발 간 등 록 번 호
11-1241140-100001-10

2025년 연구보고서

데이터과학 기술을 활용한 경제구조통계 자료수집 개선 연구

2026. 3.

https://mods.go.kr/dsri

국가데이터처
국가데이터연구원

연구보고서 2025-17

데이터과학 기술을 활용한 경제구조통계 자료수집 개선 연구

박성률 · 김민규 · 곽승우

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 3월

국가데이터연구원장

목 차

요 약

경제구조통계 특성 항목은 현재 조사를 통해 수집되고, 일부 무응답은 대체(imputation) 처리된다. 지금까지 시도하지 않았던 특성 항목 데이터 수집 방식의 다변화는 효율적인 조사 관리, 자료처리 시간 단축, 무응답 처리 등 경제구조통계의 데이터 품질을 획기적으로 높일 수 있다. 특히 데이터과학 기술을 활용한 자료수집은 국가통계 생산 방식의 새로운 패러다임을 제시하는 변곡점이 될 것이다.

본 연구는 데이터과학 기술을 활용한 경제구조통계 특성 항목의 자료수집 개선을 목표로 진행하였다.

주요 연구 내용은 선행연구, 오픈 데이터 현황, 데이터과학 기술을 활용한 자료수집, 데이터 통합을 활용한 특성 항목 보완, 실증분석으로 구성되어 있다.

스크래핑, open API, EXCEL 등 다양한 경로로 수집된 데이터는 경제구조통계 작성에 활용 가능할 것으로 보였지만, 낮은 연계율로 인해 실제 활용에는 제약이 따를 것으로 판단했다. 낮은 연계율의 주된 원인은 데이터 전처리가 이루어지지 않은 데서 비롯되었다. 추가로, 오픈 데이터는 조사 개념과 행정 개념의 차이, 산업분류 부재에 따른 불일치, 불완전한 주소 체계, 입력자의 비표준화된 처리 등의 구조적인 문제로 조사자료와의 연계가 어려운 상황이다. 그럼에도 불구하고 해당 자료는 통계 작성 시 간접 정보 제공을 위한 보조자료로 활용 가능할 것으로 판단된다.

따라서, 정부기관의 행정자료가 통계 작성에 효율적으로 활용되려면, 제도적 기반을 확립하고 자료 생산 시점부터 정합성을 확보하려는 노력이 선행되어야 한다. 이러한 노력의 첫걸음은 행정 서식의 표준화라고 할 수 있다.

주요 용어: 데이터과학, 스크래핑, API, 데이터 통합, 데이터 전처리

제 1 장

서 론

제1절 연구 배경 및 목적

국가데이터처 경제구조통계는 산업 및 지역별 사업체(기업체)의 일반 현황, 규모, 고용 구조 및 경영 상태 등을 파악하는 데 목적이 있다. 경제구조통계의 주요항목은 사업체의 고용 관련 정보와 재무제표, 손익계산서의 영업(사업) 관련 항목(매출액, 영업비용 등)이 핵심이다.

경제구조통계는 조사 환경 변화(조사 불응, 비대면 환경 조성 등)와 행정자료 및 빅데이터 활용 증가로 인해 자료수집 방법의 전환점을 맞이했다. 하지만 현재 이러한 변화는 사업체 일반 현황, 고용, 사업실적 등 주요항목에만 국한된다. 이는 해당 항목들을 대체할 행정자료 등이 제한적이기 때문이다.

주요항목은 입수된 행정자료를 통계 작성에 바로 활용할 수 없다. 이는 행정자료가 각 기관의 목적에 맞게 신고된 자료로서, 통계 작성 기준과는 일치하지 않기 때문이다. 따라서 통계 작성 시 행정자료를 이용할 때는 통계별 기준에 맞게 조정하고 품질 점검을 거친 후 통계생산에 활용해야 한다.

그 외 항목(산업별 특성 항목)은 주요항목 대비 중요도가 상대적으로 낮고, 활용 가능한 행정자료도 없다. 따라서, 특성 항목 자료수집은 조사 현장에서 응답자 인터뷰를 통해 이루어지며, 일부 무응답 항목은 대체(imputation) 처리된다.

국가데이터처의 경제구조통계는 행정자료 활용뿐만 아니라 빅데이터 등 보조 정보의 활용도가 점차 높아짐에 따라, 향후 자료수집 방식이 조사 방식에서 등록 기반으로 전환될 가능성이 충분하다.

따라서, 국가데이터처는 조사 방식에서 등록 기반으로의 자료수집 전환에 대비해야 할 시점이다. 이러한 사전 대비는 통계 작성 기준[1]뿐만 아니라 이용자의 활용 정도를 고려하여 점검할 필요가 있다.

2024년 중장기 1차 연도 연구에서는 특성 항목 자료수집 개선에 관한 선행연구를

1) 경제구조통계와 행정자료의 작성 단위(사업체 및 기업체 단위)와 작성 항목(조사 항목과 행정 및 빅데이터 입수 항목)

수행했다. 해당 연구에서는 경제구조통계의 특성 항목 자료수집에 데이터과학 기술[2] 활용을 제안하고, 일부 특성 항목을 연구 대상으로 선정했다. 연구 대상을 일부 항목으로 제한한 이유는 작성 항목별 오픈 데이터의 가용성에 차이가 있기 때문이다.

본 연구는 1차 연구의 후속 연구로, 경제구조통계의 특성 항목 자료수집 방법을 개선하기 위한 연구이다. 분석 대상은 2025년 경제총조사 시범예행조사 중 2개 지역을 선정했고, 조사표는 업종별 6종이 모두 포함되었다.

현행 특성 항목 자료수집 방법은 현장 조사(약 50%)와 현장 미조사(약 50%)로 구성되며, 현장 미조사는 대체(imputation) 처리된다. 본 연구에서 제안하는 방법은 현장 조사와 현장 미조사를 구분하지 않고, 해당 산업의 모든 사업체를 대상으로 작성 항목별 데이터과학 기술을 활용하여 자료를 수집하는 것이다.

데이터 수집 방법의 변화는 체계적이고 효율적인 조사 관리 및 자료처리 시간 단축을 가능하게 한다. 또한, 현장 미조사 대상의 자료수집도 가능해져 경제구조통계 데이터 품질 향상이 기대된다. 특히, 상시로 입수가 가능한 특성 항목은 등록 데이터베이스(DB)에 탑재할 수 있어 경제구조통계 DB의 활용성[3]을 높일 수 있을 것이다.

제2절 연구 추진 방법

본 연구의 세부 분야는 크게 다섯 가지이며, 각 분야의 과제 내용과 추진 방법은 다음과 같이 정리하였다.

첫째, 선행 연구

현재 경제구조통계의 특성 항목은 조사 현장에서 직접 자료를 수집하고 있다. 자료수집 대상은 모든 사업체(모집단)가 아닌 통계별로 선정된 작성 대상(표본)이므로, 표본으로 선정되지 않은 사업체와는 차이가 있다. 또한, 작성 대상 데이터에 결측값이 발생하면 대체(Imputation) 처리한다.

본 연구는 응답자 부담 경감, 통계 데이터 품질 향상, 결측값 대체 데이터 활용 그리고 등록 기반 경제 데이터베이스(DB)로 전환 대비 항목 확대 등을 위해 데이터과학 기술을 활용한 특성 항목 자료수집 개선 방법을 제안하였다. 이를 위해 특성 항목 자료수집에 활용 가능한 데이터과학 기술을 소개하고, 국내외 선행연구 검토와 적용 사례 등을 제시할 예정이다.

---

2) 스크래핑, 크롤링, open API 등

3) 경제구조통계 작성항목 개선·개발 지원, 경제분석 활용 정보 확대 등

둘째, 특성 항목 오픈 데이터 현황 검토

3장에서는 특성 항목별 데이터과학 기술을 적용할 오픈 데이터 현황을 살펴본다. 먼저, 분석 대상 항목은 경제구조통계(경제총조사 시범예행조사 6종 조사표) 특성 항목의 개념과 작성 기준을 검토하여 선정한다. 다음으로, 선정된 항목을 기준으로 활용 가능한 오픈 데이터(특정 웹사이트, 공공데이터포털 등) 현황을 조사하고 자료수집을 진행한다.

셋째, 데이터과학 기술을 활용한 자료수집

4장에서는 대상 항목과 오픈 데이터가 확정되면 데이터과학 기술을 활용하여 실제 자료를 수집한다. 오픈 데이터의 형태(정형, 반정형, 비정형)와 정보량에 따라 활용 및 연계 방법에 차이가 발생한다. 각 특성 항목의 오픈 데이터별 가장 적절한 활용 및 연계 방법을 검토하고, 자료수집 시 유의 사항도 함께 살펴본다.

넷째, 데이터 통합을 활용한 특성 항목 보완

5장에서는 데이터과학 기술로 수집된 자료의 데이터 통합 및 품질 점검을 진행한다. 이 과정은 '데이터 통합 방법 정립 연구(김민규와 박성률, 2025)'에서 제안한 데이터 통합 실무적 절차(6단계)를 기준으로 점검할 예정이다.

다섯째, 실증분석

경제총조사 시범예행조사 결과와 데이터과학 기술을 활용한 자료수집 결과를 비교한다. 경제총조사 시범예행조사는 현장 조사와 현장 미조사(대체 처리)로 구분되어 있으므로, 실제 조사 결과와 비교하고 대체 처리된 데이터와도 비교하여 제안 방법의 활용성을 검토한다.

제 2 장

선행연구

제1절 데이터과학 기술 소개

자료수집에 활용할 수 있는 데이터과학 기술은 사물 인터넷(IoT)을 통한 센서 데이터 수집, 자동화된 설문조사, 웹 스크래핑, 웹 크롤링 그리고 Open API(Open Application Programming Interface)를 이용하는 방법 등이 있다.

데이터과학 기술을 통해 자료를 수집할 때는 개인의 프라이버시를 침해하지 않도록 유의해야 한다. 또한, 수집된 데이터가 대표성을 갖지 못하고 편향될 가능성도 고려해야 한다. 본 연구에서는 주로 웹 스크래핑과 Open API를 활용할 예정이므로, 이에 대한 세부적인 내용을 좀 더 자세히 검토할 필요가 있다.

1. 웹 스크래핑(web scraping)

웹 스크래핑은 웹사이트의 특정 정보를 자동으로 추출하고 수집하는 컴퓨터 프로그램 기술을 의미한다. 이는 웹사이트의 HTML 코드를 분석하여 필요한 정보만 선별적으로 수집하고 가공하는 방법을 말한다.

예를 들어, 인터넷 쇼핑몰에서 특정 제품의 가격, 재고, 상품명, 상품 설명 등의 정보를 수집하는 방법이 바로 웹 스크래핑이다. 흔히 함께 사용되는 용어로 웹 크롤링(web crawling)이 있는데, 둘의 가장 큰 차이는 다음과 같다.

· 웹 스크래핑 : 원하는 정보의 위치를 정확히 파악해서 필요한 정보만 추출한다.

· 웹 크롤링 : 웹상의 정보를 광범위하게 수집한 후, 색인(INDEX)으로 분류하여 데이터베이스를 구축한다.

실제로는 이 두 방법을 명확하게 구분하기보다 혼용하여 활용하는 편이다.

2. 웹 크롤링(web crawling)

웹 크롤링은 크롤러(또는 봇)라는 자동화된 웹 문서 탐색 프로그램을 활용하여 웹 사이트의 링크를 따라가며 페이지 데이터를 다운로드하는 방식을 사용한다. 이 방식은 웹상의 데이터를 무작위로 수집한 후, 색인(INDEX)을 작성하여 데이터베이스(DB)화하여 저장한다. 대표적인 예로, 검색 포털사이트가 정보를 수집하는 방법을 들 수 있다.

3. Open API(Open Application Programming Interface)

API는 두 소프트웨어 간의 '접점' 역할을 한다. 이를 통해 다른 소프트웨어에서 제공하는 기능을 결합하여 새로운 서비스에 활용할 수 있게 해준다.

대표적인 예로 지도 서비스를 들 수 있다. 구글, 카카오, 네이버 등이 제공하는 지도 서비스의 API를 활용하면, 장소를 소개하는 블로그 콘텐츠 중간에 해당 장소의 정확한 위치 정보를 연동하여 제공할 수 있다.

Open API는 누구나 사용할 수 있도록 대중에게 공개된 API를 의미한다. 대표적인 예는 공공데이터포털이다. 이곳에서는 사용자가 서비스키와 인증 절차를 거쳐 다양한 분야의 공공 정보를 제공받을 수 있다.

제2절 국내외 선행연구

데이터 수집에 활용 가능한 데이터과학 기술로는 센서 데이터 수집, 위성 정보 수집, 웹 스크래핑 등이 있다. 실시간으로 수집되는 이 자료들은 기존 설문을 통해 수집된 자료와 자료 형태 및 포함된 정보의 양이 다르다. 따라서 기존 정보와 결합하기 위해서는 적절한 가공 방법을 고민해야 한다.

데이터과학 기술을 이용한 자료수집 방법은 설문을 통한 정보수집보다 짧은 시간에 자료를 수집 및 처리할 수 있다는 장점을 가지고 있다. 하지만 자료수집 과정에서 법적인 제약이 발생할 수 있으며, 출처에 따라 정보의 질이 일정하지 않다는 단점도 존재한다.

그럼에도 불구하고 여러 국가기관에서는 이러한 자료수집 및 처리 과정에 데이터과학 기술을 결합하여 단점을 보완하고, 정확하고 의미 있는 통계를 생산하기 위한 연구를 활발히 진행 중에 있다.

국가통계에 활용하기 위한 선행연구에서는 설문조사 외의 온라인 자료 수집 방안에 대한 연구가 진행되었다. 특히, API와 웹 스크래핑 활용 결과를 UNECE(United Nations Economic Commission for Europe, 유럽경제위원회)를 통해 공유하고 가이드라인을 제시하는 유럽의 사례가 있다.

선행연구들은 주로 수집된 자료를 활용하여 향후 수집될 자료의 품질을 높이는 모형 개발에 관심을 두고 있었다. 그 외에도 학술적인 연구 목적으로 웹 스크래핑을 통해 수집한 텍스트 데이터의 분석 및 활용 연구, 또는 웹 스크래핑이나 웹 크롤링 시 발생할 수 있는 저작권 문제 연구 사례가 있었다.

제3절 사례 소개

국가통계 작성 기관과 기업에서 이러한 정보를 실제로 어떻게 활용하고 있는지 살펴보면 다음과 같다.

1. 국가통계

가. 미국

미국에서는 미국 노동 통계국(BLS)이 대체 데이터(alternative data)와 빅데이터(big data)에서 수집한 정보를 소비자 물가지수(CPI) 계산에 포함해 왔다. 여기에서 사용된 대체 데이터는 웹 스크래핑과 API를 포함한 다양한 출처의 자료를 의미한다.

UNECE에서 2023년에 발표한 자료에 따르면, 온라인 샵 자료를 스크래핑하여 소비자 물가지수를 개선하는 데 사용한 사례가 있다. 또한, 이렇게 수집한 자료를 활용할 수 있도록 API를 제공하고 있다.

나. 유럽

영국과 유럽 연합의 국가들은 조화 물가지수(HICP) 계산 과정에 웹 스크래핑 데이터를 활용하는 연구를 진행하며 다양한 활용법을 제시하고 있다.

특히, 유럽에서는 목적별 개별소비지출분류(ECOICOP) 기준에 맞춰 조화 물가지수를 계산하기 위해, 스크래핑된 데이터를 자동으로 분류하는 모형을 학습시키는 연구가 주를 이루고 있다(영국, 노르웨이, 네덜란드, 폴란드 등).

이러한 연구들은 상품 상세 정보를 수집한 후 자연어 처리(NLP)를 통해 분류에 필

요한 주요 변수들을 생성하고, 이를 활용한 지도 학습 또는 준지도 학습으로 품목을 자동 분류하는 모형 개발에 집중하고 있다. 대상 품목은 주로 온라인 의류 판매 정보를 시작으로 유가, 항공요금 등으로 확대되는 추세다. 이 과정에서 스크래핑 시 유의사항과 개발 접근 방식에 대한 가이드라인도 함께 제시하고 있다.

한편, 그리스는 온라인 구직 포털 정보를 수집하여 고용 시장의 변화를 자동으로 파악하는 연구 결과를 발표하기도 했다.

다. 그 밖의 지역

캐나다는 2021년 소비자 물가지수 산정 시 컴퓨터와 주변 기기 품목에 대해 웹 스크래핑 자료를 활용하여 매달 가격 변동을 측정했으며, 주변 품목으로 확대를 추진하고 있다.

일본은 소비자 물가지수에 포함되는 항공, 호텔, 해외여행 상품 등의 가격을 스크래핑으로 수집하고 있다. 인터넷 예약 가격을 취합하여 기존보다 더 짧은 기간 동안의 가격 변화를 확인하는 데 도움을 받고 있다.

브라질은 유럽에서 적용한 분류 모형 기법을 전자제품 및 주방 용품 분류에 활용하여 연구를 진행하였다.

이러한 사례들은 여러 나라에서 이미 다양한 출처의 자료들을 활용하여 국가통계를 생산하고자 노력하고 있음을 보여준다. 각국은 이에 따른 문제점과 개선점을 연구하고 있으며, 일부에서는 스크래핑 정보의 신뢰성에 대한 연구도 병행하고 있다. 국내에서도 스크래핑을 통한 정보 활용이 이루어지고 있다.

2. 기업사례

기업들은 데이터 분석, 시장 조사, 가격 비교, 콘텐츠 수집 및 관리 등을 위해 데이터를 수집하며, 이와 관련된 여러 사례가 알려져 있다.

우선 웹 크롤링과 관련하여, 링크드인(세계 최대의 비즈니스와 고용 중심의 소셜 미디어 플랫폼)의 자료를 크롤링하여 잠재 고객을 발굴하는 사례(ClayAI), 또는 인공지능 학습에 필요한 데이터를 크롤링하는 사례(OpenAI)가 있었다.

다음으로 웹 스크래핑과 관련하여, 도메인별 챗봇을 구성하기 위해 웹 스크래핑 기술을 사용하는 기업(Alibaba Cloud)도 있었다.

이처럼 다양한 분야에서 많은 기업이 해당 정보를 활발하게 활용하고 있다.

제 3 장

오픈 데이터 현황

제1절 경제구조통계 특성 항목 소개

경제구조통계 작성 항목은 크게 사업체 일반사항, 고용, 사업실적, 특성 항목으로 구분된다. 여기서 일반사항, 고용, 사업실적(일부 차이)은 사업체의 업종 및 규모와 상관없이 공통으로 작성하고, 특성 항목은 업종에 따라 작성 내용에 차이가 있다.

<표 3-1>은 경제총조사 시범예행조사의 조사표별(6종) 특성 항목의 현황을 나타냈다.

<표 3-1> 경제총조사 시범예행조사 조사표 및 조사항목 현황

조사항목 \ 조사표	1								2		3		4	5	6								
조사항목 \ 산업	A	D	F	H	K	O	P	S	B	C	B	C	G	I	E	J	L	M	N	P	Q	R	S
연간 영업 개월 수	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
월간 정기 휴무일 수	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
인공지능 활용 여부	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
스마트농장 운영 여부	○																						
온라인 거래 여부	○													○	○		○		○	○		○	○
매출형태별 수입액			○																				
로봇 활용 여부				○						○		○		○									
보조금							○	○											○	○	○	○	○
스마트공장 운영 여부										○		○											
사업체 건물(사업장) 연면적												○	○	○									
일일 평균 영업 시간													○	○	○	○	○	○	○	○	○	○	○
디지털 플랫폼 거래 여부													○	○		○						○	
상품매입처별 구입액 구성비													○										
상품판매처별 매출액 구성비													○										
상품판매 유형별 매출액 구성비													○										
배달(택배 포함) 판매 여부													○										
무인매장 운영 여부													○	○								○	○
무인 결제 기기 활용 여부													○	○								○	○
객실 수														○									
월평균 객실 이용 건수														○									
편의시설 개수														○									
매출 형태별 수입액														○									
객석 수														○									
판매 유형별 매출액 구성비														○									
직능별 종사자 수																○							
이용인원(고객) 수																						○	

※ 조사표(2) : 종사자수 9인 이하, 조사표(3) : 종사자 수 10인 이상
※ 조사표(1) S : 산업중분류 94, 조사표(6) S : 산업중분류 95, 96

경제총조사 시범예행조사의 6종 조사표에 포함된 특성 항목은 총 26개로 설계되었다. 이 중에서 3개[4]는 업종에 상관없이 모든 산업에서 공통으로 조사하는 특성 항목이고, 그 외 23개 특성 항목은 전 주기('20년) 경제총조사의 특성 항목과 대부분 동일하게 설계되었다.

<표 3-2>는 '20년 기준 경제총조사 대비 '25년 기준 경제총조사 시범예행조사의 특성 항목 변경 사항을 보여준다. 주요 변경 내용은 조사항목 명칭 변경 1건, 신규 조사항목 추가 4건, 기존 조사항목 삭제 3건이다.

신규 조사항목 추가는 산업별 신기술 적용 현황을 파악하는 항목들로 구성되었다. 구체적으로 살펴보면, '인공지능(AI) 활용 여부'는 전(全) 산업의 공통 특성 항목이고, '스마트농장 운영 여부'는 산업대분류 A 작성, '스마트공장 운영 여부'는 산업대분류 C 작성, '로봇 활용 여부'는 산업대분류 C, H, I에서 작성한다.

조사 항목 변경 사항으로, '온라인 쇼핑' 항목의 명칭이 '온라인 거래'로 변경되었고, 조사표를 작성하는 대상 산업은 산업대분류(E, G, I, L, N, P, R, S)에서 산업대분류(A, E, I, L, N, P, R, S)로 변경되었다.

<표 3-2> '20년 기준 경제총조사와 '25년 기준 경제총조사 특성 항목 변동 사항

구 분	'20년 기준	'25년 기준	비고(산업)
(명칭 변경)	온라인 쇼핑	온라인 거래	A, E, I, L, N, P, R, S
(항목 추가)	-	인공지능(AI) 활용 여부	전 산업
	-	스마트농장 운영 여부	A
	-	스마트공장 운영 여부	C
	-	로봇 활용 여부	C, H, I
(항목 삭제)	부지 면적	-	C
	연구 기술직 종사자	-	J, M
	연간 생산량	-	D

본 연구의 분석 조사항목 선정은 1차 연도 연구에서 제시한 원칙을 동일하게 적용했다. 첫째, 경제총조사 시범예행조사 항목 중 상대적 민감도가 낮은 항목을 우선 고려했다. 둘째, 웹사이트 등에 공개된 데이터소스(정형, 비정형, API, 추가 데이터베이스 등)를 활용할 수 있는 항목을 선정했다. 셋째, 정책 수립의 기초가 되는 최신 트랜드를 반영한 항목(신기술 활용에 관한 추가 항목)을 대상으로 선정했다.

다음은 본 연구에서 자료수집 대상으로 선정한 조사항목의 개념과 작성 방법 등에 대한 설명이다. 이 정보는 데이터과학 기술 등을 활용하여 자료를 수집할 때 중요한 자료로 활용될 것이다.

4) 연간 영업 개월 수, 월간 정기 휴무일 수, 인공지능 활용 여부

<그림 3-1>은 영업기간 2개 항목에 대한 조사항목 설명이다.

먼저 ‘연간 영업 개월 수’는 사업체(기업체)가 작성기준연도(t년) 1년간 정상적으로 영업한 개월 수를 말하며 계절적 요인, 노동쟁의 및 기타 요인으로 휴업한 기간은 제외한다. 광업제조업 분야에서는 ‘연간 조업 개월 수’로 표현하고 있다.

다음 ‘월간 정기 휴무일 수’는 사업체가 작성기준연도(t년)에 매월 정기적으로 휴무한 날을 말하며, 국경일, 명절, 창립기념일 등 부정기적인 휴무일은 제외한다. 정기적인 휴무일이 없이 연중 계속 영업(전자상거래사업체 등)하는 경우 ‘휴무일 없음’으로 기입한다.

이 두 항목은 전(全) 산업 대상 작성하는 항목이며, 전(前) 주기 경제총조사 및 연간조사에서도 지속적으로 작성되고 있는 항목이다. 이 항목들의 정보를 분석에 활용할 뿐만 아니라 그 외 종사자(임시근로자 등) 수와 사업실적 보정에도 활용된다.

※ 출처 : 2025 경제총조사 시범예행조사 조사표

<그림 3-1> 영업기간(연간 영업 개월 수, 월간 정기 휴무일 수) 조사항목

<그림 3-2>는 일일 평균 영업시간 항목에 대한 조사항목 설명이다.

주로 서비스업 작성 항목으로 산업대분류 11개(G, I, E, J, L, M, N, P, Q, R, S(95-96))에서 자료를 수집하고 있다. 사업체의 정상적인 영업일을 기준으로 작성하며, 계절적인 요인 등으로 영업시간이 일정하지 않은 경우 평균 영업시간을 작성한다. 또한, 사업체 폐점 이후에도 거래, 서비스가 이루어지는 경우 영업시간에 포함한다.

※ 출처 : 2025 경제총조사 시범예행조사 조사표

<그림 3-2> 일일 평균 영업시간 조사항목

<그림 3-3>은 배달(택배 포함) 판매 여부 항목에 대한 조사항목 설명이다.

배달(택배 포함) 판매 여부 항목은 산업대분류 G(도소매업)인 사업체만 조사하며, 이 항목은 사업체가 판매하는 재화를 소비자에게 배달(택배) 서비스로 제공했는지 여부를 조사한다. 구체적으로 소비자가 매장에서 구매 결제하고, 사업체에서 배달을 통해 소비자에게 전달하거나, 소비자가 온라인에서 구입하고, 사업체에서 배달(택배)을 통해 소비자에게 전달하는 과정을 말한다.

반면 편의점 등에서 이용자의 물품을 이용자가 원하는 장소로 배달하거나 배달(택배)이 최종소비자가 아닌 중간 유통 지점으로 전달되고, 이후 소비자가 방문 수령하는 경우 배달(택배 포함) 범위에 포함되지 않는다.

※ 출처 : 2025 경제총조사 시범예행조사 조사표

<그림 3-3> 배달(택배 포함) 판매 여부 조사항목

<그림 3-4>는 객석 수 항목에 대한 조사항목 설명이다.

객석 수 항목은 산업분류 I(56, 음식·주점업)인 사업체만 조사하며, 사업체가 산업활동을 위해 마련한 객석 여부를 파악한다. 만약 사업체 방 안에 좌석은 없고 식탁만 설치된 경우 수용 가능 좌석 수를 객석 수에 포함하여 조사한다. 만약 객석을 다른 사업체와 공유하여 사용할 경우 공유하는 좌석 수를 사업체 수로 나누어 기입한다.

※ 출처 : 2025 경제총조사 시범예행조사 조사표

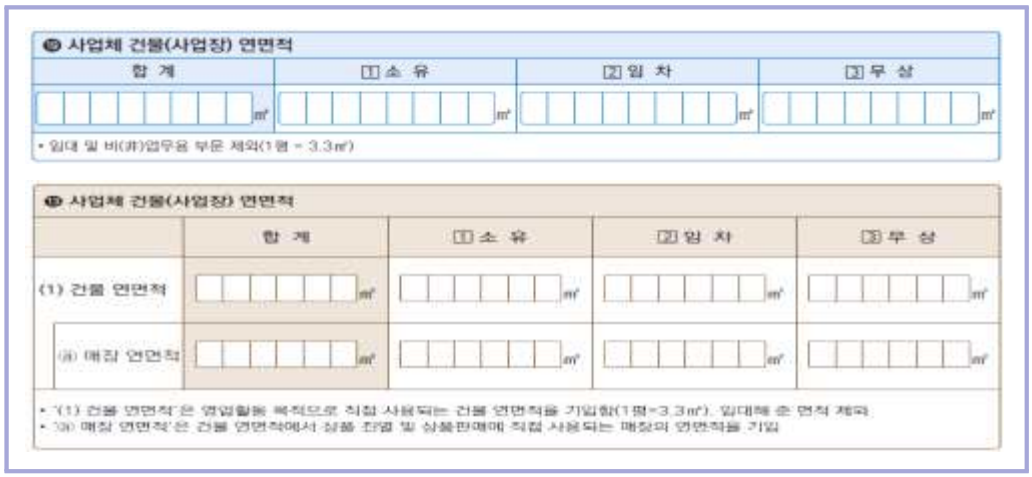
<그림 3-4> 객석 수 조사항목

<그림 3-5>는 사업체 건물(사업장) 연면적 항목에 대한 조사항목 설명이다.

사업체 건물 연면적 항목은 산업대분류 C(제조업), G(도소매업), I(숙박 및 음식점

업)인 사업체만 조사하며, 사업체가 영업 활동을 위해 직접 사용하고 있는 사업용 건물의 연면적을 파악한다. 특히 산업대분류 G(도소매업)는 사업장 연면적 내 상품 판매 목적으로 직접 사용하는 매장의 연면적을 추가로 파악한다.

사업체 건물 연면적에는 건물의 계단, 승강기, 연결통로, 사업을 위해 임차한 면적, 회의실, 사무실, 소비자 편의시설 등이 포함되고, 자기 소유 사업장을 임대한 면적, 주거용 면적, 비닐하우스 등 임시적 구축물 등은 포함되지 않는다. 매장 연면적에는 상품 판매와 직접 관련 없는 종업원 이용시설, 공용면적, 창고 등은 포함하지 않는다. 사업체 건물 연면적은 산업대분류 C, G, I 특성상 사업체의 매출액에 큰 영향을 미치기 때문에 해당 항목을 조사하는 것으로 보인다.



※ 출처 : 2025 경제총조사 시범예행조사 조사표

<그림 3-5> 사업체 건물(사업장) 연면적 조사항목

<그림 3-6>은 온라인 거래 여부 항목에 대한 조사항목 설명이다.

