

발 간 등 록 번 호

11-1241140-100001-10



2025년 연구보고서

데이터 통합 방법 체계화 연구

2026. 1.



<https://mods.go.kr/dsri>



국가데이터처
국가데이터연구원

연구보고서 2025-08

데이터 통합 방법 체계화 연구

김민규·박성률



Statistics Korea
Statistics Research
Institute

발간사

“데이터의 가치는 분석과 활용을 통해 의사결정을 지원하고, 혁신과 효율성 향상 등 구체적인 성과를 창출하는 데서 비롯됩니다.”

급변하는 불확실성의 시대에 데이터는 더 이상 단순한 숫자의 기록이 아니라, 미래를 예측하고 사회 문제를 해결하는 핵심 나침반으로 자리매김하고 있습니다. 국가데이터연구원은 이러한 시대적 요구에 부응하여 국민의 삶을 실질적으로 개선하고 AI 기반의 공공 AX 대전환을 뒷받침하기 위한 데이터 기반 연구에 지속적으로 매진해 왔습니다.

2025년 연구보고서에는 우리 사회가 직면한 환경 변화에 능동적으로 대응하고자 첨단 기술을 국가통계에 접목하기 위해 치열하게 고민한 연구 성과를 담았습니다.

첫째, 인공지능(AI) 기반 국가통계 기술혁신을 선도하고자 노력하였습니다.

생성형 AI 기술을 현장조사에 적용하기 위한 기초연구를 통해 조사자료의 내용검토 및 자동분류, 질의응답에 활용 가능성을 모색하였으며, 이는 통계 생산의 신속성과 정확성을 획기적으로 제고하는 토대가 될 것입니다. 아울러 생성형 AI를 활용한 나우캐스트 지표 서비스 제공 방안 연구는 통계서비스의 새로운 가능성을 여는 의미 있는 첫걸음이라 할 수 있습니다.

둘째, 점차 열악해지고 있는 조사환경에 대응하기 위해 새로운 통계방법론 연구와 국가통계 품질제고를 위한 연구를 강화하였습니다.

확률표본과 자원자표본을 통합한 추정 방안 연구는 응답자 조사 부담을 완화하고 비확률표본의 병행 활용 가능성을 제시하였으며, 데이터 과학기술을 활용한 자료수집 개선 연구와 데이터 통합방법 연구는 다양한 데이터의 연계·통합 방법을 보다 체계화하였습니다.

셋째, 사회적 사각지대를 조명하고 지속가능한 미래를 지원하기 위한 데이터 기반 정책 연구에 집중하였습니다.

최근 심각한 사회 문제로 대두된 ‘고립·은둔 청년’의 실태 파악을 위한 조사 문항 개발 연구를 비롯하여, 돌봄 분야 국가통계 활용 방안과 국내 최초의 기후변화 통계·지표 분석 연구는 데이터가 사회안전망 강화에 기여할 수 있음을 보여줍니다. 또한 소득이동통계 심층 분석 연구와 생애과정 이행에 대한 중·고령기 비교 연구는 관련 정책의 실효성과 활용도를 한층 높일 것으로 기대됩니다.

아울러 가계동향조사의 소비지표 작성 연구와 퇴직연금 적립금 배분 방법 연구는 국민의 체감 경기를 보다 정확히 진단하고 합리적인 경제정책 수립을 지원하는 든든한 기반이 될 것입니다.

2025년 10월부터 새롭게 출발한 국가데이터처 국가데이터연구원은 앞으로도 최신 기술과 사람을 잇는 데이터 연구를 통해 국가통계의 지평을 지속적으로 확장해 나가겠습니다.

본 연구보고서가 통계 생산자와 이용자 모두에게 실질적인 도움이 되고, 각계각층의 의사결정자에게 깊이 있는 통찰을 제공하기를 기대합니다.

많은 관심과 성원을 부탁드립니다.

2026년 1월

국가데이터연구원장

가진

목 차

제1장 서론	1
제1절 연구 배경 및 목적	1
제2절 연구 내용 및 방법	5
제2장 데이터 통합의 개념 및 유형	7
제1절 데이터 통합의 개념	7
제2절 데이터 통합의 유형	11
제3장 데이터 통합 과정 및 방법	18
제1절 전처리 및 정합성 점검	20
제2절 결합	29
제3절 결측값 대체	55
제4절 보정 및 대표성 점검	68
제5절 최종 품질 점검	84
제4장 통계생산 기관의 데이터 통합 사례	86
제1절 국제기구별 사례	86
제2절 국가별 사례	97
제5장 결론 및 제언	111
제1절 연구 요약·시사점 및 한계	111
제2절 데이터 통합 활성화를 위한 제언	116
참고문헌	118
부 록	126
Abstract	137

요 약

디지털 전환의 가속화로 공공 및 민간 부문에서 방대한 데이터가 축적되고 있으며, 동시에 통계 이용자들은 보다 시의성 높고 세분화된 고해상도 통계를 요구하고 있다. 그러나 단일 자료만으로는 이러한 수요를 충분히 충족하기 어렵기 때문에, 다양한 출처의 자료를 결합하여 활용도를 높이는 데이터 통합의 중요성이 점차 커지고 있다.

본 연구는 이러한 문제의식을 바탕으로 국가통계 품질과 신뢰성을 제고할 수 있는 데이터 통합 방법을 체계화하는 것을 목적으로 한다. 연구의 주요 내용은 다음 세 가지로 요약된다. 첫째, 데이터 통합의 개념을 정의하고, 그 범위와 특징을 명확히 규정하였다. 둘째, 데이터 통합의 유형을 목적·단위·자료 구조·주체·방법별로 체계화하였다. 셋째, 데이터 통합 과정을 「전처리 - 정합성 점검 - 결합 - 결측값 대체 - 보정 및 대표성 점검 - 품질 점검」의 단계로 구조화하여 각 단계가 상호 유기적으로 작동할 때 데이터 통합이 효과적으로 구현될 수 있음을 제시하였다. 이러한 결과를 통해 데이터 통합은 단순한 기법적 절차가 아니라 통계생산의 전(全) 주기에 걸친 관리체계이자 국가통계의 전략적 인프라임을 강조하였다.

아울러 본 연구는 국내 데이터 통합 활성화를 위해 ① 중앙 관리체계 확립, ② 데이터 거버넌스 활성화, ③ 반정형·비정형 자료의 통합 연구, ④ 실험적 데이터 통합 프로젝트 확대, ⑤ 메타데이터 작성 체계 확립을 제안하였다. 본 연구는 데이터 통합의 개념 정의와 유형 분류, 절차 및 방법 체계화를 통해 국가통계의 신뢰성과 활용도를 제고할 수 있는 기반을 제시하였다. 이는 데이터 통합이 변화하는 데이터 환경 속에서 국가통계가 직면한 과제를 해결하고, 지속 가능한 통계생산 체계로 발전하기 위한 핵심 인프라임을 시사한다.

주요 용어 : 데이터 통합, 방법 정립, 고해상도 통계, 국가통계, 인프라

제 1 장

서 론

제1절 연구 배경 및 목적

1. 연구 배경

디지털 전환의 가속화로 데이터 환경은 사회 전반에서 급격히 변모하고 있다. 인공지능(AI), 클라우드 컴퓨팅, 사물인터넷(IoT), 모바일 기기 등 정보통신 기술의 발전은 민간과 공공 부문에서 방대한 데이터를 폭발적으로 생성하고 있으며, 이는 고품질 정책통계 생산과 증거 기반 정책 결정의 중요 자산이다. 따라서, 이를 효율적으로 통합·활용하는 능력이 국가통계의 경쟁력을 좌우한다(UNECE, 2020; ESCAP, 2020).

데이터 환경 변화에 따라 이용자들 역시 더욱 시의성이 높고, 세분된 통계를 요구하고 있다. 그러나, 기존의 조사 중심 통계체계는 조사비용 증가, 응답률 저하, 시간적 제약 등 구조적 한계를 지니고 있으며, 특히 실시간 정책 대응이나 소지역 분석과 같은 고해상도 통계(high-resolution statistical data)¹⁾ 수요에 대응하기 어렵다. 이러한 배경에서 다양한 출처의 데이터를 결합하여 정합성과 활용성을 높이는 **데이터 통합(data integration)**의 필요성이 증대되고 있다.

데이터 통합은 단순한 자료 병합(merging)을 넘어, 이질적 자료 간 개념 정렬, 정합성 확보, 오류 검토, 품질 점검 등을 포함하는 복합적인 과정이다. 이러한 과정을 통해 조사 결과를 보완하고, 결측 정보를 대체하여 새로운 통계를 생산할 수 있다. 즉, 데이터 통합은 정밀하고 신뢰성 있는 통계생산을 가능하게 하여, 통계생산 기관이 **정보 통합자(information integrator)** 역할을 하도록 한다. 이는 **국가데이터처**가 기존 통계생산 기관을 넘어, 다양한 데이터를 조정·통합·활용하는 **데이터 허브(data hub)**로 발전해야 함을 시사한다.

1) **고해상도 통계**는 단위가 세분된 통계 데이터로 미시 수준에서 경제·사회적 변화를 정밀하게 파악하고, 정책 의사결정을 지원하는 통계이다. 가령, 읍면동 단위의 일별 의료 이용 통계처럼 공간·시간적으로 정밀한 단위에서 생산된 통계를 의미한다(<부록> 1. 참조).

국제기구들은 데이터 통합의 개념 정의, 유형 구분, 품질 확보 기준 등을 제시하고 있으며(ESCAP, 2020; UNECE, 2017, 2020), 주요 국가의 통계기관들 역시 이를 전략적으로 추진하기 위해 관련 제도와 인프라를 갖춘 통합 시스템을 운영하고 있다(Benzeval et al., 2020; CIID, 2021). UNSD(n.d.)는 데이터 통합을 “공식 통계생산을 위한 기본 절차”로 간주하고 있으며, UN ESCAP(2020)와 UNECE(2017)는 데이터 통합을 GSBPM(Generic Statistical Business Process Model)의 모든 단계에서 적용할 수 있는 전략적 기능으로 규정하였다. 캐나다(Frenette et al., 2025), 뉴질랜드(Statistics New Zealand, n.d.-a), 호주(ABS, n.d.-a) 등도 통합 기반 통계생산을 위한 제도적·기술적 기반을 정비하고 있다. 이는 국내에서도 데이터 통합을 제도적으로 뒷받침할 체계가 필요함을 의미한다.

국내에서는 국가데이터처를 중심으로 조사자료, 행정자료, 빅데이터를 활용한 데이터 통합 기반 통계생산이 점차 확대되고 있다. 특히, 통합 데이터 세트는 응답자 부담을 완화하고, 응답률 하락을 보완하며, 특정 영역의 정보 결손을 해소할 수 있는 효과적인 대안으로 주목받고 있다. 다만, 현재 데이터 통합 방법은 제도적 기반이 아직 충분히 정비되어 있지 않으며, 기술적·개념적·품질 관리 측면에서도 체계적이고 표준화된 지침 마련이 필요한 단계이다.

2. 연구 필요성

데이터 환경 변화에 따라 데이터 생성량은 기하급수적으로 증가하고 있으며, 정부, 공공기관, 민간 부문에서 생산되는 행정자료, 조사자료, 빅데이터 등은 각각 독립된 구조와 목적에 따라 수집되어 이질성과 복잡성을 지니고 있다. 이러한 자료(들)를 통합하여 정책 수립, 사회 현상 분석, 통계생산 등에 활용하려는 요구가 높아지고 있으나, 실제 통합 과정에서는 다양한 제약이 존재한다.

첫째, 자료 간 불일치 문제가 있다. 자료의 구조 및 기준 시점 차이, 개념 및 분류 불일치 등은 상호운용성(interoperability)³⁾을 저해한다. 예를 들면, 같은 변수라도 수집 주체, 목적, 조사 시기, 측정 단위가 달라 동일 개념을 다르게 해석하여 데이터 통합 시

2) **자료(data source)**와 **데이터(data)**는 혼용되는 개념으로 이를 명확히 구분할 필요가 있다. 자료는 현실 세계로부터 수집된 가공되지 않은 1차 정보를 의미한다. 반면, 데이터는 자료를 정의·분류·정제·구조화하여 분석 및 통계 산출이 가능하도록 가공한 정보를 의미한다(<부록> 1. 참조).

3) **상호운용성(interoperability)**은 서로 다른 시스템, 기관, 국가 간에 데이터가 원활하게 공유되고 해석될 수 있도록 하는 **구조적 호환성**을 의미하며, 메타데이터 표준화, 코드 체계 일치, 분류 기준 통일 등 데이터 구조 전반의 일관성을 요구한다(<부록> 1. 참조).

왜곡이 발생할 수 있다(Harron, 2016). 이는 통합 과정에서 빈번하게 발생하는 문제로 국제기구(ESCAP, UNECE 등)가 제시하는 표준화된 지침 도입의 필요성을 보여준다. 이러한 문제의식은 본 연구가 데이터 통합 방법을 체계화하고 품질 관리 지침을 마련해야 하는 근거가 된다.

둘째, 고유식별자 부재 및 오류 문제다. 주민등록번호, 사업자등록번호 같은 고유식별자(unique identifiers)가 있다면 자료를 쉽게 결합할 수 있지만, 실제로는 부재하거나 오류가 포함된 경우가 많다(Winkler, 2006). 특히, 반정형(semi-structured)·비정형(unstructured) 자료는 구조화된 식별정보가 없거나 부분적으로만 존재하며, 오류가 혼재되어 있다. 따라서, 이러한 자료를 구조화된 데이터와 결합하기 위해서는 식별자 대체 알고리즘이나 유사도 기반 결합과 같은 고도화된 기법이 필요하다. 이는 향후 빅데이터 기반 데이터 통합체계 구축을 위한 선결 조건이다. 이와 관련, 국제적으로는 개인정보 대체키(pseudonymization key)와 기계학습 기반 연계(machine learning-based linkage)를 활용하는 사례가 늘고 있어, 국내에서도 이러한 기술적·제도적 보완이 요구된다.

셋째, 데이터 통합 과정 전반에 걸친 표준화된 지침이 부족하다. 자료 전처리(pre-process), 결합(matching) 방법, 품질 점검(quality assessment) 등 각 단계에 대한 표준 작업절차(standard operating procedures, SOP)가 마련되지 않아 통합 데이터 세트의 신뢰성이 저하될 수 있다. 현재까지 국내 데이터 통합체계에서는 메타데이터 기반 SOP가 체계적으로 정립되어 있지 않아, 데이터 통합 결과의 일관성과 재현성 확보에 한계가 있다.

넷째, 법·제도적 제약과 기관 간 협업 부족은 데이터 통합의 장애요인으로 작용한다. 개인정보 보호법과 정보보호 관련 법령 그리고 부처 간 자료 공유에 대한 견해 차이는 데이터 통합 추진 과정에서 구조적 제약 요인으로 작용하고 있다. 데이터 통합에는 원자료 수준에서의 접근과 결합이 필요하지만, 자료 보유 기관은 행정적 부담 및 법적 위협으로 인해 이를 제한하거나 거부하는 경우가 많다(Montgomery County Department of Health and Human Service, 2022). 이와 관련, 국내에서는 제도적·기술적 기반이 아직 충분히 마련되지 않았으므로, 국가데이터처를 중심으로 한 데이터 통합 전담 기구의 설립을 검토하고 법·제도 개선과 협업 체계 구축을 병행해 추진할 필요가 있다.

이러한 한계는 기술적·방법적 문제를 넘어 제도·인식·운영 방식 전반의 개선이 요구되는 복합적 과제이다. 따라서, 데이터 통합은 단순한 기법의 나열이 아니라 이론적 근거와 실제 적용을 아우르는 종합적 접근이 필요하다.

3. 연구 목적

본 연구의 목적은 변화하는 데이터 환경에 대응하고, 통계생산 기능을 고도화하기 위해 데이터 통합 방법을 체계적으로 정립하는 데 있다. 나아가 국가데이터처가 데이터 기반 정책을 추진하고, 고해상도 통계를 생산할 수 있도록 데이터 통합 전략 수립과 제도적 기반 마련에 이바지하고자 한다. 이를 위해 본 연구는 네 가지 구체적인 목적을 설정하였다.

첫째, 데이터 통합의 개념을 정립한다. 데이터 통합의 개념을 명확히 정의하고, 유형을 분류하여, 통합 과정 및 방법의 이론적 토대를 마련한다.

둘째, 데이터 통합 과정과 방법을 체계화한다. 자료 수집부터 품질 점검에 이르는 절차를 설계하고, 단계별 방법을 제시한다. 이를 통해 데이터 통합을 전(全) 주기적 관리체계로 확립한다.

셋째, 데이터 통합의 품질 점검 체계를 제시한다. 단계별 품질 점검과 최종 산출물 품질 점검을 이원화하여, 오류 전파를 차단하고 활용 적합성을 확보하는 구조를 마련한다. 이를 바탕으로 결합 정확도 제고, 정합성 강화, 메타데이터 문서화 등을 통해 품질 관리 방안을 제시한다.

넷째, 국내외 통계기관의 데이터 통합 사례를 분석한다. 본 연구의 데이터 통합 과정을 준거로 주요 국제기구와 국가의 사례를 검토하고, 전처리 원칙과 품질 점검 기준을 적용하여 사례별 구현 수준을 고찰한다. 이를 통해 그 성과와 한계를 파악, 국내 통계환경에 적합한 데이터 통합체계 구축을 위한 시사점을 도출한다.

제2절 연구 내용 및 방법

본 연구는 이론 분석(conceptual analysis)을 통해 데이터 통합의 개념·유형·방법적 틀을 체계화하고, 다중 사례연구(multiple case study design⁴⁾)를 바탕으로 실제 적용과 한계를 비교·분석하였다. 이를 토대로 연구 범위와 내용을 다섯 가지 영역으로 구체화하였다.

첫째, 데이터 통합의 개념 정립

데이터 통합은 국내외에서 다양한 방식으로 정의되고 활용되고 있으나, 그 개념의 경계가 불명확하고 용어 혼용도 빈번하다. 본 연구는 국제기구 및 주요 국가 통계기관의 표준 문헌을 검토하여 데이터 통합의 정의와 범위를 정리하고, 공통 요소를 도출함으로써 그 개념을 체계적으로 정립하였다.

둘째, 데이터 통합의 유형 분류

데이터 통합은 활용 목적, 자료 구조, 적용 방법 등에 따라 다양한 유형으로 구분된다. 이러한 분류는 통계생산 기관이 목적에 부합하는 통합 전략을 수립하고, 자료의 품질과 활용 가능성을 사전에 점검하여, 통합의 필요성과 타당성을 검토하는 기준이 된다. 국제기구 및 주요 국가 통계기관 역시 정책적 활용성, 제도적 정비 수준 등을 고려하여 다양한 유형을 제시하였다. 본 연구는 이를 토대로 다섯 가지 기준에 따라 데이터 통합 유형을 분류하였다.

- **통합 목적:** 활용 목적 통합, 기반 마련 통합
- **통합 단위:** 미시 통합(개체 단위), 거시 통합(집계 단위)
- **자료 구조:** 정형, 반정형, 비정형 자료
- **통합 주제:** 내부 통합, 기관 간 통합, 민관 통합
- **통합 방법:** 레코드 연계, 통계적 매칭 등

셋째, 데이터 통합의 과정 및 방법 체계화

데이터 통합은 단순 절차나 단일 기법에 국한된 것이 아니라, 여러 단계가 유기적으로 구성된 체계적 과정(structured multi-stage process)이다. 각 단계는 고유의 기술적·방법적 과제를 수반하며, 체계적인 절차 설계를 통해 통합 데이터의 품질과

4) 다중 사례연구(multiple case study design)는 복수의 사례를 선택하여 공통 패턴을 분석하거나 이론을 반복적으로 검증하는 방법으로, 단일 사례연구보다 일반화 가능성과 설명력을 높일 수 있는 연구 설계이다(Yin, 2018).

활용성을 보장할 수 있다. 본 연구는 데이터 통합을 다섯 단계로 구조화하고, 단계별 적용 가능한 방법을 정리하였다.

- 1단계: 전처리 및 정합성 점검
- 2단계: 결합
- 3단계: 결측값 대체
- 4단계: 보정 및 대표성 점검
- 5단계: 최종 품질 점검

넷째, 데이터 통합 사례 분석

본 연구는 다중 사례연구 설계에 기반하여 주요 국제기구·국가의 데이터 통합 사례를 분석하였다. 특히, 데이터 통합 과정(자료 수집-전처리-결합-결측값 대체-보정-품질 점검)을 준거 틀로 삼아 각 사례를 검토하였다. 이를 통해 제도적 기반, 기술적 적용 수준, 품질 관리 방식 등을 파악하고, 국내 통계환경에 적합한 데이터 통합 전략을 모색할 수 있는 시사점을 도출하였다.

다섯째, 데이터 통합의 향후 발전 방향

본 연구는 데이터 통합의 개념 정립, 유형 분류, 과정 및 방법 체계화, 사례 분석을 종합하고, 이를 토대로 데이터 통합의 실효성을 높이기 위한 향후 발전 방향을 제시하였다.

- 중앙 관리 체계 확립
- 데이터 거버넌스 활성화
- 반정형·비정형 자료 통합 연구
- 실험적 데이터 통합 프로젝트 확대
- 메타데이터 작성 체계 확립

제 2 장

데이터 통합의 개념 및 유형

제1절 데이터 통합의 개념

1. 데이터 통합의 정의

데이터 통합(data integration)은 서로 다른 출처에서 수집한 자료⁵⁾를 결합하여 분석 가능한 구조로 만드는 일련의 절차를 말한다(ESCAP, 2020; UNECE, 2020). 이는 통계생산의 효율성과 품질을 동시에 확보하는 방법으로, 단일 자료만으로는 충족하기 어려운 다양성·포괄성·시의성·정밀성을 확보하는 데 활용된다.

국제기구들은 각기 다른 관점에서 데이터 통합을 정의하고 있지만, 공통으로 다출처(multiple data source) 활용, 구조적 정합성(structural consistency), 분석 가능성 확대(enhanced analytical potential)를 핵심 요소로 강조한다.

- UNSD(n.d.)는 데이터 통합을 “여러 출처의 자료를 결합하여 새로운 통계 산출물(statistical outputs)을 생성하는 공식 통계 시스템의 핵심 인프라”로 규정한다.
- UN ESCAP(2020)은 “정책 수요 충족을 위해 다양한 출처의 자료를 정렬(alignment)하고, 연계(linking)하여 분석할 수 있는 형태로 만드는 과정”으로 설명하면서, 메타데이터 관리와 표준화된 절차의 필요성을 강조한다.
- UNECE(2020)는 데이터 통합을 “동일 개체(entity) 또는 개념(concept)에 대한 정보를 여러 자료로부터 조합하여 통계생산의 품질과 효율을 향상하는 절차”로 정의하며, 정합성(consistency), 완전성(completeness), 시의성(timeliness)을 핵심 속성으로 제시한다.

이처럼 국제기구들은 데이터 통합을 분석 가능성과 활용성을 증대하기 위한 전략적 결합으로 이해하고 있다.

5) 초기의 데이터 통합은 동일 구조를 가진 자료 간 병합(merging) 중심이었다. 그러나 현재는 조사자료, 행정자료, 지리정보(geospatial information), 빅데이터 등 서로 다른 성격의 자료들을 결합하는 방식으로 진화하고 있다(ABS, n.d.-b).

국가 통계기관들은 데이터 통합을 정의하면서 데이터 품질 확보, 정책 활용 가능성, 행정자료의 통합과 안전한 데이터 환경 조성을 강조하고 있다.

- 뉴질랜드 통계청(Statistical New Zealand, 2013, 2022)은 데이터 통합을 “서로 다른 출처의 자료를 효과적으로 연결하여 새로운 통찰(Insight)과 정책 활용 가능성(policy utility)을 창출하는 수단”으로 정의한다. 이 정의에는 실제 통합을 위한 데이터 품질 진단과 관리체계를 포함하고 있다.
- 호주 통계청(ABS, n.d.-a)은 “다양한 출처의 자료를 결합해 신뢰성 있는 통계를 제공하는 핵심 수단”으로 설명하며, 데이터 통합을 통계 품질 제고와 정책 대응력 강화를 위한 기반으로 본다. 한편, 호주 보건 복지 정보원(Australian Institute of Health and Welfare)은 데이터 통합에 대해 명시적 정의를 내리지 않았지만, 여러 데이터 세트를 주제 중심으로 통합하는 작업을 수행하면서 데이터 통합 절차와 품질 관리를 강조하고 있다(AIHW, 2022).
- 캐나다 통계청은 데이터 통합을 “행정자료와 조사자료 등을 마이크로데이터 수준에서 결합해 포괄적 정보를 생산하는 절차”로 정의하고, 정밀한 정책 분석과 활용성 확대를 위한 전략적 수단으로 인식한다(Frenette et al., 2025; Trant & Whitridge, 1999).

국제기구와 국가 통계기관이 제시한 정의를 종합하면, 다섯 가지 공통 요소가 도출된다.

- 서로 다른 출처의 자료를 목적에 맞게 결합
- 개념 정합성과 방법적 표준화 확보
- 정확성(accuracy), 정합성(consistency), 시의성(timeliness) 등의 품질 충족
- 통계생산, 정책 분석 등 명확한 목적에 기반한 설계
- 거버넌스와 재현 가능한 절차를 갖춘 안정적 운영 체계

이상의 논의를 바탕으로 데이터 통합을 다음과 같이 정의하고자 한다.

데이터 통합은 서로 다른 출처의 자료를 개념 및 구조적으로 표준화하고, 통합 목적에 맞게 결합하여 검증된 품질의 통계 산출물을 생산하는 체계적인 절차이다.

이러한 개념 정의는 통계생산 환경의 복잡성 증가와 정책 수요의 고도화에 효과적으로 대응하기 위한 전략적 틀로 기능할 수 있다.

요컨대, 데이터 통합은 신뢰성 있는 공식 통계를 생산하고 정책 수요에 정밀히 대응할 수 있는 전략이다. 따라서 국가데이터처를 비롯한 관련 기관은 데이터 통합을

기술이 아닌 체계(structure)로 이해하고, 법적 기반, 기술적 역량, 조직 운영 체계를 유기적으로 결합하여 데이터 통합의 품질과 신뢰성을 확보해 나가야 한다.

2. 데이터 통합의 역할

데이터 통합은 기존 조사 기반 통계가 지닌 한계를 보완하고, 변화하는 정책 수요와 복잡한 데이터 환경에 대응하기 위한 핵심 전략이다. 행정자료, 조사자료, 지리정보, 빅데이터 등 다양한 자료의 활용이 가능해졌지만, 단일 자료만으로는 증가하는 통계 수요를 충분히 충족하기 어렵다. 특히, 응답률 저하, 조사 비용 증가, 실시간 통계 수요 확대 등 구조적 변화는 기존 통계생산 방식의 한계를 드러내며 데이터 통합의 필요성과 중요성을 점차 부각시키고 있다. 또한, 통계생산 기관의 역할이 자료 수집자에서 정보 통합자로 확장됨에 따라, 데이터 통합은 통계 품질 제고와 정책 활용성 강화를 위한 핵심적 역할을 담당하게 되었다.

첫째, 통계 품질 향상

데이터 통합은 통계자료의 정밀도와 신뢰성을 제고하는 직접적 수단이다. 행정자료, 조사자료, 빅데이터를 상호보완적으로 활용함으로써 단일 자료에서 발생할 수 있는 결측, 측정오차, 응답 편향을 효과적으로 완화할 수 있다(Benzeval et al., 2020; Kim & Tam, 2020). 예를 들어, 건강 관련 조사에 나타나는 자기 보고(self-reported)의 한계를 건강보험 청구자료와 결합하면 변수의 유효성과 포괄범위(coverage)가 확대된다. 나아가 통합된 데이터 세트는 교차 검증(cross-validation)을 통해 신뢰도를 검토할 수 있으며, 정합성 검증을 통해 해석 가능성(interpretability)과 일관성(coherence)을 확보할 수 있다.

둘째, 통계생산 효율성 제고

데이터 통합은 조사 부담과 비용을 낮출 수 있는 유효한 방안이다. 가령, 외부 자료를 활용한 항목 대체(item substitution)나 사전 입력(pre-filled) 방식을 적용하면 기존 조사 범위를 줄이거나 반복 조사를 대체할 수 있다. 이를 통해 응답자 부담이 완화되고 중복조사가 방지되어, 결과적으로 자원의 효율적 활용이 가능해진다(UNECE, 2019a). 이처럼 데이터 통합은 낮은 응답률과 조사 참여 회피 현상이 확산하는 현실에서 통계 품질을 유지하면서도 운영 효율성을 확보하는 중요한 전략이 될 수 있다.

셋째, 고해상도 통계생산 및 데이터 기반 환류 체계 정착

통합 데이터 세트를 활용하면 소지역·소집단·정책 대상자 단위의 고해상도 통계를

생산할 수 있게 된다. 이는 정책 수요자 맞춤형 통계의 생산과 제공을 촉진하고, 복지·노동·교육·환경 등 다분야 간 연계를 강화하며, 정책 효과성과 형평성 평가를 지원한다. 궁극적으로는 정책의 「설계-집행-평가」 전 주기에 걸친 데이터 기반 환류 체계를 정착시키는 기반이 된다.

넷째, 비정형 자료의 통계적 활용 기반 마련

모바일 신호, 위성 이미지, 기상·대기·교통 센서 자료는 시의성과 범위 면에서 강점을 지니지만, 구조적 이질성과 불완전성으로 인해 단독 활용과 해석에 제약이 따른다. 그러나 이러한 비정형 자료를 조사자료나 행정자료와 통합하면 기존 통계의 시의성을 보완할 수 있으며, 새로운 보조지표의 생산과 해석 가능성도 확보된다.

앞서 제시한 역할을 효과적으로 수행하기 위해서는 데이터 통합이 체계적인 통계생산 체계로 정착되어야 한다. 이를 위해 다음과 조건들이 충족되어야 한다.

첫째, 자료 간 형식·개념의 표준화 및 메타데이터 구축이 필요하다. 변수 정의, 분류체계, 관측 단위 등의 공통 기준을 마련하여 자료 간 차이를 해소하고, 이를 메타데이터로 작성하는 작업은 데이터 통합의 첫 단계이다.

둘째, 통합 가능성을 고려한 자료 품질의 사전 평가가 요구된다. 자료의 완결성, 시의성, 대표성, 정확성 등 핵심 품질 요소를 사전에 진단하고, 이를 바탕으로 통합의 적절성과 한계를 검토해야 한다.

셋째, 개인정보 보호 체계의 확보가 전제되어야 한다. 고유식별자(unique identifiers)를 사용할 경우, 법적 근거를 명확히 하고, 가명화·익명화와 같은 개인정보 보호 기술을 적용함으로써 데이터 보호와 활용 간 균형을 유지해야 한다.

넷째, 통합 목적에 부합하는 설계 및 운영 체계 구축이 필요하다. 이는 단순 기술 구현 차원을 넘어 정책 목적과 통계 수요에 적합한 설계와 운영 전략을 수립하는 것을 의미한다.

다섯째, 표준화된 통합 절차 마련이 요구된다. 자료 전처리, 결합 방법, 품질 검토 및 평가 등 통합 전반에 걸쳐 일관된 표준 작업절차(SOP)를 확립하여야 안정적이고 신뢰성 있는 데이터 통합이 가능해진다.

이와 같은 조건이 충족될 때, 데이터 통합은 국민의 통계 수요에 부응하고, 증거 기반 정책 결정을 지원하는 핵심 인프라로 기능할 수 있다.

제2절 데이터 통합의 유형

앞 절에서는 데이터 통합의 역할을 통계 품질 향상, 통계생산 효율성 제고, 고해상도 통계생산 등 기능적 측면에서 살펴보았다. 본 절에서는 데이터 통합 유형을 ①통합 목적, ②통합 단위, ③자료 구조, ④통합 주체, ⑤통합 방법으로 구분한다. 이러한 유형 구분은 통계생산 기관이 통합 전략을 수립하고, 자료 품질과 활용 가능성을 사전에 검토하며, 데이터 통합의 필요성과 타당성을 판단하는 중요한 준거 틀로 기능한다. 국제기구와 국외 통계기관 역시 기술적 결합뿐만 아니라, 정책적 활용성, 제도적 정비 수준 등을 종합적으로 고려하여 다양한 분류체계를 제시하고 있다 (ESCAP, 2020; Statistics New Zealand, 2013; UNECE, 2019a, 2020).

1. 통합 목적에 따른 유형

데이터 통합의 목적이 통계생산 및 정책 분석에 있는지, 통합 가능성 검토와 인프라 구축에 있는지에 따라 활용 목적 통합과 기반 마련 통합으로 구분한다. 목적에 따른 구분은 데이터 통합의 실행 여부뿐만 아니라, 지향점과 활용 방향을 명확히 함으로써 통계생산 기관의 데이터 통합 계획 수립과 자원 배분의 기준으로 기능한다.

첫째, 활용 목적 통합은 정책 수요와 통계 활용 요구에 대응하여 추진하는 통합 유형으로 분석 가능성과 정책 연계성을 동시에 충족하는 것이 목적이다. 개별 자료의 한계를 보완하기 위해 서로 다른 출처의 자료를 결합하여 정밀하고 포괄적인 통계를 생산하는 데 중점을 둔다. 이를 통해 소지역·소집단 등 미시 단위의 분석이 가능해지고, 시의성 높은 통계생산에도 기여한다.

- **공식 통계생산을 위한 통합**은 단일 자료의 한계를 극복하기 위해 다양한 자료를 결합하여 정확성과 시의성 높은 통계를 산출하는 것이 목적이다. 호주 통계청은 노동 계정을 작성하면서 사회보장, 고용, 세무, 기업 등록자료를 결합하여, 노동시장 지표 간 정합성을 확보하고 정책 대응 속도를 높였다(ABS, 2020).
- **연구 및 분석을 위한 통합**은 정책 및 학술연구를 위해 통합 마이크로데이터를 구축하거나 다양한 분석 모형을 개발·적용하는 것이 목적이다. 호주 MADIP (Multi-Agency Data Integration Project)는 인구총조사(census), 사회보장, 세무, 의료·보건 서비스, 교육 등 행정자료를 결합하여 정책 대상자별 분석이 가능한 통합 마이크로데이터를 구축하였다(ABS, n.d.-a). 또한, 일본 농림수산성은 위성 영상자료, 기상 정보, 쌀 생육 시뮬레이션 자료를 통합하여 라오스의 벼 수확량을

예측하였다(Maki et al., 2017).

둘째, 기반 마련 통합은 데이터 통합에 앞서, 체계적인 통합과 활용을 가능하게 할 제도적·기술적 기반을 구축하는 것을 목적으로 하는 통합이다. 시범 결합, 결합 가능성 검토, 자료 간 표준화, 메타데이터 체계 수립 등을 포함하는 준비 단계에서 수행하는 통합이다.

- **표본 틀 구축을 위한 통합**은 통계조사의 표본설계에 필요한 모집단을 구축하는 것이 목적이다. 아시아개발은행(ADB)은 각국의 기업등록부, 세무 자료, 고용자료 등을 결합하여 기업통계등록부(SBR)를 구축하고, 이를 활용해 기업체 표본조사에 필요한 표본 틀을 생성하였다(ADB, 2018).
- **검증 및 품질 관리를 위한 통합**은 자료 비교와 정합성 검증을 통해 사전에 품질을 점검하는 것이 목적인 통합이다. 구체적으로 ①자료 간 변수 정의와 분류체계 일치성 점검, ②고유식별자가 부재한 경우의 결합 가능성 검토, ③통합 플랫폼 구축을 위한 기술 표준 도입 등이 검증 및 품질 관리를 위한 통합에 해당한다. Eurostat(2021a)은 통계생산의 품질과 신뢰도를 제고하기 위해 개별 자료 검증, 집계 자료 검증, 결측값 대체 절차를 통해 변수 간 논리성과 정합성을 확보하고 있다.

2. 통합 단위에 따른 유형

데이터 통합에서 **단위**는 분석의 정밀도와 활용 가능성을 결정하는 핵심 요소로 통계 산출물의 품질에 직접적인 영향을 미친다. 통합의 대상이 되는 분석 단위(analysis/record unit)에 따라 크게 세 가지 유형으로 구분할 수 있다. 이러한 구분은 통합 목적, 자료 속성, 분석 가능성 등을 종합하여 설정되며, 각 통합 단위는 통계의 해석 수준과 적용 가능성에 영향을 미친다.

첫째, 미시 단위 통합(micro integration)은 개인, 가구, 기업 등 **개체 단위**에서 이루어지는 통합으로 미시적 속성에 대한 분석이 가능하여 맞춤형 정책 분석에 활용될 수 있다. 미시 단위 통합은 주로 고유식별자나 공통 변수(예: 성별, 연령, 주소 등)를 기반으로 결합이 이루어진다. 예를 들어, 국민건강보험자료와 국민건강영양조사를 개인 단위로 결합하면 의료 이용과 건강 행태 간 상관관계를 분석할 수 있으며, 고용보험과 세무 자료를 결합하면 기업별 고용·수익 구조를 동시에 파악할 수 있다. 그러나 몇 가지 한계도 존재한다. 개체 단위의 민감한 정보를 다루는 만큼 개인정보 보호에 대한 부담이 크며, 다른 출처의 자료를 결합하는 과정에서 변수 정의, 분류체계, 관측 시점 불일치 등으로 인한 정합성 문제를 해결해야 한다. 이러한 불일치가 해소되지 않을 경우, 왜곡이

발생할 수 있다. 따라서, 미시 단위 통합에서는 엄격한 품질 관리 절차가 필수이다.

둘째, 거시 단위 통합(macro integration)은 지역(예: 시군구, 읍면동 등), 산업, 국가 등 **집계 단위**를 기준으로 통합하는 방식이다. 개체 단위의 속성을 직접 확인할 수는 없지만, 경제·사회 현상의 전반적 흐름을 파악할 수 있다는 점에서 의미를 지닌다. 거시 단위 통합의 장점은 포괄성과 시의성에 있다. 광범위한 집계 단위를 활용하기 때문에 전반적인 흐름을 포괄적으로 파악할 수 있으며, 시의성 있는 지표를 신속하게 산출할 수 있다. 이러한 특성은 경제·사회 전반을 진단하고 정책 환경 변화를 적시에 반영하는 데 유리하다. 예컨대, GRDP와 산업별 생산·고용 지표를 결합하면 지역경제 구조를 종합적으로 진단할 수 있고, 무역 통계와 산업연관표를 결합하면 글로벌 가치 사슬(global value chains)의 구조적 변화를 분석할 수 있다. 그러나, 거시 단위 통합은 집계 단위 자료를 활용하기 때문에, 개체 단위의 세부 속성을 파악할 수 없으며 정책 대상자별 맞춤형 분석에는 제약이 따른다. 또한, 서로 다른 집계 자료 간 범위(coverage)와 기준 시점 불일치가 발생할 수 있어, 이를 조정하지 않으면 지표 간 비교나 통합의 정확성이 저하될 위험이 있다.

셋째, 혼합형 통합(micro-macro mixed integration)은 미시 단위 자료와 거시 단위 자료를 결합하여 다층적 분석을 가능하게 하는 통합이다. 미시 단위의 정밀성과 거시 단위의 포괄성을 동시에 추구하는 보완적 유형이다. 혼합형 통합의 특징은 상호 보완성에 있다. 개체 단위 자료를 기반으로 한 세부 분석 결과를 집계 단위 지표와 결합함으로써 통계의 정확성과 해석력을 동시에 높일 수 있다. 가령, 기업 단위의 재무·고용자료를 산업별 생산·부가가치 지표와 결합하면 개별 기업의 성과가 전체 산업 구조 속에서 어떤 의미를 갖는지 파악할 수 있다. 또한, 개인 단위의 소득·소비 지표를 GRDP와 결합하면 지역 간 분배 구조와 경제적 불평등 문제를 입체적으로 분석할 수 있다. 그러나, 서로 다른 수준의 자료를 일관되게 결합하기 위해서는 단위 간 불일치, 기준 시점 및 표본 추출 방법 차이 등을 표준화해야 한다. 이는 방법론적으로 복잡하여 불일치 조정 과정에서 새로운 오류가 발생할 가능성도 존재한다. 따라서, 혼합형 통합은 미시·거시 단위 통합에 비해 설계와 운영 단계에서 더 높은 수준의 통계적·기술적 정교함을 요구한다.

3. 자료 구조에 따른 유형

자료는 그 형식과 구조적 특성에 따라 정형, 반(半)정형, 비정형으로 구분된다. 이 구분은 데이터 통합 자체의 유형을 정의하기보다는, 통합 과정에서 전처리 방식, 결합 방법, 품질 관리 등에 영향을 미치며 통합 설계의 기초로 작용한다.

첫째, 정형 자료(structured data source)는 행과 열로 구조화된 테이블 형태로 변수 정의와 분류체계가 명확하게 규정되어 있어 분석과 결합에 유리하다. 정형 자료는 고유식별자나 공통 변수의 일치 조건 등을 활용하여 비교적 쉽게 결합할 수 있다.

둘째, 반정형 자료(semi-structured data source)는 완전한 테이블 형태는 아니지만, 일정한 구조⁶⁾를 갖춘 자료로 XML, JSON, 로그 파일 등이 이에 해당한다. 반정형 자료는 대량의 속성(attribute) 정보와 시간·위치·장치 정보가 함께 기록되어 있어 자료의 유연성과 시의성이 높다. 그러나, 데이터 통합에 활용하기 위해서는 변수 추출,⁷⁾ 단위 일치, 개념 조화 등 정형화를 위한 전처리가 필요하다. 반정형 자료는 정형 자료와 결합해 활용되는 경우가 많다.

셋째, 비정형 자료(unstructured data source)는 텍스트, 이미지, 영상, 음성 등 구조화되지 않은 자료로, 사회 현상에 대한 맥락 정보(contextual information)를 제공하고 정성적인 정보(qualitative information)를 정량적으로 분석(quantitative analysis)할 수 있는 기반을 제공한다. 다만, 비정형 자료는 통계적으로 직접 분석하거나 다른 자료와 결합하기 어렵기 때문에 데이터 통합에 활용하기 위해서는 추출, 분류, 정제 등 복잡한 전처리⁸⁾가 필요하다. 또한, 통계 품질, 반복성, 검증 가능성 측면에서 한계가 여전히 존재한다.

정형·반정형·비정형 자료 통합은 각 자료의 구조적 이질성을 고려한 정합성 확보가 필수적이다. 특히, 「정형-비정형」 또는 「반정형-비정형」 자료 통합의 경우, 스키마 정렬(schema alignment)과 개념 조화(semantic harmonization)가 선행되어야 한다. 이러한 구조적 차이를 효과적으로 극복할 수 있다면, 예측 모형구축이나 시뮬레이션 기반의 고차원 분석이 가능해져 정책 분석 및 통계생산의 정밀도를 높일 수 있다. 실제로 본 연구진은 스크래핑(scraping)⁹⁾을 통해 수집한 텍스트와 이미지 자료를

6) 구조적 태그(structural tag)와 포맷(format) 예시

<ul style="list-style-type: none"> • 태그: XML(eXtensible Markup Language) <pre> <user> <id>123</id> <name>보통씨</name> <location>대전</location> </user> </pre>	<ul style="list-style-type: none"> • 포맷: JSON(JavaScript Object Notation) <pre> { "userID": 123 "name": "보통씨", "location": "대전" } </pre>
---	---

7) 반정형 자료는 변수명이 비표준화되었거나, 동일 변수 내 여러 유형의 응답이 혼재된 경우가 많다.

8) 비정형 자료의 전처리 예시

- 자연어처리(NLP) 또는 이미지 분류 알고리즘을 통해 구조적 변수 생성
- 태깅(tagging) 및 특성 추출(feature extraction)을 통한 정형화
- 분류체계 연계 또는 해시값 처리를 통한 정합성 확보

9) 웹 페이지에 접속하여 **필요한 정보를 추출하고 수집**하는 기술을 의미한다. 크롤링(crawling)과

경제총조사의 특성 항목과 결합하는 연구를 진행하고 있으며, 이는 공식 통계의 외연 확장과 데이터 기반 정책설계의 실증 가능성을 보여주는 사례라 할 수 있다.

4. 통합 주체에 따른 유형

데이터 통합에서 ‘무엇을’ 또는 ‘어떻게’만큼 중요한 것이 “누가 통합하는가?”이다. 통합 주체에 따라 의사결정 구조, 법적 근거, 자료 접근 및 보안, 비용과 위험 분담 방식이 달라지고, 결과적으로 통합의 속도와 품질이 좌우된다.

첫째, 기관 내(intra agency) 통합은 하나의 조직 내부에서 보유·관리하는 자료를 통합하는 유형이다. 메타데이터 구축, 버전 관리, 표준화된 수집·정제·결합 절차가 통합의 성패를 좌우한다(ESCAP, 2020). 기관 내 통합은 협의·계약 비용이 적고, 자료 접근이 신속하며, 민감 자료의 안전한 결합과 품질 관리가 용이하다. 그러나, 부서 간 칸막이(silo)와 이질적 시스템(legacy)이 상호운용성을 저해할 수 있으며, 메타데이터나 표준 절차가 미흡하거나 미비할 경우 통합 데이터 세트의 품질이 낮아질 위험이 있다. 또한, 조직 내부라도 법적 근거와 윤리·투명성을 소홀히 하면 신뢰 문제가 발생한다(ESCAP, 2020).

둘째, 기관 간(inter agency) 통합은 국가통계 시스템 내 여러 기관이 보유한 자료를 통합하는 유형이다(ESCAP, 2020; UNSD, 2018). 법적 근거, 기관 간 협약, 윤리·보안 규정, 책임성 확보가 필수적이다(UNSD, 2018). 특히, 개념·분류·시점 차이, 품질 편차는 핵심 과제이며, 이를 해결하기 위해 표준 분류, 참조 데이터베이스, 메타데이터, 보안 규약이 요구된다. 통합 전 단계에서 자료 품질 개요를 작성하여 편향·결측·일관성·시의성·포괄성을 점검하고, 통합 후에는 통합 데이터 세트에 대한 품질 점검 및 보고를 의무화해야 한다(ESCAP, 2020). 기관 간 통합은 응답 부담 경감, 비용 효율, 소지역·소집단 통계의 시의성·포괄성 개선에 기여할 수 있으나, 기관 간 협의 비용과 이해 충돌, 자료 불일치, 개인정보 보호 및 자료 독점 문제를 동반한다. 따라서, 사전 법제 정비와 역할·책임 명시가 선행되어야 한다.

셋째, 민관(public-private, B2G) 통합은 민간이 보유한 자료(예: 통신·카드·플랫폼·소매·IoT 등)를 공공 자료와 결합하거나 공동 분석해 공익 목적의 통계를 생산하는 유형이다(European Commission, 2020). 민간 자료는 통계 목적과 다르게 생성되었기 때문에 대표성과 개념 적합성이 부족할 수 있다. 이에 사전 적합성 평가, 품질 점검, 시험적 적용(pilot test), 윤리적·법적 영향평가, 보상 규정, 기밀 보호조치가 필요하다. 민관 통합은 고빈도·고해상도·실시간 자료를 활용하여 시의성과 상세성을 개선하고,

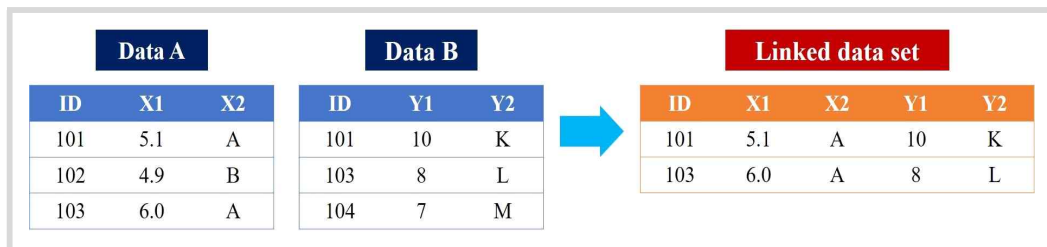
혼동되기도 하는데, 크롤링은 웹 페이지 전체를 탐색하여 정보를 추출한다.

새로운 통계생산을 가능하게 한다. 반면, 법제 미비, 계약 실패, 지식재산권 충돌, 편향·대표성 문제, 사회적 수용성 부족의 한계가 있다. 이와 관련하여, European Commission (2020)은 공익성 심사, 보상 체계, 표준계약, 중개 투명성 등 책임 있는 B2G 공유 제도화를 권고하고 있다.

5. 통합 방법에 따른 유형

데이터 통합은 데이터¹⁰⁾를 어떻게 결합·추정·보완하느냐에 따라 레코드 연계(record linkage), 통계적 매칭(statistical matching), 데이터 융합(data fusion)으로 구분된다.

첫째, 레코드 연계는 서로 다른 데이터에서 동일 개체(same entity)를 식별·연결(identifying and linking)하여 연계 데이터 세트(linked data set)를 산출하는 방법이다. 고유식별자나 공통 변수를 활용해 개체 단위에서 결합을 수행하며, 데이터 통합에서 가장 기본적이고 핵심적인 절차로 기능한다(Christen, 2012; Harron, 2016; Herzog et al., 2007; Winkler, 2006).



자료 출처: 저자 작성

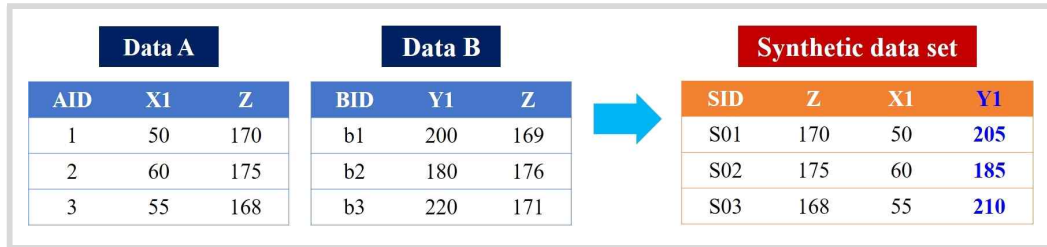
<그림 2-1> 레코드 연계 예시

<그림 2-1>은 동일 식별자(ID)를 가진 레코드를 ‘일치(match)’로 판정해 개체 단위로 결합하는 레코드 연계 과정을 도식화한 것이다.

둘째, 통계적 매칭은 데이터를 직접 결합하지 않고, 공통 보조 변수(auxiliary variables)로 미관측 변수(unobserved variables)를 추정하는 방법이다(D’Orazio, 2017). 레코드 연계가 결합을 통해 연계 데이터 세트를 산출하는 것과 달리, 통계적 매칭은 확률적 가정과 모형 기반 추정을 통해 합성 데이터 세트(synthetic data set)를 생성한다(D’Orazio et al., 2006; Rässler, 2002). <그림 2-2>에 도식화된 바와 같이,

10) 데이터 통합에서 전처리·정합성 점검은 자료 수준, 결합은 데이터 수준에서 이루어진다.

