JDFR Journal of Data Forensics Research
데이터포렌식연구, 2권 2호, 2025년 12월, 111-142면
Journal of Data Forensics Research, Volume 2, Issue 2, December 2025, Pages 111-142.
https://doi.org/10.12972/JDFR.2025.2.2.1
Print ISSN 3058-9460
Online ISSN 3058-9479

원저

범죄 수사용 합성 통신데이터(synthetic CDR) 생성 모델의 설계 및 활용 연구

박재만

대구경찰청 전문수사관

교신저자: 박재만, ajkrap@gmail.com

요약

본 연구는 개인정보와 위치정보를 포함한 통신내역의 특성상 실제 데이터를 연구·교육 목적으로 활용하기 어려운 수사 환경의 구조적 한계를 해결하기 위해, 범죄 수사용 합성 통신데이터(synthetic CDR) 생성 모델을 제안한다. 기존 해외 연구에서는 대규모 모바일 통화기록을 활용한 이동성 및 사회적 연결망 분석, 모바일 CDR 기반 네트워크 이상탐지 및 트래픽 예측, CDR 통계분포를 모사한 synthetic CDR 생성기 개발 등이 연구되어 왔으나, 피의자–공범 간 관계 구조, 기지국 기반 위치 이동, 시간·행동 패턴을 통합적으로 반영한 범죄수사용 합성 통신데이터 모델은 제시된 바 없다. 국내 연구에서도 합성데이터의 활용 필요성이 꾸준히 제기되고 있지만, 실제 구현 가능한 알고리즘이나 수사용 통신데이터 생성 연구는 부재한 상황이다. 본 연구는 이러한 공백을 해소하기 위해 한국의 번호체계, 행정구역 기반 공간 구조, 시간대별 통신 패턴, 이동 특성, 범죄 유형 기반 시나리오 등을 반영한 고현실성 통신데이터 생성 프레임워크를 설계하였다. 제안된 모델은 전화번호 생성, 시간 이벤트 모델링, 기지국 기반 위치 시뮬레이션, 관계망 구조 형성, 통신 패턴 규칙 등의 모듈로 구성되며, 실제 수사에서 관찰되는 통신 패턴과 유사한 통계적 특성을 재현하도록 설계되었다. 생성된 데이터는 다중 피의자, 다지역, 장기간 분석이 가능한 구조를 가지며, 수사 분석기법 훈련, 패턴·이상탐지 알고리즘 검증, 데이터포렌식 교육, AI 기반 수사기술 개발 등 다양한 활용이 가능하다. 연구 결과, 제안된 synthetic CDR은 시간대별 통화량, 통신 빈도, 이동 경로, 관계망 클러스터링 등 주요 분석 지표에서 실제 통신내역과 유사성을 보였다. 본 연구는 국내 최초로 한국 수사 환경에 특화된 범죄 수사용 synthetic CDR 생성 모델을 제시함으로써, 민감 데이터 활용의 제약을 극복하고 수사 R&D의 발전 기반을 마련했다는 점에서 중요한 학술적·실무적 의의를 가진다.

주제어

가상데이터, 통신내역, 합성 통신내역, 합성데이터 생성 모델, 수사용 데이터

Received: December 10, 2025
Revised: December 30, 2025
Accepted: December 31, 2025
Published: December 31, 2025

Original Article

Study on the design and application of synthetic call detail record generation models for criminal investigation

Jaeman Park

Specialized Investigator, Daegu Metropolitan Police Agency, Republic of Korea

Corresponding Author: Jaeman Park, ajkrap@gmail.com

ABSTRACT

This study proposes a synthetic call detail record (CDR) generation model that is designed to overcome the limitations of using real communication data in investigative research and training. Existing international studies have explored mobility analysis, social network structures, anomaly detection, and statistical CDR simulation; however, none have addressed suspect–accomplice relationships, cell-tower-based mobility, or temporal–behavioral patterns required for investigative applications. To fill this gap, the proposed framework incorporates the Korean numbering system, administrative spatial structure, time-dependent communication patterns, mobility characteristics, and rule-based communication behaviors. The model reproduces realistic statistical properties across call frequencies, temporal distributions, mobility traces, and network clustering, enabling its use in crime analysis training, anomaly detection research, digital forensics education, and AI-based investigative tool development. The results demonstrate the high fidelity of the model to real CDR patterns, establishing a foundational contribution to synthetic investigative data generation in Korea.

KEYWORDS

synthetic data, call detail records(CDRs), synthetic CDR, synthetic data generation model, investigative data

I. 서론

통신내역(call detail records, CDR)은 피의자 특정, 공범 관계 규명, 기지국 기반 위치 추정, 통신 패턴 분석 등 핵심적 역할을 수행하는 수사자료이다. 수사기관은 통신내역을 통해 특정 시점·지역에서의 발신자·수신자 정보를 확인하고, 반복·집중되는 통신 패턴과 번호 간 연결 구조를 분석함으로써 범죄조직의 활동 양상과 역할 분담을 추론할 수 있다.

그러나 CDR에는 전화번호, 통화 시각, 기지국 위치 등 고위험 개인정보가 포함되어 있어, 연구·교육·기술 검증과 같은 2차 활용은 법적·제도적 제약으로 인해 극히 제한적이다. 그 결과, 수사기관 외부는 물론 내부 R&D·교육훈련 분야에서도 충분한 규모와 구조를 갖춘 통신데이터를 확보하기 어렵고, 새로운 분석 기법이나 알고리즘을 시험·비교·검증할 수 있는 환경을 구축하는 데 어려움이 지속되고 있다.

특히 통신내역은 단독 분석보다 계좌거래, 교통·위치 정보, CCTV 등 다른 수사자료와의 결합을 통해 의미가 확대되지만, 실제 CDR을 연구·훈련용으로 제공하기 어렵기 때문에 이러한 융합 분석을 반복적으로 실험·검증할 수 있는 데이터 환경이 사실상 부재한 상황이다. 수사관 교육과정에서도 실무 사건에서 일부 발췌된 예시 데이터를 제한적으로 활용하는 수준에 그치는 경우가 많아, 대규모·복합 구조를 가진 통신데이터를 기반으로 한 체계적 훈련에는 한계가 존재한다.

이처럼 실제 CDR은 수사 실무에서는 필수적인 자료이지만, 개인정보보호와 제도적 규제로 인해 연구·교육·기법 검증용 데이터로서는 활용 여지가 극히 제한되는 이중적 성격을 가진다. 따라서 원본 CDR을 직접 사용하지 않으면서도 통신데이터의 구조적·통계적 특성을 재현할 수 있는 새로운 데이터 생성 방식이 요구된다.

이러한 요구에 대응하는 기술 중 하나가 합성데이터(synthetic data)이다.

합성데이터는 특정 목적을 위해 원본 데이터의 형식·구조, 통계적 분포 특성과 패턴을 학습하여 생성한 모의(simulated) 또는 가상(artificial) 데이터로서, 실제 개인의 식별정보를 직접 포함하지 않으면서도 원본 데이터의 분석적 가치는 유지하도록 설계된 데이터이다. 이러한 합성데이터는 개인정보 침해 위험을 최소화하면서 연구·산업 분야에서 데이터 활용을 가능하게 하는 기술로 주목받고 있으며, 특히 개인정보 규제로 인해 원본 데이터의 직접 활용이 제한되는 환경에서 합리적인 대안으로 평가받고 있다[1]. 그러나 범죄수사·디지털포렌식 영역, 특히 한국의 번호 체계·행정구역·기지국 구조·통신 습관을 반영한 수사용 synthetic CDR(합성 통신내역)에 대한 체계적 연구는 아직 초기 단계에 머무르고 있다.

이에 본 연구는 실제 수사에서 활용되는 통신내역의 구조와 패턴을 참고하여, 피의자와 공범, 주변 인물 간 통신 관계, 기지국 기반 발신 위치, 시간대별 통신 특성 등을 반영한 범죄 수사용 synthetic CDR 생성 모델을 제안하고, 그 활용 가능성을 수사·디지털포렌식 관점에서 검토하고자 한다. 이를 통해 실제 CDR을 직접 사용하지 않고도 수사 R&D, 교육·훈련, 알고리즘 검증 등에서 활용 가능한 범죄 수사용 합성 통신데이터 환경을 구축하는 것을 목표로 한다.

1.1. 연구배경 및 필요성

디지털 환경의 확장과 이동통신의 고도화로 인해 범죄수사에서 통신데이터의 중요성은 지속적으로 증가하고 있다. 예를 들어, 특정 지역 내 중복 발신자 탐지(기지국 수사), 조직형 범죄의 통신 네트워크 분석, 특정 통화자 또는 특정 지역에 집중되는 통신 패턴을 통한 관계 및 활동 추정 등은 이미 실무에서 널리 활용되는 분석 절차이다.

그럼에도 불구하고 실제 CDR은 통신비밀보호법, 개인정보보호법 등 관련 법령에 따라 엄격히 보호되며, 수사 목적 이외의 연구·교육·기술 검증 용도로는 활용이 거의 불가능하다. 수사기관 내부에서도 실제 사건 자료를 교육·훈련·알고리즘 개발에 반복적으로 사용하는 데에는 사건 당사자의 개인정보, 사건 내용의 민감성, 재사용에 따른 법적·윤리적 문제 등 다수의 제약이 존재한다. 이러한 환경은 통신데이터 기반 수사기법 고도화와 AI·데이터 분석 기반 수사 R&D에 구조적인 한계를 초래한다.

이러한 법적·제도적 제약으로 인해 실제 CDR을 연구·교육 목적에 활용할 수 없게 되면서, 실무에서는 개별 사건에서 확보된 제한된 기간·범위의 통신내역만으로 분석을 수행하는 경우가 많다. 이 때문에 충분한 규모와 다양한 패턴을 갖춘 데이터셋을 확보하기 어렵고, 이를 위한 표준화된 연구용·교육용 데이터 또한 거의 존재하지 않는다.

한편, 합성데이터는 실제 개인의 정보 없이도 통계적 특성과 구조적 패턴을 재현할 수 있어, 개인정보보호와 데이터 활용 요구를 동시에 충족시키는 수단으로 주목받고 있다. 그러나 국내 형사사법 분야의 합성데이터 논의는 주로 개념·정책 수준에 머물러 있으며, 실제 수사 패턴과 통신 환경을 반영한 synthetic CDR의 설계·생성·평가에 관한 실증 연구는 거의 이루어지지 않았다. 특히 한국 고유의 010 기반 번호 체계, 행정구역 단위 기지국 주소, 수사 실무에서 자주 관찰되는 중복 통신·허브 노드·주변부 노드(peripheral node) 등의 구조를 반영한 합성 통신데이터 모델은 부재한 상황이다.

따라서 실제 CDR을 직접 활용하지 않고도 수사 실무에서 사용하는 분석 관점(예를 들어, 다중 기지국 중복 발신자 분석, 조직범죄 네트워크 구조 분석)을 재현할 수 있는 범죄 수사용 synthetic CDR 생성 모델이 필요하다. 이러한 모델은 수사 R&D, 교육·훈련, 분석 도구 검증, AI 모델 개발 등 다양한 영역에서 활용될 수 있는 안전한 데이터 기반을 제공할 수 있을 것이다.

1.2. 연구 목적

본 연구의 목적은 실제 통신내역을 직접 제공할 수 없는 수사 환경을 고려하여, 한국형 통신 환경과 수사 실무에서 관찰되는 통신 패턴을 반영한 범죄 수사용 synthetic CDR 생성 모델을 설계하고, 그 활용 가능성을 실증적으로 제시하는 데 있다. 구체적인 연구 목적은 다음과 같다.

첫째, 한국의 전화번호 체계, 행정구역 기반 기지국 주소 구조, 시간대별 통신 특성 등을 반영하여, 한 건의 통신 기록이 하나의 행(row)에 대응되도록 설계된 합성 통신데이터 생성 모델을 개발한다.

둘째, 피의자와 주변 인물 간의 발신·착신 관계 구조를 모사하고, 다회 통신 상대방, 허브 노드, 주변부 노드 등 수사 실무에서 자주 분석되는 네트워크 특성을 반영한 synthetic CDR을 생성한다.

셋째, 생성된 synthetic CDR에 대해 기지국 기반 중복 발신자 탐지, 주요 대상자를 중심으로 한 통신 관계망 분석, 시간·공간 패턴 분석 등을 수행하여 실제 수사에서 사용하는 분석 관점과 구조적 특성을 어느 정도 재현하는지 검증한다.

넷째, 제안된 synthetic CDR 모델이 수사관 교육·수사용 데이터 분석 훈련, 분석 도구 및 알고리즘 검증, 수사 R&D 및 AI 모델 개발 등 다양한 실무 영역에서 활용 가능한지를 논의하고, 향후 합성 수사데이터 생태계 구축 방향을 제시한다.

본 연구의 결과는 실제 CDR을 직접 사용할 수 없는 상황에서도 수사기관이 안전하고 유연하게 활용할 수 있는 합성 통신데이터 기반 분석·훈련 환경을 마련하고, 데이터 기반 수사 역량 강화 및 공공 안전 확보에 기여할 것으로 기대된다.

II. 선행연구

Synthetic CDR 또는 통신기록 기반 데이터 모형화 연구는 해외·국내에서 지속적으로 이루어져 왔으나, 연구의 목적과 분석 관점에 따라 상이한 방향으로 발전해왔다. 본 장에서는 기존 문헌에서 통신기록이 어떻게 활용되었는지 검토하고, 이를 바탕으로 본 연구가 지향하는 범죄 수사용 synthetic CDR 생성 연구의 학술적 위치를 정립하고자 한다.

2.1. 해외의 연구 동향

해외 연구는 주로 대규모 통화기록 데이터를 활용하여 인간 이동성, 사회적 연결 구조, 이상행동 탐지 등 분석 중심의 방향으로 발전해왔다.

우선, J. Candia 등(2008)은 대규모 모바일폰 기록을 분석하여 개인 및 집단의 이동 경로와 통신 활동이 단순한 우연적 사건의 집합이 아니라 시간·공간적으로 반복되는 규칙적 패턴을 보인다는 점을 확인하였다. 예컨대, 기지국 단위 통화량은 업무 시간대·주거 시간대 등 일상생활 리듬에 따라 일관된 주기성을 보였으며, 개인의 통화 간 시간 간격(interevent time) 역시 heavy-tailed 분포를 따르되, 평균 통화빈도로 정규화하면 집단 간 차이가 사라지고 동일한 보편적 분포로 수렴하는 현상이 관찰되었다[2]. 이러한 결과는 CDR이 인간의 이동 및 사회적 연결 활동을 안정적으로 반영하는 데이터라는 점을 보여주며, 합성 통신데이터를 생성할 때도 이러한 규칙적 분포 특성을 고려하는 것이 중요함을 시사한다.

한편, K. Sultan 등(2018)은 실제 모바일 CDR을 기반으로 네트워크 이상탐지 및 트래픽 예측 모델을 제안하였으며, 시계열 기반 통신 패턴의 변동성이 네트워크 행위 분석의 핵심 지표가 될 수 있음을 확인하였다[3]. 이는 합성 CDR에서도 시간대별 통신 비율·이상행동 편차 등을 재현해야 분석 실험에 신뢰성을 부여할 수 있음을 보여준다.

