한국ITS학회논문지 J. Korea Inst. Intell. Transp. Syst.

심층 생성모델 기반 합성인구 생성 성능 향상을 위한 개체 임베딩 분석연구

Entity Embeddings for Enhancing Feasible and Diverse Population Synthesis in a Deep Generative Models

권 동 현* · 오 태 호** · 유 승 모*** · 강 희 찬****

* 주저자 : 한국과학기술원 조천식모빌리티대학원 박사과정
** 공저자 : 한국과학기술원 조천식모빌리티대학원 연구원
*** 공저자 : 연세대학교 기술정책협동과정 박사과정
**** 교신저자 : 한국교통안전공단 모빌리티플랫폼처 연구위원

Donghyun Kwon* · Taeho Oh* · Seungmo Yoo** · Heechan Kang***

* Dept. of Cho Chun Shik Graduate School of Mobility, KAIST
** Graduate Program in Technology Policy, Yonsei Univ.
*** Mobility Research Department, Korea Transportation Safety Authority

† Corresponding author : Heechan Kang, hckang@kotsa.or.kr

Vol. 22 No.6(2023)
December, 2023
pp.17~31

pISSN 1738-0774
eISSN 2384-1729
https://doi.org/10.12815/kits.2023.22.6.17

Received 23 October 2023
Revised 31 October 2023
Accepted 6 November 2023

요 약

활동기반 모델은 현대의 복잡한 개인의 통행행태를 반영한 정교한 기반의 수요예측이 가능하지만, 분석 대상지의 상세한 인구정보가 필수적으로 요구된다. 최근 다양한 심층생성 모델을 활용한 합성인구 생성 기법이 개발되었고, 설문조사를 통해 수집된 샘플 데이터에 존재하지 않는 실제 인구와 유사한 인구 특성을 모사한 데이터를 생성해내는 방법론이 제시되었다. 이는 이산형으로 이루어진 샘플 데이터를 연속형 데이터로 변환하여 분포 영역을 정의한 뒤 생성된 표본 데이터의 거리를 정교하게 계산하여, 불가능한 인구 특성 조합을 억제하는 방식으로 데이터의 확률 분포를 학습한다. 하지만 데이터 변환 과정에 활용되는 개체 임베딩이 잘 학습되지 않으면 의도와 다르게 왜곡된 연속형 분포 영역이 정의될 수 있고, 원본 데이터 표현의 오류로 인한 잘못된 합성인구를 생성할 가능성이 존재한다. 따라서 본 연구에서는 정확도 높은 임베딩을 추출하여 간접적으로 합성인구 생성 성능을 증가시키고자 한다. 결과적으로 합성인구의 다양성과 정확성 측면에서 기존 대비 약 28.87% 성능이 향상하였다.

핵심어 : 합성인구 생성, 개제 임베딩, 심층생성모델, 활동기반모형

ABSTRACT

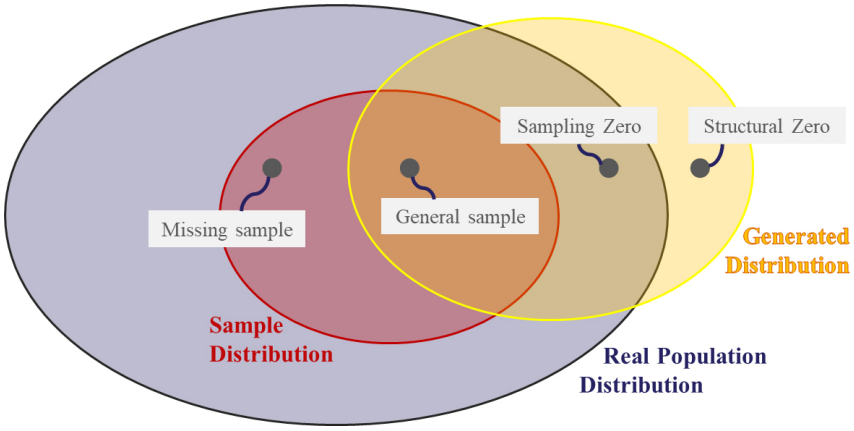
An activity-based model requires detailed population information to model individual travel behavior in a disaggregated manner. The recent innovative approach developed deep generative models with novel regularization terms that improves fidelity and diversity for population synthesis. Since the method relies on measuring the distance between distribution boundaries of the sample data and the generated sample, it is crucial to obtain well-defined continuous representation from the discretized dataset. Therefore, we propose an improved entity embedding models to enhance the performance of the regularization terms, which indirectly supports the synthesis in terms of feasible and diverse populations. Our results show a 28.87% improvement in the F1 score compared to the baseline method.

Key words : Synthetic population, Entity embedding, Deep generative models, Activity based model

Ⅰ. 서 론

1. 개요

활동기반 모델(activity-based Model)은 개인의 필요 활동 또는 욕구를 충족하는 통행행태를 도출하여 교통수요 예측을 수행하는 기법으로, 기존에 통합적으로 수행된 교통수요 모델의 한계점을 극복하고 고해상도의 공간적, 시간적 구성 요소 분석이 가능하게 만들면서도 복잡한 다차원적인 요소들(교통정책, 교통시설, 교통환경, 사회경제지표 등)을 유기적으로 반영할 수 있게 되었다(Castiglione et al., 2015). 해당 모델링 방안은 분석 초기 단계에 대상지의 실제 모집단 인구 특성에 대한 정보가 필수적으로 요구되지만, 현실에서 총인구수만큼의 개인 정보에 대한 데이터 수집이 불가능하므로 샘플 인구조사 및 조사 대상 지역의 사회경제지표 데이터를 기반으로 합성인구를 생성하는 기법들이 도출되었다. 기존의 IPF[1]와 같은 방법론은 생성된 인구 데이터의 주변확률분포(marginal distribution) 혹은 이변량 주변확률분포(bivariate marginal distribution)만을 기반으로 합성인구를 생성하고 검증하기 때문에, 실제 인구의 정확성과 다양성을 고려하지 못하는 문제점이 존재하였다. 또한 활동기반 모델 수행을 위해 도출된 합성인구는 활동 생성 및 스케줄링을 도출하는 연속적 분석과정의 기반이 되므로, 초기 데이터의 정확도를 개선하지 않을 시에 작은 오차가 누적되어 바람직하지 않은 결과를 초래할 수 있다(Garrido et al., 2020). 이렇듯, 생성된 표본의 정확성과 다양성은 잠재적으로 수요예측 결과에 큰 영향을 미치게 되므로 수집된 샘플 데이터에는 존재하지 않지만, 실제 인구에 존재하는 데이터를 생성하기 위한 방법론의 필요성이 지속적으로 논의되었다. 그러나 실제 인구 모집단 정보를 알지 못한 상황에서의 실제 인구 특성 조합과 불가능한 인구 특성 조합의 구분법은 직접적인 정답 항목이 제공될 수 없으므로 기존의 판별적인 모델(discriminative model) 활용 접근법으로는 해결되지 못한 문제로 남았다. 그러나 최근 광범위한 연구를 통해 심층 생성모델(deep generative model) 기반으로 데이터의 확률 분포를 학습하여 합성인구를 생성하는 혁신적인 방법론이 개발되었으며, 생성된 인구 샘플의 정확도와 다양성을 높이기 위해 생성 데이터를 구분하는 방안이 제시되었다 (Borysov et al., 2019).