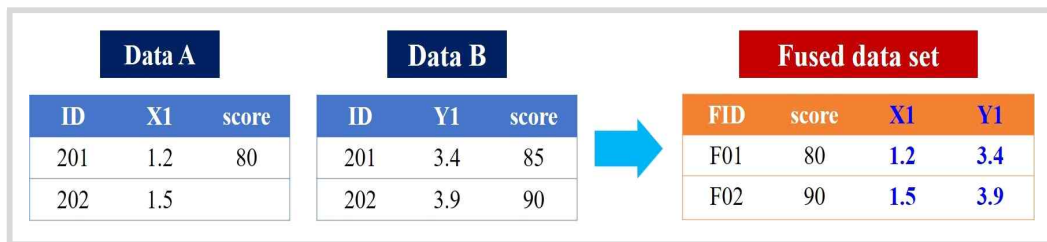
통계적 매칭은 Data A에 없는 변수 Y1을 공통 보조 변수 Z를 매개로 추정하여 합성 데이터 세트를 생성한다.



자료 출처: 저자 작성

<그림 2-2> 통계적 매칭 예시

셋째, 데이터 융합은 결합 여부와 무관하게 상충·중복 정보를 조정하고, 결측을 보완하여 목적에 맞는 일관된 데이터 세트를 구축하는 절차적 틀이다(Boonstra & del Pino, 2025; ESCAP, 2020; Han & Si, 2025). 동일 변수에 서로 다른 값이 존재하면 우선순위(prioritizing)를 정해 대푯값을 선택하거나, 결측이 발생했을 때는 대체(imputation)를 수행한다. 또한, 변수 간 분포나 대표성 차이를 바로잡기 위해 가중값 조정(weight adjustment) 또는 보정(calibration)을 적용하는데, 이러한 사후 조화(post-harmonization) 즉, 통합이 완료된 후, 남아 있는 데이터 간 불일치나 대표성 차이를 조정하는 과정 역시 데이터 융합의 한 형태로 간주할 수 있다. <그림 2-3>은 우선순위 선정과 결측값 대체를 통해 상충·결측 문제를 보완하는 데이터 융합 과정을 도식화한 것이다.



자료 출처: 저자 작성

<그림 2-3> 데이터 융합 예시

데이터 통합에서 세 방법은 독립된 기법이 아니라, 「레코드 연계의 정확한 결합 → 통계적 매칭을 통한 정보 확장 → 데이터 융합을 통한 보정과 정합성 확보」라는 연속적 과정에서 상호보완적으로 가능하다.

제 3 장

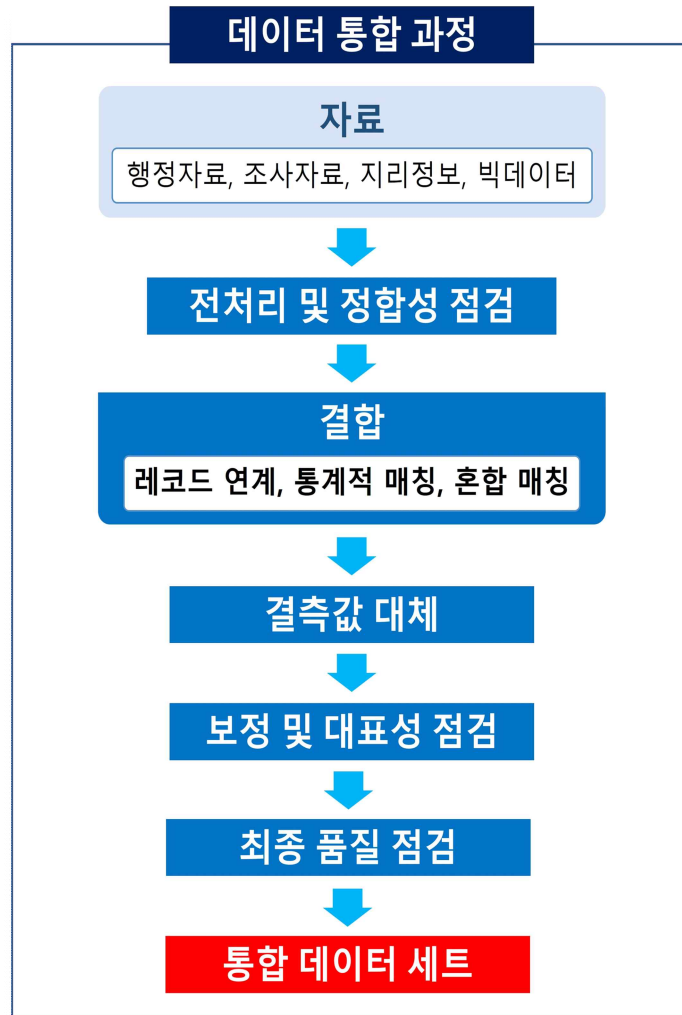
데이터 통합 과정 및 방법

데이터 통합은 자료 수집에서 품질 점검에 이르는 통계생산 과정 전반을 포괄하는 복합적 절차이다. 국내에서는 데이터 통합을 위한 과정이 아직 충분히 구체화하지 않았으며, 개념적·기술적 측면에서도 표준화된 체계가 미흡하다. 이에 본 연구는 데이터 통합 과정 — 전처리 및 정합성 점검(제1절), 결합(제2절), 결측값 대체(제3절)¹¹⁾, 보정 및 대표성 점검(제4절), 최종 품질 점검(제5절)¹²⁾ — 을 단계별로 구분하여 제시하였다. 한편, 본 연구에서는 데이터 융합(data fusion)을 데이터 통합 방법의 유형에는 포함하되, 실제로는 결합 결과에서 발생한 결측과 불일치를 보완하고 대표성을 확보 단계에서 수행되는 절차로 보았다.

각 단계는 활용 가능성과 국제 기준을 염두에 두고 구성하였으며, 통합 대상 자료의 정비, 결합, 품질 점검 등 전 과정을 아우른다. 특히, 각 단계는 통합 목적과 자료 성격, 활용 범위에 따라 필수적으로 수행해야 하는 단계와 선택적으로 적용할 수 있는 단계로 구분된다. 예를 들어, 결합 과정이 단일 자료 내 레코드 연계만으로 충분할 경우 결측값 대체나 보정 단계를 생략할 수 있으며, 반대로 복수 자료의 비정합 문제가 크면 보정 및 품질 점검 절차가 강화된다. 따라서, 본 연구에서

- 11) 결측값 대체(imputation)는 결합 과정의 일부로 수행될 수도 있다(UNECE, 2020 참조). Abowd 등(2021, p. 2)은 “record linkage as a missing data problem where true match status is unknown and must be imputed”라고 명시하며, 결측값 대체가 결합의 틀 안에서 수행될 수 있음을 제시하였다. 그러나, 본 연구에서는 데이터 통합 개념의 일관성과 단계별 과정 구분의 명확성을 위해 **결측값 대체를 결합의 하위 절차가 아닌 독립된 절로 구분하여 제시한다.**
- 12) 품질 점검은 크게 **과정 중심 품질 점검**과 **결과 중심 품질 점검**으로 구분할 수 있다. ①과정 중심 점검은 자료 수집부터 보정에 이르는 데이터 통합 전 과정에서 수행된 활동의 적절성을 검토하는 절차이다. ②결과 중심 점검에서는 최종 산출물인 통합 데이터 세트가 품질 기준을 충족하는지를 평가하는 단계로, 정확성(accuracy), 완전성(completeness), 정합성(consistency), 시의성(timeliness), 대표성(representativeness) 등 다차원적 품질 속성을 중심으로 진단한다. 따라서, 품질 점검은 전 과정 품질 관리와 최종 산출물 진단을 모두 포함하는 이중적 성격을 가지며, 두 접근을 종합함으로써 통합 데이터의 품질을 보장할 수 있다. 본 연구에서는 **데이터 통합 단계별로 과정 중심 품질 점검 절차를 내재화**하고 있으며, 마지막 단계에서 **최종 품질 점검 절차를 별도로 수행**하였다. 즉, 단계별 품질 점검은 통합 과정의 신뢰성을 확보하기 위한 중간 관리 기능, 최종 품질 점검은 통합 데이터 세트의 품질 수준을 종합적으로 진단하기 위한 성과 평가 기능으로 수행하였다.

제시하는 데이터 통합 과정은 일률적 절차가 아니라, 통합 목적과 자료 특성에 따라 반복되거나 생략될 수 있는 유연한 모형적 틀(a flexible methodological framework for data integration)로 설계되었다. 이러한 접근은 단일 기관의 통계를 넘어, 다기관 협업 기반의 통계작성 체계 구축, 공공·민간 데이터의 통합 활용 기반으로 확장될 수 있다.



<그림 3-1> 데이터 통합 과정

<그림 3-1>은 본 연구의 데이터 통합 과정을 도식화한 것으로 통합 데이터 세트가 구축되는 체계적 절차를 제시한다.

제1절 전처리 및 정합성 점검

데이터 통합의 첫 단계인 전처리 및 정합성 점검은 서로 다른 출처에서 수집한 자료의 형식·개념·논리를 일관되게 맞추는 과정이다(ESCAP, 2020; UNECE, 2019b; Statistics New Zealand, 2013). 전처리는 스키마 정렬(schema alignment)과 개념 조화(semantic harmonization) 작업에 해당하고, 정합성(consistency) 점검은 이러한 결과가 논리적 모순 없이 구현되었는지를 검증한다.

1. 자료 수집

자료 수집은 조사자료, 행정자료, 지리정보, 빅데이터 등 서로 다른 목적과 체계에서 생성된 자료를 확보하는 과정이다. 이 단계에서는 원자료(raw data source)의 제공 조건과 법·제도·기술적 환경을 검토하고, 생성 목적, 표본 틀, 참조 기간, 단위 및 개체 정의, 수정 이력 등을 포함한 메타데이터를 함께 확보해야 한다. 이는 이후 스키마 정렬, 개념 조화, 정합성 검증의 기준을 마련하는 핵심 절차이다. 따라서 자료 수집은 단순한 취득이 아니라, 전처리 전반을 좌우하는 설계 단계로서 체계적인 메타데이터 작성이 필수적이다.

2. 전처리

전처리(preprocessing)는 자료를 결합이 가능한 데이터로 전환하기 위한 절차로 자료 진단, 스키마 정렬, 개념 조화, 암호화로 구성된다(ESCAP, 2020; UNECE, 2019b). 이는 단순히 오류를 수정하거나 변수명을 정리하는 기술적 사전 작업을 넘어, 데이터 통합 전체의 성패를 좌우하는 토대가 될 수 있다. 본 연구에서는 전처리의 중요성을 고려하여 다음의 기본 원칙을 제안한다.

첫째, 최소침습성(Statistics New Zealand, 2013)은 원자료를 불필요하게 변형하지 않고, 필요한 경우에만 개입을 허용하는 원칙이다. 가령, 주민등록번호에서 오류가 발견될 경우, 임의 수정이 아니라 외부 자료 검증을 통해 보완한다.

둘째, 재현성(Statistics New Zealand, 2013)은 전처리 절차를 반복했을 때, 동일한 결과를 도출할 수 있어야 한다는 원칙이다. 이는 전처리 과정 자체의 정합성을 담보하며, 이후 단계에서 산출물의 신뢰성을 보장하는 근거가 된다.

셋째, 메타데이터 우선(Leulescu & Agafitei, 2013)은 변수 정의, 코드 체계, 분류 구조 등 메타데이터의 정비를 자료 처리에 앞서 수행하는 것을 의미한다. 행정자료의 경우 기관별 코드와 정의가 달라 메타데이터 정비 없이는 결합할 수 없는 경우가 많다. 이는 결합 단계에서 발생할 수 있는 구조적 불일치를 원천적으로 차단하는 효과가 있다.

넷째, 윤리적·법적 요건 준수(Harron, 2016)는 개인정보 보호와 접근권한 관리를 포함하여, 사회적 합의를 기반으로 데이터 통합의 지속 가능성을 담보하는 핵심 원칙이다.

다섯째, 상호운용성(ESCAP, 2020; Eurostat, 2024)은 기관·시스템·국가 간 데이터가 호환되고 교류될 수 있도록 구조적·기술적 정합성을 확보하는 원칙이다. 본 연구는 상호운용성 원칙을 전처리에 포함함으로써 국내 기관 간 통합뿐만 아니라 국제 비교 통계생산에도 적용될 수 있도록 제안한다.

가. 자료 진단

자료 진단(profiling)은 자료 간 불일치와 중복을 확인하고, 변수별 결측값과 이상값(outliers)을 점검하는 과정이다(de Waal et al., 2019; Herzog et al., 2007). 결측값과 이상값은 단순 삭제나 추정보다는 결측 코드(NA)나 부재 코드(Unclassified)를 부여하여 이후의 결측값 대체 단계에서 처리한다.

나. 스키마 정렬

스키마¹³⁾ 정렬(schema alignment)은 서로 다른 자료에서 동일 개념일 가능성이 큰 변수를 식별하고, 대응 관계(mapping relation)를 설정하는 과정이다(Rahm & Bernstein, 2001). 목적은 개념 조화 단계에서 변환과 표준화를 수행할 수 있도록 변수명, 자료 유형, 코드 체계, 단위 등의 형식적 일관성(coherence)을 확보하는 데 있다(Rahm & Do, 2000; Rahm & Bernstein, 2001). 같은 개념이라도 자료 출처에 따라 스키마가 다를 수 있다. 예를 들어, 성별이 행정자료에서는 ‘남/여’, 조사자료에서는 ‘M/F’로 기록된 경우, 성별 변수를 결합 후보(mapping candidate)로 선정하는 것이 스키마 정렬이다.

- **스키마 추출 및 구조 분석.** 통계설명자료, 행정자료 서식, 설문 조사표 등을 활용해 자료별 변수·속성, 자료 유형, 코드, 단위 등 구조적 특성을 파악한다. 예컨대, 소득 변수가 조사자료에는 월 단위, 행정자료에서는 연 단위로 기록되었으면 구조 분석을 통해 이러한 차이를 식별할 수 있다.

13) 스키마(schema)는 자료 또는 데이터의 구조를 규정하는 틀로, 변수명, 자료 유형, 코드 체계, 단위, 제약 조건 등을 포함한다(Rahm & Bernstein, 2000).

- **의미적 유사성 평가.** 문자 유사도(예: Levenshtein 거리¹⁴, Jaro-Winkler 유사도¹⁵)나 사전 및 온톨로지(ontology, 개념 간 연관성)¹⁶를 활용하여 변수 간 의미적 유사성을 확인한다(Doan & Halevy, 2005; Rahm & Do, 2000; Rahm & Bernstein, 2001). 예를 들어, 문자열 유사도 분석을 통해 ‘출생 연도’와 ‘생년’이 본딴말-준말(약어) 관계임을 탐지할 수 있다.
- **시점 변수 식별.** 동일 대상을 다루더라도 자료 수집 시점이 다르면 통합 데이터의 정확도가 저해될 수 있다. 따라서 각 자료에 포함된 시점 변수의 존재와 구조를 식별한다.¹⁷

14) 문자열 간 삽입, 삭제, 치환의 최소 연산횟수를 의미하며, 거릿값이 작을수록 두 문자열은 유사하다(Berger et al., 2021).

$$ld_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0 \\ \min \begin{cases} ld_{a,b}(i-1,j)+1 \\ ld_{a,b}(i,j-1)+1 \\ ld_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} \end{cases} & \text{otherwise} \end{cases}$$

(a, b): 비교 문자열, i는 a의 i번째 문자까지를 의미

(i, j): a, b에서 현재 비교 중인 위치(문자 인덱스)

☐ ‘정필재’를 ‘전필재’로 치환하면 $ld = 1$, ‘정으나’를 ‘정은아’로 치환하면 $ld = 2$

15) 철자 순서와 접두사 일치 등을 반영해서 유사도를 측정하는 방법으로, 0(불일치) ~ 1(완전 일치) 사이의 점수로 표현한다(Jaro, 1989; Winkler, 1990).

- Jaro similarity: $J = \frac{1}{3} \left(\frac{m}{|s_a|} + \frac{m}{|s_b|} + \frac{m-t}{m} \right)$

(s_a, s_b): 각 문자열의 길이, m: 일치 문자 수, t: 치환된 문자 쌍의 $\frac{1}{2}$

- Jaro-Winkler similarity: $JW = J + \{p \times L \times (1-J)\}$

L: 접두사의 길이(최대 4), p: scaling factor(보통 0.1)

☐ ‘조나래’ vs. ‘조나레’: $J \approx 0.9333$, $L = 3$, $JW \approx 0.9533$

16) 문자열 유사도만으로 일치하지 않는 경우(예: surname - family name - last name), 사전 기반 동의어 결합 또는 온톨로지 기반 개념 유사도(Wu & Palmer, 1994)를 활용한다.

- Conceptual similarity = $\frac{2 \times N_3}{N_1 + N_2 + (2 \times N_3)}$

C_3 : 두 개념(C_1, C_2)의 최소 공통 상위 개념

N_1 : C_1 에서 C_3 까지의 노드 수

N_2 : C_2 에서 C_3 까지의 노드 수

N_3 : C_3 에서 근본 개념(root)까지의 노드 수

☐ 정보(root) - 개인정보(C_3) - 연락처(C_2) - 전화번호(C_1), 이메일, 주소

$$\text{Conceptual similarity} = \frac{2 \times 1}{2 + 1 + (2 \times 1)} = 0.4$$

- **결합 후보 선정.** 구조 분석과 의미적 유사성 평가를 종합해 동일 개념일 가능성이 높은 변수를 식별하여 결합 후보(mapping candidate)를 선정한다.
- **대응표 작성.** 선정된 결합 후보에 대해 변수 유형, 코드 체계, 단위 등의 결합 가능성을 정리한 대응표(concordance table)를 작성한다.

다. 개념 조화

개념 조화(semantic harmonization)는 자료 간 변수의 의미와 정의를 공통 기준에 맞추어 변환·표준화함으로써 **개념적 일관성**을 확보하는 과정이다(ESCAP, 2020). 이는 단순한 외형을 일치시키는 것을 넘어, “**동일 현상을 동일하게 관측·해석·비교할 수 있도록 만드는 과정**”이다(Eurostat, 2014; Rahm & Bernstein, 2001). 개념 조화는 「기준 개념 정의-규칙 설계-변환·표준화」 순으로 진행한다.

- **기준 개념 정의.** 통합 목적에 부합하는 기준 개념(reference concept)을 설정하는 것이 개념 조화의 첫 단계이다. 기준 개념은 변수의 의미뿐만 아니라, 단위·시점·분류·범위를 함께 규정한다. 즉, 기준 개념은 변수의 전체 속성을 포괄하는 기준으로, 변환과 표준화의 모든 과정은 이 기준에 맞추어진다. 소득 변수를 예로 들면, 통합 목적이 가구의 경제 능력 평가라면 세전소득, 실제 가치분 자원 분석이 목적이면 세후소득이 기준 개념이 된다.
- **규칙 설계.** 기준 개념이 확정되면 변수를 이에 맞추어 변환·표준화하는 규칙을 설계한다. 이는 단순한 기술적 처리 절차가 아니라, 개념적 동등성(conceptual equivalence)을 확보하는 장치이다(Rahm & Bernstein, 2001). 다만, 지나치게 엄격하거나 과도한 규칙은 변수—특히, 고유식별자와 공통 변수—의 변별력을 약화할 수 있으므로 주의가 필요하다.

아래는 규칙 설계에 따라 수행한 **변환 및 표준화 예시**이다.

- **개념 환산.** 변수 정의가 다를 경우 기준 개념에 맞추어 환산해야 한다. 동일 변수명이라도 의미가 다르면 결합 시 결과가 왜곡될 수 있다. 가령, 소득의 기준

17) 자료별 시점 변수를 정의하고, 그 차이를 계량적으로 점검하여야 한다(Eurostat, 2014). 이에 본 연구는 ‘자료 간 평균 시점 차이(mean time gap)’라는 지표를 제안한다.

$$\text{평균 시점 차이} = \frac{1}{n} \sum_{i=1}^n |t_i^A - t_i^B|$$

t_i^A, t_i^B : 동일 레코드(i)에 대한 두 자료의 기준 시점

이 지표는 값이 작을수록 시점 불일치가 적음을 의미한다. 예를 들어, 동일한 A에 대해 조사자료는 ‘2020.7.1.’, 행정자료는 ‘2020.12.31.’로 기록되어 있다면, 시점 차이는 183일이다. 이러한 방식으로 개별 레코드의 차이를 계산한 후 평균을 내면 전체 자료 간 평균 시점 차이를 산출할 수 있다.

개념을 세전소득으로 정한다면, 세후소득에 세금과 사회보험 기여금을 더해 세전소득으로 환산한다.

- **코드 변환.** 명목 척도 변수는 수치적 의미가 없으므로, 서로 다른 코드 체계는 동일 코드 체계로 통일해야 한다(Rahm & Bernstein, 2001). 예를 들어, 성별 변수가 ‘M/F’로 표기되었다면 국제표준(ISO/IEC 5218)에 따라 ‘1=남성, 2=여성’으로 변환한다.
- **형식 표준화.** 동일 대상이라도 표기 형식이 다르면 형식 불일치로 품질 저하가 발생한다(Eurostat, 2014). 예를 들어, ‘세종특별자치시’와 ‘세종시’, ‘서울특별시 강남구 역삼동’과 ‘서울 강남 역삼동’은 같은 지역이지만, 형식 불일치로 다른 값으로 인식될 수 있다. 이를 방지하기 위해 날짜·시간·수치·주소 등은 표기를 통일한다.
- **문자 정규화.** 전각·반각, 대소문자, 공백, 특수문자 등 단순 표기 차이를 제거하여 일관된 문자열로 정규화한다. 예를 들어, ‘A B C’는 ‘ABC’로, ‘SEOUL’은 ‘Seoul’로, ‘(주)삼성’은 ‘삼성’으로 변환한다.
- **유사 문자 변환.**¹⁸⁾ 단순 표기 차이가 아닌 철자 변이(spelling variation)나 발음 유사성(phonetic similarity)으로 인해 유사 문자열이 발생할 수 있다. 예컨대, ‘정은아’와 ‘정은나’는 표기는 다르지만, 발음이 유사해 동일인일 가능성이 있으며, ‘soonjin’과 ‘sunjin’은 모음 철자만 다른 경우로 음운적으로는 같을 수 있다. 이럴 때, Levenshtein 거리나 Jaro-Winkler 유사도 등 의미적 유사성 결과를 활용해 두 문자열을 통일한다.
- **시점 표준화.** 기준 시점과 단위를 일치시킨다.¹⁹⁾ 예컨대, 월 단위 자료를 연

18) 문자 정규화가 표기상의 형식 차이를 제거하여 형식적 일관성을 확보하는 절차라면, 유사 문자 변환은 철자와 음운 변이까지 고려해 개념적 일관성을 확립하는 과정이다. 이는 이름, 주소와 같은 공통 변수의 식별에 특히 중요하다.

19) 시점 단위의 변환 및 표준화 예시

• 월 단위 - 연 단위 변환:
$$U_{year} = \sum_{m=1}^{12} U_m / U_m = \frac{U_{year}}{12}$$

• 회계연도(2024.4.~2025.3.) - 통상연도(2024.1. ~ 2024.12.) 변환:
$$U_{CY,2024} = \frac{9}{12} U_{FY,2024} + \frac{3}{12} U_{FY,2025}$$

• **기준일 조정:** 조사자료(t_1 : 2024.10.1., y_1 : 2000)와 행정자료(t_2 : 2024.12.31., y_2 : 2200)의 기준일이 다를 경우

- 조정 기준일(t_{adj})이 조사자료(t_1)와 행정자료(t_2)의 사이에 있다고 가정

- y_i : 시점별 레코드값, 일(day) 단위로 계산

① 중간값 보정: 두 시점 간 변화가 완만하고, 기간(gap)이 짧을 경우, t_{adj} 가 $t_1 \sim t_2$ 의 중간에 있다고 가정한다.

단위로, 회계연도(fiscal year)를 통상연도(calendar year)로 변환한다. 이 과정에서 스키마 정렬에서 산출한 ‘자료 간 평균 시점 차이’를 활용하면 보정 필요성과 그 범위를 판단할 수 있다.

- **단위 변환.** 변수의 측정 단위가 다를 경우 기준 단위로 환산한다. 가령, 천 원 단위의 조사자료와 백만 원 단위의 행정자료는 단위 조정값(adjustment factor)²⁰⁾을 적용해 통일한다.
- **분류체계 표준화.** 산업, 직업, 지역 등 범주형 변수는 분류체계가 다를 수 있으므로 국제표준 분류(예: ISIC, ISCO, ISCED)를 기준으로 표준화하거나 상위 공통 분류 수준으로 재분류한다. 예컨대, 행정자료는 국제표준산업분류(ISIC), 조사자료는 한국표준산업분류(KSIC)일 경우, 대응표를 활용해 동일 산업 단위로 재분류한다.
- **포괄범위 표준화.** 동일 변수라도 모집단 범위가 다를 수 있다. 예를 들어, ‘취업자’를 행정자료는 18세 이상 인구로 정의하고, 조사자료는 15~64세만 포함한다면, ‘18세~64세’를 기준으로 설정하거나 결측 처리로 범위를 통일한다.
- **결측값 처리.**²¹⁾ 개념 조화 단계에서 결측값 처리는 주로 변환·표준화 과정에서

$$\hat{y}(t_{adj}) = \frac{y_1 + y_2}{2} = \frac{2000 + 2200}{2} = 2100$$

② 선형 보정: 두 시점 사이에 뚜렷한 비선형적 변화가 없을 때, $t_1 \rightarrow t_2$ 의 선형 변화를 가정해 그 비율만큼 보정한다(Newsbury, 1981 참조).

$$w = \frac{day(t_1, t_{adj})}{day(t_1, t_2)}, \quad \hat{y}(t_{adj}) = y_1 + w(y_2 - y_1)$$

- 조정 기준일(t_{adj})을 2024.11.15.로 가정했을 경우,

$$w = \frac{day(2024.10.1., 2024.11.15.)}{day(2024.10.1., 2024.12.31.)} = \frac{45}{91} \approx 0.495$$

$$\hat{y}(2024.11.15.) = 2000 + 0.495(2200 - 2000) = 2099$$

③ 최근접값 보정: 레코드값이 급변할 때, 가장 가까운 시점의 값을 사용하는 방법이다. 계산이 간단하고 적용이 쉽지만, 시간 간격이 크면 정확성이 떨어지고 추세 반영에도 한계가 있다(Lepot et al, 2017 참조).

$$\hat{y}(t_{adj}) = \begin{cases} y_1, & \text{if } |t_{adj} - t_1| < |t_{adj} - t_2| \\ y_2, & \text{otherwise} \end{cases} = \begin{cases} 2000, & \text{if } 45 < 46 \\ 2200, & \text{otherwise} \end{cases} = 2000$$

20) 단위 조정값 = $\frac{\text{원자료 단위 값}}{\text{기준 단위 값}}$

가령, 조사자료(천 원 단위)의 ‘15,700천 원’을 행정자료(백만 원 단위)의 단위($\frac{1000}{1000000} = 0.001$)로

통일하면 ‘15.7백만 원’이 된다.

21) 데이터 **결합(matching)** 후, 발생한 결측값 처리는 **대체(imputation)** 단계에서 수행한다.

발생한 구조적 결측을 관리하는 것이다. 예컨대, 단위 변환이나 분류 표준화 과정에서 일부 값이 변환되지 않거나 소득 변수가 세전 기준으로 환산되지 못한 경우가 이에 해당한다. 이러한 결측은 통계적 추정보다는 자료 품질 관리 차원에서 결측(NA) 또는 부재 코드(Unclassified)를 부여한다. 또한, 자료별로 결측 표기가 다른 경우(예: 조사자료 ‘9999’, 행정자료 ‘.’), 이를 ‘NA’로 통일한다.

라. 암호화

암호화(encryption)는 개인정보 보호와 보안을 보장하기 위한 핵심 절차이다(Harron, 2016). 주민등록번호, 성명, 주소와 같은 식별자가 유출된다면 심각한 개인정보 침해로 이어져 자료 활용 자체가 불가능해질 수 있다. 따라서, 식별이 가능한 개인정보는 반드시 암호화하거나 비식별 대체 번호(pseudonymous matching keys)를 생성해야 한다. 암호화는 개념 조화 이후, 정합성 점검 이전이 가장 적절하다. 변수 정의와 코딩 체계가 표준화된 뒤 암호화를 적용해야 표준화된 변수값을 기반으로 안정적인 암호화를 진행할 수 있다. 반대로 너무 이른 단계에서 암호화하면 품질 관리와 오류 점검이 어렵고, 지나치게 늦으면 개인정보 노출 위험이 커진다. 즉, 암호화는 단순한 기술적 조치가 아니라, 데이터 통합 전체의 신뢰성과 정당성을 뒷받침하는 제도적·윤리적 장치이다. 통합 과정에서 보안이 보장되지 않으면 사회적 수용성을 얻기 어렵다.

3. 정합성 점검

정합성(consistency)은 자료의 논리적 일관성을 전제로 결합 가능성을 판단하는 기준이다(ESCAP, 2020; Eurostat, 2014; UNECE, 2019b). 스키마 정렬과 개념 조화를 통해 형식적·개념적 일관성을 확보했다더라도, 자료가 기준에 부합하지 않는 경우가 있다. 예를 들어, 변수 단위를 모두 연 단위로 변환했음에도 일부가 월 단위로 남아 있을 수 있고, 한국표준산업분류(KSIC)로 변환하는 과정에서 일부 코드가 잘못 연결될 수도 있다. 이러한 오류, 왜곡, 정보 손실을 최소화하여 통합 데이터 세트의 품질을 사전에 확보하기 위해서는 정합성 점검이 필수적이다. **정합성 점검은 전처리 단계와 연계해 반복 수행되며, 그 결과는 메타데이터에 체계적으로 기록한다(ESCAP, 2020).**

가. 개념 정합성

개념 정합성(consistency of definition)은 동일 변수가 동일한 개념으로 정의·적용되고 있는지를 검증한다(ESCAP, 2020; Eurostat, 2014; UNECE, 2019b). 가령, 가계소득을

세전소득 기준으로 변환했음에도 일부 레코드에 세후소득 값이 남아 있다면 이는 개념 정합성이 훼손된 사례라 할 수 있다. 이러한 문제는 개념 조화 단계에서의 불완전 변환이나 원자료 결측에 기인한다.

$$\text{개념 정합성} = \frac{\text{기준 개념과 일치하는 레코드}}{\text{전체 레코드}} \times 100$$

나. 시점 정합성

시점 정합성(temporal consistency)은 자료가 동일한 기준 시점(reference date) 또는 보고 기간(reporting period)을 반영하고 있는지 검증한다(ESCAP, 2020; Eurostat, 2014; UNECE, 2019b). 예를 들어, 행정자료는 2024년 연평균 임금, 조사자료는 2024년 12월 임금을 제공한다고 가정할 때, 이를 연평균으로 환산하는 과정에서 2023년 값이나 특정 월의 값이 남아 있다면 시점 정합성이 훼손된 것이다.

$$\text{시점 정합성} = \frac{\text{기준 시점과 일치하는 레코드}}{\text{전체 레코드}} \times 100$$

다. 단위 정합성

단위 정합성(unit consistency)은 동일 변수가 동일한 측정 단위를 따르고 있는지를 검증한다(ESCAP, 2020; Eurostat, 2014; UNECE, 2019b). 예를 들어, 임금 변수를 연 단위로 통합했음에도 일부 기록이 월 단위 상태로 남아 있다면 단위 정합성이 손상된 것이다. 이 경우 단위 차이를 확인하고 단위 조정값을 활용하여 단위를 통일한다(Harron, 2016; Eurostat, 2014).

$$\text{단위 정합성} = \frac{\text{기준 단위와 일치하는 레코드}}{\text{전체 레코드}} \times 100$$

라. 분류 정합성

분류 정합성(classification consistency)은 동일 개념이 일관된 분류체계를 따라 분류되어 있는지 검증한다(ESCAP, 2020; Eurostat, 2014). 가령, 한국고용직업분류(KECO)를 한국표준직업분류(KSCO)로 변환하는 과정에서 일부 코드가 상위 수준에서만 결함할 경우 세부 코드와 불일치할 수 있다. 이러한 불일치는 대응표(concordance table)를 활용하여 교정·표준화한다.

$$\text{분류 정합성} = \frac{\text{기준 분류체계와 일치하는 레코드}}{\text{전체 레코드}} \times 100$$

마. 범위 정합성

범위 정합성(coverage consistency)은 변수가 동일한 모집단 범위와 적용 조건을 일관되게 포괄·반영하고 있는지를 검증한다(ESCAP, 2020; Eurostat, 2014; UNECE, 2019b). 예컨대, 고용률 산출의 기준 개념을 ‘18세~64세’로 정의했음에도 일부 레코드에서 65세 이상이 포함되어 있다면, 범위 정합성이 훼손된 것이다. 이는 전처리 단계에서 모집단 조건이 일관되게 적용되지 않을 때 발생한다.

$$\text{범위 정합성} = \frac{\text{기준 포괄범위와 일치하는 레코드}}{\text{전체 레코드}} \times 100$$

바. 통계적 정합성

통계적 정합성(statistical consistency)은 결합할 자료들이 전체적으로 유사한 분포와 구조를 보이는지를 점검하는 절차이다(Eurostat, 2014; Harron et al., 2017; Herzog et al., 2007). 이는 개별 변수·레코드 차원의 형식적·개념적 오류 점검과 달리, 집합 수준에서의 정합성을 평가한다. 예를 들면, 두 자료가 동일 모집단을 반영한다고 가정했을 때, 연령별 분포, 성별 비율, 지역별 인구 구조, 소득 계층별 비중이 과도하게 차이가 나면 안 된다. 특정 집단의 비중이 조사자료에서는 30%인데 행정자료에서는 10%라면, 그 집단은 과대 표집이 된 것이다. 마찬가지로 지역별 인구 구조가 자료마다 현저하게 다르면 데이터 통합 결과가 왜곡될 수 있다(Harron et al., 2017).

통계적 정합성은 **기술통계(예: 평균, 중간값, 분산, 비율 등)**와 **분포 동일성 검정**으로 확인할 수 있다(Harron et al., 2017; Herzog et al., 2007). 예를 들어, 범주형 변수는 카이제곱 검정(chi-square test)으로, 연속형 변수는 Kolmogorov-Smirnov 검정으로 확인한다.²²⁾ 만약 분포 차이가 유의하였다면, 결합 전 단계에서 보정 가중값(calibration weighting)을 적용해 조정²³⁾할 필요가 있다(Deville & Särndal, 1992; Harron et al., 2017; Särndal, 2007).

22) 두 경험적 분포(누적분포함수; ECDF) 간의 최대 거리를 측정하여, 두 분포가 같은지를 검정하는 비모수적 방법이다(NIST, n.d.). 이 ‘최대 거리($D_{n,m}$)’가 클수록 두 분포의 차이가 크다.

$$D_{n,m} = \max_x |F_n(x) - G_m(x)|, \quad F_n(x) = \frac{1}{n} \sum_{i=1}^n 1\{X_i < x\}, \quad G_m(x) = \frac{1}{m} \sum_{j=1}^m 1\{Y_j < x\}$$

$D_{n,m}$: 두 누적분포함수가 가장 멀리 떨어진 거리

F_n, G_m : 표본 X, Y 의 누적분포함수

☞ $X = \{2, 4, 5\}$ ($n = 3$), $Y = \{1, 3, 5, 7\}$ ($m = 4$), $x : 1, 2, 3, 4, 5, 7$

$$\Rightarrow D_{n,m} = \frac{1}{4} = 0.25$$

제2절 결합

결합(matching)은 레코드 연계(record linkage), 통계적 매칭(statistical matching), 혼합 매칭(hybrid matching)으로 구분한다. 이러한 구분은 ①결합 목적, ②입력 변수(input variables)²⁴⁾, ③산출물(outputs)이 근본적으로 다르기 때문이다. 레코드 연계는 연결 변수(고유식별자, 공통 변수)를 활용하여 동일 개체를 식별하고 연결하는 것이 목적이며, 그 결과 개체 간 연계 결과표(linkage table)²⁵⁾를 산출한다(Herzog et al., 2007). 반면, 통계적 매칭은 공통 보조 변수를 매개로 미관측 변수를 추정하여 합성 데이터 세트(synthetic data set)를 산출하는 것이 목적이다(D’Orazio, 2017). 두 방법은 상호 대체적 관계가 아니라 보완적으로 작동한다. 일반적으로 레코드 연계를 통해 연계 틀(link frame)을 구축한 뒤, 필요한 변수만 통계적 매칭으로 보강하는 **혼합 매칭**이 활용된다. 결합 과정을 이처럼 세 범주로 구분하면, 각 결합 방법의 ①목적, ②입력 변수, ③산출물이 명확히 구분되어 데이터 통합의 논리적 일관성과 투명성을 확보할 수 있다.²⁶⁾

x	1	2	3	4	5	7
$F_n(x)$	0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{2}{3}$	1	1
$G_m(x)$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{3}{4}$	1
$F_n(x) - G_m(x)$	$\frac{1}{4}$	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{4}$	0

- 23) **예** 표본 가중합이 남성 60%, 여성 40%, 목표가 남녀 각각 50%라면 보정 가중값은 남성 가중값에 $0.83(\approx \frac{5}{6})$, 여성 가중값에 $1.25(= \frac{5}{4})$ 를 곱하여 표본 분포가 목표에 근접하도록 한다.
- 24) 입력 변수는 **연결 변수(linkage variables)**와 **조건 변수(conditioning variables)**로 구분할 수 있다(D’Orazio, 2017; Harron et al., 2017; Winkler, 2014). **연결 변수**는 **고유식별자·공통 변수**(예: 주민등록번호, 사업자등록번호, 성명, 생년월일, 주소 등)와 같이 **동일성 판별**이 가능한 변수, **조건 변수**는 **공통 보조 변수**(예: 연령대, 학력, 지역, 소득 분위 등)와 같이 추정을 위한 **예측 변수**를 의미한다. 연결용 변수와 조건용 변수는 일부 겹치기도 하지만(예: 연령대, 지역 등), 그 목적과 기준이 다를 수 있다.
- 25) **연계 결과표**는 데이터 간 연계 점수, 연계 결정 등을 기록한 **중간 산출물**로 **최종 산출물은 연계 데이터 세트(linked data set)**이다. 본 연구에서는 레코드 연계와 통계적 매칭의 구분을 위해 편의상 산출물을 ‘연계 결과표’와 ‘합성 마이크로데이터 세트’로 제시하였다.
- 26) **결합 목적에 따라 입력 변수와 산출물의 형태가 달라질 수 있다.** 정책 평가 목적 결합에서는 식별 변수(identifying variable)와 시점 변수의 정확성이 중요하지만, 모형 추정을 위한 결합에서는 변수 간 통계적 상관관계와 분포 일치가 더 중요하게 고려된다. 따라서, 결합 목적을 명확히 설정하고 이에 부합하는 입력 변수 구성과 산출물 형태를 설계하는 것이 필수적이다.

1. 레코드 연계

레코드 연계는 서로 다른 데이터에서 동일 개체를 식별하고 연결하는 과정(Herzog et al., 2007; Winkler, 2006)으로 다음의 다섯 단계로 이루어진다.

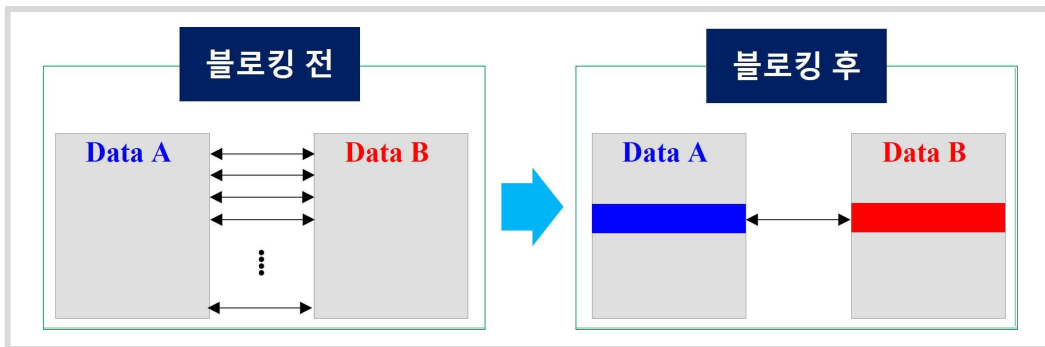
- 1단계: 블로킹. 전수 비교에서 후보군을 줄여 계산 효율성 확보
- 2단계: 비교 및 유사도 산출. 후보 레코드 쌍의 속성 유사성 정량화
- 3단계: 연계 방법 선택. 동일성 여부 평가
- 4단계: 연계 결정. 점수나 확률을 기준으로 최종 판정
- 5단계: 연계 품질 점검. 연계 품질 측정, 연계 결과표 작성 등

가. 블로킹

블로킹(blocking)은 모든 레코드 쌍을 전수 비교(full pairwise comparison)하지 않고, 일정 기준에 따라 비교 대상을 제한하여 연산 효율성을 높이는 절차이다(Christen, 2012; Herzog et al., 2007). 예컨대, 두 데이터에 각각 백만(10^6) 개의 레코드가 존재한다면, 전수 비교 시 일조(10^{12}) 개의 조합이 발생한다. 따라서, 동일 속성값을 공유하는 레코드만 비교 대상으로 남기는 블로킹이 필수적이다.

$$C = \{(i, j) : b(x_i) = b(y_j)\}$$

$b(x_i), b(y_j)$: 블로킹 함수²⁷⁾



자료 출처: 저자 작성

<그림 3-2> 블로킹의 기본 개념

27) 블로킹 함수는 후보 레코드 쌍을 줄이기 위해 레코드를 특정 속성(예: 생년월일)을 기준으로 그룹화하는 함수이다.

예) $b(x_i), b(y_j) = \text{생년월일} \Rightarrow \text{생년월일이 같은 사람끼리만 비교}$

블로킹은 연산 효율성을 높이지만, 재현율(recall)²⁸⁾이 낮아질 위험이 있다(Winkler, 2006). 따라서 효율성과 정확성 간의 균형을 고려한 적절한 블로킹 기법 선택이 중요하다. 주요 기법은 다음과 같다.

- **정확 블로킹(exact/static blocking)**. 단일 블로킹 키(blocking key)를 활용하여 후보군을 구성한다(Herzog et al., 2007). 예를 들어, ‘출생 연도+성별’을 블로킹 키로 설정하면, ‘1991년생 여자’만 상호 비교된다. 구현이 간단하고 효율성이 높으나, 블로킹 키의 입력 오류로 동일인이 다른 블록에 분류되면 누락이 발생하여 재현율이 낮아질 수 있다.
- **정렬 기반 블로킹(sorted neighborhood method)**. 특정 키(예: 이름, 생년월일 등)를 기준으로 데이터를 정렬한 뒤, 일정 크기의 창(window)을 설정하여 해당 범위 내 레코드만 비교한다(Herzog et al., 2007). 예를 들어, 이름을 가나다순으로 정렬하면 ‘안중철’은 ‘안중천’, ‘안중청’ 등 인접한 후보와만 비교한다. 결측이나 오자(typo)가 존재하더라도 레코드 비교가 가능하지만, 창의 크기 설정에 따라 누락률이 달라질 수 있다.
- **캐노피 클러스터링(canopy clustering)**. 간단한 거리 지표(예: TF-IDF²⁹⁾, Jaccard 유사도³⁰⁾)로 구역을 설정한(clustering) 후, 구역 내에서 정밀 비교(예: Jaro-Winkler 유사도)를 수행하는 방법이다(Christen, 2012). 주소나 사업체명처럼 철자 변형이 잦은 변수에 유용하며, 수억 건 이상의 대규모 데이터에 적합하다. 다만, 구역 설정 기준에 따라 재현율이 달라질 수 있다(Christen, 2012; McCallum et al., 2000).
- **다중 블로킹(multi-pass blocking)**.³¹⁾ 하나의 블로킹 기준(pass)만 적용하면 오연계

28) 재현율(recall)은 실제 일치 쌍 중 ‘일치’로 올바르게 분류한 비율이다.

29) $TF-IDF(t, d) = TF(t, d) \times IDF(t)$

• TF(term frequency): 해당 단어의 빈도

예) ‘연계’가 6번 등장했다면 $TF(“연계”) = 6$

• IDF(inverse document frequency): 전체 문서(자료)에서 해당 단어가 가지는 희소성, IDF가 높으면 출현 빈도가 낮은 단어, IDF가 낮으면 빈도가 높은 단어

$$IDF(t) = \log\left(\frac{N}{1+df(t)}\right)$$

$df(t)$: 단어(t)가 등장한 문서(d)의 수

30) Jaccard 유사도: 얼마나 많은 요소가 두 데이터에 공통으로 포함되어 있는지 측정하는 방법

$$Jaccard\ similarity(A, B) = \left| \frac{A \cap B}{A \cup B} \right|$$

31) 다양한 ‘블로킹 키(blocking key)’ 조합을 사용하는 방식은 Christen(2012)과 Winkle(2006) 등에서

(false match)는 줄일 수 있으나, 누락이 증가한다. 다중 블로킹은 서로 다른 블로킹 키를 순차적으로 적용하여 이를 보완하는 방법이다(Christen, 2012). 예를 들어, 1차는 ‘성별+출생 연도+시군구’, 2차는 ‘성별+출생 월+성(姓)’을 블로킹 키로 설정한다. 식별자 오류나 결측의 영향을 덜 받아 재현율을 높일 수 있지만, 연산량이 다시 증가하는 단점이 있다(Christen, 2012).

나. 비교 및 유사도 산출³²⁾

블로킹으로 후보 레코드 쌍의 수가 줄더라도, 동일 개체 여부를 즉시 확정할 수는 없다. 따라서 각 후보 쌍의 속성값을 비교하고, 유사한 정도를 수치화한 유사도 점수(similarity score)를 산출해야 한다(Christen, 2012; Herzog et al., 2007). 이 절차는 연계 결정을 위한 정량적 근거를 제공한다.

- **문자열 변수**는 동일성 판별의 핵심 요소지만, 오자나 약자, 철자 변형으로 인해 단순 일치 여부만으로는 동일 개체 여부를 판단하기 어렵다. 이에 Levenshtein 거리, Jaro-Winkler 유사도, 코사인(cosine) 유사도³³⁾ 등 문자열 기반 거리 함수를 이용해 유사도를 계산한다.
- **수치형 변수**는 단순한 일치보다는 차이의 허용 범위(tolerance)를 설정해 비교해야 한다. 절댓값 차이, 비율 차이, 정규화 거리(normalized distance)³⁴⁾ 등을 활용한다.

간접적으로 언급되고 있다. 이들은 다양한 블로킹 전략의 효율성과 정합성 향상 가능성을 언급하고 있으나, ‘다중 블로킹(multi-pass blocking)’이라는 용어를 명시적으로 정의하거나 직접적인 사례를 제시하지는 않았다.

32) **비교 및 유사도 산출과 전처리**는 구분되어야 한다. 전처리(스키마 정렬, 개념 조화, 정합성 점검)는 두 자료가 같은 기준으로 비교될 수 있도록 **준비하는 과정**이다. 예를 들어, 변수명이 다르면 같게 맞추고, 코드 체계가 다르면 변환하며, 불가능한 값(예: 출생 연도 2020년, 연령 50세)을 점검한다. 하지만, 전처리가 끝나도 여전히 ‘Han Soonjin’과 ‘Han Sunjin’이 동일인인지 아닌지를 알 수 없다. 이에 각 레코드 쌍의 **유사성을 수치화**하여 이후 연계 방법이 작동할 수 있는 입력값을 제공하는 ‘비교 및 유사도 산출’이 필요하다.

33) 코사인 유사도는 두 문자열을 n -그램 벡터(n -gram vector; 문자열을 연속된 n 개의 글자 단위로 나눈 것)로 변환한 뒤, 두 벡터의 각도(cosine)로 유사도를 산출한다.

$$\cos(x, y) = \frac{\text{두 문자열이 공유하는 } n\text{-그램의 수}}{\sqrt{x\text{의 전체 길이}} \times \sqrt{y\text{의 전체 길이}}}$$

예) x : ‘대전정부청사’의 2-그램은 ‘대전’, ‘전정’, ‘정부’, ‘부청’, ‘청사’

y : ‘대전통계센터’의 2-그램은 ‘대전’, ‘전통’, ‘통계’, ‘계센’, ‘센터’

$$\Rightarrow \cos(x, y) = \frac{1}{\sqrt{5} \times \sqrt{5}} = 0.2$$

34) 최대·최솟값 기준으로 정규화: $z = \frac{|x_{\max} - x_{\min}|}{\sigma}$

예를 들어, 두 연령값이 각각 43세와 44세라면 절댓값 차이는 1이며, ±1세의 오차 범위를 허용하면 동일인으로 간주할 수 있다. 또한, 여러 변수를 동시에 고려해야 할 때 변수 간 상관관계를 반영하는 마할라노비스(Mahalanobis) 거리³⁵⁾를 사용한다.

- **범주형 변수**는 일치 여부를 비교적 명확히 판단할 수 있으나, 분류체계 내 위계가 존재하면 부분 일치를 고려해야 한다. 예컨대, 산업분류 코드가 각각 ‘24111’과 ‘24123’이면 세세 분류는 다르지만, 상위 소분류(‘241’)가 일치하므로 부분 일치로 처리할 수 있다. 이러한 부분 유사도는 전문가 판단이나 경험적 기준에 따라 점수를 부여한다.
- **지리 변수**는 오탈자나 행정구역 개편 등으로 직접 비교가 어려울 때는 좌표화(geocoding)하여 Haversine 거리³⁶⁾를 계산함으로써 유사도를 산출한다.

다. 연계 방법 선택

비교 및 유사도 산출을 거치면 각 레코드 쌍은 변수별 유사도 점수로 구성된 벡터를 갖게 된다. 연계 방법은 이러한 점수 벡터를 활용하여 동일 개체 여부를 추정한다(Christen, 2012; Herzog et al., 2007). 즉, **연계 방법은 “유사도 점수를 어떻게 활용해서, 동일성을 추정할 것인가?”**의 문제이다. 연계 방법은 크게 정확 연계, 결정적 연계, 확률적 연계³⁷⁾로 구분할 수 있다.

35) 변수 간 상관관계를 고려한 표준화된 거리 척도로 다변량 이상값 탐지, 수치형 변수 비교 등에 사용한다(De Maesschalck et al., 2000).

$$D^2 = (x_1 - x_2)^T \Sigma^{-1} (x_1 - x_2)$$

Σ^{-1} : 해당 변수 공분산 행렬의 역행렬

36) 위도·경도를 이용하여 두 점 간 구면 거리(great-circle distance)를 산출한다(Sinnott, 1984).

$$\text{Haversine distance} = 2R \arcsin \sqrt{\sin^2\left(\frac{\varphi_2 - \varphi_1}{2}\right) + \cos\varphi_1 \cos\varphi_2 \sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}$$

R : 지구 반지름

φ_1, φ_2 : 두 지점의 위도

λ_1, λ_2 : 두 지점의 경도

예) R_A 주소 좌표는 36.3623°N, 127.3560°E(대전정부청사)

R_B 주소 좌표는 36.3650°N, 127.3417°E(대전통계센터)

⇒ 두 지점 간 Haversine 거리는 약 1.3km로 동일 생활권에 있다고 볼 수 있다. 즉, 두 레코드는 ‘높은 유사도’를 보인다고 판정할 수 있다.