또한, Songailaite & Krilavicius(2021)는 실제 CDR 분포를 기반으로 통화 지속시간, 발신·착신 시각, 통신 성공·실패 비율 등 통계적 특성을 모사한 Synthetic CDR 생성기를 제안하였다[4]. 다만 이 연구는 통계적 분포 기반 모델에 중심을 두고 있어, 관계형 네트워크 구조나 시간·기지국 기반 행동 특성 등 수사용 Synthetic CDR이 요구하는 관계적·행위적 패턴은 반영하지 못했다.

종합하면, 해외 연구는 실제 CDR의 사회적·공간적·행태적 분석과 일부 합성데이터 생성 연구로 발전해 왔으나, 범죄 수사 목적의 synthetic CDR 생성 연구는 여전히 부재한 상황이다.

2.2. 국내의 연구 동향

국내에서도 개인정보보호 규제에도 불구하고 통신데이터의 네트워크적 가치와 합성데이터 활용 필요성을 강조하는 연구가 등장하였다.

먼저, 이진호(2014)는 실제 CDR을 기반으로 소셜 컨택 네트워크를 구성하고, 해당 네트워크가 스케일 프리(scale-free) 특성을 갖는다는 점을 실증하였다[5]. 여기서 스케일 프리 구조란, 노드의 연결 정도 분포가 멱함수(Power-law) 형태를 띠어 소수의 노드가 매우 많은 연결을 보유하는 반면 다수의 노드는 적은 연결만을 갖는 허브 중심 비대칭 구조를 의미한다. 이는 평균값 중심으로 설명되는 일반적인 분포와 다른 특성으로, 실제 사회적 관계망이 특정 소수 노드에 연결이 집중되는 경향을 보인다는 점을 보여준다. 이 연구는 발신·착신 관계망 분석이 중심성, 허브 노드, 연결 구조 등 사회적 관계 패턴을 밝히는 데 효과적임을 확인하였으며, 실제 CDR의

구조와 패턴을 모사하여 생성되는 synthetic CDR에서도 반복적 상대방·공통 상대방(shared contact)·관계 편중 등 실제 네트워크 특성을 충실히 반영해야 함을 시사한다.

또한, 오진호·강전석·류연승(2025)은 형사사법정보의 민감성과 개인정보보호 규제가 커지는 상황에서 비식별화만으로는 데이터 활용 한계를 극복하기 어렵다고 지적하며, 이를 보완하기 위한 기술로서 재현데이터(본 연구에서 사용하는 합성데이터와 동일한 개념)의 필요성을 제시하였다[6]. 연구는 synthetic Data가 수사 R&D·교육훈련·검증환경 구축에 활용 가능한 대안임을 명확히 언급하며, 본 연구의 필요성을 직접적으로 뒷받침한다.

그 외 국내 연구들에서도 CDR 기반 행동패턴 분석, 지역특성 연구 등이 수행되었으나, 연구 데이터의 규모·기간·범주가 제한되어 있으며, 수사환경에서 활용 가능한 synthetic CDR 생성 알고리즘을 제안한 사례는 존재하지 않는다. 특히 한국 고유의 010 번호 체계, 행정구역 기반 기지국 구조, 시간대별 통신 습관, 피의자 중심 통신 패턴 등을 반영한 모델은 확인되지 않는다.

2.3. 선행연구의 한계와 본 연구의 차별성

해외 연구는 대규모 실제 CDR을 활용하여 인간 이동성·사회적 행동 패턴·이상탐지 등에 중요한 성과를 냈으나, 범죄수사 목적의 synthetic CDR 생성에는 접근하지 않았다. 즉, 공범 구조, 반복적 통신, 시간·공간 기반 행위 패턴, 기지국 기반 위치 변화 등 수사적 요구 요소는 반영되지 않았다.

국내 연구는 통신기록의 네트워크 분석 가치와 synthetic Data 필요성을 강조하고 있지만, CDR을 실질적으로 모사하는 생성 알고리즘, 수사 시나리오 기반 합성모델, 범죄유형별 관계망 패턴을 재현하는 접근은 아직 보고되지 않았다.

따라서 본 연구는 다음 측면에서 차별성을 갖는다.

첫째, 국내 수사 실무에서 실제로 활용되는 기지국 수사, 중복 발신자 탐지, 조직범죄 네트워크 분석에 직접 적용할 수 있는 수사용 synthetic CDR 생성 모델을 최초로 제안한다.

둘째, 한국의 010 기반 번호 체계, 행정구역 기반 기지국 주소, 대상자 중심 발·착신 기록 구조 등 국내 통신·수사 환경을 반영하여 synthetic CDR 생성 프레임워크를 구축하였다.

셋째, 실제 수사 시나리오를 그대로 반영할 수 있는 관계망·시간·공간·행위 패턴을 통합 모델링하였다.

넷째, 생성된 synthetic CDR을 기지국 기반 용의자 특정 실험, 조직범죄 네트워크 분석 실험 등에 적용하여, 수사관 교육·R&D·AI 모델 검증에 활용 가능한 실무형 합성 수사데이터의 가능성을 실증하였다.

이와 같이 본 연구는 기존 분석 중심의 CDR 연구와 합성데이터 정책 논의 사이의 공백을 메우고, 국내 최초로 범죄 수사용 합성 통신데이터 생성 모델을 구체적으로 설계·실험·평가했다는 점에서 학술적·실무적 의의가 크다.

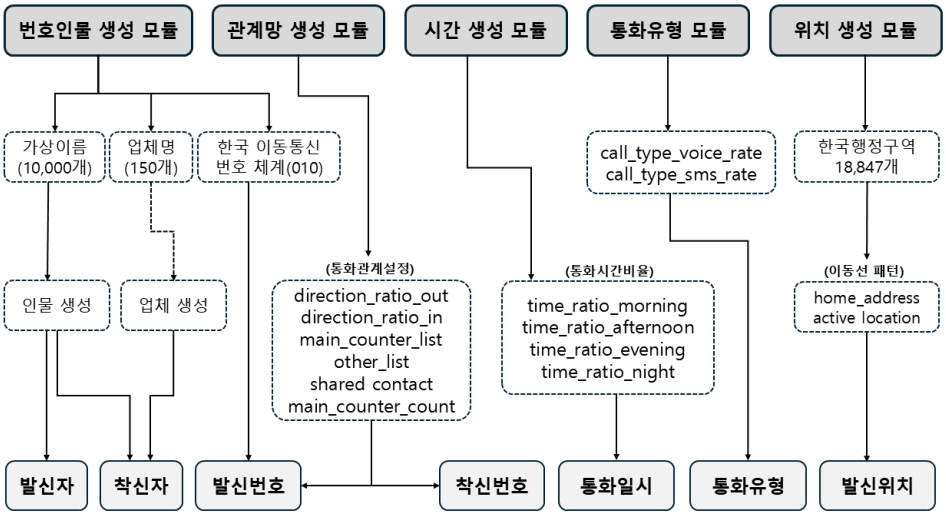
III. 연구방법

3.1. Synthetic CDR 생성 모델의 개요

본 연구에서는 실제 통신내역과 유사한 구조와 패턴을 갖는 고현실성 합성 통신데이터를 생성하기 위해 모듈 기반 synthetic CDR 생성 모델을 설계하였다. 본 모델은 한국의 전화번호 체계, 기지국 기반 위치 구조, 시간대별 통신 패턴, 피의자–참여자 간 관계망 구조, 통화 유형별 특

성 및 통신 상대방 간 상호작용 패턴 등을 반영하도록 구성되어 있다. 전체 모델은 번호 생성 모듈, 시간 이벤트 생성 모듈, 위치 생성 모듈, 관계망 생성 모듈, 통화 유형 및 상대방 관계 패턴 생성 모듈로 구성되며, 각 모듈의 결합을 통해 실제 CDR과 유사한 고품질 합성 데이터를 생성한다. 이러한 구조는 다양한 통신 패턴 조합을 현실적으로 구현함으로써 수사·포렌식 분석 환경에서의 활용성을 높인다.

이러한 모듈 간 상호작용을 보다 직관적으로 제시하기 위해, 본 연구에서 제안하는 synthetic CDR 생성 모델의 전체 아키텍처를 <Figure 1>에 제시하였다.



<Figure 1> Overall architecture of the synthetic CDR generation model

3.2. 데이터 구조 및 변수 정의

본 연구에서 생성한 합성 통신데이터는 실제 수사 실무에서 제공되는 통신내역(CDR) 형식을 참고하여, 한 건의 통화 기록이 하나의 행(row)에 대응되도록 설계하였다. 최종적으로 생성되는 합성 데이터의 각 레코드는 발신·착신 정보, 통화 시각, 통신사, 발신 위치, 통화 구분 등 수사 분석에 필요한 핵심 항목을 포함한다. 합성 통신내역의 기본 헤더 구조는 <Table 1>과 같다.

<Table 1> Field structure of the synthetic CDR

구분	변수명	설명	예시
기본 정보	발신인	통화를 발신한 사람의 성명(가상 인물 이름)	홍길동
	발신번호	발신 측 가입자 번호(가상 전화번호)	01012345678
	착신인	통화를 수신한 사람의 성명(가상 인물 이름)	이영희
	착신번호	착신 측 가입자 번호(가상 전화번호)	01098765432
시간 정보	시작일시	통화가 시작된 날짜와 시각(YYYY-MM-DD HH:mm:ss)	2025-01-01 12:00:00
	종료일시	통화가 종료된 날짜와 시각(YYYY-MM-DD HH:mm:ss)	2025-01-01 12:05:00
통신사 정보	업체명	통신 서비스 제공자(KT, SKT, LGU+)	SKT
위치 정보	발신위치	발신 기지국 위치 주소	대구시 수성구 무학로227
통신 유형	구분	통신 유형(음성통화/SMS 등)	국내음성통화
	비고	추가 정보가 기록된 메모 필드	

이와 같이 정의된 필드들은 실제 이동통신사업자의 통신사실확인자료와 수사관이 통상 분석하는 항목을 기준으로 선정한 것으로, 발신·착신 관계, 통화량, 시간대별 패턴, 지역 기반 이동성 분석 등 수사 실무적 분석에 필요한 필수 항목으로 구성하였다.

또한, 변수명을 모두 한글로 통일하여 수사 실무자들이 직관적으로 이해할 수 있도록 하였으며, 향후 AI 학습이나 추가 분석이 필요한 경우에는 동일 구조를 유지한 채 파생 변수(예: 통화시간, 요일, 시간대 구분, 상대방 관계 강도 등)를 손쉽게 추가할 수 있도록 확장성을 고려하였다. 이러한 구조는 향후 AI 기반 행위 분석 모델, 범죄 시나리오 검증 환경, 교육 훈련 데이터셋 구축 등 다양한 목적에 활용될 수 있는 기반을 제공한다.

3.3. Synthetic CDR 생성 알고리즘의 통합 파이프라인(Integrated Pipeline)

본 절에서는 synthetic CDR 생성 과정을 전체적인 흐름 차원에서 제시하고자 한다. 본 생성 절차는 번호·시간·위치·관계망·통화유형 등 통신데이터를 구성하는 모든 요소를 순차적으로 연결하는 파이프라인으로 구성되며, 수사 목적에 따라 다양한 규모와 유형의 합성 데이터를 생산할 수 있도록 설계되었다. 전체 절차는 모듈 간 상호작용을 기반으로 동작하며, 수백 건에서 수십만 건 규모까지 확장 가능한 구조를 갖는다.

3.3.1. 전체 처리 개요

Synthetic CDR 생성은 (1) 초기 설정값 로딩, (2) 통신 이벤트 생성, (3) 개별 모듈 처리, (4) 최종 레코드 통합의 4단계 구조로 이루어진다. 사용자가 입력한 대상자 정보, 시간대 비율, 통신 방향 비율, 활동지 목록, 통화유형 비율 등은 모든 모듈의 기반 파라미터로 사용되며, 이후 각 모듈은 이전 단계에서 생성된 출력값을 다음 단계의 입력값으로 활용한다.

이러한 설계는 실제 통신기록(CDR)의 생성 과정을 모사하면서도, 범죄유형별 특성을 반영한 합성 데이터 생성이 가능하도록 높은 유연성과 확장성을 제공한다.

3.3.2. 통합 파이프라인 절차 요약

Synthetic CDR 생성 과정은 다음의 순서로 수행된다.

<Table 2> Integrated execution procedure of the synthetic CDR generation pipeline

단계	처리 단계	주요 설정 및 처리 내용
1	인물 및 전화번호 초기화	한국 이동통신번호 체계(010-XXXX-XXXX)에 따라 전화번호를 생성하고, 대상자·상대방 명단을 구성한다.
2	주요 상대방 및 관계 가중치 설정	주요 통신 상대방 목록을 생성하고, 필요 시 shared_node_ratio 값을 적용하여 공통 상대방(shared contact)를 구성한다.
3	시간대 비율을 반영한 통신 이벤트 생성	time_ratio 변수를 기준으로 새벽·오전·오후·저녁 중 하나의 시간대를 선택하고, 해당 구간에서 발신 시각(start_time)을 생성한다.
4	시간대별 발신 위치 생성	home_address와 active_address_list를 기반으로, 오전에는 주거지, 오후·저녁에는 활동지를 적용하는 방식으로 발신 위치를 생성한다.
5	발신·착신자 네트워크 기반 매칭	OUT/IN 비율에 따라 발신자 또는 착신자를 대상자로 지정하고, 상대방은 주요 상대방 또는 일반 상대방 목록에서 선택한다.
6	통화 유형 및 통신사 정보 부여	음성통화와 SMS비율(call_type_voice_rate, call_type_sms_rate)에 따라 통신유형을 선택하고, 전화번호를 기반으로 통신사를 부여한다.
7	레코드 단위 데이터 통합	발신자·착신자·시간정보·위치·통신유형·통신사 정보를 하나의 CDR 행으로 통합한다.
8	Synthetic CDR 데이터 출력	모든 레코드를 리스트에 누적한 후 출력한다.

위 절차는 실제 CDR의 구조와 행동 기반 통신 패턴을 동시에 반영하여, 수사 분석·교육훈련·AI 모델 검증 등 다양한 실무 환경에서 사용 가능한 고현실성 합성 통신데이터를 생성할 수 있도록 설계되었다.

3.3.3. 통합 파이프라인의 특징

본 연구에서 제시한 합성 통신데이터 생성 절차는 다음과 같은 특성을 가진다.

첫째, 시간(temporal), 공간(spatial), 관계(network), 행위(behavioral), 유형(type)을 결합하여 실제 통신 패턴을 반영한다.

둘째, 파라미터 조정만으로 사건유형(보이스피싱, 불법도박, 조직범죄 등)별 synthetic CDR 생성이 가능하므로 높은 확장성과 재현성을 가진다.

셋째, 발신·착신 번호, 시간정보, 위치주소, 통신유형, 통신사 등 실제 CDR 제공 양식을 그대로 유지하여 실제 수사환경에 최적화된 구조를 가진다.

