



제1장

인구주택총조사 자료연계 및 활용방안 연구

- 인구주택총조사 종단자료 구축 중심으로 -

최필근 · 정미옥

제1절 서론

1. 연구배경 및 목적

최근 통계조사의 환경은 갈수록 악화되어지고 있다. 사생활 중시 및 기업비밀 보호 의식의 확산으로 조사 참여를 거부하거나 민감한 항목에 대해서 응답을 기피하는 현상이 늘어나고 있다. 또한 1인가구, 노인가구, 맞벌이가구, 소규모사업체의 증가 등 인구, 사회 및 경제 구조학적인 변화로 조사상의 어려움은 더욱 더 커질 전망이다. 하지만, 최근에 통계 이용자가 요구하는 자료의 수준은 점점 복잡해지고 다양해지는 추세로 이 모든 것을 충족시키기 위해 조사를 확대한다는 것은 한계가 있다. 조사를 통해 발생하는 비용과 시간, 그리고 조사자의 응답부담을 고려할 때 조사 확대가 근본적인 해결방안이 아닌 것만은 확실하다.

이러한 환경에서 통계 이용자의 요구를 충족시킬 수 있는 대안으로 현재 생산 가능한 자료를 활용하여 2차 자료를 생산하는 방안을 고려할 수 있다. 현재 통계청을 비롯하여 많은 통계작성 기관에서 다양한 통계를 생산 중이며, 수년간의 노력으로 통계청에서 입수하고 있는 행정자료만 해도 약 140종이다. 이 많은 자료를 연계한다고 가정하면 수많은 2차 자료의 생산이 가능할 것이다. 물론 연계가 가능한지, 연계를 위해 자료 사용이 가능한지 등에 대한 문제는 추가적으로 검토되어야 할 부분이다.

통계개발원에는 2차 자료의 생산을 위한 연구를 간헐적으로 수행해왔다. 이영섭 외(2007)는 사업체기초조사자료와 국민연금자료를 이용하여 통계조사자료와 행정자료 간의 연계를 시도하였다. 정미옥(2007)은 사업체모집단 보완을 위하여 행정자료의 활용 가능성을 검토하고, 사업체모집단의 기초자료인 사업체기초조사자료와 국민연금자료를 이용한 정확 연계를 수행하였다. 심규호·박시내(2010)는 2009년 경제활동인구조사와 가계동향조사의 공통 표본가구를 대상으로 가구번호, 가구주와의 관계, 성별, 연령 등의 변수의 조합을 통해 가구원의 정확연계를 수행하였으며, 2011년에는 기업활동조사와 광업제조업조사를 기업정보와 사업체정보를 바탕으로 정확연계를 실시하였다. 변종석 외(2013)는 2009년 경제활동인구조사와 생활시간조사 자료에 대하여 기존의 정확연계에서 벗어나 확률적 연계를 적용하였다. 기존의 연구들을 살펴보면 대부분이 중요한 연계키를 이용한 정확 연계를 실시한 연구들이며, 가장 최근 연구에서 처음으로 확률적 연계방법의 적용을 시도하였다. 하지만, 실제 활용하기까지는 추가적인 검토가 필요할 것으로 판단된다.

자료를 연계하여 사용하기 위해서는 정확연계를 기반으로 확률적 연계가 추가적으로 적용되어야 할 것이다. 모든 자료가 1대 1로 정확하게 연계되지 않으므로 연계되지 못한 자료의 처리가 중요하다고 할 수 있다. 만약 연계가 된 자료만을 이용하여 분석한다면 기존의 분포가 왜곡되어 올바른 정보를 얻을 수 없을지도 모른다. 이러한 문제를 해결하기 위해 확률적 연계를 활용한 추가적인 연계작업이 필요하다. 본 연구에서는 정확연계와 확률적 연계를 고려한 연계를 단계적으로 실시하고자 한다.

다음은 인구주택총조사 자료연계 필요성을 언급하고자 한다. 기존의 연구에서는 인구주택총조사 자료를 사용한 연계연구는 거의 찾아볼 수가 없다. 많은 시간과 비용을 들여 조사하는 센서스의 경우 자료의 가치를 강화하고 활용성을 제고할 필요성이 있다. 더 나아가 센서스 자료의 연계를 통해 종단화(패널) 자료를 구축한다면 더 많은 정보를 활용할 수 있을 것이다. 최근의 추세도 횡단면 측면뿐만 아니라 종단면 측면에서의 자료 중요성이 부각되고 있다. 일반적으로 패널조사를 실시하는데 있어 상당한 비용과 시간이 소요된다는 것은 알려진 사실이다. 따라서 이러한 비용과 시간을 고려할 때 센서스 자료를 연계하여 패널화를 할 수 있다면 이는 매우 의미 있는 작업이라 할 수 있다. 센서스에 포함된 다양한 항목들에 대하여 시간의 흐름에 따른 변화과정 및 타 자료와의 추가적인 연계에 의한 다양한 분석으로 통계이용자와 정책부서에 많은 정보를 제공할 수 있을 것이다.

본 연구에서는 인구주택총조사 자료를 연계하여 종단화 자료를 구축하는 방안과 활용 가능성을 제시하고자 한다. 현재 센서스 종단화 구축을 진행하고 있는 해외사례를 살펴보고, 한국 통계청의 현실에 맞는 종단화 자료를 구축하고자 한다. 본 연구를 통해서 한국의 인구주택총조사 자료의 종단화 구축의 기반을 마련하고, 향후 타 자료와의 연계를



실시하여 기존 통계의 활용도를 높이고 높은 품질의 2차 자료를 제공하는데 기여하고자 한다.

2. 연구내용 및 방법

본 연구는 인구주택총조사 자료의 연계를 통해 종단자료를 구축하는데 있다. 연구를 위한 자료는 2009년 인구주택총조사 시범예행조사 자료와 2010년 인구주택총조사 자료를 이용할 것이다. 센서스의 종단화 특성상 두 자료의 정확 연계율이 높을수록 의미가 있으므로 이름, 주소, 생년월일 등 연계키가 되는 항목에 대한 정비(표준화)가 필요하다. 현재 센서스 자료를 검토한 결과 2005년 센서스 자료는 연계의 핵심키인 이름 항목이 텍스트화 되어있지 않아 사용이 불가하며, 타 항목들에 대해서도 정비가 되어있지 않다. 반면에 시범예행조사 자료는 이름 등이 정비되어 연계에 사용가능하다. 따라서 센서스 종단화 연구의 초기 단계임을 감안하여 2005년 센서스 자료 대신 2009년 시범예행조사 자료를 활용하여 연계가능성을 검토하고자 한다.

첫 단계는 이름, 주소, 생년월일 등과 같은 핵심 항목을 이용하여 정확연계 비율을 높이하고자 한다. 핵심키를 적절히 사용하여 연계작업을 진행할 것이며, 필요시 직접 자료를 검토하여 정확 연계율을 향상시킬 것이다. 하지만, 시간과 인력의 한계가 있어 많은 부분 검토는 이루어지지 못할 것이다. 정확연계 후 시범예행조사 자료 전체와 정확연계된 자료를 비교분석하여 분포차이가 발생하는지 검토할 것이다. 다음 단계는 정확연계된 자료를 전체 자료로 가정하고 확률적 연계를 실시한다. 연계한 후 기존의 자료와 확률적으로 연계된 자료를 비교하여 확률적 연계 가능성을 검토할 것이다. 마지막으로 이 결과를 바탕으로 한국의 센서스 종단화 가능성을 제시할 것이다.

본 연구를 위하여 제2절에서는 센서스 자료의 종단화 구축을 실시하는 해외사례를 살펴보고, 구축과정과 방법을 정리한다. 제3절에서는 2009년 시범예행조사 자료를 기준으로 2010년 센서스 자료와 정확연계를 실시하고, 연계 및 미연계된 자료의 특성을 분석한다. 제4절에서는 확률적 연계과정 소개와 확률적 연계에 의한 결과를 비교분석한다. 마지막으로 제5절에서는 한국의 센서스 종단화 가능성 제시 등 연구의 최종적인 결론을 요약하고 본 보고서를 마무리하고자 한다.

제2절 센서스 자료의 종단화 구축 해외사례¹⁾

1. 호주의 종단센서스 자료구축

호주통계청은 센서스 자료의 활용성을 제고하기 위해 2006년 통계적 종단센서스 자료구축 (Statistical Longitudinal Census Dataset)을 제안하였으며, 2006년 센서스 자료를 기반으로 통계적 연계방법을 사용하여 2011년 자료와 연계하고 향후 2016년, 2021년 센서스 자료와 지속적으로 연계를 할 예정이다. 구축된 종단센서스 자료는 그 자체로도 유용하게 이용될 수 있으며, 더 나아가 특별한 통계적 목적을 위해 다른 자료와 연계되어 활용성을 높일 수 있다. 호주의 경우 종단센서스 자료를 활용하여 교육효과에 따른 직업의 변화 과정이나 부모의 사회적·경제적 수준이 자녀에게 미치는 세대간 전이 효과 등에 관한 정보를 얻고 있다. 또한 타 조사자료와의 연계를 통해서도 많은 정보를 획득한다. 가구조사와 연계하여 개인 또는 가족에 대한 조사 결과가 어떠한 영향으로 발생하는지 등의 영향 요소를 분석하고, 건강조사와 연계하여 사람의 과거 특성(사회적, 경제적, 육체적 환경)이 현재의 건강상태에 미치는 연관관계를 분석한다. 그리고 다양한 행정자료와도 연계될 수 있는데 출생사망 등록자료와의 연계를 통해 특별한 타입의 암으로 사망한 사람들의 공통적인 환경 요소(지역, 사회 및 경제적 환경, 직업, 수입, 교육 등) 분석이 가능하며, 이민자료와 연계하여 이민자를 위한 프로그램 종류와 이민기간 등의 요소가 이민자들의 정착에 어떠한 영향을 보이는지를 분석하여 이민정책을 위한 정보로도 활용 가능하다.

호주통계청은 2006년 자료와 2011년 자료의 종단화 구축에 앞서 사전준비 연계작업을 실시하였다. 2005년에 실시된 시험조사와 2006년 센서스 자료를 연계함으로써 연계의 정확성 및 연계방법에 대한 평가를 실시하였다. 호주는 센서스에 포함된 이름과 주소 항목은 센서스 자료처리가 모두 끝난 후에 폐기해야 하므로 센서스 간 연계를 위해 이름과 주소를 사용할 수 없다. 따라서 이름과 주소를 폐기하기 전 이름과 주소를 사용한 경우와 사용하지 않은 경우의 연계결과를 비교하여 연계의 정확성을 평가하였다. 그리고 주소 대응으로 매쉬 블록(mesh block)²⁾ 정보를 활용하여 연계의 정확성을 높이하고자 하였다. 다음으로 두 자료의 연계를 위해 연계 항목에 대한 표준화 작업이 필요하다. 이름 항목은 자동방법과 수동방법을 이용하여 정비하고 있다. 먼저 이름수정 소프트웨어와 이름사전을 이용하여 자동으로 정비하는데 사전의 이름과 정확하게 매칭되지 않는 경우에는 사전의 이름과 유사성에 대한 점수를 계산하여 일정 수준 이상인 경우에 사전의 이름으로 변환

1) Bishop, G. (2009), Guiver, T. (2011), Solon, R. and G. Bishop (2009), Trewin, D. (2006), Australian Bureau of Statistics (2013) 자료를 참고하여 작성

2) 매쉬 블록(mesh block)은 호주의 새로운 건물군으로 호주 전체에는 약 30만 개의 매쉬 블록이 있으며, 일반적으로 한 블록에는 30-60개의 거처가 포함된다.



한다. 자동으로 정비되지 않는 경우에는 수동방법을 이용하지만 비용을 고려하여 제한적으로 적용한다. 주소 항목은 조사당일 거주지, 평상시 거주지, 1년 전 거주지, 5년 전 거주지를 적절한 방법을 통해 정비(결측인 경우에는 대체)한다. 나이 항목은 다음 조사 기간(시험조사 1년, 본조사 5년)을 더하여 일치시킨다.

