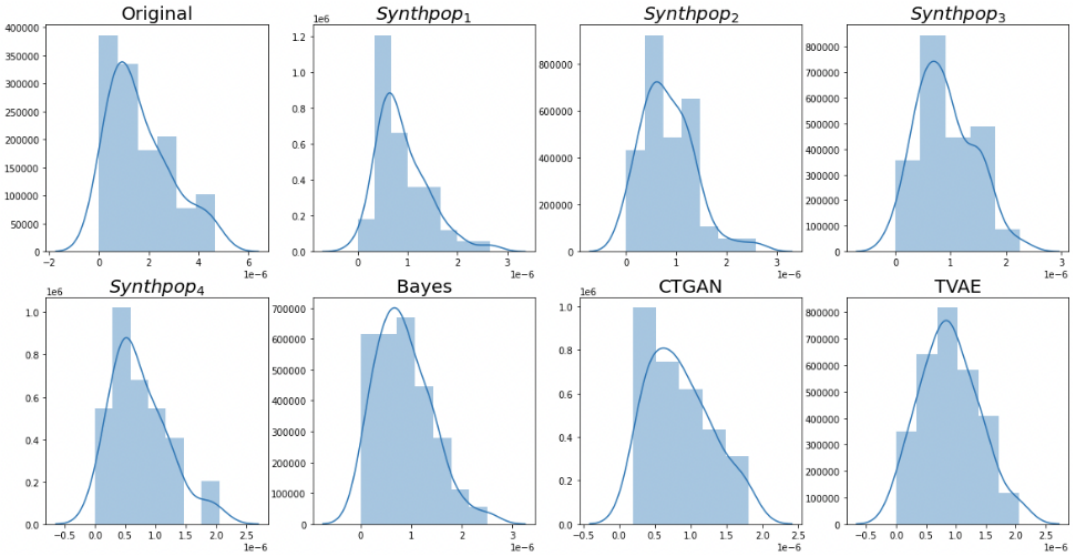

Figure 6: *Null distribution of pMSE based on logistic regression without observed pMSE. They are drawn using the first synthetic datasets.*

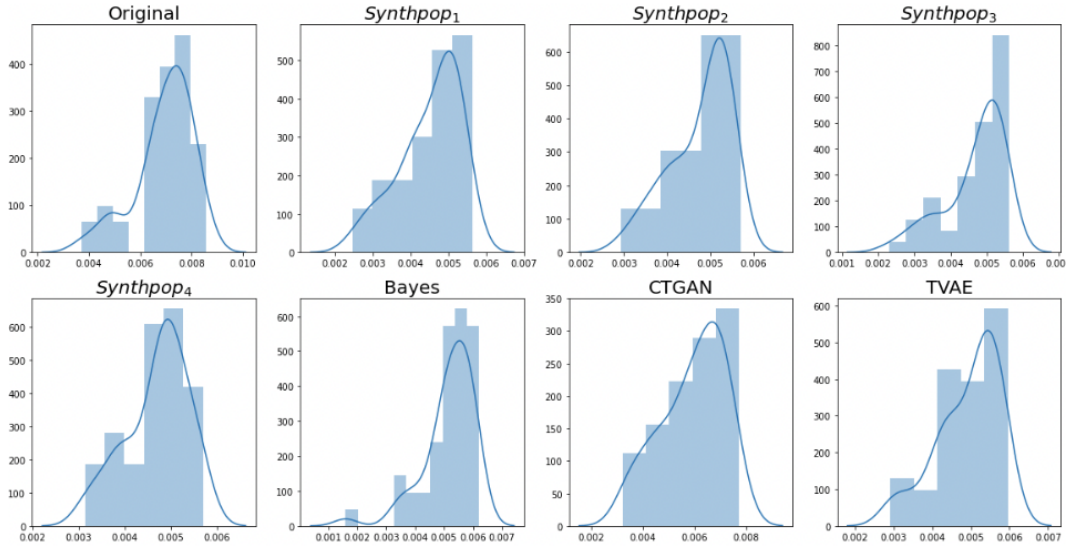

Figure 7: *Null distribution of pMSE based on CART without observed pMSE. They are drawn using the first synthetic datasets.*

