제6장

통계이용 활성화를 위한 2차 자료 생산·활용 방안 연구

심규호㊟박시내

제1절 서 론

올바른 국가 정책의 수립을 위해서는 무엇보다 대표성과 신뢰성을 확보한 국가통계의 뒷받침이 절실하다. 2010년 8월 1일 통계법 제18조(또는 제20조)에 의거하여 승인받은 통계는 총 850종으로 지정통계 90종, 일반통계 760종이며, 작성형태별로는 조사통계 342종, 보고통계[1]는 452종, 가공통계[2]는 56종이다. 승인통계 중 정부기관에 의해 작성되고 있는 통계는 702종이며, 이중 통계청에서 생산하고 있는 통계는 52종[3]으로 정부기관 작성통계 중 7.4%를 차지한다.

통계청에서 작성되는 통계는 인구 및 경제·사회 다양한 영역을 포괄하며[4], 표본의 대표성 및 신뢰성을 확보하고 있다는 장점을 갖는 반면 하나의 자료 안에 수록하고 있는 정보는 제한적이다. 반면 외부 연구기관에서 수행되는 패널조사의 경우 표본 수는 매우

---

1) '보고통계'란 국가나 지방단체가 법령에 의거해 개인이나 단체 등에서 의무적으로 제출한 보고, 신고, 등록, 신청, 인허가 등과 같은 행정업무에 수반하여 수집된 자료(administrative data) 혹은 등록 자료(register)를 토대로 작성된 통계로 등록기반통계(register-based statistics) 혹은 행정통계라고 부르기도 한다.

2) '가공통계'란 조사 또는 보도통계를 활용하여 새로운 형태의 지수, 지표 등으로 작성한 통계로 경기 종합지수나 국민계정 등을 말한다.

3) 통계청에서는 현재 52종의 통계가 작성되고 있다(2010.12.1.기준). 이 중 조사통계는 42종, 가공통계는 8종, 보고통계는 2종으로 조사통계는 전수조사 13종과 표본조사 29종이 있다. 조사통계 중 전수조사(13종)의 작성주기는 5년, 2년, 매년, 매월/분기 등 다양한데, 5년 주기로 작성되는 통계는 인구주택총조사 등 4종이며, 2년주기는 통계인력 및 예산조사, 매년 작성되는 통계는 광업·제조업조사, 전국사업체조사 등 5종, 매월/분기별로 작성되는 통계는 전자상거래동향조사 등 3종이다. 조사통계 중 표본통계는 총 29종으로 5년 단위로 작성되는 통계는 생활시간조사, 농림어업복지실태조사, 매년 작성되는 통계는 사교육비조사, 사회조사 등 16종이 있으며, 매월/분기별로 작성되는 통계는 경제활동인구조사, 가축동향조사 등 11종이다. 가공통계는 총 8종으로 5년 주기 작성통계는 장래인구추계, 장래가구추계이며, 매년 작성되는 통계는 국가자산통계, 지역소득 등 4종, 매월 작성되는 통계는 경기종합지수, 설비투자지수이다. 보고통계는 총 2종으로 매년 작성되는 국제인구이동통계와 매월 작성되는 국내인구이동통계가 있다.

적은 반면 수록정보는 매우 방대하다는 특성을 갖는다. 보다 신뢰도 높은 연구의 수행과 정책 수립을 위해서는 충분한 자료의 확보가 매우 중요하지만, 하나의 자료에서 분석에 필요한 충분한 정보를 얻는다는 것을 매우 어려운 일이다. 그러나 이러한 문제(단일자료 정보의 불충분성)는 데이터 매칭(*data matching*) 또는 데이터 통합(*data fusion*)을 통해 상당 부분 보완이 가능하다. 일반적으로 조사된 데이터에는 가구 식별번호 및 개인 식별번호, 나이, 성별 등 공통적으로 포함된 항목이 있으며, 공통 항목을 통해 다수의 데이터를 통합하면 보다 많은 정보를 얻을 수 있다.

최근 통계청 마이크로 데이터 이용에 관한 외부 수요자의 설문조사 결과에 따르면 마이크로 데이터 이용 시 어려운 점으로 상세정보의 부족과 필요 변수의 부재, 데이터 가공의 어려움, 이용 자료의 부족 등을 꼽고 있으며, 이를 해결하기 위해 담당자 문의, 보조정보 활용, 타 데이터 연계 및 예측값 적용을 시도하였다고 응답하고 있다. 또한 단일 통계자료의 정보 부족을 해결하기 위하여 통계청 내 데이터 간 연계를 시도해봤다는 응답 비중 또한 상당히 높았다. 이 같은 조사 결과는 마이크로 데이터 이용자들이 단일 데이터의 정보 한계를 극복하기 위하여 이미 다양한 노력을 시도하고 있음을 의미한다.

본 과제는 이 같은 외부 환경에 대응하고, 기존 통계 간 연계를 통해 2차 통계(연계통계)를 생산함으로서 기존 통계의 활용도를 높이고, 저비용·고효율의 통계생산을 도모하기 위한 목적을 갖는다. 또한 결합자료를 구축하고, 매칭과정에서 발생하는 다양한 문제에 대한 대비책 또한 논의하고자 한다. 갈수록 조사환경이 열악해지고, 연계통계에 대

〈표 6-1〉 통계작성 현황

(2010. 12. 1. 기준)

작성방법		작성주기	통계명칭
조사 통계 (42종)	전수조사 (13종)	5년	° 인구주택총조사 등 4종
		2년	° 통계인력 및 예산조사
		매년	° 광업제조업조사, 전국사업체조사 등 5종
		매월/분기	° 전자상거래동향조사 등 3종
	표본조사 (29종)	5년	° 생활시간조사, 농림어업복지실태조사
		매년	° 사교육비조사, 사회조사 등 16종
		매월/분기	° 경제활동인구조사, 가축동향조사 등 11종
가공통계 (8종)		5년	° 장래인구추계, 장래가구추계
		매년	° 국가자산통계, 지역소득 등 4종
		매월	° 경기종합지수, 설비투자지수
보고통계 (2종)		매년	° 국제인구이동통계
		매월	° 국내인구이동통계

4) 외부 수요자를 대상으로 조사된 결과에 의하면 통계청 마이크로 데이터 이용 분야는 인구·가구 51.2%, 고용·노동·임금 27.9%, 보건·사회·복지 27.9%, 광공업·에너지 23.3%, 농림어업 14.0%, 물가·가계 11.6% 순으로 나타난다(중복응답 허용).

한 외부 수요자들의 관심이 증대되어 온 바 이에 대비하기 위해서는 통계청 내부의 노력이 매우 필요할 것으로 보인다.

본 연구의 구성은 다음과 같다. 첫째, 연계통계 생산 및 활용에 대한 해외 사례(호주)를 검토하고, 매칭 프로그램인 febrl에 관해 소개하고자 한다. 또한 자료 연계 및 활용에 관한 기존사례를 논의하고, 2차 자료 수요에 대한 전문가 수요조사 결과를 제시한다(제2절). 둘째, 통계청 자료 중 연계 가능한 자료를 검토하고, 다양한 매칭 기법에 관해 논의한다(제3절). 셋째, 『가계동향』-『경제활동인구조사』 간 자료매칭 과정 및 결과를 설명하고, 연계자료의 주요 특성치를 요약한다(제4절). 마지막으로 본 연구의 요약과 한계점을 제시한다(제5절).

제2절 선행사례 검토

1. 해외사례

가. 호주

1) 인구보건 및 임상자료의 연계

가) 배경과 적용

호주 통계청(ABS)은 인구보건 및 임상 데이터 연계를 통한 보건 분야 연구에 있어서 선두적인 위치를 차지하고 있다. 이에는 대규모의 인구 데이터베이스, 연계 데이터 활용에 대한 호의적인 정책과 입법 환경, 그리고 우수한 연구능력에 기인한다.[5] 인구보건 및 임상 데이터 연계를 통해 얻을 수 있는 이점[6]은 기존의 종단적 연구나 역학, 의료 서비스 연구에서의 전통적인 접근과 비교하여 보다 저비용으로 효율적인 연구 성과를 얻을 수 있다는 점이다. 또한 기존 정보에 부가된 정보가 추가되면 보건 분야 연구에 대한 연구 수요 및 투자를 활성화시킬 수 있으며, 보건 분야 이외의 다른 분야로부터의 연계를 통해 데이터의 활용도를 더욱 높일 수 있다. 또한 연계 데이터 분석 결과를 통해 인구보건, 임상, 생물 의학의 각 부문 간 협력 분위기도 조성될 수 있다. 특히 *Investment of Review of Health and Medical Research*(2004)는 연계를 통한 연구와 정책 도출이 호주 보건 시스템에 중요한 이익을 가져올 수 있으며, 보건 예산부문의 효율성을 증대시킬 수 있다고

5) Australian Government. *Sustaining the virtuous cycle for a healthy, competitive Australia. Investment Review of Health and Medical Research.* Final report. Canberra: Commonwealth of Australia(2004).

6) "Data Linkage Australis. *Scoping paper: a model for a data linkage facility in New South Wales.* Sydney: the Sax Institute(2005).

지적하였다.[7] 데이터 연계를 위하여 국가적인 대규모 투자가 있었던 데에는 데이터 연계가 공공보건과 역학연구, 의료 서비스 및 임상 연구 분야 발전에 중요한 공헌을 할 것으로 기대되었기 때문이다.

호주의 인구보건 및 임상 데이터 연계의 적용 범위는 다음과 같다. 첫째, 의료 서비스의 사용과 비용에 관련된 부문(가령, 특정 조건의 환자에 대한 병원 치료와 병원 치료의 비용과 패턴의 연구)이다. 둘째, 의료 서비스 이용과 사망·질병 리스트와의 연계를 통한 의료 서비스 효과의 다양한 모델의 검증이다. 셋째, 특정 임상 및 치료법의 결과(예를 들어 수술 후 생존이나 의약품의 부작용에 관한 판매 후 조사결과)에 적용될 수 있다. 넷째, 질병 발생요인의 분석과 연구(가령 부수적인 건강 조건 이전에 존재하는 조건과 위험요인의 기준 정보를 관련시키기 위해 패널 연구 대상자들로부터의 질문 정보들을 질병 명부와 연계하여)를 수행할 수 있다. 다섯째, 보건 데이터와 다른 분야(노인치료, 지역사회 치료 서비스 등)의 데이터 통합을 통한 전반적인 건강상태와 사회적 요인과의 관계를 분석할 수 있다. 여섯째, 특정 환경에 노출된 개인자료와 건강 데이터의 결합을 통한 환경적 요인과 건강과의 관계를 연구할 수 있다. 일곱째, 질병 리스트와 유전정보와의 연계를 통해 생물학적 요인과 건강과의 관계를 알 수 있다. 여덟째, 사망 자료와 질병 리스트와의 연계를 통해 임상 실험자들을 대상으로 장시간의 추적 연구결과를 축적할 수 있다.

나) 현황

호주의 데이터 연계 시스템(*the Western Australian Data Linkage System*)은 1995년부터 운영되어 왔다. 1995년부터 2003년까지 이 시스템은 258개의 프로젝트에 데이터를 제공해왔으며, 172개의 저널 기사를 포함하여 708개의 연구 성과물을 냈다. *Western Australian Data Linkage System*의 연계 데이터를 이용한 연구는 정책과 임상 치료에서의 일련의 변화를 가져왔으며[8], 유사한 방법의 지역 기반 시스템이 *New South Wales*와 *Australian Capital Territory*의 지역 서비스를 위해 만들어졌다. 국가나 주의 연구자들은 다양한 데이터의 연계 작업을 수행한다. 예로 *Commonwealth and Victorian State Goverment*는 게놈 데이터와 임상 데이터를 연결해주는 연방 데이터 그리드(Bio 21 : MMIM[9])에 투자해왔고, *Commonwealth Scientific and Industrial Organisation(*CSIRO*)*는 데이터 연계를 위한 서비스 지향적인 소프트웨어를 개발하고 있다. 한편 *New South Wales Government*와 비정부 협력체들은 데이터 연계를 통해 패널의 건강상태를 추적한 연구 사업에 투자해왔다.[10]

---

7) Australian Government. *Sustaining the virtuous cycle for a healthy, competitive Australia. Investment Review of Health and Medical Research.* Final report. Commonwealth of Australia(2004).

8) Brook EL, Rosman DL Holman CDJ, Trutwein B. *Summary report: Research output project, WA Data Linkage Unit(1995~2003).* Perth: WA Data Linkage Unit(2005).

9) http://mmim.ssg.org.au/

호주는 데이터 연계를 통한 연구 분야에서 국제적으로 선두적인 입지 마련을 위한 기반을 잘 갖추어 왔다. 특히 호주는 국가 역량을 제고하기 위한 특별한 기회를 갖추고 있다. 호주의 보건 시스템은 매우 자세한 정보를 담고 있는 고품질의 통계이며, 호주의 개인 정보 보호법은 데이터 연계의 기초가 되는 한편 개인 정보 보호를 위한 확고한 기반을 제공한다. 또한 다른 선진국의 데이터 연계 시스템(Manitoba, British Columbia, Oxford 등)이 다루는 인구자료보다 호주의 인구자료는 상대적으로 크고 매우 다양하다.[11] 국가 보건 데이터 셋은 Australian Health Minister's Advisory Council(AHMAC)의 National Health Information Group을 통해 9개의 Australian Government에 의해 보관된다. 이렇게 수집된 메타-데이터는 National Health Data Dictionary와 Australian Institute of Health and Welfare(AIHW)의 전자보관소에서 이용이 가능하다. National Health Information Management Principal Committee의 소위원회인 Statistical Information Management Committee(SIMC)는 최근 연구목적의 보건 데이터 세트들 간의 연계를 위한 관할권의 상태를 문서화하기 위한 프로젝트를 끝마쳤고, 현재 국가적 차원의 연계를 위한 체제 공사를 진행 중에 있다. 이전에 호주통계청(ABS)은 인구조사 데이터에 제약된 연결만을 허용했지만, 2006년 인구조사부터 사회·경제 분야의 광범위한 자료의 연계를 가능케 하였다.

데이터 연계와 활용 및 자료 제공에는 원칙이 필요하다. 연계된 데이터는 오직 연구와 통계적 목적으로만 활용되어야 하며, 개인적 목적으로 사용되어서는 안 된다. 또한 개인정보는 절대로 유출되어서는 안 되며, 데이터 연계 처리 방식에 대한 합의가 있어야 한다. 또한 연계자료 이용의 활성화를 위해서는 연구자들의 광범위한 접근을 위한 기반시설을 필요로 하며, 이를 위해서는 데이터 제공과 이용자 간의 협력을 증진시킬 수 있는 적절한 관리 시스템이 필요하다.

다) 기초/기반 시설

연구자들이 이용 가능하고, 유용한 연계 가능 데이터를 만들기 위해서는 국가적인 수준의 기반시설의 지원이 매우 중요하다. 하지만 연관된 데이터 셋의 크기가 매우 크고, 내용이 방대하며, 데이터의 소유권의 관리기관이 상이함으로 인하여 많은 어려움을 겪게 된다. 앞서 언급한 대로 호주는 인구 보건 데이터와 임상 데이터 셋의 연계를 위한 고도화된(그러나 국지적인) 시스템을 갖고 있다. 그러나 전국적인 인구 데이터를 완전히 포괄하고, 국가 수준의 인구 자료와의 연계를 촉진시키기 위해서는 자료 연계와 연계된 자료의 분석을 위한 방법과 도구, 인력 투자 등 기반시설에 대한 상당한 투자가 요구된다. 연구를 위한 지지기반에의 특별한 투자는 보건 분야의 일련의 활동에 의한 지지와

10) http://www.45andup.org.au/

11) 현재 호주 이외에 데이터 연계를 이용해 국가적 차원의 광범위한 인구보건 연구를 할 수 있는 국가는 스코틀랜드가 유일하다.

*National Health and Medical Research Council*의 리더쉽을 필요로 한다. 데이터 연계를 통한 연구기반 시설 확충을 위한 향후 추진방향은 다음과 같다.

1. 국가적 보건 데이터 연계 기반시설은 국지적 수준 및 국가적 수준을 포괄한다.
   - 사망이나 병원퇴원 등 국지적으로 관리되는 자료들을 국가적으로 연계 가능한 데이터 셋으로 발전시킬 것이다.
   - 국가 당국은 데이터 관리자로부터 연계 가능한 데이터로의 접근 · 승인을 용이하게 할 것이며, 공공복지 데이터의 발표가 공공의 이익을 위한 연구 프로젝트임을 확신시킬 것이다.
   - 당국은 국가적 데이터 셋과 승인된 데이터 연계를 수행할 것이며, 이것들은 연계 가능한 형태로 이용 가능할 것이다.
   - 주정부는 소유하고 있는 특수한 관할 데이터와 지역 연구 데이터 셋과 관련된 연계 서비스를 제공할 것이며, 주정부와 국가는 기존의 연구 데이터 셋과 비정부 소유의 데이터 간 연계를 촉진할 것이다.
   - 데이터 연계뿐만 아니라 주정부와 국가는 연계 데이터로의 접근과 사용을 뒷받침하는 기반시설(승인 서비스나 툴, 쿼리와 탐색, 데이터 분석과 품질관리 등)을 제공할 것이다.

2. 보건 데이터 연계 발전을 위한 국가 기구는 다음의 내용을 포함한다.
   - 국가 보건 데이터 연계 시스템의 협력
   - 보건 데이터의 연계와 효율적인 승인과정에 기초한 연구의 적합성을 보장하기 위한 NHMRC와의 협업을 포함하는 방법론의 표준화와 발전
   - 데이터 셋에 대한 접근과 데이터 연계를 돕는 Data-Base Structure를 위한 협상
   - 국가적 소프트웨어 발전과 획득을 포함하는 방법론의 표준화와 발전
   - 데이터 셋 기술자와 저장소, 비지니스 용어들의 유지와 발전
   - 메타 데이터 기술자와 저장소, 비지니스 용어들의 유지와 발전
   - 서비스 지향적인 아키텍쳐, 개인정보보호 분석과 비정형화된 데이터의 조작 방법을 포함하는 새로운 방법이나 기술적 솔루션과 관련된 지식의 모음과 공유

3. 인재양성은 다음의 내용을 포함한다.
   - 국가적 연구 능력의 발전 및 필요한 인력을 양성하기 위한 국가적 자금 계획
   - 기존 연구진들의 데이터 연계와 연계된 데이터 셋 분석을 위한 훈련

2) Febrl[12] program 소개

레코드나 데이터 연계는 의료분야 연구 및 정책수립에 매우 중요한 요소이다. 연계 데이터의 활용을 통해 의료정책 관련 연구를 향상시키고, 의약품의 부작용을 발견하며, 비용을 절감하고, 의료 시스템의 부정을 적발할 수 있는 효율적 자원이 되기 때문이다. 최근 몇 년간 레코드 연계 기법의 많은 부분에서 중대한 발전이 일어났고, 이들 대부분은 데이터 마이닝이나 머신러닝 분야로부터 비롯되었다. 이들 새로운 방법의 대부분은 아직 현 레코드 연계 시스템에 구현되지 않았거나, 상업용 소프트웨어 내에 감춰져 있어 이해 불가하다. 이것은 사용자들에게 새로운 레코드 연계 기술에 대한 학습뿐 아니라, 기존 연계 기술을 새로운 기술과 비교하는 것 또한 어렵게 한다. 따라서 사용자들이 새로운 레코드 연계 기법을 낮은 가격에 학습하고 실험해볼 수 있는 유연한 tool이 요구된다.

여기서는 오픈소스 소프트웨어 라이센스로 이용 가능한 *Febrl(Freely Extensible Biomedical Record Linkage)* 시스템에 대해 소개하고자 한다. 이 시스템은 data cleaning과 표준화, 인덱싱(블로킹), 필드 비교, 레코드 쌍 분류를 위해 최근 개발 · 발전된 여러 기법들을 포함하며, 이것들 모두를 *GUI*로 담고 있다. 그러므로 *Febrl*은 사용자들이 기존의 레코드 연계 기법과 새로운 레코드 연계 기법 모두를 학습하고 실험할 수 있게 해주는 훈련 도구일 뿐 아니라 연구자들로 하여금 수백만 개의 데이터 세트를 포함한 연계 작업 수행을 가능토록 해준다.

의료분야의 많은 민간기관과 공공기관이 급속히 불어나고 있는 수백만 개의 레코드 데이터를 수집, 저장, 처리, 분석하고 있다. 이들 데이터의 대부분은 사람(지역 보건의 혹은 병원의 환자, 민간 의료 보험 회사의 회원 등)에 대한 정보이며, 이름과 주소, 그리고 의료와 관련된 개인 세부사항 등을 포함하고 있다. 몇 개의 데이터베이스로부터 추출한 동일인에 관련된 레코드의 수집과 연계는 날로 더 중요해지고 있다. 이는 다른 방법으로는 연계된 데이터가 이용가능하지 않은 관계로 연구가 불가능하거나, 그렇지 않으면 시간의 소요가 많고, 값비싼 조사 방법을 이용해야 하기 때문이다. 의료 분야에서 레코드 연계는 주요한 관심사이다. 연계된 데이터가 조사비용을 감소시키고, 의약품의 부작용을 발견하며, 역학 연구에서의 고가의 조사 데이터를 대신하여 사용될 수 있기 때문이다.[13]

최근 호주에서는, 연합 정부의 국가적 협동 연구 인프라 구축 전략(*National Collaborative Research Infrastrucrue Strategy - NCRIS*)[14]상의 12개의 역량 영역 중의 하나인 인구

12) Freely Extensible Biomedical Record Linkage의 약자로 호주 통계청이 인구보건 데이터와 임상 데이터의 연계를 위해 개발한 프로그램이다.

13) 예를 들어 병원 데이터, 사망 신고서와 연계된 구급차 심장마비 데이터베이스에 기초한 Western Australia의 연구는 구급차와 병동에의 제세동기 설치를 이끌었으며, 이로써 많은 생명을 구하게 되었다.

14) http://www.ncris.dest.gov.au

보건과 임상 데이터 연계(*Population Health and Clinical Data Linkage*)와 Western Australian 데이터 연계 유닛(*Data Linkage Unit, Kelman et al*., 2002)이 개발한 방법론에 기초한 2006년의 NSW와 ACT 의료 레코드 연계 센터(*Centre for Health Record Linkage - CHeREL*)[15]의 설립과 함께 의료 레코드 연계의 중요성이 인식되고 있다.

비록 이용 가능한 많은 상업용 데이터통합과 연계 시스템이 존재하지만, 이들 대부분은 연계 엔진에 구현된 기술의 세부사항이 이용 불가하다는 점에서 사용자 입장에서는 '블랙박스'에 불과하다. 또한 이러한 시스템들의 많은 수가 비즈니스 데이터의 통합 또는 고객 메일링 목록의 정제·중복 제거와 같은 특정 영역에 세분화되어 있다. 그러나 의료 분야에서 데이터 연계는 더 복잡하고, 다양한 소스로부터의 데이터를 포함한다. 의료 기록뿐만 아니라 의료 시스템 외부에서 수집된 인구 데이터(선거 명부나 경찰 사건 데이터베이스와 같은)까지 포함할 수 있다. 의료분야에서 상업용 연계 시스템이 핵심 연계 엔진으로 사용되고 있지만, 상당수의 추가적인 프로그래밍이 연계 엔진을 특정 분야에 접목시키기 위해 필요하다. 이것은 연계 환경이 상업용 연계 엔진의 목적에 의해 제한되어짐을 의미한다.

