



## 데이터 통합을 활용한 경제구조통계 개선

- 전처리와 자료 비교 가능성이 통계 활용 가능성에 미치는 영향 -

### 박성률

국가데이터연구원  
데이터방법연구실

### 김민규

국가데이터연구원  
데이터방법연구실

- **(연구 목적)** 행정자료와 조사자료를 결합할 때 전처리와 자료 간 비교 가능성이 결합 가능성 및 통계 활용성에 미치는 영향을 분석하였다.
- **(분석 자료)** 공정거래위원회 통신판매업 자료, 경제총조사 모집단 명부 및 경제총조사 시범예행조사 자료를 이용하였다.
- **(핵심 결과)** 사업자등록번호 정확 연계율은 3.86~10.41%였으며, 사업체명·대표자명·주소를 활용한 결정적 연계의 추가 성과는 거의 없었다.
- **(핵심 해석)** 연계 레코드에서도 핵심 분석 변수의 불일치율이 높아 기술적 결합이 곧 통계적 활용 가능성을 의미하지 않음을 확인하였다.
- **(함의)** 외부 자료의 국가통계 활용을 위해 결합 알고리즘 고도화뿐 아니라 자료수집, 표준화, 전처리 및 메타데이터 관리가 선행될 필요가 있다.

### 연구 배경 및 목적

- **(연구 배경)** 경제구조통계는 경제 현황과 산업 구조를 파악하고 정책 수립과 경제 분석을 지원하는 핵심 국가통계
  - 응답률 저하, 응답 부담 증가 및 조사비용 상승으로 조사 중심 통계생산 방식의 지속가능성에 대한 문제가 제기
  - 행정·민간 자료를 기존 조사자료와 결합하여 활용하려는 시도가 확대되고 있음
  - 서로 다른 목적으로 생성된 자료는 모집단 범위, 변수 정의, 자료구조, 기준 시점 및 기록 방식이 상이할 수 있음
  - 이 차이가 충분히 조정되지 않으면 자료의 양이 증가하더라도 실제로 결합할 수 있는 레코드와 통계분석에 활용할 수 있는 레코드의 범위는 제한될 수 있음
- **(연구 목적)** 행정자료와 경제총조사 자료의 결합 과정에서 전처리와 자료의 구조적 조건이 결합 가능성 및 통계 활용 가능성을 어떻게 제약하는지 실증적으로 확인

- 📌 **(핵심 질문)** 레코드가 기술적으로 결합되었다는 사실만으로 해당 결합자료를 국가통계 작성에 활용할 수 있는가를 검토함

## 데이터 및 방법

- **(활용 데이터)** 생성 목적과 모집단 범위가 서로 다른 세 가지 사업체 자료를 활용
  - 공정거래위원회 통신판매업 자료, 경제총조사 모집단 명부, 경제총조사 시범예행조사 자료
  - 통신판매업 자료에는 영업 중인 사업체와 폐업 사업체가 함께 포함될 수 있음
  - 경제총조사 자료는 조사 목적에 부합하는 활동 사업체를 중심으로 구축됨
- **(전처리)** 자료 간 구조적 차이를 확인하고 연계 가능성을 높이기 위해 자료 진단, 스키마 정렬 및 개념 조화의 세 단계로 전처리 수행
  - **(자료 진단)** 자료별 모집단 범위와 레코드 수를 비교하고 결측치, 중복 레코드, 조사 대상 외 사업체 및 식별·분석 변수의 이용 가능성 점검
  - **(스키마 정렬)** 자료별 변수명, 자료형 및 표현 형식을 비교하고 사업자등록번호, 상호명, 대표자명 및 주소의 대응 관계를 설정
  - **(개념 조화)** 변수 개념과 코드 체계를 통일하고 상호명·주소를 표준화하며 결측, 해당 없음 및 구조적 미수집 값을 구분함
- **(이력관리)** 전처리 과정의 변수 변환과 대응 관계를 변환표와 변경 이력으로 기록하여 재현성과 추적 가능성을 확보함
- **(레코드 연계)** 정확 연계와 규칙 기반 결정적 연계의 두 단계로 수행
  - **(정확 연계)** 사업자등록번호가 정확히 일치하는 레코드를 동일 사업체로 판정
  - **(결정적 연계)** 미연계 레코드에 대해 사업체명, 대표자명, 주소를 대상으로 Jaro-Winkler 유사도 0.90 이상인 경우에만 연계
- **(평가 지표)** 연계율, 추가 연계율, 일치율, 불일치율, 정밀도, 재현율 산출
- **(해석 유의)** 별도 수작업 검증자료가 없어 정밀도와 재현율은 사업자등록번호 정확 연계 결과를 참조 집합으로 한 일관성 지표로 해석함