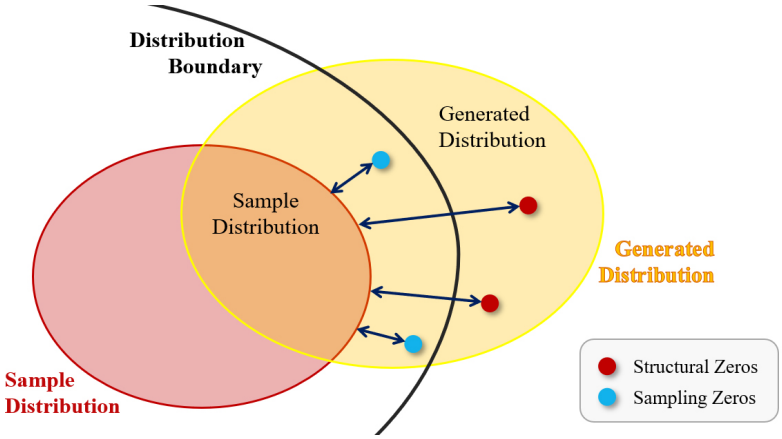


<Fig. 1> Conceptual data types of the generated samples

생성모델을 통해 도출된 인구 샘플은 일반 표본(general sample), 결측 표본(missing sample), 샘플링 제로

1) IPF : Iterative Proportional Fitting

(sampling zero), 구조적 제로(structural zero) 등 네 가지 생성된 데이터 유형으로 분류할 수 있고, 이를 개념화하여 도식화하면 <Fig. 1>와 같다. 데이터 분포 영역의 맥락에서 샘플 인구(sample distribution)는 빨간색 분포로 표현되며, 실제 인구(real population distribution)의 분포는 파란색으로 표현된다. 한편, 생성모델로부터 학습된 분포(generated distribution)는 노란색으로 표현되며, 실제 모집단 분포를 최대한 유사하게 모방하는 것을 목적으로 학습된다. 일반 표본은 샘플 인구와 실제 인구 분포 영역 모두에 존재하는 실현 가능한 표본이며, 결측 표본은 사용가능한 표본이지만 생성되지 않는 표본을 의미한다. 샘플링 제로는 주어진 샘플 인구 분포에는 존재하지 않지만, 실제 인구 분포에는 존재하는 표본이고, 이와 대조적으로 구조적 제로는 논리적으로 불가능한 속성 조합을 생성한 표본을 의미한다. 만약 실제 인구 데이터의 분포를 알고 있다면, 샘플링 제로는 해당 분포 내에 존재하는 인구 데이터로 정의할 수 있다. 또한 구조적 제로는 분포 영역 외에 존재하는 데이터로 정의된다. 예를 들어, 생성된 표본이 나이 15세의 학생이지만 전문 직업을 가진 직장인이라면, 실제 모집단이나 표본 분포에 존재하지 않으므로 모델의 학습 과정에서 최대한 생성을 방지할 수 있도록 확률 분포를 모델링해야 한다. 따라서 이상적인 합성인구 생성모델은 구조적 제로의 생성을 최소화하면서 일반 표본과 결측 표본, 그리고 샘플링 제로의 생성을 최대화할 수 있어야 하며, 데이터 구분법을 통해 생성모델의 학습 과정에서 명확히 분류되어야 한다. 샘플 데이터를 기반으로 데이터의 확률 분포를 학습하는 모델의 특성상 데이터가 완벽히 구분되는 방안은 없으나, 생성되어야 할 표본의 정확성과 다양성 측면에서 균형 관계(trade-off relationship)를 이룬 채로 확률 분포를 도출하여 일부분 해결할 수 있다. 하지만 생성모델은 주어진 인구 샘플 분포를 복제하는 것이 아닌, 학습 분포의 영역을 실제 모집단 인구의 분포와 유사하도록 의도적으로 늘려 샘플링 제로를 생성해야 하므로, 이러한 노력은 필연적으로 구조적 제로의 가능성으로 이어진다.



<Fig. 2> Conceptual data types of the generated samples

이러한 문제는 직관적으로 데이터 분포 영역의 관점에서 실제 인구 데이터와 샘플 인구 데이터를 바라봄으로써 일부 해결되었다. 샘플 데이터는 실제 인구 데이터 내에 존재하는 데이터이며, 그 데이터 분포가 서로 유사한 형태를 이루도록 신뢰성 있게 수집된 데이터임을 가정할 수 있다. 이러한 접근방식은 구조적 제로는 실제 인구 분포 영역과 상당히 먼 거리에 위치하고 샘플링 제로는 상대적으로 가까운 거리에 위치할 것으로 예상할 수 있으며, 그 거리를 정량적으로 측정할 수 있다면 생성모델의 학습 단계에서 확률 분포가 샘플 데이터의 분포와 너무 멀어지지 않도록 조정하는 과정을 추가하고, 샘플 데이터의 분포를 기점으로 실제 인구 데이터의 분포와 유사해지도록 조정하는 과정을 추가하여 샘플링 제로와 결측 표본의 생성을 유도할

수 있다. 따라서 이전 연구에서는 이를 정량적으로 측정할 수 있는 새로운 손실함수를 정의하고 학습 과정 중 정확도와 다양성 측면에서 최적의 균형을 이루는 학습 가중치를 도출하였다(Kim and Bansal, 2023). 이는 <Fig. 2>와 같이 샘플 데이터 분포 영역(sample distribution)은 거리 계산을 위한 물리적 의미를 지니며, 생성모델로부터 생성된 표본 데이터와 최소 거리를 계산하는 과정을 거치는 것으로 이해할 수 있다. 이때 분포 영역 정의를 위한 수집된 인구 특성 정보는 범주화된 데이터로 구성되어 있으므로 직접적으로 분포 영역을 정의하면 이산형 공간(discrete space)을 이루게 되어, 거리 측정 과정의 단순화로 인해 샘플링 제로와 구조적 제로의 정교한 구분이 불가능하다. 이에 따라 범주화된 데이터를 연속형 공간(continuous space)으로 변환하여 손실함수를 적용할 시에 합성인구 생성모델의 성능이 향상될 수 있음을 보였지만, 새롭게 정의된 데이터 분포가 샘플링과 구조적 제로의 생성에 어떤 영향을 미치는지 충분한 연구가 수행되지 않았다. 따라서 본 연구에서는 다양한 모델을 통해 임베딩 공간 추출한 뒤 생성모델로부터 얻어진 샘플 인구 데이터를 정확성과 다양성 측면에서의 성능을 비교하고 이전 연구에서 제시한 샘플링 제로와 구조적 제로에 대한 가정을 견고히 뒷받침하고자 한다.

