

CTGAN을 이용한 재현자료 생성 및 평가 측도에 대한 비교 연구

김정훈

2023-12-06

목차

1. 서론
2. GAN의 간단한 소개
3. CTGAN의 간단한 소개
4. 자료 설명
5. 재현자료 평가
6. 결론 및 향후 연구

1. 서론

1. 서론

- 재현자료란 실제로 측정하지 않은 임의의 데이터로 넓게 정의하기도 하지만, 통상적으로 추정된 모형에서 생성된 가상의 데이터를 의미한다.
- 실제로 공공기관에서 제공하는 데이터들은 재현자료일 경우가 많다. 보안성의 이유 때문에 재현자료는 필수적임
- ① **완전 재현 자료**는 실제데이터가 하나도 없이 가상으로만 이루어진 데이터를 의미하며, 정보보호 측면에서 가장 강력한 보안성을 가진다.
- ② **부분 재현 자료**는 공개하려는 변수들 중 일부만 선택하여 재현자료로 대체한 데이터를 의미하며, 보통 재현자료로 대치되는 변수들은 민감한 정보에 관한 변수들이다.
- ③ **복합 재현자료**는 일부 변수들의 값을 재현자료로 생성하고, 생성된 재현자료와 실제데이터를 모두 이용하여 일부 변수 값들을 다시 도출하는 방법으로 생성한다.

1. 서론

- 재현자료의 평가의 목적으로는 그 자료가 사용될 목적과 환경에 따라 달라질 수 있지만, 가장 중요한 것은 재현 자료가 원본 데이터의 중요한 특성과 패턴을 잘 반영하며, 개인정보에 대한 보호가 제공되는지를 꼭 확인해야함.
- 본 연구의 목적으로는 재현자료 생성 방법중 하나인 CTGAN 모델을 이용하여 2016년 가계금융복지데이터를 이용하여 재현자료를 생성.
- CTGAN 모델을 이용하여서 다양한 변수들의 재현자료를 만들었을 때 발생한 문제점들을 해결할 것이며, 평가 방법들을 이용해 유용성 측면과 보안성 측면에서 확인해볼 것이다.

2. GAN의 간단한 소개

2. GAN의 간단한 소개

- GAN모델을 쉽게 설명하자면, 생성자를 위조지폐를 생산하는 위조지폐범(Generator)이라고 생각하고 판별자를 위조지폐를 구분하는 경찰(Discriminator)이라고 하자, 최종 목표는 생성자에서 만들어진 위조지폐를 경찰을 속일 수 있을 정도로 만드는 것이 GAN 모델의 최종목적.
- 실제 데이터에 유사한 데이터를 생성하는 것을 목표로하는 생성자 (Generator)와 실제 데이터와 생성된 데이터의 차이를 판별해주는 판별자(Discriminator)간의 상호 경쟁을 통해 실제데이터와 생성된 데이터간의 차이를 줄이는 딥러닝 모델

2. GAN의 간단한 소개

- $\min_G \max_D V(G, D) = E_{x \sim P_{\text{data}}(x)}[\log D(x)] + E_{z \sim P_z(z)}[\log(1 - D(G(z)))]$
- 생성자 입장에서는 생성자 모델이 잘 생성된다면, 최솟값은 마이너스 무한대 , 판별자 입장에서는 최댓값이 0 일때 잘 구분했다고 볼 수 있다. 일종의 생성자와 판별자간의 min-max게임을 실시하는 것이 GAN모델의 알고리즘의 원리

3. CTGAN의 간단한 소개

3. CTGAN의 간단한 소개

1. 모드별 정규화 (Mode-Specific Normalization)
2. 조건부 생성자(Conditional Generator)
 - 1) 조건 벡터(Conditional Vector)
 - 2) 생성자 손실 함수(Generator Loss Function)
 - 3) 샘플링 훈련 전략(Training-by-sampling)
3. 신경망 구조(Neural Network Structure)

2. CTGAN의 간단한 소개

1. 모드별 정규화 (Mode-Specific Normalization)

- 원래는 (최소 - 최대 정규화 : 최대값 1, 최소값 -1로 설정) 하는 min-max normalization 방법을 사용했지만, 수치형 데이터가 가우시안 분포가 아닌 다른 분포의 형태를 가질 때, 역전파 과정에서 Gradient Vanishing이 발생 할 수 있다.
- GAN 모델이 다중 모드를 잘 발견하지 못하는 문제 발생

위와 같은 문제를 해결하기 위해 일반적인 형태의 분포를 혼합 가우시안 형태로 나타내고 특정 레코드의 모드를 명시적으로 나타내도록 전처리한다. 위 과정을, **변분 가우시안 혼합 모델**이라고 말하며 이를 통해 정규화를 진행

2. CTGAN의 간단한 소개

변분 가우시안 혼합 모델

①. k개의 혼합분포를 찾은 다음에 가중치(μ_k)가 특정 임계치 ϵ 이상인 경우의 분포를 남긴다.

$$\Rightarrow P(C_{i,j}) = \sum_{k=1}^3 \mu_k N(c_{i,j}; \eta_k, \phi_k)$$

여기서 μ_k, η_k, ϕ_k 는 모드의 평균, 가중치 및 표준편차이다.

