진짜 같은 가짜 ! 재현데이터의 개념 및 활용 사례

이재근 팀장 (jglee@kcredit.or.kr)

〈요 약〉

▶ 재현데이터(Synthetic Data)는 개인의 프라이버시를 보호하면서도 민감한 정보를 분석하고자 하는 연구자들에게 데이터를 제공 할 수 있는 대안적 개인정보 비식별 조치 기법의 하나

- ○ 실제로 측정된 데이터(Real Data)를 생성하는 모형이 존재한다고 가정하고, 통계적 방법이나 기계학습 방법 등을 이용하여 추정된 모형에서 새롭게 생성한 모의데이터(Simulated Data)
- ○ 모집단의 통계적 특성들을 유지하면서도 민감한 정보를 외부에 직접 공개하지 않으며, 개인이 제공한 데이터가 아닌 임의로 생성한 데이터로 개인정보보호 관련 법규의 규제로부터 자유로운 특징이 존재

▶ 해외에서는 재현데이터 기법을 이용하여 개인의 민감한 정보가 들어있는 데이터를 연구자들이 분석에 활용할 수 있는 체계를 지원하고 있으나, 우리나라는 도입 초기 단계

- ○ 미국(조사통계국), 영국(경제사회이사회), 독일(노동청) 등은 정부차원에서 소득, 조세 데이터 등을 재현데이터로 개발하여 관련 연구자들에게 제공하고 연구자들의 분석결과를 검증해 주는 분석 체계를 운영
- ○ 국내에서는 통계청, 일부 공공기관 등에서 기초연구 및 재현데이터 개발 등을 추진 중

▶ 신용정보원에서는 금융 빅데이터 개방시스템(CreDB) 서비스 이용자의 분석 및 교육 지원 등을 위해 개인신용정보 표본DB에 대한 재현데이터를 개발

- ○ 신용정보원에 등록된 신용정보(대출, 연체, 신용카드개설)를 기반으로 약 180만 명에 해당하는 가상 차주에 대한 25개월('16.12월말 ~ '18.12월말) 분의 재현데이터를 개발
- ○ 신용정보원에서 개발한 신용정보 재현데이터는 4개의 테이블로 분리된 시계열 형태의 대규모 원시데이터(Raw Data)를 국내 최초로 재현했다는 점에서 의의가 있음

▶ 그동안 금융업권의 신용정보는 정보의 민감성으로 인해 금융권 이외의 분야에서는 접해보기 어려워 다양한 분석과 활용이 제한되었으나 재현데이터는 데이터 활용 제한의 문제를 일부 해소 가능

- ○ 법·제도적인 규제로 인해 그동안 공개되지 않았거나, 타 업종과 융합 분석에 제약이 많았던 민감한 금융데이터를 재현데이터 형태로 개발하고 분석, 교육 등에 활용하는 방안을 모색해볼 필요 있음

▶ 신용정보원은 향후 금융분야의 다양한 데이터를 안전하면서도 통계적 정확성이 높은 재현데이터로 추가 개발하고 금융업권 및 학계 등 다양한 분야에서 활용될 수 있도록 노력할 계획

I. 배경

▶ **Deep Fake는 영상이나 사진 등을 인공지능을 통해 합성 또는 조작하여 진짜 같은 가짜 영상 등을 만드는 기술을 의미하는데, 이는 민감한 개인정보를 이용하는 분야에도 유용하게 활용될 수 있음**

○ 개인정보가 포함된 데이터를 개인의 프라이버시를 보호하면서 안전하게 분석 및 활용할 수 있도록 하는 기법으로는 개인정보 비식별 조치가 주로 이용되고 있으나, 진짜 같은 가상의 데이터를 이용하는 재현데이터(Synthetic Data) 기법도 있음

[인공지능을 이용하여 유명인의 사진을 기반으로 만들어진 가상의 인물]

자료: 엔비디아

▶ **특정 개인을 알아볼 수 없도록 조치하는 개인정보 비식별 조치 기법은 개인정보가 포함된 데이터를 안전하게 활용하기 위해 널리 쓰이고 있으나 사전 및 사후적인 관리 이슈가 존재**