Ⅱ. 문헌 검토

1. 생성모델 기반 합성인구 생성 방안

합성인구 생성 문제는 표 형식의 정형 데이터(tabular data)를 생성하는 문제와 유사하다. 이는 범주형으로 표현된 데이터 간의 상관관계를 성공적으로 포착함과 동시에 높은 확장성(scalability)을 유지하는 대규모 데이터 생성 문제에 적용할 수 있어야 한다. 이는 또한 범주형 데이터를 동일한 형태로 다양하고 정확하게 생성하는 것이므로, 활동기반 모델의 설명력과 정확성을 높이는 필수적이고 도전적인 문제이다. 해당 문제는 일반적인 생성모델 중 가장 대표적인 GAN[2]을 도입하여 해결한 사례가 존재하며 그 효과성을 입증하였다(Xu and Veeramachaneni, 2018). 기존 방법론과 비교했을 때 합성인구 생성에 적용하였을 때의 주요 장점은 모델의 확장성을 보장하면서 샘플 데이터를 직접 풀링하는 대신 인구 특성의 분포를 학습하여 새로운 샘플 데이터를 생성한다는 점이다. 따라서 표본 데이터의 분포와 가장 잘 일치하는 샘플 데이터를 생성할 뿐만 아니라 실제 모집단에 존재할 수 있지만 관측되지 않은 표본도 생성할 수 있다. 이러한 이점을 활용하여 활동기반 모델 적용을 위한 현실과 유사한 합성인구 생성 문제 해결에 상당한 진전을 보였고 심층 생성모델의 견고성(robustness)은 여러 데이터 세트를 사용한 실증 연구(Borysov et al., 2019; Aemmer and MacKenzie, 2022)를 통해 낮은 샘플링 비율의 데이터 세트에도 합성 모집단 작업을 적용할 수 있음을 보였다. 또한, 해당 방법론이 보편화되는 고차원의 서로 다른 의미를 가진 희귀한 특성 조합을 고려하여 다양성을 생성해낼 수 있는 모델로 Wasserstein GAN이 합성인구 생성에 가장 적합한 모델임이 입증되었다(Garrido et al., 2020). 위의 연구로부터 합성인구의 유효한 특성을 생성하는 데 있어 생성모델의 우수성을 보여져 왔으나, 정확하고 다양한 표본을 생성하기 위해 이들을 구분해야 한다는 지적이 반복적으로 제기되었다. 따라서 주관적인 판단을 통해 구조적 제로를 정의하는 방안이 제시되었으나 일반화에 어려움이 존재하여 해결방안으로 제시되지 못했다. 이러한 문제의 돌파구로써 생성모델로부터 도출된 데이터를 구분하고 샘플링과 구조적 제로를 구별하는 새로운 정규화 용어가 제안되었고, 구조적 제로의 생성을 억제하기 위한 손실함수와 샘플링 제로

2) GAN : Generative Adversarial Network

을 유도하기 위한 손실함수를 각각 정의한 뒤 생성자(Generator)의 손실함수에 추가하여 문제를 해결하였다(Kim and Bansal, 2023). 해당 방법론은 데이터의 정확성과 다양성을 평가하기 위해 정밀도와 재현율 종합적으로 판단한 F1 Score를 기반으로 최종 성능을 평가하였다. 이때 정밀도는 생성된 인구 특성 정보 대비 모집단 인구 특성 정보에 존재하는 비율로 정확도를 판단하고, 재현율은 생성된 데이터의 속성 조합이 모집단에 속하는 비율을 나타내어 다양성을 판단한다. 그러나 해당 방법론은 샘플 데이터로부터 얻은 분포 영역을 연속적 공간으로 표현한 뒤 생성된 표본의 거리를 측정하여 샘플링 제로와 구조적 제로의 생성을 유도하고 억제하는 방안으로 구현되었으므로 정의된 분포 영역의 형태에 따라 크게 영향받는다.

2. 개체 임베딩 추출 기법

자연어 처리(Natural language processing), 이상치탐지(Anomaly detection) 등의 분야에서는 단어 혹은 이벤트들에 대한 정보의 의미를 표현하기 위한 다양한 기법들과 다차원적인 복잡성을 줄이고 데이터의 유의미한 상관관계를 학습하여 데이터를 표현하는 효과적인 기법들이 제시되었다(Hancock and Khoshgoftaar, 2020). 가장 잘 알려진 기법으로 범주형 데이터를 0과 1의 고유한 인덱스로 구분된 이진 벡터로 변환하는 과정을 일컫는 원-핫 인코딩(One-hot encoding)이 있으나, 여전히 데이터가 고차원적으로 표현되어 대규모의 범주 유형이 있는 경우에 적합하지 않고 순서나 관계를 고려하지 못하므로 정보손실의 문제가 발생한다. 이러한 희소표현(sparse representation)의 한계점을 보완하기 위해 발전된 임베딩 기법은 분포가설(distributional hypothesis)을 기반으로 단어들의 유사성을 계산하기 위한 표현과 저차원 표현을 위한 밀집 벡터를 도출한다. 범주형 데이터는 정수로 표현된 데이터들의 집합이지만, 단어 학습과 유사하게 데이터를 직접적으로 활용한다면 학습에 비효율적일 뿐만 아니라 고차원일수록 데이터가 지닌 고유한 의미를 담고 있지 않아 모델 성능 향상에 어려움이 존재한다(Guo and Berkhahn, 2016). 따라서 범주형 데이터를 효과적으로 학습하기 위한 임베딩 기법이 연구되었고 일반적으로 개체 임베딩(Entity Embedding)으로 불리며, 해당 방법론을 적용하면 데이터가 연속적으로 표현될 뿐만 아니라 데이터의 의미적 관계와 유사성을 내포하게 된다. 또한 추출된 임베딩을 입력 데이터로 활용하여 다양한 문제에 학습 시에 성능이 비약적으로 상승함을 보였다. 따라서 합성인구 문제를 위해 적용되는 개체 임베딩 또한 인구 데이터 특성 간의 상관관계(contextual relationship)에 대한 정보를 내포함으로, 주어진 데이터를 직접 활용하는 것 보다 더욱 정교하고 정확한 연속적 공간을 정의할 수 있다. 특히 불가능한 인구 특성 조합의 생성을 방지하여 구조적 제로의 생성이 최소화될 것으로 예상된다. 이에 따라 기존 연구에서는 자연어 처리 분야에서 널리 적용된 임베딩 기법 중 하나인 BERT[3]의 사전학습(pre-training) 과정 중 MLM[4]을 채택하여 데이터를 변환하였다 (Kim et al., 2022). 이는 단방향 순서만을 고려해 데이터간의 의미를 파악하는 것이 아닌, 데이터의 배열을 모두 고려한 양방향으로 데이터 특성 간의 관계를 파악한다(Devlin et al., 2018). 따라서 데이터의 속성 조합을 마스킹 기법으로 일부 가린 뒤에 가려진 데이터 속성을 예측하는 모델을 기반으로 정확도를 도출해 성능을 평가하게 된다. 이에 따라 인구 특성 간의 관계와 의미를 온전히 반영한 연속형 데이터로 변환이 가능해졌으나, 해당 방법론의 메커니즘을 도입했을 뿐 정확성을 높이기 위한 다양한 모델이 고려되지 않았다. 이는 앞선 연구에서 가정한 샘플링 제로와 구조적 제로의 구분을 위한 손실함수의 성능을 간접적으로 그리고 잠재적으로 증가시킬 수 있으므로, 임베딩 공간을 통한 데이터 분포 영역 정의는 생성모델의 정확성과 다양성 측면에서 상당히 큰 영향을 미친다. 본 연구에서는 특히 대표적으로 다양한 연구가 진행된 자연어 처리 기법에서 활용된 기계번역 모델들을 적용하여 그 효

3) BERT : Bidirectional Encoder Representations from Transformers
4) MLM : Masked Language Model

과를 파악하고자 한다.