②. 특정값 $C_{i,j}$ 에 대해 각 가우시안 분포에서의 확률 $\mu_k N(c_{i,j}; \eta_k, \phi_k)$

$P(k) = \frac{\mu_k N(C_{i,j}; \eta_k, \phi_k)}{\sum_p \mu_p N(C_{i,j}; \eta_p, \phi_p)}$ 확률을 이용하여서 $C_{i,j}$ 가 어떤 가우시안 분포에서 나왔는지 선택

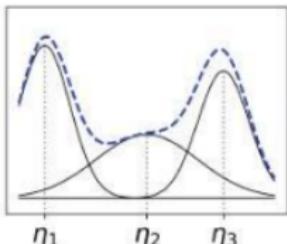
ex) $p_1, p_2, p_3 = [0.1, 0.3, 0.1]$ 이라고 했을때 1번이 $1/5 = 20\%$, 2번이 60% , 3번이 20% 로 선택될 확률임.

2. CTGAN의 간단한 소개

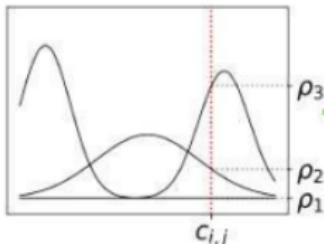
- ③. 위 단계에서 선택된 가우시안 분포의 평균과 표준편차를 이용, 특정값 $C_{i,j}$ 를 $\alpha_{i,j}$ 로 정규화, 어떤 정규분포가 선택되었는지 원핫벡터 $\beta_{i,j}$ 로 표시
- ④. j번째 레코드는 수치형 값 $C_{i,j}$ 와 범주형 값 원핫벡터 ($d_{i,j}$)로 구성된 형태에서 모드별 정규화를 이용. 수치형 데이터 ($\alpha_{i,j}, \beta_{i,j}$)와 범주형 데이터 원핫 벡터 ($d_{i,j}$)로 구성된 형태.

$$\Rightarrow r_j = \alpha_{i,j} \oplus \beta_{i,j} \oplus \dots \oplus d_{i,j} \oplus \dots d_{N_d,j}$$

Model the distribution of a continuous column with VGM.



For each value, compute the probability of each mode.



Sample a mode and normalize the value.

$$\alpha_{i,j} = \frac{c_{i,j} - \eta_3}{4\phi_3}$$
$$\beta_{i,j} = [0, 0, 1]$$

2. CTGAN의 간단한 소개

2. 조건부 생성자(Conditional Generator)

기존 GAN 모델에서의 범주형 자료에서 특정 클래스 희소성 문제를 해결하기 위해 사용하며 위 조건부 적대적 생성망을 이용해 재현하기 위해서는 3가지 조건을 만족해야한다

- 1) 조건 벡터
- 2) 생성자 손실 함수(Generator Loss Function)
- 3) 샘플링 훈련 전략(Training-by-sampling)

위 문제들을 해결하면 생성한 재현자료의 조건부 분포 $P_g(\text{row}|D_{i^*} = k)$ 는 실제 데이터 분포의 조건부 분포 $P(\text{row}|D_{i^*} = k)$ 와 같아지고 다음 식을 통해 실제 데이터 분포를 얻을 수 있게 된다.

$$P(\text{row}) = \sum_{k \in D_{i^*}} P_g(\text{row}|D_{i^*} = k_*) P(D_{i^*} = k_*)$$

2. CTGAN의 간단한 소개

① 조건 벡터

- 범주형 데이터 열에 대한 조건을 만들기 위해 원핫벡터 형태의 마스킹 벡터를 이용
- 범주형 열 중 하나의 열만 선택하고 그 열의 여러 클래스 중 하나의 클래스만 선택해서 조건으로 나타냄.
- 선택된 클래스 일 때 1로 , 나머지는 0으로 표현
- 조건벡터는 마스킹벡터를 모두 결합한 형태인 $m_1 \oplus \dots \oplus m_{N_d}$ (여기서 N_d 는 범주형 데이터 열의 갯수로 표현)

ex) 예를 들어 두개의 범주형 열에서 첫번째 열에는 1,2,3의 클래스, 두번째 열에는 1,2의 클래스가 존재하는 경우 2번째 열에서 1이 선택되었다면, $m_1 = [0, 0, 0]$, $m_2 = [1, 0]$ 으로 표현되고 조건벡터는 $[0,0,0,1,0]$ 로 표현 할 수 있다.

2. CTGAN의 간단한 소개

② 생성자 손실 함수

- 위의 예시처럼 [0,0,0,1,0] 조건이 들어왔다면, 첫번째 범주형 열은 아무 클래스나 들어가 생성해도 되지만 두번째 범주형 열은 1이라는 값이 생성되어야 한다.
- 즉, 생성된 두번째 범주형 열 데이터(\hat{d}_2)가 조건(m_2)과 같아야한다.
- 생성자가 조건과 같은 클래스를 만들도록 학습시키기 위해 생성자 손실 함수에 \hat{d}_2 와 m_2 의 크로스 엔트로피 손실을 추가한다.
- $CrossEntropy = \sum_{k=1}^{|D_i|} m_i^{(k)} \log \hat{d}_i^{(k)}$ 여기서 ($|D_i|$ 는 i번째 범주형 열의 클래스 갯수)

2. CTGAN의 간단한 소개

③ 샘플링 훈련 전략

- 생성자가 만든 재현자료의 조건부 분포 = 실제 데이터의 조건부 분포 → 식별자가 두 분포간의 거리를 정확하게 추정
- 모든 사용 가능한 조건벡터와 훈련 데이터를 사용해야만 제대로 된 학습이 가능하다. → 균일하게 조건 벡터와 훈련 데이터를 표본을 뽑는 전략을 통해 위 문제를 해결 가능.
- 구체적인 방법은 6단계로 나누어서 설명.