○ 가명처리, 총계처리, 범주화 등의 비식별 조치 기법은 실제로 수집된 데이터의 일부를 바꾸거나 변형 또는 삭제 시켜서 개인정보의 노출 위험을 줄이는 기법으로 널리 쓰이고 있음

○ 그러나 비식별 조치 기법이 적용된 데이터가 특정 개인을 알아볼 수 있는지에 대해 사전 및 사후적으로 평가하고 관리해야 하는 관리적 부담이 존재

▶ **이에 반해 재현데이터는 실제데이터의 통계적 특성을 고려하여 임의로 생성한 가상의 데이터로 모집단의 통계적 특성들을 유지하면서도 민감한 정보를 외부에 공개하지 않는 기법으로 관심 받고 있음**

○ 재현데이터 생성을 위해 실제데이터를 이해하고, 재현하기 위한 통계모형을 찾는 것은 매우 어려운 작업으로, 재현데이터를 생성하는 다양한 기법이 존재하나 이중 다중대체법이 가장 널리 알려져 있음

▶ **미국, 영국, 독일 등에서는 민감한 개인정보가 포함된 데이터를 분석하고자 하는 연구자들에게 분석 대상의 개인정보 노출 위험을 줄이면서도 관련 연구의 활성화를 위해 재현데이터를 제공**

- 국내에서도 재현데이터를 이용하여 법규 등의 규제로 인해 분석 또는 공개가 어려웠던 데이터를 다양한 분야에 활용하는 사례가 점차 나타나고 있는 상황

▶ **본 보고서[1]는 대안적 개인정보 비식별 조치 기법인 재현데이터의 개념 및 특성과 국내외 활용 사례 등을 소개하고 금융 분야에 적용할 수 있는 방안에 대해 검토해 보고자 함**

II. 재현데이터의 개념 및 특성

1. 재현데이터의 정의 및 종류

▶ **재현데이터에 대한 초기 연구는 설문 등의 결측값을 대체하기 위하여 하버드대의 통계학과 교수였던 Donald B. Rubin 교수가 1981년에 제시한 다중대체법(Multiple Imputation)에 근간을 두고 있음**

- 다중대체법은 표본조사에서 결측치가 발생하였을 때 결측된 변수의 값에 대한 예측모형을 세우고, 결측값을 여러 개의 예측값으로 대체한 후 분석을 실시하는 절차
- Rubin 교수는 1993년에 자신이 제안한 다중대체법을 이용해서 완전 재현데이터를 생성하는 방법을 최초로 제안(Journal or Official Statistics)

▶ **McGraw-Hill 과학·기술 사전에서는 재현데이터를 직접 측정으로 획득되지 않은 주어진 상황에 적용되는 모든 생산 데이터로 정의[2]**

- 위키피디아에서는 재현데이터를 실제로 측정하지 않은 임의의 데이터로 넓게 정의하기도 하지만, 통상적으로 추정된 모형에서 생성된 가상의 데이터를 의미
- 이러한 논의 아래 재현데이터의 개념을 재정의 해보면 실제로 측정된 데이터(Real Data, 이하 실제데이터)를 생성하는 모형이 존재한다고 가정하고, 통계적 방법이나 기계학습 방법 등을 이용하여 추정된 모형에서 새롭게 생성한 모의데이터(Simulated Data)라고 할 수 있음

1) 신용정보원에서 운영한 '재현데이터 연구반'의 연구 결과인 '재현자료의 국내외 활용 현황 및 금융분야 적용 가능성 검토' 보고서와 신용정보원이 재현데이터 기술을 이용하여 수행한 신용정보 교육용DB 개발의 결과 등을 기반으로 작성

2) Any production data applicable to a given situation that are not obtained by direct measurement, McGraw-Hill Dictionary of Scientific and Technical Terms

▶ **재현데이터는 데이터 생성 방법에 따라 완전 재현데이터, 부분 재현데이터, 복합 재현데이터로 구분되며, 이중 완전 재현데이터의 보안성이 가장 우수한 것으로 알려짐**