레코드 연계는 복잡한 과정이며 사용자들에게 기술적인 세부사항의 상당 부분까지 이해할 것을 요구한다. 예를 들어 사용자는 얼마나 근접하게 이름 또는 주소의 문자열 비교가 수행되는지 알 필요가 있다. 이것이 매칭 가중치의 계산에 영향을 주기 때문이다. 적당한 가격의 이용 가능한 소규모의 상업용 레코드 연계 시스템이 몇몇 존재하지만, 이들은 다양한 형태의 데이터를 다룰 수 없고, 제한된 수의 기능만을 포함하며(가령 통상적으로 이용되는 특정 문자열 비교 방법만을 구현), 적은 수의 데이터 세트만을 연계할 수 있다. 반면 대규모의 레코드 연계 시스템은 보통 매우 고가여서, 오직 대규모 조직에서만 이용가능하다. 거의 모든 상업용 시스템들의 연계 엔진 소스코드는 조사 불가능하므로 사용자들에게는 '블랙박스'에 불과하다. 따라서 기존기술과 새롭게 발전된 레코드 연계 기술의 장점과 한계의 이해를 위하여, 레코드 연계 현역 종사자들이 이들을 실험할 수 있는 tool을 갖는 것은 중요하다. 이러한 tool은 유연해야 하며, 많은 연계 방법들을 포함해야 하고, 다양한 실험 연계를 위한 여러 환경 설정이 가능해야 한다. 또한 의료 분야의 대다수의 레코드 연계 사용자들이 프로그래밍에 대한 경험이 적으므로, GUI 시스템이 레코드 연계 프로젝트를 어떻게 설정하고, 가동하는지 이해할 수 있도록 구조에 대한 양질의 정보와 논리적인 사용 방법을 제공해야 한다.

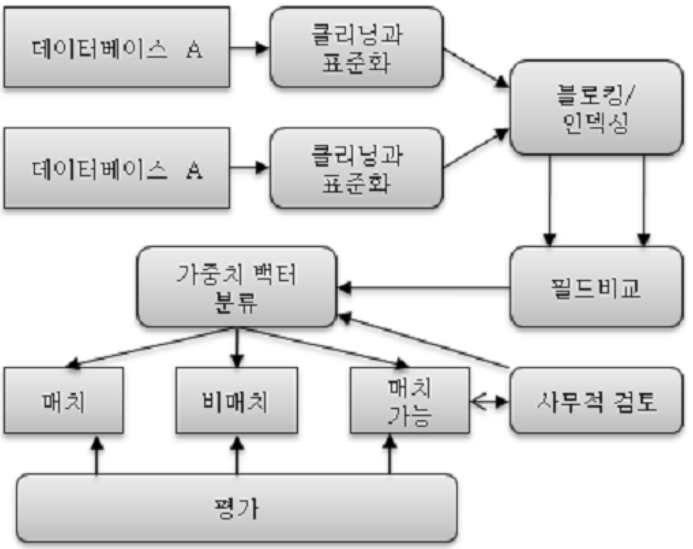
이렇듯 소스코드로의 접근을 허용하고 많은 연계 기술을 포함하는 유연성 있는 레코드 연계 시스템의 결핍 문제는 *Febrl* 연계 시스템에 의해 해결할 수 있다. 다음에는 일반적인 레코드 연계 과정에 대한 간략한 소개와 함께 *Febrl* 유저 인터페이스의 구조와 기

15) http://www.cherel.org.au

능에 대한 상세한 설명을 하고자 한다. 또한 보건의료 분야에서의 *Febrl*의 이용에 대한 논의를 개진한 후 미래 업무의 전망을 다룰 것이다.

[그림 6-1]은 레코드 연계 과정의 개략적 흐름을 보여준다. 대부분의 실제 데이터 집합들은 노이즈나 불완전하고 부정확한 형식의 정보를 포함하기고 있기 때문에 데이터 클리닝과 표준화는 성공적인 레코드 연계를 위한 중요한 사전 작업이 된다. 또한 데이터의 사후 분석, 데이터 마이닝 등을 위한 사전 작업으로서도 데이터 클리닝과 표준화 작업은 중요하다(Rahm & Do, 2000). 데이터의 질이 확보되지 못한다면 성공적인 데이터 연계와 중복 제거에 큰 장애물로 작용할 수 있기 때문이다(Clarke, 2004). 데이터 클리닝과 표준화에 있어 가장 큰 과제는 원시 입력 데이터를 명확하고 일관성 있는 형식으로 변환시키고, 정보의 표현과 인코딩에 있어서의 불일치를 해결하는 것이다(Churches et al., 2002).

두 데이터베이스 'A'와 'B'가 연계된다는 것은 데이터베이스 'A'의 개별 레코드들이 데이터베이스 'B'의 모든 레코드들과 연계되어야 함을 의미한다. 따라서 잠재적인 레코드 쌍 비교의 총 횟수는 두 데이터베이스 크기의 곱 $|A| \times |B|$이 된다($|\cdot|$는 데이터베이스 내의 레코드 수를 의미한다). 이와 유사하게, 데이터베이스 'A'의 중복 제거 시 레코드 쌍 비교의 총 횟수는, 개별 레코드가 모든 다른 레코드와 비교되므로 $|A| \times (|A|-1) / 2$ 가 된다. 레코드 쌍 간의 상세한 필드(혹은 속성) 비교는 노력이



[그림 6-1] General record linkage process

많이 소모되는 작업이므로 데이터 연계나 중복제거에 있어 병목 구간으로 작용하며, 이는 데이터베이스의 규모가 클 때 모든 쌍의 비교를 실행불가하게 만든다(Baxter et al., 2003; Christen & Goiser, 2007).

데이터베이스들 내에 중복 레코드가 없다고 가정하면(즉 데이터베이스 'A'의 하나의 레코드가 데이터베이스 'B'의 하나의 레코드로만 매치되고 그 역도 성립할 때), 올바른 매칭의 최대 횟수는 둘 중 크기가 작은 데이터베이스의 레코드 수와 같다고 볼 수 있다. 따라서 더 큰 규모의 데이터베이스 간의 연계 시에도 비록 매칭의 경우의 수는 이차식의 형태로 증가하지만, 잠재적인 매칭의 수는 오직 선형적으로만 증가하게 된다. 이것은 중복 레코드의 수가 데이터베이스상의 총 레코드의 수보다 항상 작게 되는 중복제거의 경우에도 마찬가지이다.

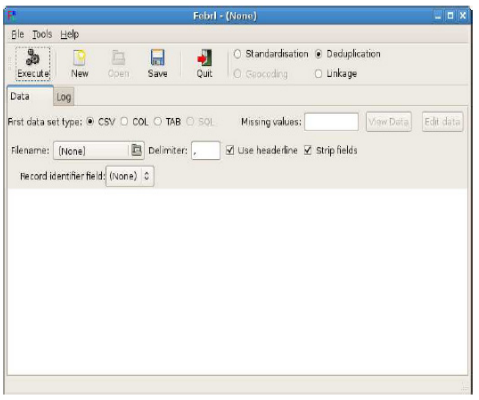
레코드 연계 방법은, 대량의 잠재 레코드 쌍 비교 횟수를 줄이기 위하여 indexing이나 filtering 기법[16] 등을 사용한다. 하나의 레코드 필드(속성)나 몇 개 필드의 조합[17]을 이용하여 데이터베이스를 몇 개의 블록으로 쪼갠다. 블로킹 키 내에서 같은 값을 가지고 있는 레코드들은 모두 하나의 블록으로 삽입되며, 그 결과 후보 레코드 쌍들은 동일 블록 내의 레코드로부터만 생성된다. 이러한 후보 쌍들은 다양한 비교 함수 -한 개의 레코드 필드나 몇 개의 필드의 조합에 적용되는- 를 이용하여 비교가 이루어진다(Baxter et al., 2003). 이러한 함수들은 정확한 문자열이나 수치의 비교처럼 간단할 수도 있지만, 지리적 위치(위도와 경도)의 참조표에 기반한 거리 비교와 같이 복잡할 수도 있으며, 더불어 오타의 식별과 분산의 계산도 가능하다. 필드 비교에서는 매칭 가중치(matching weight)라 불리는 유사한 값들을 '정규형' 형태로 산출한다. 동일한 두 필드 값은 '1'의 매칭 가중치를 가지며, 완전히 다른 두 필드 값은 '0'의 매칭 가중치를 갖는다. 어느 정도 유사한 필드 값들은 '0'에서 '1' 사이의 매칭 가중치를 갖게 된다. 다양한 비교 함수에 의해 계산된 모든 매칭 가중치를 포함하는 개별 레코드 쌍들을 위해 가중치 vector가 형성된다. 이러한 가중치 백터는 이용하는 모델의 결정에 따라 레코드 쌍들을 매치(matches), 비매치(non-matches), 매치가능(possible matches)으로 분류하는데 쓰인다(Christen & Goiser 2007, Fellegi & Sunter 1969, Gu & Baxter 2006). 블로킹 과정에서 제거된 레코드 쌍들은 비교과정의 진행 없이 비매칭으로 분류되며, 그 후 다양한 평가 방법들이 레코드 연계 쌍들의 품질 평가를 위해 사용된다(Christen & Goiser 2007).

매치 가능 등급은 최종 연계 상황의 결정을 위해 사람의 감독이 필요한 -사무적 검토로 통용되는- 레코드 쌍들을 말한다. 이론상으로는, 이러한 사무적 검토의 담당자는 연계 상태를 해결할 수 있게 해주는 추가적인 데이터에 접근할 수 있으나 현실적으로는

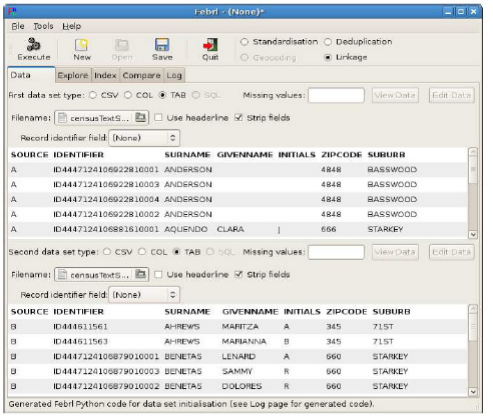
16) 합하여 블로킹 기법이라 부른다.
17) 블로킹 키로 불리기도 한다.

추가적으로 이용 가능한 데이터가 없다. 따라서 사무적 검토는 경험과 상식, 직관에 의해 결정을 내리는 과정이 된다. 사람이 검토하고 분류한 레코드 쌍들은 후속 연계의 분류 품질의 향상을 위한 트레이닝 데이터로 이용될 수 있다.

*Febrl GUI*는 레코드 연계 비전문가들의 *Febrl* 사용을 좀 더 용이하게 하고자 개발되었다. *Febrl GUI*의 구조는 *Rattle* 오픈소스 데이터 마이닝 툴의 구조를 따른다. 기본 아이디어는 레코드 연계 과정의 단계당 하나의 탭(현대의 웹브라우저 내 탭과 비슷한)을 포함한 하나의 창(윈도우)을 가지는 것이다. [그림 6-2]는 *Febrl GUI*의 초기 화면이다. 처음에는, 두 개의 탭만이 보여지지만, 입력 데이터가 발생되면 추가로 탭들이 생겨난다. 개별 탭을 통해 사용자는 방식과 그들의 변수를 선택할 수 있으며, '*Execute*' 버튼의 클릭으로 이러한 설정을 확정지을 수 있다. '*Log*' 탭에서는, 상응하는 *Febrl Python* 코드를 볼 수 있다. 모든 필요한 과정이 설정되면, 생성된 *Python* 코드는 저장되어 *GUI* 환경 밖에서도 구동 가능해진다. 또한 새로운 프로젝트를 시작하고 이 결과를 *Febrl GUI* 내에서 측정할 수도 있다. 주요 탭들의 기능은 다음의 스크린 샷을 통해 자세히 설명할 것이다.



[그림 6-2] Initial Febrl user interface after start-up



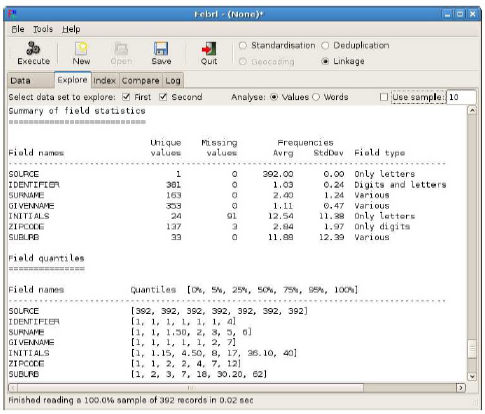
[그림 6-3] Febrl user interface for a linkage project after the 'Census' input data sets have been initialised

사용자는 먼저 실행할 프로젝트의 타입을 선택해야 한다(a, 데이터 세트의 클리닝과 표준화 b, 데이터 세트의 중복제거 c, 두 데이터 세트의 연계). *Febrl GUI*의 '*Data*' 탭은 하나 혹은 두개의 입력 데이터 세트 선택 영역을 표시하며, 이에 따라 화면이 바뀌어진다. 현재, 자주 이용되는 CSV(콤마 구분 형식)를 비롯하여 몇 개의 텍스트 기반 파일 형식이 지원된다. 파일이 선택되어지면 [그림 6-3]에서처럼 파일의 처음 몇 줄이 보인다.

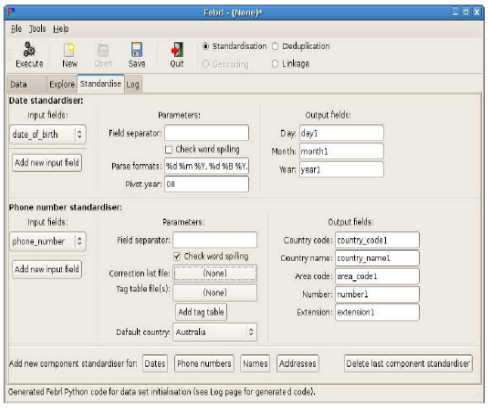
이로 사용자는 선택한 설정을 확정할 수 있고, 필요하다면 변경할 수 있다. 설정에 문제가 없으면, '*Execute*' 버튼을 눌러 설정을 확정하여, 데이터 탐색 탭과, 선택된 프로젝트 타입에 따라 표준화, 인덱싱, 비교, 분류 탭들을 볼 수 있다.

'*Explore*' 탭은 사용자가 입력 데이터의 내용과 질을 더 잘 이해할 수 있게끔 입력 데이터의 분석을 가능케 한다. '*Execute*' 버튼을 클릭하면, 데이터 세트가 읽어지고 모든 필드(또는 속성과 칼럼)가 분석된다. 필드 내 다른 값의 수, 알파벳 순서상 가장 작은 값과 큰 값, 최빈값, 사분위편차, 결측값의 수, 필드 타입의 추정(숫자 혹은 문자로만 이루어졌거나 혼합된 형태이면)에 관한 레포트가 개별 필드당 요약된다. [그림 6-4]는 이 요약을 보여준다.

*Febrl GUI*를 사용하는 데이터 클리닝과 표준화는 현재, 데이터 연계나 데이터 중복제거와는 구별되어 행해진다. 하나의 데이터 세트는 정제되거나 표준화될 수 있고, 새로운 데이터 세트에 쓰여진 후 중복제거되거나 연계될 수 있다. 현재 *Febrl*은 이름, 주소, 날짜, 전화번호의 표준화 도구를 포함하고 있다. 이름 표준화 도구는 간단한 이름(하나의 이름과 하나의 성만으로 이루어진)을 위해서는 규칙 기반의 방식을 쓰고, 보다 복잡한 이름을 위해서는 확률적 은닉 마르코프 모델(HMM) 방법을 결합하여 사용한다. 반면, 주소 표준화 도구는 전적으로 *HMM* 방식을 사용한다. 이러한 *HMM* 방식들은 현재 독립된 *Febrl* 모듈을 이용하여 *Febrl의 GUI* 환경 밖에서 이루어진다. 날짜는, 입력 데이터 세트에서 발견되기 쉬운 예상 날짜 형식을 제공하는 형식 문자열 리스트를 이용하여 표준화된다. 전화번호는 형식 기반 방식을 통해 표준화된다. '*Execute*' 버튼을 클릭하면,



[그림 6-4] Data exploration tab showing summary analysis of record fields(or attributes, columns)

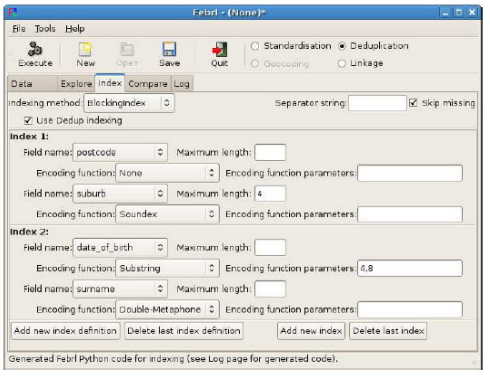


[그림 6-5] Example data and telephone number standardisers(for synthetic Febrl data set)

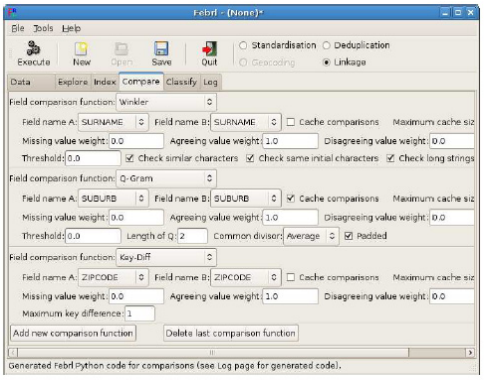
‘*Output/Run*’ 탭에서 표준화된 결과 파일의 파일 이름을 선택할 수 있고, ‘*Output/Run*’ 탭의 ‘*Execute*’ 버튼을 클릭하면 표준화 프로젝트가 시작된다([그림 6-5]).

블로킹 혹은 인덱싱은 상세 레코드 쌍의 비교 횟수를 줄이기 위해 이용된다. ‘*Index*’ 탭 상에서, 사용자는 일곱 가지 중 하나의 인덱싱 방식을 택할 수 있다. Febrl은 ‘*Full Index*’ (모든 레코드 쌍을 비교하므로 이차 복잡성을 지닌) 방식과 많은 레코드 연계 시스템에서 구현된 표준 ‘*Blocking Index*’ 방식을 비롯하여, 최근 개발된 5개의 인덱싱 방식을 담고 있다. 정렬된 이웃 방식에 기반한 ‘*Sorting Index*’, 퍼지 블로킹을 허용하는 길이 q의 부분 문자열을 사용하는 ‘*Q Gram Index*’, *TF-IDF* 또는 *Jaccard* 유사성을 이용한 중첩 캐노피 클러스트링을 적용하고 있는 ‘*Canopy Index*’, 인덱스 키 값을 다차원 공간과 연결시키며 이러한 다차원 객체에 대해 캐노피 클러스터링을 수행하는 ‘*String MapIndex*’, 인덱스 키 값으로의 효율적인 접근과 상응하는 블록의 생성을 가능케 하기 위하여 인덱스 키 값의 모든 접미사들을 생성하고 그들을 정렬된 배열에 삽입하는 ‘*Suffix Array Index*’. 인덱싱 방식이 선택되면, 실제 인덱스 키를 선택하고 그들의 다양한 변수를 설정해야 한다. 인덱스 키는 한 개의 필드 값으로 이루어져 있거나 비슷한 소리의 값들을 동일 블록으로 묶기 위해 음성학적으로 인코딩되는 몇 개 필드 값의 결합으로 이루어질 수 있다. *Febrl*은 *Soundex*, *NYSIIS*, *Phonex*, *Double-Metaphone*을 포함하여 9개의 인코딩 방식을 담고 있다.

레코드 쌍의 필드(속성) 값 비교를 위해 쓰이는 유사도 함수는 [그림 6-7]에서와 같이 ‘*Comparison*’ 탭에서 선택할 수 있다. *Febrl*은 20개의 근접 문자열 비교 기능을 포함,



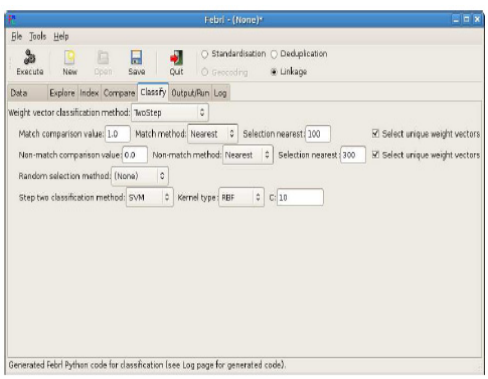
[그림 6-6] Example indexing definition using the 'Blocking Index' method and two index definition



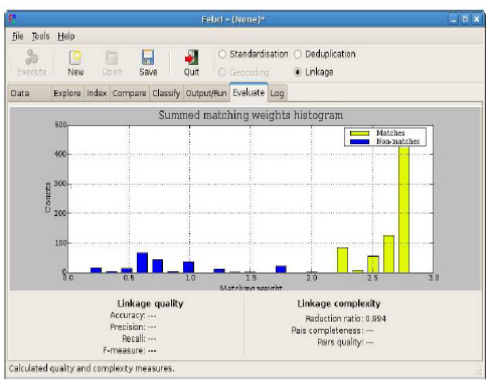
[그림 6-7] An example of three field comparison function definitions

26개의 유사도 함수를 포함하며, 날짜, 시간, 나이, 수치에 특화된 함수 또한 포함한다. 이러한 모든 유사도 함수는 0(완벽한 불일치)에서 1(정확한 일치) 사이의 값을 반환한다. 동의 혹은 비동의의 가중치를 설정함으로써 이러한 값을 조정할 수 있고, 하나 혹은 둘의 비교 값이 비어있을 때 특수값을 반환하도록 조정할 수도 있다. 개별 비교된 레코드 쌍을 위해 계산된 유사도 가중치는 가중치 벡터에 저장되며 다음 단계에서 레코드 쌍을 분류하는데 이용된다.

*Febrl*은 몇 개의 레코드 쌍 분류자와 감독 기법, 무감독 기법을 포함한다. 기존의 '*FellegiSunter*' 분류자는 두 임계치의 수동 설정을 필요로 하였으나, 감독 기법의 '*Optimal Threshold*' 분류자는 모든 비교 레코드 쌍의 트루 매치 상태를 알고 있다고 가정하므로, 상응하는 합산 가중치 벡터를 기반으로 최적 임계치가 계산된다. 감독 분류자의 또 다른 예가 '*SuppVecMachine*'이며, 이는 지지벡터기계(*support vector machine*)를 구현한다. '*KMeans*'와 '*FarthesFirst*' 분류자는 무감독 클러스터 방식이며 가중치 벡터를 매치와 비매치 클러스터로 묶는다. 다양한 중심 초기화와 거리 측정법이 *Febrl*에 구현되어 있다. 마지막으로 '*Two Step*' 분류자는 [그림 6-8]에서 보듯 무감독 방식이며, 첫 번째 단계에서는, 높은 확률로 트루 매치와 트루 비매치로 상응되는 비교 레코드 쌍으로부터 가중치 벡터를 선택한다. 두 번째 단계에서는 이진 분류자의 훈련용 예시로 이러한 벡터들을 이용한다([그림 6-8]).