온라인 거래 항목은 산업대분류 8개(A, I, E, L, N, P, R, S)인 사업체만 조사하며, 사업체(기업체)의 온라인 거래 여부와 온라인 및 모바일 매출액을 작성한다. 온라인 거래는 주문이 통신망에서 이루어지지 않으면 온라인 거래로 볼 수 없음에 유의해야 한다.

※ 출처 : 2025 경제총조사 시범예행조사 조사표

<그림 3-6> 온라인 거래 여부

본 연구에서 해당 항목의 자료수집 개선 범위는 ‘온라인 거래 여부’까지로 제한한다. 그 이유는 온라인 및 모바일 매출 정보는 사업체 영업비밀 정보(민감정보)이고, 또한 웹 페이지에서 쉽게 파악할 수 없기 때문이다.

<그림 3-7>은 스마트공장 운영 여부 항목에 대한 조사항목 설명이다.

스마트공장은 제품의 기획 및 개발부터 양산, 주문, 완제품 출하(판매)까지 모든 생산과정에 ICT(정보통신) 기술을 통합한 공장을 말하며, 여기에는 응용 시스템뿐 아니라 현장 자동화와 제어자동화 영역까지 공장 운영의 모든 부분이 포함된다.

스마트공장의 5대 요건은, 4M+1E(Main, Machinery, Material, Method, Environment)들이 실시간으로 디지털화, 지능화(알고리즘 또는 인공지능 등 솔루션을 이용하여 최적해 또는 예측 가능한 해를 제공), 통합(수평적, 수직적), 엔지니어링 지식의 창출, 스마트 시스템과의 연결이 가능해야 한다.

※ 출처 : 2025 경제총조사 시범예행조사 조사표

<그림 3-7> 스마트공장 운영 여부

또한, <표 3-3>의 스마트공장의 구성요소 중 하나라도 도입했다면 스마트공장에 해당된다.

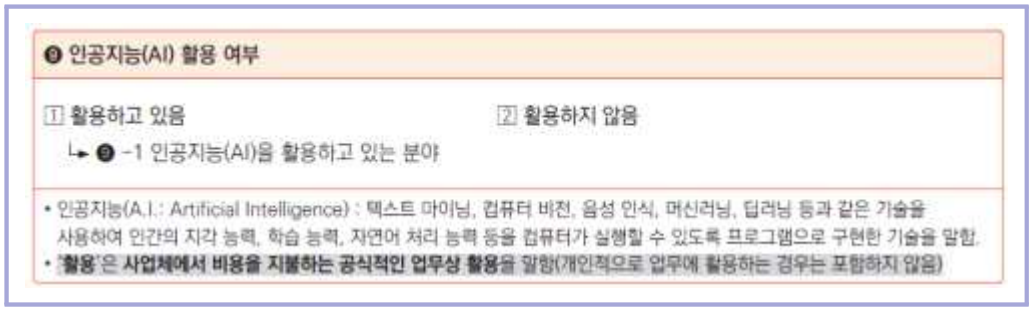
<표 3-3> 스마트공장 구성요소

구성요소	설명
ERP	경영활동 및 공장운영 데이터를 통합/관리하는 전사적 자원관리 시스템 ※ 단, 단순 회계용 경영전용 ERP는 제외
PLM	제품개발부터 폐기에 이르기까지 제품생산 전 과정의 데이터를 관리하는 시스템
SCM	제조업의 전체 공급망을 전산화하여 효율적으로 처리할 수 있는 관리 시스템
FEMS	제조공장의 에너지 이용 효율을 개선하는 에너지관리시스템
CPS/ 디지털트윈	물리 시스템과 이를 제어하는 컴퓨팅 요소가 결합된 차세대 시뮬레이션 기반 생산·운영 시스템

구성요소	설명
MES	제조 데이터를 통합하여 관리하는 시스템으로 공장운영 및 통제, 품질관리, 창고관리, 설비관리, 금형관리 등 제조현장에서 필요로 하는 다양한 기능 지원
APS	ERP와 MES 두 시스템 간 중간에 위치하여 수요계획, 생산계획, 및 스케줄을 관리하는 시스템
빅데이터/AI	입고, 생산, 재고, 납기 등 제조현장의 모든 데이터를 수집하고 분석하여 의사결정을 도와주는 데이터 관리 시스템
제조로봇	각 산업 제조현장 내 제품생산에서 출하까지 공정 내 작업을 수행하기 위한 로봇으로 자동제어 되고, 재 프로그램이 가능하고 다목적인 3축 또는 그 이상의 축을 가진 자동 조정장치
협동로봇	사람과 같은 공간에서 작업하면서 사람과 물리적으로 상호작용 할 수 있는 로봇
자율이송로봇	공장 내에서 물품의 분류, 적재, 포장, 이송 등을 수행하는 물류용 로봇으로, 로봇의 주행을 돕는 마커, 자석 등이 불필요하다는 점에서 기존 이송로봇과 차별화됨
스마트센서/머신버전	소자부품과 각종 센서, 통신기술과 영상처리 기술 등을 활용해 제조공장의 각종 데이터 측정하는 장치
IoT 기기·장비	각종 기기와 장비에 IoT 센서를 탑재하여 데이터의 수집(센싱, 전달)과 가시화(모니터링), 설비제어 등을 지원

<그림 3-8>는 인공지능(AI) 활용 여부 항목 항목에 대한 조사항목 설명이다.

이 항목은 2025 경제총조사 시범예행조사에서 신규로 작성하는 항목으로 사업체의 산업활동에 인공지능(AI)을 활용하는지 여부를 파악한다. 인공지능(AI)을 활용하는 경우 <그림 3-9>와 같이 분야별 해당되는 항목에 체크한다.



※ 출처 : 2025 경제총조사 시범예행조사 조사표

<그림 3-8> 인공지능(AI) 활용 여부 조사항목

‘영업기간’과 ‘인공지능(AI)’ 항목은 모든 산업에서 작성하는 특성 항목으로 본 연구에서 자료수집 개선에 포함된 항목이지만, 사업체별 현장조사 및 자료수집 개선에 많은 어려움이 예상된다.

❾-1 인공지능(AI)을 활용하고 있는 분야 (중복 선택 가능)

(1) 분야		활용 여부	설 명
기획	① 시장 조사·예측	□	소비자 수요, 경쟁사 동향, 매출 흐름 등을 분석하여 제품 전략이나 판매 예측 등에 활용
	② 제품·서비스 개발	□	신제품 기획이나 서비스 고도화를 위해 소비자 의견, 데이터 분석을 통한 아이디어 도출 및 제품 테스트에 활용
생산	③ 생산	□	생산 공정 최적화, 생산 라인의 불량 감지 등 자동화를 통해 효율을 높이는 데 활용
	④ 품질·판매·물류 관리	□	품질 개선, 수요 기반 공급계획, 재고 예측, 배송 경로 최적화 등 판매 및 물류 효율화를 위해 활용
마케팅	⑤ 고객 지원	□	고객 문의 자동 응답(챗봇, 보이스봇 등), 불만 사항 분류 등 고객 상담 및 서비스 대응을 자동화하는 데 활용
	⑥ 홍보	□	홍보 콘텐츠 생성, 광고 효과 분석 등 고객 맞춤형 홍보 활동에 활용
조직 관리	⑦ 안전·보안 관리	□	작업장 사고 예방, 출입 통제, 보안 감시 영상 분석 등 직원 및 시설의 안전 확보와 데이터 보안에 활용
	⑧ 경영 지원	□	회계, 세무, 인사 등 내부 관리 업무의 자동화로 활용
	⑨ 직원 교육·훈련	□	직원 맞춤형 교육 콘텐츠 제공, 역량 평가 자동화 등 사내 교육에 활용
⑩ 기타(_____)		□	

※ 출처 : 2025 경제총조사 시범예행조사 조사표

<그림 3-9> 인공지능(AI)을 활용하고 있는 분야

<그림 3-10>은 로봇 활용 여부 항목에 대한 조사항목 설명이다.

이 항목은 2025년 경제총조사 시험예행조사 신규 항목으로 사업체의 산업활동에 로봇을 활용하고 있는지를 파악한다. 로봇 활용 여부를 파악하는 산업대분류는 3개(제조업, 운수업, 숙박 및 음식점업)이고, 제조업에서는 제품 생산에서 출하까지 공정 내 작업을 수행하는 로봇을 말하고, 운수업·숙박 및 음식점업에서는 불특정 다수를 위한 서비스 제공 및 전문화된 작업을 수행하는 로봇을 말한다. 예를 들면, 서빙 로봇, 화재감시 로봇, 전문 요리용 로봇, 시설 청소용 로봇 등이 포함된다.

❿ 로봇 활용 여부

① 활용하고 있음 → ① 로봇 개수 [ ][ ][ ] 개

② 활용하지 않음

• 전문서비스용 로봇 : 불특정 다수를 위한 서비스 제공 및 전문화된 작업을 수행하는 로봇
(예) 청소 자동화 로봇, 자율 이동 로봇 등)

※ 출처 : 2025 경제총조사 시범예행조사 조사표

<그림 3-10> 로봇 활용 여부

<그림 3-11>은 무인매장 여부 항목에 대한 조사항목 설명이다.

무인매장 여부 항목은 산업대분류 G(47 소매업), I(숙박 및 음식점업), R(예술 스포츠 여가업), S(95, 96 개인서비스업)인 사업체만 조사하며, 이 항목은 항목명 그대로 사업체의 무인매장 운영 여부를 조사한다.

무인매장이란 고객이 종사자 없이 또는 최소한의 도움으로 상품 또는 서비스를 선택하고, 스스로 대금을 지불하는 시스템(무인결제 시스템)을 갖추어 상품 구매 또는 서비스를 받을 수 있는 물리적 공간이 있는 사업체를 말한다. 특히, 영업시간 중 일부만 무인으로 운영하는 경우를 포함하여 조사할 수 있다.

산업별 대표적인 무인매장은, 산업대분류 G는 아이스크림·편의점·밀키트 등, 산업대분류 I는 카페·모텔·라면가게 등, 산업대분류 R은 독서실·인형뽑기점·노래방 등, 산업대분류 S는 빨래방·세차장 등이 포함된다.

※ 출처 : 2025 경제총조사 시범예행조사 조사표

<그림 3-11> 무인매장 운영 여부

<그림 3-12>는 무인 결제 기기 활용 항목에 대한 조사항목 설명이다.

무인 결제 기기 활용 항목은 산업대분류 G(47 소매업), I(숙박 및 음식점업), R(예술 스포츠 여가업), S(95, 96 개인서비스업)인 사업체만 조사하며, 이 항목은 사업체에서 활용하고 있는 무인 결제 기기 활용 여부와 기기 대수를 조사한다.

무인 결제 기기는 키오스크(무인 정보 단말기) 중 하나로 주문, 자동 정산 등 결제 서비스를 제공하는 무인 단말기를 말하며, 카드로 결제 가능한 자동판매기기도 포함된다. 다만, 주문만 가능하고 결제 기능이 없는 키오스크는 제외한다.

※ 출처 : 2025 경제총조사 시범예행조사 조사표

<그림 3-12> 무인 결제 기기 활용 운영 여부

제2절 오픈 데이터 현황

1. 네이버지도/스크래핑/항목(5개)

지도는 지구 표면의 상태를 일정한 비율로 줄여, 이를 약속된 기호로 평면에 나타낸 그림이라고 정의한다(출처: 국립국어원 표준국어대사전). 또한, 지도는 측량 결과에 따라 공간상의 위치와 지형 및 지명 등 여러 공간정보를 일정한 축척에 따라 기호나 문자 등으로 표시한 것으로 정의한다(출처: 국토교통부 국토지리정보원).

지도의 이용 방식은 시대별 제작 목적, 기술 수준, 활용 범위 등에 따라 크게 변화해 왔다. 고대에는 사냥이나 채집 등 생존과 이동을 위한 단순한 정보 전달 수단에 불과했지만, 중세와 근대를 거치며 점차 정치, 군사, 상업 등 다양한 목적을 위해 정교해졌다. 현대에 이르러서는 디지털 기술을 기반으로 종이 지도, 디지털 지도, 위성/항공 지도 등 다양한 형태의 서비스가 제공되고 있으며, 지도에는 다양한 실시간 정보들도 포함되고 있다. 또한, 현대 지도는 정확성이 크게 향상되어 매우 정밀한 지리 정보를 제공한다. 덕분에 이제는 누구나 컴퓨터나 스마트폰을 통해 실시간 지도 정보에 쉽게 접근하고 이를 활용할 수 있다.

본 연구에서는 포털사이트에서 제공하는 대표적인 웹 지도인 네이버 지도의 사업체 정보를 수집하고자 한다. 여기서 네이버 스마트플레이스는 사업자가 네이버 검색, 지도, 앱 등에 자신의 사업체 정보를 무료로 등록하고 관리할 수 있도록 제공하는 서비스이다(출처: 스마트플레이스 홈페이지).

스마트플레이스 등록 절차는 ①업종 검색, ②사업체 정보 확인, ③기본 정보 및 ④부가 정보 입력, ⑤등록 완료의 단계로 진행되며, 각 단계별 내용은 다음과 같다.

①업종 검색은 사업자등록증에 기재된 종목을 선택한다. 따라서, 사업자등록증은 반드시 필요한 서류이고, 업종에 따라 추가 서류 제출이 필요할 수 있다. 예를 들면, 운전학원은 사업자등록증(필수 서류)과 운전면허학원등록증(추가 서류)이 등록에 필요하다. 이처럼 네이버 스마트플레이스는 사업체의 업종에 대한 객관적이고 확실한 정보를 기반으로 운영된다고 볼 수 있다.

②사업체 정보 확인은 사업자등록증 사본 등록을 통해 진행된다. 사업자등록증은 신규 등록 시 반드시 제출해야 하는 서류이며, 사업자 정보 확인이 완료되면 스마트플레이스에 이미 등록된 업체인지 자동으로 조회된다. 이때 동일한 업체가 존재할 경우, 중복 등록 신청은 불가하다.

③기본 정보는 사업체의 일반정보를 입력한다. 여기서 **업체명, 업종, 전화번호, 주소 및 지도 위치**는 **필수 정보**이며, 업체 사진과 대표 키워드는 선택 정보이다. 이 정

보들은 사업체 영업 활동과 밀접한 관련이 있어, 대부분 사업체에서 공개적으로 제공하고 있다. 또한, 부가 정보는 **영업시간, 휴무일**, 포장, **배달**, **좌석 공간(음식점업)**, 주차, 결제 수단 등을 포함한다. 이 정보들은 사업체의 선택 입력 사항이지만 기본 정보와 같이 사업체 영업 활동을 위한 정보로서 공개되어 있다.

이 일련의 과정은 사업체 등록 확인, 중복 방지, 정확한 위치(주소) 파악 등 통계 활용 가능 정보들이 데이터베이스로 체계적으로 관리되고 있음을 시사한다. 다만, 이 정보들은 전체 등록 정보를 제공하지 않기 때문에 자료 활용에는 일부 제한적이다.

본 연구에서는 네이버 지도를 통해서 입수 가능한 정보로 **(1)연간 영업 개월 수, (2)휴무일을 활용한 사업체 정기 휴무일 수, (3)영업 시간, (4)택배 여부, (5)객석 수 여부**를 선정하고, 웹 스크래핑 기술을 활용하여 자료를 수집하고자 한다.

연구 관점에서는 경제총조사 시범예행조사 2개 지역의 전체 사업체를 대상으로 개별 정보를 수집한다. 하지만 총조사 관점에서 볼 때, 민간 기관(네이버)과의 협업을 통해 작성 시점의 대량 데이터베이스(스마트플레이스[5])를 확보한다면 이는 자료수집 방법 개선에 있어 큰 혁신이라고 평가할 수 있다.

※ 출처 : 네이버 스마트플레이스 홈페이지(https://new.smartplace.naver.com)

<그림 3-13> 네이버 스마트플레이스 메인 화면 및 등록 절차

5) 2023년 기준 스마트플레이스 이용 업체 수는 235만 개, 스마트스토어(네이버에서 제공하는 쇼핑몰 구축 플랫폼) 판매자 57만 명임

<지역별 스마트플레이스 분포도 : 수도권 53%(서울 22%, 경기 26%, 인천 5%), 부울경 22%(부산 6%, 울산 2%, 경북 4%, 경남 6%, 대구 4%), 충청권 11%(충남 4%, 충북 3%, 대전 3%, 세종 1%), 전라권 8%(전북 3%, 전남 3%, 광주 2%), 강원·제주 5%(강원 3%, 제주 2%)>

(※ 출처: 네이버 디지털 생태계 리포트 2023_네이버 Agenda Research 2023. 9. 19.)

2. 공공데이터포털·건축허브/open API·EXCEL/항목(건물 연면적)

연면적은 하나의 건축물 각 층의 바닥면적 합계로 건물의 전체 크기를 나타낸다(건축법[6]상 정의). 경제구조통계는 건축법상 정의된 연면적 개념 내에서 사업체(사업장)가 실제로 점유하고 있는 연면적을 구분하여 작성한다. 다시 말하면 넓은 의미(광의)가 아닌 좁은 의미(협의)의 개념으로 자료를 수집한다.

이 정보들은 건축물의 물리적 현황과 소유자 정보를 기록하여 관리하는 공적인 장부인 건축물대장에서 파악할 수 있다. 건축물대장에는 위치, 면적, 구조, 용도, 층수 등이 포함된 건축물의 표시에 관한 사항과 소유자의 성명, 주소, 지분 등의 소유자 현황에 관한 사항 그리고 건축물의 신축, 증축, 용도변경 등 변동 사항이 포함된다.

<표 3-4>는 건축물대장에서 제공하는 면적의 종류와 정의를 나타낸다.

연면적은 건축물대장 표제부에서 제공하며, 건축물 내 모든 층의 면적을 합친 것으로 정의된다. 전유면적은 건축물대장 전유면적 항목에서 제공하며, 소유자가 독점하여 사용하는 부분의 면적, 즉 실제 사용 면적을 말한다. 공용면적은 건축물대장 공유면적에서 제공하며, 건축물 내 사람들과 공동으로 사용하는 면적의 합계를 말한다.

경제구조통계는 사업체(사업장) 건물 연면적에 전유면적과 공용면적을 합산한 면적을 포함하는 것으로 파악되었다.

<표 3-4> 면적의 종류 및 정의

종류	정의	비고
연면적	건축물 내 모든 층의 면적을 합친 것	건축물대장_표제부
전유면적	건축물에서 소유자가 독점하여 사용하는 부분의 면적	건축물대장_전유면적
공용면적	건축물 내 사람들과 공동으로 사용하는 면적의 합계 (예. 계단, 엘리베이터, 복도 등이 포함)	건축물대장_공유면적

사업체 연면적은 산업대분류 중 제조업, 도소매업, 숙박·음식점업 3개 분류에서 작성하는 항목이다. 이 세 업종 모두에게 연면적은 필수적인 요소이다. 제조업은 생산 및 물류 효율성, 도소매업은 재고 관리 및 고객 접근성, 숙박·음식점업은 고객 경험[7] 및 공간 효율성에 초점을 맞춰 최적의 면적을 확보하고 활용하는 것이 중요하기 때문에 해당 자료를 수집한다.

---

6) 건축법 시행령 제119조(면적 등의 산정방법) 제1호 제4항 연면적

7) 고객 경험(Customer Experience, CX)은 구매 전, 소비, 구매 후 단계를 포함한 모든 단계의 소비 과정 중에서 소비자가 반응하는 인식, 정동, 감각 처리, 행동의 총체를 말한다(위키백과 사전).

<그림 3-14>는 사업체(사업장) 건물 연면적의 공개 현황을 나타낸다.

자료를 제공하는 오픈 웹사이트는 공공데이터포털과 건축HUB 두 곳이고, 각 웹사이트에서 open API 혹은 EXCEL 다운로드 방식으로 정보를 제공한다.

공공데이터포털은 검색창에 건축물대장으로 검색하면 쉽게 자료를 찾아 활용할 수 있고, 건축HUB는 ①서비스 모아보기, ②원하는대로 건축데이터, ③유형별 건축데이터 제공, ④건축물대장_표제부, 전유공용면적 순으로 자료를 찾을 수 있다.

이 데이터베이스(DB)들은 건축물 전체의 정보뿐만 아니라 호실별 상세한 정보까지 제공하며, 데이터 품질관리가 잘 되어 있어 자료 활용성이 매우 높다. 주의할 점은 해당 자료를 다른 DB와 연계할 때, 주소 정보 정비에 많은 시간과 노력이 필요할 수 있다는 점이다.

<공공데이터포털_open API·EXCEL>

<건축허브_open API·EXCEL>

<그림 3-14> 건물 연면적_공공데이터포털, 건축허브

3. 공정거래위원회/EXCEL/항목(온라인 거래 여부)

온라인 거래는 사업체와 소비자 간(B2C) 상품 및 서비스 거래가 컴퓨터 통신망의 시스템을 통해 이루어지는 거래로 정의한다(경제구조통계 정의).

국내 온라인 거래는 1994년 인터넷 상용화와 1996년 한국전자거래협회 출범, 관련 정책 및 법률 등 제도적 기반 마련을 통해 시작되었다. 1997년 최초의 온라인 쇼핑몰 등장 이후, 2005년 오픈마켓(옥션, 지마켓 등)이 활성화되고 대중화되면서 온라인 유통 시장이 크게 성장하였다. 또한, 쿠팡, 티몬, 위메프 등 다양한 형태의 쇼핑몰 등장으로 온라인 시장은 더욱 확대되었다.

2010년 스마트폰 대중화는 모바일 기기를 통한 온라인 쇼핑을 주류로 만들었으며, 해외 직구와 역직구 시장이 성장하면서 온라인 쇼핑은 국내를 넘어 글로벌 시장으로 확대되었다.

국가데이터처 온라인쇼핑동향조사 결과에 따르면, 2025년 9월 기준 국내 온라인쇼핑몰 거래액은 약 24조 원으로 추정된다. 분야별로 거래액을 살펴보면, 음식료품 3.6조 원, 음식서비스 3.3조 원, 여행 및 교통서비스 2.9조 원, 생활용품 1.7조 원, 의복 1.7조 원으로 나타났다. 특히, 온라인쇼핑몰 거래액은 전년 동월 대비 13.3% 크게 상향되어 국내 경제에서 매우 중요한 분야이자 유통 산업의 핵심 동력으로 자리매김하고 있다.

온라인 거래 여부 항목은 ①온라인 거래 여부와 ②전체 매출액 중 온라인 매출액 비중, ③온라인 매출액 중 모바일 매출액 비중 세 부분으로 나눠진다. 본 연구에서는 이 세 부분 중 ①부분을 중점적으로 다룬다. ②과 ③부분은 사업체의 민감한 정보로 분류되며, 현실적으로 정보를 파악하는 데 어려움이 있어 제외되었다.

온라인 거래 관련 사업체는 경제총조사(5년 주기, 현장조사 50%)와 서비스업조사(1년 주기, 표본조사)를 통해 파악된다. 하지만 응답자의 낮은 참여율과 불성실한 응답, 조사 전문인력 부족, 예산 및 시간적 제약 등으로 인해 조사 환경은 점차 어려워지고 있다. 이에 따라 조사항목의 자료수집 방법을 개선하고자 공공데이터를 검토했다.

※ 출처 : 공정거래위원회 홈페이지(https://www.ftc.go.kr)

<그림 3-15> 공정거래위원회 통신판매사업자 다운로드 화면

공정거래위원회는 홈페이지를 통해 통신판매사업자 데이터베이스(DB)를 제공하고 있다(<그림 3-15>). 여기서 통신판매업자[8]는 인터넷, TV홈쇼핑, 카달로그, 신문잡지 등

8) 「전자상거래 등에서의 소비자보호에 관한 법률(약칭: 전자상거래법)」 (공정거래위원회)
제2조(정의) 제3항 “통신판매업자”란 통신판매를 업으로 하는 자 또는 그와의 약정에 따라 통신판매업무를 수행하는 자를 말한다.
제12조(통신판매업자의 신고 등) 제1항 통신판매업자는 대통령령으로 정하는 바에 따라 다음 각 호의 사항을 공정거래위원회 또는 특별자치시장·특별자치도지사·시장·군수·구청장에게 신고하여야 한다. 다만 통신판매의 거래횟수(50회), 거래규모(부가가치세 간이과세자) 등이 공정거래위원회가 고시

다양한 수단을 이용해 상품 등을 판매하는 사업자를 의미한다. 또한, 통신판매사업자는 원칙적으로 통신판매업 신고 의무가 있다. 만약 통신판매업 신고를 하지 않고 영업하는 경우, 법적 제재를 받게 된다.

<공정거래위원회 행정서식>

<국세청 행정서식>

<그림 3-16> 통신판매업사업자 행정서식

통신판매업은 온라인 거래와 포괄 범위에 차이가 있다. 통신판매업은 다양한 판매 방식을 활용하는 반면, 온라인 거래는 컴퓨터 통신망 시스템을 활용한 경우로 한정된다. 따라서 통신판매업이 온라인 거래보다 더 넓은 범위라고 할 수 있다.

통신판매사업자 데이터베이스(DB)는 사업체의 다양한 정보를 제공하고 있다. 이 DB는 통신판매번호, 신고기관명, **상호**, **사업자등록번호**, **법인 여부**, 전화번호(개인정보 처리), 전자우편(비식별 처리) 신고일자, **사업장 소재지**(도로명 주소까지), **업소 상태(정상영업, 폐업처리, 휴업처리 등)**, **판매 방식(인터넷, TV홈쇼핑, 카달로그, 신문잡지, 등),** 취급 품목, 인터넷 도메인, 호스트 소재지 정보가 포함되어 있다.

본 연구에서 필요한 정보는 크게 자료 간 연계를 위한 정보와 온라인 거래 정보이다. 따라서 경제구조통계 작성에 필요한 정보를 목적에 따라 명확히 구분해 정리할 필요가 있다. 자료 연계에 필요한 정보는 사업자등록번호, 상호, 법인 여부, 사업장 소재지, 업소 상태이고, 온라인 거래 여부 파악을 위해 필요한 정보는 판매 방식에 인터넷이 포함된 사업체로 정리한다.

로 정하는 기준 이하인 경우에는 그러하지 아니한다.

4. 공공데이터포털/EXCEL/스마트공장 운영 여부 항목

국내 스마트공장 관련 정부 지원은 중소벤처기업부의 '스마트 제조 혁신 지원 사업'을 통해 다양하게 이루어진다. 특히 스마트공장 구축 사업은 정부형, 부처협업형, 지역특화형(레전드 50+, 도약(Jump-Up), 자율형공장, 제조 AI특화, 대·중·소상생형(상생형 AI 트랙), 디지털협업공장 등 다양한 형태로 지원된다.

사업 신청은 스마트공장 사업관리시스템을 통한 온라인 신청이 원칙이다. 제출 서류는 사업계획서, 사업자등록증명원 등이 필요하다. 지원 사업에 선정된 기업의 데이터는 <그림 3-17> 스마트공장 사업관리시스템[9]과 <그림 3-18> 공공데이터포털 내 스마트공장 DB에서 활용할 수 있다.

<맞춤형 공급기업 검색>

<공급기업 리스트>

※ 출처 : 스마트공장 사업관리시스템

<그림 3-17> 스마트공장 사업관리시스템 지원 기업 현황

<중소기업정보진흥원>

<중소벤처기업부>

※ 출처 : 공공데이터포털

<그림 3-18> 공공데이터포털 홈페이지_스마트공장 기업

9) 스마트공장사업관리시스템 홈페이지(https://www.smart-factory.kr) > 사업안내 > 맞춤형 공급기업 검색에서 확인 가능(기업 정보 : 공급기업 유형, 종사자규모, 매출규모, 지역, 주력분야, 전문분야, 특화업종 등)

5. 그 외 항목(전문가 자문)

앞(1.~4.)에서 다룬 특성 항목은 포털사이트 검색을 통해 해당 홈페이지(네이버 지도, 공공데이터포털, 공정거래위원회 등)에 접속하여 자료를 입수했다. 반면, 그 외[10] 특성 항목은 자료수집 개선에 필요한 원천 데이터의 존재 여부를 파악하기 위해 관련 분야 전문가 자문을 먼저 수행했다.

첫째, **로봇 활용 여부** 관련하여 한국로봇산업진흥원 전문가 자문을 진행하였다.

한국로봇산업진흥원은 공급자 관점의 로봇산업실태조사(승인통계)와 수요 사업자 관점의 로봇사용실태조사('25년 시험조사)를 작성하고 있다. 본 연구와 관련이 있는 로봇사용실태조사는 로봇 사용 판별(1차)과 본조사(2차)로 구성된다. 특히 로봇 사용 판별에서는 다양한 산업에서 로봇 활용 여부를 파악했다.

본 연구는 로봇 활용 여부의 자료수집 개선을 위해 관련 기관에서 관리하는 데이터베이스(DB)나 지원 사업 등을 파악해 보았으나, 현재는 해당 자료가 없는 것으로 확인됐다. 그렇기에 로봇사용실태조사는 모집단을 파악하고자 판별(1차) 조사를 진행하는 것으로 보인다. 여기서 주목할 점은 자료수집 체계의 개선에 앞서 실제 산업 현장에서 로봇이 활용되는 범위가 훨씬 더 넓다는 사실이다. 현재 경제구조통계는 산업대분류 3개(C, H, I)에서만 로봇 활용 여부를 파악하지만, 로봇사용실태조사에서는 제조업 외 다양한 비제조업[11] 분야에서 로봇 활용 비중이 높게 나타나고 있다.

따라서, 로봇 사용 범위를 확대하여 명확한 로봇 사용 실태를 우선 파악하는 것이 국가 정책 수립의 기초가 되는 핵심 데이터 확보에 있어 매우 중요하다.

둘째, **무인 결제 기기 활용 여부** 관련하여 소상공인시장진흥공단 전문가 자문을 진행하였다.