○ 완전 재현데이터(Fully Synthetic Data)는 공개하려고 하는 데이터에 측정된 실제데이터가 하나도 없이 모두 가상으로 생성된 데이터로만 이루어진 데이터를 의미하며, 정보보호 측면에서 가장 강력한 보안성을 가짐

○ 부분 재현데이터(Partially Synthetic Data)는 공개하려는 변수들 중 일부만을 선택하여 재현데이터로 대체한 데이터를 의미하며, 보통 재현데이터로 대치되는 변수들은 민감한 정보에 관한 변수들임

○ 복합 재현데이터(Hybrid Synthetic Data)는 일부 변수들의 값을 재현데이터로 생성하고 생성된 재현데이터와 실제데이터를 모두 이용하여 또다른 일부 변수들의 값을 다시 도출하는 방법으로 생성

▶ **재현데이터를 생성하는데 사용되는 알고리즘으로는 전통적 통계 또는 베이지안 방법, 기계학습 모형 방법, 차등정보보호에 의한 방법 등이 있음**

[재현데이터 생성 기법]

전통적 통계 또는 베이지안 방법 (Bayesian Methods)	기계학습 모형 (Machine Learning Model)	차등정보보호 (Differential privacy)
- Mutiple Imputation - Bayesian Network - Perturbed Gibbs Sampler - Bayesian Method with zero-inflation - Re-sampling from Multivariate Distribution	- Semantic Graph based method - MDL(Minimal Description Length) based KRIMP algorithm - CART(Classification And Regression Tree) - Fuzzy c-regression Models - Support Vector Machine - Random Forest - Recommendation Systems - Social Network Model - Generative Adversarial Network	- Proposed Multiplicative Weights update rule with Exponential Mechanism (MWEM) - Differentially Private Data Synthesizer - Mapping Program

자료 : 신용정보원 재현데이터 연구반 결과 보고서

▶ **재현데이터는 실제데이터를 설명하는 모형에서 임의로 생성한 가상의 데이터이기 때문에 이를 생성하고 활용하는데 몇가지 특징이 있음**

○ 데이터들의 복잡한 관계를 설명할 수 있는 모형에 근거하기 때문에, 실제데이터의 여러 가지 통계적 특성들이 유지되는 특징을 가지나, 원본데이터와의 지속적인 비교 등을 통해 신뢰성을 확보해야 하는 번거로움도 존재

○ 재현대상 변수의 수가 증가할수록 원본의 통계적 특징을 유지하도록 하는 알고리즘 생성에 제약이 생기며, 원본의 크기가 크거나 시계열 데이터인 경우에는 모형의 추정이 어려움

○ 재현데이터는 생성 방법에 따라 데이터의 특성 및 제한점이 존재

- 대부분의 생성 알고리즘은 완전 재현데이터를 생성하기 위하여 개발되었으며, 부분 재현데이터를 생성하는 방법도 다양하지만 완전 재현데이터에 비하여 보안성이 약하다고 알려져 있음

○ 개인정보를 포함하는 실제데이터를 기반으로 생성한 재현데이터는 통계적 모형에 의해 임의로 생성된 가상의 데이터이기 때문에 개인정보보호 등의 규제로부터 자유로운 측면이 있음

2. 재현데이터 생성 도구

○ 재현데이터를 생성할 수 있는 다양한 소프트웨어가 오픈 소스 형태 등으로 공개되어 있으며, 통계분석 오픈소프트웨어인 R 기반의 sms, synthpop, simPop 등의 패키지 소프트웨어도 존재

[재현데이터 생성 도구]