2. CTGAN의 간단한 소개

- ① 범주형 열 중 하나를 균등한 확률로 선택
- ② 선택된 범주형 열에서 각 값의 발생빈도를 통해 확률 분포를 만든다
이 때 각 빈도에는 로그함수를 취해준다.

ex) D_2 에서 1번째 클래스가 100번 발생하고, 두번째 클래스가 50번 발생했다면, 확률분포는

$$\left(\frac{\log 100}{\log 100 + \log 50}, \frac{\log 50}{\log 100 + \log 50} \right) = (0.54, 0.46)$$

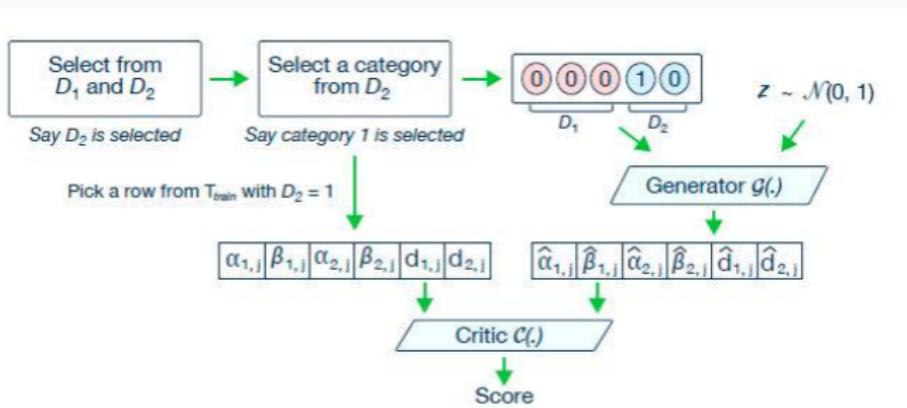
- ③ 2번에서 구한 확률분포에 따라 1개의 값을 선택, 밑의 그림을 봤을 때는 첫번째 클래스가 선택되었다.

- ④ 선택된 열과 클래스에 따라 조건 벡터를 만들고 훈련데이터를 랜덤추출한다.

ex) 위의 그림을 참조 했을 때 범주형 열 D_1 (클래스 3개), D_2 (클래스 2개) 중에서 D_2 의 첫번째 클래스가 선택되었으므로 $m_1 = [0,0,0]$, $m_2 = [1,0]$, 조건벡터는 $[0,0,0,1,0]$ 이다.

2. CTGAN의 간단한 소개

- ⑤ 생성자에 조건벡터와 변분 가우시안 분포에서 추출한 잠재변수를 입력변수로 넣어서 재현자료를 생성한다.
 - ⑥ 식별자에 실제 데이터와 조건벡터, 재현자료와 조건벡터를 각각 넣어서 나온 결과로 두 분포간의 거리(Score)를 계산해 식별자와 생성자를 업데이트한다. 식별자는 거리가 멀어지게, 생성자는 거리가 가까워지게 하는 방향으로 학습(GAN모델의 학습 방향성과 유사)



2. CTGAN의 간단한 소개

신경망 구조

① 생성자 신경망 구성

$$\begin{cases} h_0 = z \oplus cond \\ h_1 = h_0 \oplus \text{ReLU}\left(BN\left(FC_{|cond| + |z| \rightarrow 256}(h_0)\right)\right) \\ h_2 = h_1 \oplus \text{ReLU}\left(BN\left(FC_{|cond| + |z| + 256 \rightarrow 256}(h_1)\right)\right) \\ \hat{\alpha}_i = \tanh\left(FC_{|cond| + |z| + 512 \rightarrow 1}(h_2)\right) \quad 1 \leq i \leq N_c \\ \hat{\beta}_i = gumbel_{0.2}\left(FC_{|cond| + |z| + 512 \rightarrow m_i}(h_2)\right) \quad 1 \leq i \leq N_c \\ \hat{d}_i = gumbel_{0.2}\left(FC_{|cond| + |z| + 512 \rightarrow |D|}(h_2)\right) \end{cases}$$

2. CTGAN의 간단한 소개

신경망 구조

② 식별자 신경망 구성

$$\begin{cases} h_0 = r_1 \oplus \dots \oplus r_{10} \oplus cond_1 \oplus \dots \oplus cond_{10} \\ h_1 = drop\left(\text{leaky}_{0.2}(FC_{10|r| + 10|cond| \rightarrow 256}(h_0))\right) \\ h_2 = drop\left(\text{leaky}_{0.2}(FC_{256 \rightarrow 256}(h_1))\right) \\ C(\bullet) = FC_{256 \rightarrow 1}(h_2) \end{cases}$$

3. 자료 설명

3. 자료 설명

본 데이터는 2016 가계 금융 복지 조사의 가구주 자료이며, 총 레코드의 숫자는 18,273개의 레코드, 편의상 가중치는 고려하지 않았으며, 결측값이 존재하지 않는다.