첫째, **정확 연계(exact linkage)**는 주민등록번호, 사업자등록번호 등 **고유식별자의 완전 일치**를 기반으로 동일 개체를 식별하는 방법이다(Christen, 2012; Cibella et al., 2009). 이론적으로 정확 연계는 연산량이 적고 오류가 거의 없다. 그러나, 실제로는 결측, 오자, 표기 방식 차이에 민감해, 동일 개체임에도 누락(false non-match)이 발생할 수 있다(Christen, 2012). 또한, 개인정보 보호 강화로 인해 고유식별자를 직접 활용하는 정확 연계는 법적·윤리적으로 제약받기도 한다(ESCAP, 2020).

$$EL_{i,j} = \begin{cases} 1, & \text{if } x_i = y_j \\ 0, & \text{otherwise} \end{cases}$$

$x_i = (x_1, \dots, x_i)$: 데이터 X 에서 레코드 i 의 고유식별자

$y_j = (y_1, \dots, y_j)$: 데이터 Y 에서 레코드 j 의 고유식별자

둘째, **결정적 연계(deterministic linkage)**는 고유식별자가 없거나 불완전할 때, 식별력이 높은 **공통 변수**의 유사도 점수를 합하거나 특정 규칙에 따라 조합하는 규칙 기반 방법이다(Christen, 2012; Harron, 2016). 구현이 간단하고 결과 해석이 명확하지만, 결측·오자·표기 차이에 취약하다(Christen, 2012). 또한, 가중값 설계가 경험적 판단에 의존하기 때문에 확률적 연계에 비해 주관적이고 유연성이 떨어진다.

$$DL_{i,j} = \sum_{r=1}^R w_r \gamma_{ij}^{(r)} \quad 38)$$

$\gamma^{(r)}$: 변수별 유사도 점수

w_r : 가중값($w_r \geq 0$)

$DL_{i,j}$: 연계 점수

셋째, **확률적 연계(probabilistic linkage)**는 Fellegi와 Sunter(1969)의 모형에 따라 레코드 쌍이 동일 개체일 확률을 계산한다(Christen, 2012; Harron, 2016). 공통 변수별 일치·불일치 확률을 통합해 동일성을 추정하기 때문에 데이터 불안정성에 유연하게 대응할 수 있다. 특히, 공통 변수별 신뢰도를 가중값으로 반영하므로 정확·결정적 연계보다 누락률이 낮다. 다만, 변수 간 독립성을 가정하기 때문에 변수 간 상관 구조를 충분히 반영하지 못하는 한계가 있다(Christen, 2012; Winkler, 2006, 2014;

37) 베이지안 연계와 기계학습 기반 연계는 확률적 연계의 확장 형태로 볼 수 있다.

38) 예 유사도 점수: 이름=0.95, 생년월일=0.99, 주소=0.98

⇒ 가중값: $w_{name} = 4.5, w_{birth} = 3.5, w_{adr} = 1.5$

⇒ 연계 점수: $DL_{i,j} = (4.5 \times 0.95) + (3.5 \times 0.99) + (1.5 \times 0.98) = 9.21$

Herzog et al., 2007).

$$w_r = \log \frac{m_r}{u_r}, \quad PL_{i,j} = \sum_{r=1}^R w_r I\{\gamma_{ij}^{(r)} = 1\} \quad 39)$$

$\gamma^{(r)}$: 변수별 일치 여부($\gamma^{(r)} = 1 \Leftrightarrow$ 일치, $\gamma^{(r)} = 0 \Leftrightarrow$ 불일치)

m_r : 동일 레코드 쌍에서 변수 r 이 일치할 확률

u_r : 비동일 레코드 쌍에서 변수 r 이 우연히 일치할 확률

w_r : 가중값(변수별 신뢰도)

$PL_{i,j}$: 연계 점수(레코드 쌍 유사도)

넷째, **베이지안 연계**(Bayesian linkage)는 레코드 쌍의 일치·불일치 사전확률(prior probability)을 기반으로 동일 개체일 사후확률(posterior probability)을 추정한다(Marchant et al., 2023; Tancredi & Liseo, 2011). 베이지안 연계는 결측과 오자가 많은 자료에 유용하며, 불확실성을 반영한 신뢰 구간 산출이 가능하다(Marchant et al., 2023). 또한, 다대다(many-to-many) 관계를 포함한 복잡한 구조에서도 유연하게 적용된다(Marchant et al., 2023; Tancredi & Liseo, 2011). 다만, 사전 정보가 충분하지 않으면 결과의 주관적 해석 여지가 커져 투명성과 설명 가능성(explainability)에서 제약이 발생할 수 있다(Liseo & Tancredi, 2009).

$$P(M | \gamma) = \frac{P(\gamma | M)P(M)}{P(\gamma | M)P(M) + P(\gamma | U)P(U)} \quad 40)$$

39) **예** 이름: $m_{name} = 0.95, u_{name} = 0.10$

생년월일: $m_{birth} = 0.99, u_{birth} = 0.01$

주소: $m_{adr} = 0.98, u_{adr} = 0.05$

\Rightarrow 가중값: $w_{name} = \log \frac{0.95}{0.10} \approx 2.25, w_{birth} = \log \frac{0.99}{0.01} \approx 4.60, w_{adr} = \log \frac{0.98}{0.05} \approx 2.99$

\Rightarrow 연계 점수: $PL_{i,j} = w_{name} + w_{birth} + w_{adr} \approx 9.84$

40) **예** 이름: $m_{name} = 0.95, u_{name} = 0.10$

생년월일: $m_{birth} = 0.99, u_{birth} = 0.01$

주소: $m_{adr} = 0.98, u_{adr} = 0.05$

사전확률: $P(M) = 0.01$ (전체 레코드 쌍 중 약 1%가 일치 쌍이라고 가정)

$\Rightarrow P(\gamma | M) = 0.95 \times 0.99 \times 0.98 \approx 0.922, P(\gamma | U) = 0.10 \times 0.01 \times 0.05 \approx 0.00005$

\Rightarrow 사후확률: $P(M | \gamma) = \frac{0.922 \times 0.01}{(0.922 \times 0.01) + (0.00005 \times 0.99)} \approx 0.9947$

$P(M)$: 두 레코드가 일치(match)일 사전확률

$P(U) = 1 - P(M)$: 두 레코드가 불일치(un-match)일 사전확률

γ : 변수별 일치·불일치 비교 결과(비교 벡터)

$P(\gamma|M)$: 동일 레코드 쌍일 때, 해당 비교 결과(γ)가 관측될 확률

$P(\gamma|U)$: 비동일 레코드 쌍일 때, 해당 비교 결과(γ)가 관측될 확률

다섯째, 기계학습 기반 연계(machine learning-based linkage)는 비교 단계에서 생성된 레코드 쌍의 유사도 점수 벡터를 분류 문제(classification problem)로 접근⁴¹⁾한다(Reich, 2024; Reich et al., 2024). 지도학습(supervised learning)은 이미 알려진 일치·불일치 레코드 쌍을 학습 데이터로 활용하고, 준지도학습(semi-supervised learning)이나 비지도학습(unsupervised learning)은 학습 데이터 없이 구조적 패턴을 학습한다(Reich et al., 2024). 기계학습 기반 연계는 전통적 확률 모형의 독립성 가정을 완화하고, 변수 간 비선형 관계까지 학습할 수 있다는 장점이 있다(Reich, 2024; Reich et al., 2024).

라. 연계 결정

연계 결정은 연계 방법에서 산출된 점수 또는 확률을 바탕으로 레코드 쌍을 ‘일치(match)’, ‘불일치(non-match)’, ‘불확실(clerical review/possible match)’로 분류하는 단계이다(Christen, 2012; Herzog et al., 2007; Winkler, 2014).⁴²⁾ 연계 방법이 “동일 개체인지를 어떻게 추정할 것인가?”에 초점을 둔다면, 연계 결정은 “추정된 결과를 바탕으로 어떻게 최종 판정할 것인가?”의 문제라 할 수 있다. 즉, 연계 방법은 동일성

41) 기계학습 기반 연계의 주요 분류 알고리즘은 다음과 같다.

- 랜덤 포레스트(random forest)는 다수의 결정트리(decision tree)를 무작위로 학습하고, 그 예측 결과를 집계(ensemble)하여 최종 판정을 내리는 기법이다. 변수 간 복잡한 상호작용을 반영할 수 있어 안정성이 높지만, 개별 변수의 영향력을 해석하기 어렵다(Breiman, 2001). 레코드 연계에서는 변수별 유사도 점수가 다양한 상호작용을 보일 때 특히 유용하다. 가령, 이름은 다르지만, 생년월일과 주소가 일치하는 경우, 단일 규칙으로는 동일성을 판정하기 어렵지만 여러 결정트리의 조합을 통해 안정적인 결과를 얻을 수 있다.
- 서포트 벡터 머신(support vector machine)은 고차원 공간에서 동일성과 비동일성을 가장 잘 구분하는 초평면(hyperplane)을 학습하는 알고리즘이다(Cristianini & Scholkopf, 2002). 이름 유사도와 주소 거리 등 복수의 비교 지표가 존재할 때, 두 집단 간의 경계를 최적화함으로써 판정의 일관성과 정확성을 높일 수 있다.
- 그래프 신경망(graph neural network)은 레코드 간 구조적 연결 관계를 학습하는 방법으로 다중 단위의 관계 구조를 동시에 고려할 수 있다(Wu et al., 2021). 예를 들어, 동일 가구에 속한 여러 가구원을 서로 연결된 네트워크로 표현하면, 개인 단위의 유사도뿐 아니라 가구 단위의 연결 정보까지 학습하여 동일성 여부를 판정할 수 있다.

42) 정확 연계는 고유식별자의 완전 일치를 기반으로 동일 개체를 식별하는 방식이기에 연계 방법과 연계 결정이 사실상 일치한다.

추정의 과정이고, 연계 결정은 그 추정 결과를 실제 연계 여부로 확정하는 절차이다. 연계 결정은 규칙 기반 결정과 확률 기반 결정 그리고 두 방법을 결합한 혼합 운용으로 구분한다.⁴³⁾

첫째, **규칙 기반 결정(rule-based decision)**⁴⁴⁾은 유사도 점수의 가중합 또는 특정 조건의 조합 결과가 임계값 이상이면 ‘일치’, 미만이면 ‘불일치’로 분류한다.

$$RD_{i,j} = \begin{cases} match, & \sum_{r=1}^R w_k \gamma_{ij}^{(r)} \geq T \\ non-match, & \sum_{r=1}^R w_k \gamma_{ij}^{(r)} < T \end{cases}$$

T : 임계값

or

$$RD_{i,j} = \begin{cases} match, & \sum_{r=1}^R w_k \gamma_{ij}^{(r)} \geq T_u \\ non-match, & \sum_{k=1}^K w_k \gamma_{ij}^{(r)} \leq T_l \\ Possible(clerical review), & T_l < \sum_{k=1}^K w_k \gamma_{ij}^{(r)} < T_u \end{cases}$$

T_u : 상한 임계값, T_l : 하한 임계값

규칙 기반 결정은 절차가 단순하고 기준이 명확해 재현성이 높으며, 연계 근거를 설명하기도 쉬워 빠른 의사결정이 필요한 경우에 적합하다(Christen, 2012). 다만, 임계값 설정에 따라 결과가 크게 달라질 수 있고, 임계값을 경험이나 직관에 의존하는 경향이 있어 주관이 개입될 여지가 있다. 또한, 데이터가 바뀔 때마다 임계값을 재설정해야 하고, 부분 일치(불확실 구간)를 세밀하게 반영하기 어렵다(Christen, 2012; Herzog et al., 2007; Winkler, 2014).

둘째, **확률 기반 결정(probabilistic decision)**⁴⁵⁾은 확률 모형 기반 연계에서 계산된 동일성 확률(연계 점수)에 사전 정의된 두 임계값(T_u, T_l)을 적용하여 판정한다. 즉, 상한값 이상이면 ‘일치’, 하한값 이하면 ‘불일치’, 중간 구간은 ‘불확실’로 분류한다.

43) 규칙 기반 결정과 확률 기반 결정의 기본 접근은 같다. 다만, **규칙 기반 결정은 사람이** 사전에 정한 점수 기준의 합산을, **확률 기반 결정은 통계 모형이** 산출한 값을 임계값과 비교한다.

44) 규칙 기반 결정은 결정적 연계와 함께 사용되기에 용어와 계산식이 혼용되기도 한다. 하지만, 결정적 연계는 연계 방법이고, **규칙 기반 결정은 동일성 여부를 판정하는 구체적인 절차**이다. 이에 본 연구에서는 두 개념을 혼동하지 않고 명확히 구분하여 기술한다.

45) 확률적 연계, 베이저안 연계, 기계학습 기반 연계에서 사용된다.

$$PD_{i,j} = \begin{cases} match, & \sum_{r=1}^R w_k I(\gamma_{ij}^{(r)} = 1) \geq T_u \\ non-match, & \sum_{r=1}^R w_k I(\gamma_{ij}^{(r)} = 1) \leq T_l \\ Possible(clerical review), & T_l < \sum_{r=1}^R w_k I(\gamma_{ij}^{(r)} = 1) < T_u \end{cases}$$

T_u : 상한 임계값, T_l : 하한 임계값

확률 기반 결정은 규칙 기반에 비해 연계 결정을 정량적 기준에 따라 수행할 수 있다는 장점이 있으며, 세밀한 확률 정보를 반영해 판단의 일관성을 높인다. 그러나, 실제 데이터에서는 변수 간 독립성 가정이 자주 깨지므로 이론적 전제의 충족이 어렵고(Christen, 2012; Winkler, 2006), 임계값 설정 역시 여전히 주관적 판단에 의존한다는 한계가 있다.

셋째, 혼합 운용(hybrid record linkage)은 여러 연계 방법(정확, 결정적, 확률적)과 결정 방법(규칙 기반, 확률 기반)을 단계적으로 결합하는 접근법이다. 각 방법의 장점을 통합하여 연계 효율성과 정확성을 동시에 확보하고자 할 때 활용되며, 특히 고정된 기준만으로는 판단이 어려운 상황에서 효과적이다.⁴⁶⁾ 혼합 연계의 절차는 다음과 같다.

- **1단계:** 「정확 연계」 또는 「결정적 연계+규칙 기반 결정」으로 명확히 일치하는 레코드 쌍 선별
- **2단계:** 남은 레코드 쌍에 대해 「확률적 연계+확률 기반 결정」으로 일치 쌍 선별
- **3단계:** 일부 애매한 쌍은 수작업 검토⁴⁷⁾ 또는 전문가 판단을 통해 보완

혼합 운용은 유연성과 범용성이 높지만, 연계 기준 간 충돌(conflict)이 발생하면 판단의 일관성을 해칠 수 있다(Christen, 2012). 또한, 단계별 결과가 문서화되지 않고, 검증 체계가 미흡할 경우 오류 추적이 어려워진다(Herzog et al., 2007).

46) Christen(2012)은 혼합 운용을 “sequential linkage approach”로 제안하였으며, Winkler(2006)는 “하나의 시스템에서 다양한 유형의 연계 방법을 사용하는 것은 결정적 규칙의 한계를 극복하는 데 도움이 될 수 있다.”라고 복수의 연계 결정 방식 적용을 권장하였다.

47) 수작업 검토는 비용이 많이 들고 시간이 소요되지만, 연계 품질을 확보하기 위한 **안전망**이다. 특히, 다양한 출처에서 수집된 반정형·비정형 자료는 자동화 알고리즘만으로는 오류를 줄이는 데 한계가 있다. 수작업 검토를 통해 모호한 사례를 직접 판정하면, 오류 누적을 방지하고 연계 결과에 대한 신뢰를 높일 수 있다.

<표 3-1> 확률 연계 - 확률 기반 결정에 따른 연계 결과표(Linkage Table) 예시

ID_A	이름_A	생년월일_A	주소_A	ID_B	이름_B	생년월일_B	주소_B	이름_점수 (JW score)	생년월일_점수	주소_점수	가중값	일치확률	연계 결정
A001	김○○	1993.07.07.	서울 강남구 역삼동	B103	김동은	1993.07.07.	서울 강남구 역삼동	0.99	1.00	0.98	8.54	0.992	Match
A002	박◎◎	1991.08.08	부산 해운대구 좌동	B250	박슬기	1991.08.09	부산 해운대구 좌1동	0.95	0.85	0.89	6.71	0.913	Match
A003	백●●	1992.09.09	대구 북구 태전동	B178	백승규	1992.06.09	대구 북구 태전동	0.97	0.45	0.95	4.32	0.610	Possible
A004	이●●	1992.10.10.	인천 남동구 구월동	N.A.	N.A.	N.A.	N.A.	—	—	—	—	—	Non-match
A005	최●●	1993.07.08.	대전 서구 월평동	B199	최하운	1993.07.08.	대전 서구 월평1동	0.78	1.00	0.70	3.87	0.525	Non-match

주: 상한 임계값 = 0.90, 하한 임계값 = 0.60

자료 출처: 저자 작성

<표 3-2> 연계 데이터 세트(Linked Data Set) 예시

ID	이름	생년월일	성별	주소	고용상태	직종	임금(원)
A001_B103	김○○	1993.07.07.	여	서울 강남구 역삼동	재직	사무직	3,200,000
A002_B250	박◎◎	1991.08.08	여	부산 해운대구 좌동	재직	생산직	3,800,00
A003_B178	백●●	1992.09.09	남	대구 북구 태전동	재직	서비스직	3,500,00
A004	이●●	1992.10.10.	남	인천 남동구 구월동	N.A.	N.A.	N.A.
A005	최●●	1993.07.08.	여	대전 서구 월평동	N.A.	N.A.	N.A.

자료 출처: 저자 작성

마. 연계 품질 점검

연계 품질 점검(linkage quality assessment)은 연계 결과의 정확성과 신뢰성을 정량적으로 평가하는 단계이다. 이를 위해 참값(ground truth)을 포함한 기준자료(gold standard data set)⁴⁸⁾를 확보한 후, 혼합 행렬(confusion matrix)을 기반으로 품질 평가 지표를 산출한다.

<표 3-3> 품질 점검 지표 산출을 위한 혼합 행렬

		실제	
		일치	불일치
예측	일치	TP _(true positive)	FP _(false positive)
	불일치	FN _(false negative)	TN _(true negative)

- **정확도(accuracy)**. 전체 레코드 쌍 중 정확하게 분류된 비율로 전체 품질을 개괄하는 지표이다. 다만, 불균형 데이터에서는 편향이 발생할 수 있어, 정밀도(precision)와 재현율(recall/sensitivity)을 함께 고려해야 한다.⁴⁹⁾

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- **정밀도(precision)**. 일치로 분류된 레코드 쌍 중 실제로 일치한 비율을 의미하며, 오연계를 얼마나 효과적으로 줄였는지를 보여주는 지표이다.

$$Precision = \frac{TP}{TP + FP}$$

- **재현율(recall)**. 실제 일치 쌍 중 ‘일치’로 올바르게 분류된 비율을 의미하며, 누락을 얼마나 잘 줄였는지를 보여주는 지표이다.

$$Recall = \frac{TP}{TP + FN}$$

- **F-1 점수(F-1 score)**. 정밀도와 재현율의 조화평균으로 두 지표 간 균형을 평가한다.

48) 현실에서는 기준자료가 없는 경우가 많아, 수작업 검토용 표본을 활용한 오류율 추정, 변수 간 논리적 일관성 분석, 연계 전후의 분포나 시계열 변화 비교 등의 방법도 고려해야 한다.

49) 일치하는 레코드 쌍이 1%인 경우, 모든 쌍을 불일치로 예측해도 정확도는 99%가 될 수 있다.

$$F-1 \text{ score} = 2 \times \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})}$$

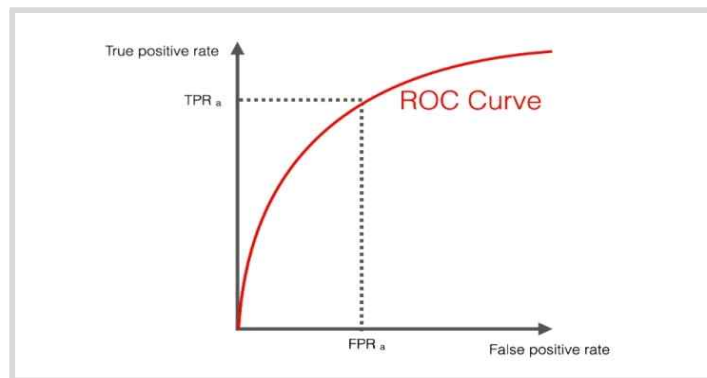
- **오연계율(false match rate, FMR)**. ‘예측 일치’ 중 잘못 연계된 비율로 개인정보 보호와 정확성 확보 측면에서 핵심적으로 관리해야 할 지표이다.

$$FMR = \frac{FP}{TP + FP}$$

- **누락률(false non-match rate, FNMR)**. ‘실제 일치’ 중 연계에 실패한 비율로 모집단 대표성을 위해 낮게 유지해야 한다(Christen, 2012).

$$FNMR = \frac{FN}{TP + FN}$$

- **ROC 곡선·AUC(receiver operating characteristic curve and area under the curve)**. 확률 기반 연계에서 다양한 임계값에 따른 위양성률(false positive rate, FPR)⁵⁰⁾과 진양성률(true positive rate, TPR)⁵¹⁾ 간의 균형을 시각화하여 연계 품질을 평가하는 지표이다.



자료 출처: encord(n.d.). ROC Curve

< 그림 3-3 > ROC 곡선

50) 실제 불일치를 일치로 잘못 분류한 비율, $FPR = \frac{FP}{FP + TN}$

51) 실제 일치를 누락시키지 않고 잡아낸 비율, $TPR = \frac{TP}{TP + FN}$

ROC 곡선은 연계 기준 변화에 따른 정밀도와 재현율의 관계를 나타내며, AUC⁵²⁾는 값이 1에 가까울수록 연계 정확도가 높음을 의미한다(Fawcett, 2006).

품질 점검 단계에서 오류가 발견되면, 그 결과는 반드시 수정(adjustment)하고 반영(implementation)해야 한다. 이는 품질 점검을 단순한 진단이 아니라 연계 품질을 지속해서 향상하는 환류 절차로 기능하게 한다는 점에서 중요하다. 오연계는 제거하고, 누락은 재식별·재연계함으로써 연계 데이터 세트의 정합성을 높이고, 집단 간 대표성 편차는 가중값 조정이나 표본 보정을 통해 교정함으로써 대표성을 확보해야 한다 (Christen, 2012).

연계 품질은 정량 지표만으로 충분히 평가하기 어렵기 때문에 수작업 검토 절차가 병행되어야 한다.⁵³⁾ 다만, 수작업 검토는 상당한 시간과 인력이 필요하므로, 본 연구에서는 검토 대상의 효율적 축소를 위한 전략을 제안한다.

- 연계 결과 중 일정 비율(예: 1~5%)을 무작위 추출하여 검토용 표본 구성
- 둘 이상의 검토자가 독립적으로 일치 여부를 판단하고, 불일치 사례에 대해서는 협의를 통해 최종 결정
- 검토 결과를 토대로 오류 유형(FP, FN 등) 및 발생원인 분석
- 분석 결과를 반영해 임곗값 또는 알고리즘 재조정

끝으로 연계 과정의 투명성, 품질 보장, 재현 가능성을 확보하기 위해 **메타데이터 기반의 품질 보고서(quality metadata report)**를 작성해야 한다. 이는 품질 점검 결과의 신뢰성과 투명성을 제도적으로 보증하는 장치이며, 산출된 연계 데이터 세트의 품질을 공식적으로 입증하는 역할을 한다.

$$52) AUC = \int_0^1 TPR(FPR^{-1}(x)) dx$$

1에 가까울수록 일치/불일치를 정확히 분류했음을 의미, 0.5는 무작위 분류와 같다고 간주한다.

53) Fellegi와 Sunter(1969)가 제시한 고전적 확률적 연계 이론에서는 연계 점수(score)가 임곗값 사이에 위치하는 경우를 ‘possible match’로 간주하고, 이를 수작업 검토 대상으로 분류하였다. Winkler(2006, p.3)는 이러한 “수작업 검토가 연계 오류를 최소화하는 데 중요한 수단”이 될 수 있음을 강조하였다. 또한 Christen(2012)은 수작업 검토의 효율성을 높이기 위해 도구와 시각화 방법의 개발, 검토 기준의 명확화가 필요함을 제안하였다.

2. 통계적 매칭

통계적 매칭(statistical matching)은 서로 다른 데이터들이 동일 변수를 직접 공유하지 않더라도, 이를 결합해 마치 하나의 데이터 세트처럼 활용할 수 있도록 하는 방법이다(D’Orazio et al., 2006; Rässler, 2002). 즉, 공통 보조 변수(auxiliary variables)를 매개로 미관측 변수(unobserved variables)를 추정하여, 합성 데이터 세트(synthetic data set)를 구축하는 절차이다. 레코드 연계가 실제 레코드를 결합해 연계 데이터 세트를 산출하는 것이라면, 통계적 매칭은 확률적 가정과 모형 기반 추정을 통해 존재하지 않는 합성 데이터 세트를 생성한다는 점에서 차별된다(D’Orazio et al., 2006; Rässler, 2002).

통계적 매칭은 접근 방법에 따라 합성 매칭(synthetic matching)과 모형 기반 매칭(model-based matching)으로 구분한다.⁵⁴⁾ **합성 매칭**은 추정된 결합확률분포를 이용해 주변 분포와 의존 구조를 반영한 합성 마이크로데이터를 생성한다. 이는 **새로운 가상 데이터**를 만드는 과정으로 볼 수 있으며, 가정에 따라 서로 다른 형태의 결과가 도출될 수 있다. 반면, **모형 기반 매칭**은 미관측 변수를 결측값으로 간주하고, 다른 데이터의 정보를 이용해 모형을 적합한 후 그 값을 **예측** 또는 **대체**하는 방법이다. 모형 기반 매칭에서는 실제 레코드가 채워져 분석에 직접 활용할 수 있는 완성 데이터 세트를 구축한다.

통계적 매칭(합성 매칭)의 기본 전제는 두 데이터가 공통 보조 변수를 공유하며, 이를 기준으로 각 데이터의 고유 변수 간 결합 관계가 독립적이라는 **조건부 독립성 가정**(conditional independence assumption, CIA)이 성립한다는 것이다(D’Orazio et al., 2006, 2024; Rässler, 2002). 예를 들어, 소득과 건강 상태 변수가 서로 다른 데이터에만 존재할 때, 성별이나 지역과 같은 공통 보조 변수를 기준으로 두 변수 간 결합 분포를 추정한다. 다만, 조건부 독립성 가정은 실제 데이터 환경에서는 자주 충족되지 않지만, 통계적 매칭을 통해 서로 다른 데이터 간 결합 분포를 추정하고 새로운 분석 가능성을 확보한다는 점에서 여전히 핵심 전제 조건으로 간주한다.

통계적 매칭은 일반적으로 다음 다섯 단계로 구성된다.

54) 본 절의 ‘2. 통계적 매칭’은 **합성 매칭**을 중심으로 다룬다. 이후 설명하는 「결합확률분포 추정 - 매칭 할당 - 합성데이터 구축」은 모두 합성 매칭의 절차이다. **모형 기반 매칭**은 실질적으로 결측값 대체(imputation)에 해당하므로, ‘제3절 결측값 대체’에서 다룰 예정이다. 따라서 현재 다루는 통계적 매칭은 좁은 의미에서 합성 매칭을 의미한다. 이러한 구분은 본 절의 논리 구조를 명확히 하고, 통계적 매칭과 결측값 대체 논의를 논리적으로 연결하는 역할을 한다.

- **1단계: 매칭 단위 설정.** 공통 변수 기준으로 매칭 단위(cell)를 설정하여, 후보군 제한
- **2단계: 결합확률분포 추정.** 공통 변수를 조건으로 변수 간 결합확률분포 추정
- **3단계: 매칭 할당.** 추정한 분포를 이용해 미관측 변수값을 예측·삽입
- **4단계: 합성 데이터 세트 구축.** 고유 변수를 결합하여 합성 마이크로데이터 세트 생성
- **5단계: 매칭 품질 점검.** 가정 충족 여부 검토 및 결합 결과의 타당성과 일관성 평가

가. 매칭 단위 설정

통계적 매칭의 출발점은 **매칭 단위 설정(definition of matching units)**이다. 이는 두 데이터가 공통으로 보유한 변수, 즉 공통 보조 변수를 기준으로 비교와 추정이 이루어질 수 있는 최소 분석 단위(cell)를 정의하는 과정이다(D’Orazio et al., 2006; Rässler, 2002). 매칭 단위 설정의 목적은 무의미한 전수 조합을 방지하고, 공통 보조 변수를 기준으로 의미 있는 비교가 가능한 집단 내부에서만 미관측 변수를 추정하도록 범위를 체계적으로 축소하는 데 있다. 이는 레코드 연계의 블로킹과 유사한 절차로, 효율성을 높이고 이질적 비교로 인한 왜곡을 줄이기 위한 단계이다. 예컨대, 데이터 A가 개인의 소득과 고용 정보를, 데이터 B가 교육과 건강 정보를 포함한다면, 성별·연령·지역과 같은 공통 보조 변수를 기준으로 ‘여자, 30~34세, 서울 거주’와 같은 매칭 단위를 설정할 수 있다. 이렇게 정의된 단위 내부에서만 매칭이 수행되므로, 서로 다른 데이터라도 모집단 특성이 유사한 집단(A, B)을 중심으로 비교할 수 있다.⁵⁵⁾ 다만, 매칭 단위를 지나치게 세분하면 단위별 표본 수가 급격하게 줄어드는 희소성 문제(sparsity)가 발생한다. 예를 들어, 성별, 연령, 지역, 학력, 직업 등 여러 변수를 동시에 조합하면 단위의 수가 과도하게 늘어나, 일부 단위에서는 표본이 거의 없거나 완전히 결측될 수 있다. 이를 방지하기 위해서는 ①변수를 적절히 병합하거나, ②연속형 변수는 구간화하거나, ③주요 변수만 선택하여 단위를 단순화하는 접근이 필요하다. 또한, 데이터 규모가 크거나 변수 조합이 복잡할 때는 블로킹(blocking) 기법을 활용할 수 있다. 이는 공통 보조 변수 중 일부만 먼저 사용해 비교 집단(블록)을 만들고, 그 블록 내부에서만 매칭을 수행하는 방식이다.⁵⁶⁾ 즉, 매칭 단위는 너무 넓으면 내부 이질성이 커져 조건부 독립성 가정이 약화하고, 너무 세분화하면 표본 부족이나 공백 셀이

55) 매칭 단위를 사전에 정의하면 계산 복잡도가 감소하고, 동일 배경·특성을 공유하는 집단 내에서 추정이 이루어지므로 조건부 독립성 가정의 현실 적합성이 강화된다(D’Orazio et al., 2006).

56) 예컨대 ‘성별+연령대’를 기준으로 1차 블록을 구성한 뒤, 그 내부에서 지역이나 학력 변수를 추가로 고려하여 매칭을 수행하는 방식이다.

발생하므로 데이터 규모, 공통 변수의 설명력, 분석 목적을 종합적으로 고려해 균형 있게 결정되어야 한다(Rässler, 2002).⁵⁷⁾

$$A_c = \{i \in A | Z_i = c\}, \quad B_c = \{j \in B | Z_j = c\}$$

Z : 공통 보조 변수 집합(예: 성별, 연령대, 지역 등)

c : 매칭 단위(예: 여자, 30~34세, 수도권)

설정된 매칭 단위의 타당성은 단위별 분포 일관성을 통해 점검할 수 있다.

$$Consistency \ Z = 1 - \frac{1}{2} \sum_{k=1}^K |P_A(Z = k) - P_B(Z = k)| \quad 58)$$

K : 공통 보조 변수의 범주 수(예: 연령대가 다섯 구간이면 $K = 5$)

$P_A(Z = k)$: 데이터 A 의 공통 보조 변수(Z)가 범주 k 에 속할 확률

$P_B(Z = k)$: 데이터 B 의 공통 보조 변수(Z)가 범주 k 에 속할 확률

나. 결합확률분포 추정

통계적 매칭은 동일 개체를 직접 식별할 수 없다는 제약에서 출발한다. 따라서, 관측되지 않은 변수 간의 관계를 **결합확률분포 형태로 추정함**으로써 새로운 데이터 세트를 생성한다(D’Orazio et al., 2006; Rässler, 2002). 즉, 통계적 매칭(특히, 합성 매칭)의 핵심은 관측되지 않은 두 변수 X, Y 의 결합확률분포($f(X, Y | Z)$)를 추정하는 데 있다. 결합확률분포는 다음 다섯 가지 접근으로 추정할 수 있다. 각 방법은

57) 매칭 단위를 결정하는 방식에는 여러 유형이 있다.

- 단순 층화(simple stratification): 1~2개의 공통 보조 변수로 단위를 정의하는 방식으로 단순하지만, 비교 가능성이 커지는 장점이 있다.

☞ 성별 × 연령대(5세 단위)

- 다단계 층화(multi-way stratification): 여러 개의 변수를 조합하여 더 세분된 단위를 정의한다.

☞ 성별 × 연령대(5세 단위) × 지역(시·도)

- 마이크로셀(micro-cell) 접근: 가능한 한 세밀하게 단위를 설정하는 방식으로, 고정밀 매칭이 가능하지만, 표본 수 부족과 공백 셀(empty cell) 발생 위험이 크다.

☞ 성별 × 연령 × 지역(시·군·구) = $2 \times 100 \times 250 = 50,000$

↳ ‘남자, 32세, 대전 서구’가 하나의 마이크로셀

- 셀 병합(cell collapsing): 공통 변수가 지나치게 세분되어 일부 단위에서 표본이 부족하거나 비어 있는 경우, 인접 범주를 병합하여 매칭 가능성을 확보한다(D’Orazio et al., 2006).

58) TVD(total variance distance; Villani, 2008)를 응용하여 본 연구에서 산출한 계산식으로 값이 1에 가까울수록 두 데이터의 공통 보조 변수 분포가 단위 수준에서 잘 맞는다.

단순성과 유연성, 가정의 강도와 불확실성 수준 사이에서 서로 다른 균형을 갖는다. 추정된 분포는 이후 합성데이터 세트 구축 단계에서 레코드 단위로 구현되며, 품질 점검과 보정을 거쳐 분석에 적합한 형태가 된다. 따라서 추정 방법 선택 시, 계산 편의성뿐만 아니라 데이터 특성, 분석 목적, 외부 정보의 활용 가능성을 종합적으로 고려해야 한다.

첫째, 조건부 독립성 가정은 “공통 보조 변수(Z)가 주어졌을 때, 데이터 A에만 존재하는 변수 X 와 데이터 B에만 존재하는 변수 Y 는 서로 독립이다.”라는 가정이다. 예컨대, 공통 보조 변수가 성별과 연령대라면, 같은 성별·연령대 집단 내에서는 소득(X)과 건강(Y)이 서로 영향을 주지 않는다고 가정하는 것이다. 비현실적이지만 계산이 단순하고, 가장 널리 사용된다(D’Orazio et al., 2006; Rässler, 2002).

$$f(X, Y|Z) = f(X|Z) f(Y|Z) \quad 59)$$

$f(X, Y|Z)$: Z 가 주어졌을 때, X, Y 의 결합확률분포

$f(X|Z)$: Z 에 조건화된 X 의 조건부 분포

(예: 성별·연령대가 같을 때의 소득 분포)

$f(Y|Z)$: Z 에 조건화된 Y 의 조건부 분포

(예: 성별·연령대가 같을 때의 건강 분포)

둘째, 셀 단위 조건부 추정(cell-level conditional distribution estimation)은 공통 보조 변수의 조합(예: 성별×연령대×지역)을 기준으로 매칭 단위를 정의하고, 각 단위 내에서 조건부 분포를 추정하여 결합 분포를 합성하는 방법이다(D’Orazio et al., 2006; Rässler, 2002). 셀 단위 조건부 추정은 동일 조건 내 집단 특성을 보존할 수 있다는 장점이 있으나(D’Orazio et al., 2006), 표본 수가 적은 단위에서는 추정값의 분산이 커져 불안정해질 수 있다. 이를 보완하기 위해, 희소 단위를 상위 집단과 병합하거나 베이지안 계층모형을 적용해 추정값을 안정화한다(Rässler, 2002).

$$f(X, Y|Z = c) = f(X|Z = c) f(Y|Z = c) \quad 60)$$

59) **예** 공통 보조 변수가 ‘30대 남자’일 때,

데이터 A(소득): ‘고소득(월 400만 원 이상)’일 확률 0.6

데이터 B(건강): ‘건강 양호(건강지수 80 이상)’일 확률 0.5

⇒ CIA를 적용: $P(\text{고소득, 건강양호} | 30대\ 남자) = 0.6 \times 0.5 = 0.3$

⇒ 합성된 데이터에서는 ‘30대 남자’의 약 30%가 고소득이면서 건강이 양호한 것으로 추정

60) **예** ‘여자, 30~34세, 수도권 거주’ 집단에서

데이터 A(변수 X): 평균 50, 표준편차 10 ($X \sim N(50, 10^2)$)

$f(X, Y|Z = c)$: $Z = c$ 일 때, X, Y 의 결합확률분포
 $f(X|Z = c)$: 단위 c 에 속하는 집단에서 X 의 조건부 확률분포
 $f(Y|Z = c)$: 단위 c 에 속하는 집단에서 Y 의 조건부 확률분포

셋째, 모형 기반 접근(model-based conditional distribution estimation)은 고유 변수 (X, Y) 가 공통 보조 변수(Z)에 따라 다변량 정규분포를 따른다고 가정하고, 평균과 분산·공분산 구조를 추정한다(D’Orazio et al., 2006; Rässler, 2002).

$$(X, Y)|Z \sim N(\mu(Z), \Sigma(Z)), \quad \Sigma(Z) = \begin{bmatrix} \sigma_X^2(Z) & \sigma_{XY}(Z) \\ \sigma_{XY}(Z) & \sigma_Y^2(Z) \end{bmatrix} \quad 61)$$

$\mu(Z)$: Z 에 따른 X, Y 의 평균 벡터
 $\sigma_X^2(Z), \sigma_Y^2(Z)$: X, Y 의 분산, $\sigma_{XY}(Z)$: X, Y 의 공분산

일반적으로 EM(expectation-maximization) 알고리즘을 활용하여 모수를 반복적으로 추정하는 방식으로 수행된다(Dempster et al., 1977). 그러나, 공분산 구조(X, Y 의 상관구조)가 완전히 식별되지 않으므로, 조건부 독립성 가정하에 추정 결과와 외부 자료를 이용한 보완 결과를 함께 제시해야 한다(Rässler, 2002).

넷째, 코플라 접근(copula approach)은 변수의 주변 분포(marginal distribution)와 의존 구조(dependence structure)를 분리하여 추정하는 방법이다(Nelsen, 2006). 즉, 각 변수의 분포 형태는 그대로 유지하되, 결합 구조만 별도의 함수로 하여 전체 결합 분포를 형성하는 방법이다. 이 방법은 정규성 가정을 벗어나 비정규·비선형 의존을 반영할 수 있으나(Nelsen, 2006), 의존 모수 추정을 위해서는 추가적인 가정이나 외부 자료가 필요하다.

데이터 B(변수 Y): 평균 30, 표준편차 5 ($Y \sim N(30, 5^2)$)

⇒ 합성 결합 분포: $(X, Y)|Z = x \sim N\left(\begin{bmatrix} 50 \\ 30 \end{bmatrix}, \begin{bmatrix} 100 & 0 \\ 0 & 25 \end{bmatrix}\right)$

⇒ ‘여자, 30~34세, 수도권 거주’ 집단에서 변수 X, Y 서로 독립적이므로, X 에 따라 Y 가 달라지는 확률구조는 반영되지 않음

61) 예) ‘여자, 30~34세, 수도권 거주’ 집단에서

데이터 A(변수 X): 평균 50, 표준편차 10, 분산 100 ($X \sim N(50, 10^2)$)

데이터 B(변수 Y): 평균 30, 표준편차 5, 분산 25 ($Y \sim N(30, 5^2)$)

⇒ $\mu(Z) = \begin{bmatrix} 50 \\ 30 \end{bmatrix}, \quad \Sigma(Z) = \begin{bmatrix} 100 & ? \\ ? & 25 \end{bmatrix}$

⇒ CIA를 적용하면 $\sigma_{XY}(Z) = 0$

⇒ 외부 자료의 상관계수가 0.3이면 $\sigma_{XY}(Z) = 0.3 \times 10 \times 5 = 15$

$$F_{X, Y|Z}(x, y | z) = C_z(F_{X|Z}(x | z) F_{Y|Z}(y | z)) \quad 62)$$

$F_{X, Y|Z}(x, y | z)$: $(X \leq x, Y \leq y)$ 일 누적확률분포

$F_{X|Z}(x | z)$: $X \leq x$ 일 조건부 누적확률분포

$F_{Y|Z}(y | z)$: $Y \leq y$ 일 조건부 누적확률분포

다섯째, 부분적 식별(partial identification approach)은 변수 간 “결합확률분포는 하나의 정확한 값으로 정해진다.”라고 가정하는 여타 방법과 달리 “결합확률분포를 정확히 알 수 없다.”라는 사실을 인정하고 가능한 확률 범위를 추정하는 방법이다(Manski, 2003; Moriarity & Scheuren, 2001).

$$\max\{0, p_X + p_Y - 1\} \leq P(X=1, Y=1 | Z) \leq \min\{p_X, p_Y\} \quad 63)$$

p_X : Z 가 주어졌을 때, $X=1$ 일 확률

p_Y : Z 가 주어졌을 때, $Y=1$ 일 확률

부분적 식별은 데이터 제약이 큰 상황에서도 가능한 결합확률의 경계(bound)를 제시해 추정 불확실성을 정량적으로 표현할 수 있는 장점이 있다.

62) 예 ‘여자, 30~34세, 수도권 거주’ 집단에 가우시안 코플라(Nelsen, 2006) 적용

데이터 A의 변수 X 가 x 이하일 확률(u): $F_{X|Z}(x | z) = 0.75$

데이터 B의 변수 Y 가 y 이하일 확률(v): $F_{Y|Z}(y | z) = 0.60$

의존 모수(ρ): 0.5

⇒ 가우시안 코플라: $C_\rho(u, v) = \Phi_\rho(\Phi^{-1}(u), \Phi^{-1}(v))$

$\Phi^{-1}()$: 표준정규분포의 역누적확률분포

$C_\rho(u, v)$: ρ 를 가진 이변량 정규분포의 누적확률분포

⇒ $\Phi^{-1}(0.75) = 0.674, \Phi^{-1}(0.60) = 0.253, \Phi_{0.5}(0.674, 0.253) \approx 0.67$

CIA일 경우 결합확률분포: $(0.75 \times 0.60) = 0.45$

⇒ ‘변수 X 가 x 이하이고 동시에 Y 가 y 이하일 조건부 확률’은 두 변수의 의존 모수가 0.5면 CIA보다 큼

63) 예 공통 보조 변수가 ‘30대 남자’일 때,

데이터 A(소득): ‘고소득(월 400만 원 이상)’일 확률 0.6

데이터 B(건강): ‘건강 양호(건강지수 80 이상)’일 확률 0.5

⇒ $\max\{0, (0.6 + 0.5) - 1\} \leq P(X=1, Y=1 | Z) \leq \min\{0.6, 0.5\}$

$0.1 \leq P(X=1, Y=1 | Z) \leq 0.5$

⇒ 30대 남자 중 고소득이면서 건강이 양호한 사람일 확률은 10%~50%

다. 매칭 할당

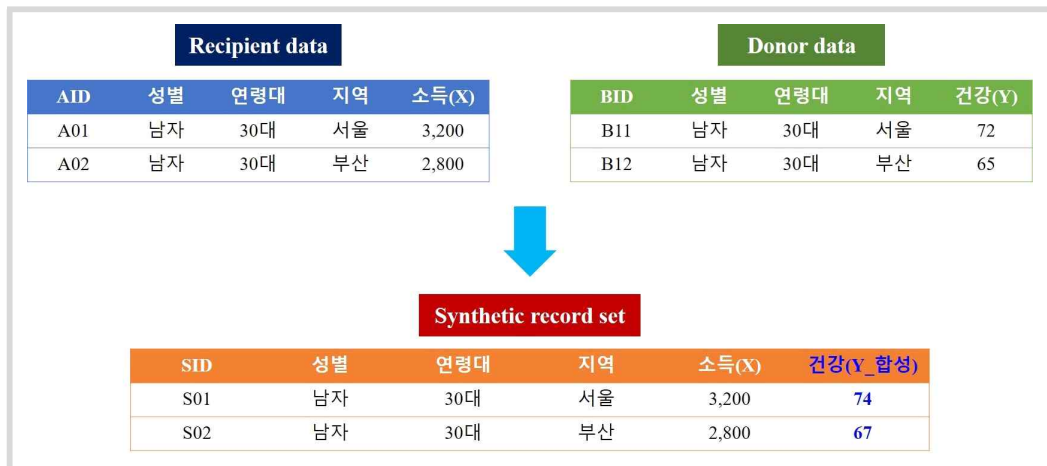
매칭 할당(matching allocation)은 추정된 결합확률분포를 각 레코드에 적용하여 변수 X 와 Y 의 값을 부여하는 과정이다. 이는 합성 매칭의 핵심 절차로, 추정된 결합 구조를 실제 합성 마이크로데이터로 구현하는 매개 단계에 해당한다. 반면, 모형 기반 매칭(=결측값 대체)은 예측 모형을 통해 변수를 직접 채워 넣기 때문에 별도의 할당 절차가 필요하지 않다. 따라서 매칭 할당은 합성 매칭과 모형 기반 매칭을 구분 짓는 중요한 기준이 된다. 매칭 할당은 “추정된 확률을 어떻게 실제 데이터값으로 전환할 것인가?”를 다루는 절차로, 결합확률분포의 추정 방법과 데이터 특성에 따라 여러 접근이 존재한다(D’Orazio et al., 2006; Rässler, 2002).

- **확률적 할당(stochastic allocation)**은 추정된 결합확률분포에 따라 난수를 이용해 X 와 Y 의 값을 무작위로 부여한다(D’Orazio et al., 2006; Rässler, 2002). 예를 들어, ‘30대 남성’ 집단에서 $P(X = \text{고소득}, Y = \text{건강 양호} | Z) = 0.3$ 이라면 해당 집단의 100명 중 약 30명을 무작위로 선택해 두 속성(고소득, 건강 양호)을 동시에 부여한다. 구현이 간단하고 분포 보존에 유리하지만, 난수 생성에 따라 매번 결과가 달라질 수 있다.
- **핫-덱 할당(hot-deck allocation)**은 공통 보조 변수가 같거나 유사한 집단의 변숫값을 ‘기증자(donor)’로 사용하여 레코드를 채우는 방법이다(Andridge & Little, 2010; D’Orazio et al., 2006). 예컨대, 소득만 있는 데이터 A에 건강 정보를 가진 데이터 B의 변숫값을 무작위로 부여하는 방법이다. 실제 변숫값을 사용하므로 해석이 직관적이지만, 표본이 작은 셀에서는 불안정할 수 있다.
- **최근접 이웃 할당(nearest neighbor allocation)**은 공통 보조 변수의 값이 가장 유사한 레코드를 찾아 결합하는 방법이다(D’Orazio et al., 2006). 예컨대, ‘30대 남성’ 집단 내에서도 학력·직업과 같은 추가 정보가 있다면, 이를 거리(예: 마할라노비스 거리, 유클리드 거리)로 계산해 가장 가까운 레코드를 결합한다. 유사성이 높은 레코드를 연결해 자연스러운 결합이 가능하지만, 거리 척도 설정에 따라 그 결과가 달라질 수 있다.
- **순위 기반 매칭(rank-preserving allocation)**은 각 변수의 값을 매칭 단위 내에서 순위화하고, 유사 순위의 레코드끼리 결합하는 방법이다(D’Orazio et al., 2006; Moriarity & Scheuren, 2001). 예컨대, 소득 상위 10% 레코드를 건강 상위 10% 레코드와 결합하는 방식이다. 순위 기반 매칭은 일관된 방향성을 가진 결합 구조(monotonic relationship)를 반영하는 데는 적합하지만, 분포의 불연속성이나 이질성이 충분히 반영되지 않을 수 있다.

라. 합성 데이터 세트 구축

합성 데이터 세트(synthetic data set)는 결합확률분포를 이용하여, 원래는 한 데이터에만 존재하던 변수를 다른 데이터에 부여함으로써 새로운 레코드를 생성한 데이터 세트를 말한다(D’Orazio et al., 2006; Rässler, 2002). 즉, 합성 데이터 세트는 단순히 두 데이터의 레코드를 직접 연결한 결과가 아니라, 공통 보조 변수에 조건화된 결합확률분포를 기반으로 새로운 변수값을 생성한다는 점에서 **합성(synthetic)**의 성격을 갖는다(D’Orazio et al., 2006; Rässler, 2002). 연계 데이터 세트가 “누가 누구와 연결되었는가?”라는 결합 정보(고유식별자, 공통 변수, 연계 레코드)를 보존한다면, 합성 데이터 세트는 이러한 정보를 남기지 않는다는 점에서 명확히 구분된다. 합성 데이터 세트 구축은 레코드 단위 결합(record-level integration)과 전체 데이터 세트 구축(full construction)의 두 단계로 구성된다.

첫째, 레코드 단위 결합은 매칭 할당 결과를 바탕으로 수혜 데이터(recipient data⁶⁴)의 레코드에 기증 데이터(donor data⁶⁵)의 변수(Y) 값을 부여하는 단계이다. 중요한 점은 Y 값을 그대로 가져오는 것이 아니라, 기증 데이터의 분포를 반영한 **결합확률분포($f(X, Y|Z)$)에 따라 새로운 값을 확률적으로 생성**한다는 것이다. 레코드 단위 결합의 결과, 변수 차원에서는 수혜 데이터의 변수 X 와 기증 데이터의 변수 Y 가 하나의 레코드에 공존하며, 행(row) 단위로는 두 데이터의 정보가 통합된 새로운 레코드가 형성된다. 이 과정을 도식화하면 <그림 3-4>와 같다.



자료 출처: 저자 작성

<그림 3-4> 레코드 단위 결합 예시

64) 수혜 데이터(recipient data)는 합성 데이터 세트의 **기본 틀(skeleton)**을 제공하는 데이터로, 공통 보조 변수 Z 와 자체 변수 X 를 포함한다.

65) 기증 데이터(donor data)는 수혜 데이터에 존재하지 않는 변수를 제공하는 데이터로, 공통 보조변수(Z)를 기준으로 연결되며, 확률결합분포에 따라 변수 Y 값이 부여된다.

둘째, 전체 데이터 세트 구축은 모든 매칭 단위(예: 성별×연령대×지역 조합)에 대해 레코드 단위 결합을 반복 수행하여, 최종적으로 하나의 완전한 합성 데이터 세트를 생성하는 단계이다. 이 단계에서 데이터별로 분리되어 있던 변수가 하나의 레코드에 통합되어 분석 가능성과 활용 범위가 확대된다. 다만, 합성 데이터 세트는 실제 변수값이 아니라, 조건부 분포 추정과 확률적 할당에 기반한 가상 데이터이므로 해석 시 그 한계와 불확실성을 고려해야 한다.