[그림 6-8] Example 'Two-Step' unsupervised weight vector classifier



[그림 6-9] Evaluation tab showing the matching weight histogram an quality and complexity measures for a linkage

'*Evaluation*' 탭에서는 [그림 6-9]와 같이 중복제거나 연계 프로젝트의 결과가 모든 비교 레코드 쌍의 합산된 매칭 가중치의 히스토그램으로 표현된다. 만약 레코드의 트루 매치와 비매치 상태를 볼 수 있다면, 수행된 연계의 품질은 정확도, 정밀성, 기억 및 F-척

도의 측정을 이용해 보여질 것이다. 중복제거나 연계 프로젝트의 복잡도(인덱싱 단계와 이것의 품질에 의해 생성된 레코드 쌍의 수)를 나타내는 척도 또한 나타나며, 감소 비율과 쌍의 완결성, 쌍의 품질 등이 이러한 척도에 해당된다([그림 6-9]).

2. 국내사례

가. 기존사례

1) 외부 자료 간 연계

2차(연계) 자료는 기존에 이미 구축된 자료인 'a' 자료 'b' 자료 간의 연계를 통해 생성된다. 이 때 자료 'a'와 자료 'b'에 모두 존재하는 변수가 key 변수가 되는데, 외부 연구기관의 패널조사는 통일 표본에 대한 종단적 조사를 원칙으로 하기 때문에 자료 간 연계를 위한 key 변수를 자료 안에 담고 있다. 즉 data set 구성 자체가 자료 간 연계가 가능하도록 설계되어 있다. 여기서 사용되는 key 변수는 조사 원년도에 부여된 일련번호(가구 및 개인번호)이며, 조사 기간 중 일련번호는 동일하게 유지하는 것을 원칙으로 한다. 여기서는 한국노동연구원이 지난 1998년부터 실시하여 지금까지 지속되고 있는 한국노동패널(KLIPS) 자료를 예시로 자료 간 연계에 관해 간략하게 살펴보고자 한다.

노동패널 데이터는 크게 가구용과 개인용 데이터로 분류되며, 개인용 자료의 일자리 정보를 토대로 가공된 자료인 직업력 자료가 있다. 가구용 데이터에는 가구 식별번호인 'HHID'가 있으며, 개인용 데이터에는 개인 식별번호인 'PID'가 있다. 개인번호(PID)는 '가구 번호×100 + 가구원 번호'로 생성된 것으로 가구자료와 개인자료 연계 시 임의로 개인번호가 생성 가능하도록 설계되어 있다. 노동패널 데이터의 자료 간 연계유형은 가구와 가구 자료 간 (시계열)연계, 개인과 개인 자료 간 (시계열)연계, 가구와 개인 자료 간 (횡단면)연계, 개인과 부가자료 간 (횡단면)연계로 유형화될 수 있다. 다음은 노동패널의 자료 간 연계 과정을 사례별로 검토한 것이다.

가) 가구자료에서 가구원 정보 추출하기

노동패널 가구용 원자료에 수록된 가구원의 성, 연령, 가구주와의 관계, 학력 등의 정보는 자체 가공되어 개인자료에 '가구정보'라는 라벨이 붙은 변수로 제공된다.[18] 따라서 개인자료를 분석하는 일반적인 상황이라면, 가구자료에서 가구원의 인적특성 정보들을 별도로 추출할 필요가 없다. 그러나 상황에 따라 개인자료에 포함되지 않은 개인의 다른 정보가 필요한 경우가 발생할 수 있다. 대부분의 통계 패키지들은 행 단위보다는 열 단

18) 본 연구에서 가계자료의 가구원 정보를 경제활동인구조사 자료와 결합할 때도 이와 같은 자료 처리과정을 거치게 된다.

위의 연산에 적합하도록 만들어져 있다. 따라서 이러한 문제를 해결하기 위해서는 자료의 간단한 조작이 필요하다.

가구자료의 성, 연령, 가구주와의 관계는 다음의 그림과 같이 1번 가구에 소속된 가구원일 경우 1개 case 내에 붙어 있다. 예컨대 1번 가구에 소속된 가구원이 15명이라면 1번째부터 15번째 가구원의 성별변수들 사이에 각각의 값이 들어가 있다. 이렇게 가로형태로 붙어있는 성별 변수를 개인 case별로 잘라내어 다시 붙이게 된다면 case는 늘어나는 반면 성별변수는 15개에서 1개로 줄일 수 있다.

<가구자료의 가구원 정보>

가구원의 인적정보(최대 15명)					
h110241	h110242	…	h110361	h110362	..
1번째 가구원의 성별	2번째 가구원의 성별	…	1번째 가구원의 만나이	2번째 가구원의 만나이	..

	성별	연령	가구주와의 관계	…
1번째 가구원				
2번째 가구원				
…				
1번째 가구원				
2번째 가구원				
…				

[그림 6-10] 가구자료의 가구원 정보의 조작

```
*=================================;
*     SAS - Transpose 문의 사용     ;
*=================================;

data a1; set h.klips11h ;
keep      hhid11
          h110221-h110235 /*pid*/  h110241-h110255 /*sex*/  h110261-h110275 /*가구주와의 관계*/
          h110361-h110375;/*만나이*/
if hwave11=1;

proc sort; by hhid11;
run;

data a1; set a1;
proc transpose data=a1 out=trans1(drop=_name_  rename=(col1=pid)); by hhid11;
          var h110221-h110235; /* 1~15 번째 가구원의 pid */
proc transpose data=a1 out=trans2(drop=_name_  rename=(col1=sex)); by hhid11;
          var h110241-h110255; /* 1~15 번째 가구원의 성별 */
proc transpose data=a1 out=trans3(drop=_name_  rename=(col1=rel)); by hhid11;
          var h110261-h110275; /* 1~15 번째 가구원의 가구주와의 관계 */
proc transpose data=a1 out=trans4(drop=_name_  rename=(col1=age)); by hhid11;
          var h110361-h110375; /* 1~15 번째 가구원의 만 나이 */

data fn1; merge trans1 trans2 trans3 trans4; by hhid11;
if pid^=.; /* 결측치 처리 */
drop _label_;

proc means; var sex rel age;
proc freq; tables sex rel age;
run;
```

[그림 6-11] SAS program 예제

제6장

[그림 6-11]은 가로로 15개씩 나열된 데이터를 하나의 변수로 세로 형태로 전치(transpose) 시켜 묶는 과정이다. 프로그램의 로직은 첫째, 노동패널 11차년도 가구자료로부터 1부터 15번째 가구원들의 성별, 가구주와의 관계, 연령 변수를 추출하여 4개의 취합된 data set으로 묶는다. 둘째, 이렇게 생성된 4개의 데이터를 case merge하여 hhid11를 기준으로 붙인다.

나) 가구와 개인자료 간 연계

가구자료와 개인자료는 기본적으로 데이터 구성의 기본단위가 다르다. 그러나 자료 사용 시 가구자료와 개인자료를 연결하여 사용해야 할 경우가 많다. 예컨대 가구 총소비를 종속변수로 하는 회귀모형을 구성한다고 가정하였을 때, 필요한 설명변수로는 가구 총소득, 자산, 부채, 총가구원 수, 가구주의 성별, 연령, 가구주의 경제활동상태 등이 된다. 이 같은 경우 다른 변수들은 가구자료 내에 존재하지만, 가구주의 경제활동상태는

<가구 data set>

VIEWTABLE: TMP1.klips01h

	hhidwon	hhid01	hwave01	hwaveent	sample98	h010141	h010142
1	1	1	1	1	1	1	23
2	2	2	1	1	1	1	23
3	3	3	1	1	1	1	23
4	4	4	1	1	1	1	23
5	5	5	1	1	1	1	23
6	6	6	1	1	1	1	23
7	7	7	1	1	1	1	23
8	8	8	1	1	1	1	23
9	9	9	1	1	1	1	23
10	10	10	1	1	1	1	23
11	11	11	1	1	1	1	23
12	12	12	1	1	1	1	23
13	13	13	1	1	1	1	23
14	14	14	1	1	1	1	23
15	15	15	1	1	1	1	23
16	16	16	1	1	1	1	25
17	17	17	1	1	1	1	16
18	18	18	1	1	1	1	23
19	19	19	1	1	1	1	24
20	20	20	1	1	1	1	24
21	21	21	1	1	1	1	24
22	22	22	1	1	1	1	24
23	23	23	1	1	1	1	24
24	24	24	1	1	1	1	24
25	25	25	1	1	1	1	24
26	26	26	1	1	1	1	24
27	27	27	1	1	1	1	24
28	28	28	1	1	1	1	24
29	29	29	1	1	1	1	24
30	30	30	1	1	1	1	24

<개인 data set>

VIEWTABLE: TMP2.klips01p

	PID	HHID01	HMEM01	JOBTYPE	HHIDWON	P010101	P010102
1	101	1	1	1	1	2	10
2	102	1	2	1	1	1	11
3	201	2	1	1	2	1	10
4	202	2	2	.	2	2	20
5	203	2	3	.	2	2	11
6	301	3	1	1	3	2	10
7	401	4	1	1	4	1	10
8	402	4	2	.	4	2	20
9	501	5	1	1	5	2	10
10	601	6	1	.	6	1	10
11	602	6	2	.	6	2	20
12	603	6	3	.	6	2	14
13	604	6	4	1	6	1	15
14	701	7	1	.	7	2	10
15	801	8	1	1	8	1	10
16	901	9	1	.	9	2	10
17	902	9	2	1	9	2	11
18	903	9	3	1	9	2	12
19	1001	10	1	1	10	2	10
20	1101	11	1	.	11	1	10
21	1102	11	2	1	11	2	20
22	1201	12	1	1	12	2	10
23	1301	13	1	.	13	1	10
24	1302	13	2	.	13	2	20
25	1303	13	3	.	13	2	11
26	1401	14	1	2	14	1	10
27	1402	14	2	2	14	2	20
28	1403	14	3	1	14	1	16
29	1404	14	4	1	14	1	17
30	1405	14	5	.	14	2	6

[그림 6-12] 가구 데이터와 개인 데이터의 구성

```
data p11; set a. klips11p;
keep hhid a116101;
proc sort; by hhid;
run;

datr h11; set b. klips11h;
proc sort; by hhid;
run;

data hp11; merge p11 h11; by pid;
run;
```

[그림 6-13] SAS program 예제

개인자료에서 별도로 구성한 다음 가구자료에 붙여야 한다. 역으로 개인자료에 가구정보(가구소득 · 소비 등)를 연계해야 하는 경우가 발생할 수도 있다.

이같이 가구자료와 개인자료를 결합하여야 하는 경우 key변수로는 'hhid(가구번호)'를 사용하게 된다. 자료 연계과정은 가구 및 개인자료를 각각 'hhid'로 정렬한 후 가구번호를 이용하여 variable merge하게 된다. 이때 가구와 개인자료는 동일 년도의 자료를 사용하게 됨으로 10차년도 가구자료와 개인자료의 연계는 'hhid10'를 사용하고, 9차년도 가구자료와 개인자료의 연계는 'hhid09' 변수를 사용한다.

다) 직업력 자료와 개인자료의 연계

직업력 자료(work history data)란 개인의 일자리 정보를 수록한 자료로 개인의 성, 연령, 교육수준 등과 같은 인구학적 특성은 포함되어 있지 않다. 직업력 자료는 기본적으로

<직업력 data set>

VIEWTABLE: Written by SAS

	PID	JOBWAVE	JOBSEQ	JOBNUM	JOBNUMC	JOBCENS	JOBTYPE
1	101	1	1	101	.	3	1
2	101	2	1	201	101	1	1
3	101	3	1	301	201	1	1
4	101	5	1	501	301	2	1
5	101	5	2	502	.	3	1
6	101	6	2	601	502	2	1
7	101	8	3	801	.	3	2
8	101	9	3	901	801	1	2
9	101	10	3	1001	901	2	2
10	101	10	4	1002	.	3	1
11	101	11	4	1101	1002	1	1
12	102	1	1	101	.	3	1
13	102	2	1	201	101	1	1
14	102	3	1	301	201	1	1
15	102	5	1	501	301	1	1
16	102	6	1	601	501	1	1
17	102	7	1	701	601	2	1
18	102	10	2	1001	.	3	1
19	102	11	2	1101	1001	1	1
20	201	1	1	100	.	0	1
21	201	1	2	101	.	3	1
22	201	2	2	201	101	1	1
23	201	3	2	301	201	1	1
24	201	4	2	401	301	2	1
25	201	4	3	402	.	3	1
26	201	6	2	601	401	2	1
27	201	6	4	602	.	3	1
28	201	7	4	701	602	1	1
29	201	8	4	801	701	1	1
30	201	9	4	901	801	1	1

<개인 data set>

VIEWTABLE: Written by SAS

	PID	HHID11	HMEM11	HHID10	HHID09	HHID08	HHID07
1	101	1	1	1	1	1	1
2	102	7258	1	7258	1	.	1
3	201	6034	1	6034	6034	6034	6034
4	202	2	2	2	2	2	2
5	203	2	3	2	2	2	2
6	401	4	1	4	4	4	4
7	402	4	2	4	4	4	4
8	801	8	1	8	.	8	8
9	901	9	1	9	9	9	9
10	902	9	4	9	9	9	9
11	903	6056	1	6056	6056	6056	6056
12	1001	10	1	10	10	.	10
13	1101	11	1	11	11	11	11
14	1102	11	2	11	11	11	11
15	1103	11	3	11	11	11	11
16	1201	12	1	12	12	12	12
17	1203	12	3	12	12	12	12
18	1301	13	1	13	13	13	13
19	1401	14	1	14	14	14	14
20	1402	14	2	14	14	14	14
21	1403	5048	1	5048	5048	5048	5048
22	1502	15	2	15	15	15	15
23	1503	15	3	15	15	15	15
24	1601	16	1	16	16	16	16
25	1801	18	1	18	18	18	18
26	1802	18	2	18	18	18	18
27	1803	18	3	18	18	18	18
28	1804	18	4	18	18	18	18
29	1805	18	5	18	18	18	18
30	1902	19	2	.	.	.	.

[그림 6-14] 직업력 데이터와 개인 데이터의 구성

```
*=============================;
*     SAS - Merge문의 사용   ;
*=============================;

data a1; set p.klips11p(keep=pid p110101 p110107);
          rename p110101=sex p110107=age;
          proc sort; by pid;

data a2; set w.klips11w; if jobwave=11;

proc sort; by pid;

data work; merge a1 a2; by pid;  if jobwave=11;
run;
```

[그림 6-15] SAS program 예제

제6장

일자리 단위로 만들어진 데이터 셋이다. 즉 한 개인이 1차 조사부터 11차 조사 시점까지 총 20개의 일자리를 가졌다면 20개의 '행(column)'이 형성된다. 만약 1차 조사 시점 이전에 이미 5개의 일자리 경력(회고적 일자리)이 존재한다면 총 25개의 일자리 '행(column)'이 생성된다. 노동패널 직업력 자료는 차수를 거듭할수록 행의 길이가 무한정으로 증가하기 때문에 수록 변수는 최소화하였다. 따라서 개인의 인적정보는 연구자가 임의로 개인자료에서 연계하여 쓰도록 설계되어 있다. 직업력 자료에 개인의 인적정보를 붙이는 작업은 직업력 자료와 개인자료를 각각 '개인번호(pid)'로 정렬 후 개인번호를 key변수로 연계하면 된다.

2) 통계청 자료 간 연계

통계청에서는 마이크로 데이터를 외부 연구자들에게 제공하지만, 자료 간 연계된 데이터는 제공하지 않으며, 개인 식별 및 자료 간 연계가 가능한 정보(가구번호, 가구원번호)는 제공하지 않는 것을 원칙으로 한다. 그러나 일부 연구자들은 통계청 데이터를 연계하여 분석과제를 수행하였다. 가장 대표적인 예가 『경제활동인구조사』(이하 경활 자료)의 패널화이다. 『경제활동인구조사』는 한국의 대표적인 노동력 조사(*labor force survey*)로 외부 연구자들이 경활 자료의 패널화에 관심을 갖은 이유는 이를 통해 노동시장에서 동태적 움직임을 포착할 수 있기 때문이다. 엄밀하게 말해 직장이동은 적절한 설문을 통해 조사가 가능하다. 미국의 CPS(*Current Population Survey*)는 1994년부터 "지난 달 당신은 'A'라는 일자리에서 일한다고 응답했습니다. 여전히 당신은 'A'라는 일자리에서 일하고 있습니까?"라는 항목을 조사하고 있다. 이러한 질문을 통해 지난 달과 이번 달의 직장의 동일성 여부를 확인할 수 있으며, 이를 통해 직장 간 이동을 정확하게 측정할 수 있다. 한편 Blanchard and Diamond(1990)에서는 CPS의 3월 부가조사에서 이 설문을 이용하고 있으며, 1976년부터 지난 1년 동안 몇 명의 고용주와 일했는지에 대한

〈표 6-2〉 시작 시점 정보의 조합에 따른 직장 간 이동 판정

<table>
<tr><td rowspan="3">t-1월
조사의
취업 연월</td><td colspan="3">t월 조사의 취업 연월</td></tr>
<tr><td colspan="2">t-2월 이전</td><td>t-1월</td><td>t월</td></tr>
<tr><td>t-2월 이전</td><td>두 값이 동일할 경우
'직장유지'로 판전
두 값이 상이할 경우
'주업변경'으로 판정</td><td></td><td></td></tr>
<tr><td>t-1월</td><td>주업 변경</td><td>불확실한 경우</td><td>직장 간 이동</td></tr>
</table>

회고적(*retrospective*) 조사를 실시하였다. 이 자료를 통해 Blanchard & Diamond(1990)에서는 직장 간 이동의 수준을 측정하였다. 그러나 우리나라의 노동력 조사는 이 같은 내용을 포함하고 있지 않다. 따라서 직장이동을 측정하기 위해서 연구자는 간접적인 방법을 취하는데, 가령 동일한 일자리 여부에 대한 판단을 위해서 현 직장의 시작 연월 정보를 이용하게 된다. 만약 지난 달의 일자리 시작 연월과 이번 달 일자리의 시작 연월이 같다면 동일한 일자리로 판단하는 것이다.

한편 월간 자료를 패널화하기 위해서는 월별『경제활동인구조사』간 연계작업이 필요하다. 통계청에서는『경제활동인구조사』마이크로 자료를 제공하지만, 개인 식별이 가능한 '개인 ID'는 외부에 제공하고 있지 않다. 따라서 외부 연구자들은 가용한 여러 변수(생년월일, 성, 가구원 지위, 교육수준)를 활용하여 '개인 ID'를 구성하여 자료를 연계하는 방식을 취하였다.

『경제활동인구조사』를 패널화하여 분석한 사례는 남재량의 연구(1997년)가 효시라고 할 수 있다.『경제활동인구조사』는 동일한 표본을 기본적으로 5년간 유지하므로 동일한 개인을 최장 60개월까지 지속적으로 관찰할 수 있다. 즉 장기 월별 패널자료의 구축이 가능하다고 본 것이다. 남재량(1997, 2005년)은 연접한 두 달을 연계하는 단기 월별 패널 형태를 탈피하여 장기 월별 패널자료를 구축함으로써 경활 자료를 통한 실업기간에 관한 정보량을 늘리고자 하였다. 이렇게 구축한 장기 월별 패널자료를 사용하여 각 개인들의 실업 시작시점과 실업 종료시점을 알 수 있고, 이로부터 실업기간을 추출할 수 있기 때문이다.

남재량(1997, 2005년)이『경제활동인구조사』를 패널화하여 분석한 것은 고용불안 측정에 유량분석의 필요성 때문이다. 통상적으로 고용불안은 '실직 가능성'과 '재취업 가능성'으로 측정하는데, 이는 노동시장 변수들(실업률 등)만으로는 불충분하다. 노동시장 변수들은 노동시장의 정태적(*static*) 측면을 측정하는 저량 변수(*stock variables*)에 불과하기 때문이다. 반면 고용불안의 정의를 구성하는 '실직 가능성'과 '재취업 가능성'은 모두 본질적으로 노동시장의 동태적(*dynamic*) 측면을 반영하는 개념이다. 즉 '실직 가능성'은 취업에서 실업이나 비경활로 이동하는 노동력 상태의 동태적 변화를 의미하며, '재취업 가능성' 역시 실업이나 비경활 상태에서 취업으로 이동하는 동태적 변화를 의미하기 때문이다. 따라서 남재량은『경제활동인구조사』의 패널화를 통해 여러 유량변수(노동력 상태별 유·출입률, 실업 지속기간 등)들을 측정하여 고용불안의 정도를 평가하고 분석하고자 하였다.

남재량(1997, 2005년)이『경제활동인구조사』를 장기 월별 패널화하여 분석한 이래 이병희(2005, 2008년) 역시 이러한 방법을 따라 노동력의 동태적 분석을 시도하였다. 이병희(2005년)는 경활 자료를 이웃하는 월별 개인별로 대응시켜 월별 패널자료

〈표 6-3〉 통계청 자료 간 연계분석 사례 : 『경제활동인구조사』 자료의 패널화

저자 및 작성기관	활용 연계자료	연구내용
남재량(1997년) (서울대학교)	『경제활동인구조사』 (1981~1994)의 패널화	- 연구내용 : 장기 월별 패널자료를 구축하여 우리나라 실업률의 장기 추세변화 분석 - 『우리나라 실업률 추세변화에 관한 연구(1997)』
남재량(2005년) (한국노동연구원)	『경제활동인구조사』 (1993~2003)의 패널화	- 연구내용 : 장기 월별 패널자료를 구축하여 실업지속 기간의 측정 및 분석 - 『고용불안계층의 실태 및 고용정책과제(2005)』
이병희·정재호(2005년) (한국노동연구원)	『경제활동인구조사』 (1996~2003)의 패널화	- 연구내용 : 장기 월별 패널자료를 구축하여 노동이동과 경력변동 실태 분석 - 『노동이동과 인력개발 연구(2005년)』
이병희 외(2008년) (한국노동연구원)	『경제활동인구조사』 (1985~2006년)의 패널화	- 연구내용 : 장기 월별 패널자료를 구축하여 노동시장 이행확률 분석 - 『노동시장의 구조변화와 고용변동(2008)』
김혜원 외(2008년) (한국노동연구원)	『경제활동인구조사』 (2006년)의 월별 패널화	- 연구내용 : 월별 패널자료를 구축하여 직장이동의 규모와 결정요인분석 - 『직장이동의 노동시장 효과 분석(2008)』
김혜원 외(2008년) (한국노동연구원)	『경활』+『경활 근로형태 부가조사』 (2004~2005년)	- 연구내용 : 『경활』과 『경활 부가조사』를 연계하여 장기 패널자료 구축, 직장이동의 선택과 임금성과 및 직장이동 유형에 따른 단기 임금 변화 분석 - 『직장이동의 노동시장 효과 분석(2008)』
이병희(2008년) (한국노동연구원)	『경활』+『경활 근로형태 부가조사』 (2003~2005년)	- 연구내용 : 『경활』과 『경활 부가조사』를 연계하여 장기 패널자료를 구축, 최저임금 인상이 직장 유지에 미치는 영향력 분석 - '최저임금의 고용유지 및 취업 유입 효과' 『산업노동연구(2008)』

(*month-to-month matched data*)를 구성하여 취업, 실업, 비경제활동 간의 노동력 상태 이동을 분석하였다. 이병희가『경제활동인구조사』를 월별 패널화하는데 사용한 방법은 원자료에 수록된 성, 생년월일을 통해 임의로 개인 식별번호를 구성한 후, 월별로 대응시켜 연계 데이터를 구성한 방식이다. 또한 이웃하는 월별로 취업 년월이 다르고, 취업 년월이 다음 조사월과 일치하는 경우는 이직한 근로자로 식별하였다. 남재량(1997, 2005년)이『경제활동인구조사』의 장기 패널화를 통해 유량변수를 통한 고용불안을 분석하고 했다면, 이병희(2005, 2008년)는 유량변수의 분석 및 노동이동과 노동시장 이행확률을 분석하였다.[19] 특히 이병희(2008년)는『경제활동인구조사』의 장기 패널자료 분석을 통해 시기별 노동시장 이행확률을 분석하였는데, 이를 통해 외환위기 이후 노동시장의 구조변화를 포착하였다.