넷째, 생성된 synthetic CDR은 네트워크 분석, 행동분석 모델 학습, 교육훈련 데이터셋, R&D 검증환경 등 다양한 목적으로 활용할 수 있다.

3.4. Synthetic CDR 생성 알고리즘 설계

본 연구에서 제안하는 synthetic CDR 생성 모델은 실제 이동통신 기록의 구조적 특징과 수사 실무에서 관찰되는 통신행동 패턴을 반영하여 설계되었다. 생성 알고리즘은 여러 개의 독립적 기능 단위(module)로 구성되며, 각 모듈은 특정 변수를 생성하거나 규칙을 적용하는 역할을 수행한다. 이러한 모듈 기반 구조는 확장성이 높으며 다양한 범죄 시나리오를 재현할 수 있는 유연성을 제공한다. Synthetic CDR 생성 과정에 사용되는 주요 파라미터는 <Table 3>과 같다.

<Table 3> Definition of parameters for the synthetic CDR generation model

구분	변수명	설명	예시
시간대 설정	time_ratio_morning	오전(06-12시) 통화 비율	0.20
	time_ratio_afternoon	오후(12-18시) 통화 비율	0.35
	time_ratio_evening	저녁(18-24시) 통화 비율	0.30
	time_ratio_night	심야(00-06시) 통화 비율	0.15
통화 방향	direction_ratio_out	발신 비율	0.55
	direction_ratio_in	착신 비율	0.45
주요 상대방	main_counter_count	주요 통신 상대방 수	3
	shared_node_ratio	다중 참여자 연결도	0.25
통화 특성	call_type_voice_rate	음성통화 비율	0.90
	call_type_sms_rate	문자메시지 비율	0.10
	call_count	총 통신 생성 건수	1,000
시간 변수	duration_min/max	통화시간 범위(초)	10-180
위치 생성	home_address	대상자의 주거지 주소	대구 중구 동인동
	active_address_list	활동지 후보 리스트	중구, 북구, 수성구
	time_based_location_rule	시간대별 위치 규칙	오전=주거지, 오후=직장
인물 정보	person_list	대상자 및 상대방 명단	내부 생성

이들 파라미터는 통신 데이터의 시간적 분포, 발신·착신 비율, 주요 상대방 구조, 통신유형 비율, 통화 지속시간, 위치 기반 이동 특성 등 실제 CDR의 핵심 속성을 제어하기 위한 입력값으로 활용된다.

각 파라미터는 이후 단계별 모듈에서 통신 이벤트 생성 규칙, 관계망 구조 형성, 위치·시간 기반 행동 패턴 모사 등에 직접 반영되며, 전체 synthetic CDR의 행위적·구조적 특성을 결정하는 핵심 요소로 기능한다.

특히 이러한 파라미터 체계는 사건유형에 따라 값만 조정해도 서로 다른 통신 패턴을 재현할 수 있다는 점에서 모델의 확장성과 시나리오 적합성을 높여준다.

이와 같은 파라미터를 기반으로 본 연구의 synthetic CDR 생성 모델은 다섯 개의 기능 모듈로 구성되며, 각 모듈은 입력된 파라미터를 조합하여 최종 통신기록을 생성하는 절차적 역할을 수행한다.

3.4.1. 번호·인물 생성 모듈(Phone & Person Module)

본 모듈은 synthetic CDR 생성 과정에서 발신·착신 정보를 구성하는 핵심 요소인 전화번호와 전화 사용자 정보를 생성하는 기능을 담당한다. 먼저, 이동통신(Mobile) 번호는 대한민국 모바일 번호 체계('010-XXXX-XXXX')를 준용하여 010 다음 네 자리·뒤 네 자리를 난수(random) 조합으로 구성하였다. 이를 통해 실제 이동통신사 번호부여 구조와 동일한 형식의 번호를 생성할 수 있도록 하였다.

전화번호에 대응되는 발·착신자명(person identifier)은 모바일 번호인지, 유선번호인지에 따라 구분하여 생성하였다. 모바일 번호의 경우, 인구 통계 기반 이름 생성 방식을 적용하였다. 즉, ① 통계청「인구주택총조사」에서 공개한 대한민국 성씨 분포 상위 10개, ② 대법원 전자가족관계등록시스템에서 제공하는 남녀 이름 상위 10,000개(남성 5,000개, 여성 5,000개)를 조합하여 총 100,000개의 한국인 실명 후보군을 구축하였다. 이후 성(姓)-이름 조합을 무작위(random sampling)로 수행하여, 실제 한국인의 명명 패턴과 동일한 수준의 현실성을 확보하였다.

유선전화(지역번호 기반)일 경우에는 인물명이 아닌 업체명으로 생성하도록 설계하였다. 이는 실제 수사 실무에서 유선번호는 대부분 업체·배달점·직장 등에 해당하는 경우가 많다는 통찰을 반영한 것이다. 특히 피의자의 배달·택배 주문 기록 추적 등 수사 영역에서 유선전화 출현 빈도가 높다는 점을 고려하여, 전국적으로 존재하는 프랜차이즈 기업명 150개를 기반으로 업체명 리스트를 구축하였다. 프랜차이즈 기업은 지역에 따라 명칭이 달라지지 않고 전국 공통 명칭을 갖기 때문에, 합성 데이터의 현실성을 유지하면서도 복잡한 지역별 가중치 없이 유선전화 발신자명을 생성할 수 있다.

또한 본 모듈은 main_counter_count 값에 따라 주요 통신 상대방 목록을 구성하고, 통화 상대 선택 시 이 목록에서 반복적(random repeated sampling)으로 선택될 가능성을 높여 특정 상대방과의 빈번한 통신 관계가 자연스럽게 나타나도록 구현하였다. 이는 실제 수사에서 확인되는 "주요 지인·거래처·공범 등 반복적 상대방 출현 패턴"을 단순하고 직관적인 방식으로 재현할 수 있도록 한다. 더불어 shared_node_ratio 값을 적용하여 여러 대상자가 동일한 상대방과 통신하는 중복 통화자(shared contact) 구조를 생성함으로써, 조직형 범죄나 대포폰 특정 구조에서 자주 관찰되는 인적 네트워크 패턴 또한 반영할 수 있도록 하였다.

본 모듈의 출력 결과는 (1) 인물·업체명 리스트, (2) 각 인물·업체에 대응되는 전화번호, (3) 주요 통신 상대방 목록, (4) 공통 통화자 후보군이며, 이는 이후 단계의 통신 이벤트 생성, 관계망

패턴 구성 등에 활용된다.

3.4.2. 시간 이벤트 생성 모듈(Time Event Module)

본 모듈은 synthetic CDR에서 개별 통신 사건의 발신·착신 시각(timestamp)을 생성하는 기능을 담당하며, 실제 수사 현장에서 사건유형별로 관찰되는 시간대 기반 행동패턴을 반영할 수 있도록 설계되었다. 이를 위해 전체 24시간을 ① 새벽(00-06시), ② 오전(06-12시), ③ 오후(12-18시), ④ 저녁(18-24시)으로 구분하여, 각 시간대별 통신 발생 비율을 사용자가 직접 옵션으로 설정할 수 있도록 구성하였다.

이러한 4분할 구조로 설정한 이유는 단순한 시간 분류를 넘어 범죄 유형별 시간 행동패턴 차이를 데이터 생성 과정에 반영하기 위함이다. 예를 들어, ① 온라인 도박·도박장소개설 사건은 주로 새벽 시간대(00-06) 활동량이 높게 나타나는 경향이 있으며, ② 보이스피싱·대출사기·투자사기는 피해자가 전화를 받을 가능성이 높은 오후(12-18시)에 피의자 발신 이벤트가 집중되는 특징을 가진다. 따라서 사건 유형에 따라 시간대 비율을 조정할 수 있어야 실제 수사 패턴을 충실히 모사할 수 있으며, 본 모듈은 이러한 수사적 니즈를 반영하도록 설계되었다.

코드 구현 측면에서 본 모듈은 time_ratio_morning, time_ratio_afternoon, time_ratio_evening, time_ratio_night 값을 기반으로 확률 가중치(probabilistic weighted sampling) 방식으로 시간대를 선택한다. 선택된 시간대가 결정되면 해당 구간 내에서 난수 기반의 시각을 생성하여 실제 발신 시각(start_time)으로 사용한다. 종료 시각(end_time)은 별도로 설정된 통화시간(duration) 범위(duration_min~duration_max) 내 난수를 더하여 산출하며, 이를 통해 시간적 연속성을 자연스럽게 부여한다.

또한, 시간대 선택은 단순한 균등 분포가 아니라 코드에서 구성한 확률 기반 누적분포(cumulative distribution)에 따라 구현되므로, 특정 시간대 비중을 0.5 또는 0.7과 같이 높게 설정할 경우 통신 이벤트가 해당 구간에 집중되는 구조를 가지게 된다. 이 방식은 실제 CDR 분석에서 관찰되는 시간대 편향(time-of-day bias)—예: 가족·지인의 통화가 주로 저녁에 몰리는 현상, 영업·사무직 업무전화가 오후에 집중되는 현상—을 자연스럽게 재현할 수 있도록 한다.

본 모듈의 최종 출력은 (1) 시작일시(start_time), (2) 종료일시(end_time), (3) 사건유형 기반 시간대 분포를 반영한 전체 통신 발생 스케줄(time distribution)이며, 이후 위치 생성 모듈에서 시간 기반 규칙(예: “오전=주거지, 오후=활동지”)을 적용하는 입력값으로 활용된다. 이처럼 시간 이벤트 생성 모듈은 단순 타임스탬프 생성 기능을 넘어, 사건유형별 활동성 차이를 고현실성으로 반영할 수 있는 핵심 요소로 기능한다.

3.4.3. 위치 생성 모듈(Location Module)

위치 생성 모듈은 synthetic CDR에서 가장 중요한 구성요소 중 하나로, 발신 시점에서 사용자가 실제로 위치했을 것으로 추정되는 기지국 기반 발신 위치 주소를 생성하는 기능을 수행한다. 위치 정보는 피의자 은신처 특정, 범죄동선 분석, 공범 간 접촉 여부 검증 등 수사 실무에서 핵심적 역할을 하므로, 본 연구에서는 실제 행정구역 공간구조를 반영한 고현실성 생성 방식을 채택하였다.

본 모듈에서 사용하는 주소 데이터는 통계분류포털에서 제공하는 대한민국 행정구역 공식 주소 18,847개 전체 목록을 기반으로 구축하였다. 연구자는 synthetic CDR 생성 시 대상자의 주거지(home_address)를 1개 입력하게 되며, 활동지는 코드 내부에서 해당 주소 목록을 기반

으로 시간대별 규칙에 따라 자동으로 선택된다. 즉, 오전 시간대에는 주거지 중심으로 위치가 생성되고, 오후·저녁 시간대에는 행정구역 목록 내 다른 지역을 확률적으로 선택하여 활동지(직장, 자주 방문하는 업소, 아지트, 범행지 등으로 가정)로 배정하는 방식이다. 본 연구는 이러한 시간대 기반 위치 변이를 통해 실제 통신 기록에서 관찰되는 반복적·주기적 이동 패턴을 모사하도록 설계하였다.

코드 기반 구현 방식은 다음과 같다.

첫째, 오전(06-12시)에는 대상자의 주거지(home_address)에서 발신되는 것으로 설정하여, 일반적인 일상생활 패턴을 반영한다.

둘째, 행정구역 전체 주소 목록에서 확률적으로 선택된 활동지(active location)를 사용하여 발신 위치를 생성한다. 이 활동지는 모델이 내부적으로 구축한 주소 데이터베이스에서 시간대 규칙에 따라 자동 선정된다. 이러한 방식은 '주거지 → 활동지 → 귀가'로 이어지는 일상적 이동 패턴을 재현할 뿐 아니라, 일상 또는 사회적 활동 과정에서 여러 장소를 이동하는 것과 유사한 공간적 패턴이 생성된다. 이러한 방식은 통신내역에서 흔히 관찰되는 이동성 기반 공간 변화를 모사하는 데 효과적이다.

셋째, 심야(00-06시)에는 사건 유형에 따라 주거지 또는 특정 활동지에서 발생하도록 옵션을 조정할 수 있으며, 도박 사건 또는 불법행위가 심야에 집중되는 사건의 경우 활동지 비중을 높여 현실성을 강화할 수 있다.

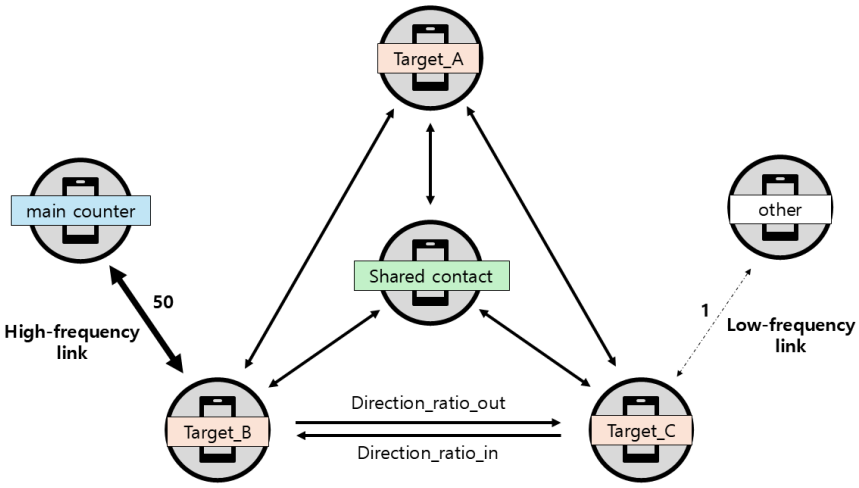
이처럼 시간대 규칙에 기반하여 주거지-활동지 두 공간이 반복적으로 등장하는 구조는 실제 수사에서 GPS·기지국 분석을 통해 파악되는 피의자의 이동 동선 패턴과 매우 유사하게 구성된다. 코드 상에서도 동일한 인물의 통신 이벤트가 home_address와 active_address_list 간 번갈아 나타나며, 결과적으로 "집-직장", "집-아지트", "주거지-활동지"와 같은 동선 특성이 자연스럽게 드러나는 것이 특징이다.

또한 모든 주소는 행정구역 목록 중 해당 지역의 기지국 인근 주소를 기반으로 선택되므로, 지역적 맥락 역시 보존된다. 예컨대 대구 수성구를 활동지로 설정한 경우, synthetic CDR에서 생성되는 발신위치는 모두 수성구 내 실제 존재하는 도로명 주소에서 랜덤 선택되며, 이는 실제 CDR 자료에서 사용되는 위치 정보의 세밀한 주소 단위와 동일한 수준의 공간 정밀도와 일관성을 확보하는 데 중요한 요소이다.