<표 3-4> 합성 데이터 세트(Synthetic Data Set) 예시

SID	성별	연령대	지역	소득(만 원)	건강(합성)
S01	남자	30대	서울	3,200	74
S02	남자	30대	부산	2,800	67
S03	여자	40대	서울	3,500	79
S04	여자	40대	부산	3,000	72

자료 출처: 저자 작성

마. 매칭 품질 점검

매칭 품질 점검(matching quality assessment)⁶⁶⁾은 합성 데이터 세트가 분석에 적합한 품질을 갖추었는지를 검증하는 단계이다. 통계적 매칭의 결과물은 동일 개체를 직접 연결한 데이터가 아니라, 확률적 가정에 기반해 생성된 합성 데이터이므로 품질 검증이 필수적이다(D’Orazio et al., 2006; Rässler, 2002). 품질 점검은 다음 다섯 가지 기준을 중심으로 수행한다.

- **공동 지지영역(common support)**⁶⁷⁾ 두 데이터의 공통 보조 변수값이 실제 **겹치는 영역**에서만 매칭이 이뤄졌는지를 확인한다(Stuart, 2010). 합성 데이터 세트는 공동 영역 내에서만 신뢰할 수 있기에 데이터 A가 $Z = z_1$, 데이터 B가 $Z = z_2$ 만 가진다면, X 와 Y 의 결합 추정은 불가능하다.

66) 합성 데이터 세트의 품질은 이후 ‘보정 및 대표성 점검’에서 계량적으로 진단하고, 문제점은 보정 절차를 통해 수정한다.

67) **예** 데이터 A: 성별 = **남자/여자**, 연령대 = 20대, **30대**, 40대

데이터 B: 성별 = **남자/여자**, 연령대 = **30대**, 40대, 50대

⇒ 합성 데이터 세트는 30대, 40대에서만 신뢰할 수 있다.

- **조건부 주변 확률분포 보존(margin preservation).**⁶⁸⁾ 합성 데이터 세트는 데이터 A와 B의 조건부 주변 확률분포를 유지해야 한다(D’Orazio et al., 2006). 매칭 과정에서 변수 X 와 Y 의 분포가 크게 왜곡되면, 추정된 결합 구조의 신뢰성이 저하된다(D’Orazio et al., 2006; Little & Rubin, 2019).
- **내적 일관성(internal consistency/coherence).** 합성 데이터 세트 내 값들은 논리적으로 일관되어야 한다. 예컨대, ‘성별=남자, 임신=예’ 또는 ‘가구원 수=1, 자녀 수=3’과 같은 조합은 실제 존재할 수 없는 값이다. 이러한 불일치는 매칭 과정의 오류나 추정 모형의 부적절한 적용으로 인해 발생한다.
- **대표성 및 편향.** 합성 데이터 세트는 모집단의 특성을 대표해야 한다. 매칭 과정에서 특정 집단의 비중이 과대 또는 과소 반영되면, 분석 결과에 편향(bias)이 발생한다. 따라서 합성 데이터 세트의 분포가 외부 기준자료(예: 통계등록부, 행정자료, 대규모 조사자료 등)와 일치하는지를 반드시 점검⁶⁹⁾해야 한다(D’Orazio et al., 2006; Little & Rubin, 2019).
- **민감도 분석(sensitivity analysis).** 합성 데이터 세트는 기본 가정에 민감하므로 그 가정의 변화에 따른 결과 변동성을 평가해야 한다(Rässler, 2002). 즉, 조건부 독립성 가정, 매칭 단위 설정, 분포 추정 방법 등이 달라질 때, 합성 데이터 세트의 구조나 통계적 결과가 어떻게 변하는지를 비교⁷⁰⁾함으로써 불확실성을 정량적으로 검증한다.

68) 조건부 주변 확률분포 보존은 절대 차이, χ^2 검정 등으로 점검한다.

• 절대 차이: $\Delta_X(z) = |P(X|Z=z) - P^*(X|Z=z)|$

$P(X|Z)$: 데이터 A에서 X 의 조건부 주변 확률분포

$P^*(X|Z)$: 합성 데이터 세트에서 X 의 조건부 주변 확률분포

☞ 데이터 A에서 ‘30대 남자’의 고소득 비율: $P(X = \text{고소득} | Z = \text{30대 남자}) = 0.40$

합성 데이터 세트에서 ‘30대 남자’의 고소득 비율: $P^*(X = \text{고소득} | Z = \text{30대 남자}) = 0.43$

⇒ $\Delta_X(z) = 0.03(3\%)$ 으로 허용 오차($\pm 5\%$) 범위 내이므로 조건부 주변 확률분포 보존

69) ☞ 인구총조사: 30대 남자 서울 거주 = 15%

합성 데이터 세트: 30대 남자 서울 거주 = 18%

⇒ ‘30대 남자 서울 거주’ 집단에서 3%p 차이 발생

⇒ χ^2 검정 결과 대표성 차이가 통계적으로 유의

70) ☞ 가우시안 코플라에서 의존모수(ρ)를 다르게 설정해서 비교

변수 X 가 x 이하일 확률(u)=0.5, 변수 Y 가 y 이하일 확률(v)=0.5

• 의존모수(ρ)가 0(두 변수가 서로 독립적)이면, $\Phi_{0.0}(0.5, 0.5) = 0.25$

• 의존모수(ρ)가 0.3이면, $\Phi_{0.3}(0.5, 0.5) \approx 0.299$

• 의존모수(ρ)가 0.5면, $\Phi_{0.5}(0.5, 0.5) \approx 0.333$

3. 혼합 매칭

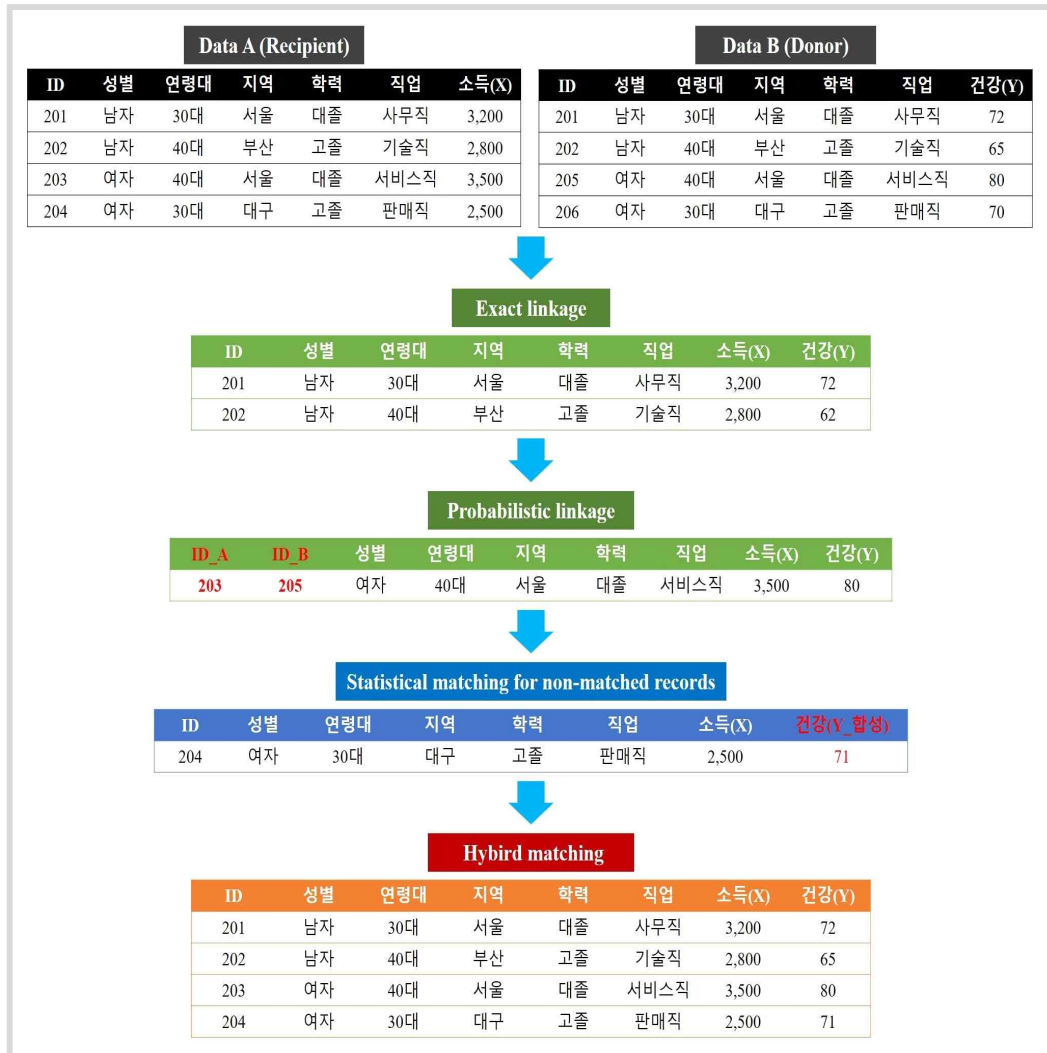
혼합 매칭(hybrid matching)⁷¹⁾은 레코드 연계와 통계적 매칭을 상호보완적으로 결합한 접근 방법이다. 현실의 데이터 환경에서는 고유식별자나 공통 변수가 충분하지 않아 모든 데이터를 완전하게 결합하기는 어렵다. 이에 레코드 연계를 통해 가능한 범위까지 동일 개체를 정확하게 연결하고, 남아 있는 비연계(non-linked) 레코드는 통계적 매칭으로 보완함으로써, 전체적으로 완전한 데이터 세트를 구축할 수 있다(D’Orazio et al., 2006; Rässler, 2002). 다시 말해, 혼합 매칭은 레코드 연계가 제공하는 높은 정확성(accuracy)과 통계적 매칭이 제공하는 데이터의 완전성(completeness)을 결합하여 데이터 통합의 실효성을 극대화하는 방법이다. 혼합 매칭의 절차는 다음과 같이 구성된다.

- **1단계: 정확 연계 또는 결정적 연계.** 고유식별자나 공통 변수를 이용하여 레코드 연계를 수행하여, 연계 데이터 세트를 생성한다.
- **2단계: 확률적 연계.** 고유식별자가 불완전하거나 불확실한 경우, 확률 모형 기반 연계를 통해 동일 개체일 확률을 추정하고 연계 여부를 결정한다.
- **3단계: 통계적 매칭.** 여전히 연계되지 않은 레코드에 대해서는 공통 보조 변수를 기준으로 결합확률분포를 추정한다. 수혜 데이터와 기증 데이터를 설정한 후, 수혜 데이터의 각 레코드에 기증 데이터의 변수를 확률적으로 부여한다.
- **4단계: 통합(integration).** 레코드 연계와 통계적 매칭 과정을 거친 데이터를 최종 데이터 세트(hybrid data set)로 통합한다. 레코드 연계에서 얻은 실제 값은 최대한 유지하고, 통계적 매칭을 통해 생성된 합성값(synthetic value)은 불가피한 결측 보완으로 취급한다.
- **5단계: 혼합 매칭 품질 점검.** 혼합 매칭은 레코드 연계와 통계적 매칭을 결합한 복합적 절차이므로, 품질 점검 또한 두 접근의 점검 요소를 함께 포함한다. ①레코드 연계 단계에서는 정확도, 정밀도, 재현율 등을 적용하고, ②통계적 매칭 단계에서는 합성 데이터 세트의 분포 보전과 대표성 유지 여부를 검증한다. ③최종 혼합 데이터 세트에서는 연계 값과 합성 값 간의 내적 일관성과 대표성을 점검한다. 또한 동일 매칭 단위 내에서 두 값의 분포 차이가 과도하지 않은지를 **균질성 검사(homogeneity check)⁷²⁾**로 확인한다.

⇨ 주변 확률분포가 같더라도 결합 구조(코플라)가 바뀌면 동시 발생 확률은 달라짐

⇨ 매칭 결과의 공동사건 비율이 가정 변화에 얼마나 민감한지 파악할 수 있으며, 이는 합성 데이터 세트의 품질과 신뢰성에 직결

71) 혼합 매칭(hybrid matching)은 **혼합 통계적 매칭(mixed statistical matching)** 또는 **보완적 연계 및 매칭(complementary linkage and matching)**으로도 불린다(Christen, 2012).



자료 출처: 저자 작성

<그림 3-5> 혼합 매칭 과정 예시

72) 균질성 검사는 χ^2 검정, t -검정, Kolmogorov-Smirnov 검정, KL 발산 등으로 확인할 수 있다.

• Kullback-Leibler divergence(Cover & Thomas, 2006): 0~1의 값으로, 0에 가까울수록 동일 분포

$$D_{KL}(P_L \parallel P_S) = \sum_y P_L(y|z) \ln \frac{P_L(y|z)}{P_S(y|z)}$$

$P_L(y|z)$: 연계 데이터에서 매칭 단위 $Z = z$ 일 때, 범주 y 에 속하는 조건부 확률분포

$P_S(y|z)$: 합성 데이터에서 매칭 단위 $Z = z$ 일 때, 범주 y 에 속하는 조건부 확률분포

예) 건강 변수의 구간 분포가 $P_L[0.2, 0.5, 0.3]$, $P_S[0.25, 0.45, 0.3]$

$$D_{KL}(P_L \parallel P_S) = 0.2 \ln \frac{0.2}{0.25} + 0.5 \ln \frac{0.5}{0.45} + 0.3 \ln \frac{0.3}{0.3} = 0.008$$

⇒ 두 분포가 매우 유사함

제3절 결측값 대체

결측값 대체(imputation)는 통합 데이터의 완결성과 내적 일관성을 확보하기 위한 절차이다. 일반적으로 결측값 대체는 관측되지 않은 항목을 추정이 가능한 값으로 치환하는 과정을 의미한다(Little & Rubin, 2019). 이는 단순히 빈칸을 채우는 것이 아니라, 관측된 변수 간의 구조와 관계를 활용하여 ‘관측되었을 법한 값(plausible value)’을 산출하는 과정이다. 이러한 접근은 단일 데이터 내 결측뿐 아니라, 데이터 통합 과정에서 발생하는 변수 누락, 불일치, 충돌의 보완에도 적용된다.

행정자료, 조사자료, 빅데이터 등 다양한 출처의 데이터는 수집 목적과 구조가 각각 다르므로 결합 과정에서 나타나는 결측이나 불일치는 불가피하다. 결측값을 그대로 두면 대표성과 신뢰성이 훼손될 뿐만 아니라, 추정 결과에도 체계적인 편향이 발생할 수 있다. 따라서 결측값 대체는 데이터 통합의 불완전성을 해소하고, 결과적으로 목적에 맞는 통합 데이터 세트를 구축하기 위한 필수 단계라 할 수 있다.

결측값 대체는 데이터 통합의 두 가지 방법론적 틀과 밀접히 연결된다.

첫째, 데이터 융합(data fusion)은 서로 다른 데이터를 일관된 데이터 세트로 조정하는 과정으로, 결측 대체는 이 과정에서 누락되거나 충돌한 값을 보완하는 실질적 수단으로 기능한다. 예컨대, 동일 변수가 행정자료와 조사자료 모두에 존재하지만, 일부 값이 결측되거나 불일치할 경우, 이를 보완하는 절차는 곧 데이터 융합이자 결측값 대체의 구현이다(Boonstra & del Pino, 2025; de Waal et al., 2019). 즉, 데이터 융합은 결측값 대체의 개념적 틀을 제공하고, 결측값 대체는 데이터 융합의 방법론적 실행 단계로 작동한다.

둘째, 통계적 매칭 중 모형 기반 매칭(model-based matching)은 결측값 대체를 구체적 방법론으로 실현한다. 공통 보조 변수를 매개로 두 데이터의 관계를 모형화하고, 이를 통해 비관측 변수를 추정·보완하는 절차는 사실상 결측값 대체와 동일한 논리 구조를 가진다(D’Orazio et al., 2006; Rässler, 2002). 예를 들어, 데이터 A에 ‘연령·소득’ 변수, 데이터 B에 ‘연령·건강’ 변수가 있다면, 공통 보조 변수인 연령을 매개로 소득과 건강의 관계를 추정해 데이터 A의 결측 변수를 보완할 수 있다.

요컨대, 결측값 대체는 개념적으로는 데이터 융합의 일부로서 결측과 충돌을 해소하는 과정이며, 방법론적으로는 모형 기반 매칭을 통해 구현되는 절차이다. 이 두 관점이 결합할 때, 결측값 대체는 통합 데이터 세트의 완결성과 신뢰성을 동시에 보장하는 핵심 기제로 작동한다.

1. 결측 유형

결측이 발생하는 메커니즘은 세 가지 유형으로 구분된다(Rubin, 1987). 이러한 구분은 단순한 이론적 구별이 아니라, 적절한 결측값 대체(imputation) 방법 선택의 근거로 기능한다.

첫째, 완전 무작위 결측(missing completely at random, MCAR)은 결측이 데이터 내 어떤 변수와도 관련 없이 전적으로 무작위로 발생하는 경우이다(Rubin, 1987). 예를 들어, 조사 과정에서 일부 설문지가 분실되거나 행정자료 전송 과정에서 기술적 오류로 일부 레코드가 누락되는 사례가 이에 해당한다. MCAR은 특정 속성과 무관하게 발생하므로, 데이터 결합 과정에서 구조적 편향을 유발하지는 않는다. 주민등록부와 세무 자료를 결합할 때 단순 전송 오류로 소득 항목 일부가 누락된 상황은 개인의 소득 수준이나 연령과는 무관하게 발생한 것이므로 MCAR에 해당한다. 이런 경우에는 단순 대체나 평균 대체도 허용될 수 있지만, 정보 손실을 최소화하기 위해 **다중 대체(mass imputation)**가 권장된다(UNECE, 2020).

둘째, 무작위 결측(missing at random, MAR)은 결측이 해당 변수 자체와는 무관하나, 다른 관측 변수와는 체계적 관계가 있다(Little & Rubin, 2019). 예컨대, 고령층의 전자 세무 신고율이 낮아 소득자료의 소득정보가 일부 누락된다면, 결측은 소득 변수와는 직접 관련되지 않지만, 연령 변수와는 밀접한 상관이 있다. 또한, 조사에서 연령대가 높을수록 소득 문항에 응답하지 않는 비율이 증가한다면, 결합 이후에도 소득 변수의 결측이 특정 연령대에 집중적으로 나타날 수 있다. 이러한 상황에서는 연령·성별과 같은 공통 보조 변수를 활용한 **모형 기반 매칭이 효과적이다**(D’Orazio et al., 2006).

셋째, 비무작위 결측(missing not at random, MNAR)은 결측이 바로 그 변수의 실제 값과 직접적으로 연관된 경우이다(Little & Rubin, 2019). 가령, 고소득자가 소득을 의도적으로 응답하지 않거나, 흡연자가 흡연 여부를 기재하지 않은 경우, 결측 발생 자체가 변수의 특성과 긴밀히 연관되어 있다. 이때 단순 대체나 회귀 기반 대체는 결측 메커니즘을 반영하지 못해 왜곡을 초래할 수 있다. 따라서, MNAR에서는 베이저안 접근이나 기계학습 기반 접근처럼 **결측 발생 메커니즘 자체를 모형화**해야 한다.

요약하면, “MCAR은 무작위 오류에 가까워 단순한 대체도 가능하지만, MAR은 관측 변수와의 관계를 모형화해야 하며, MNAR은 결측 발생 메커니즘 자체를 함께 추정해야 한다.” 데이터 통합 과정에서는 특히 MAR과 MNAR 유형의 결측이 빈번하게 나타나므로, 레코드 연계와 통계적 매칭 이후에도 결측값 대체가 독립된 절차로

수행된다. 결측 메커니즘에 대한 정확한 이해는 곧 통합 데이터 세트의 품질과 신뢰성을 좌우하는 핵심 요건이라 할 수 있다.

2. 결측값 대체 방법

결측값 대체는 적용 복잡성과 정교성에 따라 여러 가지 방법으로 구분된다. 가장 단순한 형태는 평균 등 대푯값으로 결측을 치환하는 단순 대체이며, 불확실성을 반영한 다중 대체, 통계 모형을 활용하는 모형 기반 대체 그리고 최근에는 기계학습 알고리즘을 적용하는 접근으로 발전해 왔다. 데이터 통합의 맥락에서 결측 발생 메커니즘과 데이터의 특성에 따라 적절한 방법을 선택해야 한다. 특히, 레코드 연계나 통계적 매칭을 거친 이후에 남아 있는 결측은 변수 간 구조를 보존하면서 불확실성을 반영할 수 있는 다중 대체나 모형 기반 대체가 효과적이다(D’Orazio et al., 2006; Rässler, 2002).

가. 단순 대체

단순 대체(simple imputation)는 결측값을 평균, 중앙값, 최빈값 또는 유사 사례의 관측된 값으로 채우는 가장 기본적인 방법이다. 예컨대, 가계조사에서 일부 응답자의 주거비가 누락된 경우, 동일 지역 가구의 평균 주거비로 대체하거나 유사한 소득 수준의 다른 가구 응답을 적용할 수 있다. 구현이 간단하고 계산이 빠르지만, 불확실성을 반영하지 못하고 변수 간의 상관 구조를 왜곡할 수 있는 한계가 있다(Little & Rubin, 2019). MCAR 상황에서는 비교적 유용하지만, 체계적인 결측(MAR, MNAR)에서는 편향을 초래할 수 있다.

- **평균 대체**(mean substitution). 동일 집단의 관측값 평균으로 결측값을 대체한다. 단순하고 직관적이지만, 분산이 과소 추정되고, 다른 변수와의 관계가 반영되지 않기에 통합 데이터 세트의 상관관계를 왜곡할 수 있다(Little & Rubin, 2019; Rubin, 1987).
- **중앙값·최빈값 대체**(median·mode substitution). 연속형 변수에는 중앙값, 범주형 변수에는 최빈값을 사용한다(de Waal et al., 2019). 극단값의 영향을 줄여 분포 왜곡을 완화하지만, 평균 대체와 마찬가지로 분산의 과소 추정과 상관관계 왜곡은 여전히 존재한다(Lohr, 1999).
- **선형 보간**(linear interpolation).⁷³⁾ 시계열 데이터에서 인접 시점 값을 이용해 결측을 추정하는 방법이다(Dagum & Cholette, 2006; Meijering, 2002). 선형 보간은

직관적이고, 단기간 결측에는 유용하지만, 구조적 단절이나 급격한 변동(예: IMF 위기, COVID-19 등)을 반영하기는 어렵다(Hyndman & Athanasopoulos, 2021).

- **핫-덱 대체(hot-deck imputation)**. 결측이 있는 레코드와 유사한 특성을 가진 기증자(donor)의 값을 무작위로 선택해 대체한다(Andridge & Little, 2010). 예를 들면, 동일 성별·연령·지역의 개체 중에서 소득 값을 차용하는 방식이다. 분포 보존이 가능하며, 무작위 핫-덱(random hot-deck)을 적용하면 불확실성도 반영할 수 있다. 다만, 유사도 선정 기준이나 표본 크기에 따라 결과의 안정성이 달라질 수 있다.
- **콜드-덱 대체(cold-deck imputation)**. 외부 기준자료(예: 통계등록부, 행정자료, 대규모 조사자료 등)의 값을 이용하는 방법이다(Brick & Kalton, 1996). 외부 기준자료의 시점 불일치, 품질 수준, 정의 차이를 충분히 고려해야 하며 불일치가 큰 경우 오히려 새로운 편향을 초래할 수 있다.

단순 대체는 데이터 통합의 초기 진단 단계나 완전 무작위 결측(MCAR)에서는 효율적이지만, 불확실성을 반영하지 못하고 분산을 과소 추정하는 한계로 인해 품질 점검 및 보정 단계를 반드시 병행해야 한다.

나. 모형 기반 대체

모형 기반 대체(model-based imputation)⁷⁴⁾는 변수 간의 관계를 통계 모형으로 추정하여 결측을 대체하는 방법이다(Little & Rubin, 2002; Rässler, 2002). 즉, 관측된 변수로부터 결측 변수의 값을 설명할 수 있는 ①관계를 모형화하고, ②모수(parameter)를 추정한 뒤, 그 ③추정 결과를 활용하여 결측을 대체한다. 모형 기반 대체는 변수 간 공분산 구조와 상관관계를 반영하므로 단순 대체보다 정보 손실이 적고, 추정 효율이 높다. 또한, 베이지안 접근을 적용하면 결측 처리 과정에서 불확실성을 자연스럽게 반영할 수 있다. 다만, 모형이 잘못 설정되면 편향이나 왜곡이 심화할 수 있으며, 특히 비무작위 결측(MNAR) 상황에서는 결측 발생 메커니즘 자체를 모형화하지 않으면 타당한 추정을 얻기 어렵다.

- **회귀 대체(deterministic regression imputation)**.⁷⁵⁾ 결측(Y)과 관측 변수(X)의 관계를

$$73) X_i^{miss} = X_{i-1} + \frac{X_{i+1} - X_{i-1}}{2}$$

74) 모형 기반 대체는 통계적 매칭의 모형 기반 접근(model-based matching)과 직접적으로 맞닿아 있다.

$$75) Y_i^{miss} = \hat{\beta}_0 + \sum_{p=1}^P \beta_p X_{iP}$$

회귀식으로 추정하는 방법이다(Little & Rubin, 2002). 공분산 구조를 반영하므로 평균 대체보다 현실적인 값을 생성하지만, 모든 대체 값이 회귀선 위에 위치하기에 분산이 과소 추정되는 한계가 있다.

- **확률적 회귀 대체(stochastic regression imputation).**⁷⁶⁾ 회귀 예측값에 무작위 오차항을 더해 분산 축소 문제를 보완하는 방법이다(Little & Rubin, 2002). 대체 값 간의 변동을 허용해 불확실성을 반영할 수 있고, 다중 대체와 결합하면 효과적이다. 그러나, 난수 발생으로 인해 재현성이 떨어질 수 있으며, 오차항의 분포를 잘못 추정하면 비현실적인 값이 생성될 수 있다.
- **다변량 정규 모형 기반 대체(multivariate normal imputation).**⁷⁷⁾ 데이터 전체를 다변량 정규 분포로 가정하고, 관측값을 기준으로 결측값의 조건부 확률분포를 계산하여, 그 분포에서 난수를 추출해 대체하는 방법이다(Little & Rubin, 2002; Schafer, 1997). 공분산 구조를 보존하고, 고차원 데이터에도 적용할 수 있으나, 실제 데이터가 정규분포를 따르지 않으면 왜곡이 발생하며, 비무작위 결측(MNAR) 상황에서는 부적절하다.
- **베이저안 대체(Bayesian imputation).**⁷⁸⁾ 결측값의 모수를 확률변수로 전제하고,

$$76) Y_i^{miss} = \hat{Y}_i + u_i, \quad u_i \sim N(0, \hat{\sigma}^2)$$

$$77) X_{miss} | X_{obs} \sim N(\mu_{m|o}, \Sigma_{m|o})$$

$$\mu_{m|o} = \mu_m + \Sigma_{mo} + \Sigma_{oo}^{-1}(X_{obs} - \mu_o), \quad \Sigma_{mo} = \Sigma_{mm} - (\Sigma_{mo}\Sigma_{oo}^{-1}\Sigma_{om})$$

$\mu_{m|o}$: 조건부 평균, $\Sigma_{m|o}$: 조건부 분산

예) 평균: 소득 = 3,200, 교육 = 15. 건강 = 78

ID = 2: 소득 = 3,000, 교육 = 16. 건강 = N.A.

⇒ 조건부 평균 건강 = 77.7, 조건부 표준편차 = 9

⇒ 결정적 대체 시: 77.7, 확률적 대체 시: $77.7 \pm (\text{난수} \times 9)$

$$78) Y_i^{miss} = p(Y | X, \theta), \quad \theta \sim p(\theta | Y_{obs})$$

$p(Y | X, \theta)$: 조건부 분포, θ : 모수, $p(\theta | Y_{obs})$: 사후분포

예) 모형: $Y | X, \theta \sim N(\beta_0 + \beta_1 X, \sigma^2), \quad \theta = (\beta_0, \beta_1, \sigma^2)$

사전분포: $\beta_0 \sim N(1000, 200^2), \beta_1 \sim N(200, 50^2), \sigma^2 \sim Ga^{-1}(2, 1000)$

사후분포: $\beta_0 \approx 1200, \beta_1 \approx 180, \sigma^2 \approx 200$

IF 나이 = 30, 소득 = N.A.

조건부 평균($\beta_0 + \beta_1 X$): $\mu_{Y|나이=30} = 1200 + (180 \times 30) = 6600$

⇒ 결측값 생성: $Y | 나이_{30}, 200 \sim N(6600, 200^2)$

사전분포를 바탕으로 산출한 사후분포에서 결측값을 추출하는 방법이다(Rässler, 2002; Schafer, 1997). 불확실성을 자연스럽게 반영하고, MNAR 메커니즘까지 모형화할 수 있지만, 사전분포 설정에 민감하고 결과 해석이 어렵다.

다. 다중 대체

다중 대체(multiple imputation)는 결측을 하나의 값으로 단순 대체하지 않고, 여러 번 반복 추정한 값을 결합하여 불확실성까지 반영하는 방법이다(Rubin, 1987). 단순 대체가 “하나의 확정값”을 제공한다면, 다중 대체는 “가능한 여러 대체 값의 분포와 변동을 함께 고려”하여 최종 추정값을 산출한다는 점에서 훨씬 정교하고 일관된 접근법이다. 다중 대체는 두 단계로 구성된다(Rubin, 1987; Schafer, 1997).

- **1단계: 다중 추정(multiple imputation).** 결측이 있는 변수를 대상으로 모형 기반 예측을 반복 수행하여 여러 개의 추정값을 산출한다.
- **2단계: 결과 통합(pooling).** 여러 추정값의 평균과 분산을 결합하여 최종 추정값을 산출한다.

$$\text{- 추정값 평균: } \bar{\theta} = \frac{1}{m} \sum_{j=1}^m \hat{\theta}_j$$

$$\text{- 분산 추정: } \bar{U} = \frac{1}{m} \sum_{j=1}^m U_j, \quad B = \frac{1}{m-1} \sum_{j=1}^m (\hat{\theta}_j - \bar{\theta})^2$$

\bar{U} (within imputation variance): 추정값 분산의 평균

B (between imputation variance): 추정값 사이의 분산

$$\text{- 총분산: } T = \bar{U} + \left(1 + \frac{1}{m}\right)B \quad 79)$$

다중 대체는 불확실성을 정량적으로 반영함으로써 추정값의 분산을 과소 추정하지 않고, 무작위 결측(MAR)에서는 타당한 결과를 제공한다(Rubin, 1987; Schafer, 1997). 그러나, 모형에 민감하여 잘못된 모형을 설정할 경우, 오히려 체계적 편향이 발생할 수 있으며, 비무작위 결측(MNAR)에서는 결측 발생 과정을 별도로 모형화하지 않으면

$$\Rightarrow \text{무작위 추출: IF 난수}(z) = 0.5, Y_i^{\text{miss}} = 6600 + (200 \times 0.5) = 6700$$

$$79) \text{ 예) 추정1: } \theta_1 = 100, SE_1 = 6, U_1 = 36, \quad \text{추정2: } \theta_2 = 108, SE_2 = 5, U_2 = 25$$

$$\text{추정3: } \theta_3 = 104, SE_3 = 7, U_3 = 49, \quad \text{추정4: } \theta_4 = 111, SE_4 = 5, U_4 = 25$$

$$\text{추정5: } \theta_5 = 107, SE_5 = 6, U_5 = 36, \quad \text{추정6: } \theta_6 = 106, SE_6 = 6, U_6 = 36$$

$$\Rightarrow \bar{\theta} = 106, \bar{U} = 34.5, B = 14, T = 50.83, SE_T \approx 7.13$$

$$\Rightarrow \text{최종 추정값의 평균은 106, 표준오차는 7.13}$$

적절한 추정이 어렵다(Little & Rubin, 2019). 따라서, 다중 대체는 데이터 통합 이후 남은 체계적 결측의 보정에 매우 효과적이지만, 적용 시에는 모형 선택의 타당성과 결측 메커니즘에 대한 사전 검토가 반드시 병행되어야 한다.

라. 대량 대체

데이터 통합 후, 결합 변수(고유식별자, 공통 변수, 공통 보조 변수)에는 결측이 없더라도 비결합 변수(non-matched variables)에는 여전히 결측이 남을 수 있다. **대량 대체**(mass imputation)는 통합 데이터 세트에서 특정 변수(Y)와 결합 변수(X)의 관계를 추정 후, 이를 다른 모든 **결측 단위(unit)에 일괄 적용**하여 결측을 대체하는 방법이다(Kim & Rao, 2012). 이는 레코드 수준의 항목 대체(item imputation)가 아니라, 비결합 변수의 결측을 일괄적으로 채우는 **단위 대체(unit imputation)**이다. 대량 대체는 통합 데이터 세트에서 비결합 변수를 결측 없이 확보함으로써 완전한 데이터 세트(complete data set)를 구축할 수 있어 모든 단위를 대상으로 동일 변수 집합을 이용한 분석이 가능해진다(Kim & Rao, 2012). 또한, 결합 이전의 각 데이터가 지닌 고유한 특성을 모두 활용할 수 있다는 점에서, 대량 대체는 데이터 통합의 목적에 부합한다. 가령, 행정자료와 조사자료를 결합했다면 행정자료의 포괄성(comprehensiveness)과 조사자료의 대표성(representativeness)을 동시에 확보할 수 있다. 그러나, 모형 설정에 크게 의존하기에 부적절한 모형을 적용하면 전체 데이터 세트에 체계적 편향이 확산될 수 있다. 또한, 데이터 간 변수 정의나 측정 방식의 차이가 크면 대체 결과의 해석이 불분명해질 수 있다. 이러한 한계를 보완하기 위해 다중 대체를 활용⁸⁰⁾하여 불확실성을 반영할 필요가 있다.

80) **예** 베이지안 다중 대체를 활용한 대량 대체

• 결측 구조 확인

ID(unit)	Age	Edu	Incoem(Y)
A	30	12	2,000
B	50	16	3,500
1	40	14	N.A.
2	35	12	N.A.
3	40	18	N.A.

• 대체 값 생성

- 모형 설정: $Y | X \sim N(\beta_0 + \beta_1 Age + \beta_2 Edu, \sigma^2)$
- 모수 추출: $\theta^t = (\beta_0^t, \beta_1^t, \beta_2^t, \sigma^2) \sim p(\theta | Y_{obs}, X), \quad t = 1, \dots, m$
- **대량 대체**(각 t 의 모든 결측 단위 u): $Y_u^{miss,t} \sim N(\beta_0^t + \beta_1^t Age_u + \beta_2^t Edu_u, [\sigma^t]^2)$
- $m = 3$: $t = 1: \beta_0 = 400, \beta_1 = 42, \beta_2 = 85, \sigma = 200$

마. 기계학습 기반 대체

기계학습 기반 대체(machine learning based imputation)는 기계학습 알고리즘이 학습한 변수 간 패턴을 이용해 결측을 대체하는 방법이다(Emmanuel et al., 2021; Stekhoven & Bühlmann, 2012). 회귀 대체나 평균 대체가 선형 관계와 분포 가정에 의존하는 것과 달리, 기계학습 기반 대체는 비선형 관계, 고차원 변수 간 상호작용, 연속형·범주형 혼합 데이터를 동시에 처리할 수 있다(Stekhoven & Bühlmann, 2012). 특히, 행정자료와 조사자료를 결합한 데이터 세트에서 결합 변수와 비결합 변수 간의 복잡한 관계를 모형화할 때 효과적이다. 예컨대, 선형 회귀모형이 설명하기 어려운 비선형 관계에서도 기계학습 모형은 더 높은 적합도와 예측력을 제공할 수 있다(Stekhoven & Bühlmann, 2012). 그러나, 표본 규모가 작으면 과적합(overfitting)⁸¹⁾ 위험이 크며, 해석력이 떨어져 “왜 이 값으로 대체하였는가?”를 설명하기 어렵다는 한계가 있다(Emmanuel et al., 2021; Stekhoven & Bühlmann, 2012). 따라서, 공식 통계에 적용하기 위해서는 투명성과 재현성 확보가 필수적이며, 현 단계에서는 보조적·보완적

$$t = 2 : \beta_0 = 600, \beta_1 = 38, \beta_2 = 90, \sigma = 250$$

$$t = 3 : \beta_0 = 500, \beta_1 = 40, \beta_2 = 80, \sigma = 220$$

ID(unit)	평균 ($t = 1$)	$Y^{miss, 1}$	평균 ($t = 2$)	$Y^{miss, 2}$	평균 ($t = 2$)	$Y^{miss, 3}$
1	3,270	3,370	3,380	3,330	3,220	3,220
2	2,890	2,970	3,010	2,950	2,860	2,890
3	3,820	3,940	3,930	3,830	3,740	3,810

• 결측값 대체

- 대표($t = 1$) 선택

ID(unit)	Age	Edu	Incoem(Y)
A	30	12	2,000
B	50	16	3,500
1	40	14	3,370
2	35	12	2,970
3	40	18	3,940

- 대체 간 평균

$$ID = 1 : [3370 \ 3330 \ 3220] = 3306.7$$

$$ID = 2 : [2970 \ 2950 \ 2890] = 2936.7$$

$$ID = 3 : [3940 \ 3830 \ 3810] = 3860$$

ID(unit)	Age	Edu	Incoem(Y)
A	30	12	2,000
B	50	16	3,500
1	40	14	3,306.7
2	35	12	2,936.7
3	40	18	3,860

81) 모형이 학습 데이터의 유연한 특성까지 과도하게 학습해 새로운 데이터에는 잘 맞지 않는 현상을 의미한다. 특히, 표본이 작은 상황, 즉 조사자료의 표본 크기가 작거나 통계적 매칭에서 매칭 단위의 크기가 작을 때 과적합 위험이 커진다.

수단으로만 활용하는 것이 바람직하다. 기계학습 기반 대체의 일반적 절차는 다음과 같다.

- **1단계: 결측 구조 확인.** 비결합 변수(Y)의 결측 단위를 파악한다.
- **2단계: 학습 데이터 구성.** 비결합 변수(Y)가 관측된 단위를 학습 데이터로, 결합 변수(X)를 설명 변수로 지정한다.
- **3단계: 학습.** 기계학습 알고리즘⁸²⁾을 이용해 비결합 변수(Y)와 결합 변수(X)의 관계를 학습한다.
- **4단계: 대체.** 학습된 모형을 **결측 단위** 전체에 적용하여 대체 값 생성한다.
- **5단계: 검증.** 대체 결과의 분포와 상관 구조를 점검하고, 여타 알고리즘과 교차검증(cross-validation)으로 모형성능을 확인한다.

3. 결측값 대체 품질 점검

결측값 대체 품질 점검(imputation quality assessment)은 결측 대체 과정이 ①결합 데이터 세트의 구조를 유지하고 ②추정의 불확실성을 적절히 반영하며 ③결합 목적에 부합하는지를 체계적으로 검증하는 절차이다. 즉, 대체가 잘 이루어졌는지를 평가하는 과정으로, 두 가지 측면에서 점검한다. 첫째, **데이터 적합성(data fit)**은 대체 값이 관측값과 유사한 분포, 상관, 조건부 관계를 보이는지를 검토한다. 둘째, **추론 적합성(inference fit)**은 다중 대체의 불확실성이 적절히 반영되어 추정량의 분산과 신뢰

82) 기계학습 기반 대체에서는 결합 데이터 세트의 크기, 변수 특성, 해석 가능성 등을 고려하여 적절한 알고리즘을 선택하는 것이 중요하다. 주요 알고리즘을 간략히 정리하면 다음과 같다.

- **K-최근접 이웃(k-nearest neighbor):** 결측 단위와 가장 유사한 k 개의 단위를 찾아, 평균 또는 최빈값으로 대체한다. 단순하고 분포 가정이 불필요하지만, 고차원 데이터에서는 거리 계산이 불안정하다.
- **랜덤 포레스트(random forest, RF):** 관측 단위에서 학습된 의사결정나무 앙상블(decision tree ensemble)을 통해 결측 단위를 예측한다(Stekhoven & Bühlmann, 2012). 비선형 및 상호작용을 잘 반영하고, 범주형·연속형 혼합 데이터를 동시에 처리할 수 있지만, 계산량이 크고, 예측값은 평균화되어 극단값 반영에 한계가 있다.
- **신경망(neural network, NN):** 인공신경망(ANN)을 이용하여 복잡한 비선형 관계를 학습하여 비결합 변수를 결합 변수로 예측한다. 대규모 데이터에서 강력하나 소규모 표본에서는 과적합 위험이 있으며, 결과 해석이 어렵다.
- **생성 모형 기반(generative model based):** 결측값을 포함한 데이터를 잠재 공간(latent space)에서 생성·복원하는 방법이다. 복잡한 다변량 분포 구조를 포착하여 결측 패턴까지 반영할 수 있지만, 구현이 복잡하고, 대규모 데이터 필요하다는 한계가 있다.

구간이 합리적인지를 확인한다(Little & Rubin, 2019; Rubin, 1987; Schafer, 1997). 또한, 데이터 통합에서는 대체 결과가 비결합 변수와 결합 변수(고유식별자, 공통 변수, 공통 보조 변수)의 구조적 관계를 보존하는지 그리고 외부 기준자료(예: 등록부, 공식 통계 등)와 모순되지 않는지도 함께 점검해야 한다.

가. 분포 및 특성값 정합성

분포 및 특성값 정합성(distributional and moment consistency) 점검은 대체 데이터가 관측 데이터(대체 이전 데이터)의 특성을 얼마나 잘 반영하는지를 평가하는 절차이다. 구체적으로는 평균, 분산, 표준편차, 상관관계, 회귀계수 등의 통계량이 대체 전후 얼마나 일관되는지를 비교한다. 결측값 대체에서 분포 왜곡이 발생하면 전체 데이터 세트의 신뢰성이 저하된다. 예컨대, 평균 대체는 평균은 보존하지만, 분산을 축소하는 문제가 있다(Little & Rubin, 2002). 회귀 기반 대체 또한 공분산 구조는 유지하지만, 분산을 과소 추정할 수 있다. 따라서, 대체 후 데이터의 분포 및 특성값을 체계적으로 점검하는 것은 통합 데이터 세트의 품질 확보를 위한 1차 검증 단계라 할 수 있다.

나. 관측값 모의 대체

관측값 모의 대체(over-imputation)는 실제 관측값을 임시로 결측 처리한 뒤, 이를 대체하여 원래 값과 비교함으로써 모형의 예측력을 검증하는 방법이다(Abayomi et al., 2008; Schafer, 1997). 이는 결측값만이 아니라 관측값까지 결측으로 취급해 대체 모형이 얼마나 정확하고 신뢰할 수 있는지를 평가하는 방법이다. 관측값 모의 대체 점검의 핵심은 ①예측오차의 크기와 ②불확실성 반영의 적절성이다.⁸³⁾ 특히, 베이지안 또는 기계학습 기반 대체처럼 모형이 복잡한 경우에는 필수 검증 절차이다.

- **RMSE**.⁸⁴⁾ 관측값과 대체 값 간의 평균 예측오차 크기를 나타내며, 값이 작을수록 대체 정확도가 높음을 의미한다.
- **사후 예측 커버리지(posterior predictive coverage)**.⁸⁵⁾ 각 관측값(y_i)이 대응하는 대체

83) ①대체된 값과 원래 관측값의 차이가 작아야 하며, ②다중 대체를 수행할 경우, 관측값이 대체 값의 신뢰 구간(예: 95%)에 포함되는 비율이 기댓값(95%)에 근접해야 한다.

$$84) RMSE_{over} = \sqrt{\frac{1}{M n_{obs}} \sum_{m=1}^M \sum_{i \in obs} (e_i^{(m)})^2}, \quad e_i^{(m)} = y_i - y_i^{miss, (m)}$$

n_{obs} : 임시로 결측 처리한 관측 단위의 수, M : 모의 대체 반복 횟수

$$85) \hat{p}_{cov} = \frac{1}{n_{obs}} \sum_{i=1}^{n_{obs}} I(y_i \in C_i^{(n\%)})$$

☞ 관측값이 사후 예측 95% 구간에 포함된 비율(\hat{p}_{cov})이 0.94면 불확실성을 적절히

구간($C_i^{(n\%)}$) 내에 포함되는 비율을 의미하며, 설정 대체 구간과 일치할수록 불확실성이 적절히 반영되었음을 의미한다.

만약, RMSE가 과도하게 크거나 사후 예측 커버리지가 설정 대체 구간보다 낮다면, 이는 모형의 과적합 또는 불확실성 과소 추정을 의미한다.

다. 다중 대체 진단

다중 대체 진단은 대체 결과의 안정성과 타당성을 확인하는 절차로 두 단계로 이루어진다. 첫째, 분산 분해(variance decomposition)를 통해 불확실성을 구성하는 요소인 총분산, 내부 분산, 외부 분산을 산출한다. 이를 통해 결측으로 인한 불확실성의 크기와 원인을 정량적으로 파악한다. 둘째, 효율성과 자유도 평가를 통해 대체 값의 수가 충분한지, 결측률이 결과 추정에 얼마나 영향을 미치는지를 검토한다. 대표적인 진단 지표는 다음과 같다.

- **결측 정보 비율(fraction of missing information, FMI).**⁸⁶⁾ 전체 분산 중 결측으로 인한 발생한 비율로, 그 값(λ)이 클수록 결측이 분석 결과에 미치는 영향이 크며, 불확실성 반영의 필요성이 높음을 의미한다.
- **상대 효율성(relative efficiency, RE).**⁸⁷⁾ 대체 값의 수가 무한히 많은 경우와 비교하여 현재 산출한 대체 값의 수(m)가 얼마나 효율적인지를 평가하며, 그 값이 1에 가까울수록 효율적임을 의미한다.
- **상대 증가 분산(relative increase in variance, RIV).**⁸⁸⁾ 결측으로 인해 전체 분산이 얼마나 증가했는지를 보여주는 지표이다. 결측이 많을수록 또는 대체 값 간 편차가 클수록 그 값이 커진다.
- **자유도(degree of freedom).**⁸⁹⁾ 결측 대체 횟수와 대체 결과 간의 변동성에 따라 신뢰 구간의 폭을 얼마나 조정해야 하는지를 알려주는 지표이다. 대체 횟수가

반영했다고 볼 수 있지만, 0.8이면 과소 추정, 0.99면 과대 추정했다고 볼 수 있다.

$$86) \lambda = \frac{(1 + \frac{1}{m})B}{T}$$

$$87) RE \approx (1 + \frac{\lambda}{m})^{-1}$$

$$88) r = \frac{(1 + \frac{1}{m})B}{\bar{U}}$$

$$89) df = (m-1)(1 + \frac{1}{r})^2$$

많고 결과 간 차이가 작을수록 자유도는 커져 신뢰 구간이 좁아지고(정규분포에 근접하고), 대체 횟수가 적거나 결과 간 변동이 클수록 자유도가 작아져 신뢰 구간이 넓어진다. 즉, 불확실성이 크게 반영된다.

라. 교차검증

교차검증(cross-validation)은 기계학습 기반 대체에서 모형의 예측 성능을 검증하는 방법이다(Emmanuel et al., 2021; Stekhoven & Bühlmann, 2012). 일부 관측값을 의도적으로 결측으로 처리한 후, 학습된 대체 모형을 이용해 해당 값을 복원⁹⁰⁾하여 실제 관측값과 비교함으로써 모형이 결측을 얼마나 정확하게 예측하는지를 평가한다(Stekhoven & Bühlmann, 2012). 교차검증 절차는 다음과 같다.

- **1단계:** 결합 데이터 세트 전체를 k 개의 구역(fold)으로 나눈다.
- **2단계:** 각 구역을 한 번씩 ‘가짜 결측 집합(pseudo-missing set)’으로 설정하고, 나머지 $k-1$ 개의 구역을 학습 데이터로 하여 모형을 학습시킨다.
- **3단계:** 학습 모형으로 가짜 결측값을 복원한 뒤, 실제 관측값과의 차이를 계산한다.
- **4단계:** 구역별 RMSE, MAE, R^2 산출⁹¹⁾ 후, 이들의 평균을 최종 성능 지표로 사용한다.

마. 외부 자료 기반 품질 점검

외부 자료 기반 품질 점검(external benchmarking)은 결측값 대체의 결과를 신뢰할 수 있는 외부 기준자료(예: 통계등록부, 행정자료, 대규모 조사자료 등)와 비교하여, 대체 데이터 세트의 타당성을 검증하는 절차이다(Rässler, 2002). 내부 분포 비교가 대체 모형의 적합성을 평가한다면, 외부 자료 기반 품질 점검은 대체된 값이 현실적으로 신뢰할 수 있는 수준인지를 판단하는 최종 검증 절차이다. 외부 자료 기반 품질 점검의 기준은 두 가지로 요약할 수 있다.

- **총량 정합성(consistency in totals).** 대체된 데이터의 총합이나 평균이 외부 기준자료와 얼마나 일치하는가?

90) 교차검증과 관측값 모의 대체는 모두 실제 관측값을 결측으로 처리하여, 모형을 평가한다는 점에서 유사하다. 그러나, 교차검증이 모형의 예측 성능 검증 도구라면, 관측값 모의 대체는 결측 대체 과정의 불확실성 진단 도구이다.

91) $RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$, $MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$, $R^2 = 1 - \frac{\Sigma(y_i - \hat{y}_i)^2}{\Sigma(y_i - y_i)^2}$

$$\delta = \frac{\hat{\tau} - B}{B} \times 100$$

$\hat{\tau}$: 대체 데이터 세트의 지표(합계, 평균 등)

B : 외부 기준자료의 지표

- **구조적 정합성(structural consistency)**. 세부 분포(예: 연령, 성별, 지역 등)가 외부 기준자료와 얼마나 유사한가?

$$\Delta p_g = \hat{p}_g - p_g^B$$

\hat{p}_g : 대체 데이터 세트의 세부 구조 비율

p_g^B : 외부 기준자료의 세부 구조 비율

외부 자료 기반 점검은 공식 통계와의 정합성을 평가한다는 점에서 중요하다.

바. 민감도 분석

민감도 분석(sensitivity analysis)은 결측 메커니즘에 대한 가정이 불확실할 때, 대체 결과가 이러한 가정의 변화에 얼마나 민감하게 반응하는지를 평가하는 방법이다(Little & Rubin, 2002; Schafer, 1997). 대부분의 결측 대체는 무작위 결측(MAR)을 전제로 하지만, 실제 데이터에서는 비무작위 결측(MNAR)일 가능성이 높다. 민감도 분석은 “만약 결측이 MAR이 아니라 MNAR이면, 추정 결과는 무엇이 달라질까?”라는 의문에서 출발한다. 이를 위해 시프트 파라미터(δ)를 설정하여, 결측값을 일정하게 조정하면서, 그 결과(예: 평균, 비율, 회귀계수 등)가 얼마나 변화하는지를 평가한다.