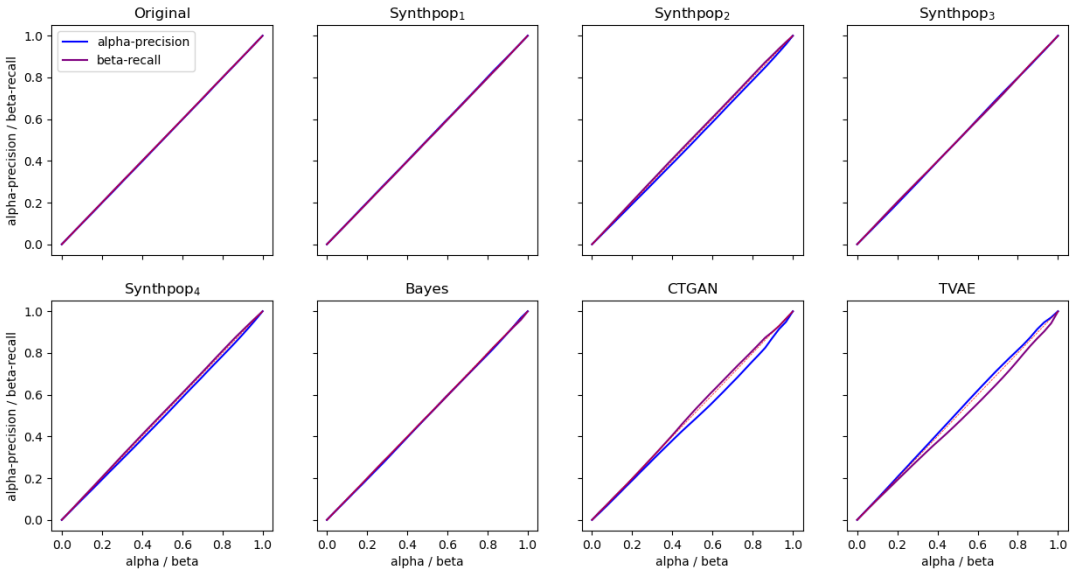

Figure 8: *α-precision (P_α) and β-recall (R_β) curves for all synthetic data methods. They are drawn using the first synthetic datasets.*

References

Alaa A, Van Breugel B, Saveliev ES, and van der Schaar M (2022). How faithful is your synthetic data? Sample-level metrics for evaluating and auditing generative models, *International Conference on Machine Learning*, 290–306, PMLR.

Arjovsky M, Chintala S, and Bottou L (2017). Wasserstein generative adversarial networks, *International Conference on Machine Learning*, 214–223, PMLR.

Arthur D and Vassilvitskii S (2007) K-means plus plus: The advantages of careful seeding, In *Proceedings of the Eighteenth Annual Acm-Siam Symposium on Discrete Algorithms*, New Orleans, Louisiana, USA, 1027–1035.

Breiman L, Friedman JH, Olshen RA, and Stone CJ (2017). *Classification and Regression Trees*, Routledge, New York.

Dhariwal P and Nichol A (2021). Diffusion models beat gans on image synthesis, *Advances in Neural Information Processing Systems*, **34**, 8780–8794.

Drechsler J and Reiter JP (2009). Disclosure risk and data utility for partially synthetic data: An empirical study using the german iab establishment survey, *Journal of Official Statistics*, **25**, 589–603.

EI Emam K, Mosquera L, and Bass J (2020). Evaluating identity disclosure risk in fully synthetic health data: Model development and validation, *Journal of Medical Internet Research*, **22**, e23139.

Elliot M (2015). Final report on the disclosure risk associated with the synthetic data produced by the sylls team, *Report 2015*, 2.

Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, and Courville AC (2017). Improved training of Wasserstein GANs, *Advances in Neural Information Processing Systems*, **30**, 1–11.

Hilprecht B, Härterich M, and Bernau D (2019). Monte carlo and reconstruction membership inference attacks against generative models, *Proceedings on Privacy Enhancing Technologies*, **2019**, 232–249.

Hu J and Savitsky TD (2018). Bayesian data synthesis and disclosure risk quantification: An application to the consumer expenditure surveys, Available from: *arXiv preprint arXiv:1809.10074*

Ishwaran H and James LF (2001). Gibbs sampling methods for stick-breaking priors, *Journal of the American Statistical Association*, **96**, 161–173.

Karr AF, Kohnen CN, Oganian A, Reiter JP, and Sanil AP (2006). A framework for evaluating the utility of data altered to protect confidentiality, *The American Statistician*, **60**, 224–232.

Khamis H (2008). Measures of association: How to choose?, *Journal of Diagnostic Medical Sonography*, **24**, 155–162.

Kingma DP and Welling M (2013). Auto-encoding variational Bayes, Available from: *arXiv preprint arXiv:1312.6 114*

Kim HJ, Drechsler J, and Thompson KJ(2021). Synthetic microdata for establishment surveys under informative sampling, *Journal of the Royal Statistical Society: Series A*, **184**, 255–281.

Kim J and Park M-J (2019). Multiple imputation and synthetic data, *The Korean Journal of Applied Statistics*, **32**, 83–97.

Kullback S and Leibler RA (1951). On information and sufficiency, *The Annals of Mathematical Statistics*, **22**, 79–86.

Lee Y (2013). Review on statistical methods for protecting privacy and measuring risk of disclosure when releasing information for public use, *Journal of the Korean Data and Information Science Society*, **24**, 1029–1041.

Lin Z, Khetan A, Fanti G, and Oh S (2018). The power of two samples in generative adversarial networks, *Advances in Neural Information Processing Systems*, **31**, 1–10.

Little RJA (1993). Statistical analysis of masked data, *Journal of Official Statistics, Stockholm*, **9**, 407–407.

Markus H, Rudolf M, and Andreas E (2020). A baseline for attribute disclosure risk in synthetic data, In *Proceedings of the Tenth ACM Conference on Data and Application Security and Privacy (CODASPY'20), March 16–18, 2020, New Orleans, LA, USA*, ACM, New York, NY, USA, 11, Available from: https://doi.org/10.1145/3374664.3375722

Murray JS and Reiter JP (2016). Multiple imputation of missing categorical and continuous values via bayesian mixture models with local dependence, *Journal of the American Statistical Association*, **111**, 1466–1479.

Nowok B, Raab GM, and Dibben C (2016). Synthpop: Bespoke creation of synthetic data in R, *Journal of Statistical Software*, **74**, 1–26.

Park MJ, Kwon SP, and Shim KH (2013). Microdata masking for Survey of Household Finances and Living Conditions, *Statistical Research Institute, Daejeon*.

Park M-J, Han J, and Park N (2020). Study on synthetic data generation methods with applications to statistics Korea RDC data, *Technical report, Statistical Research Institute*.

Raghunathan TE, Reiter JP, and Rubin DB (2003). Multiple imputation for statistical disclosure limitation, *Journal of Official Statistics*, **19**, 1–16.

Reiter JP (2003). Inference for partially synthetic, public use microdata sets, *Survey Methodology*, **29**, 181–188.

Reiter JP (2005). Using CART to generate partially synthetic public use microdata, *Journal of Official Statistics*, **21**, 441–462.

Rosenbaum PR and Rubin DB (1983). The central role of the propensity score in observational studies for causal effects, *Biometrika*, **70**, 41–55.

Rubin DB (1993). Statistical disclosure limitation, *Journal of Official Statistics*, **9**, 461–468.