## 주요 결과

- 결과는 결합 단계에서 갑자기 문제가 발생한 것이 아니라, 결합 이전의 모집단 차이와 변수 구조의 불완전성이 전처리 손실, 낮은 연계율, 결정적 연계의 제한 및 연계 후 변수 불일치로 이어지는 누적 구조를 보임
- **(전처리) 전체 자료의 약 29~44% 제외**
  - 폐업 사업체, 조사 대상 외 사업체, 중복 레코드 및 주요 변수 누락 자료를 제외한 결과, 자료별로 전체 레코드의 28.73~43.87%가 분석 대상에서 제외

**<표1> 자료별 규모 및 제외 현황**

| 지 역    | 자 료    | 전체 레코드 | 제외 레코드 | 제외율    | 활용 레코드 |
|--------|--------|--------|--------|--------|--------|
| 부산 북구  | 통신판매업  | 5,925  | 2,534  | 42.77% | 3,391  |
|        | 모집단 명부 | 23,250 | 7,455  | 32.06% | 15,795 |
|        | 시범예행조사 | 9,837  | 3,423  | 34.80% | 6,414  |
| 경북 경산시 | 통신판매업  | 10,769 | 3,094  | 28.73% | 7,675  |
|        | 모집단 명부 | 34,844 | 12,893 | 37.00% | 21,948 |
|        | 시범예행조사 | 16,418 | 7,202  | 43.87% | 9,216  |

## 주요 결과 (계속)

- 주요 제외 사유
  - 통신판매업 자료는 폐업 사업체
  - 경제총조사 모집단 명부와 시범예행조사 자료는 조사 대상 범위에 포함되지 않는 사업체가 주된 제외 사유
- 이는 자료별 모집단 범위가 동일하지 않으며, 결합 이전 단계에서 이미 실제 결합 가능한 레코드의 범위가 크게 제한되었음을 의미함

### ○ 결합 변수와 분석 변수의 완전성이 서로 다르게 분포

- 결합에 사용되는 변수와 통계분석에 필요한 변수의 완전성이 동일한 자료에 함께 확보되지 않았음
- **(결합 변수)** 통신판매업 자료의 사업자등록번호 결측률은 부산 북구 0.03%, 경북 경산시 0.29%로 낮았음
- **(분석 변수)** 시범예행조사 자료의 온라인 거래 여부 결측률은 부산 북구 37.09%, 경북 경산시 54.25%로 높았음
- 결합 변수가 완전한 자료에는 분석 변수가 부족하고, 분석 변수를 포함한 조사 자료에는 결측이 많아 레코드 연계와 통계 활용을 동시에 충족하는 자료 범위가 추가로 축소됨

<표2> 자료 및 변수별 결측 현황

| 자 료    | 변 수       | 부산 북구  | 경북 경산시 |
|--------|-----------|--------|--------|
| 통신판매업  | 사업자등록번호   | 0.03%  | 0.29%  |
|        | 주소        | 0.12%  | 0.31%  |
| 모집단 명부 | 사업자등록번호   | 4.00%  | 1.96%  |
|        | 대표자명      | 0.09%  | 0.05%  |
| 시범예행조사 | 사업자등록번호   | 3.95%  | 1.92%  |
|        | 온라인 거래 여부 | 37.09% | 54.25% |

### ○ (정확 연계) 사업자등록번호를 이용한 연계율은 3.86~10.41%에 그침

- 고유식별자인 사업자등록번호가 존재함에도 연계율이 낮게 나타난 것은 동일 사업체가 각 자료에 동시에 포함되는 비율이 낮았기 때문임
- 즉, 연계율 제한은 식별자 성능보다 자료 간 모집단 비중첩의 영향을 크게 받았음

<표3> 통신판매업 자료 기준 정확 연계 결과

| 자 료    | 결합 대상  | 연계 레코드 | 연계율    |
|--------|--------|--------|--------|
| 부산 북구  | 시범예행조사 | 163    | 4.81%  |
|        | 모집단 명부 | 353    | 10.41% |
| 경북 경산시 | 시범예행조사 | 296    | 3.86%  |
|        | 모집단 명부 | 588    | 7.66%  |