시프트 파라미터(δ)는 임의의 값이 아니라, 현실적으로 가능한 편차를 가정하기 위해 설정하는 시나리오 변수이다. 예컨대, 소득조사에서 고소득층이 응답하지 않았을 가능성을 반영해 실제 평균이 5~10% 높았을 수 있다고 δ 를 설정할 수 있고, 건강조사에서는 몸무게에 응답하지 않은 사람이 실제로는 평균 2~3kg 더 무거울 수 있다고 δ 를 부여할 수도 있다(Little & Rubin, 2002). δ 값을 \pm 로 조정하며 관심 추정값(예: 평균, 비율, 회귀계수 등)의 변화를 관찰하면 대체 결과의 안정성을 평가할 수 있다. 가령, δ 를 $\pm 200,000$ 원 조정했을 때 평균 소득이 3,000,000원에서 2,800,000~3,200,000원으로 변하는 정도라면 결론은 대체로 유지되므로 견고(robust)하다 할 수 있다. 반대로 소득 추정값이 2,000,000~4,000,000원처럼 크게 요동친다면 민감(sensitive)으로 평가한다. 견고·민감의 판단 기준은 연구 맥락에 따라 다르지만, 보통 $\pm 5\sim 10\%$ 를 허용 범위로 본다. 이러한 접근이 의미 있는 이유는 결측 메커니즘(MAR, MNAR)을 실제로 검증하는 것은 불가능하기 때문이다. 따라서 δ 를 조정해 “만약 MNAR이라면 결과는 이렇게 달라질 수 있다.”를 제시하는 것만으로도 결론의

불확실성을 직관적으로 체감할 수 있다(Schafer, 1997). 대표적 접근법으로는 패턴-믹스처 모형(pattern-mixture model)과 선택모형(selection model)이 있다.

- **패턴-믹스처 모형**(δ 조정 모형).⁹²⁾ 다양한 δ 값에 따른 결측 집단과 관측 집단의 분포 변화를 파악한다. 단순한 선형 이동으로 계산이 쉬워 기본 시나리오로 자주 사용된다.
- **선택모형**.⁹³⁾ 특정 단위(i) 결측 확률(p_i)에 δ 를 더하여 결측이 특정 집단에 더 강하게 연결되었다고 가정한 후, 추정 결과의 변화를 분석한다.

제4절 보정 및 대표성 점검

결측값 대체가 완료된 이후에도, 통합 데이터 세트가 모집단을 완전하게 대표하지 못하는 경우가 존재한다. 이는 자료 출처와 수집 방식이 달라서 발생하는 대표성 손실(loss of representativeness) 문제로 통합 데이터 세트의 일관성과 신뢰성을 저해할 수 있다. 다만, 행정자료나 전수조사 자료처럼 통합 데이터 세트가 모집단 전체를 포괄하거나 대표성 결함이 통계적으로 유의미하지 않으면 별도의 가중값 산출이나 보정 절차를 생략할 수 있다. 반면, 표본조사나 비확률적 자료(예: 빅데이터)를 포함하는 통합 데이터 세트의 경우에는 대표성 손실을 바로잡지 않으면 추정값의 편향이 발생할 가능성이 높다.

대표성 문제는 크게 세 가지 요인에서 비롯된다. 첫째, 표본조사에서는 추출 확률에 기반한 설계 가중값(design weight)을 적용하더라도 무응답 편향(non-response bias)이나 표본 추출 편향(sample selection bias)이 발생할 수 있다. 둘째, 행정자료는 표면적으로는 모집단 전체를 포괄하는 것처럼 보이지만, 실제로는 누락(under-coverage)이나 중복(over-coverage)이 존재할 수 있다. 셋째, 빅데이터는 자발적 참여(self-selection/volunteering)에 의존하기 때문에 모집단 대표성이 본질적으로 제한된다. 이러한 문제들은 통합 데이터 세트의 구조적 불일치(structural discrepancy)로 이어지며, 단순한 결합만으로는 교정할 수 없다. 따라서, 통합 데이터 세트가 모집단을 충분히 대표하지 못할 때는, **가중값 조정**이나 **분포 보정(calibration)**⁹⁴⁾을 통해 대표성을 보완한다. 다시

92) $Y_{imp}^{\delta} = Y_{imp} + \delta$

93) $\text{logit}(p_i^{\delta}) = \text{logit}(p_i) + \delta$

94) 보정은 **설계 가중값을 보정 가중값으로 변환**하여 모집단 분포와 일치시키는 과정이다(Deville & Särndal, 1992; Särndal, 2007). 즉, 보정의 본질은 “표본이 모집단을 얼마나 잘 대표하는가?”라는 문제에 해법을 제공하는 것이다. 전통적인 표본조사 맥락에서 보정은 세 가지 기본 요소로

말해, 보정은 결합 단계에서 남은 대표성 결함을 교정함으로써 통합 데이터 세트가 실제 모집단의 구조를 충실히 반영하도록 하는 절차이다. 보정은 ①초기 가중값 산출, ②보정 기준 설정, ③보정 가중값 산출의 순으로 진행된다.

1. 초기 가중값 산출

가중값은 각 개체가 모집단을 대표하는 정도로, 외부 기준자료와의 불일치를 줄이기 위한 보정 가중값 산출의 기초가 된다(Deville & Särndal, 1992). 한편, 통합 데이터 세트의 가중값은 결합 전 데이터의 설계 가중값을 그대로 가져오는 것이 아니라, ①중복(over-coverage) 조정, ②누락(under-coverage) 보완을 반영해 산출하며, 초기 가중값(initial weights)이라 한다.

가. 중복 조정

통합 데이터 세트의 기본 분석 단위는 통합 개체이다. 하나의 개체가 여러 데이터에 동시에 관측될 경우, 해당 개체는 통합 데이터 세트에 포함될 확률이 높아져 과대 표집(over-representation)을 유발할 수 있다. 중복 조정(duplication adjustment)은 각 개체의 중복 횟수(m_i)를 고려하여 가중값을 축소함으로써 이를 방지한다(Zhang, 2012). 각 데이터의 품질이나 신뢰도를 반영하기 위해, 단순 평균 대신 품질 비중(α_s)을 적용할 수 있다. 이를 통해 품질이 높은 데이터의 기여도를 상대적으로 높이고, 중복으로 인한 과대 포함을 완화할 수 있다.

$$\pi_i^{itd} = \frac{1}{m_i} \sum_{s=1}^S \alpha_s p_{i,s}$$

π_i^d : 통합 데이터 세트에 포함될 확률

m_i : 통합 개체(i)가 여러 데이터에 중복된 횟수

α_s : 데이터(s)의 품질 비중

$p_{i,s}$: 개체가 데이터에 포함될 확률

구성된다.

- 초기 가중값(initial weight)은 추출 확률과 비응답 보정을 반영해 산출된다.
- 보정 기준(calibration benchmark)은 인구총조사나 행정자료 등 모집단 참값이 알려진 변수의 총합을 의미한다.
- 보정 가중값(calibration weight)은 초기 가중값과 크게 다르지 않으면서도 보정 기준과 정확히 일치하도록 조정된 최종 가중값이다.

이를 통해 표본조사는 모집단과의 불일치를 줄이고, 추정값의 편향을 완화할 수 있다.

나. 누락 보완

일부 집단이 통합 데이터 세트에서 과소 포함되는 누락 문제를 방지하기 위해 외부 기준자료와 비교하여 보완계수(complement coefficient)를 산출한다(Deville & Särndal, 1992).

$$c_g = \frac{N_g}{n_g} \quad 95)$$

N_g : 외부 자료에서 집단(g)의 모집단 크기

n_g : 중복 제거 후, 집단별 레코드 수

보완계수는 특정 집단의 실제 크기와 통합 데이터 세트에서의 관측 크기 간 불균형을 보정하여, 각 집단의 모집단 내 비중을 정확히 반영하도록 한다. 이는 청년층, 고령층, 소상공인 등 일부 행정자료에서 대표성이 낮은 집단의 가중 보정에 특히 유용하다.

다. 초기 가중값 산출

중복 조정과 누락 보완을 거친 후, 통합 개체 단위의 초기 가중값(initial weight)을 산출한다(Valliant et al., 2018).

$$d_i = \frac{c_{g(i)}}{\pi_i^{itd}}$$

$c_{g(i)}$: 통합 개체가 속한 집단의 보완계수

초기 가중값은 통합 데이터 세트의 대표성을 확보하기 위한 절차이지만, 다음의 한계를 지닌다.

- **극단 가중값 발생.** 일부 집단에서 보완계수($c_{g(i)}$)가 과도하게 커질 경우, 극단 가중값이 발생하여 추정량의 분산이 급격히 증가할 수 있다.
- **데이터 불일치.** 전처리 과정에서 변수 정의나 포괄범위가 불일치하면, 포함확률($p_{i,s}$) 추정의 불확실성이 커져 가중값의 신뢰도가 저하된다.
- **시간 변동성.** 시계열 단절이나 기준 시점 불일치가 발생하면 데이터 품질 비중(α_s)이 달라져 일관된 가중값 산출이 어렵다.

이러한 한계는 이후 보정 단계에서 가중값을 재조정하고, 메타데이터 기반의 품질 관리 체계를 통해 완화할 수 있다.

95) $c_g > 1$ 이면 과소 포함

2. 보정 기준 설정

보정은 “무엇을 기준으로 맞출 것인가?”를 정의하는 단계이다. 보정 기준(calibration benchmark)은 통합 데이터 세트가 충족해야 할 모집단의 참값을 의미하며, 성·연령별 인구, 지역별 사업체 수, 산업별 종사자 수 등 신뢰성이 확보된 외부 자료를 근거로 설정된다. 데이터 통합에서는 데이터마다 포괄범위와 목적이 달라, 어떤 데이터를 기준으로 삼을지 결정하는 과정 자체가 대표성 확보의 핵심이 된다(Deville & Särndal, 1992; Särndal, 2007).

가. 보정 변수 선정

보정 변수(calibration variables)는 통합 데이터 세트가 모집단의 기준값과 반드시 일치해야 하는 변수이다. 일반적으로 성별, 연령, 지역과 같은 인구학적 변수가 핵심 변수로 사용되며, 목적에 따라 학력, 고용 상태, 산업분류, 가구 유형 등이 추가될 수도 있다. 보정 변수는 단순히 “맞춰야 할 변수”가 아니라, 통합 데이터 세트가 어떤 차원에서 대표성을 확보할지를 결정하는 구조적 설계 요소이다. 따라서 보정 변수는 다음과 같은 기준에 따라 선정되어야 한다.

- 통합 목적과 밀접한 변수를 선정해야 한다.
- 신뢰할 수 있는 기준자료(예: 인구총조사, 기업통계등록부 등)에서 기준값이 제공되지 않는 변수는 제외한다.
- 분석 대상 변수와 상관관계가 높을수록 보정 효과가 크다.
- 세분화 수준이 지나치게 높으면 일부 집단에서 극단 가중값이 발생하여 추정값의 불안정성이 커질 수 있다.

즉, 보정 변수의 선정은 “무엇을 맞출 것인가?”의 문제가 아니라, 통합 데이터 세트가 어떤 구조적 대표성을 확보할 것인가를 결정하는 단계이다.

나. 후보 자료원 평가

보정 변수를 선정한 후에는 해당 변수의 기준값을 제공할 외부 자료원을 선정해야 한다. 대규모 조사자료, 행정자료, 등록부, 빅데이터 등 다양한 자료가 후보가 될 수 있으나, 각각 장단점이 있기에 체계적인 평가 절차가 필요하다. 후보 자료원은 다음의 기준을 중심으로 평가한다.

- **정확성(accuracy)**. 자료원이 모집단 참값에 얼마나 근접하는지를 평가한다. 규모

차이(MAE, RMSE)와 분포 차이(TVD, JSD)⁹⁶를 함께 활용하면 모집단과의 적합성을 다면적으로 진단할 수 있다(Lohr, 1999).

- **포괄성(coverage)**. 자료원이 모집단을 얼마나 완전하게 포함하는지를 평가한다. 누락률·중복률⁹⁷이 대표적 지표이며, 특정 집단에서 체계적 누락이 발생하면, 정확한 기준값이라도 대표성이 왜곡될 수 있다(Bethlehem, 2010).
- **시의성(timeliness)**. 자료원의 기준 시점과 보정 시점 간의 시차를 의미한다. 시의성이 낮은 자료는 현재 모집단 구조를 반영하지 못하므로, 보정 기준이 아닌 보조 자료로만 활용하는 것이 적절하다.
- **일관성(coherence)**. 변수 정의 일치성⁹⁸을 평가한다(Groves et al., 2009). 변수 정의가 다르거나 분류체계가 호환되지 않으면 보정 과정에서 논리적 불일치가

96) **TVD**(total variation distance)는 두 분포 간의 전체적 차이를 0~1의 값으로 요약한다(Deville & Särndal, 1992; Särndal, 2007).

$$TVD = \frac{1}{2} \sum_e |p_e - \pi_e|$$

p_e, π_e : 표본과 모집단에서 범주(c)의 비율

JSD(Jensen-Shannon divergence)는 KL 발산의 비대칭성을 보완하여 대칭적인 분포 차이를 제공하며, 분포 유사성을 측정하는 데 안정적이고 직관적이다(Endres & Schindelin, 2003).

$$JSD(p \parallel \pi) = \frac{1}{2} KL(p \parallel \frac{p+\pi}{2}) + \frac{1}{2} KL(\pi \parallel \frac{p+\pi}{2})$$

$$97) \text{ Coverage} = 1 - \frac{D}{N}, \quad \text{Duplication} = \frac{D}{N}$$

U : 누락, D : 중복, N : 모집단 크기

98) 변수 정의 일치성 점검

- χ^2 distance(범주형 변수): 0이면 완전 일치, 값이 커질수록 불일치가 크다(Agresti, 2002).

$$\chi^2 \text{ distance} = \sum_c \frac{(p_c - \pi_c)^2}{\pi_c}$$

- *Cohen's κ* (범주형 변수): 0이면 우연한 일치, 1이면 완전 일치를 의미한다(Cohen, 1960).

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

P_o : 실제 일치할 확률, P_e : 우연히 일치할 확률

- **SMD**(standardized mean difference, 연속형 변수): $|SMD| \leq 0.1$ 이면 작은 차이, 0.3이면 중간, 0.5 이상이면 큰 차이로 해석한다(Zhang et al., 2019).

$$SMD = \frac{|\bar{y}_A - \bar{y}_B|}{s_p}, \quad s_p = \sqrt{\frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}}$$

발생할 수 있다.

- **편향 위험(bias risk)**.⁹⁹⁾ 후보 자료원이 특정 집단이나 속성에 대해 체계적 편향을 가지는 경우를 의미한다(Bethlehem, 2010; Groves et al., 2009). 편향 위험이 큰 자료원을 보정 기준으로 활용하면 통합 데이터 세트의 대표성이 심각하게 훼손될 수 있다.
- **자료 분해(disaggregation)**.¹⁰⁰⁾ 자료원이 어느 수준까지 세분된 단위로 안정적인 통계를 제공할 수 있는가를 평가한다(Kish, 1992; Lohr, 1999). 국가 단위에서는

99) 편향 위험을 진단하기 위해 여러 지표가 사용되며, 각각은 편향 발생의 다른 측면을 포착한다(Bethlehem, 2010; Groves et al., 2009).

- **선택편의 근사식(approximate formula for selection bias)**: 편향이 응답률 자체가 아니라 응답의 선택성에서 비롯됨을 보여준다.

$$Bias(\bar{y}) = \frac{Cov(R, y)}{R}$$

R : 응답률, y : 관심 변수

- **대표성 지표(representativity indicator)**: 모집단 대비 대표성을 종합적으로 측정하며, 1에 가까울수록 대표성이 높다.

$$R = 1 - 2 Var(\hat{\pi}_i)$$

$\hat{\pi}_i$: 응답할 확률 추정값

- **변동계수(coefficient of variation of weights, CV)**: 가중값이 얼마나 고르게 분포되었는지 보여주는 지표이다. 일부 단위에 과도한 비중이 실리면 값이 커진다.

$$CV = \frac{s_w}{\bar{w}}$$

\bar{w} : 평균 가중값, s_w : 가중값의 표준편차

- **극단 가중값 비율(extreme weight ratio, ER)**: 초기 가중값에 비해 일정 이상으로 커진 단위의 비율을 측정한다. ER이 높으면 보정 과정이 불안정하고, 특정 단위가 전체 가중값을 좌우할 수 있다.

$$ER = \frac{|\{i : w_i > c d_i\}|}{n}$$

w_i : 보정 후 가중값, d_i : 초기 가중값, c : 임계 배수

100) 자료 분해는 유효 표본 크기, RSE, 셀 공란 비율로 평가할 수 있다.

- **유효 표본 크기(effective sample size)**: 가중값 불균형으로 실질적 표본 크기가 줄어드는 정도를 보여준다(Kish, 1992). 값이 낮으면 자료 분해 수준이 낮음을 의미한다.

$$n_{eff} = \frac{(\sum_i w_i)^2}{\sum_i w_i^2}$$

w_i : 보정 또는 설계 가중값

- **RSE(relative standard error)**: 값이 클수록 불확실성이 커져 세부 단위 보정은 부적합하다. $RSE > 0.3$ 이면 해당 추정값을 신뢰하기 어렵다(Eurostat, 2014).

안정적이라도, 시·도나 시·군·구 단위에서는 불확실성이 급증할 수도 있다. 과도한 세분화 수준에서 자료원을 보정 기준으로 사용하면 극단값이 발생하고 추정의 안정성이 저하된다(Kish, 1992).

다. 기준값의 시계열 일관성 확보

보정은 특정 시점에서의 대표성 확보뿐 아니라, 시간의 흐름에 따른 시계열 일관성(coherence over time)을 동시에 유지해야 한다. 그러나, 외부 기준자료가 주기적으로 개편되거나 새로운 자료원으로 대체되면, 단절(discontinuity)과 수준 차(level shift)가 발생하여 시계열 연속성이 훼손될 수 있다. 이에 외부 기준값이 시계열 전반에 걸쳐 논리적으로 모순 없이 연결되도록 점검하고 조정하는 과정이 필요하다. 시계열 일관성 확보 과정은 다음 세 단계로 구성된다.

- **시계열 점검(time-series validation check)**. 외부 기준값의 연도별 증감률과 분포 변화율을 산출하여 급격한 수준 변화나 이상값(outlier) 발생 시점을 탐지한다.
- **불연속성 확인(discontinuity diagnosis)**. 시계열 단절이 경제·사회적 요인(예: 정책 변화, 경기 충격 등)에 기인한 것인지, 아니면 자료 수집 체계나 변수 정의의 변경으로 인한 것인지를 구분한다. 이를 통해 조정 대상이 데이터 문제인지 현상 변동인지를 명확히 식별할 수 있다.
- **시계열 조정(times-series adjustment)**.¹⁰¹⁾ 재기준화(re-basing), 연결계수(link factor) 등을 활용하여 시계열 연속성을 확보한다.

$$RSE = \frac{SE(\hat{X})}{\hat{X}}$$

- **셀 공란 비율(cell empty rate)**: 보정 변수의 자료 분해 수준이 과도한지를 점검하는 지표로, 빈 셀이 많으면 변수 수준을 축소하거나 병합할 필요가 있다.

$$CER = \frac{|\{c : n_c = 0\}|}{C}$$

C : 세분된 전체 셀의 수

n_c : 셀에 포함된 관측값의 수

101) 시계열 조정 방법

- **재기준화**: 기준연도를 새로 설정하여 모든 시계열을 동일 기준(=100)으로 환산한다. 비교가 단순하지만, 기준연도 변경 시 해석의 연속성이 저하될 수 있다.
- **연결계수(LF)**: 기준 변경 전후의 추정값 비율을 이용하여 시계열 수준 차를 제거해 연속성을 확보하는 방법이다(IMF, 2012a, 2017). 자료원 개편으로 단절 발생 시, 신자료 전체에 LF를 할당하여 시계열을 연결한다. 시계열을 매끄럽게 연결할 수 있지만, 구조적 차이를 단순히 LF로만 반영하는 한계가 있다.

$$LF = \frac{\hat{Y}_t^{new}}{\hat{Y}_t^{old}}, \quad \hat{Y}_{t+k}^{adj} = \frac{\hat{Y}_{t+k}^{new}}{LF}$$

3. 보정 가중값 산출

보정 가중값(calibrated weights)은 통합 데이터 세트의 초기 가중값을 외부 기준값에 맞추어 조정한 결과로, 통합 데이터 세트의 대표성과 정합성을 확보하는 절차이다 (Deville & Särndal, 1992). 보정은 기존 가중값을 폐기하고 새롭게 가중을 부여한다는 의미가 아니라, 각 데이터가 본래 가지고 있는 가중값을 출발점으로 하여 사후적으로 보정을 수행한다는 점을 명확히 해야 한다. 즉, 조사자료의 설계 가중(design weight)이나 무응답 보정(non-response adjustment) 등 기존 정보는 그대로 유지하면서, 결합 후 나타나는 대표성 손실과 편향/편의(bias)를 최소화하기 위해 외부 기준값에 따라 미세 조정하는 절차가 바로 보정이다. 보정 가중값은 다음 두 가지 조건을 충족해야 한다.

- **근접성(proximity)**. 보정 가중값은 초기 가중값과 크게 달라지지 않아야 한다 (Deville & Särndal, 1992). 이는 불필요한 변동을 최소화하여 추정값의 안정성을 유지하기 위함이다.
- **제약 충족성(constraints satisfaction)**. 보정 가중값을 적용한 추정 결과가 외부 기준값(예: 인구총조사 기준의 성·연령·지역별 분포)과 일치하거나 근접해야 한다.

한편, 보정의 실제 적용에서는 자료 유형과 결합 방법에 따라 초기 가중값 구성과 보정 방법의 선택이 달라질 수 있다.

첫째, 조사자료 기반 통합(조사자료 단독 또는 조사자료-행정자료 통합)

조사자료의 설계 가중값은 모집단을 대표하도록 설계된 기초 가중값이므로, 결합 후에도 이를 초기 가중값으로 그대로 승계한다. 다만, 결합 과정에서 일부 조사자료가 행정자료와 연계되지 않는 등 연계 실패(non-linkage)가 발생하면, 연계 확률(linkage propensity)을 추정하여 그 역수로 가중을 조정함으로써 연계 편의(linkage bias)를

예) 2024년 동일 시점 값: 구(舊)기준 1,000, 신(新)기준 1,100

2025년 신기준 값이 1,210일 때, 구기준과 연속되게 맞추면 2025년 구기준 값은 1,100

$$\Rightarrow LF = \frac{1100}{1000} = 1.1, \quad \hat{Y}_{2025}^{adj} = \frac{1210}{1.1} = 1100$$

- **스무딩(smoothing)**: 이동평균, HP 필터 등을 적용하여 단기 변동을 완화하는 방법이다. 급격한 변화를 지나치게 평탄화할 위험이 있다.
- **일관화(reconciliation)**: 장기 기준(예: 인구총조사, B_s)과 단기 기준(예: 연간 행정자료, b_s)의 총합을 일치시키는 방법이다. 장기 기준의 대표성과 단기 기준의 시의성을 동시에 확보할 수 있지만, 총합을 맞추는 과정에서 단기 기준의 세부 항목이 왜곡될 수 있다.

$$b_i^{adj} = b_i \frac{\sum_{s \in \text{기간}} B_s}{\sum_{s \in \text{기간}} b_s}$$

보정한다. 그 후, 외부 기준자료에 맞추어 사후 보정을 수행한다. 이를 통해 기존 가중값의 설계 정보를 보존하면서, 통합 데이터 세트의 대표성을 확보할 수 있다 (Särndal & Lundström, 2005; Valliant et al., 2018).

둘째, 행정자료 기반 통합(행정자료 단독 또는 행정자료-조사자료 통합)

행정자료는 표면적으로 모집단 전체를 포괄하므로, 전수자료에 가까운 경우에는 가중값 산출 자체를 생략할 수도 있다. 그러나, 실제로는 누락이나 중복이 존재할 수 있으므로, 프레임 정합성 점검(frame consistency check)을 통해 보정 제약 조건을 설정하고, 제한적 보정을 적용하여 모집단 분포를 정합시킨다.

셋째, 비확률적 자료 기반 통합

비확률적 행정자료나 빅데이터에는 설계 가중이 존재하지 않기 때문에 관측 가능 변수(observed covariate)를 이용하여 모집단과의 차이를 보정하는 의사 기초 가중(pseudo-base weight) 또는 성향 가중(propensity score weight)을 구성한다. 이는 모집단의 보조 정보(auxiliary information)를 이용하여 관측 확률(selection probability)을 추정하고, 그 역수를 가중값으로 부여¹⁰²⁾하는 절차이다(Bethlehem, 2010; Valliant et al., 2018). 그 후, 외부 기준(예: 인구 분포, 지역·성별 구성 등)을 활용하여 보정을 수행한다. 이 과정에서 발생할 수 있는 극단 가중값 문제를 완화하기 위해 트리밍(trimming)과 캡핑(capping)을 적용한다(Bethlehem, 2010; Valliant et al., 2018). 트리밍은 일정 임계값을 초과하는 가중값을 잘라내어 분산의 급증을 방지하고, 캡핑은 가중값의 상선을 설정하여 추정 안정성을 확보하는 절차이다.

요약하면, 보정은 초기 가중값을 토대로 외부 기준과의 일치를 확보하기 위한 과정이다. 즉, 조사자료의 가중값은 유지·조정되고, 행정자료는 필요시 보정되며, 비확률적 자료는 의사 가중을 새로 구성한 후 보정이 수행된다. 이러한 절차를 통해 통합 데이터 세트의 대표성과 정합성을 동시에 확보할 수 있다. 다만, 데이터 통합에서는 자료의 출처와 구조가 서로 다르므로, 보정의 기준이 되는 외부 자료가 단일하지 않을 수 있다. 따라서 복수의 기준값을 동시에 고려하거나 서로 충돌할 때는 자료원 간 우선순위를 명확히 설정하는 절차가 필요하다. 이는 보정 과정의 일관성과 통계적 정합성을 확보하기 위한 핵심 단계로, 다음과 같은 절차로 구성된다.

가. 우선순위 선정

전통적인 보정은 단일 자료원을 기준으로 수행되었다. 가령, 인구총조사의 성·

102) $w_i = \frac{1}{p_i}$ 의 가중을 통해, 특정 개체가 데이터에 포함될 확률이 낮을수록 더 큰 가중값이 부여된다.

연령별 인구 분포를 기준으로 가중값을 조정하는 방식이다. 그러나 단일 기준 접근은 데이터 통합 환경에서는 한계를 가진다.

- **유연성 부족.** 단일 자료원은 총조사에서 제공되지 않는 시의성이나 산업·직업별 세분화 정보는 반영하기 어렵다.
- **극단 가중값 문제.** 하나의 제약 조건만 맞추려다 보면 특정 세부 집단의 가중값이 비정상적으로 커져 추정 분산이 증가한다(Deville & Särndal, 1992).

데이터 통합에서는 자료의 출처와 구조가 다양하기에 복수의 외부 자료원을 함께 고려하는 접근이 필요하다.¹⁰³⁾ 이를 위해, 자료원 간 **우선순위**¹⁰⁴⁾를 명확히 하는 절차가 필수적이다. 우선순위 선정은 ①정규화, ②가중값 부여, ③종합 점수 산출로 진행된다.

- **정규화(normalization).**¹⁰⁵⁾ 후보 자료원은 정확성, 포괄성, 시의성 등 서로 다른 단위와 방향성을 가진 기준으로 평가된다. 예컨대, 정확성은 비율(%), 시의성은 지연 개월 수로 표현되며, 일부 지표는 값이 클수록(편익형) 좋지만, 일부는 값이 작을수록(비용형) 바람직하다. 따라서, 지표를 단순 합산하는 것은 불가능하며, 모든 지표를 공통 척도로 변환하는 정규화가 필요하다. 또한, 극단값(outlier)이 존재하면 중간값이 왜곡될 수 있으므로, 로그 변환이나 순위 기반 정규화를 고려해야 한다(Valliant et al., 2018).

103) Eurostat(2013)은 다원 자료 기반 보정에서 자료원의 품질 지표를 다기준 평가(MCDA) 방식으로 통합할 것을 권고한다.

104) 우선순위 선정은 외부 자료원의 장단점을 고려하여, 기여 수준을 설정하는 과정이다.

- 상위 순위(hard constraint; Deville & Särndal, 1992): 반드시 일치해야 하는 기준으로 적용
- 중위 순위(soft constraint; Valliant et al., 2018): 일정 오차 범위 내 근사 일치를 허용
- 하위 순위(reference constraint): 추세나 방향성만 반영하는 보조 기준으로만 사용

이 계층 구조는 단순 서열이 아니라, 실제 보정 모형에서 제약의 강도로 구현된다.

105) 서로 다른 척도의 지표를 0~1의 범위로 변환하는 절차이다.

- 편익형 지표(정확성, 포괄성, 자료 분해; 값이 클수록 바람직)

$$q_{jk} = \frac{x_{jk} - \min_j x_{jk}}{\max_j x_{jk} - \min_j x_{jk}}$$

q_{jk} : 자료원(j)의 기준 지표(k) 값

☞ 포괄성: 인구총조사 = 100%, 행정자료 = 95%, 빅데이터 = 60%

$$\Rightarrow \text{총조사} = \frac{100 - 60}{100 - 60} = 1.00, \text{ 행정자료} = \frac{95 - 60}{100 - 60} = 0.88, \text{ 빅데이터} = \frac{60 - 60}{100 - 60} = 0.00$$

- 비용형 지표(시의성, 일관성, 편향 위험; 값이 작을수록 바람직)

$$q_{jk} = \frac{\max_j x_{jk} - x_{jk}}{\max_j x_{jk} - \min_j x_{jk}}$$

- **가중값 부여(weighting assignment).**¹⁰⁶⁾ 정규화된 지표들은 동일 척도(0~1)로 표현되지만, 데이터 통합의 목적에 따라 각 지표의 중요도는 다르므로 이를 단순

106) 가중값 부여 방법

- 직접 할당(direct assignment): 전문가 합의를 통해 가중값을 할당하는 방법으로 신속하고 직관적이지만, 합의 과정에서 주관성이 개입될 수 있고, 일관성 검증이 어렵다.
- 계층화 분석(analytic hierarchy process, AHP; Satty, 1980): 쌍대 비교(pairwise comparison) 행렬의 고유벡터로 가중값을 산출한다.

- ① 기준들을 계층 구조로 배열
- ② 각 기준을 1~9의 척도로 쌍대 비교
- ③ 쌍대 비교 행렬 작성 후, 고유벡터를 계산하여 가중값 산출

예) 정확성, 포괄성, 시의성의 쌍대 비교 행렬

	정확성	포괄성	시의성
정확성	1	2	3
포괄성	$\frac{1}{2}$	1	2
시의성	$\frac{1}{3}$	$\frac{1}{2}$	1

⇒ 정규화 행렬 = 각 원소 ÷ 해당 열의 합

	정확성	포괄성	시의성
정확성	0.546	0.571	0.500
포괄성	0.273	0.286	0.333
시의성	0.182	0.143	0.167

⇒ 가중값(행 평균) 산출: $w_{\text{정확}} = 0.539$, $w_{\text{포괄}} = 0.297$, $w_{\text{시의}} = 0.164$

④ 일관성 지수(CI)와 일관성 비율(CR)로 검증

$$CI = \frac{\lambda_{MAX} - K}{K - 1}, \quad CR = \frac{CI}{RI}$$

λ_{MAX} : 쌍대 비교 행렬의 최대 고윳값, K : 기준의 수

RI : 동일 차수 n 의 무작위 쌍대 비교 행렬을 무수히 많이 생성해서 얻은 CI 의 평균

($n = 3 \Rightarrow RI = 0.58$, $n = 4 \Rightarrow RI = 0.90$, $n = 5 \Rightarrow RI = 1.12$)

예) 일관성 점검

	정확성	포괄성	시의성	Aw_i	w_i	$\frac{Aw_i}{w_i}$
정확성	1×0.539	2×0.297	3×0.164	1.615	0.539	2.995
포괄성	$\frac{1}{2} \times 0.539$	1×0.539	2×0.164	0.892	0.297	3.004
시의성	$\frac{1}{3} \times 0.539$	$\frac{1}{2} \times 0.297$	1×0.164	0.492	0.164	3.000

⇒ $\lambda_{MAX} \approx \frac{2.995 + 3.004 + 3.000}{3} = 3$

평균 내는 것은 타당하지 않다. 예컨대, 인구 구조 분석에서는 정확성과 포괄성이 핵심 기준이지만, 단기 정책 대응이나 실시간 모니터링에서는 시의성이 더 중요하다. 따라서, 각 평가 기준에 목적별로 적절한 가중값을 부여해야 한다 (Valliant et al., 2018).

- **종합 점수 산출.** 각 평가 기준의 중요도(가중값)와 수행도(정규화 점수)를 결합해 종합 점수를 산출한다(Valliant et al., 2018). 종합 점수가 높은 자료원일수록 대표성과 품질이 균형 있게 확보된 기준 제공원으로 평가된다.

$$S_j = \sum_{k=1}^K w_k q_{jk}$$

w_k : 기준 지표(k)의 가중값($\sum_k w_k = 1$)

q_{jk} : 자료원(j)의 기준 지표(k)에 대한 정규화 점수

이러한 우선순위 선정 결과는 이후 보정 가중값 산출 과정에서 적용할 기준값의 결합 구조(weighted benchmarking)를 결정하는 근거가 된다.

나. 제약 조건 전환 및 보정 모형 설정

종합 점수는 제약(constraint)의 강도와 유형을 결정하는 기준으로 작동한다(Deville & Särndal, 1992). 즉, 자료원의 종합 점수가 높을수록 신뢰성과 대표성이 높다고 판단되어 강제 제약(hard constraint)으로, 그보다 낮은 경우는 별점 제약(soft constraint)과 참조 제약(reference constraint)으로 설정된다.¹⁰⁷⁾ 이렇게 설정된 계층 구조는 이후 보정

$$\Rightarrow CI = \frac{3-3}{3-1} = 0, \quad CR = \frac{0}{0.58} = 0$$

$\Rightarrow CR \leq 0.1$ 이면 일관성 허용, $CR > 0.1$ 이면 판단 불일치로 재검토

107) 보정 모형은 외부 기준값을 적용하는 방식에 따라 제약 조건의 강도를 달리 설정할 수 있다(Deville & Särndal, 1992; Valliant et al., 2018).

- **강제 제약(hard constraint):** 가장 신뢰성 있는 기준에 반드시 부합하도록 하는 제약으로, 오차를 허용하지 않는다(Deville & Särndal, 1992).

$$A_1 w = b_1$$

$A_1 = (a_{jk})$: 자료원(j)과 보정 기준 지표(k)의

대응 관계를 정리한 기준 행렬

$w = (w_j)$: 보정 후 가중값 벡터

$b_1 = (b_k)$: 외부 기준값 벡터

- **별점 제약(soft constraint):** 반드시 일치할 필요는 없으며, 일정 오차 범위 내 근사 일치를 허용하고, 오차 제곱합이 최소화하도록 별점(penalty term)을 부여한다(Valliant et al., 2018).

$$\lambda_2 \| A_2 w - b_2 \|^2$$

λ_2 : 별점계수, 클수록 $A_2 w$ 와 b_2 의 차이를 더 강하게 줄임

모형(calibration model)으로 전환된다. 보정 모형은 설계 가중값의 변동을 최소화하면서, 세 가지 제약(hard, soft, reference constraints)을 동시에 충족하도록 통합된 형태로 표현된다.¹⁰⁸⁾ 즉, 가중값의 최소 조정과 외부 기준의 최대 일치라는 두 목표를 동시에

- **참조 제약(reference constraint):** 대표성은 약하지만, 추세나 방향성 등 보조 정보를 반영하기 위한 제약이다.

$$\lambda_3 \| A_3 w - b_3 \|^2$$

λ_3 : 참조계수, 작은 값으로 설정해서 참고 수준만 반영

☞ **Hard**(총조사, 성별): $b_1 = (500, 520)$

Soft(행정자료, 직업별 비율): $b_2 = (0.3, 0.5, 0.2)$, $\lambda_2 = 50$

Reference(빅데이터, 소비 증가율): $b_3 = (0.4)$, $\lambda_3 = 5$

⇒ **Hard**

- 보정 전 추정값(남 = 490, 여 = 530), 총조사 값(남 = 500, 여 = 520)
- Hard 제약 적용(남: +10, 여: -10)으로 성별 합계는 정확히 맞춰짐

⇒ **Soft**

- Hard 제약 후, 직업 비율 추정값(0.28, 0.52, 0.20), 행정자료 기준(0.30, 0.50, 0.20)
- 벌점(Soft 제약): $\lambda_2 \| A_2 w - b_2 \|^2 = 50 (-0.02)^2 + 0.02^2 = 0.04$
- 벌점을 줄이기 위해 목푼(행정자료)값에 가깝게 직업 비율 조정:
(0.28, 0.52, 0.20) → (0.295, 0.505, 0.20)

⇒ **Reference**

- Hard+Soft 제약 후, 소비 증가율 추정값(0.34), 빅데이터 기준(0.40)
- 벌점(Reference 제약): $\lambda_3 \| A_3 w - b_3 \|^2 = 15 (-0.06)^2 = 0.018$
- 벌점을 줄이기 위해 소비 증가율을 목푼값으로 이동: 0.34 → 0.3789

108) 직관적으로 설명하면 보정 모형은 다음 세 원리를 따른다.

- ① **가중값 변동 최소화:** 초기 설계 가중값에서는 최소한으로만 움직인다.
- ② **등식 제약:** 강제 제약은 정확히 일치시킨다.
- ③ **벌점 함수 최적화:** 벌점 제약과 참조 제약은 조정계수(λ)로 조절해 외부 자료원과 유사하도록 유도한다.

본 연구에서는 KKT 조건(Karush-Kuhn-Tucker conditions)을 풀어 최종 보정 가중값을 산출하였다 (Deville & Särndal, 1992; Valliant et al., 2018).

☞ **설계 가중값:** $d = (0.34, 0.33, 0.33)^\top$, $D = \text{diag}(0.34, 0.33, 0.33)$

$$\Leftrightarrow D^{-1} = \text{diag}(2.941776, 3.030303, 3.030303)$$

$$\text{Hard: } A_1 = \begin{bmatrix} 500 & 480 & 510 \\ 520 & 540 & 510 \end{bmatrix}, \quad b_1 = \begin{bmatrix} 500 \\ 520 \end{bmatrix}$$

$$\text{Soft: } A_2 = \begin{bmatrix} 0.30 & 0.28 & 0.32 \\ 0.50 & 0.52 & 0.48 \\ 0.20 & 0.20 & 0.20 \end{bmatrix}, \quad b_2 = \begin{bmatrix} 0.30 \\ 0.50 \\ 0.20 \end{bmatrix}, \quad \lambda_2 = 50$$

$$\text{Reference: } A_3 = [0.40 \ 0.34 \ 0.38], \quad b_3 = [0.40], \quad \lambda_3 = 5$$

⇒ **KKT 조건:** 라그랑주 승수(Lagrangian multiplier, ν)로 조건 정리

달성하기 위한 최적화 문제로 구조화되는 것이다.

$$\min_w \frac{1}{2}(w-d)^\top D^{-1}(w-d) + \lambda_2 \|A_2 w - b_2\|_2^2 + \lambda_3 \|A_3 w - b_3\|_2^2 \quad s.t. \quad A_1 w = b_1$$

$d = (d_j)$: 보정 전 가중값 벡터(자료원별 초기 비중)

D : 거리 척도 행렬

다. 민감도 분석

보정 모형에서 설정한 제약 우선순위¹⁰⁹⁾나 조정계수(λ)의 선택은 결과에 중대한 영향을 미친다. 특히, 조정계수의 크기나 제약 강도의 변화에 따라 극단 가중값 발생 여부 또는 추정 분산이 민감하게 달라질 수 있다(Valliant et al., 2018). 따라서, 보정 모형을 확정하기 전에는 **민감도 분석**(sensitivity analysis)을 수행해 모형의 안정성을 점검해야 한다(Valliant et al., 2018).

- **우선순위 변화.** Hard, Soft, Reference의 우선순위를 달리 설정해서 최종 보정 가중값의 변화를 비교한다.
- **조정계수(λ) 변화.** λ_2, λ_3 를 조정($\pm 10\%$, $\pm 20\%$)하여 최종 보정 가중값의 변화를 비교한다.
- **극단 가중값 발생 여부 점검.** λ_2, λ_3 조정에 따른 **가중값 변동계수**(coefficient of variation of weights)¹¹⁰⁾를 점검한다.

$$\begin{bmatrix} H & A_1^\top \\ A_1 & 0 \end{bmatrix} \begin{bmatrix} w \\ v \end{bmatrix} = \begin{bmatrix} g \\ b_1 \end{bmatrix}$$

$$H = D^{-1} + 2\lambda_2 A_2^\top A_2 + 2\lambda_3 A_3^\top A_3$$

$$g = D^{-1}d + 100A_2^\top A_2 + 10A_3^\top A_3$$

$$\Rightarrow D^{-1}d \approx (1.000, 1.000, 1.000), 100A_2^\top A_2 = (32, 35, 33), 10A_3^\top A_3 = (1.6, 1.36, 1.52)$$

$$g \approx (34.6, 37.36, 35.52)$$

$$\Rightarrow H = \begin{bmatrix} 42.541 & 39.760 & 39.120 \\ 39.760 & 43.066 & 39.212 \\ 39.120 & 39.212 & 41.754 \end{bmatrix}, g = \begin{bmatrix} 40.600 \\ 40.760 \\ 40.120 \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} 42.541 & 39.760 & 39.120 & 500 & 520 \\ 39.760 & 43.066 & 39.212 & 480 & 540 \\ 39.120 & 39.212 & 41.754 & 510 & 510 \\ 500 & 480 & 510 & 0 & 0 \\ 520 & 540 & 510 & 0 & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ \nu_1 \\ \nu_2 \end{bmatrix} = \begin{bmatrix} 40.600 \\ 40.760 \\ 40.120 \\ 500 \\ 520 \end{bmatrix}$$

$$\Rightarrow \text{최종 보정 가중값: } w = (w_1, w_2, w_3)^\top = (0.367, 0.211, 0.422)$$

109) 자료원별 신뢰도 평가가 불완전할 수 있기에, “만약 우선순위가 달라지면 결과는 얼마나 변하는가?”를 확인할 필요가 있다.

- **결과 안정성 평가.** 주요 지표(평균, 비율, 회귀계수)의 변화율이 $\pm 5\%$ 이내면 ‘안정(robust)’, $\pm 10\%$ 이상이면 ‘민감(sensitive)’으로 판단한다.

4. 대표성 점검

대표성 점검(representative assessment)은 보정 가중값이 실제 모집단의 구조와 속성을 얼마나 충실히 반영하고 있는지를 평가하는 절차이다. Deville과 Särndal (1992)은 보정을 통해 외부 기준값과의 일치성을 확보할 수 있다고 강조했으나, 보정만으로 모든 대표성 문제가 해결되지는 않는다. 가중값 분포의 불안정성, 극단값의 존재, 특정 집단의 과소·과대 대표 등은 여전히 발생할 수 있으며, 이러한 문제는 대표성 점검을 통해 확인되고 수정되어야 한다.

한편, 대표성 점검의 결과는 외부 기준자료의 품질에 따라 결과가 달라진다는 점에서 한계가 있다(Zhang, 2012). 예를 들어, 인구총조사가 시점 차이나 등록 누락 문제를 내포하고 있다면, 해당 자료를 기준으로 수행한 대표성 점검 역시 제한적일 수밖에 없다. 또한, 여러 외부 기준을 동시에 충족시키려다 보면 극단 가중값이 증가할 위험이 있다. 따라서, 대표성 점검은 단순한 확인 절차가 아니라, **보정 변수 선택과 제약 설정을 재검토하는 환류(feedback)**로 이해해야 한다.

가. 제약 충족 여부 및 극단 가중값 확인

대표성 점검의 첫 단계는 보정 과정에서 설정된 제약이 실제로 충족되었는지 검증하는 것이다. 강제 제약은 외부 기준값과 정확히 일치해야 하며, 별점 제약과 참조 제약은 사전에 설정한 허용 오차 범위 내에 있어야 한다. 가령, 인구총조사에서 20대 남성 인구가 100만 명일 때, 보정 후 가중값을 적용한 추정값도 정확히 100만 명이어야 한다. 이와 함께, 극단 가중값 발생 여부를 점검해야 한다. 일부 개체에 과도한 가중값이 부여되면 대표성을 왜곡하고 추정 분산을 급격히 증가시킬 수 있다. Valliant 등(2018)은 평균 가중값의 4배를 초과하는 단위를 극단 가중값으로 정의하고, 극단 가중값 비율(ER)을 대표성 점검 지표로 활용할 것을 제안하였다.

$$110) CV(w) = \frac{\sqrt{\frac{1}{n} \sum_i (w_i - \bar{w})^2}}{\bar{w}}$$

나. 보정 가중값 분포의 안정성 평가

보정 이후에도 가중값이 불안정하게 분포한다면, 일부 단위에 추정값이 과도하게 의존하여 대표성이 훼손될 수 있다. Kish(1999)는 이를 지적하며, 보정 후에는 가중값 분포의 안정성(weight stability)을 별도로 점검해야 한다고 강조하였다. 가중값 분포의 안정성은 **가중값 변동계수**(coefficient of variation of weights)와 **유효 표본 크기**(effective sample size)로 점검할 수 있다. 가중값 분포 안정성은 보정된 통합 데이터 세트가 전체 모집단을 안정적으로 대표할 수 있는지 검증하는 실증적 근거로 활용된다.

다. 추정값과 외부 기준 비교

보정 가중값을 적용해 산출한 추정값을 외부 기준자료와 비교함으로써 대표성 확보 여부를 평가한다. 예컨대, 보정 후 추정된 20대 고용률이 65%이고, 외부 기준자료의 값이 66%일 경우, 1%^p 차이는 허용 범위로 판단할 수 있다. 그러나, 10%^p 이상의 차이가 난다면 보정 변수 선정이나 제약 설정 과정에 오류가 있음을 의미하며 재검토가 필요하다. 이 단계에서는 단순 평균·비율 비교 외에도 분포 유사도 지표(예: TVD, JSD)를 활용해 전체 구조의 일치도를 정량적으로 평가하는 방법도 고려할 수 있다.

라. 시계열 일관성 점검

대표성은 횡단면뿐 아니라 시계열에서도 유지되어야 한다. 보정 이후 추정값이 연도 간 급격하게 변동하거나 불연속성을 보인다면, 이는 가중값 불안정이나 기준값 단절로 인한 시계열 불일치일 가능성이 크다. 이를 점검하기 위해 다음 절차를 수행한다.

- **연도별 추정값 비교.** 선형 그래프와 기술통계를 통해 급격한 변화를 탐지한다.
- **연간 변동률¹¹¹⁾ 검토.** 변동률 차이가 ±10% 이상이면, 시계열 단절일 가능성이 크다.
- **시계열 조정.¹¹²⁾** 연결계수나 재기준화 등을 활용하여 불연속 구간을 조정한다.

111) $\Delta_t = \frac{Y_t - Y_{t-1}}{Y_{t-1}}$

112) **예** 통합 데이터 세트 추정값: $int_{2022} = 100$ 만, $int_{2023} = 105$ 만, $int_{2024} = 160$ 만

외부 기준자료: $out_{2022} = 100$ 만, $out_{2023} = 104$ 만, $out_{2024} = 108$ 만

⇒ 변동률 비교: $\Delta_{int} = 52\%$, $\Delta_{out} = 4\%$

⇒ 연결계수($\frac{108}{160} = 0.675$) 적용하여 외부 기준값과 일치

시계열 일관성 점검은 통합 데이터 세트가 단일 시점뿐 아니라 장기 분석에도 신뢰성 있게 활용될 수 있도록 하는 품질 확보 절차로 기능한다.

제5절 최종 품질 점검

최종 품질 점검은 단순히 통합 데이터 세트의 오류를 식별하거나 기술적 결함을 교정하는 절차가 아니다. 이는 데이터 통합의 모든 단계를 종합적으로 검토하여, 통합 데이터 세트가 신뢰성과 활용 적합성(fit-for-use)을 충족하는지를 평가하는 최종 진단 과정이다. 즉, 품질 점검은 데이터 통합의 마지막 절차이자, 전체 통합 과정의 타당성과 일관성을 검증하는 품질 관리체계의 핵심으로 기능한다.

본 연구에서 제시한 데이터 통합 과정은 전처리 및 정합성 점검(제1절), 결합(제2절), 결측값 대체(제3절), 보정 및 대표성 점검(제4절), 최종 품질 점검(제5절)으로 구성된다. 각 단계에는 고유의 품질 점검 절차가 내재해 있으며, 이는 통합 과정의 정합성, 일관성, 논리적 타당성을 확보하기 위한 과정 중심(process oriented)의 관리 기능을 수행한다. 반면, 최종 품질 점검은 결과 중심(result oriented) 접근으로서 단계별 점검 결과를 종합하여 통합 데이터 세트 전체의 품질 수준을 평가하고 개선 방향을 제시한다. 데이터 통합 목적과 자료 성격에 따라 일부 단계는 반복되거나 생략될 수 있으나, 최종 품질 점검은 모든 통합 유형에서 공통으로 수행되는 필수 절차이다.

국제기구들은 데이터 통합 품질의 핵심 속성으로 정확성(accuracy), 완전성(completeness), 정합성(consistency), 시의성(timeliness), 대표성(representativeness)을 공통으로 제시하고 있다. UNECE(2024)는 이러한 속성들이 통합 데이터 세트의 품질 평가 시 반드시 충족되어야 한다고 강조하였으며, ESCAP(2020)은 품질 확보를 위해 사전 품질 진단(pre-assessment)과 사후 품질 점검(post-assessment)을 병행할 것을 권고하였다. Eurostat(2021b)은 품질 점검 결과를 메타데이터 기반의 품질 보고서 형태로 공개함으로써 투명성(transparency)과 재현성(reproducibility)을 확보할 것을 제안하였으며, UNSD(2022)는 다출처 기반 통계생산에서 정확성, 완전성, 정합성, 시의성, 대표성의 충족 여부를 통해 통합 데이터의 공식 통계 활용 가능성을 평가할 것을 권고하였다. 결국, 최종 품질 점검은 데이터 통합의 종결 단계이자 품질 관리 체계의 환류(feedback) 기능을 수행하는 절차로서, 단계별 점검을 통해 축적된 정보를 통합적으로 진단하고, 향후 데이터 통합 과정의 지속적 개선(continuous improvement)을 위한 근거를 제공한다.

앞서 살펴본 바와 같이, 국제기구들은 통합 데이터 세트의 품질 평가를 위해

정확성, 완전성, 대표성, 정합성, 시의성을 핵심 속성으로 제시하고 있다. 이에 본 연구에서는 각 품질 속성을 정의하고, 국제 기준과 선행연구에서 제시한 평가 기준을 중심으로 점검 지표를 제시한다.