2005년 시험조사와 2006년 센서스 연계과정을 살펴보면 다음과 같다. 먼저 두 자료의 연계를 위해 변수 가중값(field weights)과 연계 가중값(linkage weights)을 활용하였다. 변수 가중값은 각각의 자료 쪽으로부터 특정변수에 대한 값을 비교하여 각 변수에 대한 중요도를 산출하는 것으로 m-확률과 u-확률을 이용하여 구한다. m-확률은 두 자료가 동일인일 경우 각 변수의 값이 일치할 확률이고, u-확률은 두 자료가 동일인이 아닐 경우 각 변수의 값이 일치할 확률을 의미한다. 예를 들면 성별의 경우 m-확률은 1, u-확률은 0.5 근처의 값을 가지며, 거리이름의 경우 u-확률은 0에 가까운 값을 가지게 된다. 호주의 경우 m-확률과 u-확률을 구하기 위하여 센서스 조사 후에 실시되는 사후조사 자료를 활용하였다. 그리고 최종적으로 변수 가중값은 각 변수에 대하여 일치 변수 가중값($\log_2(m/u)$)과 불일치 변수 가중값($\log_2(1 - m/1 - u)$)으로 정의하여 두 자료의 항목값이 같으면 (+)값, 다르면 (-)값으로 산출하였다. 이 값을 이용하여 각각의 자료 쪽에 대해 모든 변수값의 일치여부를 판단하여 변수 가중값을 합산한 연계 가중값을 구하게 된다. 따라서 연계 가중값이 클수록 유사한 사람과 연계될 가능성이 커짐을 알 수 있다.

센서스의 경우 자료의 크기가 방대하여 모든 개체를 비교하는 것은 너무나 많은 비교 횟수가 요구되므로 블록킹(blocking) 기법을 사용하여 효율적으로 비교한다. 이것은 블록킹을 하는 항목에 한정해서 같은 개체를 찾는데 있다. 만일 성별을 블록킹 변수로 사용한다면 두 자료에서 같은 성별을 가진 사람들만 비교하여 비교대상을 감소시키는 것이다. 이때 블록을 설정하여 연계하는 경우 변수의 값이 잘못된 경우나 결측값이 존재할 경우에는 비교자체가 제한되므로 주의가 필요하다.

작성된 연계 가중값에 대하여 연계유무를 결정하기 위한 방법으로 전통적인 cut-off 방법을 이용하며, lower and upper cut-off 방법과 단일 cut-off 방법이 있다. lower and upper cut-off 방법은 두 자료의 연계 가중값이 upper cut-off 값보다 크면 연계하고, lower cut-off 값보다 작으면 비연계, 두 cut-off 값 사이에 존재하면 clerical review를 실시하여 재분류한다. clerical review는 두 자료의 값을 직접 확인하여 동일인인지 판단하는 것으로 정확 연계율을 높일 수 있는 방법이나, 많은 시간과 비용 그리고 주관적 오류발생 가능성이 높다고 할 수 있다. lower and upper cut-off 방법을 사용할 경우 clerical review를 하기 위해서는 이름과 주소처럼 유용한 정보가 필요하나, 호주의 종단센서스 자료구축에는 이름과 주소를 사용하지 않는다. 따라서 호주는 주로 단일 cut-off 방법을 사용한다. 단일 cut-off 방법은 두 자료의 연계 가중값이 지정된 cut-off 값보다 크면 연계하고, 작으면

연계하지 않는 방법으로 최소한의 clerical review를 사용하여 cut-off 값을 결정한다. cut-off 값을 결정하는 방법은 첫째, 자료의 짝에 대한 연계 가중값 순으로 정렬하여 몇 개의 구간으로 분류한다. 둘째, 각 구간에 대해 임의로 표본을 뽑아 정확연계 비율을 추정하기 위한 clerical review를 실시한다. 그리고 마지막으로 추정된 값을 근거로 연계 가중값과 정확하게 연계된 수에 대한 누적 그래프를 이용하여 cut-off 값을 결정한다.

연계결과에 대한 평가를 위해 연계할 때 사용되는 변수에 따라 3가지 연계자료를 생성하였다. gold standard 자료는 이름, 주소, mesh block, 일반변수를 사용하며, 다른 두 방법을 평가하기 위한 benchmark로 활용된다. silver standard 자료는 이름과 주소는 사용하지 않고, hash value, mesh block, 일반변수를 사용하며 bronze standard 자료는 mesh block 과 일반 변수만을 사용하는 방법으로 호주의 종단센서스 자료구축에 사용할 방법이다. 따라서 gold standard 방법에 의한 연계결과를 기준으로 bronze standard 방법의 정확성 여부를 평가하는 것이다.

두 자료의 연계를 위한 소프트웨어는 2005년 개발된 FEBRL(Freely Extensible Biomedical Record Linkage)을 사용하는데 data cleaning과 표준화, 블록킹, 자료 짝 비교를 통한 연계 등 일련의 과정을 수행한다. FEBRL은 오픈 소스로 호주통계청에서는 수시로 기능을 보완·수정하여 정확한 연계가 가능하도록 관리하고 있다.

gold standard를 기준으로 2가지 자료에 대하여 두 가지 평가방법을 사용하여 연계 정확성을 비교하였다. 평가방법은 연계 정확도(link accuracy=true links/total links)와 정확 연계 비율(match-link rate=true links/total matches)이다. 연계유무를 결정하는 cut-off 값을 높게 정할수록 연계 정확도는 높아지나 정확연계 비율은 낮아진다. 이 경우 연계 후 부차 모집단에 대한 분포의 차이가 커질 수 있으므로 주의해야 하며, cut-off 값을 조금 더 낮게 조절함으로써 분포 차이 발생 문제를 극복할 수 있다. 연계 후 gold standard 연계와의 비교분석 결과를 살펴보면 silver standard 연계는 연계 정확도가 96.3%, 정확연계 비율이 91.3%, bronze standard 연계는 연계 정확도가 94.9%, 정확연계 비율이 78.1%로 나타났다. 두 연계 모두 cut-off 값이 낮은 경우에는 모두 좋은 결과를 보이고 있다.

사전준비 연계작업 후 2006년 센서스와 2011년 센서스의 연계를 실시하였다. 두 자료의 연계에서는 이름과 주소를 사용하지 않는 대신 mesh block 정보를 이용하는 bronze standard 연계를 시도하였다. 2006년 센서스로부터 선택된 5%(979,661명)의 표본을 2011년 센서스 자료에 연계하였다. 이때 특정 변수들에 대하여 정확연계를 실시하는 결정적(deterministic) 연계방법과 연계 가중값을 이용하여 가장 유사한 개체와 연계하는 확률적(probabilistic) 연계방법을 순차적으로 사용하였는데, 2005년 시험조사와 2006년 센서스를 연계할 때 사용한 대부분의 과정을 그대로 적용하였다. 결정적 연계방법을 위한 블록킹 변수는 주로 나이, 성별, 생년월일, 토착상태, mesh block 정보를 이용하였다.



연계결과는 총 979,661명 중 800,759명(81.7%)이 연계되었으며, 연계된 800,759명 중 결정적 연계방법에 의해 정확 연계된 사람의 수는 555,844명으로 70%이며, 확률적 연계방법에 의해 정확하게 연계되었을 것으로 추정되는 사람의 수는 202,122명으로 대략 25%로 나타났다. 따라서 연계된 사람에 대해서는 95% 정도가 정확 연계되었을 것으로 추정되며, 연계 대상 979,661명 중에서는 77%가 정확 연계된 것으로 보인다. 하지만 2011년 센서스 이전에 사망한 사람과 이민자 등을 고려하면 연계비율 및 정확 연계율은 더 높아질 것으로 판단된다. 그리고 확률적 연계방법에 의해 연계된 사람 중에서 정확 연계되지 않은 사람은 연계 대상자와 동일인은 아니지만 유사한 특성을 가지므로 일반적으로 통계적인 결과에는 영향을 주지 않을 것으로 생각된다.

호주통계청은 2006년 센서스와 2011년 센서스의 연계결과에 대하여 전반적으로 만족하며, 향후 2016년, 2021년 센서스와도 연계하여 지속적으로 종단자료를 생성할 예정이다. 그리고 종단센서스 자료구축 과정에서 해당 시점의 센서스 이후에 이민을 왔거나 출생한 사람에 대해서는 5%를 임의 선택하여 다음 센서스와 추가적으로 연계를 실시할 계획이다.

2. 영국의 종단센서스 자료구축³⁾

영국의 센서스 종단화 연구(LS)는 연속되는 센서스와 vital event 정보를 연계하는 작업으로 1974년에 처음으로 제안하였으며, 이 연구는 1971년 센서스 자료를 기반으로 1981년, 1991년, 2001년, 2011년 센서스 자료를 연계하고, 연계된 자료는 vital event 정보와 다시 연계하는 과정으로 진행된다. 영국에서 종단화 연구를 시작한 주요 목적은 vital event 정보와 연계하여 센서스의 활용성을 강화하기 위함이다. vital event 정보에는 여성의 출산 및 영아 사망, 암등록 여부, 배우자의 사망여부 등의 내용을 주로 포함하고 있으며, 직업관련 사망률 분석 및 여성의 출산력, 다양한 국민 건강 정보 등에 대한 정보를 제공한다. 또한 이민관련 자료 등과의 연계를 통해 이민관련 정책 및 사회 인구학적인 영역에 대한 종단 연구에도 많이 활용되고 있다.

2000년 이전 센서스 자료의 연계과정을 살펴보면 다음과 같다. 1971년 실시된 센서스로부터 잉글랜드와 웨일즈에 속한 사람들 중 1년 365일 중 특정한 4일에 출생한 사람 전체를 표본으로 선택하며, 선택된 사람은 대략 500,000명으로 전체 국민의 1% 정도가 된다. 센서스 자료의 연계는 국민보건서비스 등록자료(National Health Service Central Register)를 이용하여 수행하는데 이 국민보건서비스 등록자료에는 NHS 번호인 개인 식별번호가 있다. NHS 번호는 국민보건서비스가 실시된 1948년부터 잉글랜드와 웨일즈

3) Blackwell, L., K. Linch, J. Smith and P. Goldblatt (2003), Hattersley, L. (1997), Smith, J. (1999) 자료를 참고하여 작성

모든 국민에게 부여하고 있으며, 새로 출생하는 사람 및 이민자에게도 NHS 번호를 부여하여 지속적인 관리를 하고 있다.

연계과정의 사전단계로 1971년 센서스 자료로부터 특정 4일(예: 1.1, 4.1, 7.1, 10.1)에 출생한 사람을 선택하여 종단화 연구(LS)를 위한 멤버를 구성한다. 추출된 멤버들에 대하여 8자리 고유번호를 할당하고, 센서스 자료를 식별(조사구, 성별, 생년월일, 결혼상태 등)할 수 있는 정보와 함께 인덱스 카드를 출력하여 이름과 주소를 기록한다. 이 인덱스 카드는 알파벳 순으로 정렬되어 국민보건서비스 등록자료(NHSCR) 센터로 보내지게 되며, 기존의 등록자료와 연계를 실시한다. 연계작업 과정에서 등록자료에 LS 멤버가 있으면 해당 NHS 번호를 인덱스 카드에 기입하고, 등록자료의 해당 사람에게는 LS라고 표시한다. 이 때 등록자료에 LS 멤버가 없는 경우에는 NHS 번호를 부여할 수 없으며, 이 경우 LS 멤버는 “not traced”로 규정한다. 또한 NHS 번호가 없는 멤버는 vital event 정보를 연계할 수 없다. 1976년 시점에서는 1971년 센서스 자료의 초기 LS 멤버의 96.8%를 국민보건서비스 등록자료에서 찾았으며, 이를 이후 센서스 자료와 연계하게 된다.