Ⅲ. 방법론

1. 데이터셋 구축 과정 및 실험적 가정

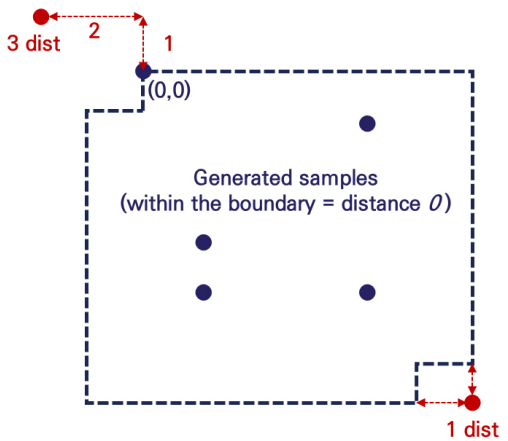
분석 대상지와 관계없이 모집단 인구 데이터에 대한 수집은 현실적으로 불가능하므로 생성된 인구표본 데이터 검증 대한 실험적 가정이 필요하다. 본 연구에서는 일반적으로 수요분석을 위한 기초자료로 활용되는 가구통행실태조사를 기반으로 연구를 수행하였고 KTDB[5]에서 전국단위로 수집한 자료의 2016년도 총 데이터를 수집하였다. 이후 모집단 실제 인구를 대변하기 위한 데이터는 수집된 총인구 데이터로 가정하였고, 그중 5%의 샘플 데이터를 랜덤하게 추출하여 현실에서 수집된 샘플 인구 데이터로 가정하였다. 이에 따른 실험적 가정은 샘플 인구 데이터를 기반으로 총인구수만큼의 데이터를 생성모델을 통해 도출해낸 후 실제 인구와 비교하여 합성인구의 정확성과 다양성을 측정한다. 만약 생성모델이 실제 인구와 유사하도록 합성인구를 생성한다면, 신뢰성이 높다고 판단하며 실제 활동기반 모델 분석을 위한 분석 대상지에서 수집한 샘플 데이터를 전부 모델의 학습에 사용한 뒤 실제 인구수만큼 합성인구를 사용하여 수요예측을 수행할 수 있다. 생성모델을 통해 도출할 합성인구의 인구 특성 정보는 모두 범주화(categorical)된 데이터로 구성된다. 본 연구에서는 인구 특성 정보의 복잡성을 최대한으로 증가시키기 위해 가구통행실태조사에서 추출 가능한 인구 특성 정보를 활용하였으며, 본 연구에서 정의된 데이터 구성은 <Table 1>과 같다. 이에 따라 총 13개의 인구 특성과 22,581개의 인구 샘플 데이터를 기반으로 총 451,613명의 실제 인구를 생성해내도록 실험환경을 구축하였다.

<Table 1> The composition of the household travel survey(HTS) dataset

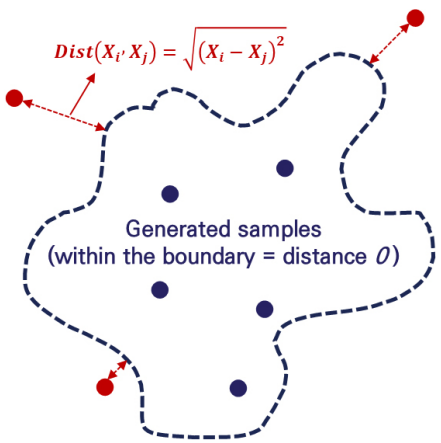
Attributes	Category	Percentage (%)	Categoty	Percentage (%)
Age	[0,10)	4.06	[50,55)	8.21
Age	[10,15)	5.09	[55,60)	9.05
Age	[15,20)	6	[60,65)	7.14
Age	[20,25)	4.52	[65,70)	5.98
Age	[25,30)	5.46	[70,75)	4.89
Age	[30,35)	7.12	[75,80)	4.33
Age	[35,40)	8.51	[80,85)	2.42
Age	[40,45)	7.68	[85	0.96
Age	[45,50)	8.57	-	-
Driver's License	Yes	58	No	42
Gender	Male	49.91	Female	50.09
Person number	1st person	43.23	4th person	7.74
Person number	2nd person	30.27	5th person	1.19
Person number	3rd person	17.58	-	-

5) KTDB : 국가교통 데이터베이스, Korea Transport DataBase

Attributes	Category	Percentage (%)	Categoty	Percentage (%)
Student types	Preschool	0.97	Univ./ Graduate school	3.91
Student types	Elementary school	6.15	None	81.78
Student types	Middle / High school	7.2	-	-
Working types	Professional	3.23	Laborers	9.74
Working types	Service	9.3	Housewife	15.28
Working types	Sales	9.56	Non-workers / Students	27.62
Working types	Office worker / Manager	16.41	Others	1.92
Working types	Agriculture / fisheries	6.95	-	-
Household income	< 1 million (KRW)	10.68	3 - 5 million (KRW)	35.39
Household income	1 - 2 million (KRW)	13.83	> 5 million (KRW)	19.34
Household income	2 - 3 million (KRW)	20.77	-	-
Household income	Apartment	53.15	Single house	29.41
Household income	Villa	9.77	Dual purpose house	0.75
Household income	multi-household	6.47	Others	0.45
Resident region	Region 1	1.97	Region 10	0.33
Resident region	Region 2	28.48	Region 11	2.12
Resident region	Region 3	4.33	Region 12	8.6
Resident region	Region 4	4.57	Region 13	2.46
Resident region	Region 5	2.83	Region 14	0.88
Resident region	Region 6	4.17	Region 15	1.4
Resident region	Region 7	3.36	Region 16	4.98
Resident region	Region 8	7.38	Region 17	3.51
Resident region	Region 9	18.64	-	-
Household relationship	house owner	43.16	Parents	2.6
Household relationship	Spouse	25.95	Others	1.95
Household relationship	Children	26.34	-	-