김혜원(2008년) 또한『경제활동인구조사』의 패널화를 통한 연구를 수행하였다. 김혜원은 경활 원자료의 생년월일, 성, 가구원 지위, 교육수준 등의 변수를 이용하여 개인 ID를 구성한 뒤 월별로 연결시켜 중복 제거를 통해 월별 패널자료를 구성하였다. 김혜원은 직장 간 이동의 규모와 결정요인의 분석을 통해 우리나라 직장 간 이동 시계열은 이론적 연구와 일치하는 경기 순행성을 띠며, 직장 간 이동자의 인적 속성 및 일자리 속성의 검토 결과, 직장 간 이동은 종사상 지위별로 큰 차이가 있음을 밝혔다.

이밖에『경활 자료』와 『경활 부가자료』간 연계를 통해 분석한 사례도 있다. 이병희(2008년)는 '이중차이법'을 통해 최저임금 인상이 직장 유지에 미치는 효과를 추적하기 위하여『경활 자료』와『경활 부가조사 자료』를 연계하여 분석에 사용하였다. 이병희가『경활 부가조사 자료』를 사용한 것은 부가조사의 '지난 3개월간의 월평균 임금'과 '평소 주당 근로시간' 정보를 통해 시간당 임금의 산출이 가능하기 때문이다. 이병희는 최저임금 효과를 분석하기 위하여 최저임금제가 시행되기 이전인 2003년 8월『경활 부가조사 자료』와 2004년 2월『경활 자료』를 결합하고, 최저임금제 시행 이후인 2004년 8월『경활 부가조사 자료』와 2005년 2월『경활 자료』를 결합하여 분석에 사용하였다. 한편 김혜원(2008년) 또한『경활 자료』와『경활 부가조사 자료』간 연계자료를 구축하여 분석에 사용하였다. 김혜원은 개인 식별변수 구성의 기본변수가 되는 생년월일이 포함된 2004년, 2005년을 분석 대상 기간으로 삼아 2004년 8월 ~ 2005년 8월『경활 자료』를 연결한 13개월 균등 패널을 구성한 후 이 자료를 각 년도의 8월 부가조사와 결합하여 노동력 상태 이행을 분석하였다.

이상 살펴본 사례(<표 6-3>)들은 경활 조사의 패널화 및 경활과 경활 부가조사를 연계하여 분석에 사용한 사례들이다. 자료 간 연계를 위해 연구자들이 개인 식별번호

19) 그러나 이 같이 연구자 임의로 개인 식별번호를 사용하여 패널자료를 구축하였을 경우, 연구자별로 개인 식별 번호 구성에 사용한 변수가 다르기 때문에 동일한 자료를 사용했다 하더라도 분석결과 간에 차이가 문제점으로 발생한다.

를 구성한 방식은 대체로 유사하나 동일하지는 않다. 따라서 동일한 자료를 연계하였다 하더라도 분석 결과는 저마다 상이하게 나타난다. 또한 표본이탈, 가중치 등 여러 가지 한계가 있음에도 불구하고, 이 같은 노력이 지속되었던 것은 노동력의 동태적 특성을 보여줄 수 있는 대표성 있는 항목이나 자료가 부재했던데 기인한다.

<표 6-4>는『경제활동인구조사』와『가계조사』의 연계분석 사례이다. 주지하다시피『경제활동인구조사』는 우리나라 고용상황을 가늠할 수 있는 대표적인 조사이며,『가계조사』는 소득을 대표하는 조사이다. 연구자들이 두 자료 간 연계분석에 관심을 갖는 것은 연계분석을 통해 고용과 소득 두 분야의 정보를 얻을 수 있으며, 분석가능 분야가 크게 확장되기 때문이다. 특히 고용과 복지의 연계는 정책적으로 매우 중요하다. 가령 근로빈곤(*working poor*) 문제라든가 한부모 가정의 빈곤 문제에 대한 올바른 정책 수립을 위해서는 현황 파악이 선행되어야 하는데, 고용과 소득을 동시에 파악할 수 있는 대표성 있는 통계자료는 아직 부재하다.[20] 가계-경활 자료 연계 분석의 대표적인 사례는

〈표 6-4〉 통계청 자료 간 연계 분석 사례 : 『경제활동인구조사』와 『가계조사』의 연계

저자 및 기관	활용 연계자료	연구내용
이병희 외(2008년) (한국노동연구원)	『경활』+『가계조사』 (2006년)	- 연구내용 :『경활』과『가계조사』의 연계자료를 통해 저소득 노동시장의 실태 및 동태 분석 -『저소득 노동시장 분석(2008)』
김혜원·윤자영(2009년) (한국노동연구원)	『경활』+『가계조사』 (2007년)	- 연구내용 :『경활』과『가계조사』의 연계자료를 통해 모자가구 여성가장의 빈곤과 고용 분석 -『여성가장 가구의 고용과 빈곤 연구(2009)』
김혜련(2009년) (통계개발원)	『경활』+『가계조사』 (2005~2009년)	- 연구내용 :『경활』과『가계조사』의 연계자료를 통해 근로빈곤 현황 결정요인, 근로빈곤의 동태적 현황 분석 -『근로빈곤의 동태적 분석(2009)』
정진호 외(2001년) (한국노동연구원)	『가계조사』의 패널화 (1998~2000년)	- 연구내용 : 『가계조사』의 패널화를 통해 근로소득의 이동 및 빈곤에 대한 동태적 분석 -『소득불평등 및 빈곤의 실태와 정책과제(2001)』

20) 해외사례에서도 높은 응답 부담 때문에 고용과 소득을 동시에 조사하여 제공하는 자료는 부재하다.

이병희(2008년)의 연구이다. 이병희는 저소득 취업자의 규모를 분석하기 위하여『한국노동패널자료』와 가계-경활 연계자료를 활용하였다.[21] 이병희는 다음과 같은 방법으로 두 조사를 결합하여 분석하였다. 첫째,『가계조사』는 가구주와 배우자, 취업한 기타 가구원이 1인인 경우에 한정하여 개인 소득 정보를 제공하고 있다. 이 정보를 경활 자료와 개인별로 결합하면 종사상 지위에 따른 개인별 소득을 알 수 있다. 그러나 두 조사 시점의 차이에 따라 가구주와의 관계가 다를 수 있기 때문에 가구주와의 관계뿐만 아니라 성, 연령이 동일한 개인에 한정하여 개인 소득 정보를 결합하였다. 둘째,『가계조사』에서 소득이 대체된 가구는 제외하였다. 셋째,『가계조사』에는 취업형태 정보가 가구주에만 담겨 있으며, 가구주의 경우에도 두 조사 간 시점의 차이로 인하여 근로형태가 변경될 수 있는 문제가 발생한다. 이러한 문제를 해결하기 위해 경활조사에서 임금근로자라고 응답하고『가계조사』에서 해당 개인의 근로소득이 있는 경우에만 결합자료에 근로소득을 부여하였으며, 비임금근로자라고 응답하고 해당 개인의 사업소득이 있는 경우에만 결합자료에 소득을 부여하였다. 이러한 방식으로 최종적으로 구성된 결합자료는 개인의 인적 특성 및 일자리 특성, 가구의 특성 및 소득·지출 등의 정보를 담게 된다. 그러나 경활조사는 매월 3만 3천 가구를 조사하는데 반해『가계조사』는 매월 9천 가구를 조사하기 때문에 결합자료의 월평균 개인 record는 경활의 20%에 불과하다는 문제가 발생한다. 따라서 가계-경활 연계자료가 우리나라 전체의 경제활동상태 및 소득 현황을 대표하지는 못한다는 점을 연구자는 명시하고 있다(이병희, 2008b; 18).

한편『가계조사』를 패널화하여 분석한 사례도 있다. 정진호 외(2001)는 1998~2000년 3개 년도의『도시가계조사』[22] 월별자료를 가구 고유번호를 이용하여 분기자료로 전환하고, 분기자료를 패널자료로 전환하여 분석자료로 활용하였다. 패널자료의 구축과정에서 원자료에는 포함되어 있지만, 매월 조사되지 않는 가구는 분석에서 제외하였다. 이 같은 방식으로 분기자료를 생성하고, 패널자료를 구축하는 과정에서 상당수의 표본이탈이 발생한다. 이 연구에 사용된 전체 연결 패널자료의 현황은 분기별 평균 표본 가구수는 분기자료가 3,839가구, 분기연결패널자료 3,437가구, 전체 연결패널자료가 1,475로 상당수의 표본이탈이 발생함을 알 수 있다.

21) 노동패널자료는 표본의 대표성이 확보되지 못한 한계가 있다.

22)『도시가계조사』는 도시지역에 거주하는 가구 중에서 가구원수가 2인 이상인 5,000여 가구를 표본 추출하여 소득 및 소비실태를 조사하였다. 따라서『도시가계조사』는 농촌지역의 모든 가구와 도시지역의 1인 가구가 조사 대상에서 제외된다는 한계를 갖으며,『도시가계조사』에서 소득은 가구주가 근로자인 가구만 조사되기 때문에 가구주가 자영업자이거나 무직자인 경우에는 소득이 조사되지 않는다는 한계를 갖는다.

나. 전문가 수요조사

1) 조사개요

연구에 들어가기 앞서 통계청 micro-data 이용자 및 통계청 용역과제 연구자를 중심으로 2차 자료 수요관련 외부 전문가 조사를 실시하였다. 조사기간은 2010.6.14 ~ 2010.7.31이며, 조사방법은 인터넷 survey로 하였다. 조사 대상자는 총 812명이며, 이중 응답자는 43명으로 5.3%의 낮은 회수율을 보였다. 조사대상자의 소속은 연구기관 34.9%, 공무원 34.9%, 교수 20.9%, 기업 7.0%, 학생 2.3%이다.

조사내용은 크게 세 가지 영역으로 구성된다. 첫째, 통계청 마이크로 데이터 사용 경험과 관련된 내용이다. 여기서는 통계청 마이크로 데이터 사용 여부와 사용 데이터 분야 및 활용 분석 분야를 조사하였다. 또한 마이크로 데이터를 사용하여 작성된 대표적 논문과 마이크로 데이터 사용 시 어려웠던 점과 이의 해결방법에 관해 조사하였다. 둘째, 통계청 마이크로 데이터 연계경험에 대해 조사하였다. 응답자의 연계분석 경험 유무를 screening하고, 연계분석 경험자에 한하여 연계에 사용된 자료와 연계분석 목적을 조사하였다. 또한 연계방법 및 연계분석에 적용된 통계적 기법을 알아보았다. 셋째, 향후 2차 통계 수요와 관련된 내용이다. 여기서는 통계청에서 제공해주었으면 하는 희망 2차 통계와 2차 통계를 활용하여 분석하고자 하는 분야를 구체적으로 응답하도록 하였다.

〈표 6-5〉 조사 내용

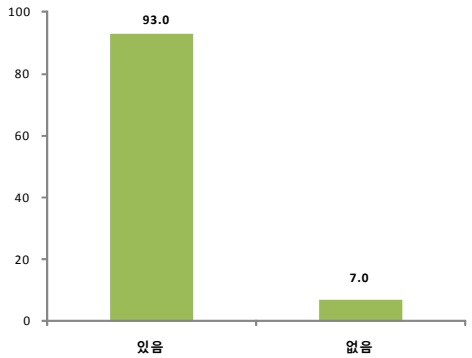
통계청 마이크로 데이터 사용 경험	◦ 통계청 마이크로 데이터 사용여부 ◦ 사용 데이터 분야 및 활용 분석 분야 ◦ 마이크로데이터 사용하여 작성된 대표적 논문 ◦ 마이크로데이터 사용 시 어려운 점 및 이의 해결방법
통계 연계분석 경험	◦ 연계분석 경험 유무 ◦ 연계분석 자료 및 연계목적 ◦ 연계방법 및 연계분석에 적용된 통계적 기법
향후 2차 통계 수요관련	◦ 통계청 제공 희망 2차 통계 ◦ 2차 통계 자료 활용계획 분야
인적사항	◦ 소속 및 소속기관 유형

2) 조사결과

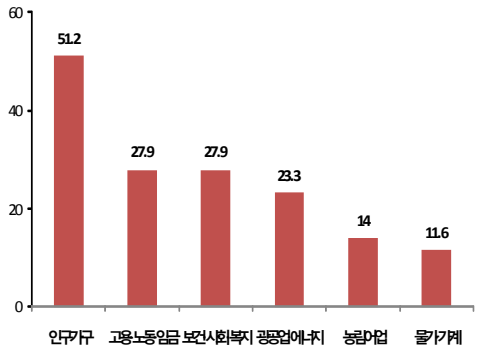
가) 통계청 micro-data 사용 경험

조사대상자에게 마이크로 데이터 사용여부와 이용분야를 조사하였다. 전체 응답자 중 93.0%가 마이크로 데이터 사용 경험이 있다고 응답했으며, 이용 분야는 인구·가구 51.2%, 고용·노동·임금 27.9%, 보건·사회·복지 27.9%, 광공업·에너지 23.3%, 농림어업 14.0%, 물가·가계 11.6%로 나타난다. 통계청 마이크로 데이터 이용 시 어려운 점으로는 자료의 상세정보 부족 60.5%, 필요한 변수의 부재 37.2%, data 가공의 어려움 20.9%, 이용 자료의 부족 16.3%로 나타난다. 한편 마이크로 데이터 이용 시 어려웠던 점의 해결방안으로는 담당자에게 문의하여 해결했다는 응답이 30.2%로 가장 높았으며, 해결 못했다는 응답은 11.6%, 보조정보를 활용하거나 다른 data 연계 및 예측값을 적용하여 해결했다는 응답은 각각 9.3%에 이른다.

이상의 응답 결과를 종합해보면 통계청 마이크로 데이터 사용에 있어서 다른 통계와의 연계 필요성을 크게 느끼고 있음을 알 수 있다. 사회·경제 현상을 분석함에 있어서 다양한 정보의 활용이 요구되지만, 단일 데이터에서 제공되는 정보만으로는 연구에 한계가 있기 때문이다. 통계청 마이크로 데이터 이용 시 어려운 점으로 상세정보의 부족, 필요변수 부재, 이용자료의 부족을 꼽은 것은 이 같은 연구자의 수요가 반영된 결과이다. 또한 이러한 한계를 극복하기 위해서 다른 data와의 연계 및 예측값, 보조정보를 활용했다는 응답 비중 또한 높아 외부 수요자들이 이미 다양한 방법으로 단일 데이터 정보 부족을 해결하려고 노력하고 있음을 알 수 있다.



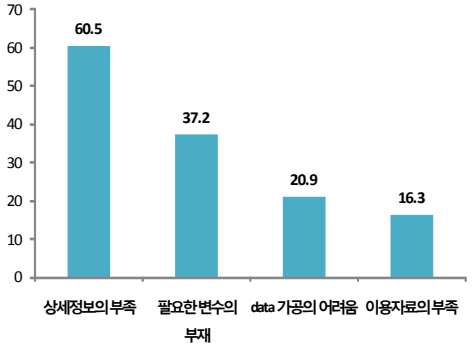
[그림 6-16] micro-data 이용 여부



주: 복수응답임

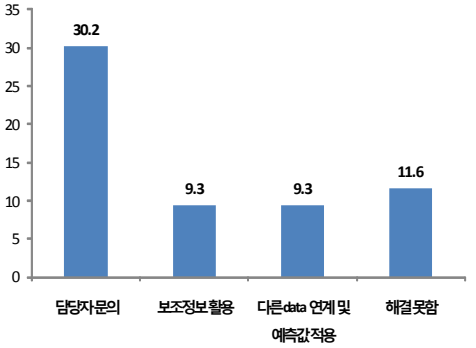
[그림 6-17] micro-data 이용 분야

제6장



주: 복수응답임

[그림 6-18] micro-data 이용 시 어려운 점



주: 복수응답임

[그림 6-19] micro-data 이용 시 어려웠던 점 해결방안

나) 자료 연계분석 경험

그렇다면 통계청 micro-data 이용자들 중 직접 데이터 연계분석 경험이 있는 비중은 어느 정도이며, 자료의 연계방법 및 분석 통계기법 종류는 어떠한가? 조사된 바에 따르면 전체 응답자 중 44.2%가 연계분석 경험이 '있다'고 응답하였다. 또한 자료 연계분야는 통계청 내 data 간 연계뿐만이 아닌 통계청 자료와 외부 연구기관의 자료 간 연계를 시도했다는 비중도 꽤 높았다. 자료 연계분야를 살펴보면 『경제활동인구조사』의 패널화, 『인구총조사』와 『사업체기초조사』, 『농어업총조사』, 『인구이동통계』의 연계 등으로 나타난다. 한편 통계청 자료와 외부 자료 간 연계는 『인구총조사』와 외부 연구기관의 패널자료, 『사업체기초조사』와 노동부의 고용보험 DB, 『사망원인통계』와 외부 임상자료, 『농어업총조사』와 농축산 관련 원시자료 간 연계 등으로 나타난다.

〈표 6-6〉 자료 연계 분야

통계청 내 data 간 연계	통계청 + 통계청 외 data 간 연계
- 『경제활동인구조사』의 패널화 - 『인구총조사』+『사업체기초조사』 - 『인구총조사』+『농어업총조사』 - 『인구총조사』+『인구이동통계』 - 『사망원인통계』+『경제활동인구조사』+『인구총조사』	- 『인구총조사』+ 연구기관 패널자료 - 『사업체기초조사』+ 고용보험 DB - 『사망원인통계』+ 암 발생자료+생활습관자료 - 『농어업총조사』+『농산물생산비조사』+농축산물 원시자료

자료 연계분석의 목적은 다양한데, 시계열 분석으로는 사업체 생멸이나 개인의 취업 상태 변화의 분석 등 사업체와 개인 노동시장 지위의 상태의 변화를 분석하기 위한 목적으로 연계하였다는 응답이 많았다. 한편 횡단면 분석으로는 인과적 모형의 검증(가령 인구의 유입이 지역 경제 활성화에 미치는 영향, 사망원인과 질병 발생원인과 관련요인과의 인과성, 사회·경제적 수준에 따른 손상 사망의 분석), 암환자의 진료관련 연구, 온실가스 통계 자료 생산, 생산액 추정, 생산성 및 효율성 분석, 모집단 추정 및 표본조사의 대표성 검정 등을 목적으로 연계하였다고 응답하였다.

자료의 연계방법은 다양한데, 자료연계의 key 변수로 활용된 정보는 가구번호, 우편번호, 사업체 정보, 사망원인, 사망일시, 생산액, 종업원 수 등을 사용하거나 지역단위의 macro-data로 변환 후 다른 자료와 연계하였다고 응답하였다. 연계자료의 분석 시 사용한 통계 기법은 단순 기술분석, 외삽법, 회귀분석, 로짓 분석, 인구추계기법, 요인분석, 군집분석, 사건사 분석 등이다.

다) 향후 2차 통계 수요

향후 통계청에서 제공을 희망하는 2차 통계에 대해 질문하였다. 이에 대한 응답을 인구, 고용 및 사회, 소득, 사업체의 네 개의 영역으로 구분하여 도식화하였다([그림 6-20]). 인구부문은 『인구총조사』와 『인구이동』, 『경제활동인구조사』 간 시계열 연계, 『사망원인』과 『경제활동인구조사』와의 연계, 『인구동향 자료』와 『사망원인』과의 연계를 희망하는 비중이 높았다.

[그림 6-20] 희망 2차 통계

특히 인구고령화로 건강에 대한 관심이 높아지면서 통계청의 인구자료와 임상자료 간의 결합에 대한 수요가 뚜렷하다. 고용 및 사회분야에서는『경제활동인구조사』자료와 다른 자료 간의 연계통계에 대한 수요가 높다. 특히 고용이동 분석을 위한『경제활동인구조사』자료의 패널화 및 다른 자료와의 연계,『사회조사』와『가계동향』,『경제활동인구조사』 등과의 연계 등을 희망하는 비중이 높다. 소득분야는『경제활동인구조사』와 마찬가지로『가계동향조사』자료의 시계열 연계,『가계동향조사』와『가계자산조사』,『가계동향조사』와 『경제활동인구조사』의 연계를 희망하는 비중이 높다. 사업체 부문은『광업·제조업조사』와『서비스업총조사』,『기업체모집단조사』와『사업체조사』간 연계를 희망하는 비중이 높다.

그렇다면 2차 통계를 활용하여 분석하고자 하는 분야는 어떠한가? [그림 6-21]은 통계청에서 제공을 희망하는 연계통계를 통해 연구자들의 예상 연구 분야를 영역별로 도시한 것이다. 인구부문의 연계분석 주제는 건강 및 보건에 관한 내용이 압도적이다. 특히 질병발생 원인, 사망자의 사회·경제적 수준, 암환자의 진단 후 생존기간 등은 통계청의 인구자료와 외부 임상자료 간의 연계를 통해 분석이 가능하다. 고용 및 사회분야는 노동공급변화 및 노동 투입량 추계, 공직자의 부정부패 추이 등에 대한 분석을 희망한다고 응답하였다. 소득 및 기업분야는 가구의 경제상태, 가구의 소득 및 소비변화, 특정산업의 규모 및 생산성 분석을 희망한다고 응답하였다. 또한 환경통계에 대한 관심도가 높아 연계통계를 통한 온실가스 통계자료 분석, 환경오염의 영향력, 지역별 비교통계 분석에 대한 관심도가 높았다.

[그림 6-21] 희망 2차 통계를 통한 분석 주제

제3절 연계가능 자료검토 및 매칭기법

1. 연계가능 자료검토

3절에서는 외부 전문가 수요조사를 바탕으로 통계청 자료 중 연계가능한 자료를 검토하고, 매칭기법을 설명하고자 한다. 앞서 2절 2.에서 연계자료 활용에 관한 외부 전문가 수요조사 결과를 제시하였다. 이를 토대로 통계청 자료 중 매칭가능한 자료를 검토해 보았다. 인구부문에서 외부 이용자들의 수요가 높은 통계는『인구총조사』와『인구이동통계』혹은『사망원인통계』를 결합한 자료이다.『인구총조사』는 전수조사이며, 성명, 성별, 연령, 교육수준 등 개인의 기본적인 특성을 파악할 수 있는 자료이다.『인구이동통계』는 행정자료를 기반으로 작성된 인구이동에 관한 통계로 두 자료를 결합하면 지역 단위의 인구특성과 인구이동에 관한 내용이 분석 가능하게 된다. 또한『사망원인통계』조사와 타 자료 간 연계에 대한 관심도 높은데, 이는 개인의 차별 사망력 분석에 대한 연구자의 수요 때문이다. 가령 평소 건강관리부터 어떠한 요인에 따라 차별적인 사망에 이르게 되는지 그 요인에 관한 연구가 가능하려면 사망 자료에 소득, 직업 등 여러 가지 사회·경제적인 변수의 결합이 요구된다.