요약하면, 위치 생성 모듈은 단순 난수 기반 주소 선택이 아니라 시간대 기반 규칙, 행정구역 기반 주소, 활동지 후보 리스트를 모두 결합하여 발신위치의 '현실적 반복성'과 '이동성 패턴'을 동시에 구현하는 구조로 설계되었다. 이는 synthetic CDR이 실제 수사에서 활용 가능한 수준의 공간적 신뢰도를 갖추도록 하는 핵심 요소이다.

3.4.4. 관계망 기반 통신 상대 생성 모듈(Network Module)

우선, 각 통신 이벤트마다 통화 방향(direction)을 결정한다. 코드에서는 direction_ratio_out과 direction_ratio_in을 기반으로 대상자가 발신 주체가 되는 OUT 통화와, 대상자로 걸려오는 IN 통화의 비율을 확률적으로 분기한다. 예를 들어, direction_ratio_out = 0.55, direction_ratio_in = 0.45로 설정할 경우, 전체 synthetic CDR 중 약 55%는 "대상자 → 상대방" 구조로, 약 45%는 "상대방 → 대상자" 구조로 생성되도록 설계된다. 이를 통해 피의자 또는 분석 대상자가 능동적으로 전화를 많이 거는 유형인지, 혹은 외부로부터 연락을 많이 받는 유형인지를 설정할 수 있다.



<Figure 2> Conceptual diagram of the relationship generation module

발신자·착신자 선택은 번호·인물 생성 모듈에서 정의한 주요 통신 상대방(main_counter_list)과 일반 상대방(other_list)을 활용해 이루어진다. 코드에서는 주요 상대방 목록을 별도의 리스트로 유지하고, 각 통화 이벤트 생성 시 해당 리스트에서 반복적으로 무작위 추출(random sampling)하는 구조를 사용한다. 그 결과, 주요 상대방들은 자연스럽게 전체 통신 기록에서 더 자주 등장하게 되며, 실제 CDR에서 관찰되는 "가족·지인·거래처·공범 등 특정 상대와의 반복적인 통신 패턴"을 모사할 수 있다. 이는 복잡한 확률 가중(weight) 모델을 쓰지 않고도 리스트 기반 반복 선택 구조만으로 통신 빈도 편차를 재현하는 방식이다.

여러 대상자(예: 피의자 A, 피의자 B, 피의자 C 등)를 동시에 설정하는 경우에는 shared_node_ratio 변수를 통해 공통 상대(shared contact)를 생성한다. 일정 비율의 통신 이벤트에서 서로 다른 대상자들이 동일한 상대방 번호와 통신하도록 구성함으로써, 조직 범죄나 공범 범죄에서 자주 나타나는 "공통 연락망", "허브 역할을 하는 중간 연락책"과 같은 네트워크 구조를 재현할 수 있다. 이는 이후 네트워크 분석(예: degree, betweenness 등)을 위한 기초 데이터로 활용될 수 있으며, 특정 번호가 여러 인물 간 관계를 매개하는 브리지 노드(bridge node) 역할을 수행하는 패턴을 synthetic CDR 수준에서부터 설계 단계에 반영하는 효과를 가진다.

구체적인 처리 순서는 다음과 같이 정리할 수 있다.

첫째, 각 통신 이벤트에 대해 통화 방향(OUT/IN)을 direction_ratio_out / in에 따라 확률적으로 결정한다.

둘째, OUT 통화의 경우, 발신번호는 대상자(또는 설정된 피의자 그룹)에서 선택하고 착신번호는 주요 상대방 목록 및 일반 상대방 목록에서 선택한다.

셋째, IN 통화의 경우, 반대로 발신번호는 주요/일반 상대방 목록에서 선택하고, 착신번호는 대상자 번호로 지정한다.

넷째, shared_node_ratio가 적용된 경우, 일정 비율의 이벤트에서 공통 상대방 번호를 사용하여 다수의 대상자와 동일 번호가 반복적으로 연결되도록 구성한다.

이와 같은 관계망 생성 모듈을 통해 synthetic CDR은 단순히 "발신번호-착신번호 쌍을 무작위로 나열한 데이터"가 아니라, 특정 인물에게 집중된 통신, 공통 상대방을 매개로 한 네트워크, OUT/IN 비율에 따른 행위 특성이 반영된 구조를 갖게 된다. 이는 실제 수사 실무에서 계좌·

통신 분석과 결합하여 핵심 허브 인물 특정, 공범 구조 파악, 역할 분담 추정 등에 활용되는 통신 관계망 분석을 모의 환경에서도 수행할 수 있도록 해준다는 점에서 중요한 의미를 가진다.

3.4.5. 통화 유형 및 부가정보 생성 모듈(Call Type & Supplemental Info Module)

통화유형 및 부가정보 생성 모듈은 synthetic CDR에서 각 통신 이벤트가 어떤 방식으로 이루어진 통신인지(음성통화·SMS)를 결정하고, 추가적으로 통신사 정보(업체명), 통신 구분코드 및 메모 필드 등 부가 변수를 생성하는 기능을 수행한다. 이 모듈은 실제 수사에서 통화유형별 패턴 분석(예: 피의자와 특정 상대가 음성통화를 거의 하지 않고 SMS만 주고받는 경우, 사건 특성상 문자발송 집중 등)이 중요한 의미를 가진다는 점을 반영하여 설계되었다.

우선 통화유형(call type)은 사용자가 사전에 설정한 call_type_voice_rate, call_type_sms_rate 변수를 기반으로 확률적으로 결정된다.

예를 들어, call_type_voice_rate = 0.90, call_type_sms_rate = 0.10으로 설정된 경우 전체 synthetic CDR의 약 90%는 음성통화로, 10%는 SMS로 생성된다. 이는 실무에서 대부분의 통신은 음성 기반이라는 일반적 특성을 반영하는 기본 설정이며, SMS를 대량 발송하는 특정 범죄 유형을 모델링하고자 하는 경우 SMS 비율을 높여 SMS 중심 통신 패턴을 강조하는 것도 가능하다. 본 모듈은 시간 이벤트 모듈과 달리 시간대와 무관하게 독립적으로 작동하는 구조를 가지며, 실제 코드에서도 시간대를 입력값으로 사용하지 않는다.

통신사 정보는 대상자의 경우 사용자 입력값을 그대로 반영하여 고정되며, 상대방 번호의 경우에는 실제 국내 이동통신사(SK텔레콤·KT·LGU+) 중에서 확률적으로 선택되는 방식으로 생성된다.

부가정보 필드(비고)는 본 연구에서 생성되는 synthetic CDR의 확장성을 고려한 요소로, 기본적으로는 비워 두어 실제 수사관이 필요 시 메모·주석·수사 메타데이터를 자유롭게 추가할 수 있도록 설계되었다. 또한, 연구자가 옵션을 추가하여 “특정 관계의 라벨링”, “특정 사건번호”와 같은 특정 태그를 자동 주입할 수도 있어 향후 AI 학습 데이터 구축 단계에서 라벨링(labeling) 목적의 활용도 가능하다.

본 모듈의 주요 처리 절차는 다음과 같다.

첫째, 각 통신 이벤트에 대해 call_type_voice_rate / sms_rate를 기반으로 음성통화 또는 SMS 유형을 확률적으로 선택한다.

둘째, 통신사는 해당 전화번호가 속한 사업자 정보를 참조하거나 사전에 정의된 목록에서 선택한다.

셋째, 통화유형에 따라 종료시각(end_time) 생성 로직이 달라지며, SMS의 경우 duration(통화시간)이 0초 또는 1초로 자동 설정된다.

넷째, 비고(memo) 필드를 생성하고, 필요 시 후처리(post-processing)에서 관계 태그·사건 태그 등을 주입할 수 있도록 구조화한다.

통화유형 및 부가정보 생성 모듈은 synthetic CDR의 전반적 구조를 완성하는 마지막 단계로, 통화유형 분포 등 실제 CDR의 통계적 특성과 일치하는 패턴을 재현하는 데 중요한 역할을 수행한다. 또한, 특정 사건 유형 기반 통신특성을 반영하거나 AI 모델이 학습할 수 있는 라벨 구조를 삽입하는 등 확장성을 고려한 구성으로 설계되어, 수사 R&D·합성데이터 표준화·범죄분석 시뮬레이션 환경 구축에서 유용한 기반 요소로 기능한다.

3.4.6. Synthetic CDR 생성 알고리즘 전체 흐름(Overall Pipeline)

이러한 절차는 모듈 간 상호작용을 통해 구현되며, 모델은 데이터 규모에 따라 수백에서 수십만 건의 CDR 데이터를 생성할 수 있다. 또한 각 단계는 파라미터 형태로 조정 가능하여 수사 목적에 따라 다양한 규모·유형의 synthetic CDR을 생산할 수 있다.

이상의 모듈은 <Table 4>와 같은 순서로 통합되어 최종 synthetic CDR을 생성한다.

<Table 4> Overall pipeline of synthetic CDR generation

단계	모듈명
1	인물 및 전화번호 초기화
2	주요 상대방 및 관계 구조 설정
3	시간대 비율을 반영한 통신 이벤트 생성
4	시간대별 발신 위치 생성
5	발신·착신자 네트워크 기반 매칭
6	통화 유형 및 통신사 정보 부여
7	레코드 단위 데이터 통합
8	Synthetic CDR 데이터 출력

이와 같은 파이프라인은 실제 CDR 구조와 행동 기반 패턴을 동시에 반영하여, 수사 분석·교육훈련·모델 검증 등 다양한 실무 환경에서 활용 가능한 고현실성 synthetic CDR을 생성할 수 있도록 한다.

IV. 실험 및 평가

본 장에서는 제안한 synthetic CDR 생성 모델이 가상의 수사 시나리오에서 기대되는 통신 패턴을 구조적으로 적절히 재현할 수 있는지를 검증한다. 실험은 두 단계로 구성되며, ① 기지국 기반 네트워크 모델의 구조적 타당성, ② 관계망 기반 네트워크 모델의 구조적 타당성을 각각 평가한다.

4.1. 실험 환경 및 데이터 구성

본 연구에서 설계한 실험은 마약 드랍 사건을 가정한 수사 시나리오를 기반으로 하여, synthetic CDR 모델이 위치·행동·관계망 측면의 구조적 패턴을 일관성 있게 생성할 수 있는지를 검증하는데 목적이 있다. 실험은 시나리오1과 시나리오2로 구분되며, 두 시나리오는 단계적으로 연결되는 구조적 검증 과정으로 구성된다.

4.1.1. 시나리오 1: 기지국 기반 네트워크 모델의 구조적 타당성 평가

시나리오 1은 3개의 지역에서 각 지역별 마약 드랍 이후 즉시 상선(Boss)에게 보고 전화를 하는 드랍퍼 3명의 행동 패턴을 가정하고, synthetic CDR이 기지국 기반 네트워크 구조를 설계 의도대로 재현할 수 있는지를 평가하기 위한 단계이다.

본 시나리오1의 개요는 다음과 같다.

첫째, 춘천 지역의 3개 장소에 드랍퍼A가 마약을 드랍한다.

둘째, 원주 지역의 3개 장소에 드랍퍼B가 마약을 드랍한다.

셋째, 양양 지역의 3개 장소에 드랍퍼C가 마약을 드랍한다.

넷째, 각 드랍퍼들은 담당 구역에 마약을 드랍한 직후, 동일한 상선(Boss)에게 즉시 전화로 보고한다.

이로 인해 각 지역의 3개 기지국에서 공통적으로 반복 등장하는 발신 번호는 단 1개가 되며, 해당 번호가 곧 각 지역을 담당한 드랍퍼의 전화번호에 해당하게 된다.

본 시나리오1에서 기지국 기반 synthetic CDR 모델이 완수해야 할 구조는 다음과 같다.

1) 각 지역(춘천·원주·양양)에 대해 3개 기지국 기반의 발신 데이터가 생성될 것

2) 각 지역의 3개 기지국에 모두 등장하는 단일 발신 노드(해당 노드는 지역별 드랍퍼로 가정)가 존재할 것

3) 각 기지국에는 단일 연결 형태의 일반 통신이용자들이 생성되어 기지국 주변부 노드를 형성할 것

이 구조가 정확하게 나타난다면, 제안한 모델은 기지국 기반 네트워크 모델을 설계 의도에 맞게 재현한 것으로 판단한다.

4.1.2. 시나리오 2: 관계망 기반 네트워크 모델의 구조적 타당성 검증

시나리오 2는 시나리오 1에서 생성·특정된 드랍퍼 A, B, C의 발착신 synthetic CDR을 기반으로, 관계망 생성 모듈이 의도한 통신 네트워크 구조를 정확하게 구현하는지를 평가하기 위한 단계이다.

본 시나리오2에서는 다음과 같은 통신 관계를 가정한다.

첫째, 드랍퍼들은 모두 동일한 상선(Boss)와 통신을 수행한다.

둘째, 일부 드랍퍼들 간에는 서로 아는 인물(상호 연결된 지인)이 존재하며, 동시에 각 드랍퍼와만 개별적으로 연결된 인물도 존재한다.

셋째, 드랍퍼의 통신 위치는 시간에 따라 동적으로 변화하며, 주로 심야·오전에는 주거지, 오후·저녁에는 활동지로 이동하는 패턴을 가진다.

본 시나리오2에서 관계망 기반 synthetic CDR 모델이 완수해야 할 구조는 다음과 같다.

1) shared_node_ratio에 의해 드랍퍼들과 모두 연결된 공통 노드(상선 노드)가 존재할 것

2) 일부 드랍퍼들 간 연결된 노드, 그리고 각 드랍퍼와만 개별적으로 연결된 노드가 동시에 존재할 것

3) 네트워크 시각화 결과에서 드랍퍼-상선-주변부 노드로 구분되는 계층적 네트워크 구조가 형성될 것

4) direction_ratio_out, direction_ratio_in 설정값이 통화 방향성의 비율이 결과에 반영될 것

5) 통화 강도의 현실적 편차를 반영하여 고빈도(high-frequency) 통신 연결과 저빈도(low-frequency) 통신 연결이 동시에 존재하는 연결 구조가 형성될 것

6) 각 드랍퍼들의 통신 위치가 시간대에 따라 주거지-활동지로 동적으로 변화할 것

이 구조가 정확하게 나타난다면, 제안한 모델은 관계망 기반 네트워크 모델을 설계 의도에 맞게 재현한 것으로 판단한다.

4.2. 실험 절차

본 절에서는 제안한 synthetic CDR 생성 모델에 대해 시나리오 1과 시나리오 2의 구조적 타당성을 검증하기 위한 실험 절차를 단계적으로 기술한다. 실험은 기지국 기반 구조 검증 → 관계망 기반 구조 검증의 흐름에 따라 순차적으로 수행되었다.

첫째, 시나리오 1에 대한 데이터 구성 설정(Data Configuration)을 수행한다.

춘천·원주·양양 지역에 각각 3개의 기지국이 존재하도록 위치 모듈을 구성하고, 각 기지국별 발신번호 수, 데이터 규모, 생성 시작·종료 시각 등을 데이터 구성 설정값으로 지정한다. 이때 동일 지역의 3개 기지국에 공통적으로 포함되는 단일 발신번호가 존재하도록 공통 연결 노드 수를 설정한다. 해당 단계는 행태 제어를 위한 파라미터 설정이 아닌, 기지국 기반 데이터 구조 자체를 정의하는 단계에 해당한다.