소상공인시장진흥공단은 소비·유통 환경의 비대면·디지털화 추세에 발맞춰, 소상공인[12] 사업장의 경쟁력을 높이기 위해 스마트 기술 도입 지원을 추진하고 있다. 여기서 스마트 기술은 소상공인이 사업장을 더 쉽고 효율적으로 운용하도록 돕는 디지털 전환 기술을 말하며, 서빙 로봇, 조리 로봇, 키오스크(무인 결제 기기 포함), 테이블오더 등이 포함된다. 소상공인 지원 방식은 일반형(키오스크, 테이블오더 등 구입비), 선도형(기술패키지, 주문제작 기술 구입비), 렌탈형(서빙·조리 로봇, 베리어프리 키오스크 렌

10) 로봇 활용 여부, 무인 결제 기기, 무인매장, AI 활용 여부

11) 제조업(C)의 비중이 매우 높으나, 비제조업 Q(보건·사회복지업), P(교육서비스업), M(전문과학기술서비스업), O(공공행정) 분야에서도 로봇 활용 비중이 높음

12) 「중소기업기본법」에 따라 매출액 기준 소기업에 해당하면서 「소상공인 보호 및 지원에 관한 법률」에 따라 상시근로자 10인 미만(제조업·건설업·운수업) 또는 5인 미만(그 밖의 업종)인 사업체

탈), SaaS형(소프트웨어 서비스 등 렌탈)로 다양하다.

소상공인시장진흥공단은 2020년부터 현재까지 지원 사업체를 선발하여 스마트 기기 설치 및 대여를 지원하고 있다. 그 결과, 결제 기기는 누적 2만 9천여 개 사업체에, 서빙 로봇은 약 7백여 개 사업체에 설치 또는 대여 지원이 이뤄졌다.

이 데이터베이스(DB)는 소상공인시장진흥공단(중소벤처기업부)에서 관리하며, 정보에는 사업체 일반정보, **무인 결제 기기·테이블오더** 설치 대수, **무인점포 여부**, **서빙·조리 로봇 렌탈 대수** 등이 포함되어 있다. 따라서 경제구조통계 작성에 활용이 가능한 것으로 판단된다. 본조사 착수 전 부처 간 협업을 통해 자료 입수를 해야 할 필요가 있다.

※ 출처 : 소상공인시장진흥공단(https://www.sbiz.or.kr)

<그림 3-19> 소상공인 스마트상점 홈페이지

다만, 여기서 고려할 점은 소상공인시장진흥공단에서 지원하는 결제기기(키오스크와 테이블오더)는 경제구조통계의 무인 결제 기기 항목과 범위가 다르다는 것이다. 소상공인시장진흥공단의 지원 기기는 사업 운영 방식(업종)에 따라 선불 결제 방식과 후불(따로) 결제 방식으로 나뉘기 때문에 포괄 범위가 더 넓다. 물론 해당 데이터베이스(DB) 내에서 선불 결제와 후불(따로) 결제를 분리 집계하는 것이 가능하다. 하지만 업종에 따라 결제 유형이 결정되므로, 통계 작성 기준에 맞춰 후불(따로) 결제 기기에 대한 포함 여부를 신중히 검토할 필요가 있다.

셋째, **인공지능(AI) 활용 여부** 관련하여 산업연구원 전문가 자문을 진행하였다.

인공지능(AI) 활용 여부 항목은 2025년 경제총조사의 신규 항목으로, 全 산업에서 조사한다. 인공지능은 산업 전반에서 생산성 향상, 비용 절감, 새로운 가치 창출을 이끌 수 있어 매우 중요하다. 이 항목은 현장 조사를 통해 자료를 수집할 예정이지만 조

사에 도움이 될 만한 보조 정보의 유무를 파악하기 위해 전문가 자문을 수행했다.

AI 활용 정보는 통계(survey 포함), 채용 플랫폼·기업정보, 정부 지원 사업 세 분야에서 확인됐다. 먼저 통계(survey 포함)는 국가데이터처 기업활동조사와 대한상공회의소와 산업연구원 서베이, 한국지능정보사회진흥원(NIA) 서베이 등에서 수행하는 통계 결과와 기업 리스트를 확보할 수 있다. 다음 <그림 3-20>과 같은 채용 플랫폼·기업정보는 기업이 공개하는 정보를 활용하여 인공지능(AI) 활용 여부를 파악할 수 있다. 스크래핑으로 데이터를 수집한 후 데이터 전처리 및 연계를 통해 활용할 수 있다. 마지막으로 AI 바우처 지원 사업 등 정부 지원 사업(과기부·정보통신산업진흥원)의 수혜 기업 리스트를 확보할 수 있다. 이 자료는 정부 지원을 받은 기업 정보이므로 가장 품질이 높은 자료라고 할 수 있다. 다만, 앞서 파악된 정보들은 주로 기업체 단위로 수집되고 있어 사업체 단위로 작성되는 경제구조통계와 차이가 있으므로 자료 활용 시 반드시 고려해야 한다.

<채용 플랫폼(잡코리아) 검색 화면>

<전자공시시스템 본문 검색 화면>

※ 출처 : 잡코리아, 전자공시시스템 웹페이지

<그림 3-20> 인공지능(AI) 활용 관련 웹사이트

제 4 장

데이터과학 기술을 활용한 자료수집

제1절 특성 항목별 자료수집

1. 스크래핑

스크래핑 과정을 이해하려면 먼저 찾고자 하는 정보의 위치를 파악해야 한다. 브라우저 화면에 정보가 표시될 때는 HTML을 이용하므로, HTML의 구조에 대한 이해가 필수적이다. HTML은 태그(tag)로 구성되어 있으며, 태그(tag)에 포함된 요소가 어떤 역할을 하는지는 태그(tag) 종류에 따라 결정된다. 이는 단순히 링크일 수도 있고, 화면에 특정 문자를 출력하는 태그(tag)일 수도 있다. 각각의 태그(tag)에는 고유한 이름(name)이나 클래스(class)가 부여된다. 따라서 원하는 태그(tag)의 이름(name)과 클래스(class)를 알고 있다면, 그 안에 있는 값에 접근하여 정보를 추출하는 것이 가능하다.

다만, 동적 웹의 경우 필요에 따라 태그(tag)가 추가로 생성되며 임의의 클래스(class)를 갖게 된다. 이 때문에 해당 값을 미리 알고 접근하기 어렵다. 따라서 기준이 되는 태그(tag) 구조를 찾은 후, 필요한 정보가 있는 구간 전체를 추출하는 방식을 사용해야 한다. HTML 구조는 브라우저에서 개발자 모드를 이용하여 브라우저 화면 요소들의 위치를 확인하여 파악할 수 있다(<그림 4-1> 참조).

<그림 4-1> edge에서 F12를 누른 후 elements를 선택하여 나온 html 결과

초기 로드 영역 외에 검색이나 상호작용으로 나타나는 화면 대부분은 IFrame 내에서 정의되고 호출된다. 따라서 데이터 조회가 정상적으로 이루어졌다면, 원하는 값을 찾기 위해서는 IFrame 내부에서 탐색 작업을 수행해야 한다.

스크래핑 과정의 큰 틀은 다음과 같다. 주어진 주소자료에서 도로명 주소와 상호명을 기준으로 지도 서비스 정보를 검색한다. 이후 검색된 자료에서 영업시간, 객석 수, 배달 여부, 택배 여부 등의 상세 정보를 수집한다. 여기서 어려운 점은 지도 서비스의 정보들이 동적 웹 방식으로 작동한다는 것이다. 이는 사용자와의 상호작용에 따라 추가 정보가 비동기적으로 제공되는 방식이다.

현재 화면에 나타나는 정보는 검색과 동시에 모두 생성되는 것이 아니라, 추가적인 사용자 상호작용을 통해서만 제공된다. 따라서 스크래핑 시, 원하는 정보가 화면에 표시되도록 마우스 클릭과 같은 동작을 코드로 구현해 주어야 한다.

파이썬에서 브라우저 상호작용을 자동화하는 대표적인 라이브러리로 selenium이 있다. 하지만 이번 연구에서는 playwright 라이브러리를 활용했다. playwright 라이브러리는 마이크로소프트에서 유지 보수하므로, 향후 더 안정적으로 지원될 것으로 기대하여 채택하였다.

실제로 selenium 라이브러리는 사용 시 몇 가지 문제점이 발견되었다. 특히, 동적 웹 환경에서 브라우저 화면을 띄우지 않고 작업을 하는 headless 모드 사용 시, 상호작용이 제대로 작동하지 않는 경우가 많았다.

반면에 playwright를 사용했을 때는 headless 모드에서도 원하는 동작이 정상적으로 이루어졌다. 자료 수집을 위해 반복적으로 브라우저 화면을 띄우는 것은 비효율적이므로, 이러한 안정적인 headless 지원이 playwright를 최종 선택한 이유가 되었다.

스크래핑의 작업을 크게 다음과 같은 세 가지 핵심 영역으로 구성된다.

(1) 스크래핑과 직접적으로 관련된 함수들

(2) 주소 파일을 읽어오고, 수집된 자료를 내보내며, 정보를 정리하는 작업, 또한 과정 중 발생하는 오류나 중요한 정보들을 기록하는 로그 작업

(3) 스크래핑에서 사용하는 정해져 있는 파라미터들의 관리

따라서 각 과정을 역할별로 명확히 분리하여 독립적인 패키지로 구성하고 처리한다.

파이썬에서 패키지는 실행 파일과 같은 위치에 있으며, __init__.py 파일을 포함하는 폴더를 의미한다. 이러한 폴더 아래에 있는 .py 파일들은 관련된 함수들을 묶어 놓은 것으로, 이를 모듈(Module)이라 부른다. 메인 실행 파일은 이 패키지 내의 모듈에 접근하여 필요하면 함수들을 가져와 사용할 수 있다.

현재 프로젝트 함수 위치는 다음과 같이 구성되어 있다. 핵심 스크래핑 로직과 관련된 함수들은 functions 패키지 아래의 scrape_info.py 모듈에 정의되어 있다. 메인(main) 함수는 이 모듈 내의 함수들을 순서에 맞춰 호출하며 자료수집 과정을 총괄한다. 자료의 입출력과 관련된 함수들은 utils 패키지에, 작업을 할 때 활용하는 파라미터들은 config 패키지의 모듈에 저장되고 있다.

주요 작업을 처리하는 scrape_info.py 모듈 내부의 함수들을 소개하고, 수집 과정을 설명하면 다음과 같다.

우선 메인(main) 함수는 주소 정보 파일 이름, 정보가 존재하는 시트 이름, 출력 파일 이름, 컬럼명 매핑(주소 정보가 제공된 파일의 컬럼명이 다른 경우), 브라우저 표시 여부, 다중 코어 사용 여부 결정, 배치 사이즈 결정, 코어 사용 개수 설정 등 다양한 실행 매개변수를 입력받는다. 이렇게 입력받은 값들을 기준으로 load_input_data를 사용해 데이터를 불러온 후, 필요한 변수들만 남기는 작업을 수행한다.

필요한 변수들을 걸러내는 작업은 extract_var 함수가 수행한다. 주요 작업은 컬럼명 매핑 정보를 활용하여 컬럼명을 통일하고, 필요하지 않은 컬럼은 제외한다. 필요한 컬럼들은 “사업체고유번호”, “사업체명”, “소재지도로읍면동명”, “소재지도로명”, “소재지도로건물본번”, “소재지도로건물부번”이다. ‘사업체고유번호’는 각 사업체별 부여된 고유번호이므로, 키 값으로 나중에 자료를 결합하는 데 사용할 수 있다. 그 외의 정보들은 검색어를 생성하는 데 사용된다. 그 후, ‘사업체명’ 변수의 값을 한 번 정제한다(normalize_korean). 상호명에 한자가 쓰여있는 경우에 우리말 발음으로 바꿔준다.

이제 자료로부터 지도 서비스에서 검색을 시행한다. 서비스를 실행하고 검색창에 입력하는 것보다 브라우저에 url을 바로 입력해서 처리한다. url 생성은 make_query_entry 함수를 통해서 이루어진다. url을 생성할 때 활용되는 컬럼값은 ‘사업체고유번호’를 제외한 정보들이다. 검색 결과에서 자료를 수집하여 하나의 자료를 만든 후, 기존의 정보와 결합하기 위해서 ‘사업체고유번호’도 함께 기록하여 다음 단계로 넘겨준다.

다음 단계에서는 worker_process 함수로 전달된 정보를 바탕으로 데이터 수집 작업이 진행한다. 생성된 url로 이동하기 위해 launch_browser를 실행하여 playwright를 실행하고, 브라우저 및 페이지 객체를 생성한다. 실제 웹 페이지가 로드된 페이지 객체와 url 정보를 데이터를 수집하는 process_store 함수에 넘겨준다.

process_store 함수는 사업체의 영업시간, 객석 정보, 배달 및 택배 정보를 수집하고 통합하는 작업을 수행한다. 페이지 객체의 goto 메서드를 이용해 상호 정보를 검색하는 url로 이동하며, 지도 검색 결과 페이지로 이동한다(goto_store_page). 원하는 정보가 페이지에 로드된 후, locator 메서드를 사용하여, 웹 페이지의 특정 HTML 태그(tag)를 찾아 해당 요소로 이동한다.

본 연구에서는 정보가 포함된 영역인 IFrame을 찾아가도록 구현하였다. 검색 결과에 영업시간 정보가 출력될 때 IFrame 태그(tag)가 동적으로 생성되기 때문이다. 이 과정에서 검색 후 충분한 대기시간을 확보하지 않으면 화면이 완전히 생성되지 않은 상태에서 지정된 태그(tag)를 찾으려 시도하게 되며, 이로 인해 원하는 결과를 얻지 못하는 문제가 발생할 수 있다.

이러한 현상은 실제로 화면이 보이지 않더라도 마찬가지므로 필요한 항목이 로드될 때까지 기다리는 시간을 설정한다. 이는 네트워크 상태에 따라 달라질 수 있으므로 환경에 따른 설정이 필요하다. 설정된 시간만큼 기다렸음에도 필요한 항목이 로드되지 않을 수 있으므로 아무런 정보도 취득되지 않으면 다시 한번 항목을 찾는 시도를 하여 거듭 로드가 늦은 위치에 접근한다. 그럼에도 불구하고 정보가 없는 경우에는 빈 값을 반환한다.

페이지가 생성되었다면, 원하는 정보가 포함된 IFrame을 찾아 실제 정보가 있는 부분을 추출한다. 이때 IFrame의 클래스(class)는 entryIframe과 searchIframe의 두 가지가 존재한다.

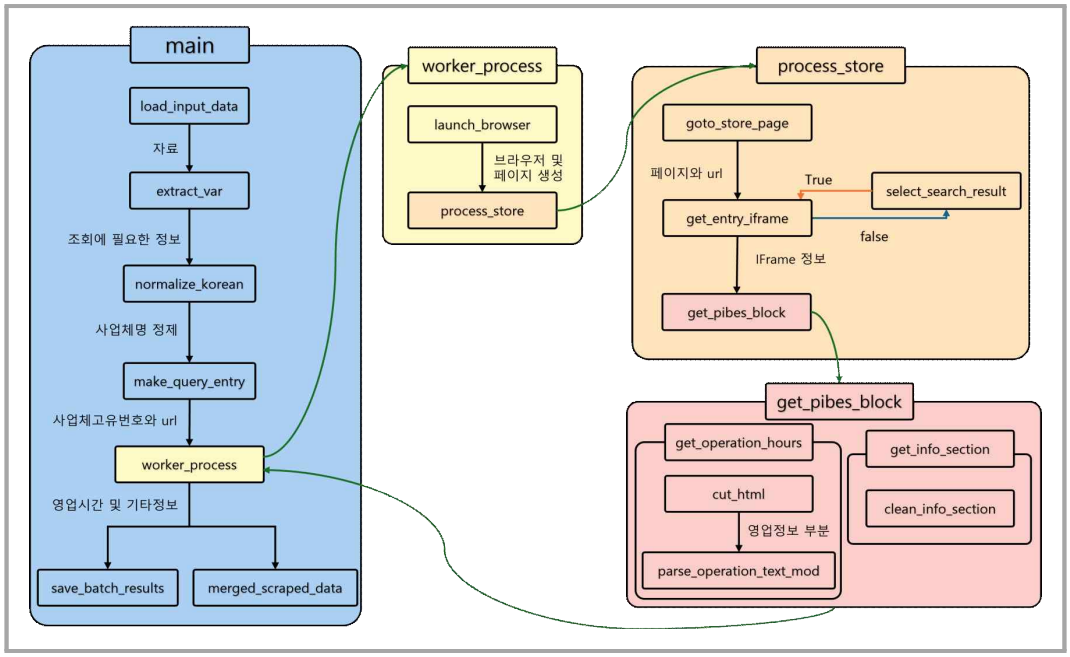
지도 검색을 통해 나온 상호의 영업시간을 비롯한 정보들은 entryIframe에서 추출할 수 있다. 반면, 검색 결과가 하나 이상인 경우에는 searchIframe이 생성된다. searchIframe이 생성될 때는 검색 결과의 상호명 검색 시 사용한 '사업체명'과 비교하여, 일치하는 값 또는 일치도가 높은 값을 선택한다(select_search_result).

원하는 결과 선택 후, entryIframe이 다시 생성된다. 이때, get_entry_iframe 함수를 사용해 해당 entryIframe에서 정보를 추출하는 작업을 수행하고, 사업체의 정보를 포함하고 있는 IFrame을 반환한다.

영업시간을 비롯한 정보는 IFrame의 pibes라는 클래스(class)에 항상 생성되므로, 해당 정보 객체를 get_pibes_block 함수로 넘겨준 후 get_operation_hours 함수 내에서 영업시간을 추출하는 작업을 처리한다. 필요한 HTML 부분을 잘라(cut_html) 보관해 두었다가, 영업시간 관련 정보를 추출하여(parse_operation_text_mod) 대응하는 컬럼들에 배치한다.

영업시간 부분의 정보는 실제로는 영업시간, 진료시간, 라스트오더, 배달시간, 워크인 주문시간, 정기휴무, 브레이크타임 등 다양한 내용이 포함되어 있다. 따라서 즉각적인 처리가 어려운 경우, 해당 정보가 있을 것으로 예상되는 HTML의 일부를 정제하지 않은 상태로 배치별로 저장한다. 이후 정제된 좌석 및 배달 정보 추출 결과(get_info_section)를 함께 결합한다.

각 함수의 작업을 통해 추출된 영업시간과 기타 정보들을 하나의 데이터로 통합한다. 이렇게 생성된 정보들은 기존 입력 자료와 결합하여 최종 하나의 자료로 저장된다. 함수들이 작동하는 구체적인 순서와 전달되는 값의 흐름은 제공된 순서도(<그림 4-2>)에서 간략하게 확인할 수 있다.



<그림 4-2> scraper의 main 함수 내에서 작업의 흐름을 나타내는 순서도

2. open API

공공데이터포털에서 제공하는 '국토교통부 건축HUB 건축물대장정보' 자료를 open API를 통해 수집하여, 각 사업체의 면적 또는 연면적 정보를 추가하고자 한다.

우선 open API의 자료에 접근하려면, 먼저 서비스 활용 신청을 통해 권한을 승인받아야 한다. 정상적으로 허가를 받으면 서비스 키가 발급되며, 이 키를 사용하여 정보 요청 및 수신이 가능해진다. open API를 통해 정보를 제공받을 수 있는 계정은 개발계정과 운영 계정 두 가지 종류가 있다.

개발 계정은 하루 10,000개, 운영 계정은 하루 1,000,000개 규모의 트래픽을 제공받을 수 있다. 건축Hub의 경우, 자동 승인이 설정되어 있어 서비스 목표와 활용 계획을 명확히 설명하면 승인을 받아 사용할 수 있다. 다만, 계정과 무관하게 제한 속도는 30tps(transaction per second)이다. 즉, 초당 30건 이상의 정보 요청 및 응답을 주고받을 수 없다. 따라서 API 호출 시 초당 처리량이 30건이 넘지 않도록 요청 간의 지연 시간을 고려해야 한다.

정보를 요청할 수 있는 권한을 얻은 후, Open API 활용 가이드에서 따라 서비스 키를 발급받아야 한다. 또한, 응답 시 활용할 행정기관 표준 코드인 시군구 코드를 행정표준코드관리시스템(http://www.code.go.kr)에서 조회하여 활용한다.

이 코드들은 API 요청 시 반드시 필수한 정보이다(<그림 4-3> 항목 구분 참고). Open API는 서비스키와 필수 항목들을 포함하여 관련 정보를 요청하면, xml 또는 Json 형태로 조회 결과를 알려준다.

- 응답메시지 명세

항목명(영문)	항목명(국문)	항목크기	항목구분	샘플데이터	항목설명
items			1..n		
itgBldGrade	지능형건축물등급	VARCHAR(100)	옵		지능형건축물등급
itgBldCert	지능형건축물인증점수	NUMBER(2,5)	옵	0	지능형건축물인증점수
crtnDay	생성일자	VARCHAR(30)	필	20220813	생성일자
naBjdongCd	새주소법정동코드	VARCHAR(30)	옵	10301	새주소법정동코드
naUgrndCd	새주소지상지하코드	VARCHAR(30)	옵	0	새주소지상지하코드
naMainBun	새주소본번	VARCHAR(20)	옵	5	새주소본번
naSubBun	새주소부번	VARCHAR(20)	옵	0	새주소부번
platArea	대지면적(㎡)	NUMBER(30,9)	옵	0	대지면적(㎡)
archArea	건축면적(㎡)	NUMBER(30,9)	옵	15403.97	건축면적(㎡)
bcRat	건폐율(%)	NUMBER(2,5)	옵	0	건폐율(%)

- 13 -

<그림 4-3> 응답 메시지 명세 예시

본 연구에서는 Open API를 통해 건축물대장 총괄표제부, 건축물대장 층별개요, 건축물대장 전유부 자료를 xml 형태로 제공받아 처리하였으며, 이를 통해 필요한 정보를 수집하였다.

xml 형태로 제공된 자료의 구조는 HTML과 유사하다. 항목명이 태그(tag)로 지정되어 있으며, 해당 항목에 대응하는 값이 태그(tag) 내부에 포함되어 있는 것을 확인할 수 있다(<그림 4-3>, <그림 4-4> 참고).

- 요청/응답 메시지 명세

REST(URI)
http://apis.data.go.kr/1613000/BldRgstHubService/getBrRecapTitleInfo?sigunguCd=11680&bjdongCd=10300&bun=0012&ji=0000&serviceKey=인증키 (단, 웹스플로러에서 확인시 파라미터 값이 한글인 경우 utf-8로 인코딩 필요)

응답 메시지

```
<response>
  <header>
    <resultCode>00</resultCode>
    <resultMsg>NORMAL SERVICE</resultMsg>
  </header>
  <body>
    <items>
      <item>
        <rnum>1</rnum>
        <platPlc>서울특별시 강남구 개포동 12번지</platPlc>
        <sigunguCd>11680</sigunguCd>
        <bjdongCd>10300</bjdongCd>
        <platGbCd>0</platGbCd>
        <bun>0012</bun>
        <ji>0000</ji>
```

<그림 4-4> 응답 메시지 예시

상위 태그(tag)에 포함된 하위 태그(tag)들에 접근하기 위해서는 상위 태그(tag)부터 순서대로 접근해 나가야 한다. 가장 상위에는 response 태그(tag)가 있으며, 그 안으로 body, items, item 순서로 정보에 접근할 수 있다.

정보수집을 위한 함수는 functions 패키지의 openAPI_collect.py 모듈에서 제공되며, 주요 함수는 다음과 같다. 시군구 및 법정동 코드를 포함한 파일을 읽어오는 load_codes 함수와 불러온 자료에서 필요한 부분만 조회하여 시군구 및 법정동 코드를 추출하고, 시군구 코드를 딕셔너리 형태로 반환해 주는 extract_add_code_gen이 여기에 포함된다.

가. collect_API_info.py

collect_API_info.py 파일은 위에서 설명한 함수들을 활용하여, 본 연구의 조사 대상 지역에 해당하는 정보 조회용 코드들을 추출한다. 추출된 이 코드들을 바탕으로 Open API에 정보 요청을 보내는 데 필요한 데이터를 구성한다.

서비스키(ServiceKey), 시군구코드(sigunguCd), 법정동코드(bjdongCd), 리스트수(numOfRows), 페이지번호(pageNo) 값들을 요청 메시지 명세에 대응하는 항목명에 맞추어 지정해 준다(<그림 4-5> 참고).

- 요청메시지 명세

항목명(영문)	항목명(국문)	항목크기	항목구분	샘플데이터	항목설명
sigunguCd	시군구코드	VARCHAR(30)	필	11680	행정표준코드
bjdongCd	법정동코드	VARCHAR(30)	필	10300	행정표준코드
platGbCd	대지구분코드	VARCHAR(30)	옵	0	0:대지 1:산 2:블록
bun	번	VARCHAR(20)	옵	0012	번
ji	지	VARCHAR(20)	옵	0000	지
startDate	검색시작일	VARCHAR(30)	옵		YYYYMMDD
endDate	검색종료일	VARCHAR(30)	옵		YYYYMMDD
numOfRows	리스트수	VARCHAR(3)	옵	10	페이지당 목록 수
pageNo	페이지번호	VARCHAR(3)	옵	1	페이지번호

※ 항목구분 : 필수(필), 옵션(옵), 복수건(복)

<그림 4-5> 응답 메시지 항목 구분표

이때, 시군구 코드와 법정동 코드는 extract_add_code_gen 함수의 결과를 이용하며, 리스트 수는 100개, 페이지 번호는 1개를 사용한다. 행의 개수는 최대 100개로 제한되어 있지만, 함께 요청 시 페이지 번호를 명시하지 않으면 항상 한 행의 값만 반환되므로 유의해야 한다. 건물의 정보는 법정동 단위로 요청 및 수신되므로, 한 번의 요청으로 받은 정보의 수가 100행을 초과하는 경우가 발생할 수 있으며, 이는 다음 절차에서 처리한다.

다음으로 get_api_url 함수를 통해 요청에 필요한 url을 생성한다. requests 라이브러리에서 url을 자동으로 생성해 주는 함수를 사용할 수도 있지만, 서비스 키 값 변환 오류 문제가 발생했다. 따라서, 필요한 모든 항목을 문자열로 인식시킨 후 결합하여 url을 생성하는 방식을 채택했다.

이렇게 생성된 값은 get_info 함수에 전달되어, requests 라이브러리의 get 함수를 통해 정보를 획득한다. xml 형태로 반환된 정보를 쉽게 사용하기 위해 xmltodict 라이브러리의 parse 함수를 사용하여 딕셔너리 형태로 변환한다. 이 딕셔너리는 키 값으로 태그(tag) 이름이 지정되며, 태그(tag) 안에 다른 태그(tag)가 포함되는 계층적 구조이므로 키에 대응하는 값이 또 다른 딕셔너리가 되는 구조이다.

딕셔너리 객체가 가진 get 메서드를 사용하여 그 내부에 있는 값에 접근하고 호출함으로써 원하는 정보를 추출할 수 있다. 응답 데이터 중 request > body > totalCount 경로에는 요청 요건을 만족하는 전체 데이터 수가 포함되어 있다. 이를 기반으로 pageNo 응답값을 변경해 가며 모든 정보를 지속적으로 수집한다. 그렇게 수집된 정보들을 하나의 csv 파일로 만들어서 저장한다.

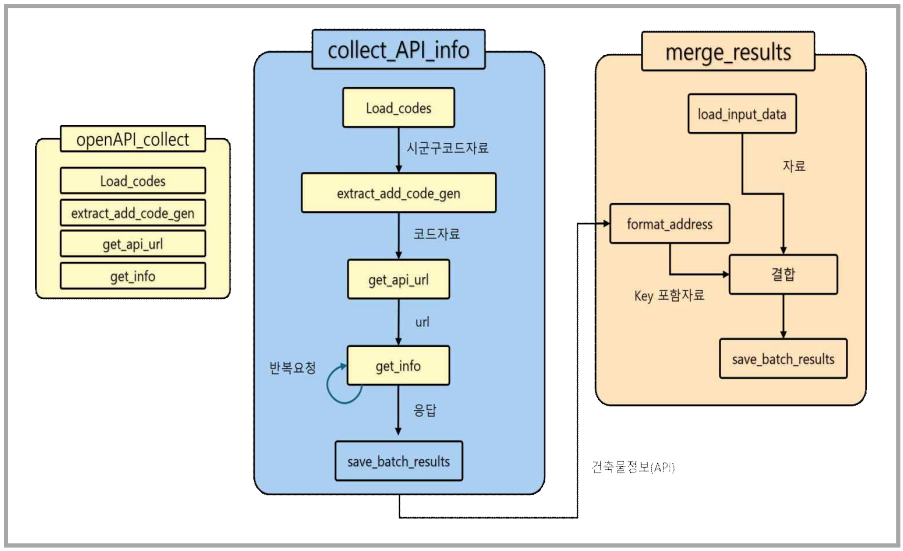
나. merging_results.py

merging_results.py 파일은 두 가지 주요 작업을 수행한다. 먼저, 조회 대상 사업체 정보가 포함된 자료를 scrape_info.py에서 사용했던 load_input_data 함수를 사용하여 불러온다. 다음으로, collect_API_info.py를 통해 생성된 결과 파일을 함께 불러와 주소지가 일치하는 사업체 자료에 층 정보와 연면적을 결합해 준다.

사업체 정보를 포함하는 데이터의 주소 정보가 '소재지도로명', '소재지도로건물본번', '소재지도로건물부번'으로 나뉘어 있다. 반면, collect_API_info.py에서 생성된 자료의 주소는 하나로 통합되어 있다. 따라서, 두 자료의 결합에 사용할 키 값을 통일하기 위해, 주소 정보를 하나로 묶어주는 format_address 함수를 정의하고, 새로운 키 컬럼을 생성한다.

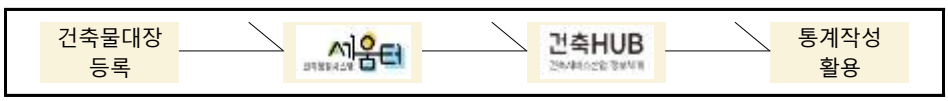
사업체의 면적을 조회할 때 키 값만으로는 특정되지 않는 경우가 있어, '소재지도로건물층명'도 함께 키 값으로 사용한다. collect_API_info.py 결과 컬럼 중 이에 대응하는 결합 컬럼 값은 'newPlatPlc_clear'와 'flrNoNm'이다. 특히, newPlatPlc_clear는 새로 생성된 값으로, 기존의 도로명 주소값에서 불필요한 '(동)'부분을 제외 처리하여 정제한 값이다.