고용 및 소득분야는 앞서 살펴본 바와 같이『경제활동인구조사』와『가계동향조사』에 대한 수요가 가장 높다.『경제활동인구조사』의 패널화를 통해 가능한 분석분야는 노동이동 등 취업상태에 대한 동태분석이며,『경제활동인구조사』와『가계조사』의 연계는 고용과 소득의 분석을 가능케 한다. 또한『가계동향조사』의 패널화는 소득 및 빈곤의 동태적 분석에 유용하다. 또한『사회조사』와 타 조사 간 연계에 대한 관심도 높은데, 이는『사회조사』는 개인의 주관적 의식에 대한 조사로 이에 대한 객관적인 지표가 부족하기 때문이다. 가령 개인의 분야별 의식수준에 경제활동상태나 소득에 관한 자세한 정보가 부가된다면『사회조사』의 활용도가 극대화될 것이다.

사업체 관련해서는『전국사업체조사』와 고용보험 DB와의 연계와『전국사업체조사』의 시계열 연계,『전국사업체조사』와『광업·제조업조사』의 연계를 검토해 볼 수 있겠다.『전국사업체조사』는 사업의 종류, 종사자 수, 연간매출액 등의 정보를 포함하고 있으나, 단일자료로 분석하기엔 수록된 정보가 매우 제한적이다. 가령 사업체 특성으로 매우 중요한 구체적인 직업 및 산업분류나 근로자 현황, 재무현황 등에 대한 정보는 포함하고 있지 않다.『전국사업체조사』는 패널화하여 생멸통계연구에도 활용할 수 있으며, 횡단면으로는 타 자료와 연계하여 분석의 활용도를 높일 수 있다는 이점이 있다.

〈표 6-7〉 연계가능 자료

분야	연계 가능 자료	key 변수	분석 내용
인구	『인구총조사』 +『인구이동통계』	지역별 집계 후 연계	- 지역별 인구특성과 인구이동
	『인구총조사』 +『사망원인통계』	지역별 집계 후 연계	- 임상보건연구
고용 및 소득	『경제활동인구조사』의 패널화	가구번호, 가구원번호 등	- 취업자의 노동이동
	『경제활동인구조사』 +『가계조사』	가구번호, 가구원번호 등	- 고용과 소득의 분석
	『경제활동인구조사』 +『경활부가조사』	가구번호, 가구원번호 등	- 비정규근로자의 특성
	『가계동향조사』의 패널화	가구번호	- 소득 및 빈곤의 동태적 분석
	『사회조사』 + 『가계조사』, 『경활조사』	가구번호	- 소득 및 취업상태에 따른 사회의식
사업체	『전국사업체조사』 +『고용보험 DB』	사업체등록번호	- 사업체 단위의 노동시장 특성 분석
	『전국사업체조사』 시계열 연계	사업체등록번호	- 생멸통계연구
	『전국사업체조사』 +『광업·제조업조사』	사업체등록번호	- 광업·제조업체의 특성

2. 매칭기법

가. 데이터 매칭의 종류

‘데이터 매칭(data matching)’이란 기존에 존재하는 자료 ‘A’와 ‘B’를 key 변수를 통해 연계하여 자료와 자료 간 정보를 통합하는 것을 말한다. 이는 ‘데이터 통합(data fusion)’이라고도 하며, 특히 통계적 방법을 이용한 데이터 매칭을 ‘통계적 매칭(statistical matching)’이라고 한다. 자료 분석을 통한 통계분석을 수행할 때, 필요로 하는 변수를 모두 포함하는 데이터 파일은 현실적으로 찾기 쉽지 않다.

이를 해결하기 위한 방안은 첫째, 필요한 변수를 포함한 데이터를 다시 수집하는 방법, 둘째, 통계적 기법을 사용해서 값을 할당(*assign*)하거나 대체(*imputation*)하는 방법, 셋째, 여러 데이터 파일을 이용해서 필요한 변수를 매칭(*matching*)시켜 사용하는 방법 등이

있다. 자료 간 매칭을 통해 추가적 정보를 얻는 방법은 시간과 비용을 절약할 수 있다는 장점 외에 분석과 추정에 있어서 더욱 신뢰성을 높일 수 있으며, 조사 응답자의 부담을 경감시키는 이점을 갖는다. 데이터 매칭은 별개의 데이터 파일을 결합하여 하나의 데이터 파일을 만드는 방법[23]으로 영국의 "*National Statistics code of Practice Protocol on Data Matching*(2003)"은 자료 매칭을 크게 5가지로 분류하였다(<표 6-8> 참조).

〈표 6-8〉 자료 매칭 방법

구분	방법
① 정확매칭 (*exact matching*)	- 주민등록번호, 국가보험번호, 사회보장번호 등 '개인 식별 ID'가 공통으로 있을 경우, 변수값 완전 일치 경우 데이터 결합 - 데이터 결합의 가장 이상적인 형태임
② 판단매칭 (*judgemental matching*)	- 정확히 일치되는 key 변수는 없지만, 자료에 대해 잘 알고 있다고 판단되는 case를 결합
③ 확률적 매칭 (*probability matching*)	- 정확 결합의 경우에 공통변수에 오류가 있을 경우 정확성에 따라 가중치를 주고 확률적으로 데이터 결합
④ 통계적 매칭 (*statistical matching*)	- 공통변수에 개인 식별 가능변수가 없을 때 수행하는 데이터 결합 방법 - 단계적 매칭, k-최근접이웃 매칭 알고리즘, 랜덤-핫덱 등
⑤ 데이터 연결 (*data linking*)	- 복수의 데이터 파일에서 변수들 간 연관성을 만들어 데이터 갱신이 가능하도록 하는 결합 방법

출처 : "National Statistics code of Practice Protocol on Data Matching(2003)"

첫째, '정확 매칭(*exact matching*)'이다. 이 방법은 주민등록번호, 국가보험번호, 사회보장번호와 같이 '개인별 식별 ID'를 나타낼 수 있는 변수가 공통으로 있을 경우, 변수값이 완전히 일치하는 경우에 데이터를 결합하는 방법이다. 같은 사람 또는 같은 물건을 완벽하게 결합하는 장점이 있고, 공통 변수에 측정오차가 없다면 이상적으로 데이터 매칭을 수행할 수 있는 장점이 있다.[24]

둘째, '판단 매칭(*judgemental matching*)'으로 공통인 변수들 사이에 정확히 일치되는 key 변수는 없지만, 자료에 대해 잘 알고 있는 경우, 적절하다고 판단되는 case를 결합하는 방법이다.

23) 데이터 간 매칭은 행정자료 활용과는 다름

24) 반면 사람과 관련된 경우에 개인의 교유한 정보를 이용해야 하므로 이러한 방법이 불가능하거나 사생활 침해의 여지가 있다는 단점이 있다.

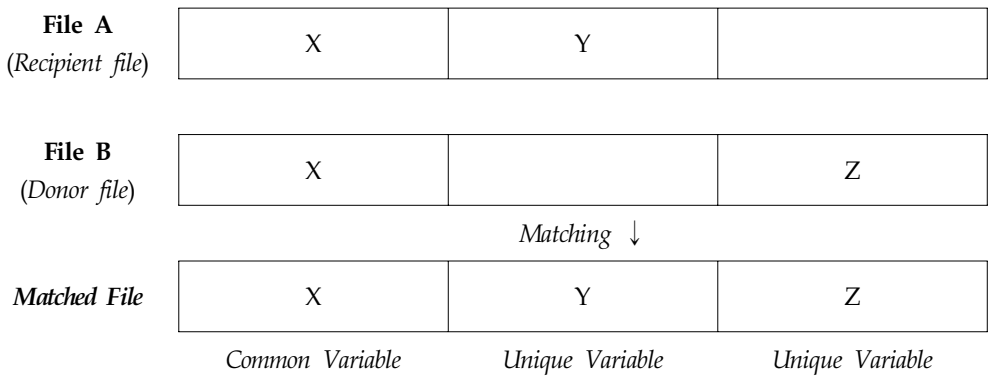
셋째, '확률적 매칭(*probability matching*)'으로 정확 결합의 경우에서 공통 변수들에 오류가 있는 경우에 정확한 정도에 따라 가중치를 주고 확률적으로 데이터를 결합하는 방법이다.

넷째, '통계적 매칭(*statistical matching*)'으로 공통변수에 개인 식별 가능한 변수가 없을 때 수행하는 데이터 결합 방법이다.

다섯째, '데이터 연결(*data linking*)'으로 둘 이상의 파일에서 변수들 간의 연관성을 만들어 내어 바로 데이터 갱신이 가능하도록 하는 데이터 결합 방법이다.

본 연구에서는 정확매칭 방법을 기본적으로 사용하되 확률적 매칭방법(*probability matching*)과 통계적 매칭방법(*statistical matching*)을 보완적으로 사용하고자 한다.

[그림 6-22]는 데이터 매칭으로 생성되는 데이터에 대한 설명이다. 서로 다른 경로로 얻어진 두 개의 파일을 고려해 보자. '*File A*'는 (*X*, *Y*)로 구성되어 있고 '*File B*'는 (*X*, *Z*)로 구성되어 있다고 하자. '*File A*'와 '*File B*'에 모두 관찰되는 변수 '*X*'를 공통변수(*common variable*)라 하고 '*File A*'에서만 관찰되는 변수 '*Y*'와 '*File B*'에서만 관찰되는 변수 '*Z*'를 유일변수(*unique variable*)라고 한다. 일반적으로 데이터 매칭을 수행하면 공통변수를 이용하여 '*File B*'에 있는 '*Z*'를 '*File A*'에 추가하게 된다. 이 때, '*File A*'를 수용파일(*recipient file*)이라 하고 '*File B*'를 제공파일(*donor file*)이라고 하며, 데이터 매칭을 수행한 후 생성된 파일을 결합파일(*matched file*)이라 한다.



출처 : 통계조사자료와 행정자료 간의 자료매칭기법 연구(2007), p.3

[그림 6-22] 데이터 매칭

나. 통계적 매칭

1) 구분과 제약조건

통계적 매칭의 접근방법을 두 가지로 나누어 볼 수 있다. 첫째, 수용파일에서 관찰되지 않은 변수를 예측하는데, 특정모형을 가정하지 않고 전적으로 데이터에 기초해서 통계적 결합을 수행하는 접근 방법이 있다. 이러한 접근 방법은 사전 준비 작업이 거의 없고 수행하기 쉽다는 장점이 있는 반면 계산 시간이 오래 걸린다는 단점이 있다.

둘째, 데이터의 특징을 잘 반영하는 모형을 사용하여 접근하는 방법으로 추상적인 모형을 만들어 관찰되지 않은 값을 예측하게 된다. 이렇게 하게 되면 일반화가 되는 장점이 있지만, 자료의 크기가 아주 큰 경우에는 자료의 형태가 매우 복잡하여 모형으로 설명하기가 어렵거나, 모형의 가정의 어긋날 경우 적절치 않다는 단점이 있다.

통계적 매칭을 수행하는 방법에 따라 '제약이 있는 결합(*constrained matching*)'과 '제약이 없는 결합(*unconstrained matching*)'으로 구분된다. '제약이 없는 결합'은 수용파일에 있는 모든 개체가 결합파일에서 나타나고, 제공파일의 모든 개체가 자료결합 과정에서 모두 사용될 필요는 없다. 이러한 결합은 결합파일에서 '*Z*'변수의 주변분포가 원래의 제공파일에서의 분포와 달라질 수 있다는 단점이 있다. '제약이 있는 결합'은 수용파일과 제공파일에 있는 모든 개체들이 한 번 이상 결합과정에서 이용되며, 자료결합을 수행했을 때 두 파일에 있는 모든 개체들이 결합파일에 나타난다. 이러한 결합은 공통변수인 'X'변수들 사이의 거리가 너무 멀어도 결합이 된다는 단점이 있다.

Van der Puttern et al.(2002)은 데이터 매칭이 유용한 결과를 도출하기 위해 다음과 같은 제약조건을 제시하였다. 첫째, '제공파일'은 '수용파일'을 대표할 수 있어야 한다. 그러나 반드시 두 데이터가 같은 모집단에서 추출될 필요는 없다. 둘째, 공통변수 'X'가 주어졌을 때, 유일변수인 'Y'와 'Z'사이에 다음과 같은 조건부 독립관계가 성립되어야 한다.

$$P(Y,Z|X) = P(Y|X) \cdot P(Z|X)$$

이러한 조건부 독립성(*CIA ; conditional independent assumption*)을 가정하는 이유는 수용파일과 제공파일 각각으로부터는 X, Y, Z의 결합확률분포함수(*joint probability distribution function*) $f(x,y,z)$를 추정할 수 없기 때문이다. 즉, $f(x,y,z) = f(y,z|x)f(x)$에서 수용파일과 제공파일 각각으로 부터는 $f(y,z|x)$가 추정 불가능하기 때문이다.

만약 CIA가 만족된다면, 즉 $f(y,z|x) = f(y|x)f(z|x)$이 성립된다면 (x,y,z)의 결합확률 분포함수는 다음과 같다.

$$f(x,y,z) = f(y|x)f(z|x)f(x)$$

여기서 $f(y|x)$는 수용파일로부터 추정 가능하고, $f(z|x)$는 제공파일로부터 추정 가능하다. 그러면 $f(x,y,z)$가 추정 가능하며 다음과 같이 $f(y,z)$도 추정 가능하다.

$$f(y,z) = \int_{-\infty}^{\infty} f(x,y,z)dx \quad \text{for continuous } x$$

이는 CIA가 만족되면 매칭 후 각각의 데이터로부터는 추정할 수 없었던 Y와 Z의 관계를 파악할 수 있게 된다는 것을 의미한다.

2) 수행과정

가) 자료의 준비

통계적 매칭 이전에 자료에 대한 검토가 필요하다. '제공파일'과 '수용파일'은 서로 다른 목적과 과정을 거쳐 얻어진 자료들이므로 '단위의 조화(*unit harmonization*)'와 '변수의 조화(*variable harmonization*)' 과정을 거쳐 통계적 결합을 효과적으로 수행할 수 있도록 해야 한다(*Marcello D'Orazio et al.*, 2006).

나) 매칭 매개 변수(matching variable)의 선택

자료가 잘 정리되어 준비가 되면 '공통변수' 중에서 매칭 '매개 변수'를 선택하여야 한다. 이 때 '수용파일'과 '제공파일'의 유일변수들 사이에 '조건부 독립성'을 가정하게 된다. 다시 말해서 매칭 매개 변수가 주어졌을 때, '수용파일'과 '제공파일'의 유일변수들은 서로 독립적이다.

조건부 독립 가정을 고려하고 나서 매칭 매개 변수를 선택하는데 있어서 주의해야 할 점이 있다. 사용가능한 모든 공통변수를 매칭 매개 변수로 하면, 변수의 차원이 높아져 표본이 공간상에 드물게 형성되며, 결과적으로 개체 간에 결합거리가 크게 측정되어 근접 결합이 힘들다. 따라서 자료에 대한 내용을 충분히 숙지하고, 일차적인 자료 분석을 한 이후에 적절한 매칭 매개 변수를 선택하고, 이에 따라 매칭을 수행하는 것이 바람직하다.

다) 근사성 측정

근사성 척도로서 거리를 사용하는 것이 일반적이다. 두 벡터를 잘 결합하기 위해서 많은 종류의 거리 측정 함수(*distance function*)가 사용된다. 예를 들면, 다음과 같은 유클리드 거리(*Euclid distance*), 마할라노비스 거리(*Mahalanobis distance*), 절대 거리(*Absolute*

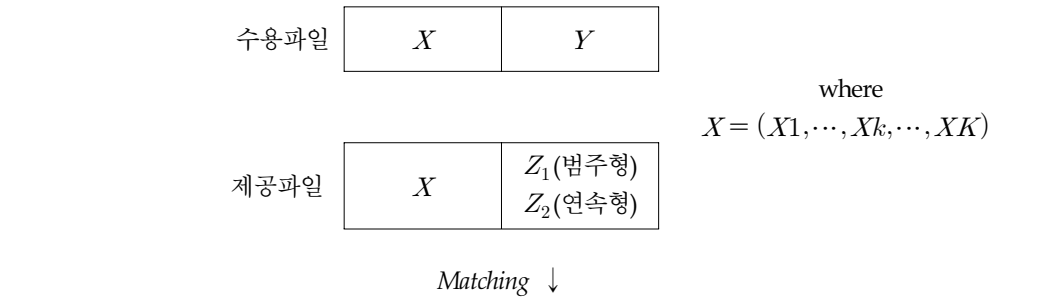
distance) 등이 있다.

◦ 유클리드 거리: $D_{ij} = \sqrt{(X_i - X_j)^2}$
◦ 마할라노비스 거리: $D_{ij} = \sqrt{(X_i - X_j)' \Sigma_{XX}^{-1} (X_i - X_j)}$,
(Σ_{XX}는 X변수들의 공분산 행렬)
◦ 절대 거리: $D_{ij} = |X_i - X_j|$

수용파일과 제공파일 간의 근사성을 측정하여 가장 유사한 개체들끼리 매칭이 이루어진다.

다. 통계적 매칭의 알고리즘

통계적 매칭 알고리즘의 설명에 앞서 이해를 돕기 위해 다음과 같은 데이터 구조를 가정한다.



[그림 6-23] 데이터 구조

1) 단계적 매칭 알고리즘[25]

가) 결합하려는 변수가 범주형인 경우

◦ Step1 : 로지스틱 회귀분석의 결과를 이용하여 '자료의 근사성'을 측정한다. 즉,

25) Van Pelt(2001)를 참조하였다.

제공파일의 결합하고자 하는 변수 'Z_1'을 종속변수로 하고 '공통변수'들을 독립변수로 하여 유의하게 나타난 변수를 중요변수로 간주하여 이들 간 근사성을 측정한다. 이 때 '수용파일'과 '제공파일'의 각 개체를 추정된 회귀식에 적합(適合)시켜 얻은 값을 근사성 측정을 위한 점수로 사용하게 된다.

추정된 회귀식을 이용한 근사성 측정 수식은 다음과 같다.

$$D_{ij}^{F} = |\widehat{Z_{1i^R}} - \widehat{Z_{1j^D}}| \quad \text{for given } i$$

$\widehat{Z_{1i^R}}$: 제공파일에서 추정된 회귀식을 수용파일에 적합시켜 구한 값

$\widehat{Z_{1j^D}}$: 제공파일에서 추정된 회귀식에 적합시켜 구한 값

D_{ij}^{F}가 작은 값을 갖는 수용파일의 i개체와 제공파일의 j개체를 결합하게 된다. Step1에서 측정한 근사성 정도(D_{ij}^{F})가 같아지게 되면 '수용파일' 하나의 개체에 여러 개의 제공파일 개체가 결합하게 된다. 이러한 경우 다음 단계로 추정된 회귀식에 포함되지 않은 다른 변수들을 이용하여 근사성을 측정한다.

◦ Step2 : 로지스틱회귀분석 결과 추정된 회귀식에 포함되지 않은 범주형 변수들을 이용하여 다음과 같이 두 번째로 근사성을 측정한다.

추정된 회귀식을 이용한 근사성 측정 수식은 다음과 같다.

$$D_{ij}^{S} = \sum_{k} I(Xk_i^R, Xk_j^D) \quad \text{for given } i$$

where $I(\cdot)$는 지시함수(indicator function): $I(a,b) = \begin{cases} 0, & \text{if } a = b \\ 1, & \text{if } a \neq b \end{cases}$

여기서 $\sum$은 범주형 변수들에 대해서만 이루어진다. D_{ij}^{S}가 작은 수용파일의 i개체와 제공파일의 j개체를 결합한다. 두 번째 단계에서 측정한 근사성이 같은 경우에도 마찬가지로 같은 값을 갖게 되면 다음 단계로 이용하지 않은 연속형 변수로 근사성을 측정한다.

◦ Step3 : 표준화한 연속형 변수의 차이로 다음과 같이 세 번째 근사성을 측정한다.

추정된 회귀식을 이용한 근사성 측정 수식은 다음과 같다.

$$D_{ij}^{T} = \sum |ZXk_i^R - ZXk_j^D| \quad \text{for given } i$$

여기서 $\sum$은 전단계에서 이용되지 않은 연속형 변수에 대해서만 이루어지며, 변수명 앞의 Z는 표준화값을 의미한다. D_{ij}^{T}가 작은 값을 갖는 수용파일의 i개체와 제공파일의 j개체를 결합한다.

나) 결합하려는 변수가 연속형인 경우

범주형 변수를 결합할 때 이용한 방법 그대로 적용한다. 단, Step1에서 결합하고자 하는 연속형 변수를 종속변수로 하고 나머지 변수를 독립변수로 하는 선형회귀분석을 수행한다.

다) 범주형 변수와 연속형 변수를 동시에 결합하는 경우

하나의 변수를 결합하는 것보다 여러 개의 변수를 한 번에 결합하는 경우가 더 일반적이다.

◦ Step1 : 결합하려는 변수가 범주형인 경우와 연속형인 경우의 첫 번째 단계의 순위 합으로 근사성을 측정한다.

$$D_{ij}^{RF} = RD_{ij}^{FZ_1} + RD_{ij}^{FZ_2} \quad \text{for given } i$$

$RD_{ij}^{FZ_1}$: 범주형 변수인 Z_1를 결합할 때, D_{ij}^{F}의 순위

$RD_{ij}^{FZ_2}$: 연속형 변수인 Z_2를 결합할 때, D_{ij}^{F}의 순위

D_{ij}^{RF}값이 작은 수용파일의 i개체와 제공파일의 j개체를 결합한다.

◦ Step2 : 결합하려는 변수가 범주형인 경우와 연속형인 경우의 두 번째 단계의 순위 합으로 근사성을 측정한다.

$$D_{ij}^{RS} = RD_{ij}^{SZ_1} + RD_{ij}^{SZ_2} \text{ for given } i$$

$RD_{ij}^{SZ_1}$: 범주형 변수인 Z_1를 결합할 때, D_{ij}^{S}의 순위

$RD_{ij}^{SZ_2}$: 연속형 변수인 Z_2를 결합할 때, D_{ij}^{S}의 순위

D_{ij}^{RS}값이 작은 수용파일의 i개체와 제공파일의 j개체를 결합한다.

◦ Step3 : 결합하려는 변수가 범주형인 경우와 연속형인 경우의 세 번째 단계의 순위 합으로 근사성을 측정한다.

$$D_{ij}^{RT} = RD_{ij}^{TZ_1} + RD_{ij}^{TZ_2}$$

$RD_{ij}^{TZ_1}$: 범주형 변수인 Z_1를 결합할 때, D_{ij}^{T}의 순위

$RD_{ij}^{TZ_2}$: 연속형 변수인 Z_2를 결합할 때, D_{ij}^{T}의 순위

D_{ij}^{RT}값이 작은 수용파일의 i개체와 제공파일의 j개체를 결합한다.

2) K-최근접이웃 매칭 알고리즘[26)]

최근접이웃 방법은 통계적 매칭에 가장 흔히 사용되는 방법으로 가장 유사한 하나의 개체를 매칭에 사용하는 방법이다. 여기서 한 단계 나아가 상대적으로 유사한 'k개'의 개체를 선택하여 매칭에 사용하는 방법이 'K-최근접이웃 방법'이다. Van der Putten et al.(2002)에 의해 제시된 데이터 매칭은 공통변수 'X'를 이용하여 가장 가까운 'k개'의 개체를 선택한 후, 이를 이용해 '통합변수'를 추가하는 방식으로 이루어진다. 이 방법을 자세히 살펴보면 다음의 단계로 이루어진다.