- 범주형 변수 : 수도권 여부, 가구주 성별, 가구주 학력, 가구주 혼인상태, 가구주 종사지위
 - 수도권 여부(urban) : G1 (수도권), G2 (비수도권)
 - 가구주 성별(sex) : 1 (남자), 2 (여자)
 - 가구주 학력코드(edu) : 1 (안받음), 2 (초등학교), 3 (중학교), 4 (고등학교), 5 (대학 (3년제 이하)), 6 (대학교 (4년제이상)), 7 (대학원 이상)
 - 가구주 혼인상태(marital) : 1 (미혼), 2 (배우자 있음), 3 (사별), 4 (이혼)
 - 가구주 종사지위코드(job) : 1 (상용근로자), 2 (임시, 일용 근로자), 3 (고용원이 있는 자영업자), 4 (고용원이 없는 자영업자), 5 (무급가족종사자), 6 (기타 종사자)

3. 자료 설명

- 연속형 변수 : 경상소득 , 소비지출 , 비소비지출
 - 경상소득 : 비교적 오랫동안 정기적으로 얻는 소득을 의미한다.
(경상소득 = 근로소득 + 사업소득 + 재산소득 + 공적이전소득 + 사적이전소득)
 - 소비지출 : 식료품비, 주거비 , 교육비 등 기본 생활에 필요한 상품과 서비스를 구입한 비용을 의미한다.
 - 비소비지출 : 소득세 , 재산세 등 각종 세금 , 건강보험료 등을 합친 경직성 비용을 의미한다.

3. 자료 설명

경상소득과 지출 변수에 대한 데이터 분포 확인

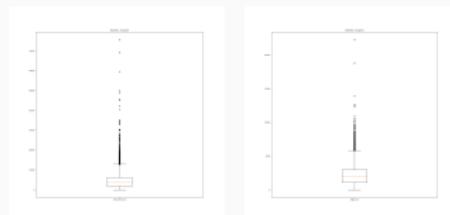
	경상소득(조사)	지출(조사)
mean	4644	2234
std	4214	1575
min	0.0	0.0
25%	1800	1086
50%	3666	1920
75%	6222	2990
max	75700	22280

Table 1: 기초통계량

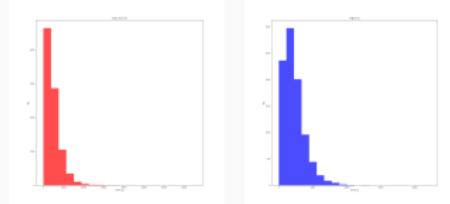
기초통계량을 통해 경상소득과 지출의 3분위수와 max값이 크게 차이나는 것을 봤을 때 이상치 값들이 많이 존재하는 것을 확인할 수 있다. 또한 상자그림과 히스토그램을 봤을 때 오른쪽으로 꼬리가 많이 치우쳐져 있는 형태임을 확인할 수 있으며, 경상소득과 지출 구성 하위변수들 또한 동일한 분포의 형태를 가지고 있는 것을 확인하였다.

3. 자료 설명

경상소득과 지출 변수에 대한 상자그림과 히스토그램



(a) 경상소득(조사)
상자
그림 (b) 지출(조사)
상자
그림



(c) 경상소득(조사)
히스토그램 (d) 지출(조사)
히스토그램

Figure 1: 경상소득(조사), 지출(조사) 상자 그림과 히스토그램

4. 재현자료 생성

4. 재현자료 생성

- CTGAN 패키지 버전
 - 첫번째 Scenario : SDV version 0.14
 - 두번째 Scenario : SDV version 0.14
 - 세번째 Scenario : SDV version 0.14 & RDT 1.4.2

4. 재현자료 생성

- CTGAN을 이용한 재현자료 생성 방법
 - Scenario 1: 아무조건 없이 기존 자료를 이용한 재현자료 생성 (epochs = 600 , batchsize = 300)
 - Scenario 2: 경상소득 , 지출 , 비소비지출에서 0의 갯수가 가장 적은 근로소득 , 식료품 소비지출 , 비소비지출에서의 세금 변수에 대해서 0 보다 큰 값과 0과 같다라는 조건을 주어 그룹화 하여, 그룹화된 데이터의 갯수마다 CTGAN의 조건을 다르게 주어 재현하는 방법. epochs는 같게 (10~100개의 경우 10으로, 100~1000개의 경우 100 으로, 1000~3000개의 경우 300으로, 3000개 이상의 경우에는 500 으로 설정하였다.)
 - Scenario 3: 위의 재현방법과 같지만, RDT라는 패키지를 이용하여서 각 그룹화된 데이터를 일정한 포맷에 맞게 변형시킨후 변형시킨 데이터를 CTGAN을 이용해 재현, 다시 원자료의 형태로 재변형시켜 재현자료를 생성하는 방법

5. 재현자료 평가

5. 재현자료 평가

- 평가 측도의 기준 : 유용성 , 보안성
 - 유용성 : 범주형 변수에 대해 파이썬에서의 SDV 패키지 및 막대그래프 , 범주형 빈도표 , R에서의 Synthpop 패키지를 이용할 것이며, 연속형 변수에 대해 상대편향 , 산점도 , 파이썬에서의 SDV 패키지 및 R에서의 Synthpop 패키지
 - 보안성 : 파이썬에서 SDV 패키지 안에 있는 CategoricalCAP 패키지를 이용

5. 재현자료 평가

- 유용성

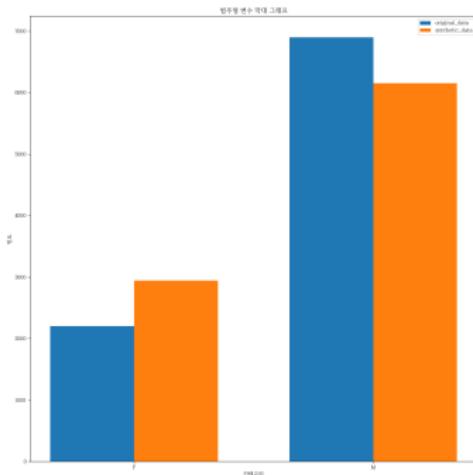
- Using Python : TVComplement , ContingencySimilarity , CategoryCoverage , KSComplement , RangeCoverage , BoundaryAdherence , MSE , MAE , KLD , CorrelationSimilarity (SDV Package)
- Using R : pMSE , MabsDD , PO50 , Bhattacharyya Distance (Synthpop Package)