<Fig. 3> Distance measure in the discrete space

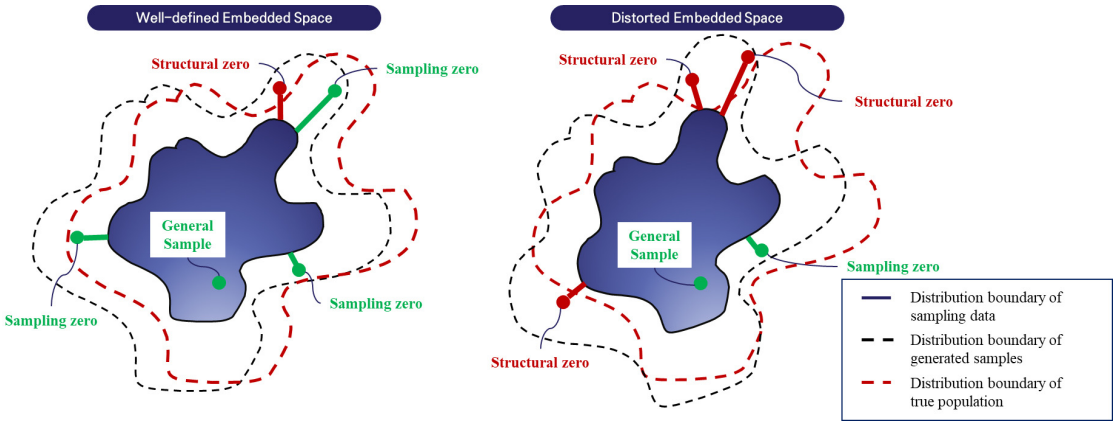


<Fig. 4> Distance measure in the continuous space

본 연구에서는 인구 샘플 데이터의 분포는 실제 모집단 인구 데이터의 분포 영역 내에 위치하며, 큰 편향 없이 유사한 확률 분포를 가질 것으로 예상한다. 또한 이전 연구에서의 가설과 같이 샘플링과 구조적 제로는 샘플 인구 데이터의 분포 영역을 정의하여, 샘플링 제로는 가깝게 위치하고 구조적 제로는 멀게 위치할 것으로 가정하고, 생성모델 학습 과정에서 최적의 표본생성 방안을 도출할 수 있도록 지원하는 손실함수를 생성모델에 도입하였다(Kim et al., 2022). 이때 생성모델을 통해 도출되는 표본 데이터의 위치는 데이터 분포 영역과의 거리가 수치적으로 계산되어야 하며, 분포 영역에 대해 이산형 공간(discrete space) 혹은 연속형 공간(continuous space)으로 정의되는지에 따라 거리 계산값의 차이가 존재하게 된다. 개념적으로 이산형 공간의 경우 <Fig. 3>과 같이 그리드 형태의 데이터로 해석될 수 있으며, 데이터 분포 영역과 생성되는 데이터 모두 그리드의 공간과 점으로 표현될 수 있다. 해당 영역에서의 거리 계산은 2차원 공간에서의 예시와 같이 일정한 간격 단위의 거리로 계산되며, 표본 데이터가 분포 영역에 포함될 수 있는 최소 거리를 탐색한다. 그러나 이산형 공간은 <Fig. 4>와 같이 생성 가능한 인구 특성 집합 내에서 제약된 조건 없이 유연한 분포 영역을 정의할 수 있으며, 데이터 표본의 위치 또한 집합 내에서 자유롭게 생성된다. 따라서 본래의 데이터 특성을 잘 표현한 연속적 공간을 기반으로 최소 거리를 계산하게 수식(1)과 같이 유클리드 거리(Euclidean distance)로 섬세한 거리 계산이 가능하고, 이는 샘플링과 구조적 제로 표본을 이산형 공간보다 더욱 정확히 구분할 수 있도록 돕는다. 또한 활동기반 모델 적용을 위한 합성인구 생성단계는 실제 인구 데이터의 분포 영역을 알 수 없고, 수집된 샘플 데이터에 의존하여 데이터를 생성하므로 더욱 정교한 모델링 과정이 요구된다. 따라서 샘플 데이터의 연속적 공간 변환 과정은 상당히 중요한 의미가 있으며, 생성모델을 통해 도출된 표본 데이터(generated samples)의 분포 영역은 <Fig. 5>의 극단적 경우와 같이 왜곡된 분포 영역을 기반으로 학습 시 샘플링과 구조적 제로의 잘못된 생성으로 이어질 수 있다. 그러나 정의된 데이터 분포 영역에 대한 검증이 현재 단계에서는 불가능하므로, 임베딩 공간 추출을 위해 사용된 모델의 정확도가 높다면 데이터의 의미를 잘 반영한 연속적 공간을 생성했다고 가정한다.

$$Dist(X_i, X_j) = \sqrt{(X_i - X_j)^2} \quad \cdots\cdots (1)$$

Where, X_i and X_j is the two different continuous vectors



<Fig. 5> The hypothesis on the miss construction of the continuous embedding space

2. 개체 임베딩 추출 및 연속형 공간 변환 방안

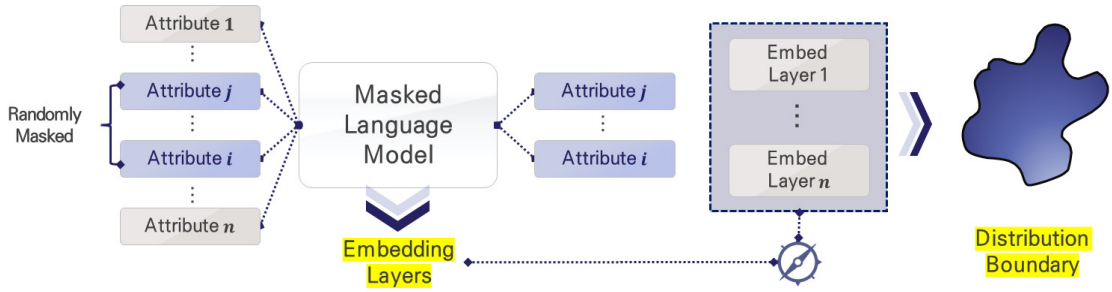
개체 임베딩 벡터 생성은 범주형 데이터를 심층망(Neural network)에 학습한 후 임베딩 레이어의 가중치 값들을 추출하는 과정이며, 범주형 특성값의 의미를 함축한 연속적인 벡터로 표현된다. 이때 활용되는 심층망 모델은 밀집 계층(Dense layer)으로 구성된 모델과 순환 신경망(Recurrent Neural Network) 계열 모델인 LSTM, 그리고 대규모 언어 모델(Large Language Models)의 근간이 되는 Transformer와 모델 아키텍쳐의 근간이 되는 Seq2Seq 모델, 그리고 Multi-head Attention 총 5가지의 모델이 적용되었다 (Sutskever et al., 2014; Palangi et al., 2016; Vaswani et al., 2017). 앞서 합성인구 문제에 적용된 MLM 과정은 샘플 데이터(학습데이터)의 인구 특성 정보인 나이, 성별과 같은 범주형 데이터값을 무작위로 특별한 마스킹 값으로 대체한 뒤, 심층망 모델을 통해 원본값에 대한 예측을 수행한다. BERT 사전학습의 원본 논문의 마스킹 과정은 초기에 주어진 데이터의 모든 특성의 배열 중 15%를 무작위로 선택하고, 80%의 확률로 마스킹으로 대체하거나, 10%의 확률로 임의의 값을 입력하거나, 나머지 10%의 확률로 마스킹하지 않는 과정으로 학습데이터를 구축한다. 합성인구 인구 특성 데이터 또한 동일한 과정을 수행하지만 실제로 마스킹을 적용하는 단계에서 마스킹 값이 아닌 임의의 값을 입력하는 과정 대신 90%의 확률로 마스킹하는 과정으로 학습데이터를 구축하였다. 이때 마스킹이 적용된 위치에 올바른 인구 특성의 실제 값이 정수 형태이므로 희소 교차 엔트로피 손실함수(sparse cross entropy loss)를 적용하여 예측값에 대한 확률과 실제 값 간의 차이를 계산하여 학습하게 된다. 또한 초기 학습된 모델을 기반으로 마스킹 조합이 다른 학습데이터를 구성하여 재학습하는 과정을 거쳐 다양한 특성 조합을 효과적으로 학습하도록 반복하였다. 해당 기법을 적용한 임베딩 벡터는 마스킹된 인구 특성 정보를 중심으로 양방향으로 위치한 인접 인구 특성 정보의 맥락을 파악할 수 있고 필수적인 관계 정보를 이해하는 것에 기여할 수 있다.