## 주요 결과 (계속)

### ○ (결정적 연계) 규칙 기반 결정적 연계의 성과는 거의 나타나지 않음

- 미연계 레코드를 대상으로 사업체명, 대표자명 및 주소를 활용, 결정적 연계 수행
- 추가 연계 건수는 경북 경산시의 모집단 명부 1건을 제외하고 모두 0건

<표4> 통신판매업 자료 기준 결정적 연계 결과

| 자료     | 결합 대상  | 미연계   | 추가 연계 레코드 | 추가 연계율 |
|--------|--------|-------|-----------|--------|
| 부산 북구  | 시범예행조사 | 3,228 | 0         | 0.00%  |
|        | 모집단 명부 | 3,038 | 0         | 0.00%  |
| 경북 경산시 | 시범예행조사 | 7,379 | 0         | 0.00%  |
|        | 모집단 명부 | 7,087 | 1         | 0.01%  |

- 이는 미연계가 단순한 표기 차이에서 발생한 것이 아니라 모집단 비중첩, 식별정보 누락 및 자료 표현의 구조적 차이에서 비롯되었음을 보여줌
- 따라서 연계 알고리즘을 정교화하는 것만으로는 결합 성과를 크게 개선하기 어려움

### ○ 연계 레코드에서도 분석 변수의 일관성이 낮음

- 시범예행조사 자료와 연계된 레코드의 온라인 거래 여부를 비교한 결과
  - (변수 일치율) 부산 북구 27.53%, 경북 경산시 16.72%
  - (불일치율) 부산 북구 72.47%, 경북 경산시 83.28%

<표5> 통신판매업 자료 기준 연계 품질 점검 결과

| 자 료    | 분석 변수 일치율 | 불일치율   | 정밀도    | 재현율    |
|--------|-----------|--------|--------|--------|
| 부산 북구  | 27.53%    | 72.47% | 88.67% | 23.93% |
| 경북 경산시 | 16.72%    | 83.28% | 90.91% | 13.56% |

- (연계 성능) 엄격한 기준을 적용함에 따라 정밀도는 비교적 높게 나타났으나, 재현율은 낮음
  - 오연계는 줄일 수 있었지만 실제로 연결되어야 할 레코드의 상당 부분을 식별하지 못했음을 의미
  - 동일한 사업체로 연계된 경우에도 온라인 거래 여부가 서로 다른 비율이 높음
- 레코드가 동일 사업체로 연계되었다는 사실만으로 결합자료의 통계적 활용 가능성을 판단할 수 없으며, 결합 이후에도 핵심 분석 변수의 일관성을 별도로 검증해야 함을 의미함

## 시사점

- **데이터 통합의 성과는 결합 방법보다 자료 조건에 크게 좌우됨**
  - 자료 간 모집단 범위와 변수 구조가 일치하지 않으면 고유식별자가 존재하거나 결합 알고리즘을 고도화하더라도 결합률을 크게 개선하기 어려움
  - 데이터 통합의 가능 범위는 결합 단계 이전의 자료 수집과 전처리 과정에서 상당 부분 결정
- **기술적 결합 가능성과 통계적 활용 가능성을 구분해야 함**
  - 동일 사업체가 기술적으로 결합되더라도 핵심 분석 변수가 누락되거나 자료별 값이 일치하지 않으면 통계분석에 활용하기 어려움
  - 품질을 평가할 때 결합률뿐만 아니라 모집단 포괄성, 분석 변수의 완전성 및 결합 후 변수 일관성을 함께 평가해야 함
- **외부 자료의 국가통계 활용을 위해 자료 표준화가 선행되어야 함**
  - 외부 자료를 국가통계에 활용하기 위해서는 사업체명, 대표자명, 주소 및 사업자 상태와 같은 핵심 변수의 정의와 표현 방식을 표준화할 필요가 있음
  - 자료 제공기관과 통계작성기관 간 메타데이터 기준을 마련하는 것도 중요함
- **전처리를 단순한 자료정제 작업이 아닌 품질 통제 단계로 관리해야 함**
  - 전처리는 결합을 위한 기술적 준비 과정에 그치지 않고 실제로 결합이 가능한 자료와 분석이 가능한 자료의 범위를 결정하는 핵심 통제 단계
  - 변수 대응 관계, 변환 규칙, 제외 기준 및 결측 처리 방법을 명시하고 변경 이력을 체계적으로 관리해야 함
- **데이터 통합의 품질관리는 전체 통계 생산 과정에서 수행되어야 함**
  - 자료수집, 개념 정의, 표준화, 전처리, 결합 및 결과 검증은 상호 독립된 절차가 아님
  - 초기 단계의 자료 불일치가 이후 결합과 분석 단계로 전이될 수 있으므로, 데이터 통합의 품질관리는 개별 결합 방법이 아니라 전체 통계 생산 과정과 데이터 관리 책임 체계(data stewardship)를 중심으로 설계할 필요가 있음