둘째, 시나리오 1에 대한 synthetic CDR을 생성한다.

앞 단계에서 설정한 기지국 위치, 데이터 생성 기간, 기지국당 데이터 규모 등의 데이터 구성 설정값을 기반으로 synthetic CDR을 생성하고, 각 지역별 3개 기지국 데이터가 서로 독립적으로 생성되되, 공통 연결 노드를 포함하는 구조로 데이터가 구성되도록 한다.

셋째, 시나리오 1의 구조적 타당성을 검증한다.

생성된 synthetic CDR에 대해 기지국별 발신 번호 집합을 구성한 후, 동일 지역의 3개 기지국에 모두 등장하는 단일 발신 번호가 존재하는지를 교차 분석을 통해 확인한다. 이를 통해 각 지역에서 드랍퍼 A, B, C가 각각 지역별 기지국에서 하나의 공통 발신 노드로 구조적으로 식별되는지를 검증한다.

넷째, 시나리오 2를 위한 데이터 구성 설정(Data Configuration) 및 통신 행태 제어 파라미터(Control Parameters)를 설정한다.

시나리오 1에서 도출된 드랍퍼 A, B, C를 기준 대상자로 설정하고, 각 대상자에 대해 대상자 이름, 기준 전화번호, 설정 지역, 통화 상대방 수, 데이터 규모, 생성 시작·종료 일시, 대상 노드들과 모두 연결되는 공통 노드의 수 등의 데이터 구성 설정값을 입력한다.

이를 통해 관계망 기반 통신 네트워크 생성을 위한 기본 데이터 구조(대상자 단위 데이터 틀)를 먼저 정의한다. 그리고 음성 통화 비율, 통화 시간대 비율, 통신 방향 비율(direction_ratio_out, direction_ratio_in), 중복 통화 비율(shared_node_ratio), 대상자 간 통화 비율 등의 행태 제어 파라미터를 기본값 또는 사용자 입력 기반의 설정값으로 지정한다. 본 단계는 시나리오 2의 통신 패턴, 네트워크 연결 구조, 시간대별 행태 특성을 결정하는 핵심 제어 단계에 해당한다.

다섯째, 시나리오 2에 대한 synthetic CDR을 생성한다.

앞 단계에서 설정한 대상자 기반 데이터 구성 설정값과 관계망·행태 제어 파라미터를 동시에 적용하여, 드랍퍼-상선-주변부 노드가 포함된 발착신 synthetic CDR을 생성한다. 이 과정에서 고빈도·저빈도 통신이 혼재된 네트워크 구조가 형성되도록 데이터가 구성되며, 이와 함께 시간대에 따른 통신 위치 변화가 반영되도록 위치 모듈을 연동하여 데이터를 생성한다.

여섯째, 시나리오 2의 구조적 타당성을 네트워크 분석 및 타임라인을 통해 검증한다.

생성된 synthetic CDR을 기반으로 통신 네트워크를 구성하고, Degree 중심성 지표를 통해 드랍퍼, 상선, 주변부 노드 간의 구조적 차이가 의도한 형태로 나타나는지를 확인한다. 또한 네트워크 시각화를 통해 드랍퍼-상선-주변부로 구분되는 계층적 구조가 형성되는지를 평가하고 타임라인을 통해 시간대에 따라 기준 노드의 발신 위치가 동적으로 변화하는 지 평가한다.

일곱째, 두 시나리오의 검증 결과를 종합하여 제안한 모델의 구조적 재현성을 최종 평가한다.

시나리오 1의 기지국 기반 네트워크 구조와 시나리오 2의 관계망 기반 네트워크 구조가 모두 각각의 설계 의도와 일치하는 경우, 제안한 synthetic CDR 생성 모델이 위치·행동·관계망 측면에서 가상의 수사 시나리오를 구조적으로 타당하게 재현한 것으로 판단한다.

4.3. 실험결과

본 절에서는 4.1 및 4.2에서 설계한 두 개의 시나리오에 따라 생성된 synthetic CDR이 의도한 구조적 패턴을 실제로 재현하고 있는지를 실험 결과를 통해 검증한다. 실험 결과는 시나리오 1(기지국 기반)과 시나리오 2(관계망 기반)로 구분하여 제시한다.

4.3.1. 시나리오 1 데이터 구성 사전 설정

시나리오 1의 기지국 기반 네트워크 구조를 생성하기 위해 사용된 synthetic CDR 생성용 데이터 구성 설정값(Data Configuration)을 정리한다. 해당 설정값들은 지역별 다중 기지국 환경에서 단일 드랍퍼 발신 노드가 형성되는 구조를 재현하기 위한 데이터 생성 조건에 해당한다.

<Table 5> Data configuration for synthetic CDR generation (scenario 1)

데이터	기지국 주소	발신번호 수	데이터규모	생성시작일시	생성종료일시	공통연결 노드
춘천 기지국	강원도 춘천시 천전리 99-13 감자밭	5,000개	5,000행	2025-04-20 18:00:00	2025-04-20 18:10:00	1
	강원도 춘천시 석사동 95-3 국립춘천박물관	5,000개	5,000행	2025-04-25 16:00:00	2025-04-25 16:10:00	
	강원도 춘천시 교동 153 춘천성심병원	5,000개	5,000행	2025-05-01 14:00:00	2025-05-01 14:10:00	
원주 기지국	강원도 원주시 일산동 162-33 원주세브란스기독병원	5,000개	5,000행	2025-04-05 21:00:00	2025-04-05 21:10:00	1
	강원도 원주시 흥업리7 강릉원주대학교 원주캠퍼스	5,000개	5,000행	2025-04-10 19:00:00	2025-04-10 19:10:00	
	강원도 원주시 반곡동 1809-1 호텔인터불고	5,000개	5,000행	2025-04-15 12:00:00	2025-04-15 12:10:00	
양양 기지국	강원도 양양군 군행리 25-3 현산공원	5,000개	5,000행	2025-04-01 22:00:00	2025-04-01 22:10:00	1
	강원도 양양군 오산리 23-4 쏠비치 양양	5,000개	5,000행	2025-04-10 17:30:00	2025-04-10 17:40:00	
	강원도 양양군 정암리 450-4 강현면사무소	5,000개	5,000행	2025-04-17 15:20:00	2025-04-17 15:30:00	

시나리오 1에서는 각 지역(춘천·원주·양양)에 대해 각각 3개의 기지국을 독립적으로 생성하고, 각 기지국마다 5,000개의 발신 번호와 5,000행 규모의 synthetic CDR이 생성되도록 데이터 구성 설정값을 지정하였다.

이는 각 기지국 데이터가 서로 독립적인 개별 통신 기록 집합으로 구성되도록 하기 위한 데이

터 생성 조건이다.

각 기지국에는 공통 연결 노드 수를 1로 설정하여, 동일 지역의 3개 기지국에 모두 등장하는 단일 발신 번호가 반드시 1개만 존재하도록 구조적 제약 조건(structural constraint)을 부여하였다. 해당 공통 연결 노드는 곧 각 지역을 담당하는 드랍퍼의 전화번호에 대응되는 노드에 해당한다.

생성 시간 범위는 실제 통신 환경처럼 기지국에 짧은 시간 내 대량의 통화가 집중되는 고밀도 발신 패턴 구조를 반영하기 위해, 각 기지국마다 약 10분 내외의 짧은 시간 구간으로 설정하였다.

이와 같은 데이터 구성 설정을 통해 시나리오 1의 synthetic CDR은 "지역별 다중 기지국-단일 드랍퍼 발신 노드" 구조가 형성되도록 사전에 제약된 데이터 생성 환경에서 구축되었다.

4.3.2. 시나리오 1 데이터 생성 결과

시나리오 1에 따라 생성된 synthetic CDR은 춘천·원주·양양 각 지역에 대해 기지국 3개씩, 총 9개 기지국을 기준으로 생성되었으며, 각 기지국마다 5,000행의 통신 데이터가 생성되어 전체 synthetic CDR의 규모는 총 45,000행으로 구성되었다.

춘천 기지국에서 생성된 synthetic CDR의 일부 샘플은 Fig. 3에 제시하였다. 해당 그림은 실제 통신사 CDR 형식과 동일한 필드 구조를 유지한 상태에서, 발신번호, 착신번호, 통화 시작·종료 시각, 통신사, 발신 위치 정보 등이 synthetic CDR로 정상적으로 생성되었음을 보여준다.

	A	B	C	D	E	F	G	H	I	J
1	발신인	발신번호	착신인	착신번호	시작일시	종료일시	업체명	발신위치	구분	비고
2	최휘찬	01040201548	엄다나	01053009178	2025-04-20 18:00:00	2025-04-20 18:03:44	SKT	강원도 춘천시 천전리 99-13 감자밭	국내음성통화	
3	박여름	01099915937	이정완	01003640214	2025-04-20 18:00:00	2025-04-20 18:04:11	SKT	강원도 춘천시 천전리 99-13 감자밭	국내음성통화	
4	최하름	01077440820	박라임	01056847355	2025-04-20 18:00:00	2025-04-20 18:05:35	SKT	강원도 춘천시 천전리 99-13 감자밭	국내음성통화	
5	박다준	01005756906	김준이	01035874600	2025-04-20 18:00:00	2025-04-20 18:02:47	SKT	강원도 춘천시 천전리 99-13 감자밭	국내음성통화	
6	임영현	01064266170	정아임	01026335223	2025-04-20 18:00:00	2025-04-20 18:04:37	SKT	강원도 춘천시 천전리 99-13 감자밭	국내음성통화	
7	조대희	01001953816	박수	01005357226	2025-04-20 18:00:00	2025-04-20 18:09:00	SKT	강원도 춘천시 천전리 99-13 감자밭	국내음성통화	
8	장영경	01053308578	장하율	01036436551	2025-04-20 18:00:00	2025-04-20 18:06:49	SKT	강원도 춘천시 천전리 99-13 감자밭	국내음성통화	
9	윤영민	01009294723	최선빈	01007547278	2025-04-20 18:00:00	2025-04-20 18:00:30	SKT	강원도 춘천시 천전리 99-13 감자밭	국내음성통화	
10	윤윤형	01008563720	김하랑	01080347443	2025-04-20 18:00:00	2025-04-20 18:07:12	SKT	강원도 춘천시 천전리 99-13 감자밭	국내음성통화	
11	윤사라	01080054107	이수하	01039591843	2025-04-20 18:00:00	2025-04-20 18:07:20	SKT	강원도 춘천시 천전리 99-13 감자밭	국내음성통화	
12	이아론	01054593805	강시찬	01010443502	2025-04-20 18:00:00	2025-04-20 18:01:26	SKT	강원도 춘천시 천전리 99-13 감자밭	국내음성통화	
13	강상아	01095921286	최우진	01099267149	2025-04-20 18:00:00	2025-04-20 18:06:04	SKT	강원도 춘천시 천전리 99-13 감자밭	국내음성통화	
14	김민규	01007990550	정한율	01010215331	2025-04-20 18:00:01	2025-04-20 18:00:19	SKT	강원도 춘천시 천전리 99-13 감자밭	국내음성통화	
15	박재진	01028851099	박로한	01029536482	2025-04-20 18:00:01	2025-04-20 18:01:32	SKT	강원도 춘천시 천전리 99-13 감자밭	국내음성통화	
16	조진헌	01072736372	코리아통신	0339478554	2025-04-20 18:00:01	2025-04-20 18:03:16	SKT	강원도 춘천시 천전리 99-13 감자밭	국내음성통화	
17	조시예	01096591126	박상연	01003248684	2025-04-20 18:00:01	2025-04-20 18:09:12	SKT	강원도 춘천시 천전리 99-13 감자밭	국내음성통화	
18	김선주	01028574146	장근혜	01032234826	2025-04-20 18:00:01	2025-04-20 18:09:08	SKT	강원도 춘천시 천전리 99-13 감자밭	국내음성통화	
19	강담희	01095944313	조효성	01075442710	2025-04-20 18:00:01	2025-04-20 18:06:59	SKT	강원도 춘천시 천전리 99-13 감자밭	국내음성통화	
20	이해든	01010551988	정희은	01091882997	2025-04-20 18:00:01	2025-04-20 18:09:57	SKT	강원도 춘천시 천전리 99-13 감자밭	국내음성통화	

<Figure 3> Sample of generated synthetic CDR data from the Chuncheon base

시나리오 1의 지역별 synthetic CDR 데이터 생성 결과 요약은 <Table 6>과 같다.

<Table 6> Summary of generated synthetic CDR for scenario 1 (base-station-based network model)

데이터	기지국 수	발신번호 수	데이터 규모	생성시작일시	생성종료일시
춘천 기지국	3개	14,998개	15,000행	2025-04-20 18:00:00	2025-05-01 14:10:00
원주 기지국	3개	14,998개	15,000행	2025-04-05 21:00:00	2025-04-15 12:10:00
양양 기지국	3개	14,998개	15,000행	2025-04-01 22:00:00	2025-04-17 15:30:00
합계	9개	44,994개	45,000행	2025-04-20 18:00:00	2025-04-17 15:30:00

각 지역별 기지국 데이터에서 고유 발신번호 수는 14,998개가 생성되었다. 기지국별 발신번호 수의 경우 파라미터 설정값은 5,000으로 설정하였으나, 이 중 공통 연결 노드 1개가 동일 지역의 3개 기지국에 중복 포함되도록 설정됨에 따라, 단순 합산값인 15,000개에서 중복분 2개가 제외된 결과이다.

각 지역별 기지국 데이터의 전체 행 수는 기지국 당 생성 행 수를 5,000행으로 설정한 파라미터 조건에 의해 결과적으로 15,000행으로 산출되었다. 이는 지역 간 데이터 규모를 인위적으로 동일하게 맞춘 것이 아니라, 기지국 단위 생성 규모를 파라미터로 제어한 결과로서 자연스럽게 도출된 수치이다. 따라서 향후 실험에서는 도심과 농촌, 유동 인구 규모 차이 등을 반영하여 지역별 기지국 생성 행 수를 상이하게 설정하는 방식으로 데이터 규모를 유연하게 조정할 수 있다.

또한 각 기지국의 생성 시간 범위는 약 10분 내외의 짧은 시간 구간으로 설정되어 있으며, 이는 실제 수사 환경처럼 기지국에 단시간 다량의 통신 이력이 집중되는 상황을 synthetic CDR 상에서 그대로 반영하기 위한 설정이다.

4.3.3. 시나리오 1 구조적 타당성 검증 결과

본 절에서는 4.3.2에서 생성된 기지국 기반 synthetic CDR에 대해, 시나리오 1에서 의도한 '다중 기지국-단일 공통 발신 노드' 구조가 실제로 구현되었는지를 구조적으로 검증한다.

검증은 기지국별 발신번호 집합의 교차 분석(set intersection)과 기지국 네트워크 시각화 결과를 통해 수행하였다.

첫째, 각 지역(춘천·원주·양양)에 대해 3개 기지국의 발신번호 집합을 각각 구성한 후, 해당 집합들 간의 교집합을 분석한 결과, 각 지역별로 정확히 1개의 공통 발신번호만이 도출됨을 확인하였다.