1971년, 1981년, 1991년 센서스 자료의 연계과정은 [그림 1-1]에 자세하게 설명되어 있다. 1971-1981년 센서스 연계와 1981-1991년 센서스 연계는 동일한 방법으로 수행하였다. 1971년 자료의 사전절차와 같이 해당 센서스가 끝난 후 각 센서스 자료로부터 특정 4일에 출생한 사람이 있는 가구 자료를 추출하고, 이 중 LS 멤버로 될 가능성이 있는 사람에 대한 인덱스 카드를 작성한다. 그리고 1971년과 마찬가지로 이 인덱스 카드는 알파벳 순으로 정렬되어 NHSCR 센터로 보내지며 기존의 LS 멤버인지 확인하는 절차를 진행한다. 이 과정에서 기존의 LS 멤버로 확인되면 “traced” LS 멤버로 연계하여 처리하게 되며, 기준이 되는 센서스 이후에 출생하였거나 이민자인 경우에는 신규 LS 멤버로 처리한다. 또한 기존의 LS 멤버가 아닌데도 불구하고 국민보건서비스 등록자료에 있는 새로운 사람도 신규 LS 멤버로 관리하게 된다.

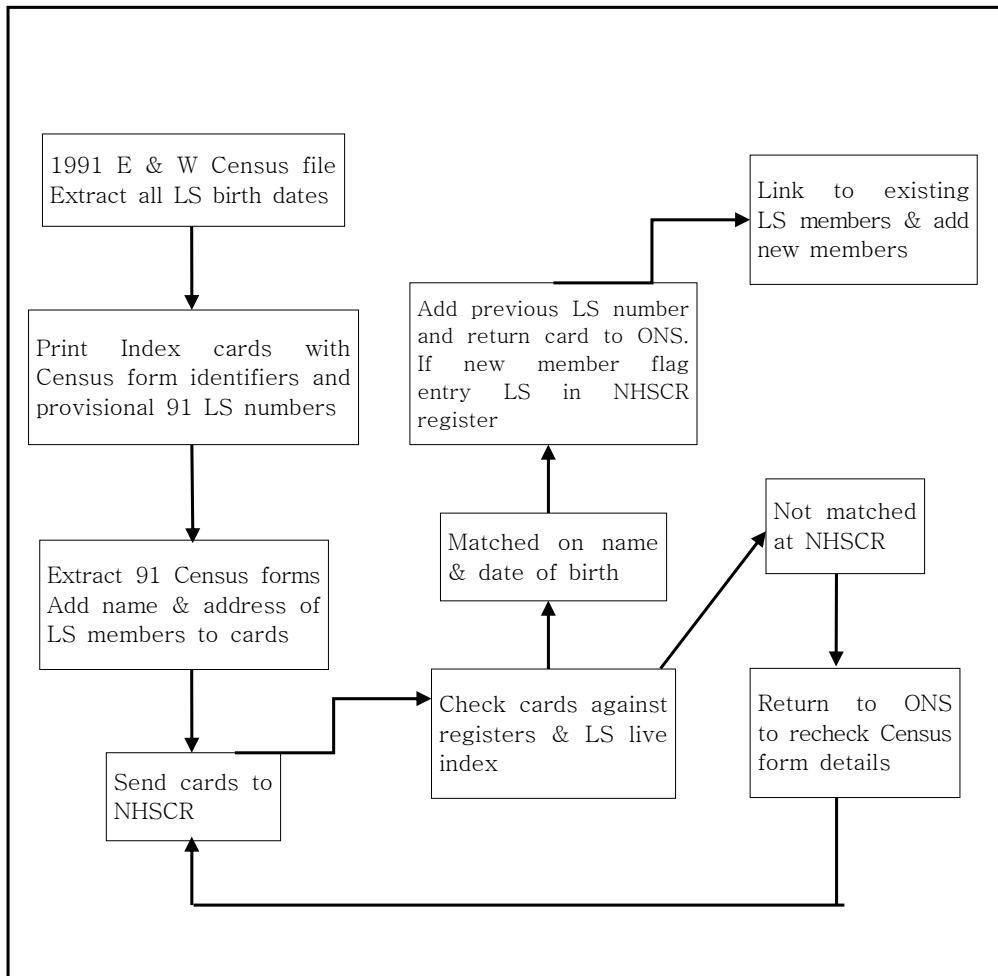
센서스 간 연계결과를 살펴보면 두 센서스의 정확연계 비율이 91%(1971-1981년), 90%(1981-1991년)로 센서스 이전에 사망한 사람 및 이민자를 고려하면 매우 성공적인 결과라고 할 수 있다. 그리고 2001년 이후에는 전산 발달로 인해 인덱스 카드 대신 컴퓨터 database를 이용하여 연계 작업을 실시하고 있다. 기존의 연계방법과는 차이가 없으며, 표준화된 이름, 주소, 생일, 성별, 지역코드 등을 이용하여 정확연계를 실시한다.

센서스 자료와 vital event 정보를 연계하는 과정은 다음과 같다. NHS 번호가 있는 멤버에 대해서는 vital event 정보를 연계할 수 있다. 센서스 자료 간 연계는 10년 주기로 이루어지지만, vital event 자료는 자료에 따라 1년 또는 6개월마다 연계하여 보완하는 작업을 수행한다.

vital event 정보의 연계과정은 발생한 vital event에 대한 정기적인 통지에 의한 방법과



ONS에서 작성한 vital event 통계 활용에 의한 방법으로 진행된다. 사망, 이민, 군복무와 관련된 사항은 주로 정기적 통지에 의한 방법에 의해서 연계된다. NHSCR 센터로 정기적인 통지가 이루어지고, LS 멤버와 관련된 정보가 있는지 확인한다. 이 때 LS 멤버로 확인된 사람이 있으면 관련된 사건 정보를 기록하여 ONS로 보내며 ONS에서는 database에 있는 LS 멤버에 대하여 기록된 사건의 정보를 연계하여 관리한다. 그리고 출생 및 사망, 영아 사망, 배우자 사망, 암등록 정보에 관련된 사항은 vital event 통계를 활용하여 연계한다. ONS에서는 작성한 vital event 통계를 활용하여 해당 출생일자를 가진 사람에 대한 자료를 NHSCR 센터로 보낸다. 이 때 NHSCR 센터에서는 확인된 LS 멤버에 대한 정보를 다시 ONS로 보내게 되며 ONS에서는 그 정보를 연계하여 관리한다.



[그림 1-1] 영국 센서스 자료의 연계과정

3. 뉴질랜드의 종단센서스 자료구축⁴⁾

뉴질랜드의 센서스 종단화 연구는 2012년 정부로부터 승인을 받아 처음으로 시작되었으며, 가장 최근의 자료인 2006년 센서스를 기반으로 2001년, 1996년, 1991년 센서스 자료를 역으로 연계하는 방법을 사용하였다. 뉴질랜드에서 센서스 종단화 연구를 시작한 주요 목적은 시간에 따라 변화하는 모집단의 현상을 이해·분석하여 센서스의 활용성을 강화하기 위함이다. 기본적인 분석내용은 호주의 센서스 자료 활용 내용과 유사한데, 교육 효과에 따른 직업의 성장과 패턴에 대한 결정요인 분석 및 세대 간의 전이 효과에 관한 다양한 분석을 통해 국가정책에 유용한 정보를 제공할 수 있다. 또한 출생·사망 등록자료에 대한 연계나 이주자료 연계를 통해서도 많은 정보의 활용이 가능하다.

뉴질랜드 통계청의 센서스 자료의 연계과정을 살펴보면 다음과 같다. 뉴질랜드의 경우 호주와 영국과는 달리 최근 센서스를 이전 센서스 자료에 연계시키는 backward 방법을 사용하고 있으며, 연계 방법은 호주와 유사하게 결정적(deterministic) 연계방법과 확률적(probabilistic) 연계방법을 순차적으로 적용하였다. 또한 종단화를 위한 표본을 추출하지 않고, 연계가 가능한 모든 국민을 대상으로 연계를 실시하였다. 2006년 센서스의 경우 전체 국민은 4,027,947명으로 집계되었으나, 결정적 연계방식을 적용시키기 위한 자료를 추가적으로 구성하였다. 이 자료에는 이전 센서스 이후에 출생한 사람과 이전 센서스 이전에 이민간 사람은 이전 센서스에 존재하지 않아 연계가 불가능하므로 제외되었으며, 정확연계에 사용되는 항목인 출생일자, 성별, 5년전 거주지가 대체된 사람도 정확연계를 위해서 제외되었다. 이러한 과정을 거쳐 결정적 연계방식 적용이 가능한 사람은 총 3,137,139명(출생, 이민, 대체자 포함하면 3,285,978명)으로 구성되었다.

먼저 출생일자, 성별, 5년전 거주지 항목을 이용하여 2006년 자료를 기준으로 2001년 자료와 결정적 연계를 실시하였으며, 총 3,285,978명 중에서 2,311,092명이 정확하게 연계되어 정확 연계율이 70.3% 수준으로 나타났다. 이 때 정확연계는 유일한 사람으로 매칭되는 경우로 한정하여 2명 이상의 매칭이 되는 경우는 정확연계에서 제외하였다.

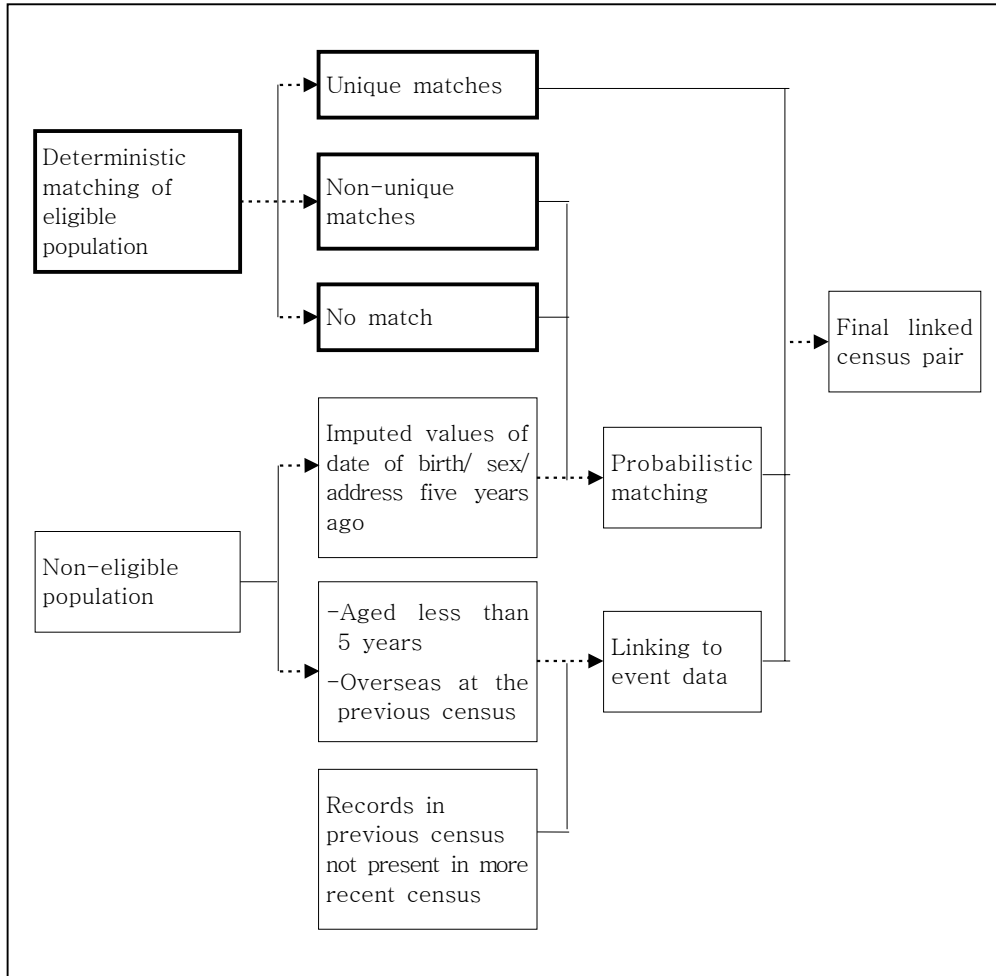
다음으로 정확하게 연계되지 않은 자료와 출생일자, 성별, 5년전 거주지 항목이 대체된 자료는 확률적 연계를 실시하였다. 확률적 연계는 전체 연계비율을 증가시킴과 동시에 기존의 분포가 왜곡되지 않도록 보완해주는 역할을 한다. 뉴질랜드의 경우 확률적 연계 방법은 호주의 경우와 유사하며, 이를 수행하기 위해 QualityStage라는 자체 소프트웨어를 개발하여 사용하고 있다.