- **정확성(accuracy)**. 통합 데이터 세트가 모집단을 얼마나 충실히 반영하고 있는지를 외부 기준자료와 비교하여 평가한다. Parker 등(2024)은 혼합자료(blended data) 기반 노동시장 통계에서 외부 기준자료와의 $\pm 5\%$ 이내 차이를 허용 기준으로 제안하였다.
- **완전성(completeness)**. 통합 데이터 세트가 모집단을 충분히 포괄하고 있는지를 평가한다. 표면적인 데이터 세트의 크기만이 아니라, 소집단의 과소 대표 여부까지 검토해야 하며, 포괄률과 결측률로 진단한다.
- **대표성(representativeness)**. 통합 데이터 세트의 구조가 모집단의 구조와 얼마나 일치하는지를 평가한다. 이는 가중값 분포 안정성, 유효 표본 크기 산출, 외부 기준과의 분포 일치도 등으로 측정한다.
- **정합성(consistency)**. 시간적·논리적 정합성을 점검한다. Eurostat(2021b)은 연도별 추정값 변동률이 $\pm 10\%$ 를 초과할 경우, 시계열 재검토를 규정하였다. 또한, 변수 간의 논리적 제약(예: 성별 = 여자, 임신 = 예) 유지 여부도 핵심 점검 항목이다.
- **시의성(timeliness) 및 접근성(accessibility)**. 통합 데이터 세트의 활용 가치를 결정하는 요인이다. UNSD(2022)는 시의성을 데이터 품질의 핵심 속성으로 정의하면서, 데이터 제공 시점이 정책 수요 시점과 얼마나 부합하는지를 측정할 것을 권고하였다. 한편, Eurostat(2021b)은 품질 평가 결과와 메타데이터 보고서를 온라인에 공개하여 이용자가 데이터의 활용 가능성을 사전에 파악할 수 있도록 하고 있다. 이는 통계 데이터 세트의 품질 점검이 내부 관리 차원이 아니라, 외부 이용자와의 신뢰 관계를 강화하는 과정임을 보여준다.

제 4 장

통계생산 기관의 데이터 통합 사례

본 연구에서 제시한 데이터 통합 과정—^①자료 수집, ^②전처리 및 정합성 점검, ^③결합, ^④결측값 대체, ^⑤보정, ^⑥품질 점검—을 기준으로 주요 국제기구 및 국가의 사례를 검토하였다.¹¹³⁾ 나아가 제3장에서 제시한 전처리 원칙¹¹⁴⁾과 품질 점검 기준(정확성, 일관성, 완전성, 시의성, 투명성 등)을 준거 틀로 삼아 각 기관이 전처리와 품질 점검을 어느 수준으로 구현했는지 고찰하였다.

제1절 국제기구별 사례

1. UN, SDG Global Database

UNSD는 지속가능발전목표(SDGs) 모니터링을 위해 전 세계 회원국으로부터 제출된 자료를 수집·조정·통합한다. 회원국이 보고하는 자료는 정의와 범위가 다르고, 일부 국가는 역량 부족으로 미제출하거나 불완전한 자료를 제출하기도 한다. 이에 UNSD는 국가별 자료와 국제기구 추정값을 종합하여 SDG Global Database를 구축하였다(UNSD, n.d.-b). 이 과정은 단순 집계에 그치지 않고, 자료 간 개념·분류 조화와 결측 대체를 포함하는 데이터 통합 과정이다.

113) 본 연구에서 제시한 데이터 통합은 국제기구(ESCAP, UNECE 등)가 제시한 핵심 구성 요소와 정합적이다. 하지만, 데이터 통합 절차는 사례별로 다양하기에(UNECE, 2020), 본 연구에서 제안한 데이터 통합 과정을 기준으로 재구성하였다.

114) 전처리 기본 원칙

- 원자료의 정보 손실을 최소화하는 **최소침습성**
- 모든 변환을 문서로 관리하여 누구나 동일 결과를 재현할 수 있도록 하는 **재현성**
- 변수 정의·단위·분류 등 변환의 근거를 기록하는 **메타데이터 우선 원칙**
- 개인정보 보호, 접근권한 등 **윤리적·법적 요건 준수**
- 「스키마 정렬 - 개념 조화 - 정합성」을 순차적 혹은 병행하여 **상호운용성**

가. 자료 수집 및 전처리

UNSD는 세 가지 경로로 자료를 수집한다. 첫째, 각국이 행정자료와 조사자료를 활용해 산출한 국가통계이다. 둘째, WHO·ILO·UNICEF 등 국제기구가 수집·보정한 자료이다. 셋째, World Bank 등 국제금융기구의 자료가 일부 지표에 활용된다(UNSD, n.d.-b). 전처리의 핵심은 국가 간 자료의 구조와 개념 차이를 최소화하는 것이다. 먼저, 변수명, 코드 체계, 단위 등을 표준화한다. 성별은 ISO/IEC 5218, 날짜는 ISO 8601, 자료 구조는 SDMX¹¹⁵⁾로 일치시킨다. 이어서, 변수 정의를 국제 기준에 맞게 통일한다. 예컨대, A 국가는 빈곤율을 소득 기준, B 국가는 소비 기준으로 측정하지만, 국제 기준에 따라 통일된 정의와 변환 규칙을 적용한다. 마지막으로 정합성 점검을 수행한다. 급격한 수치 변동, 물리적으로 불가능한 값(예: 실업률 120%), 국가별 자료와 국제기구 자료 간 과도한 차이는 불일치 사례로 기록·검토된다(UNSD, n.d.-b).

나. 결합 및 결측값 대체

UNSD는 자료를 직접 결합하지 않고, 국가별 통계를 국제 기준에 맞게 변환·조정하는 **통계적 조화**(statistical harmonization)를 수행한다(UNSD, 2021). 예를 들어, 빈곤율 지표는 국가마다 산정 기준이 다르지만, 국제 기준(1.90달러/일 기준 구매력)을 적용해 조정한다(World Bank, 2020). 이는 레코드 연계나 통계적 매칭이 아닌 정의·단위·측정 방식을 변환하는 유형의 결합으로 볼 수 있다.

해당 통계가 없거나 불완전한 경우, UNSD는 추정값을 사용한다. 예를 들어, 아동 사망률은 B3(Bayesian B-spline Bias-reduction) 모형으로, 모성사망률은 베이지안 계층모형으로 결측값을 추정한다(UNFPA, 2025; UNICEF, 2024). 지역 또는 소득별 평균을 활용하기도 하지만, 이는 공식 추정정보는 보조적 성격을 가진다. 결측값 대체는 시계열 완결성을 보장하지만, 동시에 추정값의 불확실성을 수반한다. 이를 보완하기 위해 각 지표를 **관측값**(observed)과 **추정값**(estimated)으로 **명확히 구분**해 표시한다(UNSD, 2021).

다. 보정 및 품질 점검

국가별 통계와 국제기구 추정값이 불일치할 경우, UNSD는 국가 협의 절차를 거쳐 수치를 검토·조정한다. 예를 들어, 일부 국가에서는 국내 빈곤율과 세계은행 추정값 간 괴리가 발생하는데, 이때 해당 국가 통계청과 국제기구가 협의하여 최종 수치를 확정한다(UNSD, 2021). 국제 비교 가능성을 확보하기 위해 인구 규모를 반영한 표준화

115) SDMX(Statistical Data and Metadata Exchange)는 UN, OECD, IMF, World Bank 등이 공동으로 개발한 국제표준 데이터 교환 체계/framework로 지표 코드, 변수명, 단위, 시점 등을 사전에 정의하는 방식으로, 국가별 자료를 구조적으로 정렬하는 데 유용하다(UNSD, n.d.-c).

(예: 연령표준화 사망률)와 환율 및 PPP(Purchasing Power Parity) 기반 보정이 적용된다. 이는 국가 단위의 대표성을 유지하면서도 동시에 국제 비교 가능성을 담보하기 위한 장치이다.

UNSD는 직접 품질 점검을 수행하지 않고, SDG 지표의 주관 기구(custodian agencies)에서 품질을 관리한다. 보건 지표는 WHO, 아동·교육 지표는 UNICEF, 노동 지표는 ILO가 국가별 보고자료를 검증한다. 이들은 각 지표의 개념·정의·범위가 국제 기준과 일치하는지 확인하며, 필요한 경우 보정이나 모형 기반 추정을 병행한다(UNSD, 2021; WHO, 2020).

라. 전처리 원칙 및 품질 점검 기준에 대한 고찰

UNSD는 회원국으로부터 수집한 다양한 통계를 SDG 지표에 맞추어 조정·통합한다. 이 과정에서 가장 두드러지는 점은 자료의 이질성을 국제표준에 따라 정렬하는 전처리 절차이다. 국가별로 다른 변수명, 코드 체계, 단위를 국제표준으로 통일한다. 이는 스키마 정렬과 개념 조화를 통해 상호운용성을 확보한 사례로 평가할 수 있다. 그러나, 원자료 자체의 품질 편차가 크고, 일부 국가는 불완전하거나 누락 자료를 제출하기 때문에 최소침습성 원칙과 재현성 확보에는 한계가 있다. 결측값은 모형 추정으로 대체되며, 이는 국제 비교 가능성을 높이는 동시에 원자료의 불확실성을 확대하는 요인으로 작용한다.

품질 관리 체계 또한 분산되어 있다. 지표별 주관기관이 자료를 검증하고 필요시 보정과 추정을 병행함으로써 정확성과 일관성은 확보되지만, 통합적 품질 관리 체계는 미비하다. 결측 대체와 보정 절차가 메타데이터에 표시되지만, 모든 지표가 동일 수준의 문서화와 투명성을 보장하지는 않는다. 결론적으로 SDG Global Database는 국제 비교 가능성을 위한 최소한의 품질 기준은 충족하지만, 완전성과 재현성 측면에서 구조적 제약이 존재한다.

마. 시사점

SDG Global Database는 전 세계 데이터를 통합해 SDG 모니터링의 국제 기준을 마련하고, 메타데이터와 품질 정보를 공개함으로써 데이터 활용의 투명성을 제고하였다(UNSD, n.d.-d). 이는 데이터 통합을 통한 국제 비교 가능성을 실질적으로 구현한 사례라 할 수 있다. SDG Global Database는 데이터 통합의 성공과 어려움을 동시에 보여주며, 국내 데이터 통합에도 중요한 시사점을 제공한다. 첫째, 지표별 관측값과 추정값을 병행 제공하는 방식은 통합에서 불가피하게 발생하는 불일치·결측·보정의 투명한 공개라는 측면에서 의의가 있다. 이는 통합 데이터 세트를 제공할 때 「원자료-보정 데이터-통합 데이터」의 다층적 구조를 함께 공개하는 방식으로

적용할 수 있다. 둘째, 지표별 정의, 산출 방식, 단위, 자료 출처, 담당 국제기구, 자료 가용성 등을 표준화된 형식으로 기록하는 **SDG Metadata Repository**와 같은 체계적 인프라를 도입해 데이터 통합 과정의 메타데이터를 공개함으로써 통합 데이터 세트의 신뢰성을 높일 필요가 있다.

2. OECD

OECD는 회원국의 통계청(National Statistical Institutes, NSIs), 중앙은행, 정부 부처 등에서 생산한 자료를 결합하여 데이터베이스를 구축해 왔다(OECD, 2019a, 2019b).

가. 자료 수집 및 결합

OECD는 회원국 제출 자료, 국제기구 제공 지표, 민간 데이터를 수집하여 OECD.Stat에 통합한다. 회원국은 SDMX 표준에 따라 변수 정의·분류·단위를 통일하여 메타데이터와 함께 제출한다. OECD는 이를 공통 메타데이터 지침에 맞추어 정렬한다(OECD, 2019a). 다만, 국가별 자료를 직접 결합하지 않고, 통계적 조화를 통해 비교 가능성을 확보한다(OECD, 2013, 2019a, 2023a). 예를 들어, 보건 통계에서는 OECD Health Accounts 체계를 적용하여 국가별 보건 지출 자료를 공통 분류 기준으로 변환한다(OECD, 2023a).

나. 결측값 대체

국가별 보고 지연, 조사 미 실시, 행정자료 부재 등으로 결측이 발생하면, OECD는 세 가지 방법으로 이를 대체한다. 첫째, 보간·외삽을 활용해 시계열을 연결한다(OECD, 2019b). 예를 들어, A 국가의 고용률이 2022년과 2024년에는 보고되었으나 2023년 값이 누락된 경우, 선형 보간으로 해당 연도 값을 추정한다. 둘째, GDP와 인구 구조를 독립변수로 설정한 회귀모형으로 결측값을 산출한다(OECD, 2023b). 셋째, 소득 수준이 유사한 국가 집단의 평균을 활용해 결측을 보완한다. 결측값 대체 결과는 지표 메타데이터에 기록되어 해당 값이 국가별 자료인지, OECD 추정값인지를 구분할 수 있다. 이는 데이터 통합의 투명성을 확보하는 장치로 기능한다.

다. 보정 및 품질 점검

국가별 자료는 수집 방법과 분모 차이로 인해 직접 비교가 어렵기 때문에 다양한 보정 방법이 적용된다(OECD, 2019b). 표준 분모를 사용하여 국가별 경제 규모와 인구

구조의 차이를 반영한다. 예컨대, 인구 관련 지표에는 WPP(World Population Prospects), 가격 관련 지표에는 PPP를 적용한다. 자료 간 편향을 교정하기 위해 국가별 가중값을 적용한다. 가령, 조사자료의 표본오차가 큰 경우에는 낮은 가중값을 부여한다. 국가별 자료와 보정값 사이에 괴리가 발생하면, 보정값을 공식 지표로 삼고 국가별 자료는 보조 자료로 병기한다(OECD, 2019b).

OECD는 자료 수집과 제공 전 과정에 QAF(Quality Assessment Framework)를 적용한다. QAF는 정확성, 일관성, 시의성, 접근성 등을 기준으로 평가하며, 제출 단계에서 오류 탐지 절차를 거친다. 급격한 변동이나 비정상 값이 발견되면 회원국에 재검증을 요청한다. 또한, 모든 지표에는 메타데이터를 첨부해 출처, 산출 방식, 보정 절차를 문서화한다(OECD, 2019b).

라. 전처리 원칙 및 품질 점검 기준에 대한 고찰

OECD는 회원국으로부터 제출받은 자료를 SDMX 표준과 공통 분류체계에 맞추어 정렬·변환한다. 이는 스키마 정렬과 개념 조화를 제도적으로 내재화한 것으로 최소침습성과 메타데이터 우선 원칙을 충족한다. 또한, 제출 단계에서 메타데이터를 필수적으로 포함하도록 하여 자료의 정의·단위·출처를 체계적으로 문서화함으로써 재현성과 상호운용성을 보장한다.

품질 관리 측면에서 QAF를 전 과정에 적용한다. 제출 단계에서 오류 탐지가 수행되고, 급격한 수치 변동이나 비정상 값은 회원국에 재검증을 요청한다. 이는 정확성과 일관성 확보에 이바지한다. 결측이 발생하면 보간·외삽, 회귀모형, 유사 국가 평균 등을 활용하며, 추정값 여부를 명시하여 투명성을 확보한다. 또한, WPP·PPP와 같은 국제 기준을 적용해 국가 간 비교 가능성을 높이고, 조사자료의 표본오차가 큰 경우 가중값을 조정하여 대표성을 확보한다. 이는 통합 데이터 세트의 국제 비교 가능성과 활용도를 동시에 높인다.

마. 시사점

OECD의 데이터 통합은 국제 비교 가능성을 위한 표준화와 보정체계의 제도화로 요약된다. SDMX 표준에 기반한 자료 수집, WPP·PPP 적용은 국가 간 이질성을 줄이고 비교의 신뢰도를 높였다. 그러나, 국가별 자료와 OECD 추정값 간 괴리, 저소득 국가의 높은 결측률은 보정 의존도를 높여 과소 혹은 과대평가 위험을 초래할 수 있다. 이러한 한계에도 불구하고, **메타데이터 기반의 투명한 절차 공개와 체계적 품질 점검**은 국내에 적용할 필요가 있다. 특히, QAF와 같은 품질 보증 체계를 도입한다면, 통합 데이터 세트의 신뢰성과 활용도를 높일 수 있을 것이다.

3. Eurostat

Eurostat은 각국 통계청(NSIs)이 제출한 자료를 기반으로 통계를 생산하며, ESS (European Statistical System)라는 제도적 틀 안에서 운영된다.¹¹⁶⁾ ESS는 Eurostat, NSIs, EU 기관이 협력하여 구축된 네트워크로, Eurostat이 표준화·통합·품질 관리를 총괄한다(Eurostat, 2014, 2017). Eurostat의 데이터 통합은 법적 규범, 국제표준, 품질 보증 체계를 근거로 「수집-전처리-결합-결측값 대체-보정」을 포함하는 종합적 과정이다(European Parliament & Council, 2024).

가. 자료 수집 및 전처리

Eurostat의 자료 수집은 법적 근거에 따른 의무적 제출을 기반으로 한다(Eurostat, 2014, 2017, 2021b). 예컨대, 국민계정은 ESA 2010(European System of Accounts), 노동력조사는 EU-LFS(Labour Force Survey), 가계소득과 생활 여건은 EU-SILC(Statistics on Income and Living Conditions)에 따라 수집된다. 이는 국가별 자발적 보고가 아니라 EU 규정에 따라 정기적으로 제출하는 방식이다.

전처리는 세 단계로 진행된다. 첫째, SDMX 표준을 채택하여 제출 자료를 구조적으로 정렬한다. 이 과정에서 변수명, 단위, 코드 체계는 국제표준(ISIC, ISCO, ISCED, NACE, NUTS)에 맞춰 일관되게 변환된다(Eurostat, 2017). 둘째, 공통 기준에 맞추어 변수의 개념을 통일한다. 예를 들어, 고용률 산출 시 일부 국가는 자국 노동법 정의를 따르지만, Eurostat은 ILO 권고에 따른 ‘15세 이상 인구 기준 고용률’을 적용한다. 셋째, 시계열 연속성, 논리적 일관성, 국가 평균 대비 극단값 등을 점검하여, 오류가 발견되면 회원국에 수정 보고를 요청한다(Eurostat, 2014, 2017, 2021b, 2023).

116) EU는 2024년 11월 27일 개정된 「Regulation (EU) 2024/3018」을 통해, 행정자료의 접근, 이용 및 통합 권리를 명확히 제도화하였다(European Parliament & Council, 2024). 특히, Article 17a 「Access to and use and integration of administrative data for the development, production and dissemination of European statistics」에서는 다음과 같이 규정한다. “국가법상 행정자료를 보유한 공공 및 준공공 기관은, 유럽통계의 개발·생산·배포를 위해 필요한 경우, 국가 통계 기관과 기타 국가 통계 당국이 해당 자료에 무료로 적시에 충분한 빈도와 세분성으로 접근·사용·통합할 수 있도록 허용해야 한다(National public and semi-public bodies under national law in charge of administrative data sources …shall allow the national statistical institutes (NSIs) and other national authorities to **access, use and integrate** those data free of charge in a timely manner, with sufficient frequency and granularity).” 이 조항은 데이터 통합을 단순한 행정적 재량이 아닌 법적 의무(legal obligation)로 격상시킨 것으로, 회원국의 행정자료 보유기관이 통계 목적의 통합을 거부할 수 없음을 명확히 한다. 즉, 데이터 통합은 통계 목적이라는 공익 아래 법적으로 보장된 접근·이용·통합 행위로 제도화된 것이다.

나. 결합

Eurostat은 개별 수준의 미시자료를 직접 결합하는 경우는 제한적이지만, EU-SILC와 같은 프로젝트에서는 국가별 행정자료와 조사자료를 연계한다. 예컨대, 일부 회원국은 세무 자료나 사회보장자료를 EU-SILC 표본조사와 결합하여 소득·복지 지표를 보강한다(Eurostat, 2020). 데이터 결합은 기본적으로 코드 기반의 결정적 연계로 수행한다. 국가코드, 지표 코드, 연도 단위를 연계 키로 사용한다. 그러나 대부분은 통계적 조화(statistical harmonization)를 수행한다. 국가별 조사자료의 변수를 표준 정의와 분류체계에 맞게 변환하는 개념적 통합을 지향한다.

다. 결측값 대체

Eurostat은 결측 문제를 해결하기 위해 조사별 세부 지침을 제공하고, ESS 차원에서 결측값 대체 규정을 마련하였다(Eurostat, 2013, 2014). 회원국이 특정 연도의 자료를 제출하지 못하거나 부분적으로 누락된 경우, 모형 기반 추정이나 시계열 보간을 활용한다. 예를 들어, 에너지 소비량 자료가 일부 연도에서 결측된 경우, 인접 연도의 추세나 관련 변수(예: GDP, 산업생산지수)를 활용해 대체 값을 산출한다(Eurostat, 2021b). 결측 대체 과정은 메타데이터에 기록되며, 대체 여부는 데이터 플래그로 표시된다. 이러한 절차는 데이터 통합의 투명성을 확보하고, 국가별 자료 불균형으로 인한 왜곡을 최소화한다.

라. 보정 및 품질 점검

Eurostat(2013, 2014, 2021a, 2021b)은 자료의 대표성을 확보하기 위해 가중값과 PPP를 적용한다. LFS와 SILC는 인구총조사를 기준으로 가중값을 부여하여 모집단을 대표하도록 하고, 지출 및 소득자료는 PPP를 적용해 국가 간 가격 수준 차이를 조정한다. 또한, 시계열 단절을 방지하기 위해 자료 수집 방식이나 분류체계가 개편될 경우, 과거 자료를 재작성하여 일관성을 유지한다. 이러한 보정은 국가별 다양성을 인정하면서도 EU 차원에서 일관된 비교 가능성을 확보하는 데 이바지한다.

품질 관리는 ESCP(European Statistics Code of Practice; Eurostat, 2017)와 ESS QAF(Eurostat, 2014, 2021b)에 근거한다. Eurostat은 지표별 정확성, 시의성, 일관성, 비교 가능성, 접근성을 검토하며, 급격한 수치 변동이나 합계와 부분값 간 불일치를 점검한다. 문제가 발견되면 회원국에 통보하여 재검증 후 수정된 자료를 제출하도록 한다. 최종적으로 지표별 메타데이터를 공개해 산출 방법, 자료원, 보정 절차를 명시한다.

마. 전처리 원칙 및 품질 점검 기준에 대한 고찰

Eurostat은 자료를 SDMX 표준에 따라 정렬하고, 국제표준에 맞게 표준화 후,

시계열 연속성, 논리적 일관성, 국가 평균 대비 극단값 등을 점검한다. 이는 「스키마 정렬-개념 조화-정합성 점검」을 포함해 전처리 원칙을 엄격히 적용한 결과이다. 전처리는 원자료 수집 방식을 존중하면서도 표준화된 분류체계와 코드로 변환함으로써 최소침습성을 충족한다. 제출과 검증 절차는 법령과 표준 문서로 규정되어 있어 재현성이 높으며, 지표별 산출 방식과 자료의 세부 내용이 메타데이터로 함께 공개되어 메타데이터 우선 원칙을 구현한다. 개인정보 보호는 EU 법적 규범과 ESCP를 근거로 보장되어 윤리적·법적 요건 준수에도 부합한다. 마지막으로 SDMX와 국제 분류체계 기반의 정렬은 상호운용성 확보로 이어진다.

품질 점검은 ESS QAF와 ESCP에 기반하며, 지표별 정확성, 시의성, 일관성, 비교 가능성, 접근성을 검토한다. 급격한 수치 변동이나 총합과 부분값 간 불일치가 발견되면 회원국에 재검증·수정을 요청해 정확성과 일관성을 확보한다. 결측값은 모형 기반 추정이나 보간으로 대체하며, 해당 사실은 데이터 플래그와 메타데이터에 기록되어 투명성이 보장된다. 자료는 법정 제출 주기에 따라 정기적으로 갱신되므로 시의성 역시 유지된다. Eurostat은 품질 보증 체계와 국제표준 적용이 모범적으로 제도화된 사례라 할 수 있다.

바. 시사점

Eurostat은 법적·제도적 기반 위에서 국가별 자료를 통합하여 정책 집행에 필요한 통계를 생산하고 있다. 특히, 국제 분류체계의 적용과 체계적 품질 관리는 모범적인 데이터 통합 사례로 볼 수 있다. Eurostat의 데이터 통합이 주는 교훈은 세 가지로 요약된다. 첫째, 법적 근거에 기반한 데이터 통합체계가 필요하다. 자료 제출 의무와 분류체계의 엄격한 적용은 국내 데이터 통합에서도 필수적이다. 둘째, 결측 대체에 관한 표준 지침을 마련해 통합 절차의 일관성을 강화해야 한다. 셋째, ESS QAF와 같은 품질 보증 체계를 도입하여 통합 데이터 세트의 신뢰성과 국제 비교 가능성을 높여야 한다.

4. IMF

IMF는 회원국의 통계와 국제기구·시장 자료를 결합해 데이터베이스를 구축한다(IMF, 2023a, 2023b). 또한, 회원국 자료의 품질 향상을 위해 SDDS(Special Data Dissemination Standard)와 e-GDDS(enhanced General Data Dissemination System)를 운영한다. 이는 회원국이 국제 기준에 맞춰 자료를 제출하도록 유도하고, IMF가 이를 조정·통합할 수 있는 제도적 기반을 제공한다(IMF, 2023b).

가. 자료 수집 및 결합

IMF는 회원국 중앙은행, 재무부, 통계청 등을 통해 제출되는 자료를 정기적으로 수집한다. 자료는 SDMX 표준으로 제출되며, 국제 비교 가능성을 확보하기 위해 통화 단위, 가격 수준, 기준연도 등을 통일한다. 표준화된 자료는 IFS(International Financial Statistics)와 GFS(Government Finance Statistics) 같은 데이터베이스에 구조화(schema alignment)된다(IMF, 2014, 2023a, 2023b).

IMF는 데이터를 직접 결합하지 않으며, 국제표준 매뉴얼이 규정하는 개념·분류·측정 규칙에 따라 국가별 자료를 통계적으로 조화한다. 예를 들어, 국민계정은 SNA 2008, 재정 통계는 GFSM 2014, 대외거래 통계는 BPM6에 따라 보고하도록 요구한다. IMF는 제출된 자료가 이러한 기준에 부합하는지 점검하고, 불일치가 발견되면 회원국과 협의하여 정합성을 확보한다(IMF, 2014, 2019). IMF의 데이터 통합은 이질적인 국가통계의 개념·분류·회계 규칙을 일치시키는 방식으로 달성된다.

나. 결측값 대체

IMF 데이터베이스는 회원국 자료의 불완전성이나 보고 지연으로 인해 결측이 빈번하게 발생한다. 이에 따라 IMF는 다양한 방법을 활용하여 결측값을 보완한다. 가장 일반적인 방법은 모형 기반 추정으로 국제수지나 국제투자대조표에서 결측이 발생하면, 회귀모형과 시계열 추정을 통해 보완한다(IMF, 2019). 또한, IMF는 무역 통계의 결측이나 불일치 문제를 해결하기 위해 거울 자료(mirror data)를 활용한다. 이는 세관 자료에서 거래 상대국의 수출입 정보를 활용하여 자국의 미신고 자료를 추정하는 방법이다(Geourjon et al., 2023).

다. 보정 및 품질 점검

IMF는 국가 간 비교를 확보하기 위해 환율 대신 PPP 적용, 규모 표준화, 기관 부문 경계 정렬 등의 보정을 권고·적용한다. 예컨대, 정부 재정 지표는 GFSM 2014의 범주(수입·지출·자산·부채)와 기관 부문 경계를 일치시켜 총량·비율 지표의 비교 가능성을 높인다(IMF, 2014). 국제수지 통계에서는 BPM6 표준에 따라 자산·부채·거주자 개념을 통일해 범위(coverage) 격차를 축소한다(IMF, 2019). 엄밀히 말하면 이러한 접근은 보정을 통한 대표성 확보라기보다 개념·범위 일치화에 가깝지만, 결과적으로 대표성 강화에 기여한다.

품질 점검을 제도화하기 위해 DQAF(Data Quality Assessment Framework)를 운영한다. DQAF는 정확성, 시의성, 일관성, 접근성, 신뢰성 등을 기준으로 품질을 평가한다(IMF, 2012b). 또한, SDDS와 e-GDDS를 통해 회원국이 자료를 공시하도록 요구하고, 각국의 이행 상태를 공개적으로 평가한다. 이러한 절차는 회원국이 국제

기준을 준수하도록 유도하며, 데이터베이스의 품질을 제도적으로 담보한다(IMF, 2023b).

라. 전처리 원칙 및 품질 점검 기준에 대한 고찰

IMF는 국제표준 제시와 준수 감시에 중점을 둔다. 이는 IMF의 데이터 통합이 자료를 재가공하기보다 SNA 2008, GFSM 2014, BPM6 등 매뉴얼에 맞추어 회원국 통계를 정렬·보완하는 구조로 자리 잡은 이유이기도 하다. 이러한 특수성은 전처리 원칙과 밀접하게 연결된다. IMF는 회원국이 산출한 수치를 존중하며, 원자료를 재산출하지 않고 통화 단위·가격 수준·기준연도만을 표준화하여 국제 기준에 부합하도록 한다. 이 과정은 매뉴얼에 명시되어 있어 「제출-조정-공시」 절차의 반복 가능성을 보장하며, 재현성을 확보한다. 또한, IMF는 자료 제출 시 메타데이터를 의무적으로 동반하도록 하고, 추정·보정 여부를 명기하도록 하여 메타데이터 우선 원칙을 제도화하였다. 더불어 국가마다 다른 개념과 분류를 SNA, GFSM, BPM6에 맞추어 정렬함으로써 상호운용성을 확보하고 있다.

품질 관리 기준 역시 IMF의 구조적 특성을 잘 보여준다. DQAF를 마련해 정확성, 일관성, 시의성, 접근성, 신뢰성을 종합적으로 평가한다. 이는 회원국 통계의 품질을 내부 점검이 아닌 공개 평가의 방식으로 관리한다는 점에서 독창적이다. 또한, 회귀모형과 시계열 추정을 통해 시계열 연속성을 유지한다. SDDS, e-GDDS는 시의성을 확보하는 제도적 장치이며, 추정값·보정값 여부를 명기하고 공개 평가 결과를 외부에 공유하는 절차는 투명성 확보로 이어진다.

마. 시사점

IMF의 데이터 통합에서 특히 주목할 부분은 **거울 자료 활용**이다. 거울 자료의 기본 원리는 특정 국가의 미보고 또는 불완전한 자료를 교역 상대국의 자료로 보완하는 것이다. 예를 들어, A 국가가 특정 품목의 수입 자료를 누락한 경우, 무역 상대인 B 국가의 수출 자료를 활용해 공백을 추정한다. IMF는 이러한 방식을 국제수지 및 무역 통계의 불일치를 해소하기 위한 수단으로 활용하며, 세관 행정 지침을 통해 제도화하기 시작했다.

거울 자료는 데이터 통합에 중요한 함의를 제공한다. 행정자료 간 통합에서 결측이 발생할 경우, 단순히 추정값으로 대체하는 대신 관련 기관이 보유한 상호자료를 활용함으로써 신뢰성을 높일 수 있다. 예컨대, 국세청 자료와 관세청 자료, 고용노동부 자료와 국민연금공단 자료의 상호보완적 검증은 거울 자료 활용 원리를 국내에 적용한 사례가 될 수 있다. 나아가, 거울 자료 활용은 단일 기관 중심의 통합을 넘어, 기관 간 상호검증과 자료의 교차 활용을 통한 다층적 통합으로 확장될 수 있으며, 이는 국가통계 전반의 투명성과 신뢰성을 강화하는 방향으로 이어질 것이다.

<표 4-1> 국제기구별 데이터 통합 사례 비교

	UN	OECD	Eurostat	IMF
자료 수집	<ul style="list-style-type: none"> • 국가별 통계, 국제기구 자료 • SDMX 기반 수집 	<ul style="list-style-type: none"> • 회원국 자료, 국제기구 지표, 민간 데이터 	<ul style="list-style-type: none"> • EU 규정에 따른 의무 제출 	<ul style="list-style-type: none"> • 회원국 제출 통계 • 국제기구·시장 자료
전처리	<ul style="list-style-type: none"> • 변수명, 단위, 코드 체계 표준화 • 변수 정의 표준화 • 정합성 점검 	<ul style="list-style-type: none"> • SDMX 표준 적용¹¹⁷⁾ 	<ul style="list-style-type: none"> • SDMX 표준 적용 • 변수 개념 통일 • 시계열, 극단값 등 점검 	<ul style="list-style-type: none"> • SDMX 표준 적용 • 통화 단위, 가격 수준 등 통일
결합	<ul style="list-style-type: none"> • 통계적 조화 	<ul style="list-style-type: none"> • 통계적 조화 	<ul style="list-style-type: none"> • 통계적 조화 • 제한적 결정적 연계 	<ul style="list-style-type: none"> • 통계적 조화
결측값 대체	<ul style="list-style-type: none"> • 모형 기반 대체 • 평균값 활용 	<ul style="list-style-type: none"> • 보간·외삽 • 회귀모형 기반 추정 • 평균값 활용 	<ul style="list-style-type: none"> • 모형 기반 추정 • 시계열 보간 	<ul style="list-style-type: none"> • 모형 기반 추정 • 거울 자료 활용
보정	<ul style="list-style-type: none"> • 해당 국가와 협의하여 조정 • 연령표준화, PPP 	<ul style="list-style-type: none"> • 표준 분모(WPP·PPP) 적용 • 국가별 가중값 	<ul style="list-style-type: none"> • 가중값·PPP 적용 • 시계열 단절 시, 과거 자료 재작성 	<ul style="list-style-type: none"> • PPP, 규모 표준화 등 ⇒ 개념·범위 일치에 가까움
품질 점검	<ul style="list-style-type: none"> • 자체 품질 점검 없음 • 지표별 주관기구에서 품질 관리 	<ul style="list-style-type: none"> • QAF 적용 <ul style="list-style-type: none"> - 정확성, 일관성, 시의성 등 검증 - 메타데이터 문서화 	<ul style="list-style-type: none"> • ESCP 및 ESS QAF에 근거 <ul style="list-style-type: none"> - 정확성, 시의성, 일관성 등 검증 - 지표별 메타데이터 공개 	<ul style="list-style-type: none"> • DQAF 운용 <ul style="list-style-type: none"> - 정확성, 시의성, 일관성 등 검증 • SDDS 및 e-GDDS로 공표 평가
시사점	<ul style="list-style-type: none"> • 관측값·추정값 병기 <ul style="list-style-type: none"> ⇒ 투명성·책임성 강화 • SDG Metadata Repository <ul style="list-style-type: none"> ⇒ 통합 데이터 세트 신뢰성 제고 	<ul style="list-style-type: none"> • 메타데이터 기반 절차 공개 • 체계적 품질 점검 	<ul style="list-style-type: none"> • 법적·제도적 근거에 기반한 데이터 통합체계 	<ul style="list-style-type: none"> • 거울 자료 활용의 독창성 <ul style="list-style-type: none"> ⇒ 상호검증과 자료의 교차활용을 통한 다층적 통합으로 확장

117) SDMX 표준 적용은 데이터 항목 정의, 코드, 메타데이터의 구조 등이 SDMX가 정한 개념·표준 분류체계를 따른다는 의미이며, SDMX 기반 수집은 자료를 제출하거나 교환하는 방식 자체가 SDMX 체계(예: XML, JSON 등)를 사용한다는 의미이다.

제2절 국가별 사례

1. 미국, Longitudinal Employer-Household Dynamics(LEHD)

미국은 고용보험 제도를 통해 근로자 대부분의 임금과 고용 기록을 축적하고 있다. 그러나 분산형 연방 통계체계로 인해 이 자료는 각 주(州) 정부가 관리하며, 연방 차원의 일관된 통합은 이루어지지 못하였다. 이로 인해 개인의 노동시장 이력이나 기업 고용 변화를 통합적으로 파악하기 어려워 노동시장 정책 효과 평가나, 지역 간 비교 분석에 데 한계가 있었다(Abowd et al., 2009). 이를 보완하기 위해 US Census는 행정자료와 조사자료를 연계해 노동시장 심층 분석을 가능케 하는 LEHD 프로그램을 출범하였다(Abowd et al., 2009). LEHD는 주(州) 고용보험 기록, 기업 행정자료, 인구조사 및 표본조사를 연계하여, 개인·기업·지역 단위 고용 동향 분석을 가능하게 하는 인프라를 구축한 사례이다.

가. 자료 수집 및 전처리

LEHD 자료는 세 가지 축으로 구성된다. 첫째, 주 정부가 관리하는 고용보험 기록으로 고용주·피고용인 관계, 임금 수준, 근무 이력 등이 포함된다. 둘째, 산업·지역별 고용 현황을 제공하는 QCEW(Quarterly Census of Employment and Wage)가 활용된다. 셋째, US Census 조사자료¹¹⁸⁾를 통해 개인의 연령, 성별, 교육, 인종 등 인구 특성이 추가된다.

세 자료는 형식과 개념이 달라 단순 결합이 어렵다. 따라서, 변수 정의를 통일하고, 표준화된 스키마를 마련해 주별 자료를 정렬한다. 동일 개인을 추적하기 위해 사회보장번호(SSN)를 기반으로 표준화를 수행하여(Wagner & Layne, 2014), 민감정보는 비식별 처리 후 PIK(Person Identification Key)로 변환된다(Vilhuber et al., 2013).

나. 결합 및 결측값 대체

LEHD는 다양한 데이터를 개인·기업 단위로 연계한다. SSN을 이용한 결정적 연계를 주축으로 하되, 불일치나 누락 사례에서는 확률적 연계를 병행한다. 이를 위해 PVS(Person Identification Validation System)를 운영하는데, PVS는 SSA(Social Security Administration)와 IRS(Internal Revenue Service)의 행정자료를 참조해 개인 단위 키를 부여하고, 이를 통해 노동 관련 조사자료와 고용보험 기록을 결합한다(Vilhuber et al.,

118) Current Population Survey(CPS), American Community Survey(ACS) 등

2013). 불확실한 사례는 전문가 검토(clerical review)를 통해 최종 결함 여부를 결정한다(Wagner & Layne, 2014). 다만, 고용보험 자료는 임금, 근무 시간 등의 변수가 제한적이다. 따라서, 누락된 정보는 조사자료(예: ACS, CPS)를 활용하여 보완한 후, 핫-덱 대체로 결측을 처리한다. 예를 들어, 동일 성별·연령·지역 조건을 가진 근로자의 관측값을 대체값으로 활용한다. 또한, 시계열 연속성을 위해 보간과 외삽을 적용하기도 한다(U.S. Census, 2024).

다. 보정 및 품질 점검

행정자료는 등록된 고용만을 포함하기 때문에 자영업이나 비공식 경제활동은 과소 포착되는 한계가 있다. 이를 보완하기 위해 LEHD는 행정자료와 조사자료를 교차 검증하여 보정을 수행한다(U.S. Census, 2024). 예컨대, ACS·CPS와 같은 조사자료를 연계해 대표성을 확보하고, 가중값을 부여하여 보정한다(Abowd & Stinson, 2013). 이 과정에서 행정자료와 조사자료 간 차이를 조정함으로써 결과의 대표성을 강화한다(Abowd & Stinson, 2013; U.S. Census, 2024). LEHD는 품질 관리를 위해 연계율, 오연계율, 대표성 등을 산출한다(Wagner & Layne, 2014). 또한, 원자료, 결측 대체 자료, 보정자료 간 차이를 식별할 수 있도록 메타데이터를 제공한다.

라. 전처리 원칙 및 품질 점검 기준에 대한 고찰

LEHD의 전처리 과정은 개인정보 보호와 통합 가능성을 동시에 달성하기 위한 일련의 절충으로 이루어진다. 개인 식별에는 SSN을 활용하되, 직접 노출을 피하고자 PIK라는 대체 식별자를 부여한다. 이는 분석의 연속성을 보장하면서도 민감정보 노출 위험을 최소화하여 정합성과 윤리적·법적 요건 준수를 함께 충족한다. 또한, 주별로 다른 자료를 통합하기 위해 변수 정의와 코드 체계를 표준화하고, PVS를 통해 연계 절차를 일관되게 관리한다. 이러한 구조는 재현성을 보장하며, 메타데이터에는 원자료와 보정·결측 대체 과정이 명확히 기록되어 있어 메타데이터 우선 원칙이 구현된다. 나아가 주별 행정자료와 조사자료를 공통 스키마에 맞춰 정렬한 점은 상호운용성 확보로 이어진다.

LEHD는 자료의 신뢰성을 높이기 위해 품질 지표를 정량적으로 산출·공개한다. 연계율, 오연계율, 대표성 지표를 제시함으로써 정확성과 일관성을 확보한다. 그러나, 행정자료 단독으로는 특정 집단(예: 자영업자, 이민자, 비공식 부문)을 포괄하기 어렵다는 한계가 있어, 이를 보완하기 위해 ACS·CPS와 같은 조사자료와 교차 검증을 수행하며, 이 절차는 행정자료의 한계를 보완하고 완전성을 강화하는 장치로 기능한다. 또한, 연계·보정 절차와 품질 지표를 메타데이터와 함께 공개함으로써 투명성이 제도적으로 확보된다.

마. 시사점

LEHD는 다양한 지표와 데이터를 산출한다. 예를 들어, QWI(Quarterly Workforce Indicators)는 지역별 고용·임금, 성별·연령·산업별 노동시장 동향을 제공하며, J2J(Job-to-Job Flows)는 직장이동과 임금 변화를 추적해 노동시장 유연성 분석에 활용된다. 이러한 지표들은 노동시장의 수요와 공급을 동시에 포괄하며, 노동시장 구조를 입체적으로 조망할 수 있게 한다. 이처럼, LEHD는 지역 고용정책 수립이나 임금 격차 해소 정책에 이바지하며, 데이터 기반 정책설계의 인프라로 기능한다.

LEHD의 가장 큰 특징은 주 정부의 행정자료와 연방 조사자료의 결합을 통해 국가적 노동시장 분석 체계를 구축한 점이다. 추후 다루게 될 뉴질랜드의 IDI가 전(全)영역을 포괄하는 통합 기반이라면, LEHD는 노동시장이라는 특정 정책 영역에 집중함으로써 높은 정밀성과 시의성을 확보한 통계로 차별화된다. 또한, LEHD의 공개 통계 생산물(QWI, J2J)은 데이터 통합의 성과를 사회 전체에 환류시키는 사례이다. 국내 역시 고용보험, 국민연금, 세무 자료, 인구총조사 등 다양한 자료를 결합해 노동시장 동향을 분석할 수 있는 **범정부 차원의 통합 인프라**를 구축할 필요가 있다.

2. 캐나다, T1 Family File(T1FF)

T1FF는 개인 소득세 신고자료, 출생·사망 등록, 인구총조사 및 가구 조사자료 등을 연계해 구축된 통합 데이터베이스로 가족 단위의 소득, 인구 이동, 사회경제적 변화를 장기적으로 추적할 수 있다(Statistics Canada, 2017). T1FF는 기존 표본조사 중심의 통계생산 체계가 안고 있던 비용 및 응답 부담 문제를 완화하면서도, 국가 전체를 포괄하는 전수 기반 데이터를 생산한다(Milligan, 2016).

가. 자료 수집 및 전처리

T1FF의 핵심 자료는 국세청(CRA)의 T1 개인 소득세 신고자료로, 거의 모든 캐나다 거주자가 매년 보고하기 때문에 포괄성이 매우 높다. 여기에 출생·사망 등록자료가 연계되어 가구 구성과 인구 동향을 보완하며, 인구총조사와 가구 조사자료를 결합해 교육, 인종, 이민 등 사회인구학적 변수를 추가한다(Statistics Canada, 2017).

전처리 과정은 세 단계로 이루어진다. 첫째, 세금자료와 조사자료의 변수 정의를 일치시킨다. 둘째, 주(州)마다 다른 세금 제도나 보고 체계를 표준화한다. 셋째, 사회보험번호(SIN)를 식별자로 활용해 중복을 제거하고, 오류·누락을 검증함으로써 정확성을 확보한다(Statistics Canada, 2017).

나. 결합 및 결측값 대체

데이터 결합은 주로 정확 연계를 수행한다.¹¹⁹⁾ SIN은 유일한 개인 식별번호로 세금자료, 출생·사망 등록, 인구조사 자료를 정확하게 결합할 수 있는 기반을 제공한다(Statistics Canada, 2017). 다만, 신규 이민자나 특정·소수집단의 경우 SIN 누락이나 오류가 발생하기도 한다. 이러한 경우, 이름, 성별, 생년월일 등 공통 변수를 활용한 결정적 연계가 보완적으로 수행된다(Statistics Canada, 2017).

TIFF는 모집단 대부분을 포괄하지만, 소득신고 누락, 비과세 소득 미포착, 신규 이민자의 신고 공백으로 인해 결측이 발생한다. 이러한 결측은 모의 세무 기록이나 다른 행정자료를 참조하거나, 표본조사와의 결합을 통해 보완한다(Statistics Canada, 2017). 예를 들어, 신규 이민자의 초기 소득 공백은 동일 지역·가구 구조·연령 조건을 가진 집단의 소득 분포를 참조하여 보완한다.

다. 보정 및 품질 점검

TIFF에서는 세금 신고 의무가 없는 집단(예: 최빈층, 일부 고령자)과 원주민 집단에서 대표성 부족 문제가 발생한다. 이를 보완하기 위해 인구총조사 기반의 보정 가중값을 적용한다.¹²⁰⁾ 또한, 비공식 경제활동의 누락 문제를 완화하기 위해, 가계조사 자료와의 교차 비교를 수행하며, 세금자료 기반 지표가 실제 소득 분포와 어느 정도 일치하는지 검증한다(Statistics Canada, 2017). 아울러, Statistics Canada(2017)는 TIFF의 품질 관리 절차를 체계적으로 문서화하고 있다. 연계율, 인구 대비별 포착률, 변수별 정합성, 시계열 일관성 등을 정기적으로 점검하며, 메타데이터를 공개해 투명성을 확보한다.

라. 전처리 원칙 및 품질 점검 기준에 대한 고찰

TIFF는 세금자료를 기반으로 하지만, 다른 행정자료 및 조사자료와의 결합을 위해 변수 정의와 코드 체계를 표준화하였다. 주별로 제도가 다른 세금 항목도 연방 기준에 맞춰 정렬되었으며, SIN이 고유식별자로 활용되었다. SIN 기반 연계는 누락률이 낮아 정합성 확보에 기여했으나, 신규 이민자 등 일부 집단에서는 SIN 부재 문제가 발생해 대표성이 제한되었다. 원자료는 가능한 한 보존하되, 오류 검증·중복 제거·변수 표준화 등 최소한의 가공만 수행하여 최소침습성과 상호운용성을 동시에 달성하였다. 개인정보는 비식별화 과정을 거쳐 활용됨으로써, 윤리적·법적 요건을 충족하였다.

119) 배우자, 자녀 등 가족 구성원의 정보는 신고서 간 교차 검증을 통해 연계한다. 단순한 개인 단위 연계를 넘어, 가족 단위 통합 데이터를 구성하는 특징이 있다(Statistics Canada, 2019).

120) Corak(2013)은 이러한 보정이 소득 분포의 대표성을 유지하는 데 필수적이라고 지적하였다.

또한, 세금자료, 인구총조사, 등록자료를 공통 구조로 정렬한 것은 상호운용성 강화를 뒷받침한다.

TIFF는 일부 집단의 대표성 문제를 보완하기 위해 인구총조사 기반 보정 가중값을 적용해 완전성을 강화하였다. 품질 점검은 연계율, 인구 대비 포착률, 변수별 정합성, 시계열 일관성 등을 정기적으로 검토하고, 장기 시계열 자료를 통해 일관성을 확보하였다. 세금자료의 주기성과 행정자료 특성 덕분에 시의성 있는 갱신이 가능하며, 지표와 보정 절차가 함께 공개되어 투명성도 충족되었다.

마. 시사점

TIFF는 세금자료와 인구자료를 결합하여 **가족 단위 데이터 통합**이라는 독창성을 확보하고, 거의 **전 국민을 포괄하는 장기 데이터베이스**를 구축하였다. 이를 통해 가계소득 변화를 단일 시점이 아닌 장기 시계열로 추적할 수 있게 되었으며, 이를 바탕으로 부모·자녀 세대 간 소득 연결 분석이나 조세·이전 제도의 효과 평가 연구가 수행되었다. 그러나, 세금자료는 과세 대상 소득에 국한되기 때문에 비과세 소득이나 비공식 경제활동을 포착하지 못하는 한계가 있다. 이에 따라, 저소득층의 생활 수준이 과소 추정될 수 있다. 또한, 신규 이민자나 원주민은 세금 신고 이력이 불완전하거나 누락이 많아 대표성 문제가 발생한다.

3. 호주, Multi-Agency Data Integration Project(MADIP)

호주는 2010년대 중반부터 데이터 활용을 통한 정책 개선을 목표로 **범정부 차원의 데이터 공유 체계를 강화**하였다.¹²¹⁾ 이러한 노력의 대표적 성과가 MADIP이다. MADIP는 다양한 행정자료를 통합하여 개인과 가구의 전 생애주기를 아우르는 분석을 가능하게 하는 국가적 데이터 인프라로 구축되었다(ABS, 2018).

가. 자료 수집 및 전처리

MADIP에는 인구총조사, 세금자료, 의료보험, 의약품 급여제도, 사회보장·복지자료, 교육자료, 사망 등록 등이 포함된다(ABS, 2018, n.d.-a). 이들 자료는 ABS 표준 분류체계에 따라 변수 정의를 표준화하고, 시계열 일관성을 점검한다. 개인정보 보호를

121) 생산성 위원회(Australian Government Productivity Commission)가 2017년 발표한 ‘Data Availability and Use’ 보고서를 통해 데이터 공유와 통합의 필요성이 정책적으로 뒷받침되었고, 이후 범정부 차원의 DIPA(Data Integration Partnership for Australia)가 구축되었다(PC, 2017).

위해 모든 식별자는 가명화(pseudonymisation) 처리되며, 이후 PLS(Person Linkage Spine)에 연결된다. 또한, 중복과 불일치는 정합성 점검을 통해 최소화한다(ABS, 2018, n.d.-a).

나. 결합 및 결측값 대체

PLS는 인구총조사와 주요 행정자료를 결합해 생성된 준거 틀/framework)로 MADIP 내 모든 데이터는 PLS를 통해 결합한다. 결합 방법은 결정적 연계와 확률적 연계가 혼합 적용된다. 결정적 연계는 고유식별자(세무 번호, 건강보험 번호)를 활용하며, 확률적 연계는 이름·생년월일·주소 등을 활용해 유사도를 계산하고 임계값에 따라 연계 여부를 판정한다(ABS, 2018, n.d.-a). 행정자료는 포괄성이 높으나, 보건·교육 영역에서는 일부 변수의 결측이 발생한다. 이러한 결측은 인구총조사 등 조사자료를 활용해 보완되며, 모든 절차는 메타데이터에 문서화한다(ABS, 2018, n.d.-a).

다. 보정 및 품질 점검

MADIP는 모집단 대부분을 포괄하지만, 일부 집단(예: 원주민, 신규 이민자 등)에서 대표성 부족 문제가 발생할 수 있다. 이를 보완하기 위해 인구총조사 기반 보정 가중값을 적용하여 행정자료와 조사자료 간 불일치를 조정한다. 이를 통해 모집단 대표성을 강화하고, 분석 결과의 왜곡 가능성을 줄인다(ABS, 2018, n.d.-a, n.d.-b). 품질 보증을 위해 다층적 점검 절차를 운영하며, 주요 지표는 연계율, 오연계율, 대표성, 시간적 정합성(consistency over time)이다(ABS, 2018, n.d.-a). 이러한 지표는 메타데이터와 함께 제공되어 MADIP 활용의 투명성을 높인다.