연면적의 경우에는 건물 자체의 면적에 해당하므로, newPlatPlc_clear 키 값만 사용했으며, 추가적인 키 값은 활용하지 않았다. 위의 과정을 통해 최종 결합된 자료를 저장한다.



<그림 4-6> open API를 활용한 정보수집 과정

사업체 건물 연면적 정보(건축물대장)를 수집하는 또 다른 방법은 웹사이트(건축HUB)에서 파일을 직접 다운로드(EXCEL)하는 것이다. <그림 4-7>은 건축물대장의 데이터 흐름도이다. 건축물대장 신고 자료는 세움터 홈페이지에서 정보 조회가 가능하며, 이 정보들이 건축HUB 시스템으로 집계되어 이용자들에게 제공된다.



<그림 4-7> 건축물대장 데이터 처리 흐름도

제공되는 건축물 정보는 크게 표제부, 층별개요, 전유공용면적으로 구성된다. 표제부는 건물 전체의 개요와 면적, 층별개요는 각 층별 면적 및 구조, 전유공유면적은 각 호실별로 독립 사용(전용)하는 면적과 계단·복도 등 공용면적을 구분하여 보여준다.

경제구조통계 작성 원칙에 가장 부합되는 정보는 전유공용면적이므로, 이 정보를 기준으로 우선적인 데이터 연계를 진행한다. 이 단계에서 가장 중요한 연계 정보는 상세한 주소 정보이다. 만약 이 주소 정보 부족 등의 이유로 1차 연계가 어려울 경우, 층별개요, 표제부 순으로 연계하여 보조적인 정보를 수집하고자 한다.

<그림 4-8>은 1개 사업체의 건물 연면적 정보로 세움터, 카카오지도, 조사명부, 건축허브(표제부, 층별개요, 전유공용면적의 정보를 보여준다. 이 예시에서 주목할 점은 전유공유면적에 해당 주소의 정보가 없다는 것이다. 이 경우 앞서 제시한 처리 방식(①전유공용면적, ②층별개요, ③표제부)에 따라 자료를 연계하고자 한다.

<그림 4-8> (예시) '부산시 북구 덕천동 기찰로 42 1층(355-11)' 건물 연면적

3. 그 외 DB(EXCEL) 확보

여기에서는 데이터과학 기술을 활용하는 대신, 오픈 공공데이터를 통해 수집 가능한 경제구조통계의 특성 항목에 관해 설명한다. 여기서 다룰 대상 항목은 온라인 거래 여부와 스마트공장 운영 여부이다.

먼저, 온라인 거래 여부는 연구보고서 3장 2절에서 설명한 바와 같이, '공정거래위원회 통신판매사업자 DB' 중 판매 방식에 인터넷이 포함된 사업자와 업소 상태가 정상영업인 사업체를 대상으로 선정한다.

통신판매사업자 DB에 포함된 정보는 <표 4-1>과 같다.

데이터 간 연계는 ①사업자등록번호를 활용하여 직접 연계가 가능하고, ② 그 외 미연계 사업체(사업자등록번호 오류 등)는 상호, 대표자명, 사업장소재지(일부) 등 보조정보를 활용하여 연계를 시도한다. 이 DB는 공공 오픈 데이터 중 사업자등록번호를 포함하고 있어 활용 가치가 매우 높다.

다음으로, 온라인 거래 사업체는 판매 방식에 인터넷 판매가 포함된 곳으로 정의한다. 사업체들은 재화와 용역 판매를 위해 다양한 판매 방식을 활용하며, 특히 구매자 편의성이 높은 인터넷 판매를 많이 이용하는 추세이다.

<표 4-1> 온라인 거래 DB 정보

순번	항목	구성
1	통신판매번호	연도(4자리)-시군구-일련번호(4자리) (예, 2025-경북경산-0857)
2	신고기관명	시도 시군구
3	**상호**	-
4	**사업자등록번호**	10자리 숫자
5	법인여부	개인, 법인
6	**대표자명**	-
7	전화번호	비실별화 (예, 010-개인정보)
8	전자우편	비식별화
9	신고일자	8자리 숫자 (예. 20251030)
10	**사업장소재지**	지번 주소 체계, 상세주소 비식별화 (예, ~ 옥산동 ****-*)
11	**사업장소재지(도로명)**	도로명 주소 체계, 상세주소 비식별화 (예, ~ 경산로 *길 **)
12	**업소상태**	정상영업, 폐업처리
13	신고기관대표연락처	신고기관 대표 연락처(시군구 대표 전화번호)
14	**판매방식**	인터넷, TV홈쇼핑, 카달로그, 신문잡지, 기타(중복 가능)
15	취급품목	가구, 가전, 건강/식품, 도서, 의류 등
16	인터넷도메인	종합쇼핑몰, URL, 블로그 등
17	호스트서버소재지	-

<표 4-2>는 17개 시도별 통신판매사업자 현황을 나타냈다.

전체 통신판매사업자는 230만 개로 지역별로 살펴보면, 경기 지역이 720,820개(31.3%)로 가장 많았고, 서울 지역 660,055개(26.6%), 인천 지역 143,178개(6.2%) 순으로 나타났다. 또한, 이 세 지역은 수도권(서울, 경기, 인천)으로 전체 통신판매사업자 대비 64.1%의 높은 비중을 차지하고 있다.

<표 4-2> 17개 시도별 통신판매사업자 현황

(단위 : 개. %)

시도	사업자 수	비중	시도	사업자 수	비중
서울	660,055	26.6	강원	49,865	2.2
부산	117,296	5.1	충북	43,468	1.9
대구	94,753	4.1	충남	63,187	2.7
인천	143,178	6.2	전북	41,219	1.8
광주	41,461	1.8	전남	44,016	1.9
대전	59,447	2.6	경북	76,037	3.3
울산	24,331	1.1	경남	83,847	3.6
세종	11,622	0.5	제주	29,676	1.3
경기	720,820	31.3	합계	2,304,278	100.0

이 자료는 여러 판매 채널의 통신판매 정보가 포함되어 있으므로, 조사 정의에 맞게 필터링(인터넷 포함) 작업이 필수적이다. <표 4-2>는 모든 채널이 망라된 통신판매 사업자 정보이므로, 경제구조통계의 온라인 판매 여부 항목 조사를 위해서는 인터넷 판매 방식을 포함하는 데이터만을 추출하도록 필터링해야 한다.

또한, 해당 데이터를 활용하기 위해서는 데이터 전처리 과정이 반드시 동반되어야 한다. 이 자료는 연도별 누적 사업자가 표시되어 있으며, 시군구 단위에서 EXCEL 형태로 입력되고 있어 형식이 표준화되어 있지 않다. 특히 특수문자 사용에 제한이 없어(데이터 비표준화) 데이터를 불러올 때 예상치 못한 줄 바꿈 현상이나 오류가 발생할 수 있다.

다음으로, 스마트 공장운영 여부는 '공공데이터포털 중소벤처기업부 스마트제조혁신추진단 스마트공장 공급기업 DB'에서 자료 입수가 가능하다. 이 DB는 스마트공장 사업관리시스템(smart-factory.kr, 중소기업기술정보진흥원)에서 개방된 데이터이고, 해당 DB는 부여번호, 기업명, 대표자명, 주소, 담당자, 담당자 연락처로 구성되어 있다.

<표 4-3> 스마트공장 DB 정보

순번	부여번호	기업명	대표자명	주소	담당자	담당자연락처
1						
2						
⋮						
2,277						

스마트공장은 현재 전국 기준 2,277개가 등록되어 있고, 데이터 간 연계는 사업체 식별정보(사업자등록번호)가 없어 기업명, 대표자명, 주소(상세 정보 포함) 등의 보조정보를 이용하여 경제구조통계 조사 명부와 결합할 수 있다.

<표 4-4> 17개 시도별 스마트공장 분포 현황

(단위 : 개)

시도	개수	시도	개수	시도	개수
서울	631	울산	51	전북	37
부산	168	세종	12	전남	28
대구	184	경기	523	경북	94
인천	93	강원	11	경남	193
광주	64	충북	57	제주	6
대전	81	충남	44	합계	2,277

또한, 이 DB는 중소기업 제조 현장의 경쟁력 제고를 위해 국가에서 지원하는 사업의 일환으로 생성된 것으로, 2,277개 사업체가 국내 모든 스마트공장 수라고 판단하면 안 된다. 이 정보는 통계 작성에 직접적인 활용보다는 조사의 편의를 위한 간접 정보로 조사명부에 추가하여 자료수집 시 정확한 통계 작성에 활용할 수 있다.

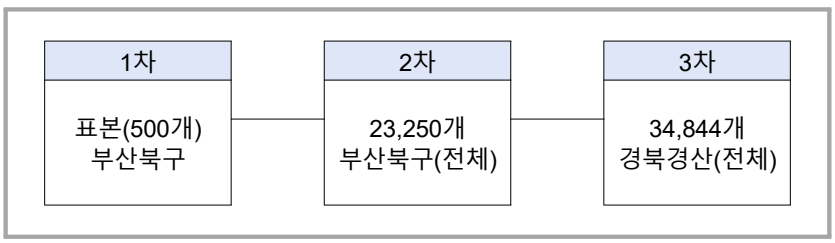
본 연구는 입수자료의 활용성 판단을 위해 분석 대상 지역[13]의 조사결과와 입수자료를 검증한다. 입수자료 결과 부산시 북구는 9개 기업, 경북 경산시는 22개 기업이 스마트공장을 운영하는 것으로 나타났다. 이 정보를 활용하여 경제총조사 시범예행조사 결과와 이 DB와 검증을 하고자 한다.

제2절 자료수집 시 유의 사항

1. 스크래핑

자료수집은 부산광역시 북구와 경상북도 경산시의 자료를 기반으로 이루어졌다. 최초에는 부산광역시의 자료 중 산업군을 고려한 500개 표본 사업체를 대상으로 하였다. 그러나 수집 결과, 극소수의 특정 산업군에 대해서만 정보가 스크래핑이 되었다. 그 이유는 해당 사업체들이 손님에게 정보를 제공할 목적이 아니라면 온라인상에서 위치 정보나 영업시간을 게시하지 않는 경우가 많았기 때문이다.

위치 정보는 제공하더라도 영업시간을 제공하지 않는 사례 또한 많았다. 따라서 더 많은 자료(부산광역시 북구 전체)를 활용하여 정보수집을 시도했다. 이 경우에도 약 전체 사업체 대비 7~10% 내외의 사업체에 대한 정보만 확보할 수 있었다.



<그림 4-9> 스크래핑 대상 사업체 확대

13) 경제총조사 시범예행조사 대상 중 부산시 북구, 경북 경산시

스크래핑할 때 연계율이 낮은 이유는,

① 조사 명부의 정보와 지도 서비스의 정보가 일치하지 않아 데이터 연계에 어려움이 있었다. 조사 명부의 사업체명에 한자나 영어가 포함되어 있지만 지도 정보에는 누락된 경우, 법적으로 등록된 사업체명과 실제 간판 등에 사용되는 명칭이 다른 경우들이다. 이러한 문제를 해결하기 위해, 지도 검색 결과로 반환된 상호명과 원본 조사 명부의 사업체명 간 유사도를 측정하고, 유사도가 높은 경우에만 해당 사업체의 정보를 스크래핑하여 데이터의 정확성을 확보했다.

② 사업체 정보의 신뢰성과 관련하여, 해당 데이터가 항상 최신 정보가 아닐 수 있다는 점에 유의해야 한다. 특히 고객들에게 최신 정보를 지속적으로 제공해야 하는 업종의 경우 정보 갱신이 비교적 잘 이루어지지만, 그 외의 사업체 정보는 마지막 업데이트 시점이 언제인지조차 확인할 수 없는 경우도 있었다.

③ 작업시간이 오래 걸리는 것 또한 큰 문제였다. HTML 요소들의 로드 시간을 예상하여 대기 시간을 설정했는데 처리해야 할 표본의 크기가 커짐에 따라(500개 → 23,250개 → 34,844개) 이 대기 시간이 누적되어 전체 작업 시간이 매우 길어졌다. 이를 해결하기 위해서 멀티 코어를 활용한 병렬 처리를 도입했으나, 여전히 예상보다 오랜 시간이 소요되었다.

④ 주소 정보 역시 문제점으로 지적되었다. 부산광역시 북구의 경우, 도로를 따라 조사를 진행하다 보면 일부 주소지가 북구를 벗어나는 사례가 발생하여, 조사된 주소가 해당 구역 내에 존재하지 않는 경우가 있었다. 이는 스크래핑 이전에 원천 자료가 오염되지 않았는지 데이터의 무결성을 확인할 필요가 있음을 시사한다. 즉, 사업체명이나 주소 같은 핵심 정보는 수집 후, 데이터 전처리 또는 정제 과정을 반드시 거쳐야 한다는 것이다.

⑤ 스크래핑 대상 사업체가 지도 서비스를 통해 제공하는 영업시간 정보는 보통 일주일 치 정도의 자료이다. 예를 들어, 정기 휴무일 외에 공휴일 여부 등 유동적인 변화에 대한 정확한 정보를 얻기 위해서는 일정 기간마다 정기적으로 자료를 수집(주기적인 스크래핑)할 필요가 있다. 하지만 서비스 제공자가 웹 페이지의 내용을 변경하면서 HTML 구조를 바꾸게 되면, 기존 스크래핑 로직으로는 정보를 더 이상 추출하지 못하게 될 수도 있다. 따라서 정기적인 데이터 수집과 함께 정보가 제대로 수집되고 있는지 지속적으로 확인하고 모니터링하는 절차가 반드시 동반되어야 한다는 것이다.

2. API

API를 통해 정보를 얻기 위해 공공데이터서비스의 건축서비스산업 정보체계(건축HUB) 시스템이 제공하는 건축 정보를 활용했다. 초기에 기획했던 방식은 조사 대상에 포함된 하나의 자료마다 API를 개별적으로 조회하여 개별 면적 및 연면적을 결합하는 것이었다. 그러나 이는 지나치게 많은 시간을 소요하는 작업이 되었다. 초당 최대 호출 횟수가 30회로 제한되어 있었고, 인증된 계정의 자격에 따라 일일 한도에 쉽게 도달하여 대기 시간이 길어졌다. 이러한 시간 지연 문제를 회피하기 위해, 해당 지역의 전체 건축물 정보를 모두 다운로드하여 저장한 후 저장된 지역 정보와 비교 결합하는 방식으로 변경하였다.

스크래핑 및 API 사용 과정에서 문제가 되었던 점은 조사된 원천 자료의 전처리 불확실성이었다. 정보를 수집하기 위해 지역자료에서 필요한 값들을 추출하여 활용하는 과정에서, 예상 범위를 벗어나는 데이터들이 자주 발견되었다. 앞서 소개했던 상호명에 한자가 섞여 있다거나 주소에 한자가 섞여 있는 경우, 지번 주소 체계에서 본번과 부번에 문자가 섞여 있는 자료들이다. 또한 현재는 일부 지역만 대상으로 했기 때문에 행정 구역상 해당 구역에 속하지 않더라도 도로가 이어져 있어 사업체가 조사 목록에 혼재되어 있는 경우들이 있었다.

실제로 처음 500개의 작은 표본 구역에서 시작하여 부산광역시 북구, 경상북도 경산시로 대상 표본 수를 확장해 나갈 때, 이전에 정상적으로 작동하던 코드를 그대로 적용하면 작동하지 않는 경우가 많았다. 이로 인해 데이터 처리 코드의 일반화 작업 난이도가 매우 높았다.

제 5 장

실증분석1 : 데이터 통합

경제총조사는 산업과 지역을 아우르는 우리나라 경제활동 전반의 구조적 특징을 파악하기 위한 핵심 공식 통계로서, 각종 정책 수립, 산업구조 변화 분석, 지역경제 진단 등에서 중요한 역할을 하고 있다. 특히 최근 사업체 규모·구조의 이질성 확대, 디지털 전환 가속화 및 AI 도입 등 경제 환경이 빠르게 변화함에 따라, 경제총조사가 포착해야 하는 정보의 범위와 깊이 역시 전통적 방식의 조사만으로는 충족하기 어려운 수준으로 복잡해지고 있다. 이는 조사 기반(survey-based) 접근만으로는 증가하는 정보 요구와 시의성을 동시에 충족하는 데 구조적 한계가 발생하고 있음을 의미한다. 특히 경제총조사의 특성 항목은 사업체의 주요 특성 외에도 AI 활용, 로봇 도입, 스마트공장 운영 여부 등 최근 산업정책에서 중요성이 높아진 항목들을 포함하고 있다. 그러나 이러한 정보는 단일 시점의 현장 조사만으로 수집하기 어렵고, 응답 부담과 비응답률 증가로 인해 일부 항목은 결측값이 대규모로 발생하는 문제도 존재한다. 이 때문에 경제총조사 특성 항목의 품질을 제고하기 위해서는 조사자료 중심의 기존 체계를 보완할 수 있는 새로운 자료원(raw data sources)을 활용하는 전략이 필요하다.

최근에는 공정거래위원회, 지자체, 공공데이터포털 등 다양한 기관에서 구축하는 행정·운영 자료와 민간 데이터가 빠르게 축적되고 있으며, 이 자료들은 특정 특성 항목과 직접적으로 연관된 정보를 세부 수준에서 제공할 수 있다. 이런 외부 자료들은 조사자료가 갖는 한계－시의성 부족, 응답 누락, 비용 및 조사 부담 증가－를 보완할 수 있는 잠재력이 있으며, 특히 특성 항목 중 일부는 외부 자료 기반으로 작성하거나 확인·대체하는 방식으로 통계 품질을 크게 향상할 수 있을 것으로 기대된다. 그러나 행정 및 외부 자료는 통계 작성 목적이 아니라 정책 집행, 사후 관리, 인허가, 사업 관리 등을 위해 구축된 자료이므로, 경제총조사 자료와 구조적·개념적 차이가 존재한다. 예컨대, 변수의 정의와 범위, 표기 방식, 포괄 범위, 시점 기준 등이 조사자료와 일관되지 않으며, 이러한 차이는 단순 병합(merging)으로는 해결할 수 없다. 따라서, 외부 자료를 경제총조사 특성 항목 작성에 활용하기 위해서는 자료 간 개념적·형식적 불일치를 해소하고, 통합이 가능한 구조를 확보하며, 통합 과정에서 발생하는 오류를 최소화하는 체계적인 절차가 필수적이다.

본 장에서는 이러한 문제의식을 바탕으로, 외부 자료와 경제총조사 시범예행조사 자료의 연계·통합을 실증적으로 제시한다. 구체적으로 공정거래위원회의 '통신판매업 자료'와 공공데이터포털의 '스마트공장 DB'를 활용하여, 경제총조사 시범예행조사 자료와 연계함으로써 외부 자료가 특성 항목의 보조 정보(auxiliary information) 또는 잠재적 대체 자료(alternative data source)로서 적정성을 갖는지를 평가한다.

제1절 데이터 통합 과정

데이터 통합(data integration)은 서로 다른 출처의 자료를 개념 및 구조적으로 표준화하고, 통합 목적에 맞게 결합하여 검증된 품질의 통계 산출물을 생산하는 체계적인 절차이다(김민규와 박성률, 2025). 이는 단순히 통계 산출물의 확장이라는 기술적 목표를 넘어, 자료 간 구조적 차이를 해소하고, 개념·범위·표기 방식에서 비롯되는 불일치가 통계에 미치는 영향을 구조적으로 최소화하는 것을 목적으로 한다. 다시 말해, 데이터 통합은 '두 자료를 연결하는 행위'가 아니라, 서로 다른 세계에서 태어난 자료를 하나의 일관된 논리적 세계로 번역하고 재맥락화하는 과정이라고 이해해야 한다.

데이터 통합을 위해서는 자료 간 차이를 올바르게 진단하고, 통계 목적에 맞는 조정 기준을 설정하며, 연계 과정에서 발생할 수 있는 오류를 최소화하는 방향으로 절차를 설계해야 한다. 이러한 맥락에서 데이터 통합 과정은 사전 점검 단계에서 시작하여, 구조 정비, 정합성 확보, 연계, 결측값 보완, 대표성 보정, 품질 검증으로 이어지는 체계로 구성된다. 데이터 통합 과정은 일반적으로 「전처리 및 정합성 점검 – 결합 – 결측값 대체 – 보정 및 대표성 점검 – 최종 품질 점검」으로 구성된다(김민규와 박성률, 2025). 이 절차는 순차적으로 진행되지만, 특정 단계에서 발견된 문제는 선행 단계로 되돌아가 조정해야 할 수도 있으며, 각 단계의 결과가 다음 단계의 품질과 타당성을 결정짓는 상호 의존적 구조를 가진다. 한편, 본 연구는 다양한 외부 자료를 활용하여 경제총조사 자료를 보완할 가능성을 검증하는 데 목적을 두고 있으므로, 데이터 통합의 핵심 기반인 전처리 및 연계를 중점적으로 다루었다.

1. 전처리

전처리(preprocessing)는 데이터 통합에서 가장 기초적이면서도 가장 중요한 단계이다. 서로 다른 자료원은 수집 목적, 구조, 변수 구성, 기재 방식 등이 서로 다르므로, 이러한 자료를 비교·연결·해석하기 위해서는 먼저 형식적·개념적 조정이 이루어져야

한다. 전처리 과정은 이러한 조정을 위해 자료를 정비하고 일관성을 확보하는 과정으로 자료 진단, 스키마 정렬, 개념 조화로 구성된다(ESCAP, 2020; UNECE, 2019).

가. 자료 진단

자료 진단(profiling)은 통합 대상이 되는 자료의 구조적 특성과 품질 상태를 파악하는 과정으로, 통합의 전제조건을 확립하는 단계이다. 이 과정에서 자료의 포괄 범위, 변수 구성, 중복 및 누락 여부, 이상값 분포 등의 기초 정보가 파악되며, 이를 바탕으로 자료 간 공통 영역을 정의하거나 조정이 필요한 요소를 식별하게 된다. 경제총조사 시범예행조사 자료와 외부 자료 모두 '사업체'를 공통 관측 단위로 하지만, 수집 경로와 구조적 형식의 이질성으로 인해 직접 결합은 불가능하다. 예를 들어, 시범예행조사 자료의 '사업체명' 변수는 사업자등록정보를 기반으로 관리되는 표준화된 값이지만, 외부 자료의 '사업체명'은 이용자 입력이나 플랫폼 표기 기준에 따라 특수문자, 괄호, 약어가 포함되는 형태로 나타날 수 있다. 또한, 주소의 경우 시범예행조사 자료는 도로명주소와 지번 주소로 구분되어 표준화되었지만, 외부 자료는 지번 주소, 도로명주소, 법정동·행정동 및 약식 표기가 혼재되어 동일 사업체임에도 다른 문자열로 인식될 수 있다. 이와 같은 구조적 불일치는 전처리 수준의 문제가 아니라, 후속 레코드 연계 단계에서 오연계 또는 누락의 주요 원인으로 작용한다. 이에 본 연구는 자료 진단에서 각 자료의 특성을 파악하고 불필요한 레코드를 정제하여, 스키마 정렬과 개념 조화의 범위를 설정함으로써 연계 과정에서 발생할 수 있는 오류 가능성을 사전에 최소화하고자 하였다.

나. 스키마 정렬

스키마 정렬(schema alignment)은 서로 다른 구조를 가진 자료를 기준틀(reference schema)에 맞추어 비교·대응시켜, 후속 단계에서 참조할 수 있는 변수 대응표(concordance table)를 구축하는 과정이다. 본 연구는 다음의 세 가지 원칙에 따라 스키마 정렬을 설계하였다.

- **기준 스키마 우선**. 경제총조사 시범예행조사 자료의 변수 체계를 기준틀(reference schema)로 설정하여 외부 자료의 변수를 이에 대응시켰다.
- **대응 관계 명시**. 각 자료의 변수명, 자료형, 범주 코드, 단위 등 구조 요소별로 일대일 대응 관계를 정의한다.
- **변경 이력 기록**. 모든 대응(mapping) 내용을 로그 형태로 저장해, 개념 조화에서 변환 및 표준화를 자동화할 수 있도록 준비한다.

다. 개념 조화

개념 조화(semantic harmonization)는 자료 간 변수의 의미와 정의를 공통 기준에 맞추어 변환·표준화함으로써 개념적 일관성을 확보하는 과정이다(ESCAP, 2020). 본 연구에서는 개념 조화를 개념 환산, 코드 변환, 형식 표준화, 문자 정규화, 결측값 처리로 구분하여 수행하였다.

- **개념 환산**. 서로 다른 기준으로 정의된 변수를 공통된 의미로 변환하는 작업으로 외부 자료에 포함된 특정 속성을 분석 목적에 맞게 재구성하였다. 예를 들어, 여러 유형이 존재하는 '통신판매방식'을 단일 속성인 '온라인 거래 여부'로 변환할 때, 인터넷 기반 판매 유형이 포함된 유형만을 '온라인 거래 = 예(1)'로 환산하였다. 개념 환산은 자료 간 의미적 호환성을 확보하는 데 핵심적이며, 이후 연계 단계에서 동일성 판단의 근거가 되는 주요 변수들이 동일한 해석·범주에 속할 수 있도록 기반을 마련한다.
- **코드 변환**. 변수의 범주값을 표준화하여 일관성을 확보하는 과정이다. 본 연구에서는 '예/아니오', 'Y/N', '1/0' 등 서로 다른 형태로 기재된 값들을 단일 코드 체계로 통일하였다. 예를 들어, 모든 이진형 변수는 '예 = 1', '아니오 = 2'의 구조로 통일하여 이후 연계 및 품질 분석에서 혼선을 방지하였다. 이는 비교, 연산, 조건 분기 등 다양한 분석 과정의 기반이 되며, 코드가 일관되지 않으면 발생할 수 있는 논리적 오류를 사전에 차단한다.
- **형식 표준화**. 텍스트 기반 변수의 구성요소를 명확히 분리하고 재구성하여 구조적 일관성을 확보하는 단계이다. 형식이 통일되지 않으면 연계 조건을 안정적으로 적용하기 어렵기 때문에, 주소와 같은 복합 텍스트는 반드시 표준화가 필요하다. 외부 자료의 주소는 지번, 법정동 주소 등 다양한 체계로 기재되어 있었기 때문에, 모든 주소를 도로명주소 체계로 변환하였다.[14)]
- **문자 정규화**. 표기 방식의 차이로 인해 동일한 개체가 서로 다르게 인식되는 문제를 해결하기 위한 절차이다. 대표적인 예로, 사업체명에서 '(주)', '주식회사', '㈜' 등 다양한 약칭과 표기들이 혼재되어 있었기 때문에, 이를 '주식회사'로 통일하였다. 대문자·소문자, 전각·반각, 특수문자 포함 여부 등 문자열 비교에서 오류를 유발할 수 있는 요소들도 모두 제거하거나 통일하였다. 문자 정규화는 연계 정확도를 높이기 위한 필수 단계이며, 결정적 연계(deterministic linkage)의 품질에 영향을 미친다.

---

14) 주소를 구성 요소별로 분리하여 「읍면동 / 도로명 / 건물 본번 / 건물 부번 / 건물명(빌딩·상가·아파트 등) / 동 / 층 / 호」의 구조로 통일하였다. 각 구성요소를 개별 변수로 분리함으로써 주소 비교가 세부 수준에서 가능해졌으며, 이후 단계에서 수행한 결정적 연계의 규칙을 설계하는 기반이 되었다.

- **결측값 처리**. 결측값은 이후 단계에서 연계 규칙이 정상적으로 작동하지 못하도록 만들거나, 품질 지표 산정 과정에서 오류를 유발할 수 있으므로, 전처리 단계에서 일정 기준에 따라 명시적 결측값(NA)을 부여하였다.

본 연구에서는 전처리 이후, 정합성 점검을 별도의 단계로 수행하지 않았다. 보다 구체적으로 말하면, 동일 변수를 기준으로 한 정의의 완전한 일치 여부, 기준시점과 참조 기간의 세밀한 조정, 관측 단위의 층위 간 차이(예: 사업체 · 법인 단위 등), 분류체계 간의 일대일 매핑 여부, 모집단 포괄 범위의 차이가 산출물에 미치는 영향 그리고 이러한 요소들을 반영한 분포 및 합계 수준의 통계적 정합성 점검을 정식 절차로 구현하지는 않았다.

첫째, 본 연구의 목적이 자료 간 정합성을 완전하게 확보하는 것이라기보다, 외부 자료와 조사자료 간 연계 가능성을 실증적으로 검증하는 데 있기 때문이다. 정합성 점검은 이상적으로는 연계 전 · 후를 모두 포괄하는 별도의 분석 과정을 필요로 하지만, 연계 가능성 자체가 충분히 확인되지 않은 상태에서 개념 · 시점 · 단위 수준의 정합성을 논의하는 것은 과도한 해석을 초래할 소지가 있다고 판단하였다.

둘째, 본 연구에서 활용한 자료의 범위와 구조가 정합성 점검을 체계적으로 수행하기에 충분히 정비된 상태가 아니었다. 외부 자료는 행정 · 운영 목적에 따라 수집된 자료로서, 조사자료와 비교할 수 있는 공통 변수의 수가 제한적이며, 각 변수에 대한 공식적인 메타데이터나 표준 정의가 완비되어 있지 않았다. 이와 같은 조건에서는 개념 · 시점 · 단위 · 분류·범위 정합성을 전면적으로 점검하더라도, 그 결과를 일반화하기 어렵고, 오히려 자료의 한계를 넘어서 과도한 결론을 도출할 위험이 있었다.