마지막으로 이전 센서스 이후에 출생한 사람 및 이전 센서스 이전에 이민간 사람이나, 최근 센서스에는 존재하지 않으나 이전 센서스에 존재(사망자 포함)하는 자료처럼 연계

4) Statistics New Zealand (2006), Statistics New Zealand (2013) 자료를 참고하여 작성



되지 않는 자료는 event 정보와 함께 최종 자료에 포함하여 관리한다. 뉴질랜드의 센서스 자료 연계과정은 [그림 1-2]에 자세하게 설명되어 있다.



[그림 1-2] 뉴질랜드 센서스 자료의 연계과정

연계결과를 살펴보면 2006년 센서스를 기준으로 2006년-2001년 2개년 센서스의 정확 연계 비율은 70.3%, 2006년-1996년 3개년은 54.5%, 2006년-1991년 4개년은 45.4%, 2006년-1986년 5개년은 38.6% 수준으로 나타났다. 그리고 일반적으로 남성의 성향변화가 더 크므로 여성이 남성보다 정확 연계율이 높게 나타났으며, 주소의 잦은 변화가 발생하는 20대의 경우 정확 연계율이 낮은 경향을 보였다.

뉴질랜드도 호주와 유사하게 자료의 보완에 관해서는 매우 엄격한 기준을 적용하고

있어 종단 센서스 자료는 허가된 직원만 접근 가능하며, 모든 자료처리가 완료되면 이름과 주소 정보는 보관하지 않고 폐기하고 있다. 그리고 향후 센서스 자료의 종단화 작업도 계속해서 진행할 예정이다.

제3절 결정적 자료연계

본 절에서는 2009년 시범예행조사 자료를 기준으로 2010년 센서스 자료와 결정적 자료연계(정확연계)를 실시하고자 한다. 자료연계의 효율성은 결정적 자료연계의 비율과 밀접한 관련이 있다. 결정적 자료연계는 두 자료로부터 일치하는 사람(or 사업체)을 연결하는 것이다. 하지만, 주민등록번호나 사업체등록번호와 같은 유일한 식별번호를 사용할 수 없는 경우에는 연계 오류가 발생할 수 있음을 명심해야 한다. 이러한 오류를 최소화하기 위해서는 이름, 주소와 같은 개인(가구)의 또 다른 식별정보를 활용할 수 있다. 따라서 본 절에서는 센서스 자료에서 활용 가능한 식별정보인 이름, 주소, 생년월일(나이), 성별 등을 이용하여 가능한 범위에서 최대한의 정확연계를 실시하고, 그 결과를 분석·평가하고자 한다.

1. 연계자료 설명

인구주택총조사 자료연계를 위해 사용되는 자료는 2009년 시범예행조사 자료와 2010년 인구주택총조사 자료이다. 서론에서 설명하였듯이 센서스 자료의 연계는 2005년 자료를 사용하여 2010년 자료와 연계하는 것이 원칙이나, 2005년 센서스 자료는 연계의 핵심키인 이름 항목 등이 텍스트화 되어있지 않아 활용도가 떨어지므로 대용으로 2009년 시범예행조사 자료를 이용한다. 2009년 시범예행조사는 2010년 총조사를 성공적으로 수행하기 위한 시범조사로 2009년 11월 1일(0시 현재)을 기준시점으로 대전 유성구와 충남 홍성군을 대상으로 조사를 실시하였다. 조사대상 14만여 가구 중 13만여 가구가 조사되었으며, 가구원 기준으로는 34만여 명으로 집계되었다. 이 인원에는 부재 및 불응으로 조사되지 못해 이웃이나 행정자료 등을 활용하여 가구원수, 성별, 나이 등 기본항목만 파악된 자료도 포함되어 있다. 본 연구에서는 연구목적에 맞게 정보를 활용하기 어려운 자료(이름 및 주소 부재, 중복, 데이터 부재 등을 포함)는 사용하지 않을 예정이므로 실제 자료의 수는 이보다 조금 적을 수 있음을 밝혀둔다.

2009년 시범예행조사는 대전 유성구와 충남 홍성군 두 지역만을 조사하여 연계작업 및 결과가 전국자료의 경우와는 차이가 발생할 수 있다. 또한 시범예행조사의 특성상 센서스에 비해 품질이 떨어짐을 감안하면 연계방법을 평가할 때 실제 결과보다 좋지 않을



가능성도 있음을 알려준다. 그리고 종단자료의 경우 정확 연계율이 매우 높기를 기대하나, 조사시점 불일치(조사간 1년의 시점차이가 발생)로 발생하는 조사대상자의 전출, 사망, 해외이민 등의 변화와 연계에 활용하고자 하는 식별정보의 부재 자료(이름, 주소 등)도 존재하여 정확 연계율을 하락시킬 가능성도 있다. 그러나 센서스 종단화 연구의 초기 단계임을 감안하여 2009년 시범예행조사 자료를 시범연계의 기회로 활용하고자 한다.

2. 결정적 자료연계 과정

두 자료의 연계과정에 이용 가능한 변수들을 살펴보면 다음과 같다. 이름, 주소와 같은 개인(가구)의 식별정보를 활용하여 정확 연계율을 높일 수 있으며, 집계와 분석을 위한 조사정보를 부가적으로 사용하여 연계할 수 있다. 이러한 부가정보로는 생년월일(나이), 성별 등이 매우 중요한 항목으로 여겨진다.

본 연구에서는 2009년 시범예행조사 자료를 기준으로 2010년 센서스 자료를 연계한다. 결정적 연계를 위해 이름, 주소, 생년월일(나이), 성별을 블록킹 변수로 사용하였다. 종단 자료의 특성상 정확 연계율을 높이기 위해 결정적 자료연계 과정을 세부적으로 나누어 진행하였다. 먼저 블록킹 변수 4개의 항목이 일치하는 자료를 우선적으로 연계하였고, 주소 및 나이의 기준을 완화해가며 동일인을 찾기 위한 시도를 순차적으로 진행하였다. 결정적 자료연계 과정의 첫 번째 단계는 <표 1-1>에서 알 수 있다. 이 단계에서는 거의 확실히 정확연계가 되는 사람에 대해서 연계하였다.

<표 1-1> 결정적 자료연계 과정의 첫 번째 단계

pass	블록킹 변수	1:1 연계수(비율)
1	이름+주소(시도, 시군구, 읍면동, 번지)+생년월일+성별	103,314(32.8%)
2	이름+주소(시도, 시군구, 읍면동, 번지)+나이(0)+성별	31,694(10.1%)
3	이름+주소(시도, 시군구, 읍면동, 번지)+나이(+1)+성별	6,670(2.1%)
4	이름+주소(시도, 시군구, 읍면동, 번지)+나이(-1)+성별	9,544(3.0%)
5	이름+주소(시도, 시군구, 읍면동, 번지)+나이(+2)+성별	1,421(0.5%)
6	이름+주소(시도, 시군구, 읍면동, 번지)+나이(-2)+성별	2,203(0.7%)
7	이름+주소(시도, 시군구, 읍면동)+생년월일+성별	20,241(6.4%)
8	이름+주소(시도, 시군구)+생년월일+성별	32,992(10.4%)
전체 자료 수(315,227)		208,079(66.0%)

주: 나이(0)는 2009년 자료에 1살을 더한 값이 2010년 자료와 일치함을 의미

pass 1은 이름, 주소(시도, 시군구, 읍면동, 번지), 생년월일, 성별이 모두 일치하는 경우로 자료상으로 100% 일치한다고 볼 수 있으며, 연계수는 103,314명으로 전체 자료 중 32.8%를 차지하고 있다. pass 2는 pass 1에서 생년월일을 사용할 수 없는 경우에 생년월일 대신 나이를 사용한 경우이다. 이 경우에도 pass 1과 같이 완전하게 일치한다고 할 수 있다. pass 3 - pass 6에서는 이름, 주소(시도, 시군구, 읍면동, 번지), 성별은 같으나, 나이에서 ± 1 , ± 2 세 차이가 나는 경우이다. 일반적으로 응답자가 누구냐에 따라 가족의 생년월일 또는 나이를 혼동하는 경우가 발생하기도 한다. 특히 노인의 경우에는 만나이, 세는나이, 주민등록상과 다른 나이를 사용하는 경우가 있어 조사 시 불일치하는 경우가 종종 발생한다. 따라서 1년전 자료와 이름, 주소(시도, 시군구, 읍면동, 번지), 성별은 같으나, 나이에서 ± 2 세의 차이가 있다면 이는 나이에서 발생하는 오류라고 판단해도 무방할 것이다. 본 연구에서도 이러한 경우에 대해서는 정확연계로 판단하고 연계를 진행하였다. 이러한 4개 과정에서 새롭게 연계된 사람의 수는 19,838명으로 전체대비 6.3%로 나타났다.

pass 1에서 pass 6까지는 모든 주소가 일치하는 자료에 한하여 연계하였으나, pass 7은 주소 중에서 번지는 제외하고 시도, 시군구, 읍면동까지만 고려한 경우이다. 그리고 번지 정보가 없는 대신 나이는 사용하지 않고 생년월일이 일치하는 경우에만 적용하였다. pass 7에서는 번지가 없는 경우, 주소 오류인 경우, 읍면동을 벗어나지 않은 범위에서의 이사로 1년 사이에 번지가 변한 경우의 자료를 연계하는 과정을 보여준다. 이는 pass 1보다 연계 조건이 조금 완화된 것으로 100% 일치한다고 단정할 수는 없으나, 1:1로 연계된 자료는 거의 일치한다고 볼 수 있다. 따라서 이를 확인하기 위해 1:1로 연계된 자료의 일부분을 내용검토(clerical review)한 결과 동일인임을 확인할 수 있었다. pass 7에서 연계된 사람의 수는 20,241명으로 전체 자료 중 6.4%를 차지하고 있다. 첫 번째 단계의 마지막인 pass 8은 주소 중에서 번지와 읍면동을 제외하고 시도, 시군구까지만 고려한 경우이다. 이 경우도 pass 7과 마찬가지로 생년월일이 일치하는 경우에만 적용하였으며, 읍면동 및 번지가 없는 경우, 주소 오류, 시군구를 벗어나지 않은 범위에서의 이사로 1년 사이에 읍면동과 번지가 변한 경우의 자료를 연계하는 과정을 보여준다. pass 8에서는 연계 조건이 더 완화되었으나, 1:1로 연계된 자료의 일부분을 내용검토(clerical review)한 결과 동일인임을 확인할 수 있었다.

첫 번째 단계에서는 연계의 정확성이 매우 높은 과정만을 선택하여 연계하였다. pass 1에서 pass 8까지 연계된 총인원은 208,079명으로 전체 자료 중 66%가 연계되었다. 그러나 이 비율은 두 자료의 정확연계 비율이 아니며, 종단자료의 특성을 고려하여 두 번째 단계에서 남은 자료에 대한 정확연계를 시도하여 정확 연계율을 더 높일 필요가 있을 것이다. 따라서 연계기준을 조금 더 완화하여 연계할 경우, 정확연계 비율을 추정할 후 적절한 수준이면 결정적 연계에 포함시키고자 한다. 결정적 자료연계 과정의 두 번째



단계는 <표 1-2>에서 볼 수 있다.