구분	관련 홈페이지 URL	세부 내용
PoPGen	http://simtravel.wikispaces.asu.edu	Arizona State University의 SimTRAVEL Research Initiative에서 개발되었으며, 상대적으로 전수 정확도가 높은 반복비율갱신(Iterative Proportional Updating) 알고리즘으로 전수인구 데이터를 생성 가능
Virtual Belgium	https://sourceforge.net/projects/virtualbelgium	인구통계, 주거선택, 활동패턴, 이동성 및 기타 정보를 시뮬레이션하여 벨기에의 인구 변화를 관찰하는 프로그램. 반복비례적합(Iterative Proportional Fitting) 알고리즘을 사용
MoSeS	https://royalsocietypublishing.org/doi/10.1098/rsta.2009.0041	실제 도시 및 지역 시스템을 위한 인구 정보를 재현하여 향후 25년 이후의 인구정보를 예측 가능. 영국의 인구정보 재현 모델은 유전 알고리즘을 기반으로 구현
TRANSIMS	https://sourceforge.net/projects/transimsstudio	미국 Los Alamos National Laboratory의 연구원이 개발한 운송 분석 시뮬레이션 시스템. 인구조사 마이크로 데이터를 기반으로 재현데이터 생성
Synthia	http://synthpopviewer.rti.org	비영리 연구 기관인 RTI에서 개발한 웹 기반 재현데이터 생성 프로그램으로, 사용자 정의 변수를 사용하여 사용자가 정의한 학습 영역에 대한 재현데이터를 생성
sms (R 기반 Lib.)	https://cran.r-project.org/web/packages/sms/index.html	주어진 영역 내 매크로 데이터로부터 마이크로 데이터를 시뮬레이션하는 기능을 제공. 계층적 구조의 데이터 처리는 불가능하나, SA(Simulated Annealing)를 단순화하여 제한된 영역에 대한 설명을 최적화하는 기능 존재
synthpop (R 기반 Lib.)	https://cran.r-project.org/web/packages/synthpop/index.html	분류회귀모형을 사용하여 재현데이터에 대한 변수를 생성. 정교한 샘플링 디자인을 필요로 하거나 가구 및 구성원 정보 등과 같은 계층 또는 클러스터 구조를 가진 데이터를 처리하는 기능은 없으나 편의성이 높음
simPop (R 기반 Lib.)	https://cran.r-project.org/web/packages/simPop/index.html	주체가 가진 속성 값에 따라 다르게 적용되는 정책의 거시적인 효과를 예측하기 위한 복잡한 구조의 데이터 재현에 매우 유용. 가구와 가구 구성원 정보 등 계층적 구조를 처리 가능. IPF와 SA를 사용한 통계량 조정, 로지스틱 회귀를 통한 모델링 등의 기능을 제공

자료 : 신용정보원 재현데이터 연구반 결과 보고서

III. 국내·외 재현데이터 활용 사례

1. 해외 사례

▶ 미국, 영국, 독일 등에서는 재현데이터 기법을 이용하여 개인의 민감한 정보가 들어있는 데이터를 생성하고, 이를 연구자들이 분석에 활용할 수 있는 체계를 지원

▶ 미국의 SIPP Synthetic Beta(SSB) 사례

○ 미국 조사통계국(Census Bureau)은 기밀성 문제로 인해 공개적으로 이용할 수 없는 SIPP (Survey of Income and Program Participation) 데이터를 연구자들에게 제공하기 위해 SIPP를 부분재현데이터로 구현하여 제공

- 세금, 수입 등의 정보를 포함하는 개인단위 데이터를 코넬대학 연구진과 협력하여 재현데이터로 구현하고, 연구자들이 통제된 분석 서버에서 분석할 수 있도록 제공

○ SSB는 대중에게 공개된 재현데이터 중 가장 광범위하게 검증된 데이터로서 재현데이터에서 얻은 주요 변수에 대한 통계치가 원데이터에서 얻은 것과 매우 유사한 것으로 알려져 있음