셋째, 정합성 점검을 형식적으로 수행하는 것 또한 지양할 필요가 있었다. 개념 · 시점 · 단위 · 분류 · 범위 · 통계적 정합성을 표면적으로 “점검했다.”라고 기술할 수도 있었으나, 메타데이터 및 준거 틀(framework)이 충분하지 않은 상황에서 그러한 표현을 사용하는 것은 연구의 투명성과 일관성 측면에서 바람직하지 않다. 본 연구에서는 정합성 점검을 독립 단계로 수행하지 않았음을 명시함으로써, 향후 후속 연구에서 정합성 분석을 위한 별도의 설계와 데이터 기반을 마련해야 한다는 점을 분명히 하고자 하였다.

이러한 이유로 본 연구는 전처리 및 개념 조화 단계에서 구조적 · 형식적 정비와 최소한의 의미 조정(개념 환산, 코드 변환, 형식 표준화 등)에 중점을 두었으며, 정합성 점검은 향후 외부 자료와 조사자료 간 통합 체계를 본격적으로 설계하는 후속 연구의 과제로 남겨두었다. 요약하면, 본 연구에서 수행한 데이터 통합은 전처리와 연계에 초점을 맞춘 기초 단계의 통합이며, 정합성 점검은 이러한 기초 작업 위에서 다음 단계에서 본격적으로 설계 · 적용되어야 할 과제로 인식하였다.

2. 연계

연계(record linkage)는 서로 다른 자료에 존재하는 동일 대상을 식별하여 단일한 분석 단위로 묶는 과정이다(Herzog et al., 2007; Winkler, 2006). 연계는 데이터 통합 과정에서 핵심적인 절차로, 연계 규칙의 설계와 적용 방식은 통합 데이터 세트의 품질을 근본적으로 규정한다. 연계 방법은 크게 고유식별자(unique identifiers)를 활용한 정확 연계(exact linkage)와 공통 변수(common variables)의 속성에 기반한 결정적 연계(deterministic linkage)로 구분되며, 본 연구에서는 두 방법을 병행하여 적용하였다.

연계 과정의 첫 출발점은 고유식별자를 활용한 정확 연계이다. 고유식별자는 동일성을 판단하는 데 가장 직접적이고 오류 가능성이 낮은 기준이며, 본 연구에서는 '사업자등록번호'가 이에 해당한다. 사업자등록번호가 두 데이터에 모두 정확히 기재되었을 때 이를 기준으로 단일한 일치 조건을 적용하여 동일 대상을 직접 연계하였다. 정확 연계는 논리적 구조가 단순하고 오연계 가능성이 낮다는 장점이 있으나(Christen, 2012), 실제 외부 자료에서는 사업자등록번호의 결측률이 높거나, 오류가 존재하는 경우가 많아 단독 기준으로 전체 대상을 포괄하기에는 한계가 있다. 따라서 정확 연계를 우선하여 적용하되, 고유식별자가 누락된 레코드에 대해서는 별도의 기준을 마련할 필요가 있었다.

고유식별자가 부재하거나 신뢰도가 낮은 경우에는 결정적 연계를 활용하였다. 결정적 연계는 공통 변수(사업체명, 대표자명, 주소 정보 등)를 조합하여 동일성을 판단한다(Christen, 2012). 이러한 공통 변수들은 전처리 단계에서 개념 환산, 형식 표준화, 문자 정규화 등을 거쳐 비교 가능한 형태로 재구성되었으며, 연계 규칙은 이러한 정제된 속성을 바탕으로 설계하였다. 특히, 주소는 도로명·지번 체계의 혼재 문제를 해결하기 위해 도로명주소 체계로 환산된 상태에서 비교하였으며, 주소의 구성요소인 읍면동, 도로명, 본번·부번, 건물명, 동·층·호를 개별적으로 평가할 수 있도록 구조화하였다. 결정적 연계의 동일성 판단에서 핵심은 문자열 유사도(string similarity)를 정량적으로 평가하는 것이다. 본 연구에서는 사업체명, 대표자명, 주소 비교를 위해 JW 유사도 점수(Jaro-Winkler similarity score)[15]를 활용하였다. JW 유사도는 두 문자열이 동일하거나 유사할 경우 명칭 표기 방식 차이, 철자 변형, 공백 등에서 발생하는 미세한 차이를 정량

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

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

  L: 접두사 길이, p: scaling factor(보통 0.1)

적으로 반영할 수 있다(김민규와 박성률, 2025). JW 점수는 단순 문자열 일치 여부가 아닌 유사한 정도를 기반으로 판단을 가능하게 하여, 고유식별자가 없더라도 합리적 수준의 동일성 판단을 할 수 있게 해준다.

본 연구에서는 결정적 연계 규칙을 JW 유사도 점수에 근거한 규칙 기반 결정(rule-based decision)으로 설계하였다. 우선, 사업체명의 JW 점수가 일정 기준 이상일 때 일치 후보로 판정하고, 주소 구성요소의 일치 여부를 추가 조건으로 사용하였다. 주소의 상위 구성요소가 일치하고 본번·부번 또는 건물명 등의 세부 요소가 일정 범위 내에서 일치할 경우, 두 레코드를 동일 개체(same entity)로 판단하였다. 반대로, JW 점수는 높으나 주소 구성요소가 크게 다를 경우 오연계(false match)로 판단하여 배제하였다. 이처럼 문자 유사도 기반 조건과 주소 기반 조건을 결합함으로써, 누락(false non-match)과 오연계의 위험을 균형 있게 관리할 수 있도록 설계하였다.[16)]

3. 연계 품질 점검

연계 품질 점검(linkage quality assessment)은 연계 결과의 타당성을 검증하고 연계 규칙의 적절성을 평가하기 위한 절차이다. 품질 진단은 기본적으로 연계율(match rate), 누락률(false non-match rate), 오연계율(false match rate), 정확도(accuracy), 정밀도(precision) 등 정량적 지표를 중심으로 수행하지만, 본 연구에서는 이러한 정량적 평가뿐 아니라 수작업 검토(clerical review)를 병행하여 연계 품질을 점검하였다.

연계율은 전체 레코드 중 실제로 연결된 비율을 의미하며, 연계 규칙의 적용 강도와 자료 간 구조적 이질성에 따라 달라진다. 그러나 연계율이 높다고 해서 곧바로 품질이 우수하다고 판단할 수 있는 것은 아니며, 오연계와 누락이 동시에 발생하는지를 평가해야 한다. 오연계율은 서로 다른 레코드를 잘못 동일 대상으로 판단한 비율로, JW 점수 임곗값이 지나치게 낮거나 주소 전처리 수준이 불충분할 때 높아질 수 있다. 반대로 누락률은 동일 대상을 서로 다른 레코드로 판단한 비율로, 연계 규칙이 너무 엄격하게 설정되었을 때 증가한다. 정밀도는 일치로 분류된 레코드 쌍 중 실제로 일치한 비율을 의미하며, 연계가 실제 동일 대상을 반영하는지를 확인하는 중요한 지표이다.

정량적 지표 외에도, 연계 규칙의 타당성을 보완하기 위해 수작업 검토를 체계적으로 시행하였다. 자동화된 규칙 기반 연계는 표기 방식의 변형, 주소의 불완전성, 문자열 정비 과정에서 발생하는 예외적 패턴 등을 완전히 포착하기 어렵기 때문에, 수작업

16) 결정적 연계는 공통 변수의 품질에 민감하기에, 앞서 수행된 개념 조화가 연계 규칙의 신뢰성을 확보하는 데 결정적인 역할을 한다. 예를 들어, ‘(주)◐◑◒’과 ‘주식회사 ◐◑◒’을 단일 표준으로 정규화하거나, 도로명 주소를 구성 요소별로 분리하는 과정은 JW 유사도 점수 산출과 주소 비교 규칙의 일관성을 확보하는 데 필수적이었다.

검토는 이러한 기계적 한계를 보완하는 데 핵심적인 역할을 한다(김민규와 박성률, 2025). 수작업 검토는 다음 단계로 수행되었다.

첫째, 연계된 레코드와 미연계 레코드 중 일정 비율을 무작위로 추출하여 표본을 구성하였다. 표본은 자료의 분포 구조를 반영하도록 계층화된 방식으로 구성하여 편향을 최소화하였다.

둘째, 표본 레코드에 대해 본 연구진이 독립적으로 동일성 여부를 판단하였다. 각 연구자는 사업체명, 대표자명, 주소 구성요소 등 연계에 활용된 모든 속성을 검토하여, 자동 연계 결과가 적절했는지 독립적으로 판단하였다.

셋째, 연구자 간 판단이 일치하지 않은 레코드를 '불일치 사례'로 분류한 뒤, 공동 검토 및 논의 절차를 거쳐 오연계 또는 누락의 원인을 분석하였다. 이 과정에서 문자열 정규화 방식의 한계, 주소 구성요소 정제의 불완전성, JW 점수의 임곗값에서 발생하는 모호 사례 등을 검토하였다.

넷째, 수작업 검토 결과를 바탕으로 JW 유사도 임곗값과 주소 일치 조건의 조합 기준을 재조정하였다. 예를 들어, JW 점수가 임곗값에 근접했음에도 불일치가 반복적으로 나타나면 해당 점수 구간의 임곗값을 상향 조정하고, 반대로 JW 점수는 낮지만, 일치성이 강한 패턴이 안정적으로 발견되면 가중값을 조정하는 방식으로 규칙을 보완하였다.

이와 같은 수작업 검토는 규칙 기반 연계 방식의 한계를 정교하게 보완하고, 자동 연계 결과의 신뢰성을 높이는 역할을 하였다. 또한, 연계 품질 지표의 해석을 정교화하는 데 필요한 경험적 근거를 제공하였다. 품질 진단은 단순한 결과 검증이 아니라, 연계 규칙과 전처리 기준을 지속적으로 개선하는 순환적 구조의 핵심 요소라고 할 수 있다.

4. 결측값 대체 및 보정 제외 이유

데이터 통합에는 결측값 대체, 보정 및 대표성 점검 등이 포함된다. 그러나, 본 연구에서는 이러한 단계들을 독립적으로 수행하지 않았다. 이는 분석의 범위를 축소하기 위한 선택이 아니라, 자료의 구조적 특성과 본 연구의 목적을 고려할 때 결측값 대체와 보정 절차를 적용하는 것이 적절하지 않다는 판단에 근거한 것이다.

첫째, 본 연구의 목적은 외부 자료가 조사자료와 연계 가능한지 그리고 어떤 품질 특성을 보이는지를 실증적으로 검증하는 데 있다. 연계 품질이 충분히 확보되지 않은 상태에서 결측값 대체를 수행할 경우, 신뢰도가 낮은 정보가 혼입되어 오히려 데이터의 정확성을 훼손할 수 있다. 결측값 대체는 속성 간 구조적 관계를 기반으로 이루어져야 하는데, 연계 품질이 안정적이지 않을 때 대체 값을 생성하면 불확실성이 증가할 뿐만 아니라, 편향이 데이터 전체에 확산될 수 있다고 판단하였다. 특히 본 연구에서

수행한 연계는 목적상 연계 가능성 검증에 해당하며, 결측값 대체가 적용될 수 있는 전제조건인 '속성값 간 관계의 안정성'이 충분히 확인되지 않았다.

둘째, 보정 및 대표성 점검은 모집단 수준의 대표성을 체계적으로 회복하기 위한 절차이지만, 본 연구에서 사용한 자료는 이러한 절차를 적용하기에 적합한 구조를 갖추지 못하였다. 보정(calibration)은 일반적으로 조사설계 정보(층화, 집락, 가중치 등), 모집단 정보, 충분한 표본 규모, 특정 속성의 사전분포 등 다양한 참조 정보가 필요하다. 그러나, 외부 자료는 통계 작성이 아닌 행정·운영상의 필요로 구축된 자료로서 이러한 설계 정보를 제공하지 않으며, 포괄 범위 역시 모집단 전체와 일치하지 않는다. 즉, 대표성 기반 조정 절차가 작동하기 위한 조건 자체가 충족되지 않았기 때문에 보정을 함부로 적용하면 왜곡을 초래할 위험이 있다.

셋째, 결측값 대체와 보정은 단일 연구 내에서 한 번 수행하고 끝나는 작업이 아니라, 장기간의 자료 축적과 반복적 검증을 통해 안정화되는 절차라는 점을 고려하였다. 예를 들어, 보정 절차를 정교하게 적용하기 위해서는 자료별 구조적 차이(예: 포괄 범위, 기재 방식, 시점 기준)를 일관된 기준으로 설명할 수 있는 메타데이터 체계가 구축되어야 하고, 결측값 대체는 자료의 분포 특성을 기반으로 수행해야 한다. 본 연구는 외부 자료 통합을 위한 기초 단계이며, 보정 기준이 존재하지 않는 상황에서 결측·보정 절차를 적용하는 것은 연구 목적에 부합하지 않는다.

넷째, 국제기구에서도 연계 품질 확보가 결측값 대체 및 보정보다 선행되어야 한다는 원칙을 명시하고 있다(UNECE, 2020; ESCAP, 2020). 즉, 자료 간 구조적 차이가 큰 초기 단계에서는 연계 규칙의 정밀화와 품질 검증을 선행해야 한다고 권고한다. 이는 초기 통합 단계에서 지나치게 적극적인 대체 및 보정 절차를 적용할 경우, 자료 간 구조적 불일치가 그대로 편향 요인이 될 수 있다는 문제를 지적한 것이다. 본 연구가 결측 대체와 보정을 수행하지 않은 것은 이러한 국제적 기준과도 부합하는 합리적 접근이다.

다섯째, 본 연구에서 활용된 자료는 일부 지역·특정 모집단의 부분 집합을 대상으로 하고 있어 대표성 기반 보정의 통계적 전제조건을 충족하기 어렵다. 표본 규모가 제한적이거나 특정 범주가 과소·과대 포함된 자료에서는 보정 절차가 오히려 오류를 확대할 수 있다. 결측값 대체 역시 동일한 문제를 가진다. 자료의 포괄 범위가 충분히 검증되지 않았고 속성 간 관계도 안정적으로 파악되지 않은 상황에서 대체 값을 생성하면, 결측을 보완하는 것이 아니라 자료의 구조적 편향을 강화하는 결과를 초래할 수 있다.

향후 외부 자료의 데이터 통합을 위한 장기적 운영 체계가 구축되고, 자료 간 개념·단위·시점 구조에 대한 정합성 분석이 축적되며, 모집단 틀(frame)이 일관되게 확보될 때 결측값 대체와 보정 절차를 본격적으로 적용할 수 있을 것이다. 따라서, 본 연구가 수행한 단계는 외부 자료 통합을 위한 기초 기반을 마련하는 단계로서 기능하며, 이후의 고도화 과정에 필요한 근거를 축적하는 데 중요한 역할을 한다.

제2절 사례 1. ‘온라인 거래 여부’ 항목 보완

1. 자료 진단

본 절에서 활용한 자료는 세 가지이다. 첫째, 공정거래위원회가 관리하는 통신판매업 자료는 통신판매업 신고를 기반으로 구축된 외부 행정자료이다. 둘째, 경제총조사 표본 명부자료는 경제총조사의 조사 틀(survey frame)로 활용되는 사업체 명부이다. 셋째, 경제총조사 시범예행조사 자료는 현장 조사를 통해 수집된 자료이다.

각 자료는 수집 목적과 관리 체계, 처리 방식, 변수 정의가 다르므로, 이를 하나의 분석 단위로 연계하기 위해서는 사전 진단을 통해 자료 간 이질성을 구조적으로 이해해야 한다. 또한, 어떤 레코드가 어떤 사유로 제외되었는지를 충분히 이해하지 못한 상태에서 연계를 수행하면, 비활동 사업체가 연계에 포함되거나, 조사 대상에서 이미 제외된 단위가 다시 분석에 사용되는 등 구조적인 오류가 발생할 수 있다.

이에 자료별로 전체 규모와 제외 규모를 제시하고, 제외 사유를 정리함으로써 이후 전처리 및 연계 절차의 전제를 명확히 하고자 한다. <표 5-1>에서 확인할 수 있듯이, 세 자료 모두 상당 비율의 레코드가 정제 단계에서 제외되었다. 중요한 것은 단순한 제외 비율 자체가 아니라, 어떤 기준과 사유에 의해 이루어졌는가 하는 점이다.

<표 5–1> 자료별 규모 및 제외 현황

(단위 : 개. %)

지역	자료 구분	총 레코드(개)	제외 레코드(개)	제외 비율(%)	정제 후 활용 레코드(개)
부산 북구	공정위 통신판매업	5,925	2,534	42.8	3,391
	경총 표본 명부	23,250	7,455	32.1	15,795
	경총 시범예행조사	9,837	3,423	34.8	6,414
경북 경산시	공정위 통신판매업	10,769	3,094	28.7	7,675
	경총 표본 명부	34.844	12.893	37.0	21,948
	경총 시범예행조사	16,418	7,202	43.9	9,216

가. 공정거래위원회 통신판매업 자료

통신판매업 자료는 신고를 기반으로 구축된 행정자료로서, 신고 이력 전체를 포괄한다는 점이 가장 큰 특징이다. 이는 곧 ‘현재 활동 중인 사업체 목록’이라기보다는 ‘과거부터 현재까지의 신고 이력 집합’이라는 의미를 지닌다. 자료 진단 결과, 휴업·폐업·직권취소·직권말소로 비활동 상태로 분류된 사업체가 상당한 비중을 차지하였다.

<표 5-2>에 나타난 바와 같이, 부산 북구의 경우 전체 5,925건 중 2,528건이 비활동 사업체로 분류되어 제외되었고, 경북 경산시의 경우 10,769건 중 3,063건이 동일한 사유로 제외되었다. 이들 비활동 사업체는 더 이상 경제활동 단위로 간주하기 어렵기 때문에, 시범예행조사 자료 및 표본 명부자료와의 연계 대상에 포함하면 연계율과 속성 일치율을 왜곡할 위험이 크다.

따라서 본 연구에서는 비활동 상태로 표기된 사업체를 일괄적으로 제거하고, 이후 연계는 '현재 활동 중인 것으로 간주할 수 있는 신고 사업체'로 한정하였다. 비활동 상태 이외에도 '통신판매방식' 변수의 결측과 지역 불일치가 일부 확인되었다. 통신판매방식은 본 연구에서 검증하고자 하는 특성 항목인 '온라인 거래 여부'와 직접적으로 대응되는 변수이므로, 이 값이 결측인 레코드 역시 제거하였다. 또한 주소 기준으로 부산 북구 또는 경북 경산시에 속하지 않는 사업체 역시 제외하였다.

<표 5-2> 통신판매업 자료의 제외 사유별 현황

(단위 : 개)

제외 사유	제외 레코드(개)		비고
	부산 북구	경북 경산시	
휴업·폐업·직권취소·직권말소	2,528	3,063	이력은 있으나 현재 활동 중인 사업체 아님
통신판매방식 변수 결측	3	31	핵심 속성 부재로 특성 항목 비교 불가능
타 시군구	3	0	

신고에 기반한 통신판매업 자료는 사업자등록번호와 주소의 기록 품질이 불완전한 경우가 적지 않다. 이 두 변수는 연계에서 핵심적인 기준이 되므로, 결측이 존재하면 어떤 연계 방법을 선택할 것인지에 대한 전략적 판단이 필요하다.

<표 5-3>에 나타난 바와 같이 두 변수의 결측률은 매우 낮다. 그러나, 결측률이 낮더라도 사업자등록번호는 정확 연계를 위한 유일한 고유식별자이므로, 해당 레코드에는 정확 연계를 적용할 수 없다. 또한, 주소 정보는 통신판매업 자료에서 가장 심각하게 비표준화된 변수로 결측률 자체는 낮지만, 주소 구성요소의 부분 결측 및 비표준 구조는 매우 광범위하게 존재한다.[17] 이는 연계 시 비교성(comparability)을 저해하는 요인이다. 그러나 주소는 사업체명, 대표자명과 함께 결정적 연계를 위한 필수 요소이기에 결측 레코드를 배제하지 않고, 개념 조화를 통해 연계가 가능해지게 하였다.

17) 통신판매업 자료의 주소는 읍면동 및 도로명, 건물명(아파트, 빌딩, 시장, 상가 등) 일부만 제공하고 있다. 또한, 지번·도로명 혼합, 행정동·법정동 혼용, 숫자·기호의 비표준화된 표기 등이 다수 존재한다.

<표 5-3> 통신판매업 자료의 결측 현황

(단위 : 개, %)

지역	활용 레코드(개)	변수명	결측 레코드(개)	결측률(%)
부산 북구	3,391	사업자등록번호	1	0.03
		주소	4	0.12
경북 경산시	7,675	사업자등록번호	22	0.29
		주소	24	0.31

나. 경제총조사 표본 명부자료

표본 명부자료는 조사표별로 포함되는 항목이 다르므로 특정 특성 항목의 결측은 단순한 누락이 아니라 조사표 설계에 따라 항목 자체가 존재하지 않는 구조적 결측(structural missingness)임을 유의해야 한다. 우선 전체 명부 규모와 정제 후 활용 레코드를 기준으로 제외 사유를 분류하였다.

<표 5-4> 표본 명부자료의 제외 사유별 현황

(단위 : 개)

제외 사유	제외 레코드(개)		비고
	부산 북구	경북 경산시	
비대상 조사표	7,237	12,811	특성 항목이 수집되지 않는 조사표 유형
중복	218	80	동일 사업체의 이중 등재 제거
타 시군구	0	3	

<표 5-4>는 명부자료의 제외 현황을 나타낸다. 표본 명부자료에서 가장 큰 제외 사유는 조사표별 구조적 제외이다. 산업대분류 및 종사자 규모에 따라 서로 다른 조사표가 설계된다. 예를 들어, '조사표 2'는 광업(B)과 종사자 9인 이하의 제조업(C), '조사표 3'은 광업(B)과 종사자 10인 이상의 제조업(C)이 대상이다. 해당 조사표 유형에서 수집하지 않는 특성 항목이 존재할 경우, 해당 레코드는 분석에서 제외된다. 부산 북구의 경우 연계 대상이 '온라인 거래 여부' 특성 항목으로 구성되지 않았기에 7,237개의 레코드를 제외하였다. 경산시 역시 동일한 구조로 12,811개가 제외되었다.

두 번째, 제외 사유는 중복 등재이다. '사업자등록번호 + 사업체명 + 대표자명 + 주소 + 전년도산업대분류' 조합을 기준으로 중복 레코드를 제거한 결과, 부산 북구에서는 218개, 경북 경산시에서는 80개가 제외되었다. 이는 명부자료 내 발생할 수 있는 이중 등재 문제를 해소하여 연계 과정에서 발생할 수 있는 이중 매칭(duplicate linkage)을 예방하는 데 필수적이다.

마지막으로 주소 불일치의 경우, 경산시에서 3개 사례가 확인되었는데, 이는 다른

지역 사업체가 경산시에 포함된 사례로 판단되어 제외하였다.

명부자료는 표본 틀 성격을 갖기 때문에, <표 5-5>에 나타난 바와 같이 결측률이 매우 낮다. 이는 표본 명부자료가 연계 기준틀(reference frame)로 활용될 수 있는 중요한 근거이다.

<표 5–5> 표본 명부자료의 결측 현황

(단위 : 개, %)

지역	활용 레코드(개)	변수명	결측 레코드(개)	결측률(%)
부산 북구	15,795	사업자등록번호	632	4.00
		대표자명	15	0.09
경북 경산시	21,948	사업자등록번호	430	1.96
		사업체명	3	0.01
		대표자명	12	0.05

다. 경제총조사 시범예행조사 자료

시범예행조사 자료는 현장을 방문해 조사한 자료로서, 사업체의 실제 활동 상태와 응답 가능성을 가장 정확히 반영한다.

<표 5-6>은 시범예행조사 자료의 제외 사유별 현황으로 폐업·휴업·전출 등은 현장에서 조사원이 확인한 정보로 행정자료나 명부자료에서 포착하기 어려운 실질적 자료이다.

<표 5–6> 시범예행조사 자료의 제외 사유별 현황

(단위 : 개)

제외 사유	제외 레코드(개)		비고
	부산 북구	경북 경산시	
비대상 조사표	2,462	6,090	특성 항목이 수집되지 않는 조사표 유형
휴·폐업	438	582	
전출(타 지역)	83	373	
기타	440	153	이중등재, 흡수합병, 주소불명확, 제외업종
타 시군구	0	4	

시범예행조사 자료는 현장 기반 조사라는 특성상 응답자 부재, 불응 등 조사 현실에 기인한 결측이 광범위하게 발생한다. 특히 본 연구에서 연계·검증하고자 하는 특성 항목은 응답 기반 항목이기 때문에, 해당 항목의 결측은 단순 누락이 아니라 조사에 따른 구조적 제약에서 비롯된 것이다.

정제 후 남은 레코드를 대상으로 결측을 확인한 결과, <표 5-7>에 나타난 바와 같이 온라인 거래 여부 항목에서 많은 결측이 나타났다.

<표 5–7> 시범예행조사 자료의 결측 현황

(단위 : 개, %)

지역	활용 레코드(개)	변수명	결측 레코드(개)	결측률(%)
부산 북구	6,414	사업자등록번호	255	3.95
		온라인 거래 여부	2,379	37.09
경북 경산시	9,216	사업자등록번호	177	1.92
		대표자명	15	0.16
		온라인 거래 여부	5,000	54.25

‘온라인 거래 여부’ 변수의 높은 결측률은 몇 가지 중요한 사실을 의미한다.

- **시범예행조사 자료는 특성 항목의 완전 원자료가 아니다.**

  온라인 거래 여부는 응답 기반으로 수집되는 항목으로 응답자 부재 및 불응 등 조사 현실 제약이 발생하면 값이 입력되지 않는다. 따라서, 시범예행조사 자료는 특성 항목의 기준자료(gold standard)가 될 수 없다.

- **외부 자료 활용의 필요성이 높아진다.**

  부산 북구의 37%, 경북 경산시의 54%라는 결측률은 시범예행조사 단독으로 온라인 거래 항목을 안정적으로 작성하는 것은 불가능하며, 외부 자료와의 연계 또는 대체가 필요하다는 점을 시사한다.

- **연계 품질에 영향을 준다.**

  이처럼 높은 결측률이면 통신판매업 자료와 연계 후, 온라인 거래 여부 비교를 수행해도 일치율이 왜곡되거나, 비교 가능한 레코드의 수가 급감하며, 속성 불일치가 과대 추정될 수 있다.

2. 스키마 정렬

경제총조사 시범예행조사 자료를 기준 스키마(reference schema)로 설정하고, 경제총조사 표본 명부자료와 공정거래위원회 통신판매업 자료의 변수 체계를 이에 대응하였다. 이는 세 자료의 품질 수준, 수집 목적, 변수의 의미적 안정성을 고려한 선택으로, 기준 스키마를 중심으로 자료 구조를 재배치함으로써 자료 간 비교 및 연계가 가능한 분석 단위를 확보하기 위함이다.

스키마 정렬은 다음의 세 단계로 진행하였다.

- 각 자료의 변수명, 자료형(data type), 범주 코드(category codes), 단위(units), 값의 표현 방식(representation)을 체계적으로 비교하였다. 변수명은 동일 변수임에도 자료원에 따라 'BRNO ― CRN ― 사업자등록번호'처럼 다르게 표기되어 있었으며, 자료형 또한 문자형 · 수치형 · 범주형이 혼재되어 있었다. 주소 정보의 경우 통신판매업 자료는 지번 · 도로명 · 약식주소가 혼재되어 있어 동일 사업체임에도 전혀 다른 문자열로 기록되는 문제가 확인되었다. 이러한 불일치는 연계 과정에서 비교 불가능성을 초래하므로, 스키마 정렬에서 반드시 조정해야 한다.
- 일대일 대응(mapping) 관계를 정의한 후, 각 자료의 주요 변수를 대응시켰다. 예를 들어, 사업자등록번호는 시범예행조사(BRNO, 문자형), 표본 명부(CRN, 문자형), 통신판매업(사업자등록번호, 수치형) 자료를 동일한 의미의 변수로 정의하고, 모두 문자형으로 통일하였다. 사업체명, 대표자명, 도로명주소 또한 동일한 구조로 대응시켰다.
- 정렬과 변환의 전 과정을 변경 이력(change log)으로 기록한 후, <표 5-8>에 제시된 대응표(concordance table)를 구축하여 개념 조화 단계에서 자동화될 수 있도록 준비하였다.

3. 개념 조화

개념 조화는 단순한 값 변환이 아니라 자료 간 의미적 호환성(semantic interoperability)을 확보하기 위한 개념적 통합 과정이며, 연계 단계에서 동일 사업체를 판단하기 위한 주요 변수의 해석을 일관화하는 단계이다.

가. 개념 환산

개념 환산은 외부 자료에 포함된 특정 속성을 분석 목적에 맞게 다시 정의하는 과정이다. 공정거래위원회의 '통신판매방식'이 다중 범주로 구성되어 있었기 때문에, 이를 단일 속성인 '온라인 거래 여부'로 환산하였다. 구체적으로, 인터넷 기반 판매 유형이 포함된 모든 유형을 '온라인 거래 = 예(1)'로 변환하고, 인터넷 판매가 전혀 포함되지 않은 유형은 '아니오(2)'로 재분류하였다. 통산판매업 자료는 '인터넷', '카탈로그', '전화', '기타' 등 다양한 조합형 응답을 포함하고 있었으며, 문자열 기반이라 동일 범주라도 다른 방식으로 기재되는 경우가 많았다. 따라서, 개념 환산 단계에서는 '인터넷'이 포함된 모든 판매 방식(예: '인터넷, 기타', '인터넷 · 신문')을 포괄적으로 인식할 수 있도록 규칙 기반 변환(rule-based transformation)을 수행하였다. 이는 연계 후 비교 가능한 특성 항목을 확보하기 위한 필수 작업이다.