<표 1-2> 결정적 자료연계 과정의 두 번째 단계

pass	블록킹 변수	1:1 연계수 (정확연계추정)
9	이름+주소(시도, 시군구, 읍면동)+나이(0)+성별	8,298 (7,053)
10	이름+주소(시도, 시군구, 읍면동)+나이(+1)+성별	2,334 (1,773)
11	이름+주소(시도, 시군구)+나이(0)+성별	11,662 (8,747)
		22,294 (17,573)

pass 9는 pass 7에서 사용한 블록킹 변수인 생년월일 대신 나이를 사용한 경우이다. 이 경우에는 생년월일을 사용한 과정보다 1:1로 연계된 사람이 동일인이 아닐 가능성은 높아진다. 따라서 실제로 1:1 연계된 사람의 자료 중 약 10%의 표본을 임의로 선택하여 내용검토(clerical review)를 실시하였다. 총 8,298명 중 15%가 동일인이 아닐 가능성이 있는 것으로 나타났다. 가구의 구성원이 변한 경우나 조사 항목별로 차이가 존재하는 경우가 있었다. 하지만, 두 조사의 시점이 1년의 차이가 있으므로 가구의 구성원이 변하거나 조사 항목이 달라질 수도 있을 것이다. 그러나, 본 연구에서는 동일인이 아닐 가능성이 있는 15%에 대해서는 정확연계의 수에서 제외하고, 연계자료로만 활용하기로 결정하였다. 따라서 pass 9에서 추가로 정확 연계된 자료의 수는 7,053명으로 추정되었다. pass 10은 pass 9에서 시범예행조사 기준에서 나이가 1살 더 많은 경우이다. 이전에 설명하였듯이 응답자에 의해 나이 오류가 발생하는 가능성도 있으므로 이를 고려하여 연계하였다. 내용검토 결과 1:1로 연계된 총 2,334명 중 24%가 동일인이 아닐 가능성이 있는 것으로 나타나 pass 10에서 추가로 정확연계된 자료의 수는 1,773명으로 추정되었다. 그리고 나이 기준을 더 완화하여 살펴본 결과 동일인이 아닐 가능성이 있는 사람의 비율이 30% 이상으로 나타나 결정적 자료연계 과정에서 제외하였다. 마지막으로 pass 11에서는 pass 8에서 사용한 블록킹 변수인 생년월일 대신 나이를 사용한 경우이다. 같은 방법으로 내용검토를 한 결과 1:1로 연계된 총 11,662명 중 25%가 동일인이 아닐 가능성이 있는 것으로 나타나 pass 11에서 추가로 정확연계된 자료의 수는 8,747명으로 추정되었다. 그리고 나이 기준을 더 완화하여 살펴본 결과 동일인이 아닐 가능성이 있는 사람의 비율이 50% 이상으로 나타나 결정적 자료연계 과정에서 제외하였다. 결론적으로 본

연구에서의 추정된 정확 연계율은 총 315,227명 중 225,652명으로 71.6%이며, 결정적 연계과정에서 연계된 비율은 230,373명으로 73.1%로 나타났다.

결정적 자료연계 과정에서 보듯이 정확 연계율을 더 높이기 위해서는 이름과 주소 등 블록킹 항목으로 이용할 수 있는 식별정보를 정확하게 작성하는 것이 무엇보다 중요하다. 또한 주소의 변동 등으로 동일인이 아닐 가능성이 있는 사람들에 대하여 실제 확인 작업을 통해 정확 연계율을 높이는 것도 하나의 방법이라 할 수 있다. 하지만 비용과 시간의 제약은 항상 존재하므로 이에 대한 지원이 있어야 할 것이다.

3. 결정적 연계자료 분석 및 평가

결정적 자료연계 과정의 첫 번째 단계 완료 후 연계된 208,079명에 대한 자료와 2009년 시범예행조사 자료의 특성을 비교해 보고자 한다. 앞서 설명한 것처럼 첫 번째 단계 후의 정확 연계율은 66%로 나타났다. 연계율이란 원자료 인원 대비 연계자료 인원의 비율을 의미하며, 연계율이 높을수록 더 많은 사람이 연계되었음을 의미한다. 먼저 원자료와 연계자료의 항목 특성별로 연계율을 확인해 보면 [그림 1-3]과 같다.

먼저 연령*성별 연계율을 보면, 전 연령대 중 20대와 60대 이후의 사람들에게서 연계율이 낮게 나타났다. 노인의 경우 연령대가 높아질수록 연계율이 낮아지는 현상을 보이며, 여성 노인이 남성 노인보다 연계율이 더 낮은 것을 확인할 수 있다. 이는 20대와 30대 초반까지는 대학진학, 군입대(군대를 가는 남성이 여성보다 확연히 연계율이 낮음), 사회진출(사회진출 빈도가 높은 남성이 여성보다 연계율이 낮음), 결혼 등으로 이동이 많은 시기이므로 다른 연령대에 비해 연계율이 낮은 것으로 보인다. 또한 60대 이후의 사람들이 연계율이 낮은 이유는 연계를 위해 사용한 블록킹 변수인 생년월일과 나이에서 젊은 연령대에 비해 응답 시 변화가 심한 것으로 판단된다. 반면에 어린 학생들과 중년의 경우에는 전체 평균 연계율보다 높게 나타나는데 타 연령대보다 이동요인이 상대적으로 적기 때문인 것으로 보인다. 특히, 나이가 어릴수록 생년월일과 나이를 혼동없이 쓰고 있는 것으로 생각할 수 있을 것이다.

혼인상태별 연계율은 사별이 50.7%로 가장 적게 연계되고, 그 다음으로 이혼, 미혼, 배우자 있음 순이다. 혼인상태는 연령과 높은 연관성이 있는데 사별이 많은 60대 이상과 미혼이 많은 20대~30대 초반의 사람들이 연계율이 낮은 것과 일치한다고 볼 수 있다. 교육정도별로 살펴보면, 대학원졸 이상과 초졸 이하의 연계율이 높은 반면, 중졸과 고졸은 평균 연계율에 미치지 못하고 있다. 1-12세 정도의 초등학교를 졸업하지 못한 아이들이 연계율이 높고, 고학력의 사람들일수록 이동 없이 안정적인 정착 생활을 하고 있는 것으로 보인다. 이러한 현상은 직업에서도 확인할 수 있는데, 고학력이 많은 전문직과 사무직이



[그림 1-3] 첫 번째 단계 후 주요 특성별 결정적 연계율

농어업과 기능노무 관련 직업보다 연계율이 10%p 이상 높은 것을 볼 수 있다.

다음으로 점유형태별로 살펴보면 자기집이나 전세가 월세나 무상임대보다 연계율이 높은 것도 직업 및 학력과 연관해서 알 수 있을 것이다. 또한 자기집의 경우에는 이사 등 이동이 타 점유형태보다 낮기 때문인 것으로 생각할 수 있다. 전세, 보증금 있는 월세, 보증금 없는 월세 순으로 연계율이 떨어지는 것도 인구이동 측면에서 타당한 결과로 보인다. 거처의 종류별로는 아파트에 거주하는 경우 다른 거처의 종류보다 연계율이 훨씬 높게 나타났다. 연립·다세대주택이나 기타 거처(오피스텔, 고시원 등)에 거주하는 사람들의 연계율이 상대적으로 낮은 이유는 잦은 전출입으로 인한 이동요인도 있지만, 불분명한 주소(주소오류) 등으로 인해 연계되지 못했기 때문일 가능성도 크다. 또한 단독주택의 연계율이 낮은 것은 홍성군에 거주하는 노인들의 비율이 많은 것으로 판단된다. 마지막으로 지역별(유성구, 홍성군) 연계율을 살펴보면, 유성구가 홍성군보다 10%p 높게 나타나는데 이는 홍성군에 연계율이 낮은 노인들의 비율이 높다는 것을 알 수 있다.

결정적 자료연계 과정의 두 번째 단계 완료 후 연계된 사람은 230,373명으로 연계된 비율은 73.1%이며, 첫 번째 단계보다 연계비율은 7.1%p가 올랐다. 하지만, 원자료 대비 연계율이 70%대 수준이므로 원자료와 연계자료의 항목 특성별 연계율 차이는 일정하게 유지되고 있음을 볼 수 있다. [그림 1-4]를 참조하기 바란다.

앞에서 설명한 결과는 2009년 시범예행조사를 기준으로 연계된 자료와의 비교이므로 전체 센서스 자료로 작성된 결과와는 연계율의 차이가 발생할 수 있음을 알려준다. 하지만 본 연구에서 보여주고자 하는 것은 원자료와 일정한 비율로만 연계된 자료와는 항목별로 특정 그룹에서 연계율 차이가 발생할 수 있다는 것이다. 이러한 결과는 2005년 센서스가 아닌 2009년 시범예행조사를 이용하더라도 변화가 없을 것으로 확신한다.

이처럼 특정 그룹에서 연계율이 상대적으로 떨어진다면, 연계된 자료의 분포는 원자료와 다르게 왜곡될 수 있다. <표 1-3>은 2009년 시범예행조사를 기준으로 결정적 연계자료와의 분포 동질성을 검정하기 위해 주요 특성별 카이제곱검정(Chi-square Test)을 실시한 결과이다. 카이제곱검정 결과를 먼저 살펴보면 제시된 특성들에 대해 p-값이 모두 유의수준 0.01보다 작다. 즉, 각 특성별로 연계자료의 분포와 원자료의 분포는 동일한 분포를 갖는 모집단으로부터 추출되었다고 볼 수 없으며, 연계자료의 분포와 원자료의 분포는 통계적으로 유의미한 차이가 있다고 볼 수 있다.

원자료와 연계자료의 성별 분포를 비교해보면, 연계자료에서 남자의 비율은 50.9%에서 50.1%로 줄고, 반면에 여자의 비율은 49.1%에서 49.9%로 커졌다. 이는 원자료에서는 남녀간의 비율이 약 2%p 차이가 났으나, 연계자료로만 분석한다면 남녀 간의 비율차이는 거의 나타나지 않는다고 볼 수 있다. 즉, 연계자료에서 남자의 연계율이 여자보다 낮기 때문에 분포의 차이를 발생시켰다고 할 수 있다. 연령별 분포에서는 두 자료 간의



〈표 1-3〉 첫 번째 단계 후 원자료와 결정적 연계자료의 분포 비교

특성	범주	원자료	연계자료	Chi-square	P-value
성별	남자	50.9	50.1	27.319	<.0001
	여자	49.1	49.9		
연령	1~ 9세	10.4	12.1	2510.288	<.0001
	10~19세	15.2	17.6		
	20~29세	14.5	11.1		
	30~39세	16.4	16.6		
	40~49세	17.4	18.6		
	50~59세	12.1	12.0		
	60~69세	6.5	6.0		
	70~79세	5.3	4.3		
	80세이상	2.3	1.6		
혼인상태	미혼	31.2	27.8	1009.852	<.0001
	배우자있음	60.9	65.6		
	사별	6.0	4.8		
	이혼	1.9	1.7		
교육정도	초졸 이하	32.0	33.7	759.740	<.0001
	중졸	8.9	8.5		
	고졸	26.2	23.2		
	대졸	26.2	26.9		
	대학원졸	6.8	7.7		
직업	전문관리	39.5	42.2	49.487	<.0001
	사무	13.2	14.5		
	서비스판매	15.2	14.4		
	농어업	13.8	12.3		
	기능노무	16.4	15.5		
거처의 종류	단독주택	22.6	18.1	2195.815	<.0001
	아파트	66.1	72.5		
	연립주택	4.4	3.6		
	다세대주택	2.7	2.4		
	기타	4.2	3.5		
점유형태	주거 전용	자기 집	59.5	1360.726	<.0001
		전세	18.6		
		보증금 있는 월세	12.5		
		보증금 없는 월세	3.4		
		무상	3.1		
	영업겸용	2.8	2.5		