5. 재현자료 평가

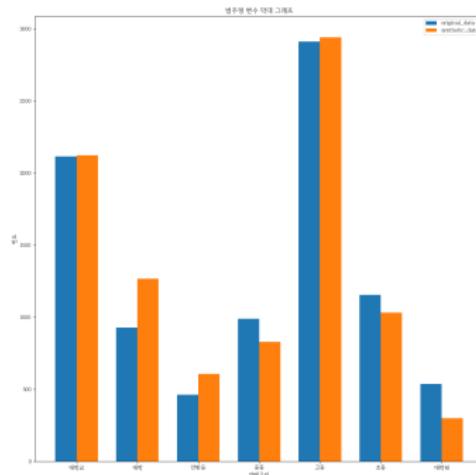
- 보안성
 - Using Python : CategoryZeroCAP , CategoricalGeneralizedCAP , NewRowSynthesis (SDV Package)

5. 재현자료 평가

Scenario 1의 막대그래프 및 범주형 척도 결과



(a) Scenario 1의 성별 변수 막대그래프



(b) Scenario 1의 교육 변수 막대그래프

5. 재현자료 평가

		원자료		재현자료	
		Count	Ratio	Count	Ratio
성별	남자	6898	76%	6153	68%
	여자	2199	24%	2944	32%
교육정도	고등	2912	32%	2941	32%
	대학교	2115	23%	2122	23%
	초등	1154	13%	1032	11%
	중등	988	11%	829	9%
	대학	927	10%	1268	14%
	대학원	538	6%	299	3%
	안받음	463	5%	606	7%

Table 2: Scenario 1 범주형 변수들의 재현 결과

5. 재현자료 평가

	CategoryCovages	TVComplement	ContingencySimilarity
sex	1.0	0.92	0.90
edu	1.0	0.94	0.90

Table 3: Scenario 1 범주형 측도

	직업	혼인	가구원	(직,혼)	(혼,가)	(직,가)	(직,혼,가)
GC	0.66	0.35	0.73	0.77	0.79	0.89	0.91
ZC	0.66	0.35	0.73	0.77	0.79	0.89	0.91

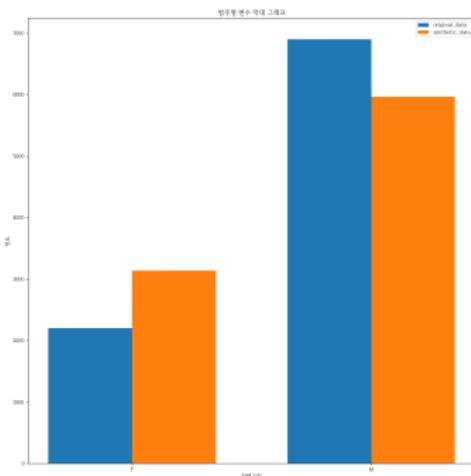
Table 4: Scenario 1 CategoricalGCAP , CategoricalZeroCAP

	pMSE	PO50	MabsDD	dBhatt
sex	0.002067	4.094757	0.163790	0.064430
edu	0.000573	2.149060	0.085962	0.033856

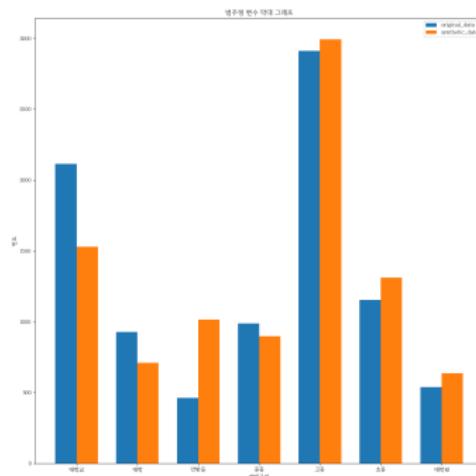
Table 5: Scenario 1 Synthpop을 이용한 범주형 측도

5. 재현자료 평가

Scenario 2의 막대그래프 및 범주형 척도 결과



(a) Scenario 2의 성별 변수 막대그래프



(b) Scenario 2의 교육 변수 막대그래프

5. 재현자료 평가

		원자료		재현자료	
		Count	Ratio	Count	Ratio
성별	남자	6898	76%	5964	66%
	여자	2199	24%	3133	34%
교육정도	고등	2912	32%	2994	33%
	대학교	2115	23%	1531	17%
	초등	1154	13%	1312	14%
	중등	988	11%	898	10%
	대학	927	10%	710	8%
	대학원	538	6%	637	7%
	안받음	463	5%	1015	11%

Table 6: Scenario 2 범주형 변수의 재현 결과

5. 재현자료 평가

	CategoryCovarages	TVComplement	ContingencySimilarity
sex	1.0	0.897329	0.85116
edu	1.0	0.902056	0.85116

Table 7: Scenario 2 범주형 측도

	직업	혼인	가구원	(직,혼)	(혼,가)	(직,가)	(직,혼,가)
GC	0.66	0.42	0.72	0.79	0.81	0.89	0.92
ZC	0.66	0.42	0.72	0.79	0.81	0.89	0.92