◦ Step1: 공통변수를 수치형으로 변환하고, 이를 이용하여 '수용파일'의 각 개체에 대해 제공파일의 모든 개체와의 거리를 계산한다. 거리계산은 유클리디안 거리를 흔히 사용한다.

◦ Step2: 계산한 거리 중 수용파일의 각 개체와 가장 가까운 제공파일의 'k개'의 개체를 선택한다.

◦ Step3: 선택된 'k개' 개체에 해당하는 제공파일의 '유일변수'를 이용하여 수용파일의 각 개체에 '통합변수'를 추가시킨다. 이 때, 유일변수가 '연속형'이면 'k개'의 '평균*(mean)*'을, '범주형'이면 '최빈값(*mode*)'을 이용한다.

실제 사례로 *D'Orazio et al.*(2006)의 "*Survey on Household Income and Wealth*(SHIW)" 자료와 "*Household Budget Survey*(HBS)"의 자료 매칭연구가 있다. 연구 내용을 요약하면

26) Van der Putten et al.(2002)을 참조하였다.

다음과 같다.

각 가구의 소비와 관련된 정보를 포함하는 'HBS 자료'와 각 가구의 소득과 관련된 정보를 포함하는 'SHIW 자료'를 사회·경제적 특성 정보를 이용하여 통계적 매칭을 한다. 이때 통계적 매칭방법은 최근접이웃 방법으로 'k=1'인 경우에 해당한다. 매칭과정은 다음의 세 가지 단계에 의해 이루어졌다.[27)]

(i) 두 조사 자료의 조화(*harmonization*)를 통해 자료의 일치성을 확인한다.
(ii) 두 자료의 통계적 프레임을 정의('유일변수'의 정의)하고 보조 정보로 활용가능한 변수를 정의('공통변수' 정의)한다.
(iii) 적합한 통계적 매칭 방법을 적용한다.

3) 회귀분석 매칭 알고리즘[28)]

회귀분석을 적용하여 매칭을 하는 방법은 먼저 하나의 데이터 파일에서 회귀모형을 추정한 후, 추정된 회귀모형을 이용하여 두 개의 데이터 파일에서 예측치를 구한다. 그리고 두 파일의 예측치 사이의 거리가 가장 짧은 개체를 찾음으로써 매칭이 이루어진다. 이 방법을 자세히 살펴보면 다음의 단계로 이루어진다.

◦ Step1: 제공자 파일의 유일변수 'Z'중 임의의 's'번째 변수를 목표변수로, 제공파일의 공통변수 'X'를 설명변수로 하여 회귀모형을 추정한다.

◦ Step2: 추정된 회귀모형을 수용파일과 제공파일에 적용하여 각 파일에서 's'번째 유일변수 'Z_s'의 예측치를 계산한다.

◦ Step3: 두 파일에서의 예측값을 이용하여 수용파일의 각 개체에 대해 모든 제공파일 개체와의 거리를 계산한다.

◦ Step4: 계산된 거리를 이용하여 수용파일의 각 개체에 가장 가까운 제공파일에 해당하는 개체의 유일변수 'Z_s'를 수용파일의 해당 개체에 추가한다. 이 때, 수용파일에 추가되는 값은 예측값 '$\widehat{Z_s}$'가 아니고 관측값 'Z_s'이다.

회귀분석에 의한 데이터 매칭 접근방법은 단순히 공통변수의 거리함수를 이용한 최근접이웃 방법과는 다르다. 최근접이웃 접근방법은 데이터 매칭이 이루어질 때 공통변

27) 각 단계별로 구체적인 내용은 D'Orazio et al.(2006)의 Application을 참조하시오.
28) Ingram et al.(2000)을 참조하였다.

수 'X'만을 이용하지만, 회귀분석 접근방법은 공통변수 'X'뿐만 아니라 제공파일의 유일변수 'Z'를 이용한다는 데 그 차이가 있다.[29)]

4) 회귀분석과 K-최근접이웃 방법의 결합 매칭 알고리즘[30)]

회귀분석 방법을 이용한 통계적 매칭방법은 추정치의 거리가 가장 가까운 하나의 개체만을 사용함으로써 상대적으로 유사한 다른 개체들의 정보를 무시하게 된다. 상대적으로 유사한 개체에 대한 정보손실을 줄여 데이터 통합기법의 성능을 높이고자 회귀분석기법에 'K-최근접이웃 접근법'을 결합하여 가장 가까운 하나의 개체가 아니라 'k개'의 개체를 이용하여 '통합변수'를 추가시키는 방법이다. 이 방법을 다음의 단계로 이루어진다.

◦ Step1: 제공파일의 유일변수 'Z'중 임의의 's'번째 변수를 '목표변수'로, 제공파일의 '공통변수' 'X'를 '설명변수'로 하여 회귀모형을 추정한다.

◦ Step2: 추정된 회귀모형을 수용파일과 제공파일에 적용하여 각 파일에서 's'번째 유일변수 'Z_s'의 예측치를 계산한다.

◦ Step3: 두 파일에서의 예측값을 이용하여 수용파일의 각 개체에 대해 모든 제공파일 개체와의 거리를 계산한다.

◦ Step4: 계산한 거리를 이용하여 수용파일의 각 개체에 가장 가까운 제공파일에 해당하는 'k개'의 개체를 선택한다.

◦ Step5: 선택된 제공자 파일의 'k개' 개체들의 유일변수 'Z_s'들의 평균(연속형인 경우)이나 최빈값(범주형인 경우)을 구한 후 이 값을 수용파일의 해당 개체에 추가한다.

5) 랜덤 핫덱 방법(Random Hot Deck)

랜덤 핫덱은 수용파일의 각 관측치에 대해 제공파일의 관측치를 랜덤하게 선택하여 매칭시키는 방법이다. 특히 수용파일과 제공파일의 관측치들은 대개 주어진 일반적인 특성(지리적 특성, 사회적 특성 등)에 따라 동질적인 부분집합으로 그룹화될 수 있다. 따라서 각각의 수용자 관측치에 대해 주어진 지형적 특성 내에서 동일지역의 관측치만

29) Ingram et al.(2000)은 실제로 현실에서 데이터 매칭 접근방법에 회귀분석과 같은 기법이 좋은 성능을 나타낸다고 보고한다.
30) 정성석 외(2004)를 참조하였다.

이 가능한 제공자로 고려된다. 일반적으로 하나 혹은 몇몇의 범주형 공통변수가 대체군(donation class)이 된다. 예를 들어 '파일 A'에는 6개의 관측치($n_A = 6$)와 3개의 변수 '성별', '연령', '연소득'이 있다고 하자(<표 6-9> 참조). 그리고 '파일 B'에는 10개의 관측치($n_B = 10$)와 3개의 변수 '성별', '연령', '연지출'이 있다고 하자(<표 6-10> 참조). 이때 파일 'A'를 수용자라 하고 '파일 B'를 제공자라고 하자. 그러면 2개의 공통변수 $\boldsymbol{X}=\{X_1 = '성별',\ X_2 = '연령'\}$와 각각의 유일변수 $Y='연소득'$과 $Z='연지출'$이 존재하게 된다.

'파일 A'의 각각의 관측치들은 '파일 B'의 10개의 관측치들로부터 랜덤하게 선택하여 제공자 값을 할당받게 된다. 만약 단위 'b'가 단위 'a'로 할당된다면 'a'에 존재하지 않는 'Z'값은 'b'의 관측된 'Z'값으로 매칭되게 된다. 즉, 최종 데이터의 'a'번째 관측치는 $(\boldsymbol{X}_a,\ y_a,\ z_b)$가 된다.

〈표 6-9〉 파일 'A'의 관측치

a	X_1	X_2	Y
1	F	27	22
2	M	35	19
3	M	41	47
4	F	61	41
5	F	52	17
6	F	39	26

〈표 6-10〉 파일 'B'의 관측치

b	X_1	X_2	Z
1	F	54	22
2	M	21	17
3	F	48	15
4	F	33	14
5	M	63	13
6	F	29	15
7	M	36	19
8	M	55	24
9	F	50	26
10	F	27	18

이론적으로 매칭결과는 $n_B^{n_A} = 10^6$가지 가능한 조합이 있다. 즉, '연지출'에 대한 10^6가지 가능한 분포가 있다. 예를 들어 다음의 <표 6-11>과 같은 매칭결과를 생각해볼 수 있다.

〈표 6-11〉 랜덤 핫덱 방법에 의한 파일 'A'와 'B'의 매칭결과

a	b donor	X_1^A	X_1^B	X_2^A	X_2^B	Y	Z
1	2	F	M	54	21	22	17
2	8	M	M	21	55	19	24
3	5	F	M	48	63	47	13
4	6	F	F	33	29	41	15
5	4	M	F	63	33	17	14
6	2	F	M	29	21	26	17

만약 공통변수 '성별'을 대체군을 정의하는데 사용한다면 '파일 B'에서 제공자는 수용파일의 각 관측치들에 대해 동일한 성별을 가진 관측치들 중에서 랜덤하게 선택될 것이다. 그러면 가능한 제공자 배열은 다음과 같이 급격하게 줄어든다.

$$(n_M^B)^{n_M^A} + (n_F^B)^{n_F^A} = 6^4 + 4^2 = 1312$$

다음의 <표 6-12>는 동일한 성별 계급내에서 랜덤하게 제공자가 선택된 결과이다.

〈표 6-12〉 동일 '성별' 내에서 파일 'A'와 'B'의 매칭결과

a	b donor	X_1^A	X_1^B	X_2^A	X_2^B	Y	Z
2	5	M	M	35	63	19	13
3	7	M	M	41	36	47	19
1	3	F	F	27	48	22	15
4	6	F	F	61	29	41	15
5	9	F	F	52	50	17	26
6	3	F	F	39	48	26	15

제4절 자료의 연계

1. 매칭과정 및 결과

가. 매칭과정

1) 결합자료 설명

『경제활동인구조사』(이하 경활조사 또는 경활자료)는 매월 15일, 전국 3만여 가구

의 15세 이상 7만여 가구원을 대상으로 조사되는 대표적인 노동력 조사로 표본틀로 인구총조사 10%표본 조사구 중 섬, 시설 단위 조사구를 제외한 27,011개 조사구를 사용하였다. 표본규모는 전국 32,000가구의 15세 이상 약 7천만 개인이며, 조사항목은 인적사항 및 개인의 경제활동상태 등이다.『가계동향조사』(이하 『가계조사』 또는 가계자료)는 매월 전국의 가구를 대상으로 조사되는 대표적인 소득 관련 조사로 전국의 9천 가구를 조사 대상으로 하여 가구주의 특성, 가구의 수입 및 지출, 가구구성 및 주거 특성을 조사한다.

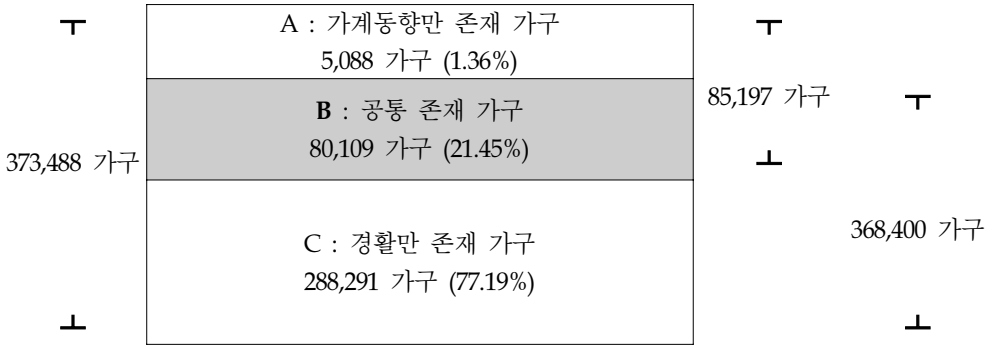
〈표 6-13〉『경제활동인구조사』와 『가계동향조사』 개요

	『경제활동인구조사』(2010년 기준)	『가계동향조사』(2010년 기준)
조사주기	매월	매월
조사대상	매월 15일 대한민국에 상주하는 만 15세 이상인 자*	전국의 일반가구**
조사기간	매월 15일이 포함된 1주간(일~토)	매월 1일~말일
표본설계 (모집단)	인구총조사(2005년) 10% 표본조사구 중 섬, 시설단위 조사구를 제외한 27,011조사구	인구총조사(2005년) 10% 표본조사구 중 섬, 시설단위 조사구를 제외한 27,011조사구
표본규모	32,000가구 약 7천만 개인	9,000가구
조사표 항목	인적사항(6개), 확인항목(5개), 취업자 항목(6개), 실업자 항목(7개), 비경제활동인구 항목(4개), 기타(7개)	가구주 특성, 가구의 수입 및 지출, 가구구성 및 주거특성

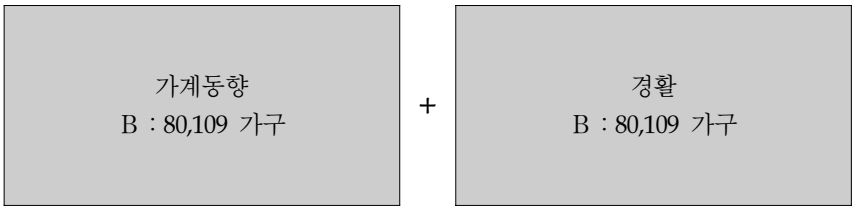
주 : * 군인 제외함
** 농림어가, 외국인가구, 비혈연가구 등 가구의 소득과 지출 파악이 곤란한 가구는 제외함

본 연구에서 자료의 매칭은 경활자료와 가계자료를 대상으로 하였으며, 분석 시기는 2009년으로 한정하였다. 자료의 결합방법과 결합결과를 논의하기 이전에 결합데이터에 대해 소개하고자 한다. [그림 6-24]는 경활자료와 가계자료의 가구단위 매칭 현황이다. 2009년도 누적자료 현황은 경활자료 368,400가구, 가계자료 85,197가구로 표본가구수의 차이로 인하여 경활자료의 가구수가 가계자료의 가구수보다 4.3배 가량 더 많다. 두 자료 간 표본가구수의 차이로 인하여 연계자료는 가계동향에만 존재하는 가구인 'A'와 두 자료 모두에 공통으로 존재하는 가구인 'B', 경활자료에만 존재하는 가구인 'C'가구로

구분된다. 각 영역별 가구수 및 비중은 'A'는 5,088가구로 연계가구의 1.4%를 차지하며, 'B'는 80,109가구로 연계가구의 21.5%, 'C'는 288,291가구로 연계가구의 77.2%를 차지한다. 이 중 본 연구의 최종분석 가구는 『가계조사』와 『경활조사』에 모두 존재하는 영역인 'B'의 80,109가구이며, 두 자료 간 가구단위 매칭률은 **21.5%**이다.[31]



[그림 6-24] 『경제활동인구조사』와 『가계조사』의 가구단위 매칭 현황: 2009년



[그림 6-25] 최종분석자료: 가구

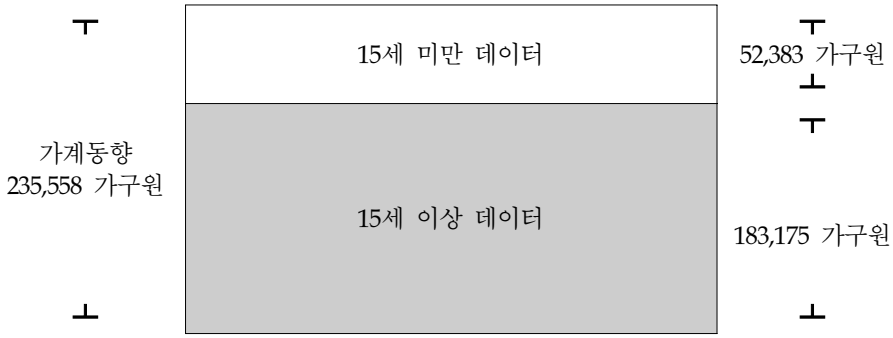
[그림 6-25]의 최종분석가구인 'B'의 경활자료 80,109가구의 가구원수는 182,679명이다. 그러나 가계자료는 개인단위가 아닌 가구단위로 구성되어 있기 때문에 개인별 자료의 매칭을 위해서는 가구단위 일치 자료인 80,109가구를 대상으로 가구원 정보를 세로로 정렬하는 데이터 변환과정을 요한다.[32] 데이터 변환 결과, 가계자료의 개인단위 총 record수는 80,109(가구단위)에서 235,558(개인단위)로 증가한 것을 알 수 있다([그림 6-26]). 개인 단위로 변환된 가계자료를 경활자료와 연계하기 위해 만 15세 이상 가구원만 선별하는 작업을 거쳤다. 개인단위로 변환한 235,558가구원 중 만 15세 이상 가구원

31) 두 자료 간 매칭되는 가구는 가계자료 85,197가구 중 80,109가구이며, 경활자료 368,400가구 중 80,109가구이다. 가계자료의 표본수가 경활자료에 비해 훨씬 적기 때문에 가계자료의 가구는 거의 연계자료에 포함되나, 경활자료의 가구는 288,291가구의 손실이 발생한다.

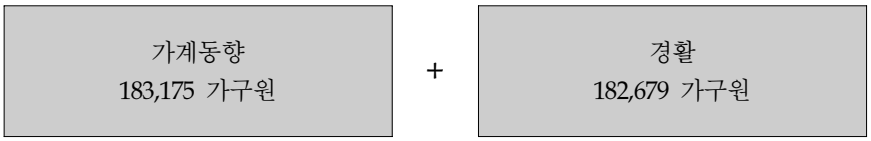
32) 데이터 변환 방법은 p.251의 [그림 6-10]과 [그림 6-11]에 한국노동패널 가구자료의 가구원 정보의 조작방법과 동일하다.

은 183,175명임을 알 수 있다([그림 6-26]참조).

[그림 6-27]은 최종 분석자료이다. 가계자료과 경활자료 중 일치하는 80,109가구를 최종 분석 자료로 사용하였으며, 해당 가구의 가구원은 가계자료는 183,175가구원이며, 경활조사는 182,679가구원으로 두 자료 간 496명의 가구원 차이가 발생함을 알 수 있다.



[그림 6-26] 가계자료의 변환 현황



[그림 6-27] 최종분석자료 : 개인

2) 결합방법

<표 6-14>는 가계자료와 경활자료 간 자료 매칭에 사용된 변수들이다. 데이터 매칭 방법 중 정확 매칭(exact matching)은 주민번호, 가구번호 등 '개인 ID'를 나타낼 수 있는 변수가 양 자료에 완전히 일치하는 경우 이를 통해 자료를 결합하는 방법이다. 이 방법은 동일 가구 혹은 동일 개인을 완벽하게 결합할 수 있다는 장점이 있고, 공통 변수에 측정오차만 없다면 데이터를 매칭할 수 있는 가장 이상적인 방법이다. 그러나 이 같은 매칭 방법을 사용하려면 매칭 예상 자료 간 '개인 ID'가 확보되어야 한다. 외부 연구기관의 패널자료의 경우 가구 자료에는 가구번호, 개인자료에는 개인번호가 존재하며, 이 변수는 차수를 거듭할수록 불변하는 것을 원칙으로 하고 있다.

그러나 경활 및 가계자료에는 원칙적으로 고유의 '개인 ID'는 존재하지 않는다. 따라서 두 자료 간 연계를 위해서는 두 자료에 공통으로 존재하는 변수를 조작하여 임의의 key 변수를 구성하여 매칭하는 작업을 거치게 된다. <표 6-14>는 매칭 가용 변수이다. 가계자료와 경활자료에 공통으로 존재하면서 key 변수로 사용가능한 변수는 조사년, 월,

가구번호, 가구주와의 관계, 성별, 연령, 교육수준 등이다.[33] 이러한 변수를 서로 다른 방법으로 조합하여 최적의 매칭 방법을 찾게 되는데, <표 6-15>는 매칭 방법에 따른 key 변수의 조합이다. '방법 1'은 조사년 및 가구번호, 가구주와의 관계, 성별, 연령을 통한 key 변수의 구성이다. '방법 2'는 조사년 및 가구번호, 가구주와의 관계, 성별, count[34]를 통한 key 변수의 조합이다. '방법 3'은 조사년 및 가구번호, 성별, 연령의 조합이다. '방법 4'는 조사년 및 가구번호, 가구주와의 관계,[35] 성별을 통한 key 변수의 조합이다.

〈표 6-14〉 자료 매칭 가용 변수

	가계동향 변수	경활 변수
조사 년	year	year
조사 월	month	month
시리얼번호(가구번호)	serial	serial
가구주 및 가구주와의 관계	hh_rel	rel
가구원 성별	hh_sex	sex
가구원 나이	hh_age	age
교육 정도	hh_edu	edu

〈표 6-15〉 매칭 방법별 key 변수의 구성

방법	key 변수의 구성
method 1	'조사년', '가구번호', '가구주와의 관계', '성별', '연령'
method 2	'조사년', '가구번호', '가구주와의 관계', 'COUNT*', '성별'
method 3	'조사년', '가구번호', '성별', '연령'
method 4	'조사년', '가구번호', '가구주와의 관계', '성별'

주 : * 조사년도, 가구번호, 연령을 정렬시킨 후 순번을 매겨 구성한 임의의 변수임

나. 매칭결과

이제 앞서 제시된 매칭 방법별 매칭결과를 제시하고, 한계점을 논의하고 한다.

<표 6-15>에서는 매칭방법으로 네 가지 방법을 제시하였으나, 여기서는 '방법 1(조사년, 가구번호, 가구주와의 관계, 성별, 연령)'과 '방법 2(조사 년, 가구번호, 가구주와의

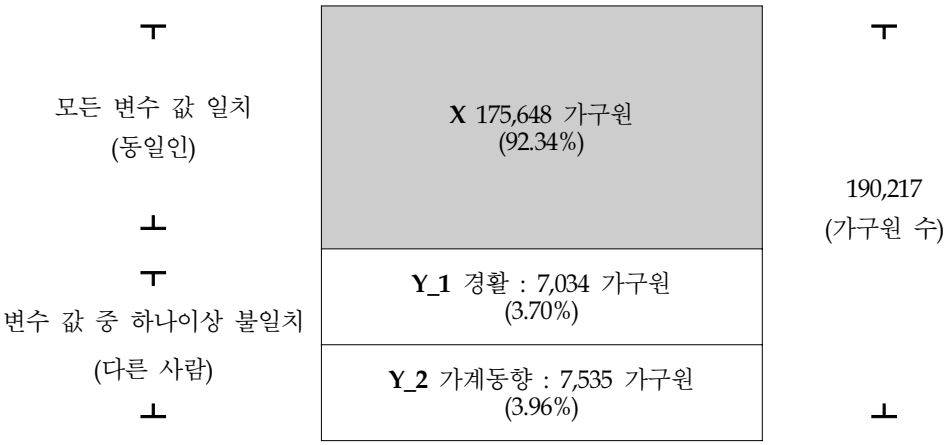
33) 외부 연구자들도 이같은 변수를 통해 key 변수 구성 후 자료를 매칭하는 방법을 취했다.
34) 'count'란 조사년도, 가구번호, 연령을 정렬시킨 후 순번을 매겨 구성한 임의의 변수이다.
35) 사실상 가구주와 배우자가 바뀐 경우가 존재하기 때문에 가구주와의 관계 '1'과 '2'를 통합하여 key 변수를 구성하였다.