라. 전처리 원칙 및 품질 점검 기준에 대한 고찰

MADIP는 출처별 자료를 ABS 표준 분류체계에 맞추어 변수 정의와 코드 체계를 통일하고, PLS를 활용하여 중복과 불일치를 줄였다. 이는 정합성과 상호운용성 확보를 제도적으로 내재화한 사례라 할 수 있다. 개인정보는 수집 단계에서 가명화 처리된 후 PLS와 연결되며, 이는 윤리적·법적 요건 준수를 충족하는 동시에 원자료를 가능한 한 보존하는 방식이므로 최소침습성 원칙에도 부합한다. 연계 과정은 일관된 규칙과 절차에 따라 문서화가 되어 있어 재현성이 보장된다. 또한, MADIP는 연계 품질 지표와 메타데이터를 함께 제공하여 데이터의 범위와 한계를 파악할 수 있도록 하고 있으며, 이는 메타데이터 우선 원칙을 충실히 반영한다.

MADIP는 품질 지표를 정량적으로 산출한다. 이를 통해 데이터의 정확성과 일관성을 확보하며, 품질 지표가 메타데이터와 함께 공개되므로 투명성도 충족된다. 다만, 특정 집단의 포착률이 낮다는 한계가 존재한다. 이를 보완하기 위해 인구총조사

기본 보정 가중값을 적용하고 있으며, 이러한 조치는 대표성 확보와 완전성 강화에 기여한다.

마. 시사점

MADIP는 국민의 전 생애주기를 포괄하는 데이터 통합을 통해 정책 효과를 정밀하게 평가할 수 있는 기반을 마련하였다. 그러나, 원주민, 저소득층 등 일부 집단의 과소 대표 문제와 행정기관 간 표준화 미비로 인한 변수 불일치 문제는 여전히 한계로 지적된다. 그러나, **무엇보다 주목할 점은 범정부 차원의 협력 구조가 제도적으로 뒷받침되었다는 사실이다.** 호주 정부는 부처 간 데이터 공유와 협력을 제도화하고, 통계청(ABS)을 중추 기관으로 지정하여 「조정-품질 관리-보안 체계」를 총괄하도록 하였다. 이를 통해, 보건, 교육, 복지, 세금 등 다양한 영역의 행정자료를 통합할 수 있었으며, 정책 집행에 있어 범부처적 시각을 확보할 수 있게 되었다. 또 하나의 핵심 시사점은 PLS 체계이다. 모든 데이터 결합을 PLS 중심으로 수행함으로써 동일인이 여러 자료에 중복해서 등장하는 문제를 최소화하고, 결합의 효율성과 정확성을 동시에 제고하였다.

MADIP는 데이터 연계·통합·활용 확대를 추진 중인 국내 정책 환경에도 중요한 시사점을 제공한다. 우선, **국가데이터처를 중심으로 한 범정부 협력 구조의 제도화를 통해 행정자료의 분산적·분절적 관리체계를 극복할 필요가 있다.** 동시에, PLS 같은 데이터 연계·통합체계를 구축해 주민등록, 세금, 사회보장 기록 등을 하나의 인프라에 정렬하고, 이를 기반으로 행정자료와 조사자료뿐만 아니라 **민간의 빅데이터까지 결합할 수 있는 체계를 마련해야 한다.**

4. 뉴질랜드, Integrated Data Infrastructure(IDI)

뉴질랜드는 다문화 사회로서 원주민·이민자의 사회적 포용, 고령화, 지역 불평등 등 복합적인 과제에 직면해 왔다. 특히, 교육 성과, 노동시장 참여, 건강 불평등, 범죄율, 복지 수급 등 다양한 분야에서 다부처 협력 기반의 대응이 필요했다. 이러한 문제의식 속에서 뉴질랜드는 2011년 IDI를 구축하였다. IDI는 행정자료와 조사자료를 연계하여 개인·가구·기업 단위의 통합 분석이 가능한 데이터 인프라이다(Statistics New Zealand, n.d.-a).

가. 자료 수집 및 전처리

IDI는 세 계층으로 구성된다(Statistics New Zealand, n.d.-a). 첫째, PS(Population

Spine)는 출생·사망 등록, 이민·비자 기록, 고용·소득 기록을 기반으로 한 준거 틀로 IDI의 중심축 역할을 한다. 둘째, 연계 데이터 세트(linked data set)는 교육부, 보건부, 사회개발부, 사법부 등 다양한 부처에서 수집된 수백 개의 행정자료가 PS에 연계된 것이다. 셋째, 조사자료에는 인구총조사, 노동력조사, 가계조사 등이 포함되며, 이는 행정자료의 한계를 보완한다.

전처리 과정에서는 출처별로 상이한 변수 정의, 수집 주기, 단위를 표준화한다. 특히, 동일 속성이 기관마다 다를 경우 신뢰성과 시의성을 평가하여 최적 추정값(best estimates)을 제공한다. 예를 들어, 성별, 출생 연월 등의 기본 속성이 기관별로 다를 경우 이를 종합해 단일화된 기준값을 생성한다(Statistics New Zealand, n.d.-b). 이러한 절차는 데이터 간 비교 가능성과 일관성을 보장한다. 또한, 개인정보는 가명화 처리된 후, PS를 통해 결합된다(Statistics New Zealand, n.d.-a, n.d.-b, n.d.-c).

나. 결합

뉴질랜드는 모든 국민을 포괄하는 단일 식별번호가 존재하지 않기 때문에 IDI에서는 결정적 연계와 확률적 연계를 병행한다(Black, 2016; Gibb, 2016). 세무 번호, 출생·사망 등록번호와 같은 고유식별자가 있으면 결정적 연계를 적용한다. 반면, 고유식별자가 없거나 불완전하면 이름, 생년월일, 주소 등 공통 변수의 유사도 점수를 계산하여 확률적 연계를 수행한다. 이 과정에서 PS는 결합의 준거 틀 기능을 한다. PS는 동일 개인이 여러 자료에 중복되는 문제를 최소화하고, 결합 품질을 높이는 핵심 장치이다. Black(2016)은 PS가 다양한 행정자료를 연결할 수 있도록 설계되었으며, 특히 단일 식별번호가 없는 국가에서 연계의 기반 역할을 한다고 평가하였다.

다. 결측값 대체 및 품질 점검

IDI는 원칙적으로 각 행정기관이 제출한 원자료를 가능한 한 원형 그대로 보존하여 연계하며, 결측값은 그대로 유지된다(Statistics New Zealand, n.d.-a, n.d.-b). 다만, 일부 행정자료는 원자료 단계에서 해당 기관의 행정 처리 규칙에 따라 결측 대체가 이루어졌을 수 있다. 예컨대, 세무 자료 부분 누락은 원천기관 수준에서 대체될 수 있다. 이는 IDI 내부 절차가 아니라 행정기관 차원의 처리에 해당한다. IDI는 자체적으로 대체를 수행하지 않는 대신, 분석 단계에서 활용이 가능한 메타데이터와 품질 지표¹²²⁾를 제공한다. 이를 통해 이용자는 결측 처리의 필요성을 판단하고, 목적에

122) IDI의 품질 점검은 내부 오류를 정정하는 품질 관리라기보다, 결합 결과를 측정하고 투명하게 공개하는 절차에 가깝다. IDI는 이용자에게 데이터 활용 시 반드시 품질 한계를 고려하라고 안내하며, 이를 위해 메타데이터와 품질 보고서를 함께 제공한다(Statistics New Zealand, n.d.-a, n.d.-b).

적합한 기법을 선택할 수 있다(Black, 2016; Gibb, 2016).

라. 보정

IDI는 행정자료 기반이기 때문에 특정 집단에서 포괄성 문제가 발생한다. 예를 들어, 원주민(Māori, Pacific peoples), 저소득층, 행정자료에 반영되지 않는 비공식 경제활동 집단은 충분히 포착되지 않을 수 있다(Black, 2016; Gibb, 2016). 그러나, IDI는 대표성 확보를 위해 가중값 보정을 적용하지 않는다. 대신, 외부 조사자료와 IDI를 비교하여 포괄범위 차이를 Data Quality Statements에 공표한다(Statistics New Zealand, n.d.-a, n.d.-b). 해당 문서에는 특정 변수와 집단에서 발생하는 과소 포착 문제, 자료별 한계가 구체적으로 명시된다.

마. 전처리 원칙 및 품질 점검 기준에 대한 고찰

IDI의 전처리 과정은 최소침습성을 엄격히 준수한다. 자료 결합 과정에서 결측값 대체나 보정은 수행되지 않으며, 원자료는 가능한 한 그대로 유지된다. 연계 절차는 PS를 통해 일관되게 관리되며, 절차와 품질 지표는 재현성 확보를 위해 문서화되어 공개된다. 또한, 각 자료의 범위, 누락 가능성, 품질 한계 등은 Data Quality Statements를 통해 제공되며, 이는 메타데이터 우선 원칙을 충실히 반영한다. IDI는 내부 보정을 수행하지 않기 때문에 완전성에는 한계가 있으나, 이러한 한계를 명시적으로 보고함으로써 오히려 투명성을 강화하였다. 행정자료는 PS를 통해 정기적으로 갱신되어 시의성도 유지된다. 요컨대, IDI는 품질을 직접 점검하기보다, 품질 정보의 공개와 투명한 문서화에 중점을 두는 체계라 할 수 있다.

바. 시사점

IDI는 데이터 통합 방식에서 다른 국가 및 국제기구 사례와 뚜렷하게 구별되는 특징을 보여주며, 데이터 인프라 설계, 제도적 거버넌스, 정책적 활용성 측면에서 중요한 시사점을 제공한다.

첫째, **모든 국민을 포괄하는 단일 식별자 부재를 극복했다.** 호주의 MADIP는 세무·건강보험 번호와 같은 안정적 고유식별자를 축으로 PLS를 구축하였으나, 뉴질랜드는 전 국민 고유번호가 존재하지 않는 상황에서 여러 출처를 결합해 PS를 설계하였다. 이는 주민등록번호나 사회보장번호 같은 단일 식별자를 활용하기 어려운 국가에서도 고품질 데이터 통합 인프라를 설계할 수 있음을 증명한 것이다.

둘째, **최적 추정값을 통한 개념 조화**이다. 대부분 국가는 자료 간 변수 표준화를 시도한다. 하지만, 뉴질랜드처럼 상충하는 정보를 종합해 단일 기준값을 공식 제공하는

체계를 마련한 사례는 드물다. 이는 행정자료와 조사자료 간 괴리를 줄여 데이터 비교 가능성을 제도적으로 강화한 것이다.

셋째, **제도적 거버넌스의 독창성**이다. IDI는 Five Safes 원칙¹²³⁾을 적용해 데이터의 접근과 활용을 통제할 뿐만 아니라, 마오리(Māori) 데이터 윤리 준거인 Ngā Tikanga Paihere를 반영하였다. 이는 데이터가 단순한 개인정보 집합이 아니라, 공동체 권리라는 맥락 속에서 존중받아야 한다는 사회적 합의를 제도화한 사례이다. 원주민 등 취약 집단의 데이터 주권을 국가 인프라 운영 원리에 포함한 점에서 뉴질랜드의 독창성이 드러난다.

넷째, **품질 지표의 체계적 공개와 투명성 확보**이다. Data Quality Statements를 발간하여 데이터의 범위·누락 가능성·품질 한계를 명시한다. 이를 통해 이용자가 데이터 품질을 직접 검토할 수 있도록 하며, “자료를 제공한다.”를 넘어 “자료를 신뢰할 수 있다.”는 점을 제도적으로 입증한다. Eurostat 역시 품질 보고서를 운영하지만, IDI처럼 데이터 통합 환경에서 상세 품질 지표를 체계적으로 공개하는 사례는 드물다.

IDI는 데이터 연계·통합·활용 확대를 추진 중인 국내에도 직접적인 합의를 제공한다. 한국은 주민등록, 세무, 복지자료 등 풍부한 행정자료를 보유하고 있으나, 기관별 분산 관리로 인한 통합 활용의 한계가 존재한다. 이에 뉴질랜드처럼 PS 기반 체계를 마련하고, ‘최적 추정값’ 개념을 제도화하며, 품질 지표 공개를 통한 데이터 신뢰성 확보가 필요하다. 아울러 Five Safes와 같은 원칙을 제도화하여 취약 집단 데이터 거버넌스를 병행할 장치를 고려할 필요가 있다.

5. 대한민국 국가데이터처, 행정자료 연계

국가데이터처는 대한민국 통계생산의 중심기관으로서 조사 환경 악화, 응답 부담 증가, 조사비용 상승이라는 복합적 도전에 대응하기 위해 행정자료를 적극적으로 활용하는 데이터 통합 전략을 추진해 왔다. 이는 단순한 비용 효율화 차원을 넘어 정책 수요에 부합하는 정밀 통계생산을 가능하게 하는 핵심 수단으로서, 국제 통계의 발전 방향과도 일치한다.

「통계법」 제24조는 행정자료의 제공과 활용에 대한 법적 근거를 명확히 규정함으로써, 국가데이터처가 중앙·지방정부 및 공공기관의 행정자료를 수집하고 이를 국가통계 생산에 활용할 수 있는 제도적 틀을 마련하였다. 이러한 법적 기반을

123) Safe people, Safe projects, Safe settings, Safe data, Safe outputs(Statistics New Zealand, n.d.-c)

통해 과세자료, 주민등록자료, 외국인등록자료, 국민연금·건강보험, 교육행정정보시스템 등 다양한 행정자료의 안정적 확보가 가능해졌으며, 그 결과 ‘2015년 인구주택총조사’의 등록 기반 전환이라는 혁신적 성과를 달성할 수 있었다.

가. 자료 수집 및 전처리

행정자료 입수 절차는 「계획 수립-업무협의-공식 입수요청-자료 입수」의 단계로 진행된다. 2024년 기준 118개 기관으로부터 319종의 행정자료가 입수되었으며, 작성 통계 66종 중 50종의 통계가 행정통계작성(17종), 항목 대체(21종), 자료보완 및 검증(12종) 등에 활용하고 있다. 이는 행정자료가 단순한 보조 수단을 넘어 통계생산의 핵심 인프라로 기능하고 있음을 시사한다.

입수된 자료는 명칭 표준화, 관리 코드 도입 등을 통해 일관성과 추적성이 확보된다. 개인정보 보호는 최우선 원칙으로, 주민등록번호 등 고유식별자는 암호화 알고리즘에 따라 비식별 처리된 후 관리되고 있다. 이는 개인정보 보호법과 통계법의 규정을 충족하면서도, 데이터 통합의 정확성을 확보하기 위한 필수 조치이다.

행정자료는 원자료 그대로 통계작성에 활용하기 어려운 경우가 많다. 변수 간 불일치, 결측, 이상값, 중복 등이 빈번하게 발생하기 때문이다. 이에 전처리는 세 단계로 이루어진다. 첫째, 자료 진단을 통해 불일치율과 중복률을 확인하고, 변수별 결측과 이상값을 점검한다. 둘째, 자료 진단 후에는 변수명, 코드, 단위, 시점 등의 구조를 표준화한다. 셋째, 자료 간 정의와 분류체계 등의 불일치는 정합성 점검 과정에서 보완한다. 예를 들어, 성별이 남·여 이외의 값이거나 개업일이 폐업일보다 늦은 날짜로 입력된 사례 등은 논리적 불일치로 간주하여 정합성 점검의 주요 대상이 된다.

나. 결합, 결측값 대체 및 보정

행정자료 연계는 정확 연계와 결정적 연계를 통해 수행되며, 다양한 식별정보(통계목적고유번호, 사업자등록번호, 법인등록사업체명, 주소 등)를 기반으로 자료 간 정확한 일치성을 확보한다. 특히, 통계목적고유번호는 개인 식별이 불가능한 숫자로, 개인정보 보호와 통계 데이터 결합의 효율성을 동시에 확보하는 핵심 수단이자 연계키로 활용되는 안전한 관리 번호이다.

행정자료 연계는 원칙적으로 결측값 대체를 최소화하고, 원자료를 가능한 한 그대로 제공하는 것을 지향한다. 이는 원자료의 신뢰성을 보존하고, 이용자가 자료 특성을 명확히 이해한 상태에서 자체적으로 적합한 방법을 선택하도록 하기 위함이다. 다만, 일부 부서에서는 분석 목적과 통계작성 필요에 따라 제한적으로 결측값 대체

기법을 적용하기도 한다. 보정은 행정자료 자체에는 적용되지 않으나, 일부 표본조사에서는 대표성 확보를 위해 가중값 조정을 활용한다.

다. 품질 점검

품질 점검은 「기초 점검-유효성 점검-논리 점검-시계열 점검」 단계로 운영된다.

- **기초 점검.** 행정 DB의 구조적 일관성과 자료 누락 여부를 확인하는 단계이다. 열(column) 개수, 자료형(type), 길이(length) 등의 자료 구조를 점검하고, DB 간 구조 일치 및 정상 이관 여부를 확인한다.
- **유효성 점검.** 입력된 값의 정확성과 타당성을 검토하는 단계이다. 사업자등록번호, 주소 등 주요항목의 표준화를 통해 행정구역분류코드를 부여하고, 기본 항목값의 누락, 중복, 결측, 날짜 형식 오류, 이상값 등을 점검한다. 또한, 중복 제거 및 이상값 처리를 수행하며, 행정 DB 간 상호 보완을 통해 항목별 일관성을 강화한다.
- **논리 점검.** 변수 간 관계의 논리적 정합성을 검토하는 단계이다. 주요 항목별 값들 사이의 논리적 일관성, 조건 간 충돌 여부 등을 확인하고, 불합리한 조합이나 오류를 탐지한다. 예를 들어, 나이 음수값, 생년월일 5~6번째 자리가 월(01~12)을 나타내지는 않는 비정상 형식 등을 검토하여 이상값을 확인하고 보완한다.
- **시계열 점검.** 주요항목의 과거 연도 자료와의 연속성 및 항목 구성비를 비교하여 시계열 일관성을 점검하는 단계이다. 이를 통해 항목값이 정상적으로 적재되었는지, 연도 간 비정상적인 변동이나 분포의 왜곡이 존재하는지를 확인한다.

품질 점검은 단순 오류 탐지를 넘어 행정자료가 국가통계로 활용될 수 있는 수준의 신뢰성을 확보하는 역할을 한다. 이는 행정자료를 **국가 자산(national asset)**으로 관리하려는 국가데이터처의 의지를 반영한 것이다.

라. 전처리 원칙 및 품질 점검 기준에 대한 고찰

전처리 과정은 다단계 절차로 이루어진다. 변수명, 코드, 단위, 시점은 표준화되고, 자료 간 충돌을 줄이기 위해 관리 코드가 부여된다. 주민등록번호는 암호화 알고리즘으로 비식별 처리 후, 관리되고 있다. 이는 원자료를 가능한 한 보존하면서 필요한 조정만 수행하는 방식으로 최소침습성 원칙을 충족한다. 또한, 변수 정의와 코드 체계를 정비하는 표준화 절차 덕분에 재현성이 보장되며, 암호화와 비식별화 처리로 윤리적·법적 요건을 준수한다. 국가데이터처의 전처리 체계는 국제 기준과 비교해도 손색이 없을 만큼 제도적 강점이 두드러진다. 다만, 국가데이터처는 자료

연계 이후 다단계 품질 점검을 수행하고 있으나, 연계율·오연계율 등 정량적 품질 지표가 공식적으로 공개되지 않아 투명성 측면에서 개선 여지가 있다.

마. 시사점

국가데이터처는 행정자료 기반의 데이터 통합을 제도·운영·기술적 측면에서 체계화하는 성과를 달성했다. 첫째, 법적·제도적 기반 확립이다. 「통계법」 제24조를 근거로 중앙·지방정부 및 공공기관의 행정자료를 통계 목적으로 활용할 수 있는 제도가 마련되었으며, 이를 통해 과세자료, 주민등록자료, 외국인등록자료, 국민연금·건강보험, 교육행정정보시스템 등 다양한 행정자료의 안정적 확보가 가능해졌다. 둘째, 자료 입수·관리체계의 정례화이다. 입수계획 수립부터 품질 점검에 이르는 일련의 절차가 제도화되었다. 2024년 319종의 행정자료가 입수되었으며, 이 중 50종은 승인통계 생산에 직접 활용되고 있어, 행정자료가 조사통계의 보완을 넘어 통계생산의 핵심 인프라로 자리 잡고 있음을 보여준다. 셋째, 전처리 및 품질 점검 체계 확립이다. 자료 입수 직후, 기초·유효성·논리·시계열 점검, 결과 보고에 이르는 다단계 품질 점검을 수행하고 있다. 이를 통해 행정자료의 통합적 활용을 위한 기반이 마련되었다. 넷째, 데이터 보안·비식별화 절차가 강화되었다. 고유식별정보는 비식별화 처리 후, 개인을 식별할 수 없는 고유키로 관리되며 관리 코드 부여를 통해 자료의 추적성과 관리 효율성이 제고되었다.

국가데이터처의 행정자료 연계는 자료 수집 단계부터 결합까지 명확하게 정립되어 체계적으로 수행되고 있다. 다만, 몇 가지 한계도 존재한다. 첫째, 통합 데이터 세트의 활용 및 품질 점검 결과의 환류 체계가 아직 충분히 정립되지 않아 데이터 통합의 완결성을 저해한다. 데이터 통합이 단순한 결합에 그치지 않고 정책 수요에 부응하는 정보 생산으로 이어지기 위해서는 개선이 필요한 부분이다. 둘째, 분산형 통계체계 속에서 국가데이터처가 모든 자료를 직접 관리·통합할 수 없다는 구조적 제약이 있다. 이는 데이터 통합 범위와 심도를 제약하며, 기관 간 협력과 거버넌스 체계가 뒷받침되지 않으면 국가데이터처 차원의 통합은 필연적으로 한계에 부딪힐 수밖에 없다. 이는 통계생산 기관인 동시에 데이터 통합 기관으로서 국가데이터처의 위상을 강화하기 위해 개선해야 할 부분이다.

<표 4-2> 국가별 데이터 통합 사례 비교

	미국	캐나다	호주	뉴질랜드	대한민국
자료 수집	<ul style="list-style-type: none"> • 주별 고용보험 기록 • 산업·기업별 고용 현황 • 조사자료 	<ul style="list-style-type: none"> • 개인 소득세 신고자료 • 출생·사망 등록자료 • 인구총조사, 가구 조사 	<ul style="list-style-type: none"> • 인구총조사 • 행정자료 <ul style="list-style-type: none"> - 세금, 의료, 복지, 교육 등 	<ul style="list-style-type: none"> • PS(등록자료 기반) • 연계 데이터 세트 • 조사자료, 행정자료 	<ul style="list-style-type: none"> • 행정자료 319종 <ul style="list-style-type: none"> - 행정통계작성(17종) - 항목 대체(21종) - 자료 보완(12종) 등
전처리	<ul style="list-style-type: none"> • 변수 정의 통일 및 정렬 • 비식별 처리 	<ul style="list-style-type: none"> • 변수 정의 일치 • 주별 세금 제도 표준화 • 중복 제거, 오류·누락 검증 	<ul style="list-style-type: none"> • 변수 정의 표준화 • 시계열 일관성 점검 • 가명화 처리 	<ul style="list-style-type: none"> • 변수 정의, 수집 시기, 단위 표준화 • 최적 추정값 제공 • 가명화 처리 	<ul style="list-style-type: none"> • 변수 및 코드 표준화 • 정합성 점검 • 비식별 처리
결합	<ul style="list-style-type: none"> • 결정적 연계(SSN 기반) • 확률적 연계 	<ul style="list-style-type: none"> • 정확 연계(SIN 기반) • 결정적 연계 	<ul style="list-style-type: none"> • 결정적 연계(PLS 기반) • 확률적 연계 	<ul style="list-style-type: none"> • 결정적 연계(PS 기반) • 확률적 연계 	<ul style="list-style-type: none"> • 정확 연계 • 결정적 연계
결측값 대체	<ul style="list-style-type: none"> • 핫-덱 대체 • 보간·외삽 	<ul style="list-style-type: none"> • 모의 세무 기록 참조 • 표본조사와 결합해 보완 	<ul style="list-style-type: none"> • 조사자료 활용 누락 보완 	<ul style="list-style-type: none"> • 결측 대체 미적용 	<ul style="list-style-type: none"> • 원자료 보존 지향 • 제한적 대체
보정	<ul style="list-style-type: none"> • 조사자료 교차 검증 • 가중값 보정 	<ul style="list-style-type: none"> • 인구총조사 기반 가중값 • 조사자료 교차 검증 	<ul style="list-style-type: none"> • 인구총조사 기반 가중값 	<ul style="list-style-type: none"> • 보정 미적용 • 외부 자료 비교 및 문서화 	<ul style="list-style-type: none"> • 보정 미적용
품질 점검	<ul style="list-style-type: none"> • 연계율, 오연계율, 대표성 지표 등 점검 • 메타데이터 제공 	<ul style="list-style-type: none"> • 연계율, 포착률, 정합성, 시계열 일관성 점검 • 메타데이터 공개 	<ul style="list-style-type: none"> • 연계율, 오연계율, 대표성, 시간적 정합성 점검 • 메타데이터 제공 	<ul style="list-style-type: none"> • 메타데이터 및 품질 지표 제공 	<ul style="list-style-type: none"> • 기초·유효성·논리·시계열 점검의 다단계로 운영
시사점	<ul style="list-style-type: none"> • 특정 정책 영역에 집중 	<ul style="list-style-type: none"> • 가족 단위 데이터 통합 • 전 국민을 포괄 	<ul style="list-style-type: none"> • 통계청(ABS)을 중추로 한 범정부적 데이터 통합 • PLS 중심의 데이터 결합 	<ul style="list-style-type: none"> • 단일 식별자 부재 극복 • 최적 추정값을 통한 개념 조화 및 시점 차이 해결 	<ul style="list-style-type: none"> • 제도·운영·기술적 체계화 • 데이터 보안·비식별화

제 5 장

결론 및 제언

제1절 연구 요약·시사점 및 한계

통계생산 기관은 전통적으로 조사를 통해 자료를 수집하고, 단일 자료를 기반으로 통계를 산출해 왔다. 그러나, 행정자료의 확산, 민간 빅데이터의 활용 가능성 그리고 정책 수요의 다양화는 단일 자료만으로는 충분하지 않다는 한계를 드러냈다. 이러한 맥락에서 등장한 것이 바로 데이터 통합이며, 이는 서로 다른 자료원을 하나의 체계 안으로 엮어내는 전략으로 이해할 수 있다.

1. 데이터 통합의 개념 및 유형

데이터 통합의 개념은 초기에는 단순히 동일 구조를 가진 자료를 병합(merging)하는 수준에 머물렀다. 예컨대, 동일 조사에서 여러 해에 걸쳐 생산된 자료를 합쳐 시계열을 구축하거나, 비슷한 설계로 수집된 두 조사자료를 병합하는 정도였다. 본 연구에서 국내외 사례와 국제기구의 지침을 검토한 결과, 오늘날 데이터 통합은 서로 다른 구조와 정의를 가진 자료를 **표준화(alignment and harmonization)**하고, 통합 목적에 맞게 **결합(matching)**하여 **검증된 품질(quality assured)**의 **통계 산출물(integrated data set)**을 생산하는 일련의 절차를 포괄하는 의미로 발전했음을 확인하였다. 이에 본 연구는 이를 국내 통계환경에 적용할 수 있는 개념적 토대로 파악하였고, 데이터 통합의 유형을 다음과 같이 구분하였다.

첫째, 데이터 통합의 궁극적 지향점이 통계생산 및 정책 분석에 있는지, 아니면 통합 가능성 검토와 인프라 구축에 있는지에 따라 활용 목적 통합과 기반 마련 통합으로 구분하였다. 전자는 행정자료와 조사자료를 결합해 소지역 단위 통계를 산출하거나, 빅데이터를 연계하여 실시간 지표를 생산하는 경우와 같이 단기적 성과에 초점을 둔다. 후자는 메타데이터 정비, 코드와 분류 기준 정렬, 품질 관리체계 구축과 같이 당장 산출물은 없지만 향후 안정적인 통합을 가능하게 하는 기반을 마련하는

작업이다.

둘째, 데이터 통합에서 단위는 분석의 정밀도와 활용 가능성을 결정하는 요소로 통계 산출물의 품질에 직접적인 영향을 미친다. 이에 통합의 대상이 되는 분석 단위에 따라 미시 단위, 거시 단위, 미시·거시 단위 통합으로 분류하였다. 이러한 구분은 통합 목적, 자료 속성, 분석 가능성 등을 종합하여 설정되며, 각 통합 단위는 통계의 해석 수준과 적용 가능성에 영향을 미친다.

셋째, 통합 대상 자료의 구조와 형식에 따라 정형·반정형·비정형 자료로 구분하였다. 정형 자료는 구조화된 자료로 변수 정의와 코드 체계가 명확해 결합이 용이하다. 반면, 반정형 자료는 구조는 있으나 표준화가 부족한 자료로 전처리와 정합성 확보가 필수적이다. 비정형 자료는 구조화되지 않은 형태로 사회 현상의 맥락 정보를 담고 있지만 통계적 활용을 위해서는 고도의 전처리와 표준화가 요구된다. 이러한 유형의 구분은 데이터 통합 절차에서 전처리와 품질 점검의 중요성을 더욱 크게 부각시킨다.

넷째, 통합 주체에 따라 의사결정 구조, 법적 근거, 자료 접근 및 보안, 비용 및 위험 분담 방식이 달라지고, 결과적으로 통합의 속도와 품질이 좌우된다. 본 연구는 통합 주체에 따라 기관 내부 통합, 기관 간 통합, 민관 통합으로 구분하였다. 특히, 빅데이터 활용에 따른 민관 통합의 필요성과 위험 요인을 동시에 제시하여, 거버넌스와 제도적 기반 구축의 중요성을 강조하였다.

다섯째, 통합 방법에 따라 레코드 연계, 통계적 매칭, 데이터 융합으로 분류하였다. 이는 단순히 기법을 나열하거나 특정 기법의 우열을 따지는 것이 아니라 데이터 통합 과정의 연속적 구조 속에서 각 방법이 어떤 위치와 기능을 가지는지 체계화하기 위함이다. 레코드 연계는 정밀성의 토대를, 통계적 매칭은 불완전성 보완과 대체의 수단을, 데이터 융합은 최종적 일관성 확보와 보정의 장치를 제공한다. 즉, 세 방법은 「결합-대체-보정」이라는 일련의 흐름 속에 모두 녹아 있으며, 각각의 방법이 특정 단계에 한정되지 않고 절차 전체를 관통하면서 상호보완적으로 작동하고 있음을 제시하기 위함이다.

2. 데이터 통합 과정 및 방법

본 연구는 데이터 통합을 단순 자료 병합이 아니라, 이질적 자료를 정렬하고 조화하여 분석이 가능한 하나의 체계로 구성하는 과정으로 정의하였다. 이는 단편적 기법 중심 논의와 달리, 통계생산 전반의 품질 관리와 활용성을 동시에 고려한 전(全)주기적 관리체계로서의 성격을 강조한 것이다. 이러한 접근은 각 단계가 고립적으로

존재하는 것이 아니라, 하나의 흐름 안에서 서로를 전제하고 보완하는 연결 구조임을 보여준다. 따라서 본 연구가 제시한 절차는 단순한 방법적 제안이 아니라, 향후 데이터 통합을 제도적으로 관리하고 품질을 보증하기 위한 거버넌스 틀이라는 점에서 의의가 있다.

가. 전처리

본 연구는 전처리를 단순한 사전 준비가 아니라, 데이터 통합 전체 절차의 성패를 좌우하는 규범적·기술적 토대로 접근하였다. 이를 위해 전처리 원칙으로 ①원자료의 정보 손실을 최소화하는 최소침습성, ②모든 변환을 문서로 관리하여 동일 결과를 재현할 수 있도록 하는 재현성, ③변수 정의·단위·분류 등 메타데이터를 확보한 뒤 변환의 근거를 기록하는 메타데이터 우선, ④개인정보 보호, 접근권한 등 윤리적·법적 요건 준수, ⑤데이터의 호환성을 확보하는 상호운용성을 제안하였다. 나아가 원칙 확립을 위해 ①스키마 정렬과 ②개념 조화를 전처리의 필수 절차로 정립하였다.

나. 정합성 점검

본 연구는 정합성 점검을 단순 확인 작업이 아닌, 데이터 결합 가능성을 판정하는 독립적 절차로 격상하였다. Eurostat은 제출 자료에서 급격한 값의 변동이나 불일치가 발견되면 즉시 수정·재작성을 요구한다. IMF 역시 SDDS 기준에서 통계 제출국이 반드시 정합성 검증 절차를 거치도록 규정한다. 이는 정합성 점검이 단순 보조가 아니라 데이터 신뢰성을 담보하는 핵심 단계임을 보여준다. 국내에서도 행정자료와 조사자료 간 단위 불일치, 조사 시점의 어긋남, 코드 체계 불일치 같은 문제가 자주 발생한다. 이를 전처리 단계에서 모두 해결하기는 어렵다. 따라서 본 연구는 정합성 점검을 별도 관문으로 설정하여, 결합 여부를 판정하고, 오류를 조기에 발견해 수정하도록 제안하였다.

다. 결합 단계

결합 과정을 레코드 연계(record linkage), 통계적 매칭(statistical matching), 합성 매칭(synthetic matching)의 세 축으로 구조화하였다. 결합은 단순히 서로 다른 자료를 이어 붙이는 기술적 과정이 아니다. 결합 단계의 성패는 이후 절차인 결측값 대체와 보정의 부담과 성과에 직접적으로 영향을 미치기 때문에, 전체 절차에서 중추적 관문의 역할을 한다. ①레코드 연계는 강력하고 우선적인 결합 방법이지만, 동시에 식별자 오류·누락·불일치가 존재하면 한계에 부딪힌다. 따라서, 레코드 연계는 최선의 시도인 동시에 다른 방법을 호출하게 만드는 절차적 분기점이다. 연계 성공률이

높다면 이후 결측값 대체의 부담이 줄고, 연계 성공률이 낮다면 통계적 매칭과 합성 매칭에 크게 의존하게 된다. 즉, 레코드 연계는 통합 절차에서 부하 조절기(load controller)와 같은 기능을 수행한다. ②통계적 매칭은 레코드 연계가 한계에 도달했을 때 나타나는 공백을 메우는 완충 장치(buffer)이다. 만약 통계적 매칭이 없다면, 결합 후 남는 공백이 전부 결측값 대체 단계로 넘어가 버려, 대체 과정이 지나치게 복잡해지고 불확실성이 증가한다. 반면 통계적 매칭을 도입하면, 결측값 대체 단계가 다뤄야 할 불확실성의 규모가 줄어들어 절차 전체의 안정성이 강화된다. ③합성 매칭은 레코드 연계와 통계적 매칭을 혼합하여, 각 방식이 가진 장점을 살리고 단점을 보완하는 접근이다. 합성 매칭은 두 접근을 결합해 잔여 오차를 최소화하고, 결합률과 대표성 간 균형을 맞추는 역할을 한다. 이처럼 결합 단계를 세 가지 축으로 구조화한 이유는, 단순히 기법을 병렬적으로 설명하기 위함이 아니다. 레코드 연계, 통계적 매칭, 합성 매칭은 각각 다른 장점을 가지며, 데이터의 특성과 조건에 따라 선택적으로 활용될 수 있다. 하나의 기법만으로도 충분할 수 있고, 또 다른 경우에는 두 가지 이상을 조합해야 할 수도 있다. 중요한 것은 본 연구가 이 세 가지 방법을 모두 수행하라는 지침을 제시한 것이 아니라, 통합 목적과 데이터 특성에 맞추어 적합한 방식을 체계적으로 선택·설계할 수 있도록 하는 틀을 마련했다는 점이다.

라. 결측값 대체

결합 후에도 공백은 존재하며, 이 공백이 크면 이후 단계에서 대표성을 보장하기 어려워진다. 결측값 대체를 제대로 수행하지 않으면, 보정 단계는 모집단과의 차이를 다루는 본래의 역할에 집중할 수 없다. 즉, 결측값 대체는 보정의 부담을 줄이고 전체 절차를 안정화하는 장치이다. 또한, 결측값 대체는 불확실성 관리 장치라는 점에서도 중요하다. 결측을 대체하는 과정에서 임의성이 개입될 수밖에 없는데, 이를 어떻게 통제하느냐가 통합 데이터 세트의 품질을 좌우한다. 본 연구는 이러한 불확실성 관리 측면을 강조하며, 결측값 대체가 빈칸을 채우는 기술이 아니라, 불확실성을 구조적으로 통제하는 절차임을 명확히 했다.

마. 보정 및 대표성 확보

보정은 단순한 가중값 조정이 아닌, 제약 조건을 활용한 체계적 절차이다. 즉, 강제 제약, 벌점 제약, 참조 제약을 함께 고려하여, 초기 가중값과 외부 기준 간의 균형을 맞춘다. 이를 통해 통합 데이터 세트는 대표성을 확보하면서도, 지나친 왜곡이나 극단값 발생을 방지할 수 있다. 보정은 앞선 단계와 긴밀히 연결된다. 전처리와 정합성 점검에서 오류와 불일치를 줄였더라도, 결합과 결측값 대체를 거치는 과정에서 발생하는 편향은 피하기 어렵다. 보정은 이 편향을 최종적으로 교정하여 통합 데이터

세트가 모집단 기준과 일치하도록 한다. 즉, 보정은 「전처리-정합성 점검-결합-결측값 대체」라는 일련의 절차를 거쳐 축적된 결과를 전제로 작동한다. 보정 이후에는 대표성 점검이 뒤따라야 한다. 이는 보정 결과가 단순히 수학적 제약을 충족하는 데 그치지 않고, 실제 극단 가중값이 과도하게 발생하지 않았는지, 가중값 분포가 안정적인지, 외부 기준과 비교했을 때 허용 범위 내에 있는지를 검증하는 과정이다. 보정은 대표성을 확보하는 관문, 대표성 점검은 그 결과의 타당성을 검증하는 최종 안전판이다. 이는 데이터 통합 결과를 단순히 ‘결합 데이터’가 아니라 ‘대표성과 완결성을 갖춘 통합 데이터 세트’로 정의하도록 한다.

바. 품질 점검

품질 점검은 단순히 산출된 결과를 확인하는 절차를 넘어, 전 과정에 걸쳐 내재한 오류와 왜곡 가능성을 차단하고, 최종적으로 활용할 수 있는 결과물을 보장하는 이중 장치로 이해되어야 한다. 본 연구는 이를 통합 단계별 품질 점검 요소와 최종 산출물로서의 품질 점검이라는 두 축으로 구분하여 제시하였다. 전자는 통합 과정의 각 단계가 적절하게 수행되었는지를 점검하여 오류 전파를 막는 과정 중심의 장치이며, 후자는 완성된 통합 데이터 세트가 실제로 활용에 적합한지를 판정하는 결과 중심의 장치이다. 이처럼 품질 점검을 이원화한 것은 데이터 통합 품질 관리의 제도적 토대를 제공한다. 단계별 품질 점검이 오류 전파를 차단해 품질의 하한선을 보장하고, 최종 산출물 점검이 사용 적합성을 판정해 상한선을 확보한다는 구조는, 향후 국가 통계생산에서 품질 보증 체계를 설계할 때 직접적인 준거가 될 수 있다.

3. 연구 한계

본 연구는 데이터 통합의 개념과 방법을 정립하고 체계화하는 데 초점을 두었으나, 몇 가지 중요한 한계를 지닌다.

첫째, 연구 성과가 개념적·방법론적 수준에 머물러 있다는 점이다. 데이터 통합 방법을 종합적으로 정리했지만, 세부 매뉴얼이나 시뮬레이션 기반 검증은 담아내지 못했다. 이를 보완하기 위해 본 연구진은 별도로 「경제총조사 특성 항목 개선 연구」를 진행 중이다. 해당 연구는 웹 스크래핑 기술을 활용하여 통합 데이터 세트를 구축 후, 이를 경제총조사에 활용하는 방안을 제안하는 연구이다. 즉, 본 연구가 개념적 틀을 제공했다면, 후속 연구는 이를 실증적으로 보완하는 방향으로 이어지고 있다.

둘째, 본 연구는 기술 발전 속도를 충분히 반영하지 못했다. 최근 데이터 통합

영역에서는 기계학습 기반의 레코드 연계, 딥러닝을 활용한 결측값 대체, 자동화된 품질 관리 도구 등이 빠르게 도입되고 있다. 그러나, 데이터 통합의 전체 절차를 조망하는 연구 설계로 인해 특정 기술 발전의 세부 적용 가능성을 깊이 다루지 못했다. 향후 연구에서는 이러한 데이터 과학 기술을 통합 절차에 어떻게 배치하고 제도적으로 수용할 것인지 탐색할 필요가 있다.

셋째, 사례 적용의 제약을 지닌다. 해외 사례는 공개 문헌을 통해 개괄적으로 파악했을 뿐, 실제 운영기관 내부의 세부 절차나 품질 관리 방식까지는 확인하기 어려웠다. 국내 사례 또한 자료 접근성과 제도적 한계로 인해, 제시한 절차별 점검 항목을 실무적으로 검증하는 수준에는 이르지 못했다. 비록 절차 체계의 타당성은 확보했으나, 그 실효성은 추가적인 사례연구와 검증을 통해 보완해야 한다.

제2절 데이터 통합 활성화를 위한 제언

데이터 통합 활성화를 위해서는 제도적 기반, 기술적 인프라, 거버넌스 체계, 품질 관리 체계가 유기적으로 작동해야 한다. 이를 위해 다음 다섯 가지 전략을 제안한다.

첫째, 중앙 관리체계 확립이다. 데이터 통합을 국가 차원에서 안정적으로 수행하기 위해서는 중앙 집중형 관리체계(centralized coordination system)의 확립이 필요하다. **국가데이터처를 중심으로** 각 부처와 지방정부, 공공기관이 생산·보유하는 행정자료의 관리 현황을 통합적으로 파악하고, 결합이 가능한 항목을 주기적으로 점검해야 한다. 또한, 데이터 표준화·보안·품질 관리 지침을 통합 관리하는 전담 조직을 운영하여 통합 절차의 일관성과 투명성을 확보해야 한다. 이와 함께, 통계 생산기관 간 데이터 접근 권한과 이용 절차를 명확히 규정하여 기관 간 협력의 효율성을 높일 필요가 있다.

둘째, 데이터 거버넌스 활성화가 필수적이다. 효율적 데이터 통합을 위해서는 다기관 협력체계(data governance)가 필수적이다. 현재 각 기관은 자체 목적에 따라 자료를 관리하고 있으나, 데이터 통합을 위한 협의 및 의사결정 구조는 아직 미흡하다. 따라서, **「중앙-지방-공공기관-민간」 간의 협의체 기반 거버넌스를 구축**하여, 데이터 표준화, 접근권한, 품질 기준 등에 관한 합의적 기준을 마련해야 한다. 또한, 데이터 거버넌스는 단순 관리 차원을 넘어, 윤리적·법적 준수, 보안 및 개인정보 보호 체계와 연동된 통합적 관리 메커니즘으로 발전해야 한다.

셋째, 반정형·비정형 자료의 통합 연구를 확대해야 한다. 기존의 데이터 통합이 구조화된 행정자료와 조사자료 중심으로 이루어졌다면, 앞으로는 반정형(semi-structured) 및 비정형(unstructured) 자료의 통합 활용 연구가 필요하다. 예를 들어, 텍스트, 이미지, 위치 정보, 로그 데이터 등은 정책 분석과 사회경제 현상 파악에 필수적인 새로운 데이터 자원이다. 따라서, 인공지능(AI) 기반 데이터 정제·분류 알고리즘과 자연어처리(NLP) 기술을 접목하여, 반정형 자료는 표 형태(data table)로 표준화하고 비정형 자료를 구조화된 데이터로 변화하여 메타데이터 체계에 포함하는 연구가 확대되어야 한다. 이는 데이터 통합의 기술적 영역을 확장함과 동시에, 미래 통계환경 변화에 대응하기 위한 기반이 될 것이다.

넷째, 실험적 데이터 통합 프로젝트의 확대도 필요하다. 예컨대, 본 연구진이 진행 중인 「경제총조사 특성 항목 개선 연구」는 웹 스크래핑 기술을 활용하여 통합 데이터 세트를 구축하는 사례로, 연구에서 제시한 절차 체계를 실증적으로 검증한 시도라 할 수 있다. 이러한 실험은 절차 체계의 실효성을 점검하고, 세부 지침을 구체화하는 데 이바지한다. 향후, 다양한 영역에서 실험적 데이터 통합 프로젝트(pilot data integration projects)를 확대하고, 그 결과를 문서화 및 메타데이터로 축적하여 통합 매뉴얼 또는 가이드라인으로 발전시킬 필요가 있다.

다섯째, 메타데이터 작성 체계의 정립이 필요하다. 데이터 통합의 투명성(transparency)과 재현성(reproducibility)을 확보하기 위해서는 메타데이터 작성 및 관리체계 구축이 필수적이다. 메타데이터는 단순한 설명 자료가 아니라, 자료 수집부터 품질 점검에 이르는 데이터 통합의 전 과정을 체계적으로 기록함으로써 데이터 통합 절차를 추적하고 일관되게 관리할 수 있는 기반이 된다. 특히, 결합·결측값 대체·보정 과정과 그 전후의 값을 명확히 기록함으로써 데이터 세트가 어떤 과정을 거쳐 생성·수정·검증되었는지를 투명하게 제시할 수 있다. 이는 데이터 통합의 검증가능성(auditability)을 강화하고, 통합 결과의 신뢰성과 타당성을 실증적으로 뒷받침한다. 나아가 메타데이터 작성 체계를 품질 관리체계와 통합하여 운영할 필요가 있다. 메타데이터가 품질 점검 결과와 자동으로 연계되어 갱신된다면, 데이터의 신뢰성과 관리 효율성이 동시에 제고될 것이다. 이를 통해 메타데이터는 단순한 기록 수단을 넘어, 데이터 통합 품질 관리의 핵심 인프라로 기능하게 될 것이다.

참고문헌

「국내 문헌」

- 법제처. (2024). 통계법.
통계청(현 국가데이터처). (2023). 기업통계등록부(SBR) 직무편람.
통계청. (n.d.). 통계용어사전.

「해외 문헌」

- Abayomi, K. et al. (2008). "Diagnostics for multivariate imputations". *Journal of the Royal Statistical Society Series C: Applied Statistics*, 57(3), 273-291.
- Abowd, J. M. et al. (2009). *The LEHD infrastructure files and the creation of the quarterly workforce indicators*. U.S. Census Bureau.
- Abowd, J. M., & Stinson, M. H. (2013). "Estimating measurement error in annual job earnings: A comparison of survey and administrative data". *Review of Economics and Statistics*, 95(5), 1451-1467.
- Abowd, J. M. et al. (2021). *Multiple-imputation record linkage using machine learning (CES-WP-21-35)*. U.S. Census Bureau.
- ABS. (2020). *Labour Account Australia: Methodology*. Australian Bureau of Statistics.
- ABS. (2017). *Census of Population and Housing: Understanding the Census and Census Data, Australia, 2016*. Australian Bureau of Statistics.
- ABS. (2018). *About the Multi-Agency Data Integration Project (MADIP)*. Australian Bureau of Statistics.
- ABS. (n.d.-a). *Multi-Agency Data Integration Project (MADIP)*. Australian Bureau of Statistics.
- ABS. (n.d.-b). *Data integration*. Australian Bureau of Statistics.
- ADB. (2018). *User Guide for ADB Statistical Business Register*. Asian Development Bank.
- Agresti, A. (2002). *Categorical data analysis (2nd ed.)*. Wiley.
- AIHW. (2022). *Data integration on Australia's health and welfare statistics: Technical report*. Australian Institute of Health and Welfare.
- Andridge, R. R. and Little, R. J. A. (2010). "A review of hot deck imputation for survey non-response". *International Statistical Review*, 78(1), 40-64.
- Bethlehem, J. (2010). "Selection bias in web surveys". *International Statistical Review*, 78(2), 161-188.
- Benzeval, M. et al. (2020). *Integrated data: Research potential and data quality (Understanding Society Working Paper No. 2020-02)*. University of Essex.

- Berger, B. et al. (2021). "Levenshtein distance, sequence comparison and biological database search". *IEEE Transactions on Information Theory*, 67(6), 3287-3294.
- Besag, J. et al. (1991). "Bayesian image restoration, with two applications in spatial statistics". *Annals of the Institute of Statistical Mathematics*, 43(1), 1-20.
- Black, A. (2016). *The IDI prototype spine's creation and coverage*. Statistics New Zealand.
- Boonstra, P. S. and del Pino, P. O. (2025). "A comparison of some existing and novel methods for integration historical models to improve estimation of coefficients in logistic regression". *Journal of the Royal Statistical Society Series A: Statistics in Society*, 188, 46-67.
- Breiman, L. (2001). "Random forests". *Machine Learning*, 45(1), 5-32.
- Brick, J. M. and Kaltion, G. (1996). "Handling missing data in survey research". *Statistical Methods in Medical Research*, 5(3), 215-238.
- Chen, Y. et al. (2020). "Doubly robust inference with nonprobability survey samples". *Journal of the American Statistical Association*, 115(532), 2011-2021.
- Christen, P. (2012). *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer Science and Business Media.
- Cibella, N. et al. (2009). *Theory and practice in developing a record linkage software*. ESSnet-ISAD workshop, Eurostat.
- CIID. (2021). *Data Integration Toolkit*. Center for Innovation in Infrastructure and Development.
- Cohen, J. (1960). "A coefficient of agreement for nominal scales". *Educational and Psychological Measurement*, 20(1), 37-46.
- Comber, A. and Zeng, W. (2022). *Areal interpolation*. The Geographic Information Science & Technology Body of Knowledge.
- Corak, M. (2013). "Income inequality, equality of opportunity, and intergenerational mobility". *Journal of Economic Perspectives*, 27(3), 79-102.
- Cover, T. M. and Thomas, J. A. (2006). *Elements of information theory (2nd ed.)*. Wiley.
- Cristianini, N. and Scholkopf, B. (2002). "Support vector machines and kernel methods: The new generation of learning machines". *AI Magazine*, 23(3), 31-41.
- Dagum, E. B. and Cholette, P. A. (2006). *Seasonal Adjustment and Time Series*. Springer.
- De Maesschalck, R. et al. (2000). "The Mahalanobis distance". *Chemometrics and Intelligent Laboratory System*, 50(1), 1-18.
- Dempster, A. P. et al. (1977). "Maximum likelihood from incomplete data via the EM algorithm". *Journal of the Royal Statistical Society Series B: Methodology*, 39(1), 1-38.
- Deville, J.-C. and Särndal, C.-E. (1992). "Calibration estimators in survey sampling". *Journal of the American Statistical Association*, 87, 376-382.
- de Waal, T. et al. (2019). "Multi-source statistics: Basic situations and methods". *International Statistical Review*, 87(S1), 179-204.
- Doan, A. and Halevy, A. (2005). "Semantic integration research in the database community: A brief survey". *AI Magazine*, 26(1), 83-94.
- Dong, X. L. and Srivastava, D. (2021). *Big data integration*. Morgan & Claypool Publishers.
- D'Orazio, M. et al. (2006). *Statistical Matching: Theory and Practice*. Wiley.