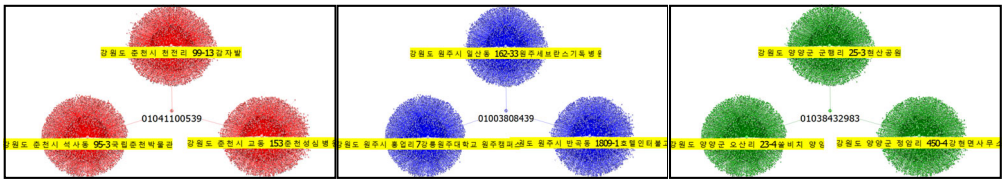
이는 시나리오 1에서 사전에 설정한 공통 연결 노드 수 1개 조건이 synthetic CDR 상에서 정확히 반영되었음을 의미한다.

둘째, 지역 간 교차 분석을 수행한 결과, 춘천 지역에서 도출된 공통 발신번호는 원주 및 양양 지역의 기지국 데이터에는 등장하지 않았으며, 원주 및 양양 지역 역시 동일한 특성을 유지하였다.

즉, 각 지역의 공통 발신 노드는 서로 독립적으로 존재하는 구조임이 확인되었다.

이는 시나리오 1에서 가정한 '지역별 단일 드랍퍼' 구조가 데이터 차원에서 명확히 분리되어 구현되었음을 의미한다.

셋째, 기지국 기반 통신 네트워크를 시각화한 결과, 각 지역의 3개 기지국은 중앙의 단일 공통 발신 노드를 중심으로 방사형으로 연결된 구조를 형성하였으며, 해당 공통 노드 외의 나머지 발신번호들은 각 기지국에만 단일 연결 형태로 분포하는 주변부 노드로 나타났다. 이를 통해 '단일 핵심 노드-다수 주변부 노드' 구조가 공간적으로도 명확히 재현되었음을 확인하였다.



<Figure 4> Base-station-level communication network for scenario 1

넷째, 공통 발신 노드와 주변부 노드 간의 연결 형태를 비교한 결과, 공통 발신 노드는 동일 지역의 3개 기지국 모두와 연결된 유일한 노드로 나타난 반면, 주변부 노드들은 특정 기지국에 연결된 단일 연결 구조(single-link structure) 를 유지하였다.

이는 시나리오 1이 교집합 탐색 구조 검증 실험으로 설계되었음을 데이터 차원에서 그대로 반영한 결과이다.

이상의 검증 결과를 종합하면, 시나리오 1의 synthetic CDR은 다중 기지국 발신 데이터의 교차 분석을 통해 단일 공통 발신 노드를 정확히 도출할 수 있는 구조를 설계 의도에 맞게 구조적으로 재현하고 있음이 확인되었다.

즉, 제안한 synthetic CDR 생성 모델은 기지국 기반 네트워크 환경에서 중복 발신 탐지를 통한 특정 사용자(드랍퍼) 식별 구조를 정량·정성적으로 모두 만족하는 형태로 구현되었음이 검증되었다.

4.3.4. 시나리오 2 데이터 구성 사전 설정 및 통신 행태 제어 파라미터 사전 설정

본 절에서는 시나리오 2의 관계망 기반 synthetic CDR 생성을 위해 사용된 데이터 구성 사전 설정값(Data Configuration)과 통신 행태 제어 파라미터(Control Parameters)를 정리한다. 시나리오 2는 대상자 단위의 통신 네트워크 구조와 시간대·방향성·중복 통화 특성이 동시에 반영되는 시나리오이므로, 먼저 데이터 구조를 정의하는 설정값을 구성한 후, 그 위에 통신 행태를 제어하는 파라미터를 적용하는 이중 구조로 설계된다.

첫째, 시나리오 1에서 도출된 드랍퍼 A, B, C를 기준 대상자로 설정하고, 각 대상자에 대해 대상자 이름, 기준 전화번호, 설정 지역, 통화 상대방 수, 데이터 규모, 생성 시작·종료 일시, 공통 노드 수(대상 노드들과 모두 연결되는 노드의 수) 등의 데이터 구성 설정값을 입력하였다. 이를 통해 시나리오 2의 관계망 기반 synthetic CDR은 대상자 단위의 통신 데이터 생성 틀을 먼저 정의한 상태에서 생성되도록 구조가 설계되었다.

시나리오 2에서 사용된 데이터 구성 사전 설정값은 <Table 7>과 같다.

<Table 7> Data configuration for synthetic CDR generation (scenario 2)

대상자	대상 전화번호	설정 지역	통화상대 수	데이터 규모	생성시작일시	생성종료일시
조우주	010-4110-0539	강원도 춘천시	50	500	2025-04-01 00:00:00	2025-05-31 23:59:59
임리암	010-0380-8439	강원도 원주시	50	500	2025-04-01 00:00:00	2025-05-31 23:59:59
강윤철	010-3843-2983	강원도 양양군	50	500	2025-04-01 00:00:00	2025-05-31 23:59:59
대상자 3명과 연결된 노드의 수				대상자 2명과 연결된 노드의 수		
1				15		

둘째, 데이터 구성 설정과는 별도로, 시나리오 2에서는 통신 패턴과 관계망 구조의 특성을 제어하기 위한 행태 제어 파라미터를 추가로 설정하였다. 음성 통화 비율, 통화 시간대 비율, 통신 방향 비율(direction_ratio_out, direction_ratio_in), 중복 통화 비율(shared_node_ratio), 대상자 간 통화 비율 등의 값은 기본값 또는 사용자 입력값을 기반으로 설정되었다. 이들 파라

미터는 드랍퍼-상선-주변부로 구분되는 네트워크 구조, 고빈도·저빈도 통신 연결의 혼재, 그리고 시간대별 행태 특성이 synthetic CDR에 반영되도록 제어한다.

시나리오 2에서 사용된 통신 행태 제어 파라미터 사전 설정값은 <Table 8>과 같다.

<Table 8> Control parameters for communication behavior and network structure (scenario 2)

구분	항목	설정값(%)
음성통화 비율	음성	80
	문자	20
통화시간대 비율	오전(06-12)	15
	오후(12-18)	40
	저녁(18-24)	35
	새벽(00-06)	10

구분	항목	설정값(%)
통신방향 비율	대상→상대	40
	상대→대상	40
	대상→신규	10
	신규→대상	10
중복통화 비율		30
대상자 간 통화 비율		10

시나리오 2의 synthetic CDR은 먼저 대상자 단위 데이터 구성 설정을 통해 기본 통신 데이터 생성 틀을 정의한 후, 그 위에 통신 행태 제어 파라미터를 적용하여 관계망 구조와 통신 패턴이 동시에 반영되는 방식으로 생성되었다.

즉, 시나리오 2의 데이터 생성과정은 "데이터 구조 정의 → 통신 행태 제어 → 관계망 기반 synthetic CDR 생성"의 다단계 구조를 가진다.

4.3.5. 시나리오 2 데이터 생성 결과

본 절에서는 시나리오 2에 따라 관계망 기반 통신 구조를 반영하여 생성된 synthetic CDR 데이터의 정량적 생성 결과와 그 구조적 특성을 분석한다. 시나리오 2에서는 시나리오 1을 통해 특정된 드랍퍼(조우주, 임리암, 강윤철)를 기준 노드로 설정하고, 관계망 생성 모듈 및 통신 행태 제어 파라미터를 적용하여 <Table 9>와 같이 발착신 통신 데이터를 생성하였다.

<Table 9> Summary of generated synthetic CDR for scenario 2 (relationship-based network model)

구분	항목
가상 인물(노드) 수	499명
전체 통신 건수	1,522건
문자 발신	284건
음성통화	1,238건
데이터 생성기간	2025-04-01 03:50:09 ~ 2025-05-31 23:41:37
발신 위치 수	641개

4.3.5.1. 전체 통신 데이터 규모

시나리오 2에서 생성된 전체 synthetic CDR은 총 1,522건의 발착신 통화 내역으로 구성되었다. 대상자 3인에 대해 각 500건씩의 통화를 생성하도록 기본 데이터 규모가 설정되어, 기본 대상자 단위 발착신 통화는 총 1,500건이 생성되었다.

본 모델의 관계망 기반 생성 구조에서는 기본 대상자 통화 외에도 공통 노드(shared node)

기반 결합 통화 및 대상자 간 상호 통화가 구조적으로 자동 포함되도록 설계되어 있다. 이에 따라 shared_node_ratio 설정에 의해 드랍퍼들과 공통으로 연결되는 상선 노드와의 통화 이벤트가 추가 생성되었고, cross_call_ratio 설정에 의해 대상자 간 상호 통화 구간이 별도의 통화 이벤트로 자동 반영되었다.

이와 같은 관계망 확장 생성 구조에 따라 기본 대상자 통화 1,500건 외에 추가적인 통화 이벤트가 포함되어, 최종적으로 1,522건의 통화 데이터가 생성되었다. 이는 시나리오 2에서 의도한 허브 노드-주변 노드-대상자 간 상호 연결 구조를 네트워크 수준에서 보다 현실적으로 반영하기 위한 설계 기반 자동 확장 생성 결과에 해당한다.

4.3.5.2. 가상 인물(노드) 수

시나리오 2에서 대상자는 총 3명으로 설정되었으며, 각 대상자에 대해 고정 통화 상대방 수(peer_count)는 50명으로 설정되었다. 이 설정값은 각 대상자에게 반복적으로 등장하는 주요 통화 상대 집단, 즉 가족·지인·거래처 등 고정적 통화 관계 집단을 형성하기 위한 내부 파라미터로 사용된다.

한편, 본 모델의 통신 데이터 생성 구조는 고정 상대방 집단 외에도 통신 방향 파라미터(target_to_third, third_to_target)에 따라 매 통화 이벤트마다 신규 제3자 번호가 확률적으로 추가 생성되는 구조로 설계되어 있다. 이로 인해 전체 통화 참여 인물 수는 고정 상대방 수의 단순 합을 초과하며, 네트워크 외곽부(periphery)가 지속적으로 확장되는 구조를 갖는다.

그 결과, 본 시나리오에서는 총 499명의 가상 인물이 최종 발착신 통신에 참여한 것으로 확인되었으며, 이는 관계망 기반 synthetic CDR 생성 모델이 고정 관계망과 유동 관계망을 동시에 포함하는 확장형 네트워크 구조를 갖도록 설계된 결과에 해당한다.

4.3.5.3. 문자·음성 통화 비율

본 실험에서 통신 유형 비율은 음성통화(80%), 문자(20%)로 설정되었다. 이에 따라 생성된 실제 결과는 음성통화 1,238건(81.3%), 문자284건(18.7%)으로 나타나, 사전 설정한 8:2 통신 비율이 실제 데이터 생성 결과에 매우 근접하게 반영되었음을 확인할 수 있다. 이는 통신 행태 제어 파라미터가 구조적으로 정상 작동하고 있음을 정량적으로 입증하는 결과이다.

4.3.5.4. 데이터 생성 기간

시나리오 2에서 사전 설정된 데이터 생성기간은 2025년 4월 1일 00:00:00부터 2025년 5월 31일 23:59:59까지이며, 실제 생성된 데이터는 2025년 4월 1일 03:50:09부터 2025년 5월 31일 23:41:37까지로 분포되어 있다. 생성된 모든 통화 데이터는 사전 설정된 기간 범위 내에서 자연스럽게 분포되었으며, 시간 축 상의 이상치 없이 안정적인 데이터 생성이 이루어졌음을 확인하였다.

4.3.5.5. 발신 위치 데이터 생성결과

본 시나리오에서 생성된 고유 발신 위치 수는 총 641개로 확인되었다. 이는 대상자의 주거지(Home) 고정 위치, 활동지(Activity) 생성, 시간대(새벽·오전 vs 오후·저녁)에 따른 동적 위치 전환 로직 적용, 상대방 발신 시에는 전국 단위 랜덤 지역 기반 위치 생성 등 위치 생성 로직이 동시에 적용된 결과이다.

이를 통해 동일 번호라도 시간대에 따라 발신 위치가 자연스럽게 변화하며, 현실 범죄 수사에서 관찰되는 이동성 기반 통신 패턴이 synthetic CDR 상에서 효과적으로 재현되었음을 확인할 수 있다.

4.3.5.6. 최종 결론

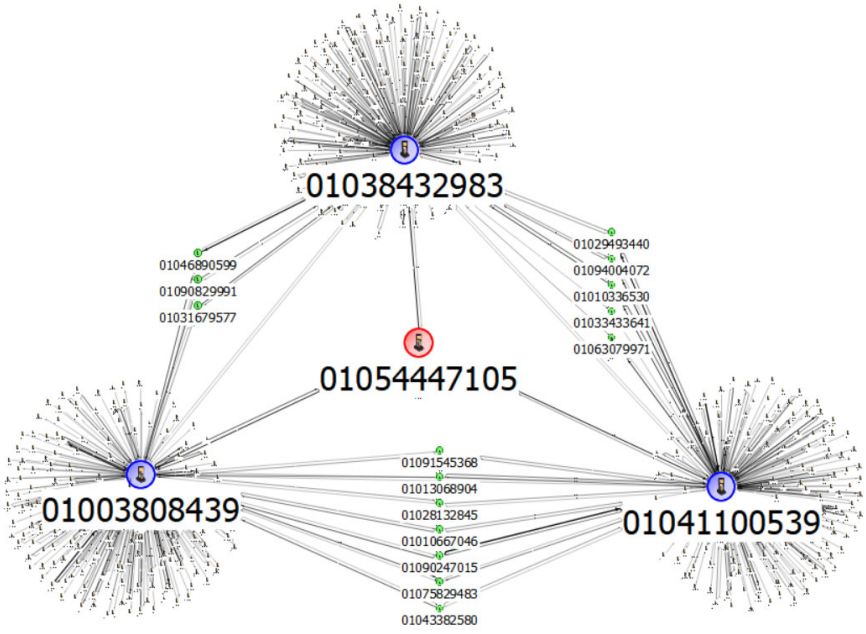
시나리오 2에서는 총 1,522건의 발착신 통신 데이터가 생성되었으며, 음성 통화 1,238건(81.3%), 문자 발신 284건(18.7%)으로 사전 설정한 8:2 비율이 구조적으로 안정적으로 반영되었다. 전체 통화 참여 인물 수는 499명으로 나타났으며, 이는 고정 통화 상대방 외에 이벤트 기반 신규 제3자가 지속적으로 생성되는 이중 인물 생성 구조에 따른 결과이다. 또한 약 2개월의 생성 기간 동안 총 641개의 발신 위치가 동적으로 생성되어 시간대별 이동성과 공간적 분산 특성이 함께 반영된 고현실성 synthetic CDR 데이터셋이 구축되었음을 확인하였다.