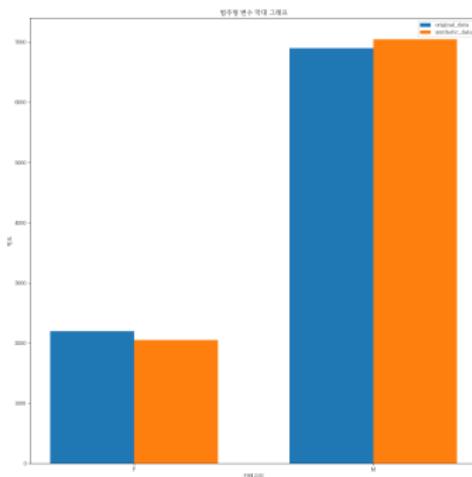
Table 8: Scenario 2 CategoricalGCAP , CategoricalZeroCAP

	pMSE	PO50	MabsDD	dBhatt
sex	0.003180	5.133561	0.205342	0.079985
edu	0.003551	3.407717	0.136309	0.085546

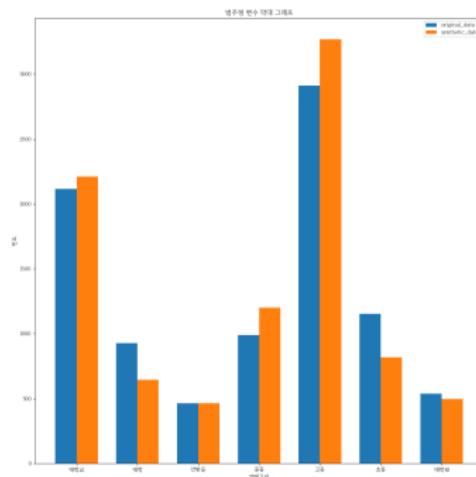
Table 9: Scenario 2 Synthpop을 이용한 범주형 측도

5. 재현자료 평가

Scenario 3의 막대그래프 및 범주형 척도 결과



(a) Scenario 3의 성별 변수 막대그래프



(b) Scenario 3의 교육 변수 막대그래프

5. 재현자료 평가

		원자료		재현자료	
		Count	Ratio	Count	Ratio
성별	남자	6898	76%	7047	77%
	여자	2199	24%	2050	23%
교육정도	고등	2912	32%	3268	36%
	대학교	2115	23%	2210	24%
	초등	1154	13%	817	9%
	중등	988	11%	1200	13%
	대학	927	10%	645	7%
	대학원	538	6%	495	5%
	안받음	463	5%	462	5%

Table 10: Scenario 3 범주형 변수의 재현 결과

5. 재현자료 평가

	CategoryCovages	TVComplement	ContingencySimilarity
sex	1.0	0.983621	0.894471
edu	1.0	0.927119	0.894471

Table 11: Scenario 3 범주형 측도

	직업	혼인	가구원	(직,혼)	(혼,가)	(직,가)	(직,혼,가)
GC	0.64	0.38	0.70	0.77	0.78	0.87	0.90
ZC	0.64	0.38	0.70	0.77	0.78	0.87	0.90

Table 12: Scenario 3 CategoricalGCAP , CategoricalZeroCAP

	pMSE	PO50	MabsDD	dBhatt
sex	0.000094	0.818951	0.032758	0.013689
edu	0.000841	1.857755	0.074310	0.041073

Table 13: Scenario 3 Synthpop을 이용한 범주형 측도

5. 재현자료 평가

연속형 변수들의 Scenario 별 측도 비교 및 산점도

		근로소득 변수 상대편향						
		평균	표준편차	최소값	20%	50%	80%	최댓값
Scenario1	-1.12	-3.63	0	0	-11.82	9.28	-30.2	
Scenario2	-5.27	-3.19	0	0	-10.10	0.69	-4.90	
Scenario3	12.21	11.92	0	0	2.29	26.23	-3.82	

		식료품 지출 변수 상대편향						
		평균	표준편차	최소값	20%	50%	80%	최댓값
Scenario1	-16.92	-8.86	0	-14.16	-28.89	-24.06	-39.21	
Scenario2	-7.17	-14.78	0	-2.08	-4.26	-8.23	-41.19	
Scenario3	1.75	4.49	0	2.08	-0.93	0.188	-39.52	

		세금 변수 상대편향						
		평균	표준편차	최소값	20%	50%	80%	최댓값
Scenario1	20.30	-13.65	0	112.5	-27.86	79.05	-74.14	
Scenario2	-4.76	-30.35	0	-50	4.91	22.04	-65.47	
Scenario3	10.01	-17.08	0	-12.50	-6.56	23.23	-83.13	

Table 14: 상대편향 비교: 근로소득, 식료품 지출, 세금

5. 재현자료 평가

근로소득 변수 연속형 측도1					
	KSC	RC	BA	MSE	MAE
Scenario1	0.94	0.70	1.0	2.3e+07	3500
Scenario2	0.94	0.95	1.0	2.3e+07	3476
Scenario3	0.91	0.96	1.0	2.77e+07	3835

식료품 지출 변수 연속형 측도1					
	KSC	RC	BA	MSE	MAE
Scenario1	0.80	0.60	1.0	4.1e+05	465.99
Scenario2	0.91	0.58	1.0	4.07e+05	474.03
Scenario3	0.92	0.60	1.0	4.75e+05	514.09

세금 변수 연속형 측도1					
	KSC	RC	BA	MSE	MAE
Scenario1	0.90	0.26	1.0	5.3e+05	330.72
Scenario2	0.95	0.35	1.0	4.4e+05	287.83
Scenario3	0.95	0.16	1.0	5.1e+05	319.97