<표 5-8> 주요 변수별 대응표

변수	경제총조사 시범예행조사 자료		경제총조사 표본 명부 자료		공정거래위원회 통신판매업 자료		
	변수명	자료형 (표기)	변수명	자료형 (표기)	변수명	자료형 (표기)	비고
사업자등록번호	BRNO	문자형	CRN	문자형	사업자등록번호	수치형	3-2-5형식 (예: 123-45-67890)
사업체명	BZENT_NM	문자형	ESTM_NM	문자형	상호	문자형	표기 불일치
사업체 대표자명	BZENT_RPRSV_NM	문자형	ESTM_RPRS_NM	문자형	대표자명	문자형	—
소재지 도로 읍면동	LCTN_ROAD_EM_NM	문자형	LOC_ROD_EMD_NM	문자형	소재지	문자형	행정동·법정동 혼용
소재지 도로명	LCTN_RDN	문자형	LOC_ROD_NM	문자형			도로명·지번 혼용
소재지 도로 빌딩시장상가명	LCTN_ROAD_BMASC_NM	문자형	LOC_ROD_BMASC_NM	문자형			표기 불일치
온라인 거래 여부 코드	ONLN_DLNG_YN_CD	범부형	—	—	판매방식	문자형	다중응답형 비정형 범주

나. 코드 변환

일관되지 않은 범주형 값은 비교·연산 단계에서 오류를 발생시키므로, 코드 변환은 개념 조화의 핵심 절차 중 하나이다. '예/아니오(Y/N)', '1/0', 문자열 기반 응답이 혼재되어 있어 모든 이진형 값을 '예=1, 아니오=2'의 통일된 코드 체계로 변환하였다. 이는 통신판매업 자료에서 '판매방식'의 비정형 응답을 범주화하는 데 중요한 역할을 하였다. 코드 변환의 목적은 단순 통일이 아니라 논리적 오류를 사전에 차단하는 데 있다. 범주 번호가 자료원마다 다르면, 동일 범주라 하더라도 연계 시 오해석될 수 있기 때문이다.

다. 형식 표준화

주소 정보와 같은 복합 텍스트 변수는 일관된 구조가 확보되지 않으면 연계 규칙을 안정적으로 적용할 수 없다. 통신판매업 자료는 행정동·법정동, 지번·도로명·약식주소가 혼재되어 있어 동일 사업체라도 서로 다른 문자열로 인식되는 사례가 발생할 수 있다. 이에 본 연구에서는 모든 주소를 「읍면동 / 도로명 / 건물 본번 / 건물 부번 / 건물명(아파트·빌딩·상가 등) / 동 / 층 / 호」의 구성요소로 분리하여 표준화하였다. 주소 구성요소를 개별 변수로 분리하면, 이후 결정적 연계에서 주소 비교가 층위별로 가능해진다. 예컨대, '가일길 99−3'과 '가일길 99'와 같은 유사 주소의 비교에서도 더 높은 정확도를 확보할 수 있다.

라. 문자 정규화

문자 정규화는 동일 개체가 문자열 표기 차이로 인해 다른 개체로 인식되는 문제를 제거하기 위한 절차이다. 대표적으로 사업체명에서 '(주)●■▲', '주식회사 ●■▲', '㈜●■▲' 등 다양한 표기가 존재하였기 때문에, 이를 모두 '주식회사 ●■▲'의 형식으로 통일하였다. 또한, 대문자·소문자, 전각·반각, 공백·특수문자 유무 등 문자열 비교 오류를 유발할 수 있는 요소를 모두 표준화하였으며, 통신판매업 자료에서 특히 심했던 약식 상호 표기나 특수문자 포함 문제 또한 모두 정규화하였다.

마. 결측값 처리

결측값은 연계 규칙이 정상적으로 작동하지 못하게 하며, 품질 지표 산정 과정에서도 오류를 유발할 수 있기 때문에 일관된 기준에 따라 명시적으로 처리하였다. 즉, ①기재가 불완전하거나 모호한 경우 'NA' 부여, ②주소 구성요소 등 일부 정보가 누락되어 비교가 불가능한 경우 'NA' 부여, ③구조적 결측(예: 명부·조사표 유형에 따라 항목 자체가 존재하지 않는 경우)은 별도 플래그(Flag=1) 처리하여 분석 시 동일 수준에서 비교되지 않도록 하였다.

4. 연계 및 품질 점검

연계(record linkage)는 일련의 독립된 작업의 단순한 나열이 아니라, 각 단계가 이전 단계의 결과를 기반으로 다음 단계의 신뢰성을 결정하는 계층적 구조(hierarchical process)를 가진다. 본 연구에서는 공정위 통신판매업 자료, 경총 표본 명부 및 시범예행조사 자료를 통합하는 과정에서 발생할 수 있는 오류를 최소화하고 비교 가능성을 확보하기 위하여, 연계 절차를 세 단계로 구성하였다. 각 단계는 자료의 구조적 이질성을 해소하고, 동일성 식별의 근거를 점차 강화하는 방식으로 설계되었다.

- **잠재적 연결 대상 발굴**(candidate generation).

  고유식별자 기반의 정확 연계가 일부에서만 가능하므로, 실제 동일 사업체임에도 고유식별자 부재·주소 불일치·표기 방식 차이 등으로 인해 연계에서 누락되는 레코드가 존재한다. 이를 보완하기 위해 사업체명·대표자명·주소 간의 구조적 유사성을 바탕으로 잠재적 동일성을 가진 조합을 폭넓게 확보하였다. 후보군 생성은 무작위적 확대가 아니라, 정제된 문자열 기반으로 동일성 가능성을 체계적으로 탐색하여 이후 단계에서 정밀한 판단을 가능케 한다.

- **연계 기준에 따른 최종 연결 판단**(deterministic linkage).

  연계 후보군 중 동일 사업체일 가능성이 높은 조합에 대해 사업체명·대표자명·주소를 종합적으로 검증한다. 기준 설계의 핵심은 보수성이다. 사업체명이 유사하더라도 대표자명이 일치하지 않거나, 읍면동이 다르면 서로 다른 사업체로 판정하였다. 반대로 일부 주소의 부번·건물명 등 비핵심 요소에서 차이가 있더라도 읍면동과 도로명, 본번이 일치하면 동일 사업체로 인정하였다. 이 계층적 판단 방식은 오연계 위험을 구조적으로 차단하는 기능을 한다.

- **연계 품질 점검**(quality assessment).

  단순히 연계 결과를 확인하는 수준을 넘어 연계율과 누락률, 오연계율, 속성(온라인 거래 여부) 일치율, 정확도·정밀도·재현율 등을 종합적으로 평가하여, 연계 과정 전반의 품질을 검증하였다. 품질 점검은 자료 구조의 제약으로 인한 한계를 명확히 드러내는 동시에, 외부 자료(통신판매업 자료)가 조사자료(표본 명부 및 시범예행조사 자료)의 보조 자료로 활용될 수 있는지 판단하는 근거를 제공한다.

이와 같은 계층적 연계구조는 자료 간 차이를 인정하고 그 차이를 조정해 나가는 과정이다. 각 단계가 상호 유기적으로 연결되어 있기 때문에, 어느 하나의 단계라도 미흡할 경우 전체 연계 결과의 신뢰성이 훼손될 수 있다. 본 연구는 이러한 위험을 최소화하기 위해 각 단계의 목적과 기능을 명확히 설정하고, 실제 자료 특성에 기반한 보

수적 판단을 통해 최종 통합 데이터 세트를 구축하였다.

앞서 언급한 것처럼, 본 연구는 “외부 자료가 특성 항목 보완에 실제로 활용 가능한지”를 판단하는 데 필요한 핵심 단계에 연구 자원을 집중하였다. 이때 중요한 방법론적 결정은 연계 기준자료(reference dataset)를 무엇으로 설정할 것인가이다. 본 연구에서는 기준자료를 공정거래위원회 통신판매업 자료로 설정하였다. 이는 단순한 기술적 선택이 아니라, 본 연구의 분석 목표에 부합하는 개념적·방법론적 정당성을 갖는 결정이다. 우선 통신판매업 자료는 통신판매업 신고를 기반으로 구축되기 때문에, 경제총조사 특성 항목 중 ‘온라인 거래 여부’를 보완할 수 있는 자료이다. 따라서, 외부자료 활용 가능성을 평가하려면, “통신판매업 신고 사업체가 경제총조사에서 어느 정도 포착되는지”를 확인하는 것이 가장 자연스럽고 적절한 접근이다. 이는 통신판매업 자료를 기준으로 ‘통신판매업＋시범예행조사’, ‘통신판매업＋표본 명부’의 연계를 수행해야 연계율을 통해 “외부 자료가 조사자료를 어느 정도 보완할 수 있는가?”를 직접적으로 확인할 수 있기 때문이다. 반대로 ‘시범예행조사＋통신판매업’ 또는 ‘표본 명부＋통신판매업’ 연계를 설정할 경우, 연계율은 ‘조사자료 전체 중 통신판매업 신고 사업체의 비율’이 되기 때문에 외부 자료 활용 가능성을 평가하는 데 의미 있는 지표가 될 수 없다. 따라서 본 연구는 통신판매업 자료를 기준으로 하여 시범예행조사 자료와 표본 명부자료 간의 연계를 단계적으로 수행하였고, 이를 통해 외부 자료 기반 특성 항목 보완 가능성을 실증적으로 검증하였다.

가. 정확 연계

정확 연계(exact linkage)는 가장 기본적이면서도 신뢰도가 높은 연계 방법이다. 사업자등록번호는 국내 사업체를 식별하는 공식적이고 고유한 식별자이기 때문에, 이 번호가 두 자료에서 동일하게 존재하는 경우 동일 개체로 판정하는 데 이견의 여지가 없다. 따라서, 정확 연계는 전체 연계 과정에서 가장 오류 가능성이 낮은 동일성 판별의 기준층(baseline layer)을 구성하며, 이 단계에서 확보된 레코드는 후속 결정적 연계의 품질 평가를 위한 참조 기준(reference set) 역할을 한다.

정확 연계 수행 결과, <표 5-9>에 나타난 바와 같이 부산 북구와 경북 경산시 두 지역 모두에서 낮은 연계율을 보였다. 부산 북구의 경우, 통신판매업 자료와 시범예행조사 자료를 연계했을 때, 전체 3,391개의 레코드 중 163개의 레코드가 일치(연계율 4.8%), 통신판매업 자료와 표본 명부자료는 353개의 레코드가 일치(연계율 10.4%)하였다. 경북 경산시에서도 유사한 결과를 보였다. 전체 7,675개의 레코드 중 통신판매업 자료와 시범예행조사 자료는 296개의 레코드가 일치(연계율 3.9%), 통신판매업 자료와 표본 명부자료는 588개의 레코드가 일치(연계율 7.7%)하였다.

<표 5-9> 통신판매업 자료 기준 정확 연계 결과

(단위 : 개, %)

지역	기준자료 레코드 수(개)	연계 자료	연계 레코드(개)	연계율(%)
부산 북구	3,391	시범예행조사 자료	163	4.8
		표본 명부자료	353	10.4
경북 경산시	7,675	시범예행조사 자료	296	3.9
		표본 명부자료	588	7.7

이러한 결과는 단순한 수치 비교를 넘어 다음의 중요한 의미를 전달한다.

첫째, 정확 연계율이 낮게 나타났다는 사실은 연계 실패나 절차적 문제를 반영하는 것이 아니라, 세 자료 간의 구조적 차이가 반영된 자연스러운 결과이다. 통신판매업 자료는 통신판매업이라는 특수 목적의 모집단을 구성한다. 반면, 시범예행조사 자료와 표본 명부자료는 경제총조사는 전체 사업체를 포괄하므로, 모집단이 부분적으로만 중첩되는 구조를 갖는다. 즉, 통신판매업 자료에서 사업자등록번호를 기반으로 연계될 수 있는 사업체는 본질적으로 전체 신고 사업체 중 일부에 한정되며, 이는 정확 연계율의 상한을 자연스럽게 제한한다. 또한, 통신판매업 자료에는 휴업·폐업·말소된 사업체가 포함되어 있을 수 있어, 조사 시점 기준 활동사업체만 포함하는 시범예행조사 자료와의 접점(교집합)은 더 좁아질 수 있다. 자료별 시점 차이(train misalignment) 역시 정확 연계율을 낮추는 요인으로 작용한다. 결과적으로 <표 5-9>에 제시된 연계율은 자료원 간 모집단 차이, 활동성의 불일치, 시점 차이라는 구조적 요인의 복합적인 결과로 이해해야 한다.

둘째, 정확 연계는 후속 결정적 연계의 품질을 검증하기 위한 기준층을 제공한다. 즉, 정확 연계에서 확보된 일치 레코드 쌍은 사실상 동일 사업체로 확정된 레코드이며, 이후 단계에서 수행되는 사업체명·대표자명·주소 기반 결정적 연계가 어느 정도의 오류를 포함하는지를 평가하는 참조 기준으로 기능한다.

셋째, 정확 연계의 연계율은 전체 연계율의 상한을 결정한다는 점에서 의미가 있다. 사업자등록번호가 완전하게 일치하는 경우가 전체 레코드의 3~10% 수준이라는 것은, 이후 수행되는 결정적 연계가 연계율을 아무리 높여도 최종적으로 활용이 가능한 연계 레코드의 신뢰성에는 본질적 한계가 존재함을 의미한다. 이는 외부 자료를 경제총조사 특성 항목의 보조 자료로 활용할 때 반드시 고려해야 하는 구조적 요인이다.

나. 결정적 연계

미연계 레코드를 대상으로 공통 변수(예: 사업체명, 대표자명, 주소 정보)를 활용하

는 결정적 연계(deterministic linkage)를 수행하였다. 결정적 연계는 외형적으로 문자열 유사도 기반의 비교 방식처럼 보이지만, 실질적으로는 후보군 생성, 유사도 계산, 임곗값 적용, 규칙 기반 판단이라는 여러 단계를 포함한다. 이는 단순한 문자열 매칭이 아니라, 자료 간 개념적 차이와 표기 방식의 불일치로 인해 발생할 수 있는 오연계 위험을 최소화하기 위한 일련의 통제된 통합 절차라고 할 수 있다.

결정적 연계의 첫 단계는 후보군 생성(candidate set generation)이다. 전처리에서 형식 표준화 및 문자 정규화를 수행하였으나, 자료 간 표기 방식 차이는 여전히 존재하였다. 이에, 후보군 생성에서는 전체 문자열의 단순 비교가 아니라 사업체명 길이, 자음·모음의 유사성, 대표자명 존재 여부, 읍면동 일치 등 복수 조건을 활용해 '충분히 비교 가능한 레코드'만을 대상으로 삼았다. 이 단계는 불필요한 비교로 인해 발생하는 오연계를 방지하는 동시에 계산 효율성을 확보하기 위한 사전 필터 역할을 한다.

후보군이 정의된 이후에는 JW 유사도 점수(Jaro-Winkler similarity score)를 활용하여 문자열 유사도를 산출하였다. 본 연구에서는 국제적으로 통용되는 기준(0.85 ~ 0.90)을 참고하였으나, 자료별 특성을 고려해 조금 더 보수적인 접근을 채택했다. 즉, 동일성 판단을 위해 JW ≥ 0.90을 기본 임곗값으로 설정하고, 0.85 ~ 0.90 구간의 레코드들은 후보군으로만 유지하였다. 이는 "유사도만으로는 동일성을 확정할 수 없다."는 경험에 기초한 것으로, 통신판매업 자료의 주소 표기 특성 때문에 유사도 점수가 높게 나타나더라도 실제 동일하지 않은 사례가 존재함을 확인한 데 따른 조치였다.

본 연구는 유사도 점수만으로 최종 연계를 결정하는 방식을 의도적으로 지양하였다. 이는 문자열 유사도가 정보 구조 전체의 일치성을 대변하지 못하는 한계를 보완하기 위한 것으로, 유사도 기반 판단이 오연계 가능성을 높일 수 있다는 점을 고려한 결과이다. 따라서, 최종 연계 여부는 규칙 기반 결정(rule-based decision)을 통해 확정하였다. 구체적으로 ①사업체명 및 건물명 유사도(JW ≥ 0.90) 충족, ②대표자명의 완전 일치 또는 경총 자료와의 실질적 동일성 판단 가능, ③읍면동 일치의 세 조건을 모두 충족하는 경우에만 동일 사업체로 확정하였다. 이와 같은 다중 조건 충족 방식은 연계의 보수성을 높이면서도 결정적 연계의 신뢰도를 확보하기 위한 필수적인 절차였다.

이러한 연계 전략을 적용한 결과, <표 5-10>에 나타난 바와 같이 매우 보수적인 경향을 보였다. 부산 북구에서는 추가 연계된 레코드는 존재하지 않았으며, 경산시에서는 표본 명부자료와의 연계에서 1개의 레코드가 확인되었다. 이처럼 결정적 연계 레코드가 극히 제한적으로 나타난 것은, 두 자료 간 구조적 이질성이 매우 커 동일성 판단에 필요한 핵심 속성들이 동시에 충족되는 경우가 거의 없었기 때문이다. 이는 외부 자료를 활용한 연계에서 전처리만으로 해결할 수 없는 구조적 한계가 존재한다는 사실을 실증적으로 보여준다.

<표 5-10> 통신판매업 자료 기준 결정적 연계 결과

(단위 : 개, %)

지역	기준자료 기준 미연계 레코드 수(개)	연계 자료	연계 레코드(개)	연계율(%)
부산 북구	3,228	시범예행조사 자료	0	0
	3,038	표본 명부자료	0	0
경북 경산시	7,379	시범예행조사 자료	0	0
	7,087	표본 명부자료	1	0.01

다. 품질 점검

연계 품질 점검(linkage quality assessment)은 연계된 레코드가 실질적인 분석 단위로서 적절한 신뢰성과 의미를 갖는지를 평가하는 단계로, 단순한 연계율만으로는 파악할 수 없는 정보의 일치 구조, 자료 간 개념적 불일치 그리고 잠재적 오류 가능성을 다각도로 점검하는 데 목적이 있다. 특히, 통신판매업 자료와 시범예행조사 자료를 결합하는 경우, 자료별 목적·구성·시점의 차이가 존재하므로, 품질 지표는 알고리즘의 성능보다는 자료 구조의 차이를 반영하는 지표로 해석될 필요가 있다. 본 연구에서는 연계 품질을 다면적으로 평가하기 위해 ①온라인 거래 여부 일치율(attribute agreement), ②오연계율(false match rate), ③누락률(false non-match rate), ④정밀도(precision), ⑤재현율(recall)과 같은 정량적 지표와 수작업 검토를 병행하였다.

<표 5-11> 통신판매업 자료 기준 연계 품질 점검 결과

(단위 : %)

지역	연계 자료	온라인 거래 여부 일치율	오연계율	연계율	누락률	정밀도	재현율
부산 북구	시범예행조사 자료	27.53	72.47	4.81	95.19	88.64	23.93
	표본 명부자료	—	—	10.41	89.59	—	—
경북 경산시	시범예행조사 자료	16.72	83.28	3.86	96.14	90.91	13.56
	표본 명부자료	—	—	7.67	92.33	—	—

<표 5-11>에 나타난 바와 같이, 온라인 거래 여부 일치율은 전반적으로 낮게 나타났다. 이는 자료 간 개념과 수집 방식 자체가 다르다는 구조적 사실을 반영한다. 또한, 시범예행조사의 경우 온라인 거래 여부가 조사불능·부재·불응 등으로 인해 부산 북구 약 37%, 경북 경산시 약 54%의 높은 결측률을 보이므로, 비교 가능한 레코드의 규모가 제한될 수밖에 없다. 이에 따라 온라인 거래 여부 일치율은 낮은 수준에서 형성되며, 이는 연계된 레코드가 서로 다른 개념적 정의 하에서 생성된 정보라는 점에서

자연스러운 결과로 해석해야 한다.

오연계율은 표면적으로는 매우 높게 나타나지만, 이는 오연계 자체가 많아서가 아니라, 품질 지표 산출 방식과 자료 간 개념 차이를 반영한 결과이다. 해당 방식에서는 온라인 거래 여부가 서로 다를 경우 이를 모두 오연계로 분류하고 있기 때문이다. 그러나, 통신판매업 자료의 '온라인 판매 여부'와 시범예행조사 '온라인 거래 여부'는 동일한 개념이 아니며, 사업체의 실제 영업 방식과 신고·응답 방식 간의 시점 차이도 존재한다. 따라서, 불일치는 반드시 오연계를 의미하지 않는다. 실제로 정확 연계된 레코드를 대상으로 수작업 검토를 수행한 결과, 명확한 오연계 사례는 확인되지 않았으며, 결정적 연계로 추가된 경북 경산시의 '통신판매업+표본 명부자료' 연계 레코드 1개 역시 실제 동일 사업체로 판정되었다. 이는 오연계율이 자료의 구조적 차이에 의해 과대 추정된 값이라는 점을 명확히 보여준다.

누락률이 매우 높게 나타난 것도 연계의 실패라기보다 자료의 모집단 구성 차이를 반영한 자연스러운 결과이다. 앞서 언급한 것처럼, 통신판매업 자료는 통신판매업이라는 특정 업종 기반의 좁은 범위(coverage frame)를 가지고 있으며, 경제총조사 자료는 전체 사업체를 포괄한다. 이러한 차이는 통신판매업 자료에 존재하는 많은 사업체가 시범예행조사 및 표본 명부에 존재하지 않는 구조를 만들어낼 수 있으며, 그 반대의 경우도 발생할 수 있다. 즉, 누락률은 모집단 차이로 인한 구조적 불일치의 크기를 나타낸 지표로 해석해야 한다. 정밀도는 연계된 레코드 중 실제 동일한 사업체의 비율을 의미하며, 정확 연계의 특성상 90% 이상일 수밖에 없다. 반면, 재현율은 실제 동일 사업체 중 연계를 통해 포착된 비율을 의미하므로, 자료 구조 차이로 인해 행정자료·조사자료 연계에서는 낮게 나타날 수밖에 없다.

연계 및 미연계 레코드 중 일정 비율을 무작위 추출하여 본 연구진이 독립적으로 동일성 여부를 수작업 검토한 결과, 정확 연계에는 오연계가 존재하지 않았으며, 결정적 연계에서 추가된 1개 레코드도 실제 동일 사업체로 확인되었다. 반면, 온라인 거래 여부 등 속성 수준의 불일치는 자료 간 개념 정의의 차이, 시점 불일치, 응답 기반 수집의 한계 등으로 설명이 가능하였다.

5. 시사점

본 절은 통신판매업 신고 자료와 경제총조사 표본 명부 및 시범예행조사 자료를 연계하여 외부 자료가 조사 기반 통계 생산에 어떤 역할을 할 수 있는지를 검토한 실증분석 사례이다. 본 연구의 실증적 결과는 다음의 시사점을 함의한다.

첫째, 외부 자료를 활용하여 조사자료의 특성 항목을 직접 대체하는 것이 현실적으

로 어렵다는 점을 명확히 보여주었다. 통신판매업 자료는 신고를 기반으로 온라인 판매 여부가 기록되는 반면, 시범예행조사 자료는 응답 기반으로 온라인 거래 여부 항목을 수집한다. 유사한 속성을 다루고 있음에도, 그 개념적 정의와 수집 목적이 달라 두 자료 간 일치율이 낮고 오연계율이 높은 것이다. 이는 외부 자료를 특성 항목의 단일 원자료로 활용하기보다는, 보조지표(reference information) 또는 품질 진단 도구(quality check)로 활용하는 것이 적합하다는 점을 의미한다.

둘째, 품질 지표가 보여준 높은 누락률과 속성 불일치는 연계 방법의 문제라기보다, 자료 간 모집단 차이에서 비롯되는 것으로 해석해야 한다. 통신판매업 자료는 작은 모집단을 대상으로 하며, 경제총조사 시범예행조사 자료는 활동사업체 전체를 포괄하는 모집단을 대상으로 한다. 이러한 불일치는 연계 가능성을 구조적으로 제한하며, 자료 간 교집합의 규모를 축소하는 원인이 된다. 따라서 외부 자료를 국가통계 생산에 활용하기 위해서는 자료 모집단의 속성과 활용 목적을 명확히 구분하고, 모집단 일치 여부를 사전에 검토하는 절차가 필요하다.

셋째, 본 연구는 전처리의 중요성과 동시에 그 한계를 분명하게 드러냈다. 사업체명 정규화, 주소 파싱, 대표자명 정비, 코드 정렬 등 다양한 전처리 도구를 활용해 자료 간 구조적 차이를 최소화했음에도 불구하고, 이는 어디까지나 사후적 보정(ex-post correction)의 성격을 띠므로 근본적인 이질성을 제거하는 데에는 한계가 있었다. 이는 데이터 통합이 기술적 정비만으로 해결될 수 없으며, 자료 생성 단계에서부터 표준화된 서식, 통일된 개념 정의, 일관된 코드와 분류체계가 마련되지 않는다면 통합의 효과가 크게 제한된다는 점을 의미한다. 특히 외부 행정자료를 경제총조사와 같은 공식 통계에 본격적으로 활용하기 위해서는 **행정자료 서식의 표준화**가 중요한 선결 조건이다. 현재 행정자료는 기관·업무 목적마다 자료 항목·코드·주소 표기 방식이 상이하고, 동일 속성이라도 정의가 다르게 설정되어 있어 전처리 단계에서 막대한 정합성 조정 비용이 발생한다. 이러한 상황에서는 데이터 통합이 반복될수록 오히려 자료 간 불일치와 정보 손실이 증가할 가능성이 크다. 향후, 데이터 통합체계를 확립하기 위해서는 행정기관 간 메타데이터 체계의 일원화, 표기 표준 설정, 코드 체계 공동 관리 등 **국가 차원의 표준화 체계 구축이 필수**이다.

넷째, 본 연구 결과는 외부 자료가 특성 항목을 '대체'하는 용도보다는 조사자료의 보완과 품질 진단에서 더 높은 실질적 활용 가능성을 가진다는 점을 보여 주었다. 예를 들어, 시범예행조사 자료의 온라인 거래 여부는 조사 현장의 제약으로 인해 구조적 결측률이 높으므로, 외부 자료를 활용하여 결측의 분포와 원인을 파악하거나 특정 유형의 사업체에서 결측이 집중적으로 발생하는지 확인하는 데 의미 있는 근거가 될 수 있다. 또한, 외부 자료와 조사자료의 불일치 양상(pattern)은 조사설계, 응답 관리, 표본 추출 방식 등 조사 기반 통계 생산 체계 전반의 개선을 위한 진단 도구로 활용될 수 있다.

제3절 사례 2. ‘스마트공장 운영 여부’ 항목 보완

1. 자료 진단

본 절에서 활용한 자료는 ①공공데이터포털의 스마트공장DB, ②경제총조사 표본 명부자료, ③경제총조사 시범예행조사 자료이다. 제2절에서 상술한 바와 같이, 각 자료는 수집 목적과 관리 체계, 처리 방식, 변수 정의가 다르므로, 이를 하나의 분석 단위로 연계하기 위해서는 사전 진단을 통해 자료 간 구조적 이질성을 파악해야 한다. 또한, 어떤 레코드가 어떤 사유로 제외되었는지를 충분히 확인하지 못한 상태에서 연계를 수행하면, 비활동 사업체가 연계에 포함되거나, 조사 대상에서 이미 제외된 레코드가 다시 분석에 사용되는 등 오류가 발생할 수 있다. 이에 자료별로 전체 규모와 제외 규모를 제시하고, 제외 사유를 정리함으로써 이후 전처리 및 연계 절차의 전제를 명확히 하고자 한다. <표 5-12>에서 확인할 수 있듯이, 경제총조사 표본 명부자료와 시범예행조사 자료는 상당 비율의 레코드가 정제 단계에서 제외되었다.

<표 5–12> 자료별 규모 및 제외 현황

(단위 : 개, %)

지역	자료 구분	총 레코드(개)	제외 레코드(개)	제외 비율(%)	정제 후 활용 레코드(개)
부산 북구	스마트공장DB	10	0	0.00	10
	경총 표본 명부	23,250	22,514	96.83	736
	경총 시범예행조사	9,837	9,342	94.97	495
경북 경산시	스마트공장DB	8	0	0.00	8
	경총 표본 명부	34.844	30,619	87.87	4,225
	경총 시범예행조사	16,418	13,263	80.78	3,155

가. 공공데이터포털 스마트공장DB

스마트공장DB는 ‘스마트 제조 혁신 지원 사업’을 통해 구축된 행정자료로서, 제조업 중심의 기술지원·고도화 사업 참여 기업을 대상으로 한다. 사업자등록번호가 제공되지 않는다는 한계가 있지만, 모든 레코드에서 도로명 기준으로 주소 정보가 상세하게 수록되어 있어 원활한 문자열 기반 주소 비교가 가능하다. 부산 북구는 총 10개, 경북 경산시는 총 8개의 레코드로 구성되어 있으며, 자료 특성상 모집단이 작아 연계 가능성은 제한적일 수밖에 없다.

나. 경제총조사 표본 명부자료

표본 명부자료는 조사표별로 포함되는 항목이 다르므로 특정 특성 항목의 결측은

단순한 누락이 아니라 조사표 설계에 따라 항목 자체가 존재하지 않는 구조적 결측이다. 전체 명부 규모와 정제 후 활용 레코드를 기준으로 제외 사유를 분류하였다. <표 5-13>은 명부자료의 제외 현황으로, 가장 큰 제외 사유는 조사표 유형에 따른 구조적 제외이다. 부산 북구의 경우 연계 대상이 '스마트공장 운영 여부' 특성 항목으로 구성되었기에 22,514개의 레코드를 제외하였다. 경산시 역시 동일한 사유로 30,617개가 제외되었다. 두 번째는 주소 불일치로 경산시에서 2개 사례가 확인되어 제외하였다.

<표 5-13> 표본 명부자료의 제외 사유별 현황

(단위 : 개)

제외 사유	제외 레코드(개)		비고
	부산 북구	경북 경산시	
비대상 조사표	22,514	30,617	특성 항목이 수집되지 않는 조사표 유형
타 시군구	0	2	

<표 5-14>에 나타난 바와 같이 표본 명부자료의 결측률은 매우 낮다. 제2절에서 언급한 것처럼 명부자료가 표본 틀 성격을 갖기 때문이다.