4.3.6. 시나리오 2 구조적 타당성 검증 결과

본 절에서는 시나리오 2에서 생성된 관계망 기반 synthetic CDR을 대상으로, 관계망 생성 모듈이 설계 의도에 부합하는 통신 네트워크 구조를 실제로 형성하였는지를 네트워크 분석 및 시각적 구조 검증을 통해 평가한다. 검증은 연결 중심성 분포, 허브 노드 형성 여부, 주변부 노드 확산 구조, 대상자 간 상호 연결성, 통화 방향성과 공간 이동성 반영 여부를 중심으로 수행되었다.

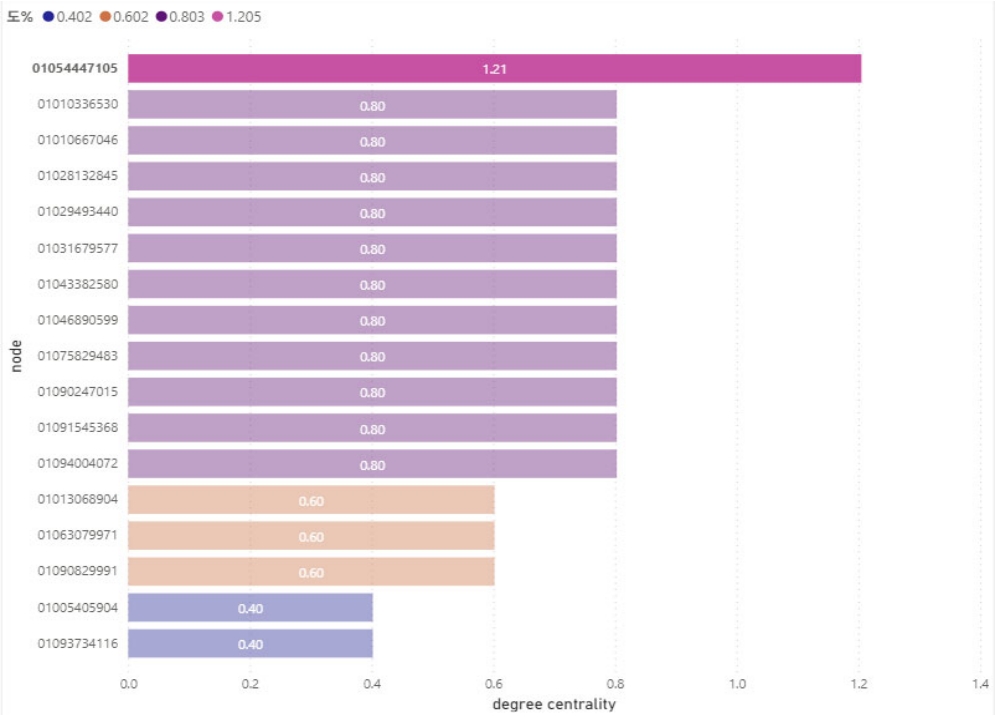
4.3.6.1. 상선(공통 노드) 기반 허브 연결 구조 검증(Degree 중심성 기반)

시나리오 2에서는 shared_node_ratio 파라미터를 통해 드랍퍼 3명과 모두 연결되는 공통 노드(상선 노드)가 형성되도록 설계되었다. 생성된 통신 데이터를 기반으로 네트워크를 시각화한 결과, 단일 노드가 세 대상자 모두와 직접 연결되는 허브(hub) 구조가 명확히 형성됨을 확인할 수 있었다<Figure 5>.



<Figure 5> Network visualization of the shared (Boss) node–centered communication structure in scenario 2

또한 <Figure 6>의 Degree 중심성 분석 결과, 상선 노드의 Degree 중심성 값이 1.205로 가장 높게 나타났으며, 일반 주변부 노드들은 0.201~0.803 수준으로 분포하였다. 실제 사이버 범죄조직 통신 네트워크를 분석한 문병훈·곽기영(2024)의 연구에서도 높은 Degree 중심성을 갖는 노드가 조직 내 최상위 통제 역할을 수행하는 핵심 노드로 식별된 바 있다[7]. 본 시나리오에서도 상선 노드가 모든 드랍퍼와 연결된 핵심 허브로 재현됨으로써 수사적 해석이 가능한 조직 네트워크 구조가 합성 데이터 상에서도 안정적으로 구현되었음을 확인할 수 있다.



<Figure 6> Degree centrality distribution of key nodes in scenario 2 (relationship-based network model)

4.3.6.2. 부분결합 노드 및 주변부 확산 구조 검증

시나리오 2에서는 드랍퍼 3인을 기준 노드로 설정하고, 이들 중 둘 이상의 드랍퍼와 동시에 연결되는 부분결합 노드(partially coupled node)와 단일 드랍퍼와만 연결되는 개별 주변부 노드가 동시에 형성되도록 관계망 생성 구조를 설계하였다. 이를 통해 현실 범죄 수사에서 자주 관찰되는 핵심 인물-중간 공유 접점-다수의 단일 접촉자로 이어지는 확산형 통신 구조를 synthetic CDR 상에서 재현하고자 하였다.

부분결합 노드는 두 명의 드랍퍼와 동시에 연결되는 중간 접점 노드로 정의되며, 사전 설정 단계에서 해당 유형의 노드를 총 15개 생성하도록 설정하였다. 실제 생성된 네트워크 시각화 결과(초록색 프레임 노드)에서도 두 명의 드랍퍼와 동시에 연결된 부분결합 노드가 정확히 15개 식별되어, 부분결합 노드에 대한 사전 설정 값이 네트워크 구조상 정확히 반영되었음을 확인하였다. 이들 노드는 중간 매개 접점, 공동 거래·상대, 상호 연결 가능성이 있는 중첩 인물군을 모사하는 역할을 수행한다.

한편, 드랍퍼와 단일 연결된 주변부 노드는 총 480개가 생성된 것으로 확인되었다. 이들 노드는 각 드랍퍼에 대해 개별적으로 연결된 단일 접촉자로 구성되며, 고정적인 대인관계, 단기 접

촉자, 일회성 거래 상대, 피해자, 단순 참고인 등 다양한 현실 수사 대상자 유형을 포괄할 수 있는 구조를 형성한다. 특히 주변부 노드의 수가 부분결합 노드에 비해 압도적으로 다수로 형성된 점은, 현실 통신 네트워크에서 소수의 핵심 노드와 다수의 개별 접촉자가 방사형으로 연결되는 구조적 특성이 synthetic CDR 상에서도 자연스럽게 재현되었음을 의미한다.

이와 같이 시나리오 2의 네트워크 구조는 ① 연결의 수 2를 갖는 부분결합 노드 15개, ② 단일 연결 기반 주변부 노드 480개, ③ 이를 매개하는 드랍퍼 핵심 노드 3개, ④ 드랍퍼 핵심 노드들과 모두 연결된 상선 노드로 구성된 이중 확산형 네트워크 구조(dual diffusion structure)를 형성하고 있으며, 이는 실무 수사에서 반복적으로 관찰되는 중첩 중심-개별 확산 혼합형 통신 네트워크 패턴을 구조적으로 타당하게 재현한 결과로 평가할 수 있다.

4.3.6.3. 통화 방향성·유형·시간대 비율 반영 여부 검증

<Table 10> Summary of generated synthetic CDR for scenario 1 (relationship-based network model)

통신 방향성 비율			음성통화/문자 비율		시간대 비율	
다회성 통화자	상대	건수	음성통화	418 (82.12%)	오전(06-12)	77
	63	407			오후(12-18)	202
일회성 통화자	상대	건수	문자	91 (17.88%)	저녁(18-24)	179
	100	100			새벽(00-06)	51

통신 방향성·유형·시간대 비율 반영 여부 검증하기 위해 조우주(010-4110-0539)의 통신내역을 기준으로 확인하였다.

통신 방향성 비율 설정은 전체 통화의 다수가 기존에 형성된 고정 통화 상대방 집단 내에서 반복적으로 발생하고, 나머지 소수가 신규 제3자와의 일회성 통화로 발생하도록 설계된 것이다. 이는 실제 현실 통신 환경에서 소수의 고정 상대와의 반복 통화가 대부분을 차지하고, 다수의 일회성 접촉자가 함께 존재하는 구조적 특성을 반영하기 위한 설정에 해당한다.

실제 생성 결과에서도 대상자별 통화 상대 분석 결과, 각 대상자에 대해 고정 상대방 집단은 소수 인원에 통화가 집중(63명→407건)되는 반면, 신규 제3자 노드는 대부분 1회성 통화로 분산 생성되는 구조가 확인되었다. 이는 본 모델의 방향성 기반 통화 생성 로직이 현실 통신 네트워크의 반복 접촉-일회성 접촉 비대칭 구조를 구조적으로 안정적으로 재현하고 있음을 의미한다.

또한 통신 유형 비율은 음성통화 80%, 문자 20%로 설정되었으며, 실제 생성 결과는 음성통화 418건(82.12%), 문자 91건(17.88%)으로 확인되어 사전 설정한 8:2 비율이 실제 데이터 생성 결과에서도 매우 근접하게 유지됨을 확인하였다.

시간대 비율 역시 오전·오후·저녁·새벽 구간에 서로 다른 가중치를 부여하여 생성되도록 설정되었으며, 생성된 통신 시각 분포 또한 특정 시간대(오후,저녁)에 통화가 집중되는 현실적 패턴을 보이며, 시간대 기반 통신 행태 제어 파라미터가 정상적으로 반영되고 있음을 확인하였다.

4.3.6.4. 공간 이동성 여부 검증

<Table 11> Summary of spatiotemporal transmission patterns of droppers (scenario 2)

대상자	대상자 전화번호	주요 발신기지국	통화횟수	주요 발신 시간대
조우주	010-4110-0539	강원도 춘천시 사북면 지촌리 805-99	58	오후, 저녁
		강원도 춘천시 동내면 사암리 934-93	58	오후, 저녁
		강원도 춘천시 신북읍 용산리 131-19	52	새벽, 오전
임리암	010-3808-8439	강원도 원주시 효저면 고산리 811-26	60	오후, 저녁
		강원도 원주시 소초면 장양리 964-16	50	오후, 저녁
		강원도 원주시 신림면 신림리 945-48	44	새벽, 오전
강윤철	010-3843-2983	강원도 양양군 강현면 전진리 424-41	50	오후, 저녁
		강원도 양양군 양양읍 임천리 146-66	50	오후, 저녁
		강원도 양양군 손양면 금강리 209-66	46	새벽, 오전

드랍퍼 3인의 발신 기지국 위치 패턴을 분석한 결과, 시간대별 활동성과 공간적 이동성이 결합된 현실적 이동 패턴이 확인되었다. <Table 11>에서 제시한 주요 발신 기지국 분포는 무작위 위치생성이 아니라, 모델 내에서 사전 정의된 시간대 기반 위치 규칙(time-dependent location rule)에 의해 "심야·오전(주거지 중심)"과 "오후·저녁(활동지·범죄지 중심)"으로 구조화되어 생성된 결과이다.

구체적으로, 세 명의 드랍퍼 모두 심야(00-06시) 및 오전(06-12시) 시간대에는 소수의 특정 기지국에서 반복적으로 통신하는 경향을 보였다. 이는 해당 시간대 활동이 상대적으로 제한적이고 일정한 위치(주거지, 숙소)에서 이루어지는 전형적 통신 패턴과 일치한다. 이러한 접속 양상은 실제 수사에서 특정 인물이 야간 거점(주거지, 은신처)을 추정할 때 활용되는 분석 기준과 구조적으로 동일한 특성을 보인다.

반면 오후·저녁 시간대에는 동일한 대상자가 서로 다른 활동 지역의 기지국을 중심으로 반복적인 통신을 수행하는 패턴이 관찰되었다. 이는 실무 수사에서 마약 전달, 물품 이동, 공범 접촉 등 범죄 관련 활동이 주로 오후·저녁 시간대에 집중되는 현실적 시간대 분포 특성을 반영한 것으로 해석할 수 있다.

종합적하면, 본 시나리오에서 생성된 synthetic CDR의 위치 데이터는 단순한 무작위 좌표 할당 방식이 아니라, 시간대 구분에 따른 이동 규칙과 주거지-활동지 이중 구조 기반 위치 전환, 대상자별 반복 거점 형성 모델이 결합된 방식으로 생성되었으며, 실제 수사에서 수행되는 거점 식별(anchor point detection), 활동 반경 분석(activity radius), 범죄지 접근성 분석(crime-site accessibility) 등에 즉시 활용 가능한 수준의 공간적 현실성과 분석 유효성을 동시에 갖추고 있음을 확인하였다.

4.4. Synthetic CDR 품질 평가

본 연구에서는 제안한 synthetic CDR 모델의 품질을 통계적 특성, 패턴 재현도, 관계망 구조 적합성, 실무 활용성 등 네 가지 기준에 따라 평가하였다. 평가 결과는 다음과 같다.

4.4.1. 통계적 특성 재현도(Statistical Consistency)

Synthetic CDR은 시간대별 통신량 분포에서 일반 이용자의 통신행태와 유사한 구조를 보였

다. 전체 통신 이벤트는 오후(12-18시)와 저녁(18-24시)에 집중되는 전형적 패턴을 나타냈으며, 이는 모델의 시간대 비율 설정값에 따라 현실적 통신 분포를 반영한 결과이다. 또한 통화시간(duration) 범위, 발신·착신 비율 등 기본적인 통계 분포가 모델의 파라미터에 따라 일관되게 생성되는 것을 확인하였다.

4.4.2. 범죄행동 패턴 재현도(Pattern Reproduction)

마약 드랍 시나리오 적용 결과, 기지국 기반 탐지에서 지역별 3개 기지국의 교집합을 통해 실제 드랍퍼 역할을 수행한 인물을 정확히 식별할 수 있었다. 또한 synthetic CDR 내 시간·공간 패턴은 심야·오전의 제한적 이동(주거지 추정), 오후·저녁 시간대의 활동지 중심 이동과 같은 범죄활동 관련 행동 시그니처를 재현하였다.

이는 모델이 단순한 무작위 생성 방식이 아니라 시간·공간·관계적 요소가 결합된 행동 기반 통신 패턴을 생성할 수 있음을 시사한다.

4.4.3. 관계망 구조 적합성(Network Structural Validity)

Synthetic CDR 기반으로 구성한 드랍퍼 3인의 통신 네트워크는 중심-주변 구조(core-periphery)가 분명히 나타나는 형태로 생성되었다. 이는 사회연결망에서 핵심 노드(core)와 주변 노드(periphery)가 구별되는 구조적 특성을 이론적으로 정의한 Borgatti and Everett(2000)의 core-periphery 모델과 구조적으로 일치한다[8].

특히 중심 노드(드랍퍼), 드랍퍼 3명 모두와 연결된 허브 노드(1명), 드랍퍼 2명과만 연결된 부분 결속 집단(15명), 단일 통신으로만 연결된 주변부 노드(480명)으로 구분되는 구조가 형성되었는데 이 구조에서 각 노드의 Degree Centrality 값은 중심 노드에서 가장 높은 것으로 확인된다. 이러한 중심성은 Wasserman and Faust(1994)가 제시한 Degree Centrality의 해석 기준과 일치한다[9]. 이는 synthetic CDR이 수사적 해석이 가능한 현실적인 관계망 구조를 생성할 수 있음을 실증적으로 보여주는 결과이다.