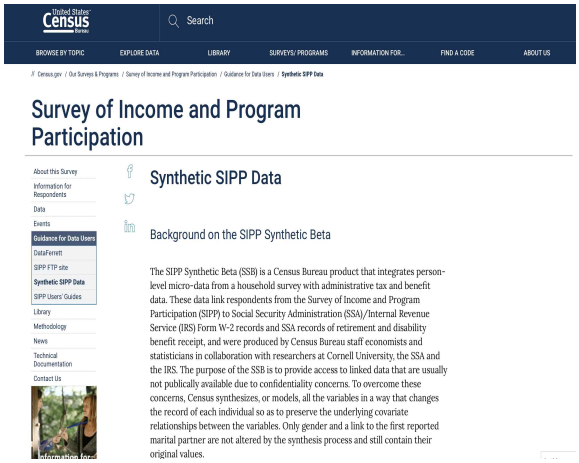
- 성별(Gender)과 배우자(Spouse Link) 정보를 제외한 모든 변수를 재현한 600개 이상의 변수로 구성되어 있으며 16개의 다중 재현데이터 세트를 제공
- SSB는 총 9개의 SIPP 패널(1984, 1990, 1991, 1992, 1993, 1996, 2001, 2004, 2008) 데이터로 구성

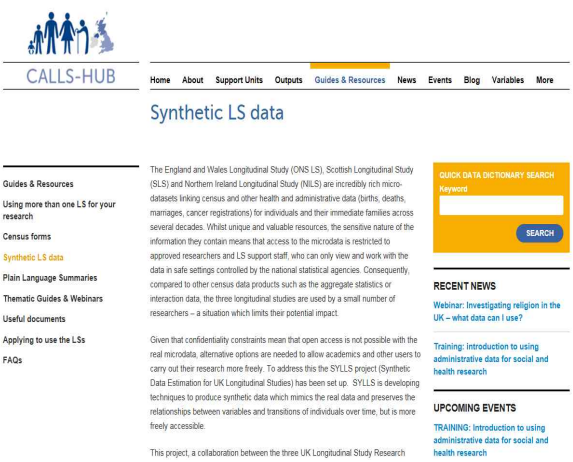
○ SSB는 1) 이용 신청서 제출 2) 이용 승인, 3) 분석 계정 할당, 4) 분석결과 공유 및 타당성 검증 등 4단계의 이용 절차를 거쳐 제공

[SSB 이용 절차]

1단계	2단계	3단계	4단계
이용 신청서 제출	이용 승인	분석 계정 할당	분석결과 공유 및 타당성 검증
- 신청자의 정보, 프로젝트 개관, 필요한 변수를 작성하여 제출 - 이 때 필요한 변수는 SIPP데이터 변수 설명을 참조하여 작성	- 조사통계국의 심사를 거쳐 연구자에게 접근 가능 여부를 통보	- 코넬대학의 Synthetic Data Server(SDS)를 통해 분석 - 연구자는 SDS에서 SAS, Stata를 활용하여 분석 - 데이터에 대한 접근만 가능하고 다운로드는 불가	- 이용자가 제공한 분석 코드(SAS 및 Stata Code)를 활용하여 원자료를 분석한 결과를 이용자의 분석 결과와 비교하여 타당성 검증

[미국의 SIPP와 영국의 CALLS-HUB 홈페이지]





▶ 미국의 Synthetic Longitudinal Business Database(SynLBD) 사례

- SynLBD는 미국의 조사통계국과 듀크대학, 코넬대학, 국립통계과학원(NISS), 미국과학재단(NSF) 등이 공동으로 미국의 사업체(Business Establishments)들을 대상으로 하는 센서스에 대한 부분 재현데이터를 개발 및 제공하는 프로그램
- 1976년부터 2000년까지 미국의 약 2,100만 개 업체들에 대한 센서스 자료이기 때문에 샘플링에 기반한 일반 서베이 자료에 비하여 노출위험이 있어, 공개되는 변수 중 산업코드(Industry Code)를 제외한 모든 변수를 순차회귀-다중대체방법을 이용하여 생성한 재현데이터로 제공
- 이용자는 SSB와 같은 절차를 통해 SynLBD 데이터를 분석 가능