관계, count, 성별)'에 의한 매칭결과만 제시하고자 한다.

'방법 1'은 조사 년, 가구번호, 가구주와의 관계, 성별, 연령을 조합하여 key 변수를 구성하였다. 따라서 이 정보가 동일하다면 동일인으로 간주하게 된다. [그림 6-28]에 제시된 매칭결과에 의하면, 사용된 key 변수가 일치하는 사례는 175,648명으로 매칭률은 92.34%에 이른다. 반면 변수 값 중 하나 이상 불일치하는 사례는 총 14,569명(비매칭률 7.66%)으로 경활 자료에서 7,034명(3.7%), 가계자료에서 7,535명(3.96%)으로 나타난다

[method 1]

◦ key 변수 구성 : '조사년', '가구번호', '가구주와의 관계', '성별', '연령'

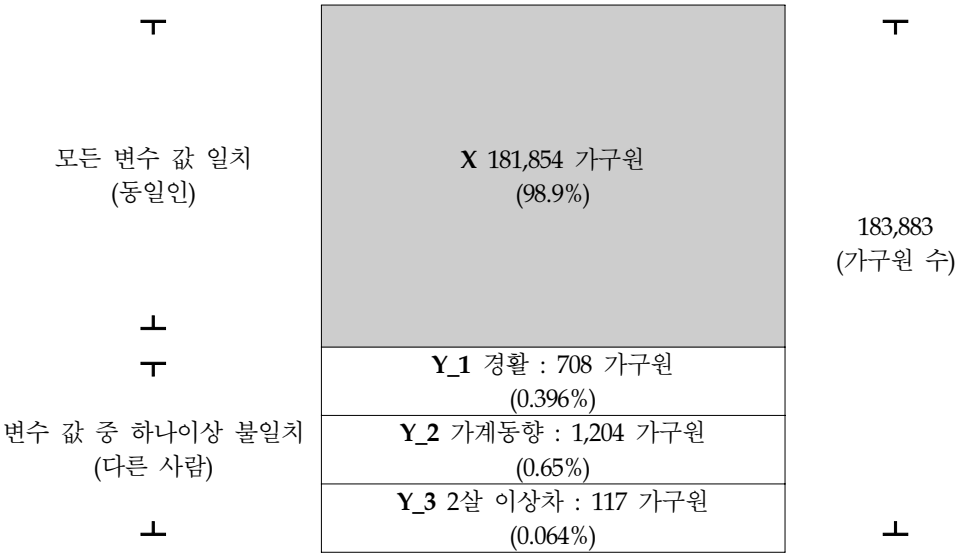


[그림 6-28] '방법 1'의 매칭 결과

'방법론 2'는 조사 년, 가구번호, 가구주와의 관계, COUNT, 성별을 조합하여 key 변수를 구성하였다. 따라서 이 정보가 동일하다면 동일인으로 간주하게 된다. [그림 6-29]에 제시된 매칭결과에 의하면 사용된 key 변수가 일치하는 사례는 181,854명(매칭률 98.9%)으로 [그림 6-28]의 '방법론 1'의 매칭률(92.3%)보다 6.6%p 높아진 것을 알 수 있다. 반면 변수 값 중 하나 이상 불일치하는 사례는 총 2,029명(비매칭률 1.1%)으로 경활자료에 708명(0.396%), 가계자료에 1,321명(0.71%)이다.

[method 2]

◦ key 변수 구성 : '조사 년', '가구번호', '가구주와의 관계', 'COUNT', '성별'



[그림 6-29] '방법 2'의 매칭 결과

2. 연계자료의 특성

가. 변수 설명

가계-경활 연계자료는 두 자료의 거의 모든 변수를 포함하고 있으므로 연계자료와 원자료 간의 주요 통계치를 비교하는 작업이 가능하다. <표 6-16>은 가계-경활 연계자료에 수록된 변수이다. 먼저 가구관련 변수로 가구의 기본특성 관련 변인은 조사년도, 가구유형 및 세대유형, 가구주 특성이다. 가구유형 관련 변인은 노인가구 여부, 비동거 배우자 및 비동거 자녀 여부이다. 가구주의 특성 관련 변인은 배우자 유무, 가구주의 직·산업 및 종사상 지위이다. 가구원 특성 관련 변인으로는 가구원의 인적특성 및 가구원의 직·산업 및 종사상 지위를 담고 있다. 또한 주택특성 관련 변인으로는 거처구분, 입주형태, 월세 평가액, 전세보증금, 사용면적, 주택소유 구분 등의 내용을 담고 있다. 가구정보 중 가장 중요한 변인은 역시 가구단위의 소득과 지출이다. 가구소득은 경상소득과 비경상소득으로 구분되어 있으며, 지출은 소비 및 비소비 지출로 구분된다. 그 밖에 가구의 자동차 소유대수 등의 정보를 담고 있다.

개인관련 변수는 주로 개인의 경제활동상태에 관한 내용을 담고 있다. 개인공통 응답문항은 주된 활동과 취업여부 및 휴직여부, 구직여부이다. 개인 중 취업자에게는 직·산업, 근로시간 및 종사상 지위, 부업여부 등을 조사하였다. 실업자는 구직여부와 구직가능성, 구직경로, 희망고용 형태 등을 조사하였다. 비경제활동인구는 구직의사와 취업가능성, 비

〈표 6-16〉 연계자료의 변수

대분류	중분류	소분류
가구사항	기본사항	· 조사년도 · 가구유형 · 세대유형 · 가구주 특성
	가구유형	· 노인가구 · 비동거 배우자 · 비동거 자녀
	세대유형	· 세대구분
	가구주 특성	· 배우자유무 · 가구주 직·산업 및 종사상 지위
	가구원 특성	· 가구원 인적특성 · 가구원 직·산업 및 종사상 지위
	비동거 배우자	· 취업배우자 · 학업배우자 · 기타배우자
	비동거 자녀	· 취업자녀 · 학업자녀 · 기타자녀
	자동차 소유대수	· 자동차 소유대수
	주택특성	· 거처구분 · 입주형태 · 월세평가액 · 전세보증금 · 월세(사글세) · 사용면적 · 주택소유 구분
	소득	· 경상소득(근로, 사업, 재산, 이전) · 비경상소득(경조, 퇴직금 및 연금)
	지출	· 소비지출 · 비소비지출
개인사항	주된 활동	· 지난 1주간 활동상태
	취업여부	· 지난 1주간 취업여부
	일시휴직여부	· 일시휴직 여부
	구직여부	· 1주간 구직여부 · 1개월간 구직여부
	취업자	· 부업여부 · 주업시간 · 부업시간 · 총취업시간 · 취업구분 · 36시간 미만 일하는 이유

개인사항		· 추가취업 유무 · 추가취업 가능성
	실업자	· 지난 주 구직여부 · 구직 가능성 · 구직 주요경로 · 구직방법 · 구직활동 기간 · 희망고용 형태
	비경제활동인구	· 구직의사 유무 · 취직가능성 유무 · 비구직 사유 · 구직활동 유무 · 구직활동 시기
	일자리 특성	· 이직시기 · 이직사유 · 직·산업 · 종사상 지위 · 취업시기 · 근로기간 유무

구직 사유 등을 조사하였다. 일자리 특성은 취업자에게만 조사된 것으로 이직시기 및 사유, 직·산업 및 종사상 지위, 취업 시기, 근로기간 유무 등의 내용을 담고 있다.

가계자료와 경활자료는 자료의 *record* 단위가 서로 상이하다. 전자는 가구단위이며, 후자는 개인단위로 조사되기 때문이다. 두 자료를 연계하기 위하여 가계자료를 가구원 단위로 변환한 후 경활자료와 연계하였다. 따라서 연계자료의 *record*는 개인단위로 구성되어 있으며, 개인단위의 경제활동 정보에 부가적으로 가구단위의 소득, 지출 등의 정보를 동시에 얻을 수 있다는 장점이 있다. 자료 생산자의 입장에서 보면 자료의 연계에 의한 정보의 확장이 큰 의미가 없는 것으로 여겨질 수 있겠지만, 자료 이용자 입장에서는 연계자료를 통해 자료의 활용성이 극대화되는 효과를 얻을 수가 있다. 특히 통계청의 마이크로 데이터 중 유난히 가계-경활 연계자료에 대한 수요가 많았던 이유는, 고용과 소득 정보를 동시에 제공하면서, 대표성이 확보된 자료가 사실상 존재하지 않기 때문이다.[36] 고용과 소득 자료는 정부의 사회 및 경제 정책수립의 근간이 되는 매우 중요한 자료이다. 가령 저소득 노동시장 분석, 근로빈곤(*working poor*), 여성가구주 가구의 빈곤, 맞벌이 가구 분석 등은『가계조사』나『경제활동인구조사』의 단일자료로의 분석만으로는 근본적으로 실태파악이 불가능하다. 이 같은 한계를 극복하기 위하여 외부 연구자들은 자료의 연계과정에서의 한계(*attrition, weight* 등)에도 불구하고 연계데이터를 통한 연구를 수행하였다. 이러한 연구들은

36) 해외통계청 역시 이러한 조사는 실시하지 않고 있다.

여러 가지 문제점을 안고 있는 한편, 관련 연구 및 정책적 수요가 존재하는 한 자료 이용자들은 앞으로도 자료의 연계와 분석을 끊임없이 시도할 것이다.

나. 주요 특성치 요약

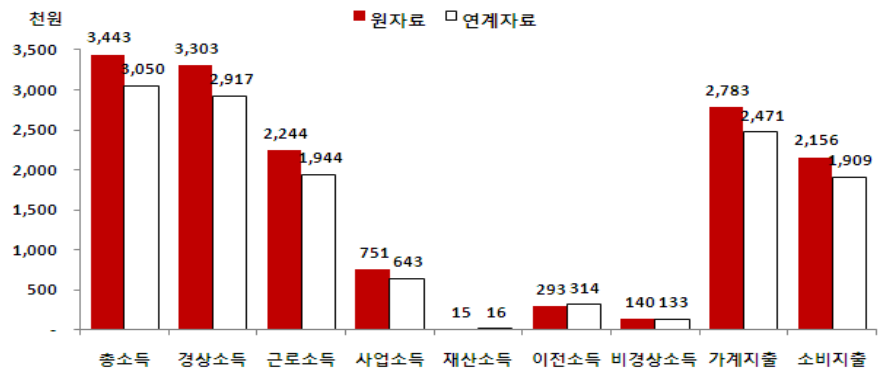
1) 가계소득 및 소비

<표 6-17>은 원자료와 연계자료의 가구소득과 소비의 평균값을 비교한 것이다. 연계자료에 포함된 가구는 『가계조사』와 경제활동인구조사에 공통으로 포함된 80,109가구로 가계동향 원자료 중 경제활동인구조사 자료와 일치하지 않는 5,088가구는 분석에서 제외된다. 2009년 원자료에서 나타나는 월평균 소득은 약 344만 원이며, 연계자료의 월소득

〈표 6-17〉 소득 · 소비 비교(전체)

(단위 : 천 원, %)

구분	원자료	연계자료[37]	원자료 대비 연계자료 비율
총소득	3,443	3,050	89%
경상소득	3,303	2,917	88%
근로소득	2,244	1,944	87%
사업소득	751	643	86%
재산소득	15	16	105%
이전소득	293	314	107%
비경상소득	140	133	95%
가계지출	2,783	2,471	89%
소비지출	2,156	1,909	89%



[그림 6-30] 소득 · 소비 비교(전체)

37) 연계자료에 사용된 가구수는 80,109가구로 연계자료의 소득에는 가중치가 적용되지 않았다.

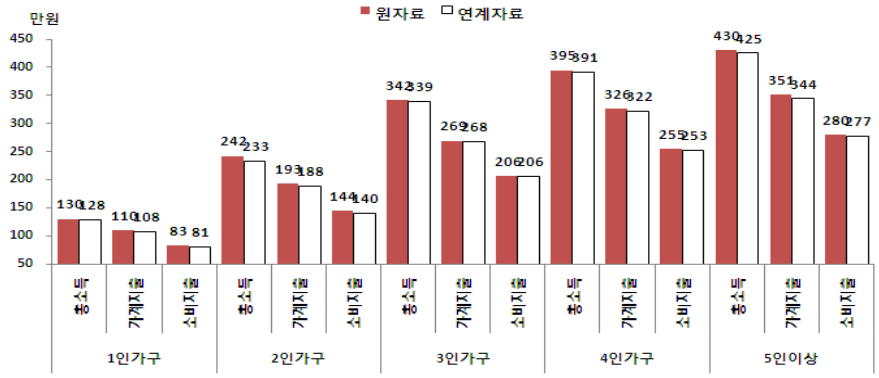
평균은 약 305만 원으로 원자료보다 약 39만 원이 적다. 한편 연계자료의 경상소득은 292만 원이며, 근로소득은 194만 원, 사업소득은 64만 원, 재산소득과 이전소득은 33만 원으로 재산소득과 이전소득은 연계자료의 값이 큰 반면, 나머지 소득항목은 원자료의 값이 연계자료보다 높게 나타난다.

<표 6-18>은 가구원수별 소득과 소비이다. 연계자료에서 나타나는 가구원수별 월평균 소득은 1인가구는 약 128만 원, 2인가구는 약 233만 원, 3인가구는 약 339만 원, 4인가구는 약 391만 원, 5인 이상 가구는 약 425만 원으로 원자료보다 연계자료의 월평균 소득이 적다. 소득 항목별 평균 소득을 살펴보면, 일부 항목에서 연계자료의 소득이 높은 구간이 발견된다. 1인가구의 경우 원자료의 비경상소득은 7만 원, 연계자료는 7만 5천 원으로 연계자료의 소득의 높으며, 4인가구 역시 원자료의 비경상소득은 10만 4천 원, 연계자료는

〈표 6-18〉 가구원수별 소득 · 소비 비교

(단위 : 천 원, %)

구분	항목	원자료	연계자료	원자료 대비 연계자료 비율
1인	총소득	1,299	1,279	98%
	경상소득	1,229	1,203	98%
	비경상소득	70	75	107%
	가계지출	1,102	1,083	98%
	소비지출	832	807	97%
2인	총소득	2,417	2,334	97%
	경상소득	2,248	2,173	97%
	비경상소득	170	161	95%
	가계지출	1,926	1,875	97%
	소비지출	1,438	1,395	97%
3인	총소득	3,416	3,393	99%
	경상소득	3,262	3,242	99%
	비경상소득	155	151	98%
	가계지출	2,690	2,682	100%
	소비지출	2,064	2,059	100%
4인	총소득	3,947	3,907	99%
	경상소득	3,843	3,798	99%
	비경상소득	104	108	105%
	가계지출	3,255	3,218	99%
	소비지출	2,552	2,530	99%
5인 이상	총소득	4,303	4,245	99%
	경상소득	4,151	4,082	98%
	비경상소득	152	163	107%
	가계지출	3,508	3,441	98%
	소비지출	2,797	2,765	99%



[그림 6-31] 가구원수별 소득 · 소비 비교

10만 8천 원, 5인 이상 가구 원자료의 비경상소득은 15만 2천 원, 연계자료는 16만 3천 원으로 1인가구와 4인가구 및 5인 이상 가구의 연계자료의 비경상소득은 원자료보다 높게 나타난다.

<표 6-19>는 가구주의 연령별 소득·소비이다. 연계자료의 가구주 연령 39세 이하 가구의 총소득은 약 333만 원이며, 40~49세는 약 367만 원, 50~59세는 약 341만 원, 60세 이상은 약 165만 원으로 총소득액은 원자료에서 나타나는 값보다 적다. 일부 소득항목에서 연계자료의 소득액이 높은 구간이 발견되는데, 40~49세의 비경상소득은 원자료 9만 4천 원, 연계자료 9만 5천 원으로 연계자료의 소득액이 약간 많으며, 50~59세의 비경상소득은 원자료 14만 3천 원, 연계자료는 15만 7천 원으로 연계자료의 소득액이 1만 3천 원 높다.

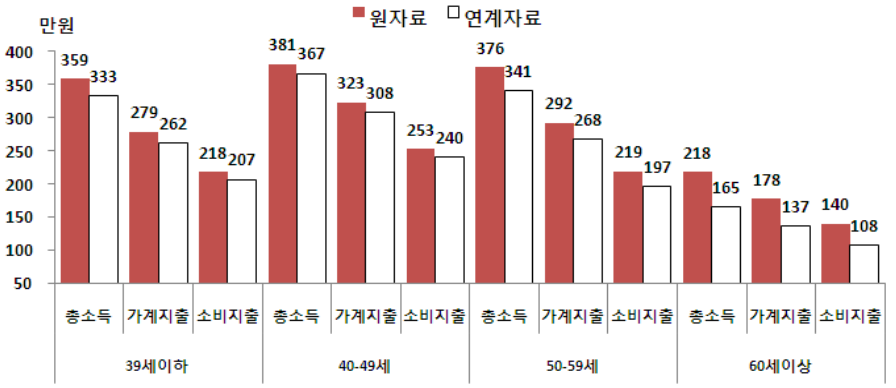
비경상 소득의 일부 구간을 제외하면 연계자료의 소득액은 원자료보다 낮은 수준인데, 연계자료의 경우 가중치가 적용되지 않은 값으로 두 자료 간 수치를 단순비교하기엔 무리가 있다. 또한 원자료 대비 연계자료의 가구 탈락률은 1.36%로 두 자료 간 결과값의 차이가 표본탈락에 의해 발생하였다고 보기 어렵다.

<표 6-20>은 맞벌이 여부별 소득·소비 비교이다. 원자료의 맞벌이 가구의 총소득은 약 426만 원, 비맞벌이 가구는 약 299만 원으로 맞벌이 가구의 월평균 소득이 약 127만 원 높다. 연계자료의 맞벌이 가구 월소득은 약 419만 원, 비맞벌이 가구는 약 254만 원으로 두 가구유형 간 소득 격차는 약 165만원으로 나타난다. 소득항목별 소득액을 살펴보면 대부분의 소득항목에서 원자료의 소득이 높다. 일부 예외는 경상소득 중 재산소득과 이전소득, 비경상소득에서 발견되는데, 맞벌이 가구의 재산소득 및 이전소득, 비경상소득은 연계자료의 소득액이 크며, 비맞벌이 가구의 이전소득 및 비경상소득 역시 연계자료의 소득액이 원자료보다 높다.

〈표 6-19〉 가구주 연령별 소득 · 소비 비교

(단위 : 천 원, %)

구분	항목	원자료	연계자료	원자료 대비 연계자료 비율
39세 이하	총소득	3,585	3,330	93%
	경상소득	3,411	3,177	93%
	비경상소득	174	153	88%
	가계지출	2,794	2,621	94%
	소비지출	2,180	2,067	95%
40~49세	총소득	3,808	3,666	96%
	경상소득	3,713	3,571	96%
	비경상소득	94	95	100%
	가계지출	3,233	3,080	95%
	소비지출	2,528	2,396	95%
50~59세	총소득	3,763	3,413	91%
	경상소득	3,620	3,256	90%
	비경상소득	143	157	109%
	가계지출	2,921	2,680	92%
	소비지출	2,193	1,974	90%
60세 이상	총소득	2,178	1,651	76%
	경상소득	2,012	1,517	75%
	비경상소득	166	135	81%
	가계지출	1,780	1,367	77%
	소비지출	1,397	1,078	77%

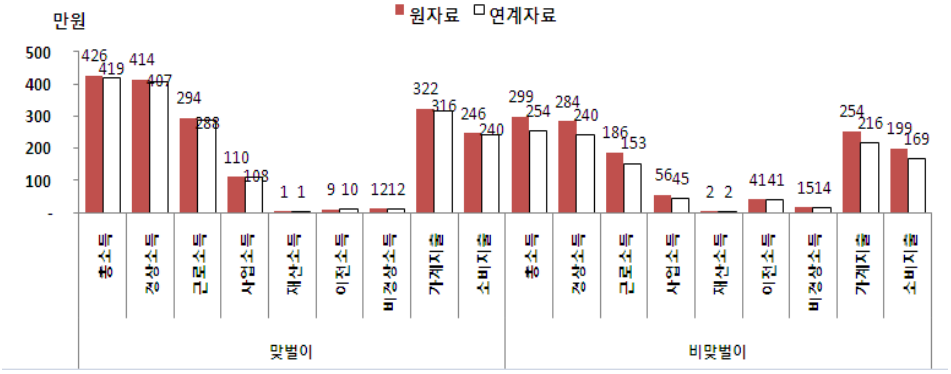


[그림 6-32] 가구주 연령별 소득 · 소비 비교

〈표 6-20〉 맞벌이 여부별 소득·소비 비교

(단위 : 천 원, %)

구분	항목	원자료	연계자료	원자료 대비 연계자료 비율
맞벌이	총소득	4,264	4,191	98%
	경상소득	4,144	4,067	98%
	근로소득	2,941	2,875	98%
	사업소득	1,103	1,081	98%
	재산소득	12	13	109%
	이전소득	89	99	111%
	비경상소득	120	124	104%
	가계지출	3,223	3,157	98%
	소비지출	2,460	2,402	98%
비맞벌이	총소득	2,989	2,539	85%
	경상소득	2,838	2,402	85%
	근로소득	1,859	1,528	82%
	사업소득	556	447	80%
	재산소득	17	17	101%
	이전소득	406	410	101%
	비경상소득	151	137	91%
	가계지출	2,540	2,164	85%
	소비지출	1,988	1,689	85%



[그림 6-33] 맞벌이 여부별 소득·소비 비교

제6장

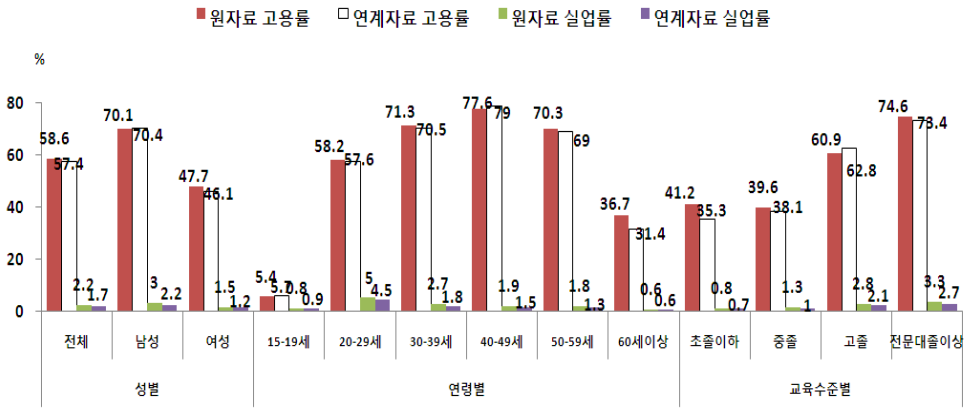
2) 경제활동상태

<표 6-21>은 원자료와 연계자료에 나타나는 경제활동상태를 개인특성별로 제시한 것이다. 원자료의 취업자는 약 2,351만 명으로 고용률은 58.6%이며, 실업자는 약 89만 명,

〈표 6-21〉 개인특성별 경제활동상태

(단위 : 천 명, %)

구분	항목	원자료			연계자료		
		경제활동인구		비경활	경제활동인구		비경활
		취업자	실업자		취업자	실업자	
성별	전체	23,506(58.6)	889(2.2)	15,698(39.2)	104,799(57.4)	3,090(1.7)	74,790(40.9)
	남성	13,734(70.1)	584(3.0)	5,278(26.9)	59,595(70.4)	1,888(2.2)	23,138(27.3)
	여성	9,772(47.7)	304(1.5)	10,420(50.8)	45,204(46.1)	1,202(1.2)	51,652(52.7)
연령별	15~19세	178(5.4)	25(0.8)	3,082(93.8)	913(5.7)	140(0.9)	15,049(93.5)
	20~29세	3,779(58.2)	323(5.0)	2,395(36.9)	13,028(57.6)	1,024(4.5)	8,552(37.8)
	30~39세	5,837(71.3)	220(2.7)	2,130(26.0)	28,083(70.5)	709(1.8)	11,038(27.7)
	40~49세	6,524(77.6)	163(1.9)	1,715(20.4)	31,826(79.0)	622(1.5)	7,840(19.5)
	50~59세	4,498(70.3)	114(1.8)	1,782(27.9)	20,011(69.0)	375(1.3)	8,599(29.7)
	60세 이상	2,690(36.7)	45(0.6)	4,595(62.7)	10,938(31.4)	220(0.6)	23,712(68.0)
교육 수준별	초졸 이하	2,717(41.2)	54(0.8)	3,816(57.9)	11,011(35.3)	209(0.7)	19,976(64.0)
	중졸	2,329(39.6)	76(1.3)	3,479(59.1)	11,179(38.1)	298(1.0)	17,848(60.9)
	고졸	9,486(60.9)	437(2.8)	5,649(36.3)	45,200(62.8)	1,511(2.1)	25,302(35.1)
	전문대졸 이상	2,810(74.6)	126(3.3)	833(22.1)	12,607(73.4)	470(2.7)	4,108(23.9)



[그림 6-34] 개인특성별 경제활동상태

비경제활동인구는 1,570만 명임을 알 수 있다. 한편 연계자료의 취업자는 약 10,480만 명으로 고용률은 57.4%이며, 실업자는 309만 명, 비경제활동인구는 7,479만 명이다. 가구소득과 마찬가지로 개인의 경제활동상태 역시 가중치가 적용되지 않은 값으로 두 자료 간 수치를 단선 비교하기에는 무리가 따르나, 대체적으로 연계자료의 취업자 비율은 원자료에서 비해 과소 포착되는 한편, 비경제활동인구는 원자료에 비해 과다 포착된다.