생성모델 학습을 위한 인구 특성 데이터는 특성별로 담긴 범주형 데이터를 그대로 학습하지 않고 원-핫 인코딩을 수행하여 이진 형식으로 변환하여 활용된다. 이는 인구 특성 정보가 종종 연속형 데이터로 집계되어있는 경우 동일한 데이터 형식으로 데이터를 처리하기 위해 특성 정보의 균일한 구간을 설정하여 이산화하기 위한 것과 동일한 특성 정보가 다르게 표현되었을 때 이를 혼합하기 위한 과정으로 활용된다. 또 다른 효과로 해당 과정을 통해 각 범주형 특성 정보가 고유한 벡터로 표현되어 인코딩 기법을 적용할 때 발생하는 편향 혹은 혼동을 방지하는 효과가 존재한다(Borysov et al., 2019). 그러나 앞서 언급한 대로 분포 영역이 여전히 이산형 공간으로 표현되므로 정확한 거리 계산을 위해 <Fig. 6>과 같이 개체 임베딩을 활용하여 연속형 데이터로 변환하는 과정이 요구되며, 이를 수행하기 위한 방법론은 식(2)과 같이 표현된다. 샘플 인구는 고유특성 정보의 개수를 나타내는 C와 N 개의 표본으로 구성된 데이터인 $X \in Z^{N \times C}$로 나타낼 수 있다. 이때 원-핫 인코딩을 수행하면 j번째 인구 특성의 고유 범주형 값을 나타내는 K_j 를 C개만큼 합한 형태인 K개의 특성 정보가 나열된 $X_{one_hot} \in \{0,1\}^{N \times K}$으로 나타낸다. 연속형 데이터로 변환하는 과정은 잘 학습된 개체 임베딩 레이어의 가중치가 추출된 $E_j \in R^{K_j \times d_j}$와 X_{one_hot}와 내적을 수행하는 과정을 통해 모든 범주형 특성을 연속형으로 변환하고 $X_{continuous} \in R^{N \times d}$와 같이 나타낸다.

$$X_{continuous} = X_{one\_hot} \bullet [E_1, E_2, \cdots, E_j, \cdots, E_C] \quad \cdots\cdots (2)$$

Where, C : The number of categorical attributes
E_j : The embedding weight matrix for the j-th attribute
X_{one_hot} : The one-hot encoded representation of the sample dataset
$X_{continuous}$: The continuous representation of the sample dataset

이때 학습을 위해 활용되는 연속형 공간이 샘플링과 구조적 제로의 구분을 위한 고정된 분포 영역을 이루도록 설정해야 하며, 이후 합성인구 생성모델에 동일한 데이터가 입력값으로 활용되어야 하므로 샘플 인구 데이터를 학습하여 연속형 공간을 추출하였다.



<Fig. 6> Entity embedding for converting the discrete distribution boundary to a continuous

3. 합성인구 생성 모델 프레임워크

합성인구 생성을 위한 심층 생성모델은 주어진 샘플 데이터를 기반으로 결합확률분포(joint probability distribution)를 추정하기 위해 학습하며, 전반적인 모델 구조는 <Fig. 7>과 같다. 이는 인구의 전반적인 특성을 가정하는 것이 아닌, 분석 대상지 고유의 인구 특성을 생성해내는 것이 주요 목적이므로 실제 인구에 존재하는 합성인구를 만들어 내는 것과 희귀하게 존재하는 합성인구를 만들어 내는 것이 주요 목적이다. 합성인구 GAN은 다양한 생성모델의 유형 중 Implicit generative model로 분류되며 주어진 데이터의 확률 분포를 그대로 학습할 수 있는 모델을 만들어 내는 것이 주요 목적이다. 그러나 판별자 D와 생성자 G를 순차적으로 반복 학습시킬 때, 학습 안정성 측면에서 종종 불안정하거나 수렴에 실패하는 경우가 있고, 목표로 한 생성 데이터의 속성 조합이 매우 복잡하고 희귀하다면 실제 인구와 같이 유사한 인구 특성 정보를 추정하는것에 어려움이 존재할 수 있다. 따라서 이러한 문제를 해결하기 위해 학습의 안정성을 보장하고 다양하고 희귀한 데이터 생성이 가능한 Wasserstein GAN(WGAN)을 채택하여 합성인구 생성에 적용하였다(Goodfellow et al., 2014; Gulrajani et al., 2017). 또한 식 (3)부터 식 (5)까지 표현된 바와 같이 기본적인 GAN 모델은 Min-Max 알고리즘 형식으로 수렴할 때, 학습을 더욱 안정적으로 수행하기 위해 식 (6)와 같이 Gradient penalty를 적용한 WGAN을 합성인구 생성을 위한 모델로 구축한다.

$$\min_G \max_D \underset{X\sim P_{data}(X)}{E}\left[\log(D(X))\right] + \underset{z\sim P_G(z)}{E}\left[\log(1-D(G(z)))\right] \quad \cdots\cdots (3)$$

$$L_D = \frac{1}{m}\sum_{i=1}^{m} -D(X_i) + D(G(z_i)) \quad \cdots\cdots (4)$$

$$L_g = \frac{1}{m}\sum_{i=1}^{m} -D(G(z_i)) \quad \cdots\cdots (5)$$

$$L_{GP} = \frac{1}{m}\lambda\left[\left(\left|\nabla_{\tilde{x}} D(\tilde{x})\right|_2 - 1\right)^2\right] \quad \cdots\cdots (6)$$

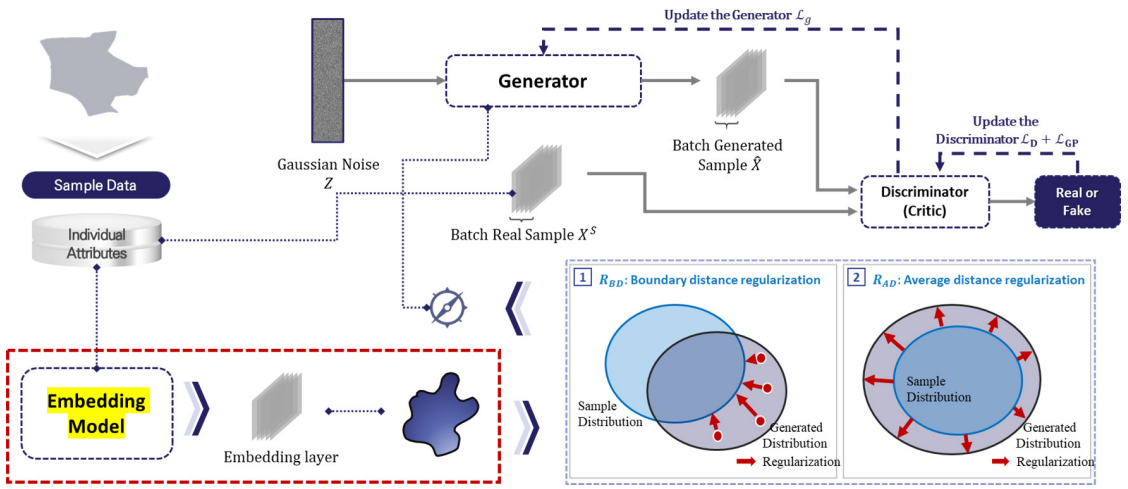
Where, m : The size of the mini-batch from the training dataset
z : The random noise distribution
$\tilde{X}_i$: The weighted average data of the generated data $\hat{X}_i$ and real data X_i