▶ 영국의 Synthetic Longitudinal Studies Data 사례

- 영국의 경제사회이사회(ESRC, Economic and Social Research Council)는 세인트앤드류스대학, 에딘버러대학, 런던대학 등과 함께 UK Longitudinal Studies[3] 자료에 대한 개별 맞춤형 재현데이터를 완전 재현데이터 형태로 제공
- SynLBD과 같이 순차회귀-다중대체방법을 이용하여 재현데이터를 생성하고, 각 변수들에 대한 회귀모형으로 CART(Classification And Regression Trees)를 기본 모델로 하여 적용
- 미국과 달리 개별 사용자의 연구 목적에 따른 맞춤형 재현데이터를 제공하고 있으며 이를 지원하기 위해 재현데이터를 생성하고 분석하는 R 패키지인 synthpop을 개발하여 공개

3) UK Longitudinal Studies는 England and Wales Longitudinal Study(ONS LS), Scottish Longitudinal Study(SLS), Northern Ireland Longitudinal Study를 통합한 데이터로서 개인의 출생, 이민, 교육 및 건강 관련 데이터를 포괄하는 방대한 마이크로데이터

▶ 독일 노동청의 사업장 패널(IAB Establishment Panel) 사례

- 독일 노동청(Institute of Employment Research)은 독일의 사업장 패널(IAB Establishment Panel)[4]에 대하여 미국 조사통계국의 SIPP와 SynLBD 사례를 참고하여 완전 재현데이터를 생성
- 이를 기반으로 직업교육과 기업의 생산성과의 관계에 대한 연구, 기업의 임시직 고용과 사업장의 크기, 인사조직 구조 등과의 관계에 대한 연구 등을 수행

2. 국내 사례

▶ 통계청을 중심으로 재현데이터 방법론에 대한 연구가 진행

- 국내에서의 재현데이터에 대한 연구는 통계청을 중심으로 방법론적인 기초 연구가 수행됨
  - 통계청은 마이크로 데이터 공개에 따라 민감한 개인정보 노출 위험을 줄이면서, 정보손실을 최소화하기 위한 방안으로 재현데이터 작성 방법론과 해외 사례 등에 대한 기초 연구 수행[5]

▶ 한국정보화진흥원과 KCB는 제주도 주민의 정보를 재현데이터로 구현하는 프로젝트를 추진 중[6]

- 개인정보 노출 위험 및 열람 권한 설정 등의 제약으로 인해 수집 및 공개가 어려웠던 데이터를 재현데이터로 구현하고 공공 데이터 플랫폼을 이용하여 공개하여 관련 연구자가 이용하도록 추진 예정
  - 제주도에 거주하는 주민에 대한 인구분석이나 사회경제 분석을 위해 직업, 소득, 소비 등의 데이터를 재현데이터로 구현하고 이를 공개하는 프로젝트를 추진 중

▶ 신용정보 교육용DB 개발을 위한 개인신용정보 재현데이터 개발

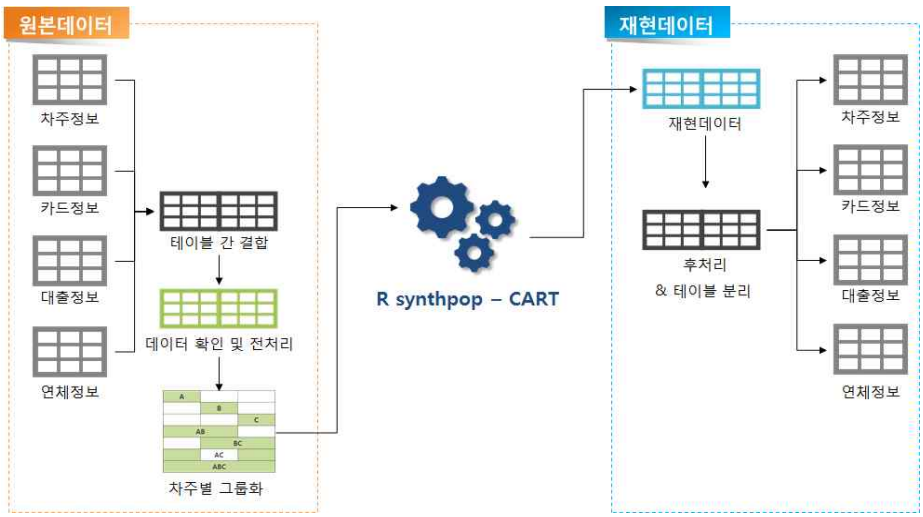
- 신용정보원은 금융 빅데이터 개방시스템(CreDB) 이용자에 대한 분석 및 교육 지원 등에 활용하기 위해 세종대학교 인공지능-빅데이터 연구센터와 함께 국내 최초로 개인신용정보를 재현데이터로 개발
  - 민감한 신용정보를 이용자 교육 등에 직접 활용하기에는 관련 법규 및 개인정보보호 차원에서 많은 제약이 존재하여, 이를 극복하기 위해 신용정보 재현데이터를 개발하고 활용하는 방안을 추진
- 개인신용정보DB 내 차주정보, 대출 및 연체 정보 등을 시계열적으로 분석할 수 있도록 지도학습 알고리즘인 CART(Classification And Regression Tree) 기법을 이용하여 재현데이터를 개발
  - 범주형과 연속형 데이터가 혼재되어 있고, 각 변수별 해석을 목적으로 하는 분석보다는 기간별로 사회적 또는 경제적 현상을 전반적으로 분석하는 목적이 크다는 점을 고려하여 CART 기법을 이용