- D’Orazio, M. (2017). *Statistical matching and imputation of survey data*. ESSnet Project on Data Integration.
- D’Orazio, M. et al. (2024). *What is the state of play on statistical matching with a focus on auxiliary information*. Joint Research Centre, European Commission.
- Emmanuelr, T. et al. (2021). “A survey on missing data on machine learning”. *Journal of Big Data*, 8, 140.
- Emmenegger, J. et al. (2023). “Evaluating data fusion methods to improve income modeling”. *Journal of Survey Statistics and Methodology*, 11(1), 643-667.
- Enamorado, T. et al. (2019). “Using a probabilistic model to assist merging of large-scale administrative records”. *American Political Science Review*, 113(2), 353-371.
- encord. (n.d.). ROC Curve. Retrieved July 7, 2025, from <https://encord.com/glossary/roc-definition/>
- Endres, D. M. and Schindelin, J. E. (2003). “A new metric for probability distributions”. *IEEE Transactions of Information Theory*, 49(7), 1858-1860.
- ESCAP. (2020). *Asia-Pacific guidelines to data integration for official statistics*. United Nations Economic and Social Commission for Asia and the Pacific.
- European Commission. (2020). *Towards a European strategy on business-to-government data sharing for the public interest*. European Union.
- European Parliament & Council. (2024, November 27). *Regulation (EU) 2024/3018 of the European Parliament and of the Council of 27 November 2024 amending Regulation (EC) No 223/2009 on European statistics (Text with EEA relevance)*. Retrieved from October 28, 2025, <https://eur-lex.europa.eu/eli/reg/2024/3018/oj>
- Eurostat. (2009). *Insights on data integration methodologies*.
- Eurostat. (2013). *Handbook on precision requirements and variance estimation for ESS household surveys*.
- Eurostat. (2014). *ESS Handbook for quality reports*.
- Eurostat. (2017). *European Statistics Code of Practice and Quality Assurance Framework*. Retrieved September 17, 2025, from <https://ec.europa.eu/serostat/web/products-catalogues/-/ks-02-18-142>.
- Eurostat. (2020). *Methodological guidelines and description of EU-SILC target variables*.
- Eurostat. (2021a). *Quality assurance of statistical integration: Macro and micro validation*.
- Eurostat. (2021b). *Quality assurance framework of the European Statistical System*.
- Eurostat. (2023). *VTL 2.0*. Retrieved September 17, 2025, from <https://cros.ec.europa.eu/book-page/vtl-20>.
- Eurostat. (2024). *Integration of statistical and geospatial information*.
- Eurostat and UN-GGIM Europe. (2023). *Geospatial data quality requirements: Update frequency and timeliness*. In *Recommendations for Geospatial Quality Reporting*. European Union.
- Fawcett, T. (2006). “An introduction to ROC analysis”. *Pattern Recognition Letters*, 27, 861-874.
- Fellegi, I. P. and Sunter, A. B. (1969). “A theory for record linkage”. *Journal of the American Statistical Association*, 64, 1183-1210.
- Frenette, M. et al. (2025). *Leveraging Statistics Canada data integration opportunities for program evaluation*. Statistics Canada.

- Gao, P. A. and Wakefield, J. (2022). "A spatial variance-smoothing area level model for small area estimation of demographic rates". *International Statistical Review*, 91(3), 493-510.
- Geourjon, A. -M. et al. (2023). *The Use of Mirror Data by Customs Administrations: From Principles to Practice (IMF Technical Notes and Manuals No. 2023/005)*. International Monetary Fund.
- Gibb, S. (2016). *Identifying the New Zealand resident population in the Integrated Data Infrastructure (IDI)*. Statistics New Zealand.
- Groves, R. M. et al. (2009). *Survey methodology (2nd ed.)*. Wiley.
- Han, P. and Si, Y. (2025). "Frontiers in data integration". *Journal of the Royal Statistical Society Series A*, 188, 24-26,
- Harron, K. L. (2016). *An introduction to data linkage*. Administrative Data Research Network.
- Harron, K. et al. (2017). "A guide to evaluation linkage quality for the analysis of linked data". *International Journal of Epidemiology*, 46(5), 1699-1710.
- Herzog, T. N. et al. (2007). *Data Quality and Record Linkage Techniques*. Springer.
- Hyndman, R. J. and Athanasopoulos, G. (2021). *Forecasting: Principles and Practice (3rd ed.)*. OTexts.
- IMF. (2012a). *Quarterly National Accounts Manual: Chapter 8. Chain-linking in the QNA*.
- IMF. (2012b). *Data Quality Assessment Framework (DQAF)*.
- IMF. (2014). *Government Finance Statistics Manual 2014*.
- IMF. (2017). *Consumer Price Index Manual: Chapter 9. Updating weights and linking series*.
- IMF. (2019). *Balance of Payments and International Investment Position Manual (BPM6) Compilation Guide*.
- IMF. (2023a). *International Financial Statistics (IFS)*. Retrieved September 1, 2025, from <https://data.imf.org/IFS>
- IMF. (2023b). *Dissemination Standards Bulletin Board (DSBB)*. Retrieved September 2, 2025, from <https://dsbb.imf.org/>
- Jaro, M. A. (1989). "Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida". *Journal of the American Statistical Association*, 84, 414-420.
- JRC. (2021). *Sharing and using geospatial data across borders: Spatial data infrastructures for the digital economy*. Joint Research Centre, European Union.
- Kiesl, H. and Rässler, S. (2009). *The validity of data fusion*. European Union.
- Kim, J. K. and Rao, J. N. K. (2012). "Combining data from two independent surveys: A model-assisted approach". *Biometrika*, 99(1), 85-100.
- Kim, J. K. and Tam, S.-M. (2020). "Data integration by combining big data and survey sample data for finite population inference". *International Statistical Review*, 89(2), 382-401.
- Kish, L. (1992). "Weighting for unequal P". *Journal of Official Statistics*, 8(2), 183-200.
- Lepot, M. et al. (2017). "Interpolation in time series: An introductory overview of interpolation methods". *Water*, 9(10), 796.
- Leulescu, A. and Agafitei, M. (2013). *Statistical matching: A model based approach for data integration*. Eurostat Methodologies and Working Papers.
- Liseo, B. and Tancredi, A. (2009). *Model based record linkage: A Bayesian perspective*. European Union.

- Little, R. J. A. and Rubin, D. B. (2019). *Statistical analysis with missing data (3rd ed.)*. Wiley.
- Lohr, S. (1999). *Sampling: Design and Analysis*. Duxbury.
- Maki, M. et al. (2017). “Estimation of rice yield by SIMRIW-RS, a model that integrates remote sensing data into a crop growth model”. *Journal of Agricultural Meteorology*, 73(1), 2-8.
- Manski, C. F. (2003). *Partial identification of probability distributions*. Springer.
- Marchant, N. G. et al. (2023). “Bayesian graphical entity resolution using exchangeable random partition priors”. *Journal of Survey Statistics and Methodology*, 11(3), 569-596.
- McCallum, A. et al. (2000). *Efficient clustering of high-dimensional data sets with application to references matching*. Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 169-178). Association for Computing Machinery.
- Meijering, E. (2002). “A chronology of interpolation: From ancient astronomy to modern signal and image processing”. *Proceedings of IEEE*, 90(3), 319-342.
- Milligan, K. (2016). “The Canadian tax and credit system: Recent developments and policy challenges”. *Canadian Tax Journal*, 64(2), 471-500.
- Montgomery County Department of Health and Human Service. (2022). *Electronic Integrated Case Management System (eICM): Data sharing for program operations while protecting privacy*.
- Moriarity, C. and Scheuren, F. (2001). “Statistical matching: A paradigm for assessing the uncertainty in the procedure”. *Journal of Official Statistics*, 17(3), 407-422.
- Nelsen, R. B. (2006). *An introduction to copulas (2nd ed.)*. Springer.
- NIST. (n.d.). *Kolmogorov-Smirnov Goodness-of-Fit Test: Kolmogorov-Smirnov Two-Sample Test*. National Institute of Standards and Technology.
- Newbury, J. (1981). *Linear interpolation*. Palgrave.
- Newhouse, D. (2023). *Small area estimation for poverty and wealth using geospatial data: What have we learned so far?*. World Bank.
- OECD. (2013). *OECD Skills Outlook 2013: First Results from the Survey of Adult Skills*.
- OECD. (2019a). *The Path to Becoming a Data-Driven Public Sector*.
- OECD. (2019b). *A data-driven public sector: Enabling the strategic use of data for productive, inclusive and trustworthy governance*.
- OECD. (2023a). *OECD Health Statistics 2023*. Retrieved September 12, 2025, from <https://www.oecd.org/health/health-data.htm>
- OECD. (2023b). *Education at a Glance 2023: OECD Indicators*.
- Parker, C. et al. (2024). “Evaluating data quality for blended data”. *Journal of Official Statistics*, 40(2), 123-145.
- PC. (2017). *Data Availability and Use, Inquiry Report No. 82*. Australian Government Productivity Commission.
- Rahm, E. and Do, H. H. (2000). “Data cleaning: Problems and current approaches”. *IEEE Data Engineering Bulletin*, 24(4), 3-13.
- Rahm, E. and Bernstein, P. A. (2001). “A survey of approaches to automatic schema matching”. *The VLDB Journal*, 10(4), 334-350.
- Rässler, S. (2002). *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*. Springer.

- Reich, V. F. M. (2024). *Machine learning based linkage of company data for economic research*. ifo Institute.
- Reich, N., Scholz, P. and Wenzig, C. (2024). *Linkage of company data using machine learning*. German Council for Social and Economic Data (RatSWD).
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley.
- Saaty, T. L. (1980). *The analytic hierarchy process*. McGraw-Hill.
- Särndal, C. E. et al. (2003). *Model assisted survey sampling*. Springer.
- Särndal, C. E., & Lundström, S. (2005). *Estimation in surveys with nonresponse*. Wiley.
- Särndal, C. E. (2007). “The calibration approach in survey theory and practice”. *Survey Methodology*, 33(2), 99-119.
- Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Chapman & Hall/CRC.
- Schiavina, M. et al. (2023). “A smart and flexible approach for aggregation of adjacent polygons to meet a minimum target area or attribute value”. *Scientific Reports*, 13(1), Article 4367.
- Sinnott, R. W. (1984). “Virtues of the Haversine”. *Sky and Telescope*, 68(2), 159.
- SRI International, AnLar and Actionable Intelligence for Social Policy. (2024). *Data Linkage and Integration for Research and Statistical Purposes: an Annotated Bibliography*. Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.
- Statistics Canada. (2017). *T1 Family File (TIFF), reference guide*. Retrieved September 16, 2025, from <https://www150.statcan.gc.ca/n1/en/catalogue/72-212-X>
- Statistics Canada. (2024). Education and Labour Market Longitudinal Platform (ELMLP): User guide.
- Statistics Canada. (n.d.). *Longitudinal Administrative Databank (LAD)*.
- Statistics Netherlands. (2016). *System of social statistical datasets (SSD)*.
- Statistics New Zealand. (2013). *Data Integration Manual (2nd ed.)*.
- Statistics New Zealand. (n.d.-a). *Integrated Data Infrastructure (IDI)*. Retrieved September 21, 2025, from <https://www.stats.govt.nz/integrated-data/integrated-data-infrastructure/>
- Statistics New Zealand. (n.d.-b). *Data in the IDI*. Retrieved September 21, 2025, from <https://www.stats.govt.nz/integrated-data/integrated-data-infrastructure/data-in-the-idi/>
- Statistics New Zealand. (n.d.-c). *How we keep integrated data safe*. Retrieved September 21, 2025, from <https://www.stats.govt.nz/integrated-data/how-we-keep-integrated-data-safe/>
- Stekhoven, D. J. and Bühlmann, P. (2012). “MissForest-non-parametric missing value imputation for mixed-type data”. *Bioinformatics*, 28(1), 112-118.
- Stuart, E. A. (2010). “Matching methods for causal inference: A review and a look forward”. *Statistical Science*, 25(1), 1-21.
- Tancredi, A. and Liseo, B. (2011). “A hierarchical Bayesian approach to record linkage and population size problems”. *Annals of Applied Statistics*, 5(2B), 1553-1585.
- Tiecke, T. G. et al. (2017). *Mapping the world population one building at a time*. World Bank.
- Trant, M. and Whitridge, P. (1999). *Integration of administrative data with survey and census data*. Statistics Canada.
- UNECE. (2014). *In-depth review of big data*. United Nations Economic Commission for Europe.

- UNECE. (2017). *Quality indicators for the Generic Statistical Business Process Model (GSBPM): For statistics derived from surveys and administrative data sources (Version 2.0)*. United Nations Economic Commission for Europe.
- UNECE. (2018). *Guidelines on the use of statistical business registers for business demography and entrepreneurship statistics*. United Nations Economic Commission for Europe.
- UNECE. (2019a). *Guidance on modernizing statistical legislation*. United Nations Economic Commission for Europe.
- UNECE. (2019b). *Guidelines on data integration for measuring SDGs*. United Nations Economic Commission for Europe.
- UNECE. (2020). *A guide to data integration for official statistics (Version 2.0)*. High-Level Group for the Modernisation of Official Statistics (HLG-MOS), United Nations Economic Commission for Europe.
- UNECE. (2024). *In-depth review of timeliness, frequency and granularity of official statistics (ECE/CES/2024/6)*. United Nations Economic Commission for Europe.
- UNECE. (n.d.). *Generic Statistical Business Process Model (GSBPM), Version 5.1*. United Nations Economic Commission for Europe.
- UNFPA. (2025). *Trends in maternal mortality 2000-2023: Estimates by WHO, UNICEF, UNFPA, World Bank Group and UNDESA Population Division*. United Nations Population Fund.
- UN-GGIM Europe. (2022). *Core Spatial Data Theme Statistical Units: Recommendation for Content (Version 1.1)*. United Nations Committee of Experts on Global Geospatial Information Management.
- UNICEF. (2024). *Levels and trends in child mortality: Report 2023*. United Nations International Children's Emergency Fund.
- UNSC and UNECE. (2006). *Statistical data editing: Volume 3. Impact on data quality*. United Nations Statistical Commission and United Nations Economic Commission for Europe.
- UNSD. (2018). *Use of administrative data for official statistics: Steps to integrate administrative data*. United Nations Statistics Division.
- UNSD. (2021). *The Sustainable Development Goals Report 2021*. United Nations Statistics Division.
- UNSD. (2022). *Guidance and Toolkit for Quality Assessment of Administrative Data for Official Statistics*. United Nations Statistics Division.
- UNSD. (n.d.-a). *Glossary of statistical terms: Data integration*. United Nations Statistics Division.
- UNSD. (n.d.-b). *SDG Global Database*. United Nations Statistics Division. Retrieved September 13, 2025, from <https://unstats.un.org/sdgs/indicators/> database
- UNSD. (n.d.-c). *SDMX for the Sustainable Development Goals*. Retrieved September 3, 2025, from <https://statistics.unsdglearn.org/courses/sdmx-for-the-sustainable-development-goals/>
- UNSD. (n.d.-d). *SDG Indicators: Metadata repository*. United Nations Statistics Division. Retrieved September 4, 2025, from <https://unstats.un.org/sdgs/metadata>
- US Census. (2024). *Design and Methodology Report*. Retrieved September 19, 2025, from <https://census.gov/programs-surveys/acs/methodology/desing-and-methodology.html>
- Valliant, R. et al. (2018). *Practical tools for designing and weighting survey samples (2nd ed.)*. Springer.

- Viljuber, L. et al. (2013). *LEHD Infrastructure files and the creation of the Quarterly Workforce Indicators*. U.S. Census Bureau.
- Viljanen, M. et al. (2022). “A machine learning approach to small area estimation: Predicting the health, housing and well-being of the population of Netherlands”. *International Journal of Health Geographics*, 21(1), 4.
- Villani, C. (2008). *Optimal Transport: Old and New*. Springer.
- Wagner, D., and Layne, M. (2014). *The person identification validation system (PVS): Applying the Center for Administrative Records Research and Applications (CARRA) record linkage software*. U.S. Census Bureau.
- WHO. (2020). *World health statistics 2020: monitoring health for the SDGs*. World Health Organization.
- Winkler, W. E. (1990). *String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage*. U.S. Census Bureau.
- Winkler, W. E. (2006). *Overview of record linkage and current research directions*. U.S. Census Bureau.
- World Bank. (2020). *Poverty and Shared Prosperity 2020: Reversals of Fortune*.
- Zheng, Z. et al. (2025). *Dynamic, high-resolution wealth measurement in data-scarce environments (Policy Research Working Paper No. 11058)*. World Bank.

부 록

1. 데이터 통합 관련 용어 정리

데이터 통합에는 통계적·행정적·기술적 배경이 다양한 자료가 활용되며, 이에 따라 관련 용어가 사용되는 맥락 또한 다양하다. 이러한 다양성은 용어의 혼용과 개념적 오해를 초래할 수 있으며, 실제 통계 현장에서는 기술적 구현 방식과 이론적 정의가 일치하지 않는 용어들이 병렬적으로 사용되는 경우도 많다. 따라서, 일관된 개념 정의와 명확한 용어 기준을 확립하는 것이 필수적이다. 이에 데이터 통합과 관련된 주요 개념 및 용어를 정의하고, 상호 비교함으로써 통계 생산자와 이용자 간의 공통 이해 기반을 마련하고자 한다. 이는 향후 통합 시스템 설계, 품질 평가, 법제 정비 등 제도적 확장 과정에서 개념적 준거(conceptual framework)로 활용될 수 있다. 모든 용어는 국가데이터처의 「통계용어사전」을 중심으로, 주요 국제기구(UNSD, ESCAP, UNECE, Word Bank 등) 문헌을 종합적으로 검토하여 정리하였다.

가. 자료, 데이터, 메타데이터

통계생산에서 혼용되는 개념 중 하나가 자료, 데이터 그리고 메타데이터이다. 세 용어는 모두 데이터 통합의 핵심 요소이지만, 각각 담당하는 역할과 수준이 다르다. 정책적·기술적 판단이 필요한 데이터 통합에서는 이를 엄밀히 구분하여야 한다.

자료(data source)는 현실 세계로부터 수집된 관측값(observed value)이나 레코드(record)를 의미한다. 이는 가공되지 않은 1차 정보로, 설문 응답, 행정기록, 센서 관측값 등이 이에 해당한다. 자료는 데이터의 출발점이자, 현실로부터 취득한 원천 정보이다.

데이터(data)는 이러한 자료를 일정한 기준에 따라 정의·분류·정제·구조화한 결과물이다. 자료가 “현상을 담은 원본”이라면 데이터는 “분석이 가능한 형태로 변환된 정보”라 할 수 있다. 변수(variable), 통계 단위(unit) 등으로 구성되어 실제 통계 산출물과 분석에 활용된다. 예컨대, 통합 대상의 전처리 및 정합성 점검은 ‘자료’ 수준에서, 결합, 대체 등은 ‘데이터’ 수준에서 이루어진다.

메타데이터(metadata)는 자료와 데이터를 설명하고 관리하기 위한 정보이다. 즉,

자료나 데이터의 출처, 작성 목적, 변수 정의, 분류체계, 품질 수준, 처리 과정 등을 기술한 일종의 “데이터에 대한 데이터”이다. 메타데이터는 통계생산 전 과정에서 투명성과 재현성을 확보하는 기반이 되면, 자료의 품질과 데이터의 일관성을 유지하기 위한 핵심 관리 도구로 기능한다.

요컨대, 세 개념은 유기적으로 연결되어 있다. 자료는 데이터의 원천이며, 데이터는 통합의 기반이고, 메타데이터는 이 둘을 설명하고 관리하는 틀이다.

나. 단위, 개체, 레코드, 관측값

단위(unit)는 통계나 자료 수집의 최소 대상이자, 표본설계, 가중값 부여, 결과 해석 등 통계생산 전 과정의 기초 기준으로 기능한다. 따라서, 통합 데이터 세트의 정합성을 확보하기 위해 단위의 정의는 명확해야 한다.

- **통계 단위(statistical unit)**는 관측 대상이 되는 개체(entity)를 개념적으로 정의한 것이다. 인구통계에서는 개인(person)이나 가구(household), 경제통계에서는 사업체(business)나 법인(enterprise), 환경통계에서는 지역(region)이나 측정 지점(measurement site)이 이에 해당한다. 통계 단위는 조사 범위를 규정하고, 표본설계와 분석의 기본 틀을 제공한다. 데이터 통합에서는 서로 다른 자료에서 동일 통계 단위를 명확히 식별해야 하며, 단위 정의가 불명확할 경우 동일 개체를 중복으로 집계하거나, 서로 다른 개체를 혼합하는 오류가 발생할 수 있다(Eurostat, 2009).
- **변수 단위(variable unit)**는 각 변수가 측정되는 기준과 범위를 나타낸다. 연령 변수의 단위는 ‘살(years)’, 소득 변수의 단위는 ‘원(KRW)’이며, 이 단위가 통일되지 않으면 통합 과정에서 오류가 발생한다. 예를 들어, 한 자료의 소득이 월(月) 단위로, 다른 자료의 소득이 연(年) 단위로 기록되어 있다면 이를 일관된 단위로 변환하지 않으면 결합 후, 평균이나 분포가 왜곡된다. 변수 단위의 명시와 통일은 데이터 통합에서 변수 불일치를 방지하는 필수 조건이다.

개체(entity)는 통계 단위에 속하는 개별 실체를 의미한다. 예를 들어, 통계 단위가 개인이면 ‘김동은’, ‘최하은’ 각각이 하나의 개체이고, 통계 단위가 지역이면 ‘광주’, ‘대전’이 각각의 개체가 된다. 개체는 고유한 속성(attribute)을 가지며, 이를 통해 동일 개체(same entity)를 서로 다른 자료에서 식별하고 결합할 수 있다. 데이터 통합의 핵심 과제는 바로 이러한 동일 개체를 정확히 식별하는 것이다. 식별이 부정확하면 중복, 오연계, 누락이 발생하여 결과의 신뢰성이 훼손될 수 있다(Statistics New Zealand, 2013).

레코드(record)는 하나의 개체에 대해 수집된 관측값의 집합으로, 특정 개체의 정보를 구조화된 형태로 저장한 단위이다. 데이터 테이블¹⁾에서 한 행(row)이 하나의 레코드를 구성하며, 각 열(column)은 해당 개체의 변수(variable)에 해당한다. 예를 들어,

한 개인의 성별, 연령, 거주지, 직업, 소득 등 여러 변수값이 포함된 한 행이 곧 하나의 레코드이다. 레코드는 통계 분석의 기초단위이자 데이터 통합에서 동일 개체를 식별하고 연결하는 실질적 단위로 기능한다. 하나의 개체가 여러 자료에 중복되어 있을 수 있으므로, 통합 과정에서는 중복 제거와 우선순위 선정을 통해 하나의 대표 레코드로 일관되게 관리해야 한다.

관측값(observed value)은 자료를 구성하는 가장 기초적인 정보 단위이다. 이는 하나의 변수에 대해 수집된 실제 수치 또는 속성값으로, 데이터 테이블에서 행과 열이 교차하는 셀(cell)에 해당한다. 예를 들어, 한 개인의 ‘연령 = 47세’, ‘성별 = 남성’, ‘소득 = 4천만 원’이라면, 각각의 값(47세, 남성, 4천만 원)이 개별 관측값이다.

다. 병합, 연계, 결합, 통합

병합(merging)은 구조가 같거나 호환이 가능한 자료를 수평(행) 또는 수직(열)으로 합치는 과정이다. 행 병합(row-wise merge)은 동일 변수 구성을 가진 여러 시점·지역·조사 주기의 데이터를 결합해 표본 크기를 확장하거나 시계열 데이터를 구축하는 방식이다. 반면, 열 병합(column-wise merge)은 동일 개체를 기준으로 서로 다른 변수의 정보를 합치는 방식으로, 예컨대 동일 응답자의 설문조사 결과와 행정자료의 소득정보를 합치는 경우가 이에 해당한다. 병합은 구조적 일관성을 전제로 하므로, 변수명·변수 정의·코드 체계·형식의 일관성을 사전에 확보해야 한다. 그렇지 않으면 분석의 신뢰성이 저하될 수 있다(UNECE, 2019b; Statistics New Zealand, 2013).

연계(linking/linkage)²⁾는 서로 다른 데이터에서 같거나 유사한 개체를 식별하고, 그

-
- 1) 데이터 테이블은 통계 단위에 기반하여 구성된다. 각 행(row)은 하나의 단위(unit)에 속한 개체의 레코드(record)를 나타내며, 각 열(column)은 변수(variable)에 해당한다. 따라서, 행과 열이 만나는 셀(cell)은 해당 개체의 특정 변수에 대한 관측값(observed value)을 담고 있다. 즉, 데이터 테이블의 구조는 「단위 - 개체 - 레코드(행) - 관측값(셀)」의 순으로 연결된다.
 - 2) **연계(linking/linkage)**와 **결합(matching)**은 학문적·실무적 맥락에서 동일 의미로 사용된다. Winkler (2006)는 연계(linkage)와 결합(matching)을 동의어로 사용하며, “then the record linkage or matching procedures may be intended to identify duplicates.”(p. 1)라고 명시하였다. Christen(2012, p. 1) 역시 “Data matching, also known as record linkage”라고 기술하였고, Harron(2016, p. 4) 또한 연계와 결합을 “a number of synonyms for record linkage depending on the field of application, including ‘record matching’, ‘entity resolution’ and ‘merge-purge’.”로 설명하며 동의어로 간주하였다. 실무에서도 Statistics New Zealand(2013, p. 32)는 “Matching, record linkage, or simply ‘linking’ is the process of comparing records and deciding which ones are links.”로 정의하고 있으며, UNECE(2019a, p. 2) 역시 미시 데이터 통합(micro-data integration)을 “integration of data based on linkage/matching of individual-level records.”로 정의하여 두 용어가 동일 개념임을 명시하고 있다. 종합하자면, 학술 문헌과 실무 지침 모두에서 연계와 결합은 개념적으로 구분되지 않는다. 다만, **본 연구에서는 데이터 통합 과정의 체계적 구분을 위해 결합을 레코드 연계와 통계적 매칭을 포괄하는 상위 개념으로 정의한다.**

개체 간의 대응 관계를 설정하여 하나의 데이터로 연결하는 과정이다. 이 과정에서는 개념 불일치, 변수 분류, 시점 차이 등을 조정하여 정합성(consistency)을 확보하며, 개체 식별을 위한 연계키(linking key) 또는 연계 결과표(linkage table)가 생성된다. 연계는 병합처럼 구조적 동일성을 전제로 하지 않고, 개체 식별을 중심으로 동일 또는 유사한 개체의 속성을 합치는 데 초점을 둔다(Christen, 2012; Harron, 2016).

결합(matching)은 개체 간의 정보를 실제로 하나의 데이터 세트로 합치는 과정이다. **본 연구에서는 결합을 레코드 연계(record linkage)와 통계적 매칭(statistical matching)을 포괄하는 개념으로 접근하였다.** 즉, 결합은 동일 개체를 식별하여 연결하는 레코드 연계의 범위를 넘어, 데이터 간의 불일치·누락·중복을 교정함으로써 분석 가능한 수준의 일관된 데이터 세트를 확보하는 절차이다.

통합(integration)은 데이터 간 형식적·개념적·논리적 일관성을 확보하기 위한 전 과정을 포괄한다. 즉, 통합은 결합 단계를 넘어 전처리, 결합, 결측값 대체, 보정, 품질 점검, 메타데이터 작성 등 데이터 품질을 체계적으로 관리하는 절차를 포함한다. 이러한 일련의 과정을 통해 서로 다른 출처의 자료를 개념적·구조적으로 표준화하고 통합 목적에 맞게 결합함으로써 검증된 품질의 통계 산출물을 생산하는 체계적인 절차가 완성된다.

라. 상호운용성, 정합성, 조화, 정렬

상호운용성(interoperability)은 서로 다른 시스템, 기관 또는 국가 간에 자료가 의미 손실 없이 원활하게 공유·해석될 수 있는 **호환 상태**를 의미한다(ESCAP, 2020). 이는 단순한 형식적 호환을 넘어, 자료가 일관된 논리(정합성), 공유된 개념(개념 조화), 합의된 형식(스키마 정렬) 위에서 운용될 수 있도록 보장하는 것을 말한다. 상호운용성은 「스키마 정렬-개념 조화-정합성」이 순차적 또는 반복·병행적으로 충족될 때 확보된다.

정합성(consistency)은 서로 다른 자료가 동일 기준과 논리하에서 모순 없이 유지되는 상태를 의미한다. 이는 자료 간 변수의 개념·시점·단위·분류체계·범위가 **논리적 일관성**을 확보했음을 의미한다(ESCAP, 2020; Eurostat, 2014).

개념 조화(semantic harmonization)는 변수 개념을 공통 기준에 맞추어 변환·표준화함으로써 **개념적 일관성**을 확보하는 과정이다(ESCAP, 2020). 이는 단순히 변수의 외형을 일치시키는 수준을 넘어, “**동일 현상을 동일 의미로 관측·해석·비교할 수 있도록 만드는 과정**”이다(Eurostat, 2014; Rahm & Bernstein, 2001).

스키마 정렬(schema alignment)은 서로 다른 자료에서 동일 개념일 가능성이 높은 변수를 식별하고, 대응 관계를 설정하는 절차이다(Rahm & Bernstein, 2001). 스키마

정렬의 목적은 개념 조화가 가능하도록 **형식적 일관성**을 확보하는 데 있으며, 이는 상호운용성 달성의 출발점이다(Rahm & Do, 2000; Rahm & Bernstein, 2001).

마. 고해상도 통계

고해상도 통계(high-resolution statistical data)는 국제기구의 문헌을 바탕으로 본 연구에서 새로이 제안하는 개념이다.³⁾ 고해상도 통계는 통계 단위가 시간적·공간적으로 세분된 통계 데이터로 미시적 수준에서 경제·사회적 변화를 정밀하게 파악하고, 정책 의사결정을 지원하는 통계이다. 특히, 소지역·소집단·정책 대상자 단위⁴⁾와 같이

3) 고해상도 통계(high-resolution statistical data)는 국제기구의 문헌을 바탕으로 본 연구에서 제시하는 개념이다. 국제기구들은 데이터의 세분성(granularity), 시의성(timeliness), 정밀성(precision) 향상을 새로운 과제로 제시하고 있다. UNECE(2024)는 통계가 더욱 자세하고 빈번하며 신속하게 생산되어야 함을 강조하였다(“continued demand for more timely, frequent and granular official statistics.” p. 1). 또한, 새로운 자료원과 기술이 전통적인 통계생산 체계를 넘어서는 정밀한 데이터 구축의 가능성을 제시하였다(“The new sources...; but they may allow for obtaining new data and achieving timeliness, frequency and granularity that were not possible before.” p. 12). 이러한 논의는 통계의 공간 단위 세분화(spatial disaggregation)로 확장된다. UN-GGIM Europe(2022)은 통계와 지리정보의 통합을 단위 수준에서 구현해야 한다고 명시하였다(“the Global Statistical Geospatial Framework (GSGF) principles encourage the link between statistical data and geography to be made at the unit record level by storing geometry against the unit record, typically on a point-based framework,...” p. 12). Tiecke 등(2017)은 “The resulting high-resolution population datasets have applications in infrastructure planning, vaccination campaign planning, disaster response efforts and risk analysis such as high accuracy flood risk analysis. (p. 1)”이라 언급하며, 고해상도 데이터가 인프라 계획, 보건 대응, 재난 관리 등 미시 단위 의사결정에 직접 활용될 수 있음을 실증적으로 보여주었다. 또한 Zheng 등(2025)은 “Accurate, up-to-date, and highly resolved measurements of economic well-being are essential for monitoring and achieving international goals of poverty alleviation. Granular estimates of household poverty and wealth are critical for understanding whether these goals are being met, as well as for targeting and evaluating anti-poverty interventions in regions where progress is lagging. (p. 2)”이라 밝히며, 지역 수준의 경제 복지 변동을 포착하기 위해 정밀한 시공간 측정(highly resolved measurement)의 필요성을 강조하였다. 이상을 종합하면, 고해상도 통계의 핵심은 단순한 데이터 정밀도 향상을 넘어, ① 시간적·공간적 세분화, ② 미시 단위 기반, ③ 국지적 변동 탐지, ④ 정책 활용 가능성 확보에 있다.

4) 고해상도 통계 예시

• 소지역 단위

- 국민건강영양조사와 건강보험 청구자료를 통합해 읍·면·동 단위의 만성질환 유병률 선출
- 위치기반 소매결제 데이터와 인구 데이터 통합을 통해 지역별 소비 패턴 분석

• 소집단 단위

- 학교급별 교육 이력과 소득자료를 통합하여 저소득 가구 아동의 학업 성취도 분석
- 장애 등록 정보와 고용보험 데이터를 통합해 장애인 고용 유지율 및 근속기간 분석

• 정책 대상자 단위

- 기초생활보장 수급자 행정정보와 고용패널 자료를 통합하여 복지제도의 취업 유인 효과 분석

세분화된 수준에서 데이터를 통합·활용함으로써 지역별·집단별 변동을 구체적으로 파악할 수 있다.

2. 데이터 통합 대상 자료별 특성

가. 행정자료

행정자료(administrative data source)는 통계생산이 아닌 행정 목적으로 수집된 행정기록 또는 관리기록을 말한다(통계청⁵⁾, n.d.). 「통계법」 제3조에서는 이를 “공공기관이 직무상 작성·취득하여 관리하는 문서·대장 및 도면과 데이터베이스 등 전산 자료로서 통계등록부 및 통계자료를 제외한 것”으로 정의하고 있다(법제처, 2024).

행정자료는 대체로 정형화된 구조를 지니며, 반복적이고 지속적으로 생성·관리된다는 점에서 시계열적 연속성과 포괄성을 지닌다. 이러한 특성 덕분에 행정자료는 통계생산의 효율성과 정확성을 높이는 핵심 자료원으로 활용될 수 있다. 특히 전수 기반이므로 조사자료보다 넓은 범위를 포괄할 수 있으며, 조사 비용과 응답 부담이 발생하지 않는다는 장점이 있다. 다만, 행정자료를 데이터 통합에 활용하기 위해서는 다음의 사항들을 유의해야 한다.

- **개념 불일치 문제.** 행정 목적에 맞춰 정의된 변수나 단위가 통계 목적과 일치하지 않을 수 있다(ESCAP, 2020). 예를 들어, 사업체 통계의 경우 행정자료에서의 사업자 단위가 행정 단위 기준으로 정의되었다면 동일 사업체라도 통계 단위와 불일치할 수 있다. 이러한 불일치는 스키마 정렬, 개념 조화 등을 통해 조정해야 한다.
- **분류체계 차이.** 행정자료의 업종, 직업, 교육 수준 등의 분류체계는 국가데이터처의 표준 분류 체계와 다를 수 있다. 이 경우, 대응표를 작성해 재분류하거나 공통 코드 체계로 재코딩하는 과정을 거쳐야 하며, 이를 통해 정합성을 확보하여야 한다.
- **결측값과 오류.** 조사자료의 결측이 주로 무응답에 기인하는 것과 달리 행정자료의 결측은 행정 시스템상의 누락, 입력 오류 또는 자료 이전 과정에서의 손실 등에 기인할 수 있다. 결측 발생 메커니즘이 체계적(MNAR)인지, 무작위(MCAR, MAR)인지에 따라 처리 방법이 달라지며, 논리적 불일치가 있는 경우 검증 및 편집 절차가 필요하다.

- 출산·양육 관련 서비스 이용 데이터와 출생 통계를 통합해 다자녀 가구의 정책 수요 예측
5) 현 국가데이터처

- **시의성(timeliness).** 행정자료는 수집 시점과 통계 기준 시점이 일치하지 않거나 제출 지연으로 인해 시계열 분석이 어려운 경우가 많다. 따라서, 자료의 최신성을 확보하기 위해 업데이트 주기 관리 및 시점 기준의 일관화가 필요하다.
- **식별자 기반 통합.** 고유식별자가 존재하면 정확 매칭(exact matching)이 가능하지만, 법적·기술적 이유로 사용이 제한되거나 존재하지 않는 경우, 이름, 성별, 생년월일, 주소 등의 공통 변수를 기반으로 결합해야 한다. 이 과정에서 오연계 또는 누락이 발생할 수 있다. 따라서, 연계 품질지표를 통해 오류를 평가하고 보완하는 절차가 필요하다(Enamorado et al., 2019; Harron, 2016).

나. 조사자료

조사자료(sample survey data source)는 통계 목적을 위해 설계된 표본조사를 통해 수집된 자료이다. 조사자료는 표본설계, 확률 추출, 가중값 등 통계학적 기반 위에서 산출되므로, 데이터 통합 과정에서 기준자료(reference data source) 또는 보정자료(calibration source)로 중요한 역할을 한다. 데이터 통합에서 조사자료는 두 가지 방식으로 활용된다.

첫째, **조사자료 간 통합**이다. 유사한 목적과 구조를 가진 다수의 조사자료를 결합하여 장기적 또는 횡단적 비교가 가능한 데이터 세트로 재구성한다. 동일 개체를 대상으로 한 다년간의 패널조사나 기관별로 수행한 유사 주제 조사 간의 결합이 이에 해당한다. 조사자료 간 통합은 표본설계의 일관성을 유지하면서도 공통 변수 기반의 구조적 일관성을 확보하는 것이 핵심이다.

둘째, **행정자료 및 빅데이터와 통합**이다. 조사자료는 행정자료나 빅데이터의 보정(calibration) 또는 결합의 기준점으로 활용될 수 있다. 이는 확률 표본이 가지는 통계적 대표성과 외부 자료의 시의성·규모를 결합함으로써 통합 데이터 세트의 품질을 제고하는 접근으로 다음과 같은 방식으로 활용된다.

- **변수 보완.** 조사자료에 존재하지 않는 변수를 외부 기준자료에서 보완하여 변수 확장
- **모형 기반 추정.** 조사자료를 가중값으로 삼아 외부 자료의 추정 모형에 반영
- **무응답 보정.** 행정자료나 빅데이터를 활용하여 응답자와 무응답자의 특성 차이 조정

여기서 중요한 것은 조사자료의 설계 가중값과 모집단 기준값의 정합성 확보이다. 특히, 비확률 자료의 활용이 늘어남에 따라 조사자료 기반의 보정 또는 사후층화(post-stratification)는 데이터 통합의 필수 절차로 볼 수 있다. 가중값의 불일치나 오차가 발생하면 추정값의 왜곡을 초래할 수 있으므로 설계 기반의 교정 절차가 반드시

병행되어 한다(Chen et al., 2020; D’Orazio et al., 2024).

조사자료는 데이터 통합에서 기준자료 또는 보정자료로 활용할 수 있기에 통합 시 허브 역할을 할 수 있다. 이를 위해서는 조사자료 고유의 정합성과 품질을 유지하면서 자료 간의 구조적·개념적 차이를 조화하는 과정이 필수적이다.

다. 지리정보

지리정보(geospatial data) 또는 **공간정보(spatial data)**는 행정구역, 위치 좌표, 거리, 경계 등 지리적 위치를 나타내는 속성 정보를 포함하는 자료 유형으로, 통계생산과 정책 분석의 정밀도와 실효성을 향상할 수 있는 핵심 자료로 주목받고 있다(Eurostat, 2024). 위치기반 속성이 포함된 모든 정보⁶⁾는 지리정보로 분류될 수 있으며, 위도·경도 좌표, 행정구역 코드, 공간 객체 간 거리 및 연결성 그리고 공간 범위와 관련된 시계열 데이터를 포함한다(Eurostat, 2024; ESCAP, 2020).

지리정보는 ‘어디(where)’라는 공간적 맥락을 부여함으로써 기존 조사·행정자료가 설명하기 어려운 위치기반의 의미를 제공한다. 조사자료나 행정자료가 개체의 속성과 특성을 설명한다면, 지리정보는 그 개체가 존재하거나 사건이 발생한 공간적 배경을 연결함으로써 통계의 해석력과 정책 활용 가능성을 크게 확장한다. 지리정보의 주요 특징은 다음과 같다.

- **공간적 해상도(spatial resolution)**. 행정동, 격자, 도로망 등 공간 단위를 기준으로 다양한 속성 정보와 결합 가능
- **위치기반 분석(location based analysis)**. 공간 패턴 탐색, 근접성 분석, 네트워크 분석 등을 통해 정책 분석의 정밀도 향상

데이터 통합은 시간과 공간을 동시에 고려하는 시공간 통계(geo-statistics)로 발전하고 있다(Eurostat, 2024; ESCAP, 2020). 소지역 통계에 대한 수요 증가, 정책 대상의 위치기반 맞춤 분석 필요성, 환경·보건·복지 정책 간 연계 분석의 확대는 모두 지리정보 기반 통계생산을 촉진하는 요인이다. 특히, UN-GGIM Europe(2022)은 통계와 지리정보의 통합을 단위 레코드(unit record) 수준에서 구현해야 함을 권고하며,

6) 지리정보는 다양한 출처에서 생성되며, 다음 유형으로 분류한다.

- 공공기관의 행정구역 자료: 국가데이터처, 국토지리정보원(NGII), 지방자치단체 등에서 제공하는 행정동, 시군구, 읍면동 등의 공간 경계 정보
- 격자 기반 자료(grid-based data): 공간을 일정 크기의 셀(grid)로 나누어 위치기반 정보를 수집·저장하는 방식(예: 100m×100m 격자 단위 인구통계)
- 센서·위성 기반 자료: GPS, 드론, IoT 센서, 위성사진에서 수집된 위치 및 환경 관련 데이터
- 주소 기반 자료: 도로명주소, 지번 주소, 우편번호 등과 연계된 위치 정보
- 위치 로그 데이터: 모바일, 카드결제, SNS 등에서 생성되는 사용자 위치기반 이력 정보

지리정보가 고해상도 통계의 기반임을 명확히 하고 있다. 지리정보를 데이터 통합에 활용하기 위해서는 다음 사항을 고려해야 한다.

- **공간 단위 불일치.** 자료별 공간 단위(예: 행정구역, 격자 등)가 서로 다를 경우, 공간적 변환(spatial transformation)이 필요하다. 가령, 읍면동 단위 자료와 격자 자료를 결합하려면, 공간 재배열 알고리즘(spatial realignment algorithm)을 적용해야 한다(Comber & Zeng, 2022; Schiavina et al., 2023).⁷⁾
- **위치 좌표 정렬.** 주소, 명칭, 행정코드 등의 위치 속성이 일관되지 않으면 통합이 어렵다. 이를 해결하기 위해 지오코딩(geocoding), 공간 식별자 부여, 코드 정렬 등의 사전 작업이 수행되어야 한다. 특히, 주소 데이터는 도로명·지번 체계 차이, 행정구역 개편에 따른 시계열 불일치를 반드시 고려해야 한다.
- **공간 통합 및 분해.** 서로 다른 공간 단위를 통합하거나 분해할 경우, 왜곡(modifiable areal unit problem)이 발생할 수 있다. 예를 들어, 시군구 단위의 소득 불균형이 읍면동 단위에서는 완전히 다른 양상을 보일 수 있다. 데이터 통합 시 이러한 왜곡을 완화하기 위해 공간 보정(spatial smoothing)이 필요하다(Besag et al., 1991; Gao & Wakefield, 2022).⁸⁾
- **갱신 주기와 시의성.** 위성·센서 기반 지리정보는 주기적 갱신이 가능하나, 행정 경계 자료는 갱신 주기가 불규칙하거나 제도 개편으로 변경될 수 있다. 따라서 시계열 분석의 일관성을 유지하기 위해서는 공간 버전 관리(spatial versioning) 체계가 필요하다(Eurostat & UN-GGIM Europe, 2023).

라. 빅데이터

빅데이터(big data)는 단순히 데이터의 양이 많다는 의미를 넘어 통계적·분석적 가치를 지닌 대규모 자료로서 여러 ‘V’로 대표되는 속성⁹⁾을 가진다(ESCAP, 2020).

7) 재배열 알고리즘은 서로 다른 공간 단위 간 결합을 위해 위치 좌표를 기반으로 변환(interpolate)하거나 재분배(reallocate)하는 방법으로 비례 가중법, 회귀 및 평활화, 회귀 기반 대체 등이 있다.

8) 공간 보정은 인접 지역 간의 공간적 상관 구조를 활용하여 변동성을 완화하여 일관된 지역 추정값을 생성하는 방법이다. 주로 소지역 추정(small area estimation)이나 질병 지도 등에서 국지적 편차를 완화하기 위해 사용된다.

9) Volume(대용량), Velocity(신속성), Variety(다양성), Variability(변동성), Veracity(진실성/신뢰성), Validity(타당성/정합성), Vulnerability(취약성), Volatility(불안정성/휘발성), Visualization(시각화), Value(가치) 등을 포함하며 데이터 기술의 발전에 따라 새로운 속성이 계속 추가되고 있다(ESCAP, 2020). UNECE(2014)는 빅데이터를 다음 세 가지 범주로 분류하였다.

- 인간 기인 3정보: SNS, 영상, 인터넷 검색, 모바일 콘텐츠 등 인간 활동 기반 데이터
- 과정 매개 정보: 의료 기록, 전자상거래 등 거래 및 사건 중심 데이터

대부분의 빅데이터는 민간 영역에서 수집·보유되며, 클라우드 및 플랫폼 기술의 발전으로 그 축적과 활용이 급속히 확대되고 있다. 한편, 정부 및 공공영역에서 수집하는 행정자료 중에서도 ‘V’의 특성이 있으면, 빅데이터로 분류할 수 있다.

빅데이터는 통계생산 과정에서 두 가지 방식으로 활용될 수 있다. ① **통계생산의 원천**으로서 기존 조사자료를 대체하거나 보완하거나 ② **데이터 통합의 재료**로서 행정자료 또는 조사자료와 결합하여 통합 데이터 세트를 보완·보강하는 수단으로 활용된다. 빅데이터를 통계 목적에 적합하게 통합하기 위해서는 다음을 고려해야 한다.

- **정합성과 품질 확보.** 빅데이터는 수집 주기의 불일치, 개념적 상이성, 비표준화된 분류체계, 시계열 단절 등 여러 장애 요소가 내재해 있기에 통합의 실효성을 확보하기 위해서는 스키마 정렬, 개념 조화 등의 사전 준비가 필수적이다.
- **메타데이터 구축.** 빅데이터는 반정형 또는 비정형이 많아 통합을 위해서는 구조화(structuring), 파싱(parsing), 정제(cleaning) 과정이 선행되어야 한다. 이 과정에서의 변환 규칙과 정보 손실을 명확히 기록한 메타데이터를 구축해야 품질 평가의 투명성을 확보할 수 있다.
- **자료 제공자 및 구조의 다양성.** 빅데이터의 활용 가능성과 통합 가능성은 자료 제공자에 따라 달라질 수 있다. 민간 플랫폼이 수집한 로그 데이터나 API 데이터는 그 구조가 비표준화되었거나 주기적 제공이 보장되지 않을 수 있다. 따라서, 데이터 표준화 및 거버넌스 체계 구축, 민간 자료와의 기술적 호환성을 확보해야 한다.
- **자료 유형별 통합 전략.** 정형 빅데이터는 행정·조사자료와 직접 결합이 가능하지만, 비정형 빅데이터는 예측 모형이나 보조변수 형태로 전환해 간접적으로 활용하는 것이 적합하다. 즉, 빅데이터는 통합 목적과 방식에 따라 유형별 역할을 부여해야 하며, 기존 자료와의 관계 속에서 기능적 위치를 정립하는 전략이 필요하다.

• 기계 생성 정보: 모바일 신호, 위성 이미지, 기상·대기·교통 센서, 웹 로그 등의 센서 기반 데이터

<부표 2-1> 행정자료, 조사자료, 지리정보, 빅데이터의 주요 특성 비교

구분	행정자료	조사자료	지리정보	빅데이터
목적	행정 절차 수행	통계생산	공간 단위 분석·관리	비즈니스, 이용기록 등 비통계 목적
구조	정형	정형	정형, 반정형	정형, 반정형, 비정형
대표성	높음(전수 중심)	높음(표본설계 기반)	단위에 따라 다름	사용자 편향 존재
시의성	일반적으로 낮음	통계생산 주기에 의존	주기적으로 갱신	실시간
접근성	법적 제한 존재	통계생산 기관 보유	공공 접근 가능 자료 존재	민간 소유, 제한적 접근
식별자	고유식별자 존재	(부분적) 고유식별자 존재	위치기반 식별자	일반적으로 없거나 약함
통합 가능성	비교적 용이	기초조사일 경우 결합 용이	주소, 격자 등 기준 필요	전처리, 가공, 해석 필요
활용 사례	표본설계, 편향 보정, 모집단 틀	통계생산	소지역 통계, 시공간 분석	소비행태, 여론, 행동 분석
주요 과제	개념 불일치, 오류, 시점 차이	무응답, 표본오차	단위 불일치, 공간 왜곡	식별자 부재, 품질 검증, 민감성

자료 출처 : ESCAP(2020), Eurostat(2024), JRC(2021), UNECE(2019a) 등에서 재구성

Abstract

Establishing a Methodological Framework for Data Integration

Mingyu Kim, Seongryul Park

The acceleration of digital transformation has led to the accumulation of vast data sets across both public and private sectors, while users of statistical data increasingly demand timely, granular, and high-resolution statistical data. However, such expectations cannot be fully met by relying on single data sources alone. This limitation has underscored the growing importance of data integration, which combines multiple data sets to enhance analytical utility and has emerged as a strategic priority for modern national statistical systems. The present study seeks to establish a methodological framework for data integration to enhance the quality, credibility, and reliability of official statistics. The current study is organized around three interrelated objectives. First, it defines the concept of data integration and elucidates its scope and distinctive features. Second, it classifies integration practices across the dimensions of purpose, unit, data type, stakeholder, and methodological approach, thereby consolidating diverse approaches into a coherent analytical taxonomy.

Third, it presents the integration process as a six-stage procedural framework — comprising pre-processing, consistency verification, matching, imputation, calibration, and quality assurance — and demonstrates that effective integration depends on the coherent interconnection of these stages. In this regard, data integration is conceptualized as a comprehensive management framework and a strategic infrastructure for official statistics, transcending the boundaries of a mere technical process. Building on these findings, this study advances five strategic directions: (a) establishing a centralized coordination and management framework; (b) strengthening data governance; (c) expanding integration research on semi-structured and unstructured data; (d) fostering experimental and pilot integration initiatives; and (e) institutionalizing a metadata documentation and management system. By articulating the conceptual foundations, structuring typologies, and systematizing procedural methods, the current study provides a coherent and enduring foundation for improving the reliability, transparency, and practical utility of national statistics. Ultimately, the present study posits that data integration will evolve into a cornerstone of sustainable, adaptive, and high-quality statistical infrastructure in an increasingly data-driven era.

Key words : data integration, methodological framework, high-resolution statistical data, national statistical systems, statistical infrastructure

연구진

- 김민규(국가데이터처 국가데이터연구원 통계방법연구실 연구사)
 - 박성률(국가데이터처 국가데이터연구원 통계방법연구실 연구관)
- * 연구진의 소속 및 직급은 연구과제 완료 시 기준임을 알려드립니다.

연구보고서 2025-08

데이터 통합 방법 체계화 연구

인 쇄 2026년 1월
발 행 2026년 1월
발 행 인 김 진
발 행 처 국가데이터처 국가데이터연구원
35220 대전광역시 서구 한밭대로 713
TEL.(042)366-7100 Fax.(042)366-7123
홈페이지 <https://mods.go.kr/dsri>
ISSN(Online) 2733-4120





국가데이터처
국가데이터연구원