이후 위의 생성모델을 기반으로 샘플링 및 구조적 제로를 구분하기 위한 정규화 손실함수를 정의한 뒤 기존의 손실함수에 일정 가중치를 부여하여 생성자의 손실함수에 추가하게 된다. 먼저 구조적 제로를 최소화하기 위한 정규화 손실함수는 식 (7)과 같이R_{BD} (boundary distance regularization)으로 정의하고, 생성모델로부터 도출된 표본이 기준 분포 영역으로 정의된 X^S와의 거리와 멀어지지 않도록 제한하여 정확성을 높인다. 샘플링 제로를 최대화하기 위한 정규화 손실함수는 식 (8)와 같이 R_{AD} (average distance regularization)로 정의되며, 생성된 표본들의 평균 거리가 기준 분포 영역과 멀어지도록 유도함으로써 샘플링 제로뿐만 아니라 결측 표본의 생성도 유도하여 다양성을 높인다. 해당 정규화 손실함수들은 서로 상이한 목적을 가지고 균형을 맞춘 최적의 상태로 수렴해야 하므로, 본 연구에서는 이전 연구를 참고하여 잘 정의된 정규화 손실에 대한 가중치와 하이퍼파라미터의 최적화된 설정을 유지했다. 따라서 α와 β는 각각 0.5와 0.005로 설정되었고, 생성모델의 총손실함수는 수식 (9)의 L과 같이 정의된다(Kim and Bansal, 2023).

$$R_{BD}(\hat{X},X^S) = \frac{1}{m}\sum_{j=1}^{m} \min_{i\in 1:N, j\in 1:m} (Dist(\hat{X}_j, X_i^S)) \quad \cdots\cdots (7)$$

$$R_{AD} = -\frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{N} Dist(\hat{X}_j, X_i^S) \quad \cdots\cdots (8)$$

$$L = L_D + L_g + L_{gp} + \alpha R_{BD} + \beta R_{AD} \quad \cdots\cdots (9)$$



<Fig. 7> The overall framework of the population synthesis with the entity embedding

최종적으로 생성된 합성인구는 정확성과 다양성 측면에서 실제 인구 데이터와 성능을 비교하게 된다. 일반적으로 기존에 평가되는 방식인 주변확률분포(marginal distirbution)의 검증 방식은 범주형 데이터의 생성개수만을 판단하므로 일률적인 측면에서의 검증이 이루어지며, 특정 인구 특성 데이터를 반복적으로 생성했을

가능성이 존재하므로 과적합 여부의 판단이 불가능하다. 따라서 머신러닝 분류 성능 평가지표로 주로 사용되는 정밀도(precision)와 재현율(recall)을 도입하여 실제 인구수 M만큼 생성된 합성인구의 성능을 평가하였다. 이때 정밀도는 식 (10)과 같이 실제 인구 데이터 대비 합성인구 데이터의 인구 특성 조합 비율을 판단하고, 재현율은 반대로 식 (11)과 같이 합성인구 데이터 대비 실제 인구 데이터의 인구 특성 조합 비율을 판단한다. 최종적으로 정밀도와 재현율을 모두 고려해야 하므로 F1 Score 값을 식 (12)와 같이 계산하여 성능을 평가하였다.

$$\mathrm{Precision} = \frac{1}{M}\sum_{j=1}^{M} 1_{\hat{X}\in X_i} \quad \cdots\cdots (10)$$

$$Recall = \frac{1}{M}\sum_{j=1}^{M} 1_{X_i\in \hat{X}} \quad \cdots\cdots (11)$$

$$F1\,Score = \frac{2\times \mathrm{Precision}\times Recall}{\mathrm{Precision}+Recall} \quad \cdots\cdots (12)$$

Where, M : The total number of the real population and generated sample

Ⅳ. 결과 및 고찰

앞서 언급한 연속형 공간 생성의 문제는 MLM의 정확도가 높을수록 더 정확한 분포 영역을 정의할 것으로 가정하였다. 이를 위해 개체 임베딩이 다양한 인구 특성 조합을 고려한 정확한 예측을 수행할 수 있도록 새롭게 마스킹된 조합을 순차적으로, 그리고 반복적으로 학습하는 과정을 10번 수행하였다. 이때 모델별로 랜덤하게 생성된 마스킹 조합에 따라 예측된 인구 특성 정보의 정확도는 <Table 2>와 같으며, 밑단에 최상의 결과 모델에서 관측된 검증 데이터 실험의 손실 값이 제시되었다. 이때 모든 모델이 반복 학습을 수행하면서 초기 마스킹 조합보다 정확도가 개선되었다. 그러나 예상과 다르게 Seq2Seq 모델이 가장 낮은 정확도와 높은 검증 손실 값을 보였다. Dense, Seq2Seq, Multi-head Attention, 그리고 Transformer 모델의 경우 큰 폭 없이 일정 단계에서 모델의 정확도가 더 이상 개선되지 않았으나, LSTM의 경우 반복 학습이 거듭될수록 Masking 8을 기점으로 정확도가 감소했다. 따라서 해당 모델은 정확도가 높은 예측 임베딩 공간 추출이 가능하지만, 과적합의 가능성이 남아있음을 시사한다. 초기 마스킹 조합부터 안정적인 학습과 높은 정확도를 보인 모델은 Transformer이다. 최근 ChatGPT와 같은 대규모 언어 모델의 근간이 되는 모델인 만큼, 합성인구 특성 예측을 위해 적용된 MLM에서도 가장 낮은 성능을 보인 Seq2Seq 대비 약 21% 증가한 수치를 보였다.

<Table 2> Masking prediction result of the proposed models

Result type	Dense	LSTM	Seq2Seq	Multi-head Attention	Transformers
Masking 1	0.789	0.789	0.682	0.798	0.821
Masking 2	0.783	0.790	0.713	0.801	0.816
Masking 3	0.804	0.813	0.714	0.813	0.819
Masking 4	0.800	0.807	0.702	0.810	0.817
Masking 5	0.801	0.812	0.721	0.812	0.818

Result type	Dense	LSTM	Seq2Seq	Multi-head Attention	Transformers
Masking 6	**0.806**	0.819	0.723	**0.817**	0.816
Masking 7	0.797	**0.820**	0.730	0.813	0.821
Masking 8	0.801	0.814	0.713	0.810	0.810
Masking 9	0.794	0.818	**0.733**	0.810	0.819
Masking 10	0.801	0.808	0.731	0.813	**0.822**
Test Loss	7.416	5.934	10.504	6.061	**5.433**