〈표 1-4〉 두 번째 단계 후 원자료와 결정적 연계자료의 분포 비교

특성	범주		원자료	연계자료	Chi-square	P-value
성별	남자		50.9	49.9	45.432	<.0001
	여자		49.1	50.1		
연령	1~ 9세		10.4	11.6	1735.758	<.0001
	10~29세		15.2	17.0		
	20~29세		14.5	11.5		
	30~39세		16.4	16.8		
	40~49세		17.4	18.5		
	50~59세		12.1	12.1		
	60~69세		6.5	6.2		
	70~79세		5.3	4.6		
	80세이상		2.3	1.7		
혼인상태	미혼		31.2	28.0	865.325	<.0001
	배우자있음		60.9	65.2		
	사별		6.0	5.1		
	이혼		1.9	1.7		
교육정도	초졸 이하		32.0	33.2	445.491	<.0001
	중졸		8.9	8.6		
	고졸		26.2	23.9		
	대졸		26.2	26.8		
	대학원졸		6.8	7.4		
직업	전문관리		39.5	40.9	22.893	<.0001
	사무		13.2	14.3		
	서비스판매		15.2	14.6		
	농어업		13.8	12.9		
	기능노무		16.4	15.8		
거처의 종류	단독주택		22.6	19.6	1223.487	<.0001
	아파트		66.1	70.8		
	연립주택		4.4	3.7		
	다세대주택		2.7	2.4		
	기타		4.2	3.5		
점유형태	주거 전용	자기 집	59.5	62.4	992.742	<.0001
		전세	18.6	18.8		
		보증금 있는 월세	12.5	11.2		
		보증금 없는 월세	3.4	2.4		
		무상	3.1	2.6		
	영업겸용	2.8	2.5			



[그림 1-4] 두 번째 단계 후 주요 특성별 결정적 연계율

차이가 더 확연하게 나타난다. 1~9세와 10대는 연계자료에서의 비율이 약 2%p 정도 올랐으나, 반면에 20대는 3.4%p 낮아진 것을 볼 수 있다. 다른 연령대에서도 조금씩 차이를 나타내고 있다. 혼인상태의 경우에도 미혼의 비율은 연계자료에서 3.4%p가 떨어졌으나, 배우자가 있는 사람은 4.7%p나 오른 것을 알 수 있다. 교육정도에서도 초졸 이하의 비율이 연계자료에서 1.7%p 올랐으며, 중졸 및 고졸의 비율은 낮아진 것을 알 수 있다. 직업의 경우 전무관리 및 사무직에서 연계자료의 비율이 올랐으며, 서비스 및 농어업에서는 다소 낮아졌다. 거처의 종류는 아파트에서 비율이 6.4%p로 많이 높아졌으며, 점유형태의 비율은 자기집에서 3.5%p가 높아졌다. 이 결과들은 이전에 살펴보았던 연계율과 같은 의미로 해석할 수 있을 것이다.

결정적 자료연계 과정의 두 번째 단계 완료 후의 자료를 사용하여 주요 특성별 카이제곱검정을 실시한 결과는 <표 1-4>에 주어져 있다. 검정 결과를 살펴보면 제시된 특성들에 대해 p-값이 모두 유의수준 0.01보다 작은 것은 첫 번째 단계 후와 변화가 없으나, 카이제곱 값이 감소한 것을 볼 수 있다. 이는 연계자료의 분포와 원자료의 분포는 통계적으로 유의미한 차이가 있지만 그 차이의 폭은 많은 부분 줄어들었다는 것을 의미한다. 하지만 현재의 연계율에서는 연계된 자료만을 이용하기에는 부적합한 것으로 판단된다.

지금까지 살펴본 바와 같이 70% 수준의 연계율에서는 특정 그룹에서의 연계자료가 원자료와는 다른 분포의 형태를 갖게 된다. 따라서 확률적 연계과정을 추가함으로써 원자료와 연계자료의 분포를 줄이려는 시도가 필요할 것이며, 제4절에서는 확률적 연계과정과 연계결과를 제시하고자 한다.

제4절 확률적 자료연계

본 절에서는 이름, 주소와 같은 식별정보 대신 혼인상태, 교육정도와 같은 개인정보를 활용한 확률적 자료연계를 실시한다. 이전의 결정적 자료연계 단계에서 정확연계 된 자료는 2009년 시범예행자료 기준으로 230,373명인 73.1%이나, 이 중 약 90%는 전수조사 자료이다. 이 때 확률적 자료연계를 위해 사용하는 항목들은 대부분 전수항목으로 구성되므로 확률적 연계방법을 평가하기 위한 항목이 마땅치 않다. 따라서 본 연구에서는 확률적 연계방법의 평가를 위해 정확연계 된 자료 중 표본 항목으로 선정된 사람들만의 자료를 이용하여 실험하고자 한다. 표본 항목을 이용하여 평가할 경우 자료의 수가 9,645개로 줄어들어 특정 셀(도수가 적은 셀)에서는 오차가 크게 나타날 수도 있을 것이다. 또한 2009년 시범예행조사는 2010년 센서스를 조사하기 위한 연습으로 자료의 품질은



다소 떨어져(내용검토 부족으로 인한 오류, 연계 시 무응답으로 인한 항목 사용 제한, 지역적 한계 등) 평가할 때 실제보다 좋지 못한 결과가 나올 가능성도 있다는 것을 알려 준다. 마지막으로 평가 후의 결과를 바탕으로 결정적 자료연계 단계에서 연계되지 못한 나머지 26.9% 자료들을 확률적 연계방법에 의해 연계시킨다. 따라서 동일인이 아닐지라도 주요한 개인정보를 사용하여 유사한 특성을 갖는 자료와 연계함으로써 결정적 연계과정에서 미연계된 자료를 보완하고, 원자료의 분포와 유사한 형태의 연계자료를 구축하고자 한다.

1. 확률적 자료연계 과정

확률적 자료연계를 실시할 경우 사용하는 연계변수의 선택도 중요하다. 변수들 중에는 일반적으로 시간이 흘러도 값이 변하지 않는 성별과 같은 정보가 있고, 시간에 따라 변할 수 있는 혼인상태, 교육정도와 같은 정보도 있다. 2009년 시범예행조사와 2010년 센서스는 1년이라는 시점의 차이가 있으므로 동일인 여부를 판단하기 위한 연계변수는 시간이 지나도 값이 크게 변하지 않는 변수를 선택하는 것이 적합할 것이다. 이를 위해 결정적 방법으로 정확 연계된 자료를 이용해 공통항목에 대한 값의 일치여부를 <표 1-5>에서 확인하였다. 이때 시범예행조사의 특성상 항목의 무응답이 존재하므로 이러한 자료들은 제외하고 비교 가능한 자료만을 선택하여 일치율을 구하였다.

<표 1-5> 결정적 연계자료의 공통항목 일치여부

공통항목	일치 여부	빈도	비율	공통항목	일치 여부	빈도	비율
성별	일치	230,373	100.0	가구주 관계	일치	203,576	92.2
	미일치	0	0.0		미일치	17,342	7.8
혼인상태	일치	172,701	96.5	거처종류	일치	194,737	91.9
	미일치	6,251	3.5		미일치	17,215	8.1
교육정도	일치	196,794	85.5	점유형태	일치	184,174	83.1
	미일치	33,289	14.5		미일치	37,359	16.9
연령	-1	2,390	1.0	연령	1	196,396	85.3
	0	15,906	6.9		±2	15,681	6.8

성별의 경우에는 결정적 자료연계 과정에서 블록킹 변수로 이용하였으므로 연계된 자료는 모두 일치한다. 혼인상태 96.5%, 가구주의 관계 92.2%, 거처종류 91.9%로 매우 높은 일치율을 보이고 있다. 이는 두 자료의 시점차이가 1년으로 매우 짧음에 기인한다.

그리고 교육정도는 85.5%, 점유형태는 83.1%로 다소 낮아지는데 1년 사이에도 입학과 졸업이 많이 발생하므로 일치율에서 차이가 나며, 점유형태 역시 이사 등으로 인해 다소 변화가 발생하는 것으로 판단된다. 나이는 자료가 1년의 시점이 있으므로 2010년 나이에서 2009년 나이를 뺀 연령 차이를 구해보면, 일치하는 경우인 +1 차이는 85.3%로 나타났다. 그 외의 경우는 조사오류로 판단되며 이전에 설명하였듯이 만나이, 세는나이, 주민등록상과 다른 나이를 사용하는 경우가 있어 조사 시 불일치하는 경우가 종종 발생하고 있다. 각 항목의 일치율을 검토한 결과 모든 항목에서 80% 이상의 일치율을 보이며 이 항목들을 확률적 연계에 사용해도 무리가 없을 것으로 판단된다. 그리고 성별, 나이, 지역은 블록킹 변수로 활용할 것이다.

확률적 연계방법에 대한 평가를 위해서 표본 항목으로 선정된 사람들만의 자료(9,645명)를 이용한다. 그 이유는 이전에 설명하였듯이 전수항목은 대부분 확률적 연계과정에서 연계변수로 사용하였으므로 평가를 위한 자료로는 적절하지 못한 것으로 판단된다. 2009년 시범예행조사 기준(315,227명)으로 약 73.1%의 자료가 결정적 연계과정에서 연계되었으므로 실제 자료의 경우 약 26.9%에 해당하는 자료는 확률적 연계에 의해서 2010년 센서스 자료와 연계되어야 한다. 따라서 이러한 상황을 고려하여 확률적 연계방법에 대한 평가는 정확연계 자료 중에서 표본 항목으로 선정된 자료를 모집단으로 가정하고, 모집단에서 임의로 25% 자료를 확률적 방법에 의해 연계하여 실제 모집단 자료와 25%가 확률적으로 연계된 자료 간의 비교를 통해 실시할 것이다. 이 때 모집단 자료와 확률적 연계 자료 간의 전체적인 분포, 방향성 등의 차이를 고려하여 연계방법의 가능성을 검토할 것이다. 확률적 연계방법으로는 해외사례에서 설명하였듯이 센서스 종단 자료 구축시 호주와 뉴질랜드에서 사용하고 있는 방법을 적용하고자 하며 자세한 내용은 다음과 같다.

먼저 두 자료의 연계를 위해서 변수 가중값(field weights)과 연계 가중값(linkage weights)을 활용하였다. 변수 가중값은 각각의 자료 쪽으로부터 특정변수에 대한 값을 비교하여 구한 각 변수에 대한 중요도를 의미하는 것으로 이 값은 다음에 주어진 m-확률과 u-확률을 이용하여 구할 수 있다.

- $m\text{-확률} = P(\text{fields agree} \mid \text{records belong to the same entity})$
- $u\text{-확률} = P(\text{fields agree} \mid \text{records belong to different entity})$

m-확률은 두 자료가 동일인일 경우 각 변수의 값이 일치할 확률이고, u-확률은 두 자료가 동일인이 아닐 경우 각 변수의 값이 일치할 확률을 의미한다. 예를 들면 성별의 경우 m-확률은 1, u-확률은 0.5 근처의 값을 가지며, 주소의 경우 동일인이 아닌데도 불구하고 주소가 동일한



경우를 나타내는 u -확률은 0에 가까운 값을 가지게 된다. 일반적으로 m -확률과 u -확률을 구하기 위해서 다양한 방법을 이용할 수 있다. 센서스 연계에서는 센서스 조사 후에 실시되는 사후조사 자료를 활용할 수 있으며, 기존에 연구되어져 있는 값을 사용한다든지 EM 알고리즘을 이용하여 구한 추정값을 활용하기도 한다. 본 연구에서는 2009년 시범예행조사와 2010년 센서스의 결정적 연계된 자료의 변수별 일치비율을 참고하여 m -확률을 구하였으며, 이는 앞에서 설명한 사후조사에 의해 구한 방법과 유사한 것으로 해석할 수 있다. 또한 u -확률의 경우 호주는 동일인이 아닌 사람의 자료를 임의로 추출(반복실험)하여 변수의 값이 일치할 확률을 구하였으나, 우리는 뉴질랜드에서 손쉽게 적용하고 있는 1을 항목의 범주수로 나눈 값을 사용하였다. m -확률과 u -확률을 구하는 방법은 국가마다 조금의 차이는 있으나, 구한 값의 확률적 의미는 동일한 것임을 밝혀둔다.

m -확률과 u -확률을 구한 후에는 두 자료의 항목이 일치한 경우에 사용하는 일치 변수가중값과 불일치한 경우에 사용하는 불일치 변수가중값을 다음의 정의에 의해서 계산한다.