제5절 결 론

본 연구는 통계이용 활성화를 위한 2차 자료의 생산 · 확대 방안을 위해 수행되었다. 2차 자료란 연계자료 혹은 결합자료라고 불리는데, 이는 기존에 생산된 다수의 마이크로 데이터를 연계하여 새로운 통계를 생산하는 것을 말한다. 통계청에서 기 작성된 통계들은 인구 및 경제 사회의 다양한 영역을 포괄하며, 표본의 대표성 및 신뢰성을 확보하고 있다는 장점을 갖는다. 반면 하나의 자료 안에 수록하고 있는 정보는 매우 제한적이다. 따라서 통계 생산자가 아닌 외부 수요자(연구자)의 입장에서 본다면, 단일 자료로 일정수준 이상의 연구 성과를 올리기엔 한계가 따른다. 이미 사회는 급변하고 있으며, 통계 수요자의 욕구 또한 다양화되고 있다. 또한 최근의 일련의 정책 흐름은 저출산 · 고령화, 다문화, 사회취약계층의 복지 등으로 요약되는데, 이러한 정책적 흐름을 뒷받침하기 위해서는 무엇보다 신뢰있는 관련 통계의 확충이 필요하다. 또한 외부 수요자 요구에 대한 대응차원으로 연계통계의 생산 및 활용은 통계청의 매우 시급한 과제가 아닐 수 없다.

연계통계의 생산 및 제공은 호주에서 이미 활발히 진행되어 왔다. 특히 호주 통계청(ABS)은 인구보건 및 임상 데이터 연계를 통한 보건 분야 연구에 있어서 선두적인 위치를 차지하고 있다. 호주는 데이터 연계시스템(*the Western Australian Data Linkage System*)을 1995년부터 운영하여 왔는데, 이 시스템을 활용하여 많은 연구와 이를 토대로 한 정책 입안이 이뤄져왔다. 또한 외부 연구기관의 패널 데이터는 데이터 셋이 복수로 구분되어 있으며, 특정 key 변수를 통해 자료 간 연계가 가능하도록 설계되어 있다.

본 연구는 본격적인 연구 설계에 앞서 외부 전문가를 대상으로 통계청 자료 간 연계 경험에 대한 수요조사를 실시하였다. 이 조사 결과에 따르면 통계청의 마이크로 데이터 이용 시 어려운 점으로 상세정보의 부족과 필요변수의 부재, 데이터 가공의 어려움, 이용 자료의 부족 등을 꼽고 있으며, 이를 해결하기 위하여 담당자 문의, 보조정보 활용, 다른 데이터 연계 및 예측값 적용을 시도했다고 응답하고 있다. 또한 단일 통계자료의 정보 부족을 해결하기 위하여 통계청 내 데이터 간 연계를 시도했다는 응답의 비중 또

한 상당히 높았다. 이 같은 조사 결과는 외부 이용자들이 단일 데이터의 정보 한계를 극복하기 위하여 이미 다양한 노력을 하고 있음을 의미한다.

본 연구는 외부 전문가 수요조사 결과를 바탕으로 통계청 자료 중 연계가능한 자료를 검토하고, 외부 이용자들의 수요가 가장 높은『가계동향조사』와『경제활동인구조사』간 연계를 시도하였다. 두 자료의 2009년도 누적자료 현황은 경활자료 368,400가구, 가계자료 85,197가구로 표본가구수의 차이는 경활자료의 가구수가 가계자료보다 4.3배 가량 더 많다. 두 자료 간 표본가구수의 차이로 인하여 연계자료는 가계자료에만 존재하는 가구인 'A'와 두 자료 모두에 공통으로 존재하는 가구인 'B', 경활자료에만 존재하는 가구인 'C'가구로 구분된다. 각 영역별 가구수 및 비중은 'A'는 5,088가구로 연계가구의 1.4%를 차지하며, 'B'는 80,109가구로 연계가구의 21.5%, 'C'는 288,291가구로 연계가구의 77.2%를 차지한다. 이 중 본 연구의 최종 분석 가구는『가계조사』와『경활조사』에 모두 존재하는 영역인 'B'의 80,109가구이며, 두 자료 간 가구 단위 매칭률은 21.5%이다. 한편 최종 분석가구인 80,109가구의 가구원 수는 가계자료는 183,175명이며, 경활자료는 182,679명으로 나타난다.

연계자료를 구축한 후, 가계수지와 경제활동상태의 주요 특성치를 요약하였다. 연계자료는 기본적으로 고용과 소득의 모든 자료를 포괄한다. 그러나 최종자료에 가중치를 적용할 수 없는 한계 때문에 원자료와 연계자료 간 수치를 단선 비교하기에는 무리가 있다. 가계수지 및 경제활동상태의 주요 특성치를 비교한 결과, 일부 소득항목을 제외하고는 대부분의 소득에서 연계자료의 소득은 원자료의 소득보다 낮은 수준인 것으로 나타난다. 한편 두 자료 간 개인의 경제활동상태를 비교한 결과, 연계자료는 원자료에 비해 취업자는 과소 포착되는 반면 비경제활동인구는 과다 포착됨을 알 수 있었다.

『가계조사』와『경활조사』간 연계자료는 고용과 소득의 양쪽의 정보를 동시에 얻을 수 있다는 이점 때문에 매우 활용도가 높은 자료임을 분명하다. 그러나 연계자료의 활용도를 높이기 위해서는 향후 검토해야 할 부분이 많다. 가장 먼저 제기될 수 있는 문제는 표본탈락 과정에서의 *bias* 문제이다. 이에 대해서는 우선적으로 두 자료는 연동표본으로 *rotation sample*되면서 시차에 의해 비매칭되는 것이 아닌지에 관한 검토가 필요하다. 또한 두 자료 간 *record* 단위가 상이하기 때문에 연계자료에 대한 회의적인 시각이 있을 수 있다. 이에 대한 해결방안으로는 두 자료의 모든 정보를 활용하도록 데이터를 구성하기보다는 분석자의 요구에 따라 유형별로 자료를 구성하는 방안이 있을 수 있다. 가령 가구소득을 메인정보로 하되, 경활자료의 특정정보(가구주 및 배우자의 경제활동특성)를 추출하여 보완하는 방식이다. 이러한 방식으로 자료를 구성하게 된다면 가계자료를 중심으로 하여 가구주 및 가구원의 경제활동상태를 보조적으로 활용하는 방식이 가장 활용도 높은 자료의 구성이 될 것이다.

참고문헌

1. 국내문헌

고은애(2004),「통계적 매칭을 이용한 데이터 통합에 관한 연구」, 석사학위논문, 동국대학교 대학원
김혜련(2009),「가계동향조사와 경제활동인구조사의 연계분석」, 통계개발원
김혜련(2009),「근로빈곤의 동태적 분석」, 통계개발원 2009년 하반기 연구보고서 Ⅲ
김혜원 외(2008),『직장이동의 노동시장 효과 분석』, 한국노동연구원
김혜원·윤자영(2009),『여성가장 가구의 고용과 빈곤 연구』, 노동부 정책연구용역
남재량(1997),『우리나라 실업률 추세변화에 관한 연구』, 서울대학교 박사학위논문
남재량(2005),『고용불안계층의 실태 및 고용정책과제』, 한국노동연구원
안일호(2003),「혼합형 데이터의 통계적 결합에 관한 연구」, 석사학위논문, 고려대학교 대학원
이병희·정재호(2005),『노동이동과 인력개발 연구』, 한국노동연구원
이병희 외(2008a),『노동시장의 구조변화와 고용변동』, 한국노동연구원
이병희 외(2008b),『저소득 노동시장 분석』, 한국노동연구원
이영섭 외(2007),『통계조사자료와 행정자료 간의 자료 매칭기법 연구』, 통계개발원 정책연구용역
정성석, 김순영, 김현진 (2004),「데이터 보강을 위한 데이터 통합기법에 관한 연구」, 응용통계연구, 제 17권 3호, pp. 605-617.
정진호 외(2001),『소득불평등 및 빈곤의 실태와 정책과제』, 한국노동연구원
한상훈, 안일호, 하덕주, 최종후 (2004),「데이터퓨전과 평가」, 한국데이터마이닝학회 2004 추계학술대회, pp. 238-254.

2. 국외문헌

Aizawa, A. & Oyama, K.(2005), A fast linkage detection scheme for multi-source information integration, *in 'Web Information Retrieval and Intergration'*, Tokyo, pp. 30-39.

Baxter, R., Christen, P. & Churches, T.(2003), A comparison of fast blocking methods for record linkage, *in* 'ACM SIGKDD Workshop on Data Cleaning, Record Linkage and Object Consolidation', Washington DC, pp. 25-27.

Brook, E.L., Rosman, D.L., Holman, C.D.J. & Trutwein, B.(2005), 'Summary report: Research output project, WA Data Linkage Unit (1995~2003)', Western Australian Data Linkage Unit Perth.

Chang, C-C. & Lin, C.J.(2001), LIBSVM; A library for support vector machines, manual. Department of Computer Science, National Taiwan University.

Christen, P., Zhu, J.X., Hegland, M., Roberts, S., Nielsen, O.M., Churches, T. & Lim, K.(2002), High-performance computing techniques for record linkage, *in 'Ausralian Health Outcomes Conference'*, Canberra.

Christen, P., Churches, T. & Hegland, M.(2004), *Febrl* - A parallel open source data linkage system, *in 'Pacific-Asia Conference on Knowledge Discovery and Data Mining'*, Sydney, Springer LNAI 3056, pp. 638-647.

Christen, P.(2005), Probabilistic data generation for deduplication and data linkage, *in* 'International Conference on Intelligent Data Engineering and Automated Learning, Brisbane, Springer LNCS 3578, pp. 109-116.

Christen, P. & Belacic, D.(2005), Automated probabilistic address standardisation and verification, *in* 'Australasian Data Mining Conference'(AusDM '05), Sydney.

Christen, p., Willmore, A. & Churches, T.(2006), A probabilistic geocoding system utilising a parcel based address file, *in* 'Selected Papers from AusDM', Spring LNCS 3755, pp.130-145.

Christen, P.(2006), A comparison of personal name matching: Techniques and practical issues, in 'Workshop on Mining Complex Data'(MCD '06), held at IEEE ICDM'06, Hong Kong.

Christen, P. & Churches, T.(2006), Secure health data linkage and geocoding: Current approaches and research direction, *in* 'National e-Health Privacy and Security Symposium', Bris-bane, Australia.

Christen, P. & Goiser, K.(2007), Quality and complexity measures for data linkage and deduplication, *in* F. Guillet & H. Hamilton, eds, 'Quality measures in Data mining', Springer Studies in Computational intellingence, vol. 43, pp. 127-151.

Christen, P.(2007), A two-step classification approach to unsupervised record linkage, *in* 'Australasian data mining conference', Gold coast, conference in research and practice in *information technology*, vol. 70.

Church, T., Christen, P.(2004), 'Some methods for blindfolded record linkage', Biomed Central *Medical Informatics and Decision Making*, vol. 4, no.9.

Clarke, D. E.(2004), 'Practical introduction to record linkage for injury research', *Injury Prevention,* vol. 10, pp. 186-191.

Cohen, W.W. & Richman, J.(2002), Learning to match and cluster large high-dimensional data sets for data integration, *in* 'ACM International Conference on knowledge Discovery and data Mining', Edmonton, pp. 475-480.

Cohen W.W., Ravikumar P. & Fienberg S. E.(2003), A comparison of string distance metrics for name-matching tasks, *in* 'IJCAI-03 Workshop on information integration on the Web', Acapulco, pp. 73-78.

D'Orazio, Marcello, Di Zio, Marco and Scanu, Mauro(2006), Statistical Matching Theory and Practice, Wiley.

Fellegi, I. P. & Sunter, A. B.(1969), 'A theory for record linkage', *Journal of the American Statistical Society*, vol. 64, no. 328, pp. 1183-1210.

Goiser K. & Christen, P.(2006), Towards automated record linkage, *in* 'Australasian Data Mining conference', Sydney, conferences *in* Research and Practice in Information Technology, vol. 61, pp. 23-31.

Gu, L. & Baxter, R.(2004), Adaptive filtering for efficient record linkage, in 'SIAM international conference on data mining', Lake Buena Vista, Florida.

Gu, L. & Baxter, R.(2006), Decision models for record linkage, *in* 'Selected Papers from AusDM', Spring LNCS 3755, pp. 146-160.

Hernandez, M.A. & Stolfo, S.J.(1995), The merge/purge problem for large databases, *in* 'ACM international conference on management of data', San Jose, pp. 127-138.

Ingram, D., O'Hare, J., Scheuren, F. and Turek, J.(2000), Statistical matching: a new validation case study, Proceedings of the survey Research Methods Section, American Statistical Association.

Jin, L, Li, C. & Mehrotra, S.(2003), Efficient record linkage in large data sets, *in* 'International Conference on Database Systems for Advanced Applications', Tokyo, pp. 137-146.

Kelman, C, W., Bass, J. & Holman, C.D.J.(2002), 'Research use of linked health data-A best practice protocol', *Aust NZ Journal of Public Health*, vol. 26, pp. 251-255.

National Statistics(2003), National Statistics code of Practice-Protocol on Data Matching, London: TSO.

Rahm, E. & Do, H.H.(2000), 'Data cleaning: Problems and current approaches', IEEE Data Engineering Bulletin, vol. 23, no. 4, pp. 3-13.

Rässler, S.(2002). Statistical Matching: A frequentist theory, practical applications, and alternative Bayesian approaches, Springer Verlag, New York.

Rässler, S.(2004). Data fusion: identification problem, validity, and multiple imputation. Austrian Journal of Statistics 33(1-2), 153-171.

Rodgers, W.L.(1984), An Evaluation of Statistical Matching, *Journal of Business and Economic Statistics* 2, 91-102.

Saporta, G.(2002). Data fusion and data grafting, Computational Statistics & Data Analysis, 38, 465-473.

U.S. Department of Commerce(1980), Report on exact and statistical matching techniques, *Statistical Policy Working Paper* 5. Washington, DC: Federal Committee on Statistical Methodology.

Van Der Putten, Peter , Kok, Joost N., and Gupta, Amar(2002), Why the information explosion can be bad for data mining, and how data fusion provides a way out, Second SIAM International Conference on Data Mining, Arlington, April, 11-13.

Van Der Putten, Peter , Kok, Joost N., and Gupta, Amar(2002), Data Fusion through Statistical Matching, Technical Paper 185, Center for eBusiness@MIT, MITSloan (ebusiness.mit.edu)

Van Pelt, X.(2001), The Fusion Factory: A Constrained Data Fusion Approach. Master of Science. Thesis, Leiden Institute of Advanced Computer Science, The Netherlands.

Williams, G.J.(2007), 'Data Mining with Rattle and R', Togaware, Canberra. Software available.

Winkler, W.E.(2000), 'Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage', Technical report RR 2000/05, US Bureau of the Census.

Yancey, W.E.(2002), 'BigMatch: A program for extracting probable match from a large file for record linkage', Technical report RR2002/01, US Bureau of the Census.

Yoshizoe, Y. and Araki, M.(1999), Use of statistical matching for household surveys in Japan, In 52nd Session of the International Statistical Institute, Helsinki, Finland.

<부 록> 외부 전문가 수요 조사 설문지

통계이용 활성화를 위한 "2차 자료 생산" 전문가 의견 청취

설문조사기간	2010.06.14 ~ 2010.07.31
전체응답자	43명
보 기	전체 / 부분 / 응답자별

1 [********** 기존의 통계청 데이터 사용 경험에 대하여 **********]

2 다음의 통계청 마이크로 데이터를 사용해본 경험이 있으십니까?

(1) 예 (2번으로 이동)	93.0%	40명
(2) 아니오 (13번으로 이동)	7.0%	3명
총응답자수		43명

3 사용한 경험이 있으시면 다음의 어떤 통계자료였습니까?

(1) 인구·가구	51.2%/100%	22명/43명
(2) 고용·노동·임금	27.9%/100%	12명/43명
(3) 물가·가계	11.6%/100%	5명/43명
(4) 보건·사회·복지	27.9%/100%	12명/43명
(5) 환경		0명/43명
(6) 농림어업	14.0%/100%	6명/43명
(7) 광공업·에너지	23.3%/100%	10명/43명
(8) 건설·주택·토지	9.3%/100%	4명/43명
(9) 교통·정보통신	7.0%/100%	3명/43명
(10) 도소매·서비스	11.6%/100%	5명/43명
(11) 경기·기업경영	11.6%/100%	5명/43명
(12) 국민계정·지역계정·국가자산	7.0%/100%	3명/43명
(13) 재정·금융·보험		0명/43명
(14) 무역·외환·국제수지		0명/43명
(15) 교육·문화·과학	4.7%/100%	2명/43명
(16) 행정		0명/43명

4 마이크로 데이터를 활용하여 분석한 분야는 어떤 분야입니까?

0000(예 : 경제, 사회, 노동, 복지 등 상세히 적어주세요.)

상세내용보기

5 마이크로 데이터를 활용하여 작성한 대표적 논문제목을 적어주세요.

0000(하나만 적어주세요)(제목, 발간연도)

상세내용보기

6 마이크로 데이터 사용에 있어서의 어려운 점은 무엇이었습니까?

0000(2가지만 선택하여 주십시요)

(1) 이용자료의 부족	16.3%/100%	7명/43명
(2) 상세 정보의 제한	60.5%/100%	26명/43명
(3) 데이터 가공의 어려움	20.9%/100%	9명/43명
(4) 담당자 접촉의 어려움	7.0%/100%	3명/43명
(5) 어떠한 자료가 있나 몰라서	11.6%/100%	5명/43명
(6) 필요한 변수의 부재	37.2%/100%	16명/43명
(7) 기타	7.0%/100%	3명/43명

7 위 문항에서 기타로 응답하였다면 기타에 대해 서술해 주십시요.

상세내용보기

8 이용시 어려운 점들을 어떠한 방법으로 해결하였습니까?

0000(보조정보 활용, 담당자 문의 등)

상세내용보기

9 [********** 통계 연계 분석 경험 유무 **********]

10 [********** 통계청과 타기관 마이크로 데이터 연계분석에 관한 질문입니다. **********]

11 마이크로 데이터 이용시 2가지 이상의 통계나 하나의 통계의 시계열 자료를 가공하여 연계 분석한 경험이 있으십니까?

0000(통계청과 타기관 마이크로 데이터 포함)

(1) 예 (9번으로 이동)	44.2%	19명
(2) 아니오 (13번으로 이동)	48.8%	21명
불응답	7.0%	3명
총응답자수		43명

어떠한 통계와 어떠한 통계를 연계분석하였습니까? 연계분석한 모든 자료를 적어주세요.
12
0000(예 : 인구총조사 + 경제활동인구조사, 경제활동인구조사 + 노동패널)
상세내용보기

13 연계분석을 하였다면 어떠한 목적이었는지 서술해 주십시요.
상세내용보기

연계작업시 어떠한 방법을 사용하셨습니까? 상세히 설명해 주세요.
14
0000(예 : 가구번호, 월 변수 등의 키 변수 사용)
상세내용보기

연계분석시 적용한 통계적 기법은 무엇입니까?
15
0000(예 : 회귀분석, 다변량 분석, 계절 조정 등)
상세내용보기

16 [********** 향후 2차 통계 수요 관련 **********]

통계청에서 제공해 주었으면 하는 2차 통계(연계 통계)는 무엇입니까?

0000(예 : 구체적으로 어떤 통계의 어떤 변수와 어떤 변수가 필요한데 없었다)

0000❤ 통계청 마이크로 데이터 예시 ❤

0000가계동향조사
0000가계자산조사
0000가구소비실태조사
0000건설업조사
0000경제활동인구조사
0000광업제조업조사
0000국내인구이동통계
0000기업체모집단
17 0000기업활동조사
0000농가경제조사
0000농산물생산비조사
0000농어업법인조사
0000농업총조사
0000도소매및서비스업조사
0000사망원인통계조사
0000사회조사
0000산업총조사
0000생활시간조사
0000서비스업총조사
0000어업총조사
0000운수업조사
0000인구동향조사
0000인구주택총조사
0000인력실태조사

0000전국사업체조사

상세내용보기

18 2차 통계(연계통계)를 활용하여 연계분석한 자료는 어떤 분야에 활용하실 계획입니까?

0000(예 : 가계동향자료와 노동부 자료를 연계하여 어떤 분석에 활용하겠다.)

상세내용보기

19 [********** 일반 문항 **********]

20 귀하의 소속은 어디십니까?

0000(소속과 함께 직함도 적어주세요.예 : 소속, 직위)

상세내용보기

21 귀하의 소속 기관의 유형을 선택하여 주십시요.

(1) 공공기관	41.9%	18명
(2) 교육기관	16.3%	7명
(3) 연구기관	27.9%	12명
(4) 기업	2.3%	1명
(5) 기타	11.6%	5명
총응답자수		43명

22 [********** 질문에 응답해 주셔서 감사합니다. **********]

23 [********** 귀하의 설문 내용은 통계법에 의해 보호되며 국가 통계 발전에 큰 도움이 될 것입니다. **********]