Vanilla WGAN 모델을 활용하여 위의 다양한 MLM 모델로부터 얻어진 임베딩 레이어를 기반으로 정의된 연속형 공간을 정의한 뒤 정규화 손실함수를 적용한 실험 결과는 <Table 3>와 같다. 본 연구의 비교기준 모델은 기준은 앞서 가정한 샘플 인구 데이터 속성 조합으로 구성된 훈련 데이터로 구성하였고, 실제 인구 데이터와 동일한 주변확률분포를 갖도록 실제 인구수만큼 샘플 데이터를 복제하여 합성인구를 생성하므로 개인 특성 조합의 다양성을 고려하지 못한다. Vanilla WGAN의 경우 샘플링과 구조적 제로를 구분하기 위한 정규화 손실함수가 적용되지 않은 표준형 WGAN을 의미한다. 이때 이산형 공간을 정의하여 정규화 손실을 고려한 학습을 진행한 뒤 생성된 합성인구는 Vanilla WGAN 보다 정확도 측면에서 월등히 높은 성능을 보였으나 인구 특성 조합의 다양성 측면에서는 상당히 저하되었다. 이는 모델 학습 시 분포 영역과 생성된 표본 데이터의 거리 측정이 정밀히 수행되지 않아 다양성을 생성하기 위한 생성모델의 확률 분포 영역을 효과적으로 늘이지 못하였고, 학습 시에 R_{BD} 의 손실이 R_{AD} 보다 더욱 낮아지도록 수렴된 것을 확인할 수 있다. 앞서 우려한 대로 Dense, Seq2Seq, 혹은 Multi-head Attention 모델과 같이 임베딩 추출 시 본래의 데이터를 잘 표현하지 못한 채 변형된 연속형 공간을 적용한 실험한 결과는 Vanilla WGAN 보다 같거나 더 낮은 성능을 보였다. LSTM 모델의 경우 높은 재현율을 달성하였지만, 여전히 낮은 F1 Score 값을 보였다. 하지만 Transformer 모델을 통해 잘 정의된 연속형 공간을 적용하였을 때 정밀도와 재현율 측면에서 각각 약 7.6%와 95.94%의 성능 개선율을 보였으며 최종적으로 F1 Score 값이 Baseline 대비 28.87% 증가하였다. 특히 이산형 공간을 정의하는 것 대신 Transformer 기반으로 연속형 공간을 변형했을 때 다양성의 측면의 개선율이 24.40%로 크게 증가하였으므로, 생성모델 기반 다양성을 생성해내는 데 큰 영향을 미치는 것으로 분석되었다. 이는 또한 기존에 제시되었던 정규화 손실함수가 적절한 환경에서 생성모델에 정확하고 다양한 합성인구를 만들어 내는 것에 매우 효과적임을 보였다.

<Table 3> Population synthesis performance comparison

Methods	Precision	Recall	F1 Score
Baseline	1	0.320	0.485
Vanilla WGAN	0.579	0.616	0.597
Discrete space	**0.777**	0.504	0.611
Continuous space (Dense)	0.590	0.599	0.594
Continuous space (LSTM)	0.602	0.619	0.610
Continuous space (Seq2Seq)	0.597	0.597	0.597
Continuous space (Multi-head Attention)	0.589	0.591	0.590
Continuous space (Transformers)	0.623	**0.627**	**0.625**

V. 결 론

이 논문에서는 합성인구를 위한 심층 생성모델 학습 과정 중 정규화 손실함수의 성능을 향상하기 위해 활용되는 연속형 공간을 정교하게 정의하기 위한 개체 임베딩 기법을 연구하였다. 정규화 손실함수는 학습 시에 범주화되어있는 샘플 데이터의 분포 영역을 기준으로 거리를 계산하고 생성된 데이터의 정확성과 다양성을 향상할 수 있다. 이때 해당 과정을 정교하게 수행하기 위한 연속형 데이터가 필요하지만, 왜곡된 형태로 변환할 시에 데이터 구분의 오류로 인한 성능 저하가 발생할 수 있다. 따라서 이를 방지하기 위한 과정으로 이산형 데이터를 연속형 데이터로 변환하는 것뿐만 아니라 인구 특성 정보의 관계와 의미를 잘 학습할 수 있게 개발된 자연어 처리 모델을 도입하여 개체 임베딩 성능을 증가시켰고, 이를 통해 변환한 연속형 데이터를 활용하여 간접적으로 합성인구 생성모델의 성능을 증가시킬 수 있었다. 결과적으로 이산형 샘플 데이터를 연속형 데이터로 변환하는 과정이 합성인구 생성에 큰 영향을 미칠 수 있음을 발견하였고, 기존 연구에서 제시된 정규화 손실함수가 매우 효과적임을 입증하였다. 그러나 여전히 현실에서 실제 인구정보를 얻는 것이 불가능하고, 이에 따라 샘플링과 구조적 제로를 명확히 구분하는 것은 쉽지 않은 문제이다. 또한 본 연구에서 제시한 생성모델은 두 가지 손실함수에 대한 균형 관계를 유지한 학습이 필요하므로 섬세한 파라미터 조정이 요구된다. 또한 생성된 합성인구를 공간적으로 분배하는 과정이 필수적이지만, 데이터의 복잡성이 비약적으로 증가하여 생성모델 적용에 한계점이 존재한다. 추가로, 본 연구에서 제시한 방법론은 실제 합성인구 적용 시에 수집된 인구 샘플 데이터가 실제 인구 데이터와 얼마나 유사한지 확인할 수 없으므로, 수집된 데이터에 이미 큰 편향이 존재할 때 적합하지 않을 수 있다. 따라서 실제 활동 기반 모델에 적용하기 위한 데이터 퓨전 기법 등의 심층적인 연구가 추가로 진행되어야 한다.

ACKNOWLEDGEMENTS

본 연구는 국토교통부 및 국토교통과학기술진흥원의 지원으로 수행되었음(과제번호 RS-2022-00141102)

REFERENCES

Aemmer, Z. and MacKenzie, D.(2022), “Generative population synthesis for joint household and individual characteristics”, *Computers, Environment and Urban Systems*, vol. 96, 101852.

Borysov, S. S., Rich, J. and Pereira, F. C.(2019), “How to generate micro-agents? A deep generative modeling approach to population synthesis”, *Transportation Research Part C: Emerging Technologies*, vol. 106, pp.73-97.

Castiglione, J., Bradley, M. and Gliebe, J.(2015), *Activity-based travel demand models: A primer*, Transportation Research Board.

Chen, T., Tang, L. A., Sun, Y., Chen, Z. and Zhang, K.(2016), *Entity embedding-based anomaly detection for heterogeneous categorical events*, arXiv preprint arXiv:1608.07502.

Devlin, J., Chang, M. W., Lee, K. and Toutanova, K.(2018), *Bert: Pre-training of deep bidirectional transformers for language understanding*, arXiv preprint arXiv:1810.04805.

Garrido, S., Borysov, S. S., Pereira, F. C. and Rich, J.(2020), “Prediction of rare feature combinations in population synthesis: Application of deep generative modelling”, *Transportation Research Part C: Emerging Technologies*, vol. 120, 102787.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y.(2014), “Generative adversarial nets”, *Advances in Neural Information Processing Systems*, vol. 27.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V. and Courville, A. C.(2017), “Improved training of wasserstein gans”, *Advances in Neural Information Processing Systems*, vol. 30.

Guo, C. and Berkhahn, F.(2016), *Entity embeddings of categorical variables*, arXiv preprint arXiv:1604.06737.

Hancock, J. T. and Khoshgoftaar, T. M.(2020), “Survey on categorical data for neural networks”, *Journal of Big Data*, vol. 7, no. 1, pp.1-41.

Kim, E. J. and Bansal, P.(2023), “A deep generative model for feasible and diverse population synthesis”, *Transportation Research Part C: Emerging Technologies*, vol. 148, 104053.

Kim, E. J., Kim, D. K. and Sohn, K.(2022), “Imputing qualitative attributes for trip chains extracted from smart card data using a conditional generative adversarial network”, *Transportation Research Part C: Emerging Technologies*, vol. 137, 103616.

Palangi, H., Deng, L., Shen, Y., Gao, J., He, X., Chen, J., Song, X. and Ward, R.(2016), “Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval”, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp.694-707.

Sutskever, I., Vinyals, O. and Le, Q. V.(2014), “Sequence to sequence learning with neural networks”, *Advances in Neural Information Processing Systems*, vol. 27.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I.(2017), “Attention is all you need”, *Advances in Neural Information Processing Systems*, vol. 30.

Xu, L. and Veeramachaneni, K.(2018), *Synthesizing tabular data using generative adversarial networks*, arXiv preprint arXiv:1811.11264.