- 일치 변수가중값 = $\log_2(m/u)$
- 불일치 변수가중값 = $\log_2(1 - m/1 - u)$

위의 식을 살펴보면 m -확률은 u -확률보다 큰 값을 가지므로 일치 변수가중값은 항상 (+)값을 갖게 되며, m -확률은 클수록 그리고 u -확률은 작을수록 더 큰 (+)값을 갖게 된다. 이 값이 클수록 두 자료를 연계할 때 더 중요한 역할을 하는 변수가 됨을 알 수 있다. 반면에 불일치 변수가중값은 항상 (-)값으로 구해진다.

<표 1-6>에는 시범예행조사 연계에 사용되는 항목들과 이전에 설명한 방법으로 구한 m -확률, u -확률, 일치 및 불일치 변수가중값이 주어져 있다. 성별, 연령, 지역 항목은 블록킹 변수로 사용할 예정이므로 따로 계산하지 않았다. 블록킹 기법은 자료 연계 시 많은 비교횟수가 요구될 때 효율적으로 사용할 수 있는 방법이다. 예를 들어 성별, 연령, 지역을 블록킹 변수로 사용한다면 두 자료에서 같은 성별, 연령, 지역을 가진 사람들만 비교하므로 비교대상이 많이 감소되어 시간의 절약을 가져온다. 블록킹 기법은 변수의 값이 잘못된 경우나 결측값인 경우에는 비교가 제한된다는 단점도 있으나, 자료연계를 위해서는 꼭 필요한 기법이라고 판단된다. 그리고 해당 블록킹 변수 내에 빈 셀이 발생할 경우에는 인접 셀과의 병합으로 조정하는 과정이 필요할 것이다.

다음으로 작성된 변수가중값을 근거로 연계 가중값을 계산하게 된다. 연계 가중값은 각각의 자료 짝에 대해 모든 변수 값의 일치여부를 판단하여 일치하는 변수는 일치 변수가중값 적용, 불일치하는 변수는 불일치 변수가중값 적용하는데 모든 변수를 비교한 후 변수가중값을 합산하여 구할 수 있다. 따라서 연계 가중값이 클수록 유사한 사람과

〈표 1-6〉 확률적 연계 변수에 대한 m-확률, u-확률, 일치 및 불일치 변수가중값

항목(범주 수)		역할	m-확률	u-확률	일치 변수가중값	불일치 변수가중값
성별	2	블록킹 변수				
연령 ¹⁾	17	블록킹 변수				
지역 ²⁾	2	블록킹 변수				
혼인상태 ³⁾	4	연계 변수	0.965	0.250	1.95	-4.42
교육정도 ⁴⁾	5	연계 변수	0.855	0.200	2.10	-2.46
가구주관계 ⁵⁾	4	연계 변수	0.922	0.250	1.88	-3.27
거처종류 ⁶⁾	5	연계 변수	0.919	0.200	2.20	-3.30
점유형태 ⁷⁾	6	연계 변수	0.831	0.167	2.32	-2.30

주: 1) 연령(0~4, 5~9, 10~14, ..., 70~74, 75~79, 80세 이상)

2) 지역(대전광역시 유성구, 충청남도 홍성군)

3) 혼인상태(미혼, 배우자있음, 사별, 이혼)

4) 교육정도(초졸 이하, 중졸, 고졸, 대졸, 대학원 이상)

5) 가구주관계(본인, 배우자, 자녀, 기타)

6) 거처종류(단독주택, 아파트, 연립주택, 다세대주택, 기타)

7) 점유형태(자기집, 전세, 보증금 있는 월세, 사글세, 무상, 영업겸용)

연계될 가능성이 높아진다. 본 연구에서는 가장 연계 가중값이 높은 사람과 연계를 하되, 연계 가중값의 최고점에서 동점자가 발생한 경우엔 랜덤하게 한 명을 선택하여 연계하는 방식으로 확률적 연계를 진행하였다. <표 1-7>에 주어진 하나의 예제를 통해 연계과정을 자세하게 설명하고자 한다.

〈표 1-7〉 확률적 연계과정 설명을 위한 예제

항목(변수)	2009년 시범예행조사 (수용자파일)	2010년 인구주택총조사(제공자 파일)			
		제공자1	제공자2	제공자3	제공자4
성별	여	여	여	여	여
연령	30-34	30-34	30-34	30-34	30-34
지역	대전 유성구	대전 유성구	대전 유성구	대전 유성구	대전 유성구
혼인상태	배우자있음	미혼	배우자있음	배우자있음	배우자있음
교육정도	대졸	고졸	대학원	대졸	대졸
가구주관계	배우자	자녀	배우자	배우자	배우자
거처종류	아파트	단독주택	아파트	아파트	아파트
점유형태	전세	자기집	자기집	보증금(월세)	전세



2009년 수용자 파일을 연계하기 위하여 2010년 제공자 파일과 비교를 실시한다. 먼저 블록킹 변수로 성별, 연령, 지역을 선택하였으므로 같은 성별, 연령, 지역을 가진 사람들로만 비교를 실시하면 된다. 예를 들어 이와 같은 블록에 해당하는 제공자가 4명 있다고 하면 각각의 자료에 대해 연계 가중값을 구할 수 있다. 첫 번째 제공자는 모든 항목이 수용자와 다르므로 불일치 변수가중값으로 합산되며, 연계 가중값은 다음과 같다.

$$\text{연계가중값(1)} = (-4.42) + (-2.46) + (-3.27) + (-3.30) + (-2.30) = -15.75$$

제공자2, 제공자3, 제공자4에 대해서도 일치하는 변수는 일치 변수가중값, 불일치하는 변수는 불일치 변수가중값을 합산하여 연계 가중값을 구하면 다음과 같다.

$$\text{연계가중값(2)} = (1.95) + (-2.46) + (1.88) + (2.20) + (-2.30) = 1.27$$

$$\text{연계가중값(3)} = (1.95) + (2.10) + (1.88) + (2.20) + (-2.30) = 5.83$$

$$\text{연계가중값(4)} = (1.95) + (2.10) + (1.88) + (2.20) + (2.32) = 10.45$$

위의 결과를 보면 제공자4의 경우에는 모든 항목에서 수용자와 일치하며 연계 가중값은 10.45로 가장 높게 나타나게 된다. 따라서 수용자는 확률적 연계를 통해 제공자4와 연계가 됨을 알 수 있다. 이 때 가장 높은 연계 가중값이 2명 이상일 경우에는 랜덤하게 1명을 선택하여 연계를 실시하였다. 마지막으로 임의로 선택된 25%의 수용자 모두가 연계될 때까지 이러한 과정을 반복하여 수행하였다.

2. 확률적 연계자료 분석 및 평가

확률적 연계자료의 평가는 결정적 연계자료 중 표본항목으로 선정된 사람의 자료(A)와 그 중 25%에 해당하는 사람은 확률적 자료연계에 의해 연계하고 나머지 75%는 결정적 연계자료로 구성된 자료(B)에 대해 연계결과를 비교한다. 연계자료에서 경제활동상태, 직업, 종사상지위 항목을 교차분석하고 그 결과를 비교함으로써 확률적 연계자료를 평가하고자 한다.

<표 1-8>은 2009년과 2010년의 경제활동상태 항목을 교차분석 하였다. 결정적 연계자료(A) 결과를 보면, 2009년에 일을 하고 있었던 사람 중 2010년에도 일을 하고 있는 사람은 비율은 78.8%이며, 일을 하지 않게 된 사람의 비율은 21.2%로 나타났다. 반면에 2009년에 일을 하지 않던 사람 중 2010년에 일을 하게 된 사람의 비율은 12.7%이며, 계속적으로 일을 하지 않고 있는 사람의 비율은 87.3%로 나타났다. 이 자료에 대하여

25%의 확률적 연계 후의 자료(B)와 비교한 차이를 살펴보면 2009년부터 2010년까지 계속해서 일을 하고 있었으나, 확률적 연계 후 2010년에 일을 하지 않는 것으로 나타난 비율이 3.9%p로 다소 늘어났다. 결과상으로는 조금의 차이가 발생하였으나, 전체적인 분석 결과의 패턴과 방향성에 대해서 큰 영향을 미칠 정도의 차이는 아닌 것으로 판단된다.

〈표 1-8〉 2009년과 2010년의 경제활동상태의 교차분석

(단위: %, %p)

분석 자료	경제활동 상태('10) 경제활동 상태('09)	일하였음	일하지 않았음
결정적 연계자료 (A)	일하였음	78.8	21.2
	일하지 않았음	12.7	87.3
결정적 +확률적 연계자료 (B)	일하였음	74.9	25.1
	일하지 않았음	12.3	87.7
차이 (C=B-A)	일하였음	-3.9	3.9
	일하지 않았음	-0.4	0.4

다음은 2009년과 2010년의 종사상지위 항목을 교차분석 한 결과가 <표 1-9>에 주어져있다. 2009년 임금근로자 중 2010년 계속적으로 임금근로자인 사람의 비율은 73.6%이며, 자영업으로 지위가 변한 사람은 4.0%, 무급가족종사자로 지위가 변한 사람은 0.4%, 일을 하지 않게 된 사람의 비율은 21.9%로 나타났다. 2009년 자영업업을 하던 사람이 2010년에도 자영업업을 하고 있는 경우는 60.2%밖에 되지 않으며, 임금근로자로 15.3%, 무급가족종사자로 5.7%, 일을 하지 않게 된 사람으로 18.9%가 지위가 변하였다. 그리고 2009년 일을 하지 않았던 사람 중 2010년에도 87.1%가 일을 하지 않고 있으며, 임금근로자로 지위가 변한 사람은 10.2%로 나타났다. 이 결과와 확률적 연계 후의 자료(B)와 비교한 차이를 살펴보면 임금근로자에서 임금근로자로의 비율이 4.0%p 감소, 자영업자에서 자영업자로의 비율이 4.3%p 감소, 무급가족종사자에서 무급가족종사자로의 비율이 5.8%p 감소한 것을 확인할 수 있었다. 이러한 이유 중에는 연계변수가 부족하여 연계의 정확도가 떨어질 수 있고, 내검요류 및 무응답 등으로 연계변수의 효과가 감소할 수도 있을 것이다. 그리고 최고 연계가중값이 동점인 경우 비율(도수)이 높은 부분으로 연계될 가능성이



높아질 수도 있다(일하지 않았음의 비율이 종사상지위 각각의 항목 비율보다 큼).

<표 1-9> 2009년과 2010년의 종사상지위의 교차분석

(단위: %, %p)

분석 자료	종사상 지위 (10) 종사상 지위(09)	임금 근로자	자영업자	무급 가족 종사자	일하지 않았음
결정적 연계자료 (A)	임금 근로자	73.6	4.0	0.4	21.9
	자영업자	15.3	60.2	5.7	18.9
	무급 가족 종사자	8.3	12.1	56.8	22.8
	일하지 않았음	10.2	1.7	0.9	87.1
결정적 +확률적 연계자료 (B)	임금 근로자	69.6	3.8	0.4	26.2
	자영업자	18.5	55.9	4.6	20.9
	무급 가족 종사자	10.2	12.1	51.0	26.7
	일하지 않았음	9.8	1.6	1.0	87.6
차이 (C=B-A)	임금 근로자	-4.0	-0.2	0.0	4.3
	자영업자	3.3	-4.3	-1.0	2.1
	무급 가족 종사자	1.9	0.0	-5.8	3.9
	일하지 않았음	-0.5	0.0	0.0	0.5

다음은 2009년과 2010년의 직업 항목을 교차분석 한 결과가 <표 1-10>에 주어져있다. 2009년 전문관리자 중 2010년 계속적으로 전문관리자인 사람의 비율은 72.2%이며, 사무자로 지위가 변한 사람은 4.4%, 서비스판매자로 지위가 변한 사람은 1.1%, 일을 하지 않게 된 사람의 비율은 20.6%로 나타났다. 2009년 사무를 하던 사람이 2010년에도 사무를 하고 있는 경우는 48.5%밖에 되지 않으며, 전문관리자로 17.5%, 서비스판매자로 6.1%, 일을 하지 않게 된 사람으로 20.8%가 직업이 변하였다. 그리고 2009년 농어업을 하던 사람은 2010년 79.6%가 그대로 유지하고 있으며, 다른 직업으로의 변화는 적으며 11.7%가 일을 그만하게 된 것으로 확인되었다. 이 결과와 확률적 연계 후의 자료(B)와 비교한 차이를 살펴보면 종사상지위의 경우와 유사하게 전문관리자에서 전문관리자로의 비율이 5.4%p 감소, 사무자에서 사무자로의 비율이 4.3%p 감소, 서비스판매자에서 서비스판매자로의 비율이 4.4%p 감소, 기능노무자에서 기능노무자로의 비율이 8.1%p 감소한 것을 확인할 수 있었다. 이 문제 또한 종사상지위와 같은 맥락에서 해결되어야 할 것이다.