Table 15: 연속형 측도1 비교: 근로소득, 식료품 지출, 세금

5. 재현자료 평가

	근로소득	변수	연속형	측도2
	KLDS		CS	
Scenario1	0.87		0.94	
Scenario2	0.96		0.96	
Scenario3	0.93		0.98	

	식료품	지출	변수	연속형	측도2
	KLDS		CS		
Scenario1	0.98		0.96		
Scenario2	0.98		0.95		
Scenario3	0.94		0.99		

	세금	변수	연속형	측도2
	KLDS		CS	
Scenario1	0.91		0.93	
Scenario2	0.97		0.97	
Scenario3	0.96		0.97	

Table 16: 연속형 측도2 비교: 근로소득, 식료품 지출, 세금

5. 재현자료 평가

NewRowSynthesis							
	inc	food	tax	inc&food	food&tax	inc&tax	all
Scenario1	0.21	0.37	0.11	0.85	0.93	0.78	0.99
Scenario2	0.19	0.36	0.14	0.81	0.90	0.70	0.97
Scenario3	0.14	0.33	0.08	0.82	0.90	0.72	0.98

Table 17: 연속형 측도3 비교: 근로소득, 식료품 지출, 세금

5. 재현자료 평가

		근로소득 변수 Synthpop 측도			
		pMSE	PO50	MabsDD	dBhatt
Scenario1		0.001263	2.682203	0.107288	0.050358
Scenario2		0.000899	2.165549	0.086622	0.042447
Scenario3		0.004777	4.908211	0.196328	0.098328

		식료품 지출 변수 Synthpop 측도			
		pMSE	PO50	MabsDD	dBhatt
Scenario1		0.009572	8.953501	0.358140	0.139354
Scenario2		0.002332	4.133231	0.165329	0.068491
Scenario3		0.007199	7.029790	0.281192	0.120840

		세금 변수 Synthpop 측도			
		pMSE	PO50	MabsDD	dBhatt
Scenario1		0.014063	11.135539	0.445422	0.169451
Scenario2		0.002341	4.078268	0.163131	0.068617
Scenario3		0.001809	3.825437	0.153017	0.060237