4.4.4. 실무 활용성 평가(Practical Utility)

Synthetic CDR은 실제 CDR과 동일한 구조(발신·착신 번호, 시간 정보, 통신유형, 기지국 주소 등)를 갖추고 있어 수사관·분석가가 기존 분석 방식 그대로 활용할 수 있다. 특히 시나리오 1에서는 지역별 기지국 교차 분석을 통해 드랍퍼 3명이 모두 정확히 식별되어, 기지국 기반 용의자 특정 절차의 구조적 재현성이 확인되었다. 또한 시나리오 2에서는 중심-허브-주변부로 구분되는 조직형 네트워크 구조에서 상선 노드가 명확히 도출되어, 관계망 기반 상선 특정 절차 역시 실무적 관점에서 타당하게 재현됨이 검증되었다. 이 결과는 synthetic CDR이 수사 R&D, 분석 알고리즘 검증, 모델 테스트 등 다양한 분야에서 활용 가능한 수준의 품질을 갖추고 있음을 시사한다.

추가적으로, 본 연구에서 생성된 synthetic CDR은 수사자료분석 교육 과정의 실무 훈련에 적용하여 현직 수사관을 대상으로 검증되었다. 훈련 참여자들은 사전 정답이 주어진 상태가 아닌 상황에서 독립적으로 분석을 수행하였으며, 이후 비교 결과 동일한 핵심 대상자와 상선 후보가 높은 일치율로 도출되었다. 수사관들은 데이터가 “실제 통신내역과 매우 유사하며”, “실제 수사 상황과 같은 분석 흐름·난이도를 제공한다”고 평가하였다.

이러한 교육 기반 검증 결과는 본 synthetic CDR이 단순 시뮬레이션 데이터가 아니라 수사

실무·교육 환경에서도 즉시 활용 가능한 고현실성 데이터셋임을 뒷받침한다.

4.5. 종합 분석 및 시사점

본 연구에서는 고현실성 synthetic CDR 생성 모델을 구축하고, 기지국 기반 용의자 특정·관계망 구조 분석·시간·공간 패턴 검증을 포함한 두 가지 실험을 통해 모델의 유효성을 평가하였다. 그 결과, 본 모델이 실제 통신데이터 분석 연구와 수사 실무에서 관찰되는 통신 패턴과 구조적 특징을 다각적으로 재현할 수 있음을 확인하였다.

첫째, 기지국 분석 실험(시나리오 1)에서는 지역별 다중 기지국의 교집합 분석을 통해 드랍퍼를 정확히 특정할 수 있었으며, 이는 synthetic CDR이 실제 수사 절차(특히 기지국 기반 용의자 탐색)를 충실히 모사할 수 있음을 보여준다.

둘째, 관계망 분석 실험(시나리오 2)에서는 드랍퍼 중심의 중심-허브-주변부 구조가 명확하게 형성되었고, 부분 결속 집단 및 중복 통화 상대방 등 실제 조직범죄에서 특징적으로 나타나는 구조가 적절히 재현되었다.

셋째, 시간·공간 패턴 분석에서는 드랍퍼의 활동시간대, 야간 정착지 패턴, 주거지·아지트 추정이 가능하여, synthetic CDR이 행동 기반 수사 모델링에 활용될 수 있음을 확인하였다.

이상의 결과는 본 연구의 synthetic CDR이 단순한 랜덤 데이터가 아니라, 수사 절차·범죄행동·조직 구조를 설명하는 실증적 패턴을 재현하는 고품질 시뮬레이션 데이터임을 의미한다. 특히 실제 수사관을 대상으로 한 교육 훈련에서 동일한 대상자와 상선 후보가 일관되게 특정된 점은 실무 적용 가능성을 뒷받침하는 중요한 근거로 작용한다.

다만, 본 모델은 '대상자 중심의 발·착신 데이터 구조'를 기반으로 하므로, 수집되지 않은 제3자의 관계가 네트워크에 반영되지 않는 구조적 한계를 가진다. 이는 실제 수사데이터 역시 동일한 제약을 가지므로 현실성 측면에서는 오히려 타당하나, 향후 다양한 수집 범위·관찰 기간을 모사한 확장 모델 개발이 필요하다.

향후 연구에서는 다른 범죄유형(보이스피싱, 다중사기, 뇌물 등)의 특성을 반영한 통신 패턴 확대, 머신러닝 기반 통신행동 이상 탐지 실험, 대규모 synthetic CDR을 이용한 AI 모델 훈련 등으로 확장할 수 있을 것이다.

종합하면, 본 연구의 synthetic CDR 생성 모델은 수사 R&D, 교육훈련, 알고리즘 평가, AI 모델 개발, 분석도구 검증 등 다양한 실무·학술 영역에서 활용될 수 있는 실질적 기여를 제공하며, 향후 디지털 수사 분야에서 synthetic Data의 활용 가능성을 제시한 기반 연구로 의의가 있다.

V. 결론 및 논의

5.1. 연구의 의의

본 연구는 국내에서 최초로 범죄 수사용 synthetic CDR 생성 모델을 체계적으로 설계하고 그 활용 가능성을 실증적으로 검토하였다는 점에서 중요한 의의를 가진다. 기존의 합성데이터 연구가 개인정보 비식별화 또는 산업·학술 연구 중심의 합성데이터 생성에 머물렀던 반면, 본 연구는 실질적인 수사 프로세스와 범죄특성 기반 행동 패턴을 반영한 데이터 생성 모델을 제시하였다. 이는 합성 통신데이터가 단순한 무작위(random) 기반 생성 데이터를 넘어, 실제 수사 현장에서 활용 가능한 '행동 기반 사건 데이터'로 발전할 수 있음을 보여주는 중요한 사례이다.

또한 본 연구는 기존의 비식별화 데이터 활용 방식이 가진 구조적 한계를 극복할 수 있는 대

안을 제시한다. 비식별화된 통신자료는 개인정보 보호 요건을 충족하지만, 범죄조직의 관계망·행동 패턴·시간적·공간적 변화 등을 담기 어렵다는 문제점이 있다. 반면 synthetic CDR은 오히려 이러한 구조적 요소를 의도적으로 심을 수 있어, 범죄 패턴 연구·수사 훈련·도구 검증 등 다양한 목적에 유연하게 활용될 수 있다는 점에서 범죄 수사 R&D 분야에서 높은 가치를 가진다.

해외의 synthetic CDR 연구와 비교하더라도 본 연구의 독창성은 명확하다. 해외 연구는 인간 이동성·사회적 행동 패턴·이상탐지등 민간·산업 중심의 활용에 초점이 맞추어져 있었으며, 범죄 수사 관점에서의 통신 행동 모델링이나 조직형 범죄 구조 재현을 직접적으로 다룬 연구는 매우 제한적이었다. 반면 본 연구는 한국의 전화번호 체계, 기지국 기반 공간 구조, 범죄유형별 행동 패턴 등 실무 기반 요구사항을 종합적으로 반영하여 수사 맥락에 맞춘 데이터 생성 모델을 현실적으로 구현했다는 점에서 연구적 기여가 크다.

더불어 synthetic CDR은 개인정보를 포함하지 않으면서도 실제 수사 분석에 필요한 구조적 특성을 유지할 수 있어, '안전한 개방형 수사 데이터 생태계' 구축의 핵심 인프라로 기능할 수 있다. 이는 향후 수사기법 연구, 경찰 교육훈련, AI 기반 수사도구 개발 과정 전반에서 기존의 실데이터 활용 제약을 해소할 수 있는 실질적인 대안을 제공한다.

5.2. 연구의 한계

그러나 본 연구에도 몇 가지 한계가 존재한다.

첫째, 본 모델은 범죄유형별 행동 패턴이 정형화되어 있는 범죄에는 효과적이지만, 돌발성·비정형성 범죄의 복잡한 행동 양상을 모두 반영하기에는 한계가 있다. 예컨대 휴대전화 외 텔레그램 등 익명 메신저 사용, 일회성 대포번호 등을 활용하는 경우 일정한 규칙 기반 생성 방식만으로는 충분히 모사 하기 어렵다. 이러한 유형의 범죄는 사전 정의된 패턴보다 행위자의 순간적 판단과 환경적 요인에 크게 의존하므로, 합성 모델이 이를 완전하게 반영하는 데에는 구조적 제약이 따른다.

둘째, 기지국 기반 위치 시뮬레이션은 실제 수사에서 활용되는 다중 기지국 중복 발신자 분석(intersection-based identification)의 절차적 구조를 충실히 재현하고 있으나, 현실의 CDR에서 나타나는 위치 불확실성까지는 완전히 모사하지 못한다는 한계가 있다. 실제 통신내역에서는 통화량 집중에 따른 인접 기지국 전환(overflow handover), 기지국 커버리지의 비대칭성, 도심 외 지역의 광범위한 셀 범위 등으로 인해 이용자의 물리적 위치가 정밀하게 반영되지 않는 경우가 빈번하다. 반면, 합성 데이터는 이러한 인프라-기반 제약 없이 지나치게 정확한 위치 패턴을 생성할 가능성이 있어, 실제 CDR에서 관찰되는 오차 특성(error characteristics)과 차이를 발생시킬 수 있다. 향후 연구에서는 이러한 현실적 위치 노이즈(location uncertainty)를 모델링하여 합성 데이터의 수사적 활용성을 더욱 높일 필요가 있다.

셋째, 범죄조직 내부의 심리·의사결정 요인과 같은 고차원 행동 특성은 관측 가능한 데이터만으로 완전하게 모델링하기 어렵다. 범죄조직의 의사결정은 단순한 통신 빈도나 네트워크 구조뿐 아니라, 조직 내 위계 관계, 리더의 판단 성향, 위험 감수 성향, 내부 신뢰 수준, 외부 수사 압박 인지에 따른 전략 변화 등 비가시적(invisible) 요소에 의해 크게 좌우된다. 이러한 요인은 일반적인 CDR 데이터로는 직접 관측이 불가능하며, 시점별 판단 변화나 심리적 동요에 따른 비선형적 행동 변동 또한 규칙 기반 생성 모델에서 정밀하게 반영하기 어렵다는 한계를 가진다.

넷째, 본 연구에서 생성된 synthetic CDR은 범죄조직 내 기능적 관계뿐만 아니라 피의자의 일상적·사적 관계까지 동시에 포함된 통신 네트워크 구조로 구현된다. 이러한 사적 관계망은 중심성, 하위 집단 탐색 등 네트워크 지표 분석 과정에서 실제 범죄조직 구조와 구분되지 않은 채

함께 반영될 가능성이 있으며, 이 경우 수사적 관점에서는 노이즈로 작용할 수 있다. 현실의 CDR 분석에서도 피의자의 가족·지인·업무상 연락처 등 수사배제 대상자가 혼재되어 나타나듯, 합성 데이터에서도 일정 비율의 비범죄 관계가 자연스럽게 포함될 수밖에 없다. 따라서 사적 관계로부터 발생하는 연결망 구조상의 영향을 정량적으로 구분·평가하는 작업은 본 연구에서 충분히 다루지 못한 한계로 남는다.

5.3. 향후 연구 방향

향후 연구에서는 다음과 같은 확장 방향이 필요하다.

첫째, 범죄 유형별 행동 모델을 보다 정교화하여, 조직형 범죄뿐 아니라 우발·비정형 범죄 영역까지 포괄할 수 있는 하이브리드 생성 구조의 도입이 요구된다.

둘째, 기지국 전파 특성, 지역별 통신 인프라 차이, 기지국 간 중첩 효과 등을 반영한 확률 기반 위치 노이즈 모델을 도입하여 위치 시뮬레이션의 현실성을 더욱 고도화할 필요가 있다.

셋째, synthetic CDR과 계좌 거래, 교통 정보, 위치 정보, 가상자산 데이터 등 이종 데이터를 결합한 융합형 합성 데이터 생성 모델의 개발이 요구된다. 이러한 융합 데이터는 실제 수사 환경과 유사한 복합 데이터 구조를 형성함으로써 AI 기반 예측·이상탐지 모델의 학습 성능을 실질적으로 향상시킬 수 있다.

넷째, 합성 단계에서 사적 관계와 범죄 관계를 구분하여 생성하거나, 분석 단계에서 사적 관계를 체계적으로 식별·배제할 수 있는 필터링 모형을 결합함으로써, 범죄조직 구조에 초점을 맞춘 정교한 관계망 분석이 가능하도록 발전시킬 필요가 있다.

5.4. 종합 결론

결론적으로 본 연구는 범죄 수사용 고현실성 synthetic CDR 생성 모델의 구조적 방향성과 실무적 활용 가능성을 실증적으로 제시한 기초 연구로서, 향후 수사정보 분석, AI 기반 수사 환경을 구축하는 데 중요한 학술적·실무적 토대를 제공한다. 본 연구에서 제시한 합성 통신데이터 생성 모델은 수사 데이터 확보의 현실적 한계를 극복하고, 안전하면서도 실효성 있는 데이터 기반 수사 연구 생태계 구축을 가능하게 하는 핵심 기반 기술로 발전할 수 있을 것으로 기대된다.

참고문헌 (References)

[1] Personal Information Protection Commission. 2024. Guidelines for generating and using synthetic data [합성데이터 생성·활용 안내서]. Personal Information Protection Commission, Seoul, Korea.

[2] Candia J, González MC, Wang P, et al. 2008. Uncovering individual and collective human dynamics from mobile phone records. Journal of Physics A: Mathematical and Theoretical, 41, 224015. https://doi.org/10.1088/1751-8113/41/22/224015

[3] Sultan K, Ali H, Zhang Z. 2018. Call detail records driven anomaly detection and traffic prediction in mobile cellular networks. IEEE Access, 6, 41728-41737. https://doi.org/10.1109/ACCESS.2018.2859756

[4] Songailaite M, Krilavičius T. 2021. Synthetic call detail records generator. CEUR Workshop Proceedings, Information Society and University Studies 2021 (IVUS 2021), Kaunas, Lithuania, pp. 1-10.

[5] Lee J. 2014. Constructing a social contact network based on cellphone call records and analysis of its scale-free property. Journal of the Korean Institute of Industrial Engineers, 40(1), 1-7. https://doi.org/10.7232/JKIIE.2014.40.1.001

[6] Oh J, Kang JS, Ryu YS. A study on balancing privacy protection and utilization of criminal justice information: Focusing on synthetic data as a de-identification technology. Criminal Investigation Studies, 11(2), 179-203. https://doi.org/10.46225/CIS.2025.8.11.2.179

[7] Moon BH, Kwahk KY. 2024. A case study on the investigation of cyber gambling criminal organizations using social network analysis. Information Systems Review, 26(3), 1-22. https://doi.org/10.14329/isr.2024.26.3.001

[8] Borgatti SP, Everett MG. 2000. Models of core/periphery structures. Social Networks, 21(4), 375-395. https://doi.org/10.1016/S0378-8733(99)00019-2

[9] Wasserman S, Faust K. 1994. Social network analysis: Methods and applications. Cambridge University Press. https://doi.org/10.1017/CBO9780511815478