지금까지 살펴본 교차분석 결과들을 고려하면, 결정적 연계자료 중 25%를 확률적 연계자료로 대체한 경우 실제 자료와 특정부분에서의 차이는 발생하지만, 자료의 전체적인 패턴과 방향성에는 큰 변화가 발생하지 않는 것으로 판단된다. 하지만, 차이가 발생하는 원인 중 하나는 시범예행조사의 품질에 기인한 것으로 추측되므로 향후 2010년 센서스를 기준으로 2015년 센서스 자료와 연계할 때 다시 확인되어야 할 것이다.

〈표 1-10〉 2009년과 2010년의 직업의 교차분석

(단위: %, %p)

분석 자료	직업 (’09) \ 직업 (’10)	전문 관리	사무	서비스 판매	농어업	기능 노무	일하지 않았음
결정적 연계자료 (A)	전문관리	72.2	4.4	1.1	0.1	1.6	20.6
	사무	17.5	48.5	6.1	0.6	6.4	20.8
	서비스판매	5.9	4.5	52.2	1.8	5.6	30.0
	농어업	0.3	0.0	3.2	79.6	5.2	11.7
	기능노무	5.2	6.6	9.5	2.9	58.1	17.6
	일하지 않았음	6.8	1.4	2.1	0.9	1.5	87.2
결정적 +확률적 연계자료 (B)	전문관리	66.8	4.6	1.4	0.2	1.8	25.2
	사무	18.1	44.2	5.0	1.2	7.0	24.6
	서비스판매	7.1	4.7	47.8	2.1	5.6	32.6
	농어업	0.6	0.3	2.9	75.5	7.0	13.7
	기능노무	6.6	8.1	9.8	4.3	50.0	21.1
	일하지 않았음	5.9	1.5	2.6	1.0	1.4	87.6
차이 (C=B-A)	전문관리	-5.4	0.3	0.3	0.1	0.1	4.6
	사무	0.6	-4.3	-1.2	0.6	0.6	3.8
	서비스판매	1.2	0.3	-4.4	0.3	0.0	2.7
	농어업	0.3	0.3	-0.3	-4.1	1.7	2.0
	기능노무	1.4	1.4	0.3	1.4	-8.1	3.5
	일하지 않았음	-0.9	0.0	0.5	0.1	-0.1	0.4

최종적인 연계자료는 3절에서 생성한 결정적 연계자료(230,373명)와 결정적 연계되지 못한 자료(84,854명)를 결합하여 작성하면 된다. 따라서 최종 연계자료에는 2009년의 모든 자료가 포함되기 때문에 이전에 언급했던 연계자료의 분포 왜곡은 발생하지 않을 것이다. 그리고 확률적 연계자료를 결합함으로써 교차분석 시 언급했던 차이는 발생하나, 결합된 정보를 사용하기에는 큰 문제가 없을 것으로 생각된다.



제5절 결 론

본 연구는 2009년 시범예행조사와 2010년 센서스를 연계하여 센서스 종단자료를 시범적으로 구축해봄으로써 센서스 자료의 분석 및 활용가치를 제고하고자 시작되었다. 해외사례에서 언급하였듯이 호주, 영국, 뉴질랜드에서는 이미 센서스 종단자료를 구축하고 있으며, 센서스 자료의 가치와 활용성 강화를 위해 많은 노력을 하고 있다. 본 연구에서는 2010년과 2015년 센서스 종단자료 구축을 위한 시범적 연계로 향후 자료 구축을 위해 연구결과를 활용할 예정이다.

본 연구는 결정적 연계 후 연계되지 못한 자료에 대해서는 확률적 연계를 실시하는 방법으로 진행되었다. 결정적 연계를 위해 이름, 주소, 생년월일(나이), 성별을 블록킹 변수로 사용하였고, 정확 연계율을 높이기 위해 2단계로 나누어 진행하였다. 첫 번째 단계에서는 총 315,227명 중 208,097명인 66%가 정확 연계되었다. 추가로 실시한 두 번째 단계에서는 22,294명이 더 연계되어 총 정확 연계된 자료(추정포함)는 230,373명(73.1%)으로 결정되었다. 하지만 특정한 특성그룹에서 연계율이 상대적으로 떨어지면서 연계된 자료의 분포가 원자료의 분포와 달라 현재의 연계율에서는 연계된 자료만을 이용하기에는 부적합한 것으로 판단되었다. 따라서 확률적 연계과정을 추가함으로써 원자료와 연계 자료의 분포를 줄이려는 시도가 필요하였다. 이에 따라 이름, 주소와 같은 식별정보가 아닌 혼인상태, 교육정도와 같은 개인정보를 이용한 확률적 자료연계를 실시하고, 결정적으로 연계된 자료를 기준으로 삼아 두 방법을 비교·평가하고자 하였다.

확률적 연계방법은 센서스 종단자료 구축 시 호주와 뉴질랜드에서 사용하는 방법을 적용하였다. 확률적 연계방법에 대한 평가는 결정적 연계자료 중 표본항목으로 선정된 자료(9,645명)를 모집단으로 가정하고, 모집단에서 임의로 25% 자료(2,411명)를 확률적 방법에 의해 연계하여 실제 모집단 자료와 25%가 확률적으로 연계된 자료 간의 교차분석 결과를 비교하였다. 그 결과 경제활동상태 항목의 경우 확률적 연계 후에 2010년에 일을 하지 않는 것으로 나타난 비율이 3.9%p로 다소 늘어났으나 전체적인 분석 결과의 패턴과 방향성에 대해서 큰 영향을 미칠 정도의 차이는 아닌 것으로 판단된다. 하지만, 종사상 지위와 직업 항목의 경우에는 확률적 연계결과가 일을 하지 않은 비율이 조금씩 더 늘어나는 것을 확인할 수 있었다. 앞에서 언급하였듯이 이러한 이유 중에는 연계변수가 부족하여 연계의 정확도가 떨어질 수 있고, 내검오류 및 무응답 등으로 연계변수의 효과가 감소할 수도 있을 것이다. 그리고 최고 연계가중값이 동점인 경우 비율(도수)이 높은 부분으로 연계될 가능성이 높아질 수도 있을 것이다. 하지만, 기존 자료가 유성구와 홍성군에만 한정된 시범예행조사가 아닌 2010년 센서스 자료를 기준자료로 이용한다면 이러한 문제는 해결될 수 있을 것이라 판단된다. 평가결과를 살펴보면 결정적 연계자료

중 25%를 확률적 연계자료로 대체한 경우에는 실제 자료와 특정부분에서의 차이는 발생하지만, 자료의 전체적인 패턴과 방향성에는 큰 변화가 발생하지 않는 것으로 보인다. 그리고 실제 2010년 센서스를 기준으로 2015년 센서스와 종단자료를 구축했을 경우에는 이 결과보다 차이는 더 작아질 수 있을 것이므로 확률적 자료연계는 충분히 사용가능하다고 생각된다.

연구자의 목표는 본 연구 내용을 바탕으로 2010년 센서스를 기준으로 2015년 센서스와 종단자료를 구축하는 것이다. 본 연구에서의 정확 연계율은 73.1%이나, 종단자료의 특성상 정확 연계율이 더 높아야 정확성 높은 분석결과를 제공할 수 있으므로 무엇보다도 정확 연계율 향상에 많은 노력을 해야 할 것이다. 결정적 연계에 사용되는 식별정보에 대한 표준화 작업 및 주민등록번호를 이용한 연계 등에 대한 노력도 필요할 것이다. 영국의 사례를 활용한 연계가 이에 해당되며 2015년 등록센서스가 진행된 후에 충분한 검토가 필요할 것으로 생각된다. 그동안 센서스는 타 조사의 모집단이 될 뿐만 아니라 전체규모(인구, 주택 등) 통계와 관련하여 많이 활용되어 왔다. 이러한 연속적인 자료가 종단화가 된다면 그 활용가치는 한층 더 높아질 것이다. 또한 다른 조사자료 또는 행정 자료와 연계함으로써 더 유용한 정보를 제공할 수 있도록 추가적으로 연계 가능한 자료 발굴에 더 많은 노력이 필요할 것이다. 마지막으로 본 연구가 타 자료연계 연구에도 많은 도움이 되기를 기대한다.



<참고문헌>

- 변종석 외(2013), 『다양한 출처 자료 처리 및 통계 생산방안 연구』, 통계개발원 정책연구용역
 심규호·박시내(2010), 「통계이용 활성화를 위한 2차 자료 생산·활용 방안 연구」, 통계개발원
 2010년 하반기 연구보고서 제2권 p. 236~301
- 이영섭 외(2007), 『통계조사자료와 행정자료 간의 자료매칭기법 연구』, 통계개발원 정책연구용역
- Australian Bureau of Statistics (2013), "Australian Census Longitudinal Dataset: Methodology and Quality Assessment 2006-2011", ABS cat. no. 2080.5
- Bishop, G. (2009), "Assessing the Likely Quality of the Statistical Longitudinal Census Dataset", ABS cat. no. 1351.0.55.026
- Blackwell, L., K. Linch, J. Smith and P. Goldblatt (2003), "Longitudinal Study 1971 - 2001: Completeness of Census Linkage"
- Guiver, T. (2011), "Sampling-based Clerical Review Methods in Probabilistic Linking", ABS cat. no. 1351.0.55.034
- Hattersley, L. (1997), "Record Linkage of Census and Routinely Collected Vital Events Data in the ONS Longitudinal Study", Record Linkage Techniques - 1997: Proceedings of an International Workshop and Exposition, Arlington
- Iams, H. M. and I. Dushi (2009), "Pension Coverage and Pension Type: Using W-2 tax Records to Correct SIPP Survey Reports", Federal Committee on Statistical Methodology Conference 2009
- Kaminski, M. W. and V. Kapani (2011), "Record Linkage Methodology in Longitudinal Database of Quarterly Census of Employment and Wages", Proceedings of Joint Statistical Meetings 2011
- Peters, P. A. and M. Tjepkema (2010), "1991-2011 Canadian Census Mortality and Cancer Follow-Up Study, Proceedings of Statistics Canada Symposium 2010
- Smith, J. (1999), "The History and Future of Record Linkage in the ONS Longitudinal Study", Joint ECE/Eurostat Work Session on Registered Administrative Records for Social and Demographic Statistics, Geneva
- Solon, R. and G. Bishop (2009), "A Linkage Method for the Formation of the Statistical Longitudinal Census Dataset", ABS cat. no. 1351.0.55.025
- Statistics New Zealand (2006), "Data Integration Manual"
- Statistics New Zealand (2013), "Developing a historical longitudinal census dataset in New Zealand"
- Trewin, D. (2006), "Enhancing the Population Census: Developing a Longitudinal View", ABS cat. no. 2060.0
- Wilkins, R., M. Tjepkema, C. Mustard and R. Choiniere (2008), "The Canadian census mortality follow-up study, 1991 through 2001, Health Reports, Vol 19, no. 3