4) 독일 노동청 산하 노동시장·직업연구소(IAB) 에서 수행하는 표본조사로 서독에서는 1993부터 동독에서는 1996년부터 매년 수행. German Social Security Data(GSSD)에 등록되어있는 고용인을 가진 사업장에 대하여 건강보험, 연금, 실업보험 등의 정보를 포함. 표본의 크기는 초기 4,000-5,000개 정도에서 2007년 15,000여 사업장으로 지속적으로 증가하여 왔고 조사되는 항목은 회사의 직원구조, 인사 정책 등에 관련된 내용을 포함
5) 재현자료 작성 방법론 검토(박민정, 김정연, 2017, 통계개발원)
6) 다중대체 방법을 적용한 재현데이터 생성 사례(PPT 자료, 2019.11, KCB)

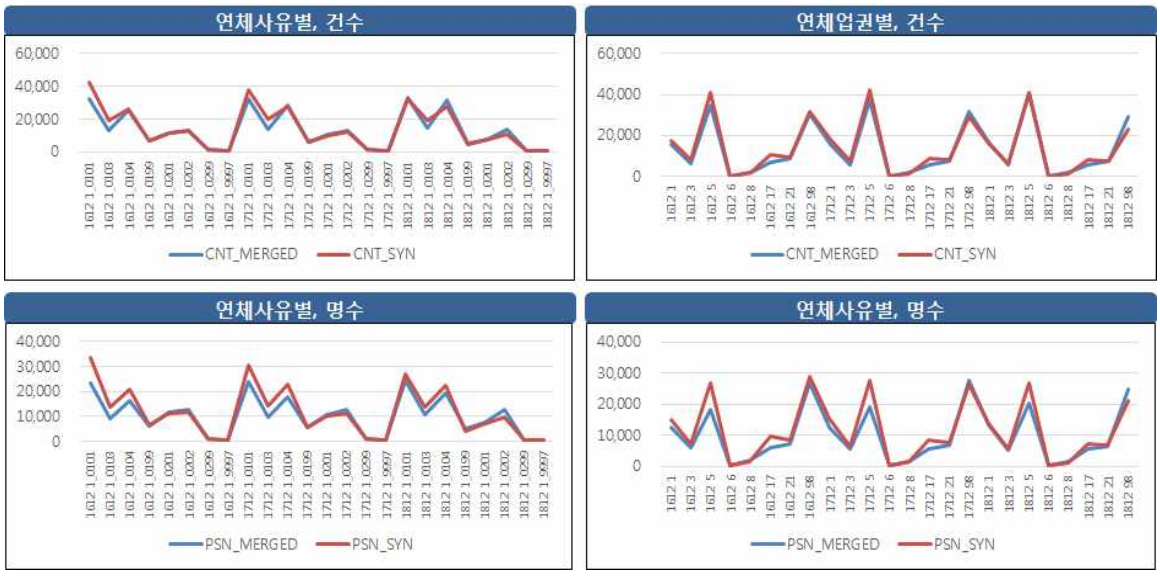
○ 차주정보, 대출정보, 연체정보 등의 전체 내용 중 특정 시점의 데이터를 대출 및 연체의 등록연월과 종료연월로 압축한 뒤, 각각의 테이블을 연결하여 모집단의 특정 패턴을 따르는 통계 모형을 도출하고 재현데이터를 생성 및 검증하는 절차를 수행