Table 18: Synthpop을 이용한 연속형 측도 비교: 근로소득, 식료품 지출, 세금

5. 재현자료 평가

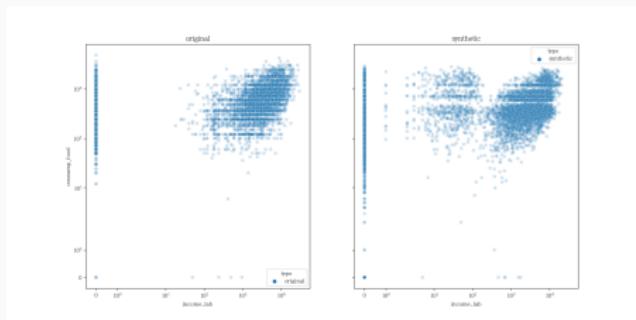


Figure 5: Scenario 1에서 근로소득과 식료품 지출의 산점도

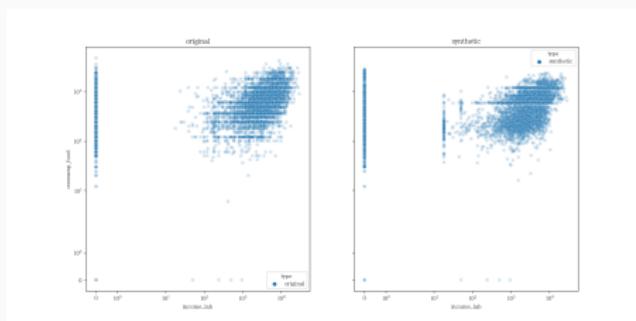


Figure 6: Scenario 20에서 근로소득과 식료품 지출의 산점도

5. 재현자료 평가

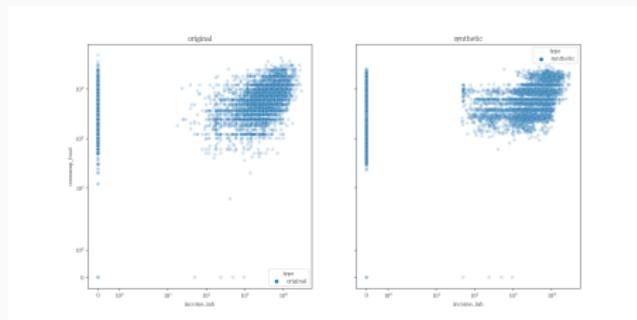


Figure 7: Scenario 3에서 근로소득과 식료품 지출의 산점도

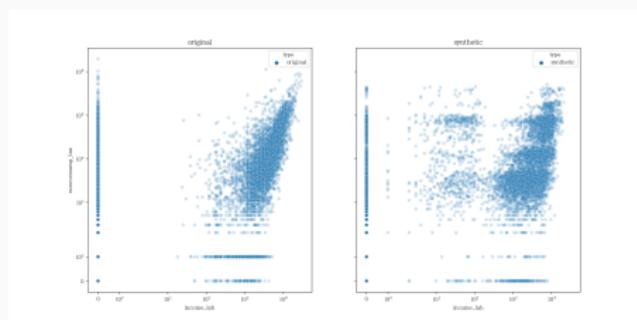


Figure 8: Scenario 1에서 근로소득과 세금의 산점도

5. 재현자료 평가

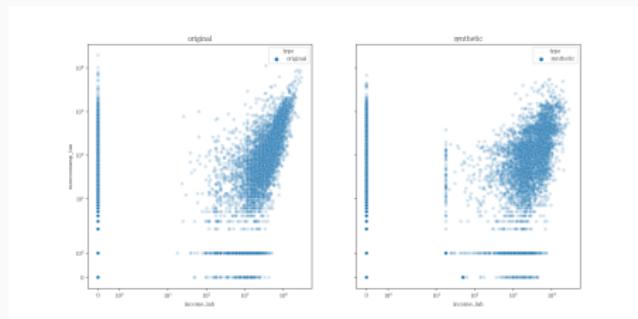


Figure 9: Scenario 2에서 근로소득과 세금의 산점도

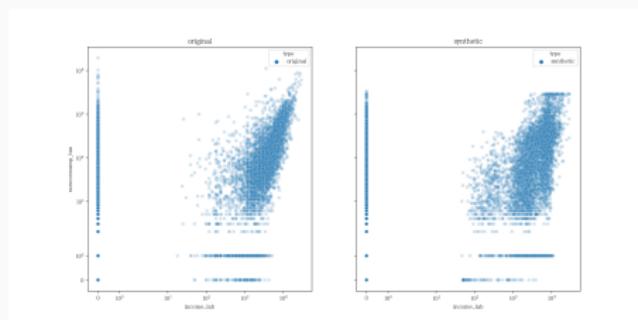


Figure 10: Scenario 3에서 근로소득과 세금의 산점도

5. 재현자료 평가

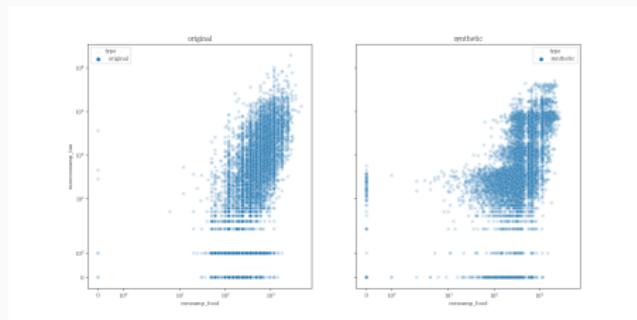


Figure 11: Scenario 1에서 식료품 지출과 세금의 산점도

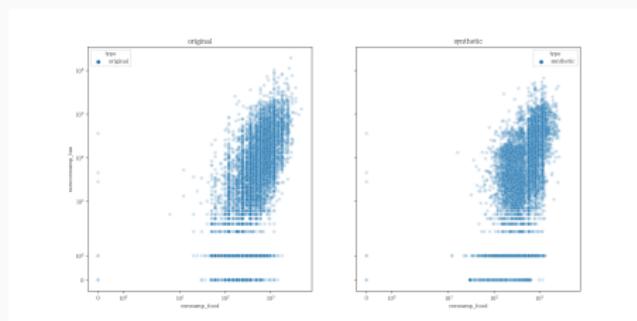


Figure 12: Scenario 2에서 식료품 지출과 세금의 산점도

5. 재현자료 평가

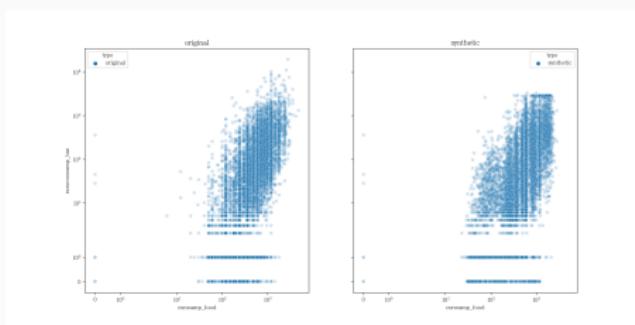


Figure 13: Scenario 3에서 식료품 지출과 세금의 산점도

5. 재현자료 평가

Scenario들의 상관계수 표 및 회귀계수

근로소득		
	식료품 지출	세금
원자료	0.46	0.45
Scenario1	0.34	0.38
Scenario2	0.38	0.35
Scenario3	0.51	0.44

식료품 지출		
	근로소득	세금
원자료	0.46	0.39
Scenario1	0.34	0.52
Scenario2	0.38	0.45
Scenario3	0.51	0.44

세금		
	근로소득	식료품 지출
원자료	0.45	0.39
Scenario1	0.38	0.52
Scenario2	0.35	0.45
Scenario3	0.44	0.44

Table 19: 원자료와 Scenario들의 상관 계수 표

5. 재현자료 평가

	회귀계수			
	원자료	Scenario 1	Scenario 2	Scenario 3
Intercept	4648	5044	4916	5385
남자	145	175	187	330
여자	-145	-175	-187	-330
안받음	-156	-162	-284	-192
초등	-182	-194	-263	-260
중등	-70	-80	-191	-237
고등	74	-63	57	179
대학	47	-78	89	-36
대학교	118	166	315	323
대학원	77	619	278	-81
근로소득	1313	71	717	730
식료품 지출	838	1993	1388	1539
세금	2111	1940	900	1237

Table 20: 원자료와 Scenario들의 회귀계수 표

6. 결론 및 향후 연구

6. 결론 및 향후 연구

- CTGAN과 RDT 패키지를 이용한 재현자료가 유용성 측면에서 가장 잘 재현이 되었음을 확인하였음.
- 회귀계수표에서 봤을때 변수간의 설명성을 잘 재현하지 못하는점을 발견하였음.
- 자료에 이상치가 많은 경우에 잘 재현하지 못함. 이상치가 많지않은 범주형 자료를 포함한 데이터를 재현할 때 좋은 재현자료를 재현할 수 있을 것이라고 봄.