<표 5-14> 표본 명부자료의 결측 현황

(단위 : 개, %)

지역	활용 레코드(개)	변수명	결측 레코드(개)	결측률(%)
부산 북구	736	주소	2	0.27
경북 경산시	4,225	주소	21	0.50

다. 경제총조사 시범예행조사 자료

<표 5-15>는 시범예행조사 자료의 제외 사유별 현황으로 폐업, 휴업, 전출 등은 조사원이 직접 확인한 정보이며, 이는 행정자료나 명부자료에서 포착하기 어려운 실질적 자료이다.

<표 5-15> 시범예행조사 자료의 제외 사유별 현황

(단위 : 개)

제외 사유	제외 레코드(개)		비고
	부산 북구	경북 경산시	
비대상 조사표	9,294	12,902	특성 항목이 수집되지 않는 조사표 유형
휴·폐업	19	179	
전출(타 지역)	8	47	
기타	21	122	이중등재, 흡수합병, 확인불가, 주소 불명확, 제외업종

정제 후 남은 레코드를 대상으로 결측을 확인한 결과, <표 5-16>에 제시된 바와 같이 주소 정보는 두 지역 모두에서 결측이 존재하지 않았다. 이는 ‘스마트공장 운영 여부’ 조사 대상이 제조업(산업대분류 C)으로 제한되며, 제조업체는 대체로 사업체 위치 정보가 안정적으로 관리되기 때문이다.

이러한 주소 정보의 안정성은 통신판매업 자료와 달리, 스마트공장DB의 결정적 연계 과정에서 주소 정합성의 신뢰도를 높여주는 요인으로 작용한다. 스마트공장 운영 여부 항목 또한 경산시에서만 결측률이 2.73%로 온라인 거래 여부 결측률(부산 북구 37.09%, 경북 경산시 54.25%)과 비교할 때 매우 낮은 수준이다.

<표 5–16> 시범예행조사 자료의 결측 현황

(단위 : 개, %)

지역	활용 레코드(개)	변수명	결측 레코드(개)	결측률(%)
부산 북구	495	주소	0	0.00
		스마트공장 운영 여부	0	0.00
경북 경산시	3,155	주소	1	0.03
		스마트공장 운영 여부	86	2.73

2. 스키마 정렬 및 개념 조화

공공데이터포털의 스마트공장DB와 경제총조사 표본 명부 및 시범예행조사 자료를 연계하기 위해 수행한 스키마 정렬 및 개념 조화는 기본적으로 제2절에서 제시한 절차와 동일하다. 즉, 사업체명 · 대표자명 · 주소 구성요소 등 공통 변수에 대한 형식 표준화, 문자 정규화, 코드 및 범주 정렬(예: 스마트공장 운영 여부를 ‘1 = 예, 2 = 아니오’로 통일) 등을 적용하였다. 다만, 스마트공장 운영 여부의 경우, 스마트공장DB와 경제총조사 시범예행조사 자료가 동일 개념을 직접적으로 측정하지는 않는다는 점이 중요하다.

스마트공장DB는 특정 지원 사업에 참여한 사업체의 ‘참여 여부’를 나타내는 행정적 속성이지만, 시범예행조사 자료에서의 스마트공장 여부는 실제 생산설비 · 시스템 수준에서의 ‘스마트공장 운영 여부’를 응답 기반으로 측정한다. 따라서, 두 자료 간 스마트공장 운영 여부 값을 속성 차원에서 직접 비교하거나 일치율을 산출하는 것은 개념적으로 타당하지 않으며, 스키마 정렬과 개념 조화의 목적은 어디까지나 ‘동일 사업체를 식별할 수 있는 최소한의 공통 구조’를 확보하는 데 한정된다. 이러한 이유로 스키마 정렬과 개념 조화의 세부 절차 설명을 반복하지 않고, 제2절과 동일한 절차가 적용되었음을 전제로 연계 및 품질 점검 결과에 초점을 맞추어 서술한다.

3. 연계 및 품질 점검

스마트공장DB와 경제총조사 표본 명부 및 시범예행조사 자료의 연계는 제2절과 달리 사업자등록번호를 활용한 정확 연계가 불가능한 환경에서 수행되었다. 스마트공장DB에는 사업자등록번호가 제공되지 않기 때문에 동일성 판단을 고유식별자에 의존하는 방식으로는 구현할 수 없다. 이에 공통 변수(사업체명, 대표자명, 주소 구성요소)를 활용한 결정적 연계에 한정하여 연계 가능성을 검토하였다.

결정적 연계에서 주소 정보는 후보군 생성과 동일성 판단을 보조하는 핵심 역할을 맡았다. 스마트공장DB의 주소는 읍면동, 도로명, 본번, 부번, 건물명, 동, 층, 호 등 여러 구성요소로 분해가 가능할 정도로 상세하게 기록되어 있다. 이를 활용하기 위해 주소 전체를 하나의 문자열로 비교하는 방식 대신 구성 요소별로 문자열 유사도 또는 완전 일치 여부를 산출한 뒤, 읍면동과 도로명에 가장 높은 가중값, 본번·부번·건물명에 중간 수준의 가중값, 동·층·호에 낮은 가중값을 부여하여 가중합 형태의 최종 주소 유사도 점수를 산출하였다.[18] 이와 같은 방식은 주소 구성요소의 중요도를 반영하여 동일성 판단의 정밀도를 높이고, 도로명·지번 혼합, 건물명 누락 등 현실적 문제를 고려하면서도 구조적 일치성 여부를 최대한 반영할 수 있도록 설계한 것이다. 그러나 경제총조사 자료의 경우 건물명·동·층·호가 누락된 사례가 상당수 존재하였다. 이에 따라 주소 유사도가 높게 산출되더라도 사업체명이나 대표자명이 일치하지 않으면 동일성 판단에 이르기 어려워, 실제 연계 여부를 결정하는 데 있어 사업체명 유사도(0.90 이상)와 대표자명 완전 일치 여부가 여전히 핵심 기준으로 작동하였다. 주소 유사도는 후보군의 타당성을 검증하는 보조지표로 유용했지만, 연계 건수를 실질적으로 늘리는 역할을 하지는 못했다.

<표 5-17>에 제시된 바와 같이, 부산 북구에서는 스마트공장DB와 경제총조사 표본 명부 및 시범예행조사 자료 간 연계는 단 한 건도 발생하지 않았으며, 경산시에서도 표본 명부자료와의 연계가 1건에 그쳤다. 이는 스마트공장 DB의 모집단이 경제총조사에서 스마트공장 항목을 조사하는 모집단과 본질적으로 다르다는 점을 의미한다.

<표 5-17> 스마트공장DB 기준 결정적 연계 결과

(단위 : 개, %)

지역	기준자료 레코드 수(개)	연계 자료	연계 레코드(개)	연계율(%)
부산 북구	10	시범예행조사 자료	0	0.00
		표본 명부자료	0	0.00
경북 경산시	8	시범예행조사 자료	0	0.00
		표본 명부자료	1	12.50

18) 주소 유사도 $= \sum_{i \in \text{주소구성요소}} w_i sim_i$

수작업 검토 결과가 이를 뒷받침해 준다. 스마트공장 DB에 포함된 사업체들을 대상으로 표본 명부 및 시범예행조사 자료 전체를 대조해 본 결과, 부산 북구의 경우 스마트공장DB에 기록된 10개 사업체 중 6개가 경제총조사 표본 명부자료 전체에서 확인되었고, 3개는 시범예행조사 자료의 전체 조사 대상에 포함되어 있었다. 경산시 역시 스마트공장DB의 8개 사업체 중 4개가 표본 명부 및 시범예행조사 자료 전체에 공통으로 포함되어 있었다. 중요한 점은, 이들 사업체가 모두 경제총조사 체계 내에서는 '조사표 6번(스마트공장 운영 여부가 수집되지 않는 조사표)'과 '전년도 산업대분류 J(정보통신업)'에 속해 있었다는 사실이다. 다시 말해, 스마트공장DB는 이들 사업체를 '스마트공장 관련 사업 참여 기업'으로 관리하고 있으나, 경제총조사에서는 제조업(C) 기반의 스마트공장 조사 대상이 아니라 정보통신업(J)으로 관리·조사하고 있었다. 이는 자료 생성 단계에서부터 서로 다른 모집단과 산업분류체계를 적용해 온 결과임을 보여준다. 동일 사업체가 행정자료에서는 제조업 기반 스마트공장 지원 대상 사업체로, 경제총조사 체계에서는 정보통신업 조사 단위로 관리되는 상황에서는 두 자료를 통합하여 '스마트공장 운영 여부'라는 단일 특성 항목으로 정합성 있게 작성하는 것은 불가능에 가깝다.

4. 시사점

스마트공장DB와 경제총조사 자료 연계 결과, 부산 북구 0건, 경북 경산시 1건에 불과할 정도로 제한적이었지만, 이 과정에서 도출된 시사점은 외부 행정자료를 활용한 특성 항목 보완의 가능성과 한계를 매우 선명하게 드러낸다. 특히, 스마트공장DB에 포함된 사업체 중 상당수가 경제총조사 표본 명부자료와 시범예행조사 자료에 이미 포착되어 있음에도, 서로 다른 조사표와 산업분류로 관리되고 있었다는 사실은 향후 행정자료와 조사자료를 결합하는 전략을 설계하는 데 있어 반드시 고려해야 할 몇 가지 중요한 함의를 지닌다.

첫째, 스마트공장DB와 경제총조사 자료 간의 낮은 연계율은 동일 사업체가 두 자료에 동시에 존재하지 않아서가 아니라, 두 자료가 동일 사업체를 서로 다른 방식으로 정의하고 분류하기 때문이다. 스마트공장DB에서는 스마트공장 지원 사업 신청 또는 수행 실적을 기준으로 제조업 관련 기업으로 간주하고, 경제총조사에서는 주사업 기준의 산업분류를 적용하여 정보통신업(J)에 배정하는 경우, 한 사업체가 두 자료에서 서로 다른 산업·속성을 갖게 된다. 이 경우, 전처리와 연계 알고리즘을 아무리 정교하게 설계하더라도 '스마트공장 운영 여부'라는 특성 항목으로 두 자료를 통합하는 데에는 근본적인 한계가 존재한다.

둘째, 스마트공장DB의 사업체가 경제총조사 체계 밖의 '완전히 새로운' 모집단이라

기보다 경제총조사라는 자료 틀 안에 존재하지만, 주사업·부사업 구조와 행정 목적에 따라 다른 분류를 부여받은 동일 모집단일 수도 있다. 예를 들어, 주사업은 정보통신업(J)에 해당하지만, 부사업으로 제조업(C) 부문을 보유한 기업이 스마트공장 지원 사업을 신청할 경우, 행정자료에서는 제조업(C) 기반 스마트공장 기업으로 관리될 수 있다. 반면, 경제총조사는 주사업을 기준으로 산업을 분류하므로, 동일 기업이 정보통신업(J)으로 조사될 수 있다. 이와 같이 주·부사업 구조와 행정·통계 목적의 차이에서 발생하는 분류 불일치는 향후 행정자료를 활용할 때 필연적으로 마주하게 될 구조적 문제이다.

셋째, 스마트공장 운영 여부와 같은 특성 항목을 행정자료로 대체하거나 보완하는 전략을 설계할 때, 단순히 “외부 자료에 해당 변수가 존재하는가?”만을 기준으로 판단해서는 안 된다. 외부 자료에 원하는 특성 항목이 존재하더라도, 그 변수가 어떤 산업 범위, 어떤 사업체 정의, 어떤 사업 구조를 전제로 수집되었는지가 명확히 통계 작성 체계와 연결되지 않는다면, 연계 후 데이터의 일관성은 오히려 저해될 수 있다.

넷째, 행정자료를 활용한 데이터 통합의 관점에서 보면, 본 연구 결과는 전처리와 연계만으로는 해결할 수 없는 ‘자료 생성 단계의 표준화 문제’를 다시금 부각한다. 스마트공장DB는 주소 정보가 상세하게 정비되어 있고, 사업체명 등 기본 정보도 일정 수준 이상으로 관리되고 있음에도, 사업자등록번호 부재, 산업분류 부여 기준의 차이 등의 불일치로 인해 통합 데이터를 산출하지 못했다. 이는 행정자료를 통계적으로 활용하기 위해서는 수집·관리 단계에서부터 표준화가 이루어져야 하며, 그렇지 않으면 데이터 통합은 제한적인 성과에 그칠 수밖에 없다는 점을 시사한다.

마지막으로, 스마트공장 운영 여부 항목 자체에 대한 시사점도 도출할 수 있다. 시범예행조사에서 스마트공장 운영 여부의 결측률은 낮게 나타났고, 조사표 설계와 조사 대상 정의가 명확하게 설정된 만큼, 당분간 스마트공장 운영 여부는 조사 기반으로 작성하는 것이 합리적이다. 외부 자료는 경제총조사에서 포착되지 않는 일부 사업체의 디지털 전환 특성을 파악하거나, 정책 사업 참여 사업체의 특성을 분석하는 데 보조적으로 활용될 수 있으나, 경제총조사의 스마트공장 운영 여부 항목을 직접 대체하거나 결측을 기계적으로 보완하는 자료로 활용하기에는 구조적 한계가 크다. 향후, 행정자료를 활용한 특성 항목 보완을 추진하기 위해서는, 자료 구조·산업분류·조사표 체계의 정합성을 우선적으로 확보하는 제도적 정비가 필수적이며, 그 이후에야 데이터 통합이 통계 품질 개선으로 이어질 수 있을 것이다.

제 6 장

실증분석2 : 데이터 병합

제1절 분석 대상

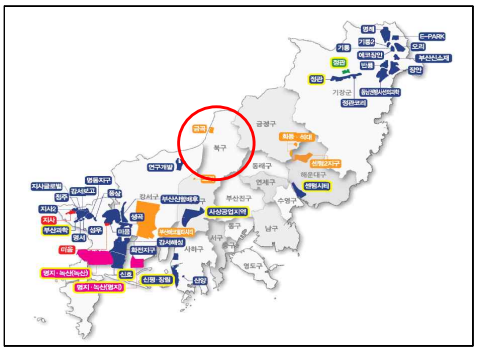
본 연구는 경제구조통계 특성 항목 자료수집 개선 방안을 마련하기 위한 연구로 2024년 기초연구를 통해서 데이터과학 기술 활용 및 공공데이터(DB) 추가 확보 방안을 제안하였다. 따라서, 제안방법 검증은 경제구조통계의 실제 조사자료를 기준으로 하여 자료수집·연계·일치 과정을 통해 개선 가능성에 대한 실증분석으로 진행한다.

실증분석 분석자료는 2025년 경제총조사 시범예행조사 데이터로, 조사 대상 3개 지역 중 2개 지역(부산시 북구, 경북 경산시) 전체 사업체를 대상으로 한다.

<그림 6-1>은 부산시 북구와 경북 경산시 산업단지공단 현황을 나타냈다.

부산시 북구는 1개 산업단지(금곡 도시첨단)가 있고, 경북 경산시는 7개 산업단지가 위치하고 있어 2개 지역의 주요 산업에 차이가 있음을 확인할 수 있다. 즉, 부산시 북구는 서비스업 중심, 경북 경산시는 제조업 중심의 도시로 볼 수 있다.

<부산시 북구>



<경북 경산시>

※ 출처 : 산업단지공단 지역별 현황지도(2025년)

<그림 6-1> 산업단지공단 현황_부산시 북구, 경북 경산시

이 두 지역은 교육 환경에서도 차이를 보인다. 부산시 북구는 종합 4년제 대학교는 없고 주로 전문대학 및 기타 교육기관이 있는 반면, 경북 경산시에는 종합대학교, 전문

대학교, 사이버대학 등을 포함하여 10여 개 대학이 위치하고 있다.

또한 인구학적 특징을 살펴보면, 두 지역 모두 주민등록인구는 26만 명~28만 명 수준으로 비슷하지만, 부산시 북구는 주거 중심 지역으로 신시가지 개발 지역 등에 따라 인구 증감을 반복하고 있고, 경북 경산시는 대구권 위성도시이고 다수의 산업단지와 대학교가 위치하고 있어 인구가 꾸준히 증가하는 추세이다.

본 연구에서 두 지역을 분석 대상으로 선정한 이유는 지역별 특징[19]에 따라 특성 항목별 자료수집(스크래핑, open API)에 차이가 발생하는지 비교하기 위해서이다.

<표 6-1>은 분석 대상 전체 사업체 수를 나타냈다.

<표 6-1> 분석 대상 전체 사업체 수 현황

(단위 : 개, %)

산업대분류	현장조사		현장 미조사		합계	
합계	23,087	(55.6)	18,407	(44.4)	41,494	(100.0)
A	40	(100.0)		-	40	(100.0)
B	7	(100.0)		-	7	(100.0)
C	3,569	(93.2)	260	(6.8)	3,829	(100.0)
D	219	(23.8)	701	(76.2)	920	(100.0)
E	119	(92.2)	10	(7.8)	129	(100.0)
F	1,253	(38.8))	1,974	(61.2)	3,227	(100.0)
G	4,293	(38.0)	7,001	(62.0)	11,294	(100.0)
H	693	(43.9)	886	(56.1)	1,579	(100.0)
I	3,761	(56.9)	2,850	(43.1)	6,611	(100.0)
J	222	(60.7)	144	(39.3)	366	(100.0)
K	300	(87.0)	45	(13.0)	345	(100.0)
L	1,160	(69.0)	521	(31.0)	1,681	(100.0)
M	732	(67.3)	355	(32.7)	1,087	(100.0)
N	483	(69.5)	212	(30.5)	695	(100.0)
O	100	(100.0)		-	100	(100.0)
P	1,245	(51.6)	1,166	(48.4)	2,411	(100.0)
Q	1,466	(96.8)	48	(3.2)	1,514	(100.0)
R	812	(73.4)	295	(26.6)	1,107	(100.0)
S	2,613	(57.4)	1,939	(42.6)	4,552	(100.0)

※ 산업대분류 운수업(H) 중 개인택시, 용달화물, 개별화물 제외

19) 지리적 특징, 산업단지공단, 대학교 등의 차이로 유동 인구에 큰 차이가 있음
(경북 경산시 〉 부산시 북구)

전체 사업체(41,494개)에서 현장 조사는 23,087개(55.6%), 현장 미조사(행정자료 및 imputation)는 18.407개(44.4%)로 나타났다. 또한, 전체 사업체 기준 산업대분류별로 살펴보면, G가 11,294개로 가장 많았고, I가 6,611개, S가 4,552개 순으로 나타났다.

다만, 산업대분류 H(운수업) 중 개인택시, 용달화물, 개별화물은 웹페이지에서 특성 항목 정보를 파악하기 어렵기 때문에 분석에서 제외하였다.

<표 6-2>는 2개 지역 분석 대상 사업체 수 현황을 나타냈다.

먼저 부산시 북구는 서비스업 대상 사업체 수가 더 많은 지역으로, 산업대분류 G가 4,822개로 가장 많았고, 산업대분류 I가 2,777개, S가 1,965개 순으로 나타났다.

<표 6-2> 2개 지역 분석 대상 사업체 수 현황(최종)

(단위 : 개, %)

산업대분류	부산시 북구			경북 경산시		
	현장조사	현장미조사	소계	현장조사	현장미조사	소계
합계	8,647	7,327	15,974	14,440	11,080	25,520
A				40		40
B				7		7
C	495	92	587	3,074	168	3,242
D	3	4	7	216	697	913
E	25	6	31	94	4	98
F	522	619	1,141	731	1,355	2,086
G	1,795	3,027	4,822	2,498	3,974	6,472
H	223	348	571	470	538	1,008
I	1,601	1,176	2,777	2,160	1,674	3,834
J	77	76	153	145	68	213
K	137	17	154	163	28	191
L	546	206	752	614	315	929
M	217	176	393	515	179	694
N	174	80	254	309	132	441
O	40		40	60		60
P	572	521	1,093	673	645	1,318
Q	696	25	721	770	23	793
R	398	115	513	414	180	594
S	1,126	839	1,965	1,487	1,100	2,587

※ 산업대분류 운수업(H) 중 개인택시, 용달화물, 개별화물 제외

반면, 경북 경산시는 서비스업과 제조업 대상 사업체 수가 고르게 분포하는 지역으로, 산업대분류 G가 6,472개로 가장 많았고, 산업대분류 I가 3,834개, C가 3,242개 순으로 나타났다. 두 지역을 비교하면, 경북 경산시 사업체 수(25,520개)는 부산시 북구 사업체 수(15,974개)보다 9,546개 더 많았고, 현장조사 비율도 경북 경산시(56.6%)가 부산시 북구(54.1%)보다 2.5% 높게 나타났다.

특성 항목은 경제총조사 지침의 산업대분류에 따라 수집 여부가 결정[20]된다.

본 연구에서 수집된 정보는 '일일 평균 영업 시간', '배달(택배 포함) 판매 여부', '사업체 건물 연면적' 3개 항목이다. 실증분석은 조사명부와 수집자료 연계율, 현장조사와 수집자료 일치율, 현장 미조사(대체값)와 수집자료 비교를 통해 수집자료의 활용 가능성(직접 활용, 보조 정보 활용)을 살펴봤다.

제2절 비교 및 분석

1. '일일 평균 영업시간' 항목

<표 6-3>은 부산시 북구 '일일 평균 영업시간' 항목의 스크래핑 자료수집 현황을 나타냈다.

부산시 북구의 전체 사업체(15,974개) 중 스크래핑(scraping)을 통해 수집된 '일일 평균 영업시간' 자료는 1,453개로 연계율(9.1%)이 낮게 나타났다. 현장조사(surveyed)는 8,647개 중 904개(10.5%)가 연계, 현장 미조사(unsurveyed)는 7,327개 중 549개(7.5%)가 연계되었다.

조사 대상 산업대분류 11개 기준에서 살펴보면, 전체 12,793개 사업체 대비 1,299개(10.2%)가 연계되었고, 현장조사(surveyed)는 6,571개 중 766개(11.3%)가 연계, 현장 미조사(unsurveyed)는 6,222개 중 533개(8.6%)가 연계되었다.

---

20) '연간 영업 개월 수, 월간 정기 휴무일 수' 항목은 全 산업, '일일 평균 영업 시간' 항목은 산업대분류 G, I, E, J, L, M, N, P, O, R, S, '배달(택배 포함) 판매 여부' 항목은 산업대분류 G, '객석 여부' 항목은 산업대분류 I, '사업체 건물 연면적' 항목은 산업대분류 C, G, I

<표 6-3> '일일 평균 영업시간' 항목 수집(연계) 현황_부산시 북구

(단위 : 개)

KSIC_S		부산시 북구					
		total	scraping				
				surveyed	scraping	unsurveyed	scraping
합계		15,974	1,453	8,647	904	7,327	549
	미대상	3,181	154	2,076	138	1,105	16
	A						
	B						
	C	587	67	495	65	92	2
	D	7		3		4	
	F	1,141	48	522	42	619	6
	H	571	18	223	10	348	8
	K	154	1	137	1	17	
	Q	721	20	696	20	25	
	대상	12,793	1,299	6,571	766	6,222	533
	E	31		25		6	
	G	4,822	367	1,795	178	3,027	189
	I	2,777	499	1,601	258	1,176	241
	J	153	1	77	1	76	
	L	752	64	546	41	206	23
	M	393	25	217	22	176	3
	N	254	26	174	25	80	1
	O	40		40			
	P	1,093	105	572	74	521	31
	R	513	36	398	11	115	25
	S	1,965	176	1,126	156	839	20

※ '일일 평균 영업시간' 항목을 미조사하는 산업대분류

<표 6-4>는 11개 산업대분류('일일 평균 영업시간' 조사 대상)의 스크래핑 연계율을 나타냈다. 산업대분류 I(숙박 및 음식점업)의 연계율은 18.0%로 가장 높았고, 산업대분

<표 6-4> 조사 대상 산업대분류별(11개) 스크래핑 연계율_부산시 북구

(단위 : %)

산업대분류	전체	현장조사	현장 미조사
E			
G	7.6	9.9	6.2
I	18.0	16.1	20.5
J	0.7	1.3	
L	8.5	7.5	11.2
M	6.4	10.1	1.7
N	10.2	14.4	1.3
O			
P	9.6	12.9	6.0
R	7.0	2.8	21.7
S	9.0	13.9	2.4

류 N(사업시설관리, 지원서비스)이 10.2%, P(교육서비스업)가 9.6%, S(수리 및 기타 개인서비스업)가 9.0% 순으로 나타났다. 그 외 산업대분류 E(수도, 하수 및 폐기물 처리, 원료 재생업)와 O(공공 행정)는 대상 웹사이트(네이버 지도에서 조사명부 주소 정보를 활용) 스크래핑을 활용한 자료수집이 어려웠다.

<표 6-5>는 부산시 북구 '일일 평균 영업시간' 항목의 연계 자료에서 현장조사 결과와 스크래핑 결과를 비교한 결과표이다. 결과 일치율은 (1,1), (2,2), (3,3), (4,4), (5,5)인 사업체(yellow color)를 의미하며, 현장조사 null을 제외한 연계 사업체 763개 중 326개의 데이터가 일치(42.7%)하는 것으로 나타났다. 또한, 스크래핑 결과(red color)는 현장조사보다 평균 영업시간이 긴 것으로 나타났고, 항목 무응답(null)은 스크래핑 결과로 대체가 가능하다.

<표 6-5> 현장조사 결과와 스크래핑 결과 비교_부산시 북구

(단위 : 개)

부산시 북구		스크래핑 결과					합계
		1	2	3	4	5	
현장조사 결과	1	55	49	33	22	16	175
	2	18	96	112	26	32	284
	3	2	25	93	42	14	176
	4	2	2	17	23	18	62
	5		3	2	2	59	66
	null		1	1		1	3
합계		77	176	258	115	140	766

<표 6-6>은 부산시 북구 '일일 평균 영업시간' 항목의 연계 자료에서 현장 미조사 결과(대체)와 스크래핑 결과를 비교한 결과표이다. 두 자료는 대상 사업체 500개 중 145개가 일치(29.0%)하는 것으로 나타났고, 스크래핑 결과는 대체 결과보다 큰 값(red color)을 가지는 것으로 보인다. 이 결과는 현장조사 결과와 동일한 특징을 보인다.

<표 6-6> 현장 미조사 결과(대체)와 스크래핑 결과 비교_부산시 북구

(단위 : 개)

부산시 북구		스크래핑 결과					합계
		1	2	3	4	5	
현장 미조사 결과 (대체)	1	36	25	36	9	14	120
	2	14	42	64	21	25	166
	3	11	26	37	20	15	109
	4	7	13	21	9	5	55
	5	2	8	9	10	21	50
	null	15	4	3	10	1	33
합계		85	118	170	79	81	533

<표 6-7>은 경북 경산시 '일일 평균 영업시간' 항목 스크래핑 자료수집 현황이다.

경북 경산시는 전체 사업체(25,520개) 중 스크래핑(scraping)을 통해 수집된 '일일 평균 영업시간' 자료는 1,009개로 연계율(4.0%)이 낮게 나타났다. 현장조사(surveyed)는 14,440개 중 547개(3.8%)가 연계, 현장 미조사(unsurveyed)는 11,080개 중 462개(4.2%)가 연계되었다.

'일일 평균 영업시간'을 조사하는 11개(대상) 산업분류 기준에서 살펴보면, 전체 17,240개 사업체 대비 813개(4.7%)가 연계되었고, 현장조사(surveyed)는 8,969개 중 428개(4.8%)가 연계, 현장 미조사(unsurveyed)는 8,271개 중 385개(4.7%)가 연계되었다.

<표 6-7> '일일 평균 영업시간' 항목 수집(연계) 현황_경북 경산시

(단위 : 개)

KSIC_S		경북 경산시					
		total	scraping				
				surveyed	scraping	unsurveyed	scraping
합계		25,520	1,009	14,440	547	11,080	462
	미대상	8,280	196	5,471	119	2,809	77
	A	40		40			
	B	7		7			
	C	3,242	46	3,074	43	168	3
	D	913	1	216	1	697	
	F	2,086	86	731	16	1,355	70
	H	1,008	4	470	2	538	2
	K	191	15	163	15	28	
	Q	793	44	770	42	23	2
	대상	17,240	813	8,969	428	8,271	385
	E	98	2	94	2	4	
	G	6,472	189	2,498	85	3,974	104
	I	3,834	279	2,160	165	1,674	114
	J	213	2	145	1	68	1
	L	929	36	614	25	315	11
	M	694	13	515	12	179	1
	N	441	13	309	10	132	3
	O	60	5	60	5		
	P	1,318	59	673	22	645	37
	R	594	36	414	27	180	9
	S	2,587	179	1,487	74	1,100	105

※ '일일 평균 영업시간' 항목을 미조사하는 산업대분류

<표 6-8>은 11개 산업대분류의 스크래핑 연계율을 나타냈다. 산업대분류 O(공공 행정)의 연계율이 8.3%로 가장 높았고, I(숙박 및 음식점업)가 7.3%, 산업대분류 S(수리

및 기타 개인서비스업)가 6.9% 순으로 나타났다. 부산시 북구와 다르게 산업대분류 E(수도, 하수 및 폐기물 처리, 원료 재생업)와 O(공공 행정)의 자료도 연계가 되었다. 다시 말하면 연계에 산업대분류별 차이는 없고, 연계 정보인 주소 정비(상세주소, 현재정보 포함)의 여부에 따라 결정된다는 것이다.

<표 6-8> 조사 대상 산업대분류별(11개) 스크래핑 연계율_경북 경산시

(단위 : %)

산업대분류	전체	현장조사	현장 미조사
E	2.0	2.1	
G	2.9	3.4	2.6
I	7.3	7.6	6.8
J	0.9	0.7	1.5
L	3.9	4.1	3.5
M	1.9	2.3	0.6
N	2.9	3.2	2.3
O	8.3	8.3	
P	4.5	3.3	5.7
R	6.1	6.5	5.0
S	6.9	5.0	9.5

<표 6-9>는 경북 경산시 '일일 평균 영업시간' 항목의 연계 자료에서 현장조사 결과와 스크래핑 결과를 비교한 결과표이다. 두 자료 결과 일치율은 현장조사 null을 제외한 연계 사업체 423개 중 179개의 데이터가 일치(42.3%)하는 것으로 나타났다. 또한, 스크래핑 결과(red color)는 현장조사보다 평균 영업시간이 긴 것으로 나타났다.