Scholkopf B, Platt JC, Shawe-Taylor J, Smola AJ, and Williamson RC (2001). Estimating the support of a high-dimensional distribution, *Neural Computation*, **13**, 1443–1471.

Snoke J, Raab GM, Nowok B, Dibben C, and Slavkovic A (2018). General and specific utility measures for synthetic data, *Journal of the Royal Statistical Society: Series A*, **181**, 663–688.

Song Y and Ermon S (2019). Generative modeling by estimating gradients of the data distribution, *Advances in Neural Information Processing Systems*, **32**, 11895–11907.

Song Y, Sohl-Dickstein J, Kingma DP, Kumar A, Ermon S, and Poole B (2020). Score-based generative modeling through stochastic differential equations, *International Conference on Learning Representations*, Available from: https://arxiv.org/abs/2011.13456

Stan M, Jordi N, Morvarid S, and Tomasz S (2015). A review of attribute disclosure control, *Advanced Research in Data Privacy*, **567**, 41–61.

Villani C (2008). *Optimal Transport: Old and New*, Springer, New York.

Woo M-J, Reiter JP, Oganian A, and Karr AF (2009). Global measures of data utility for microdata masked for disclosure limitation, *Journal of Privacy and Confidentiality*, **1**, 111–124.

Xu L, Skoularidou M, Cuesta-Infante A, and Veeramachaneni K (2019). Modeling tabular data using conditional GAN, *Advances in Neural Information Processing Systems*, **32**, 7333–7343.

Yoon J, Jarrett D, and Van der Schaar M (2019). Time-series generative adversarial networks, *Advances in Neural Information Processing Systems*, **32**, 5509–5519.

Received November 24, 2022; Revised January 11, 2023; Accepted January 12, 2023

유용성과 노출 위험성 지표를 이용한 재현자료 기법 비교 연구

안성빈[a], 트랑 도안[b], 이주희[c], 김지우[d], 김용재[e], 김윤지[a], 윤창원[a], 정성규[e], 김동하[d], 권성훈[b], 김항준[f], 안정연[1,a], 박철우[2,g]

[a]한국과학기술원 산업및시스템공학과; [b]건국대학교 응용통계학과; [c]경북대학교 통계학과; [d]성신여자대학교 통계학과; [e]서울대학교 통계학과; [f]신시내티 대학교 통계 데이터사이언스 분과; [g]한국과학기술원 수리과학과

요 약

재현자료를 생성하여 배포하는 것은 데이터 공개에 따른 정보 유출의 위험을 방지하는 대표적인 방법이다. 최근 산업에서 데이터의 활용이 중요해진 만큼 한국을 포함한 많은 국가 및 기관에서 재현자료에 관한 연구가 활발히 진행되고 있다. 본 논문에서는 대표적인 재현자료 생성 기법들과 평가 지표들을 소개한다. 전통적인 재현자료 생성 방법인 다중대체와 최근 제시된 인공신경망 기반의 재현자료 생성 방법 등을 활용하여 재현자료를 생성하는 과정을 기술함에 따라 재현자료 생성 방법에 대한 전반적인 이해를 돕는다. 이에 더해 다양한 재현자료 평가 지표를 바탕으로 생성된 재현자료들을 분석 및 비교함에 따라 앞으로의 연구에 대한 방향을 제시하고 그에 대한 토대를 마련하고자 한다.

주요용어: 노출 위험성, 비모수 베이지안, 순차적 회귀모형, 심층 생성 모형, 유용성, 재현자료

이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2022-0-00937, 통계데이터 재현자료기법의 활용성과 유용성을 높여야 하는 문제 해결)

[1]교신저자: (34141) 대전광역시 유성구 대학로 291, 한국과학기술원 산업및시스템공학과. jyahn@kaist.ac.kr

[2]교신저자: (34141) 대전광역시 유성구 대학로 291, 한국과학기술원 수리과학과. parkcw2021@kaist.ac.kr