## 한계 및 향후 연구

- **본 연구 결과를 전국 또는 다른 산업 분야에 직접 일반화하는 데 한계가 있음**
  - 별도의 수작업 검증자료가 확보되지 않아 사업자등록번호를 이용한 정확 연계 결과를 참조 자료로 사용
  - 제시한 정밀도와 재현율은 실제 결합 정확도보다는 참조 자료와의 일관성을 나타내는 지표로 해석해야 함
  - 통신판매업 자료의 수집 시점, 갱신 주기 및 운영상 오류가 결합 결과에 미치는 영향도 충분히 분석하지 못함
- **향후 연구에서는 다음 사항을 검토할 필요가 있음**
  - 전국 단위 및 다양한 산업 분야로 분석 범위 확대
  - 수작업 검증 표본을 활용한 연계 정확도 평가
  - 연계 임계값에 대한 민감도 분석
  - 확률적 연계 및 기계학습 기반 연계 방법과의 비교
  - 자료수집 시점과 갱신 주기를 고려한 시점 정합성 평가
  - 메타데이터 표준화 방안 마련
  - 데이터 이력 및 처리 과정을 관리하는 데이터 계보(data lineage) 체계 구축
  - 결합 이후 결측 대체와 통계적 보정의 적용 가능성 검토

## 결론

- 본 연구는 낮은 연계율 자체보다 그 원인이 결합 이전의 모집단 차이, 변수 완전성 부족 및 자료 표현의 구조적 차이에 있다는 점을 실증적으로 확인
  - 또한 연계된 레코드에서도 핵심 분석 변수의 불일치가 높아, 결합 성공과 통계적 사용 가능성을 동일하게 간주해서는 안 됨
- 외부 자료를 국가통계에 활용하기 위해서는 결합 알고리즘의 고도화만을 우선할 것이 아니라, 자료수집 단계의 모집단 정의, 변수 표준화, 전처리 규칙, 메타데이터 및 처리 이력, 결합 후 변수 일관성 검증을 포함하는 전 과정 품질관리 체계를 구축해야 함
- 데이터 통합의 핵심은 자료를 결합하는 데 있는 것이 아니라, 결합된 자료가 통계적으로 해석 가능하고 재현 가능한 방식으로 관리되었는지를 확인하는 데 있음

## 참고 문헌

- Christen, P. (2012). Data matching: Concepts and techniques for record linkage, entity resolution, and duplicate detection. Springer
- Eurostat. (2019). Quality assurance framework of the European Statistical System: Version 2.0. Publications Office of the European Union
- Herzog, T. N., Scheuren, F. J., & Winkler, W. E. (2007). Data quality and record linkage techniques. Springer
- Rahm, E., & Bernstein, P. A. (2001). A survey of approaches to automatic schema matching. The VLDB Journal, 10(4), 334-350
- Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. IEEE Data Engineering Bulletin, 23(4), 3-13
- United Nations Economic Commission for Europe. (2019). Generic Statistical Business Process Model: GSBPM version 5.1
- United Nations Economic Commission for Europe. (2020). A guide to data integration for official statistics: Version 2.0
- Winkler, W. E. (2006). Overview of record linkage and current research directions (Research Report Series No. RRS2006/02). U.S. Census Bureau

⇒ (연구보고서 원본) 국가데이터연구원홈페이지 > 연구간행물> 연구보고서 > 데이터과학 기술을 활용한 경제구조통계 자료수집 개선 연구

출처를 밝히지 않고 국가데이터연구원 데이터 리서치 브리프의 내용을 무단전재 또는 복제하는 것을 금합니다.  
본 리서치 브리프의 내용은 필자들의 개인적 의견이며, 국가데이터연구원의 공식적인 의견이 아님을 밝힙니다.

발 행 일 2026년 7월 31일

발 행 처 국가데이터처 국가데이터연구원

발 행 인 김 진

주 소 35220 대전광역시 서구 한밭대로 713 통계센터 6~8층

전화번호 042-366-7118

홈페이지 [mods.go.kr/dsri](https://mods.go.kr/dsri)