<표 6-9> 현장조사 결과와 스크래핑 결과 비교_경북 경산시

(단위 : 개)

경북 경산시		스크래핑 결과					합계
		1	2	3	4	5	
현장조사 결과	1	20	17	15	11	5	68
	2	24	54	53	18	21	170
	3	6	13	37	30	8	94
	4		4	3	19	9	35
	5	2	1	1	3	49	56
	null		5				5
합계		52	94	109	81	92	428

<표 6-10>은 경북 경산시 '일일 평균 영업시간' 항목의 연계 자료에서 현장 미조사 결과(대체)와 스크래핑 결과를 비교한 결과표이다. 두 자료는 대상 사업체 345개 중

105개가 일치(30.4%)하는 것으로 나타났고, 스크래핑 결과는 대체 결과보다 큰 값(red color)을 가지는 것으로 보인다. 이 결과는 현장조사 결과와 동일한 특징을 보인다.

<표 6-10> 현장 미조사 결과(대체)와 스크래핑 결과 비교_경북 경산시

(단위 : 개)

경북 경산시		스크래핑 결과					합계
		1	2	3	4	5	
현장 미조사 결과 (대체)	1	3	16	25	9	7	60
	2	18	44	44	15	17	138
	3	11	21	37	11	8	88
	4	2	5	9	7	6	29
	5	1	6	5	4	14	30
	null	17	10	2	8	3	40
합계		52	102	122	54	55	385

경북 경산 지역은 부산시 북구와는 달리 대학교와 산업단지가 분포되어 있어 상권 활성화 및 유동 인구가 많아 사업체뿐만 아니라 홈페이지 등록률이 높아 연계율이 높을 것으로 기대했지만 반대의 결과가 나타났다.

다시 말하면, 조사명부 연계 정보(사업체명, 주소)의 전처리 과정 없이 해당 정보를 활용(스크래팽 시 조사명부의 주소와 사업체명 바로 활용)했을 때 연계율이 현저히 낮음(상세주소 누락, 주소 비표준화)을 알 수 있었다. 5장 데이터 통합 과정을 적용한 결과와의 차이로 확인할 수 있다.

2. '배달(택배 포함) 판매 여부' 항목

<표 6-11>은 부산시 북구와 경북 경산시 '배달(택배 포함) 판매 여부' 항목의 스크래핑 자료수집 현황을 나타냈다.

부산시 북구와 경북 경산시는 전체 사업체(41,494개) 중 스크래핑(scraping)을 통해 수집된 '배달(택배 포함) 판매 여부' 항목 자료는 770개로 연계율(1.9%)이 낮게 나타났다. 현장조사(surveyed)는 23,087개 중 440개(1.9%)가 연계, 현장 미조사(unsurveyed)는 18,407개 중 330개(1.8%)가 연계되었다. 이 항목은 해당 웹사이트(네이버 지도)에서 5개 항목을 동시에 가져온 정보이고, 전체 산업대분류 중 조사 미대상 산업대분류 I(숙박 및 음식점업)가 463개로 가장 많았음을 알 수 있다.

조사 대상 산업대분류 G(도소매업) 기준에서 살펴보면, 전체 11,294개 사업체 대비 258개(2.3%)가 연계되었고, 현장조사(surveyed)는 4,293개 중 151개(3.5%)가 연계, 현장 미조사(unsurveyed)는 7,001개 중 107개(1.5%)가 연계되었다.

<표 6-11> '배달(택배 포함) 판매 여부' 항목 수집(연계) 현황

(단위 : 개)

KSIC_S		전체 사업체(부산시 북구, 경북 경산시)					
		total	scraping				
				surveyed	scraping	unsurveyed	scraping
합계		41,494	770	23,087	440	18,407	330
	미대상	30,200	512	18,754	289	11,406	223
	A	40		40			
	B	7		7			
	C	3,829	29	3,569	28	260	1
	D	920		219		701	
	E	129		119		10	
	F	3,227	3	1,253	1	1,974	2
	H	1,579	4	693	1	886	3
	I	6,611	463	3,761	250	2,850	213
	J	366		222		144	
	K	345		300		45	
	L	1,681	1	1,160		521	1
	M	1,087	1	732		355	1
	N	695	2	483	2	212	
	O	100		100			
	P	2,411		1,245		1,166	
	Q	1,514	1	1,466	1	48	
	R	1,107	1	812		295	1
	S	4,552	7	2,613	6	1,939	1
	대상	11,294	258	4,293	151	7,001	107
	G	11,294	258	4,293	151	7,001	107

※ 'scraping' 항목을 미조사하는 산업대분류

<표 6-12>는 부산시 북구와 경북 경산시의 '배달(택배 포함) 판매 여부' 항목의 연계 자료에서 현장조사 결과와 스크래핑 결과를 비교한 결과표이다.

현장조사는 null을 제외한 연계 사업체 139개 중 46개 데이터가 일치(33.1%)하는 것으로 나타났고, 현장 미조사(대체값)는 mull을 제외한 연계 사업체 92개 중 22개 데이터가 일치(23.9%)하는 것으로 나타났다.

현장조사에서는 택배(배달 포함) 판매 여부에 '판매하지 않음'이라고 응답했지만 웹페이지 공개 정보에는 '판매하고 있음'으로 파악되기 때문에 스크래핑 정보를 기준으로 재조사를 확인할 필요가 있다. 현장 미조사는 대체된 값(추정값)이기 때문에 스크래핑 연계 자료에 한해서 통계 작성에 활용 여부를 검토할 필요가 있다.

<표 6-12> 현장조사-스크래핑, 현장미조사-스크래핑 결과 비교

(단위 : 개, %)

		스크래핑	
		1	비율
현장조사	1	46	(30.5)
	2	93	(61.6)
	null	12	(7.9)
합계		151	(100.0)

		스크래핑	
		1	비율
현장미조사	1	22	(20.6)
	2	70	(65.4)
	null	15	(14.0)
합계		107	(100.0)

3. '사업체 건물 연면적' 항목(데이터 통합)

경제구조통계의 건물 연면적은 건축물대장 전유공용면적 자료를 활용하여 연계하고, 그 정보가 없으면 차선책으로 층별개요, 표제부 순으로 연계한다(4장 제1절 2).

전유공용면적 기준에서, 부산시 북구 전체 사업체(23,250개) 대상 연계는 2,894개로 연계율(12.4%)이 매우 낮았다. 그 이유는 조사명부 상세주소를 기준으로 전유공용면적 자료와 대조한 결과 전유공용면적의 누락으로 확인할 수 없었기 때문이다.

따라서, 차선책으로 층별개요 정보를 기준으로 조사명부와 건축물대장 정보와의 연계를 살펴봤다. <표 6-13>은 경북 경산시 진량읍 '사업체 건물 연면적' 항목의 자료수집 현황을 나타냈다(연구 기간을 고려하여 1개 읍면동을 추가 분석한 결과).

경북 경산시 진량읍 전체 사업체 4,194개 중 수집자료는 4,154개(99.0%)로 연계율이 높게 나타났다. 현장조사(surveyed)는 2,335개 중 2,327개(99.7%)가 연계, 현장 미조사(unsurveyed)는 1,859개 중 1,827개(98.3%)가 연계되었다.

조사 대상 산업대분류 3개(C, G, I) 기준에서 살펴보면, 전체 2,201개 사업체 대비 2,195개(99.7%)가 연계되었고, 현장조사(surveyed)는 1,483개 중 1,481개(99.9%)가 연계, 현장 미조사(unsurveyed)는 718개 중 714개(99.4%)가 연계되었다.

두 자료 간 일치율은 추가로 분석하지 않았다. 그 이유는 조사명부 기준과 건축물대장 층별개요 기준이 다르기 때문이다. 조사명부는 상세 주소(호실별) 면적인 반면, 층별개요는 건축물 내 층별 면적이기 때문이다.

건축물대장과 조사명부 간 연계는 대부분 가능할 것으로 보인다. 다만, 경제구조통계에서 활용할 정보는 층별개요가 아닌 전유공용면적 정보이다. 현재 공개되고 있는 전유공용면적 정보는 건축물 내 호실별 정보 누락이 있기에 이 부분에 대한 원인과 보완이 가능한지에 대한 확인이 필요하다. 전유공용면적 정보가 보완된다면 이 항목은 행정자료로 대체가 가능한 항목으로 보인다.

<표 6-13> '사업체 건물 연면적' 항목 수집(연계) 현황_경북 경산시 진량읍 (단위 : 개)

KSIC_S		전체 사업체(경북 경산시 진량읍)					
		total	EXCEL	surveyed	EXCEL	unsurveyed	EXCEL
합계		4,194	4,154	2,335	2,327	1,859	1,827
미대상		1,993	1,959	852	846	1,141	1,113
	A	7	5	7	5		
	B						
	D	147	141	38	38	109	103
	E	25	25	25	25		
	F	191	191	68	68	123	123
	H	739	721	143	143	596	578
	J	17	17	12	12	5	5
	K	29	29	25	25	4	4
	L	114	114	64	64	50	50
	M	76	74	59	59	17	15
	N	74	72	61	59	13	13
	O	4	4	4	4		
	P	112	112	59	59	53	53
	Q	80	80	79	79	1	1
	R	70	70	45	45	25	25
	S	308	304	163	161	145	143
대상		2,201	2,195	1,483	1,481	718	714
	C	922	922	890	890	32	32
	G	786	780	324	322	462	458
	I	493	493	269	269	224	224

※ ‘사업체 건물 연면적’ 항목을 미조사하는 산업대분류

제3절 시사점

본 연구는 6개 특성 항목을 대상으로 실증분석을 수행했다. 네이버 지도를 통한 스크래핑(3개 항목)과 공공데이터포털에서 제공하는 open API(1개 항목)를 이용해 수집한 데이터와, 일반 조사를 통해 수집된 데이터(조사 및 대체(imputation)) 간 결과를 비교했다. 수집된 데이터는 데이터 간 연계에 필수적인 고유 식별값(사업자등록번호) 정보 대신, 사업체의 일반정보인 사업체명, 대표자명, 주소 정보로 구성되었다. 이 정보들을 활

용하여 두 개의 상이한 데이터베이스(DB) 간 연계 작업이 진행되었다.

앞서 설명했듯이, 데이터 연계율이 매우 저조한 이유는 입수자료의 데이터 품질 문제, 필수정보 누락 그리고 사업체별 식별을 위한 상세 정보의 미제공을 주요 원인으로 꼽을 수 있다. 따라서 이렇게 수집된 데이터는 통계 작성에 직접적으로 활용하기보다는 조사 명부의 보조 정보 제공 등 보조적인 활용만이 가능할 것으로 판단된다.

첫째, 데이터과학 기술을 통한 자료수집의 한계를 개선하기 위해서는 민간 데이터를 확보하려는 노력이 필요하다. 현재 스크래핑(3개) 방식은 네이버 스마트플레이스에 등록된 정보를 네이버 지도 웹사이트에서 추출해 왔기 때문에, 등록자별로 표준화되지 않은 비정형화된 데이터베이스(DB)를 사용하게 되어 연계율이 낮다. 이를 개선하기 위해서는 네이버 스마트플레이스에 등록된 정형화된 DB(2023년 기준 230만 개 사업장)를 입수하여 활용해야 한다. 이 정형화된 DB를 활용한다면 5개 특성 항목의 통계 작성에 직접적으로 기여할 수 있을 것으로 판단된다.

둘째, open API(1개) 항목은 건축물대장에서 제공하는 DB를 활용할 수 있다. 이 DB는 매우 정형화된 자료이지만, 일부 정보(전유공용면적 부분)의 누락이 의심된다. 또한, 이 DB는 건물 내 층과 호실 면적은 제공하지만, 사업체를 식별할 수 있는 고유값(사업자등록번호)은 제공하고 있지 않다. 따라서 데이터 연계는 사업체의 일반정보인 주소(도로명 혹은 지번 주소), 건물명 등을 활용할 수밖에 없는 한계가 있다.

반면, 경제구조통계는 조사 명부 작성 시 전년도 조사 명부와 행정자료(분기)를 활용한다. 경제구조통계의 목적은 사업체의 규모(종사자, 매출) 및 분포를 파악하는 데 있으므로, 사업체 일반정보(상세 주소)보다는 고용 현황과 실적에 더 중점을 두고 조사한다. 이러한 이유로 상세 주소는 상대적 중요도가 낮게 취급되어, 과거 데이터의 오류를 확인하지 않고 넘어가는 경우도 종종 발생한다.

향후 데이터 기반 분석이 중요해지는 시점에서, 사업체 주소 정보는 경제 분야 고유 식별값을 대체하는 중요한 키가 될 것이다. 따라서 앞으로 조사될 경제구조통계에서는 상세 주소에 대한 엄격한 내부 검증(내검)이 필요해 보인다.

제 7 장

결론 및 제언

제1절 결론

국가데이터처의 경제구조통계는 사업체 일반정보, 종사자, 사업실적, 특성 항목으로 구성되어 있다. 이 중 특성 항목은 현장 조사를 통해 자료를 수집하며, 대규모 전수조사인 경제총조사에서는 현장에서 조사가 이루어지지 않는 사업체에 대해 대체(imputation) 방법을 활용하여 처리한다.

본 연구는 데이터과학 기술(웹 스크래핑, open API 등)을 활용하여 경제구조통계 특성 항목의 자료수집 방법을 개선하는 것을 목표로 진행하였다. 자료수집은 데이터 병합 과정 적용과 데이터 통합 과정 적용 등의 상황을 고려하여 그에 맞는 수집 및 처리 방식을 적용했다.

데이터 병합 과정 적용은 ①웹사이트를 선정(네이버 지도, 건축HUB)한 후 ②조사명부 정보(사업체명, 주소 등)를 기준(merging)으로, 분석 대상 전체 혹은 일부 지역을 대상으로 스크래핑, open API, EXCEL 등의 방식을 활용하여 자료를 수집하고, ③연계 및 일치율을 살펴보았다.

데이터 통합 과정 적용은 ①행정기관별 공개자료(공공데이터포털, 공정거래위원회) 검색 후 ②공개된 관련 자료(EXCEL)를 수집하고, ③연계에 활용할 자료들을 데이터 통합 기준에 따라 데이터를 처리한 후 ④연계 및 일치율을 살펴보았다.

연구 기간 내 자료 입수가 어려웠던 항목들에 대해서는 관련 분야 전문가 자문을 수행했다. 이 자문을 통해 특성 항목별 자료의 존재 여부, 입수 방법, 자료 내용, 자료 관리 주체, 정부 지원 사업 여부 등을 파악했다. 이 정보들은 향후 경제구조통계 개선 시 추가로 검토해야 할 과제로 남겨두었다.

본 연구에서 검토한 부분은 크게 두 가지로 나뉜다.

첫째, 조사명부를 기준으로 데이터과학 기술(스크래핑(3개), open API(1개))을 활용하여 데이터를 직접 수집하였다. 여기서 문제는 수집된 자료에 사업체 고유 식별값

인 사업자등록번호가 포함되어 있지 않았으며, 수집에 활용된 보조 정보[21]와 특성 항목 정보를 포함하는 공개 홈페이지(네이버 지도, 건축HUB)의 보조 정보 간에 차이가 있다는 점이다. 그 차이는 경제총조사 조사명부에 상세 주소(층, 호 등)의 일부 누락과 공개 홈페이지에서 제공하는 비표준화된 DB(사업체명, 주소(법정동, 행정동 혼재) 등)로 발생되었고, 결과적으로 두 데이터베이스 간의 연계율이 낮게 나타났다.

이 한계를 극복하기 위해서는 두 가지 개선 방안이 필요하다.

먼저, **경제총조사 조사명부 작성 시 상세 주소 누락을 방지하고 정확한 주소 체계를 정비해야 한다.** 외부 공공 및 민간 데이터베이스(DB)를 활용할 때 연계 정보로 가장 많이 사용되는 정보는 주소 정보이다. 특히 상세 주소는 사업체를 식별하는 유일한 키 값으로 활용 가능성이 높다. 따라서 주소 정보에 대한 내부 검증(내검)은 데이터 통합 과정에서 매우 중요한 작업이 될 것이다.

다음으로 **공공의 목적으로 민간 DB 입수를 확대**해야 한다. 네이버는 네이버스마트플레이스 등록(무료)을 통해 사업체의 정보(영업시간, 휴무일 수, 배달(택배) 여부, 객석 여부 등)를 제공하고 있다. 이 정보는 사업자등록증 확인을 통해 등록되므로 정확도가 매우 높다. 또한 많은 사업체(2023년 기준 230만 개)가 등록하고 있어 자료의 포괄 범위도 높아 안정적인 통계 작성이 가능할 것으로 보인다. 따라서 이 DB에 대한 입수는 경제구조통계를 작성하는 국가데이터처에서 기관 간 협의를 통해 추진해야 한다.

둘째, 공공데이터를 통해 추가 입수한 DB(2개)를 데이터 통합 절차에 따라 자료 처리하여 실증분석을 실시하였다. 이 연구에서 수행한 실증분석은 단순히 두 자료를 연계하는 기술적 실험을 넘어, 통계생산 체계가 직면한 구조적 조건과 제도적 맥락을 재조명하는 분석적 장치로서 기능했다.

공공데이터와 국가통계(경제총조사)를 연계하는 과정은 자료 생성 목적, 개념 체계, 범위 정의가 서로 다를 때 데이터 통합이 어떠한 제약을 받는지 그리고 어떠한 조건에서 통계적 활용 가능성이 확보되는지를 실증적으로 보여주었다. 이는 데이터 통합의 핵심 과제가 더 이상 연계 방법의 정교화나 자동화 기술에 있지 않음을 분명히 한다. 오히려 자료가 생성되는 제도적 환경, 식별자, 분류체계, 주소와 같은 기반 구조의 일관성 그리고 행정자료 생산 단계에서의 품질 관리 등 '상위 구조의 정합성'이 통합 가능성을 결정짓는 본질적 요소임이 확인되었다.

통신판매업 사례는 단순한 데이터 품질 문제를 넘어, 대량의 비활동 사업체 포함,

---

21) 경제총조사 시범예행조사의 조사명부에서 사업체 연계를 위한 보조 정보로 사업체명, 대표자명, 상세 주소를 활용함

불완전한 주소 체계, 수집 목적 차이로 인한 시간 구조의 불일치 등 신고 기반 행정자료가 갖는 구조적 문제를 보여준다. 이는 애초에 서로 다른 체계를 전제로 생성된 두 자료의 교집합(일치 범위)이 매우 좁다는 사실을 드러낸 실증적 근거라 할 수 있다.

스마트공장 사례는 정책 사업 참여라는 행정 개념과 실제 생산설비 운영이라는 조사 개념 간의 불일치가 데이터 통합의 성립을 어떻게 원천적으로 제약하는지를 명확히 보여준다. 이러한 개념적 불일치는 단순한 데이터 전처리 과정만으로는 해소될 수 없으며, 데이터 통합의 실질적인 가능성은 결국 '자료 생성 단계의 제도적 설계'에 의해 규정됨을 다시 한번 확인시켜 준다.

데이터 전처리는 이러한 구조적 제약 속에서도 연계 가능성을 최대한 확장하는 데 기여했다. 문자열 정규화, 유사도 기반 후보군 생성, 주소 구성요소의 재정렬 등은 연계의 정확성을 높였으며, 이는 데이터 전처리 작업이 데이터 통합에서 결정적 역할을 한다는 사실을 재확인하게 한다.

그러나, 이 데이터 전처리 역시 자료의 본질적 한계를 뛰어넘지는 못하였다. 데이터 전처리는 자료 간 비정합성을 조정하는 사후적 보정에 지나지 않으며, 그 실질적인 성과는 자료 생성 단계의 일관성이라는 기반 구조가 뒷받침될 때만 발휘될 수 있다. 이와 관련하여, AI 기반 전처리 알고리즘은 전처리의 효율성을 크게 향상할 수 있으나, 기술적 보완은 어디까지나 조정 능력을 확대하는 것일 뿐, 근본적 제약을 해소하는 대체재가 될 수 없음 또한 확인하였다.

종합하면, 외부 자료는 조사자료의 특성 항목을 직접 작성하거나 대체하는 자료로 활용되기에는 명확한 개념적, 구조적 제약을 가지고 있다. 특히 모집단을 구성하는 방식과 외부 자료의 생성 목적이 서로 다른 경우 데이터 통합의 실효성은 근본적으로 제한될 수밖에 없다. 그러나 동시에 외부 자료는 조사자료의 품질 점검, 결측 구조 진단, 특정 사업체의 특성 탐색 등 보조적인 활용 측면에서는 중요한 분석적 가치를 지닌다. 이러한 활용 방식은 외부 자료의 본질적 속성과 통계적 역할을 균형 있게 이해한 접근이며, 향후 국가통계 전반에서 데이터 통합을 전략적으로 활용하기 위한 현실적 방향성을 제시한다.

본 연구에서 강조하는 핵심 메시지는 "**지속 가능한 데이터 통합체계는 '기술의 정교화'가 아니라 '자료 생성 단계의 정합성'이라는 제도적 기반 위에서만 성립한다.**" 이다. 향후 외부 자료 활용을 극대화하기 위해서는 고유식별자 관리 체계의 안정화, 행정 서식의 표준화, 표기 방식의 일관성 확보 등이 반드시 구축되어야 할 최소 조건이다. 이러한 기반 구조가 강화될 때야 비로소 '데이터 기반 통계 체계'로의 성공적인 전환이 실현될 수 있을 것이다.

제2절 제언

조사 환경이 악화된 이유는 조사 응답 피로도 증가, 개인 정보 보호에 대한 우려, 기술 발전에 따른 난해한 질문 등 다양하다. 이러한 이유로 조사 통계는 점차 축소되고 다출처 데이터 연계를 활용하는 통계는 확대되고 있다.

데이터 연계를 활용한 통계는 자료 간 개념 정의, 단위, 품질이 수반되어야 활용 가치가 높다. 그러나 데이터 연계 분야에서 더 중요한 핵심 단계는 데이터 전처리이다. 사업체의 식별정보인 사업자등록번호는 데이터 연계 시 제공받지 못하는 경우가 더 많아 사업체명, 상세 주소 등이 자주 활용된다. 현재까지 이 정보들은 사업체의 일반 정보로 위치 정보를 제공했다면, 이제는 연계를 위한 매우 중요한 정보로 생각해야 한다. 본 연구에서 살펴본 사례와 실증분석 결과에서 그 중요성을 제시했다.

연구를 마무리하는 제언에서 두 가지를 살펴볼 필요가 있다.

첫째, 민간·행정기관이 보유한 품질 높은 DB의 발굴 및 입수는 국가통계 작성 개선에 반드시 필요하다. 본 연구에서 파악된 자료는 네이버(스마트플레이스), 중소벤처기업부-소상공인시장진흥공단(스마트상점), 과학기술정보통신부-정보통신산업진흥원(AI 바우처 지원 사업)이 해당한다. 이 자료들은 기관에서 자체 품질이 점검[22]된 자료이고, 자료의 포괄 범위가 넓어 데이터 통합 절차를 통해 직접 활용이 가능하다.

특히 아쉬운 점은 AI 활용 여부 항목에 대한 검토가 진행되지 못한 점이다. 본 연구에서 파악한 바로는 웹 페이지(잡코리아 등 채용 플랫폼, 전자공시시스템 등) 스크래핑과 정부 지원 사업 리스트 확보 등 활용 가능 검토를 통해 경제총조사 신규 항목의 응답 보조 정보로써 활용 가능성이 높아 보이기 때문이다.

둘째, 행정기관에서 관리하는 품질이 낮은 데이터의 체계적인 관리 및 품질 향상 지원을 위한 사업이 필요하다. 행정기관에서 공개하는 행정자료는 정보의 투명성 확보와 책임 행정을 구현하기 위한 공공재로 간주한다. 그러나 현실은 활용할 수 없는 행정자료가 비일비재하다. 이 행정자료는 통계 작성 목적이 아니라 별도의 관리를 하지 않고, 여러 행정기관별 입수 후 취합한 자료로 비표준화된 경우가 많다. 따라서 자료의 품질이 낮아 활용도가 매우 낮다. 행정기관들은 이 자료의 통계 작성 필요성이 낮고, 관리를 위한 인력 및 예산이 부족한 실정이다. 따라서 국가 데이터 중심 부처인 국가데이터처가 인프라 지원을 통해 데이터를 쉽게 관리할 수 있는 체계를 마련해 주고, 나아가 국가통계의 질적 향상을 위한 데이터 수집 방식의 다변화 전략이 필요하다.

---

22) 네이버 스마트플레이스(등록 시 사업자등록증 스캔), 스마트상점·AI바우처 지원 사업은 기업의 신고 자료를 기준으로 해당 리스트 관리

참고문헌

김민규, 박성률. (2025). **데이터 통합 방법 체계화 연구**. 통계개발원(현 국가데이터연구원).

윤영근, 오태근. (2022). “웹 스크래핑과 텍스트마이닝을 이용한 공공 및 민간공사의 사고유형 분석.” **문화기술의 융합**, 8(5), 729-734.

윤영근. (2024). “웹 스크래핑 및 텍스트마이닝에 기반한 중소규모 건설현장 사고유형 분석.” **문화기술의 융합**, 10(1), 609-615.

통계청(현 국가데이터처) 빅데이터통계과. 온라인 수집 가격 정보(2024).

Belchev, P. and Lamboray, C. (2021). “How to start with web scraping in the HICP: Evidence from EU member states”. UNECE Online meeting.

Brown, C. and Smucker, J. (2024). “Alternative data sources for high-tech products in the CPI”. *Monthly Labor Review*.

Christen, P. (2012). *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer Science and Business Media.

European Commission, (2020). *Practical guidelines on web scraping for the HICP*.

ESCAP. (2020). *Asia-Pacific guidelines to data integration for official statistics*. United Nations Economic and Social Commission for Asia and the Pacific.

Foerderer, J. (2023). “Should we trust web-scraped data?”. 10.48550/arXiv.2308.02231.

Greenhough, L. (2025). “Classification progress in the UK CPI”. ONS.

Herzog, T. N. et al. (2007). *Data Quality and Record Linkage Techniques*. Springer.

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.

Konny, C., Williams, B., and Friedman, D. (2022). “Big Data in the US Consumer Price Index”. *National Bureau of Economic Research*, p. 69-98.

Kruczek-Szepel, M. and Piatkowska, K. (2020). *ECOICOP classification*. Report of the project developed under the UNECE working group on ML in 2020 round. url:http://data.kostat.go.kr/sbchome/pageLink.do?link=/content/socialPageLink&p_num=2&curMenuNo=OPT_02_03_00_0

Williams, B. (2025). “Progress on Adopting Big Data in the US Consumer Price Index”. Meeting of the Group of Experts on Consumer Price Indices.

Sands, H. (2020). *Automated classification of web-scraped clothing data in consumer price statistics.*

Tanimichi, S. and Shibata, T. (2023). “Expanding the use of Big Data for CPI in Japan”. Meeting of the Group of Experts on Consumer Price Indices.

Taylor, L. and Keshishbanoosy, R. (2021). “Estimating computers and peripherals price indices using web-scraped data”. Meeting of the Group of Experts on Consumer Price Indices.

Thompson, B. (2018). “Web Scraping and BLS”. Federal Committee on Statistical Methodology Research Conference.

Tzimas, G., Zotos, N., Mourelatos, E., Giotopoulos, K. C., & Zervas, P. (2024). “From Data to Insight: Transforming Online Job Postings into Labor-Market Intelligence”. *Information*, 15(8), 496. https://doi.org/10.3390/info15080496

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. (2019). *Guidelines on data integration for measuring SDGs*. United Nations Economic Commission for Europe.

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.

Abstract

A Research on Improving Data Collection for Economic Structural Statistics Utilizing Data Science Technologies

Seong-ryul Park, Min-gyu Kim, Seung-woo Kwak

Currently, characteristic items in economic structural statistics are primarily collected through surveys, and non-response problems are addressed through imputation. Diversifying data collection methods for these characteristic items, which has not been attempted until now, will substantially enhance the quality of economic structural statistics. Such diversification would enable more efficient survey management, reduce data processing time and improve the treatment of non-responses. In particular, the application of data science technologies to data collection will mark a turning point, presenting a new paradigm for the production of national statistics.

This study aimed to improve data collection of characteristic items for economic structural statistics through the application of using data science technologies.

This research comprises of a review of prior studies, an assessment of the current status of open data, data collection methods utilizing data science technologies, the supplementation of characteristic items through data integration, and an empirical analysis.

Datasets collected through various channels such as scraping, open APIs, and Excel seemed to be applicable when compiling economic structural statistics, but owing to the low linkage rate, practical application had some constraints. These low linkage rates are attributable to insufficient data preprocessing. In addition, open data pose structural challenges, including conceptual discrepancies between survey and administrative data, inconsistencies arising from the absence of standardized industrial classifications, incomplete address systems, and non-standardized data entry processing by enumerators. All these problems make it difficult to link with survey data. Nevertheless, open data are considered as supplementary data by providing indirect information when producing statistics.

Therefore, to efficiently utilize administrative data from government agencies when compiling statistics, institutional foundation must be established, and data consistency must be ensured from the initial stage of administrative data production. The first step in this effort is the standardization of administrative forms.

Key words: data science, scraping, API, data integration, data preprocessing

연구진

- 박성률(국가데이터처 국가통계연구원 통계방법연구실 연구관)
- 김민규(국가데이터처 국가통계연구원 통계방법연구실 연구사)
- 곽승우(성공회대학교 미래융합학부 인공지능전공 조교수)

* 연구진의 소속 및 직급은 연구과제 완료 시 기준임을 알려드립니다.

연구보고서 2025-17

데이터과학 기술을 활용한 경제구조통계 자료수집 개선 연구

인 쇄	2026년 3월
발 행	2026년 3월
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