※ 개인신용정보 재현데이터 생성 현황
- ○ (개발목적) CreDB 서비스 이용자에 대한 분석 및 교육 지원, 신용정보 교육 등
- ○ (개발기준) 개인신용정보 표본DB의 대출, 연체 및 카드개설 정보 등 4개 테이블 26개 정보항목
- ○ (개발규모) 약 180만 명에 해당하는 가상 차주에 대한 25개월('16.12월말 ~ '18.12월말) 분의 대출, 연체 및 카드개설 정보

[개인신용정보 재현데이터 생성 과정]



[원본데이터와 재현데이터간 통계적 특성 비교]



V. 결론

▶ **개인정보가 포함된 민감한 데이터를 안전하게 활용할 수 있는 재현데이터의 개념 및 특성과 국내외 활용 사례를 살펴본 결과 민감한 데이터의 분석, 공개 및 교육 등의 분야에서 활용되는 사례가 다수 존재**

- 재현데이터는 정보주체가 제공한 실제데이터가 아닌 가상으로 만들어진 데이터로 살아있는 개인의 정보를 보호하기 위한 관련 규제로부터 자유롭게 활용 가능
- 미국, 영국, 독일 정부는 국가 차원에서 민감한 개인정보가 포함된 데이터를 재현데이터로 구현하고 이를 분석 등에 활용하는 사례가 다수 있었으며, 우리나라에서도 통계청, 공공기관 등을 중심으로 재현데이터에 대한 관심이 증가하고 있는 단계

▶ **재현데이터는 데이터를 분석하려는 연구자들에게 좋은 탐색과 학습의 수단이 될 수 있으며, 분석 결과에 대한 신뢰성 검증 프로세스까지 지원이 된다면, 보다 다양한 분석에 활용될 가능성을 내포**

- 재현데이터를 이용한 분석의 정확성 등을 논하기에는 아직 위험이 따를 수 있기 때문에, 초기에는 데이터 소개, 모델설계 실습 등의 교육용, 민감 데이터에 대한 분석 코드 사전 개발, 데이터 분석 공모전 등의 분야에서 활용될 수 있음
- 그리고 재현데이터를 이용한 분석 결과에 대한 신뢰성 검증 프로세스가 뒷받침 된다면, 민감한 데이터를 다루는 금융, 조세, 보건, 의료 등 다양한 분야에서 기존에 수행하기 어려웠던 다양한 분석을 수행해 볼 수 있을 것임

▶ **그동안 금융업권의 신용정보는 정보의 민감성으로 인해 금융권 이외의 분야에서는 접해보기 어려워 다양한 분석과 활용이 제한되었으나 재현데이터는 데이터 활용 제한의 문제를 일부 해소 가능**

- 법·제도적인 규제로 인해 그동안 공개되지 않았거나, 타 업종과 융합 분석에 제약이 많았던 민감한 금융 데이터를 재현데이터 형태로 개발하고 분석, 교육 등에 활용하는 방안을 모색해볼 필요 있음

▶ **신용정보원은 향후 안전하면서도 통계적 정확성이 높은 신용정보 재현데이터를 추가 개발하고, 금융계, 학계 등에서 활용할 수 있는 방안에 대해 지속적으로 검토할 계획**

- 신용정보원이 개발한 신용정보 재현데이터는 4개의 테이블로 분리된 시계열 형태의 대규모 원시 데이터(Raw Data)를 국내 최초로 재현하였다는 점에서 의의가 있으며, 데이터 분석 인력 양성 등을 위해 금융업권, 학계 등에서 활용될 수 있도록 추진할 예정