원넘버 센서스 방법론 연구

– 영국·미국·일본의 사례를 중심으로 –

박영실 · 정남수

목 차

표 목 차

그림 목차

원넘버 센서스 방법론 연구
– 영국·미국·일본의 사례를 중심으로 –

제1절 서 론

1. 연구배경

센서스[1]는 조사 지역에 대한 완전한 집계를 목표로 하고 있으나,[2] 이러한 이상이 실현되기란 쉽지 않다. 센서스 전 과정에서 발생하는 커버리지 오차(coverage error)로 인해서 센서스에서 조사된 인구와 실제(true) 인구와는 항상 차이가 나게 마련이다. 커버리지 오차란 모집단의 구성원을 누락 혹은 중복, 잘못 조사함으로써 발생하는 오차를 뜻하며, 조사 지도나 조사 지역의 가구주소 리스트가 불완전하거나 부정확해서, 또는 조사 대상의 범주를 잘못 적용함으로써 발생하게 된다(Wachter, 2008).

조사환경이 어려워지면서 센서스 커버리지 오차의 문제가 심화되는

1) 본 연구에서는 원넘버 센서스라는 제목과의 일관성 유지를 위해 미국과 영국의 센서스, 일본의 국세조사, 그리고 우리나라의 인구주택총조사를 센서스로 통일해서 부르고자 한다.

2) 유엔의 권고안에 따르면, 센서스는 주기성(defined periodicity), 동시성(simultaneity), 개별성(individual enumeration), 보편성(universality within a defined territory)의 기본원칙 하에서 실시된다. 센서스는 일정한 주기를 갖고, 정기적으로 조사기준시점을 정하여 정해진 기간 동안 동시에, 모든 인구를 개별적으로 조사한다. 이 때 조사하기로 계획된 영역내의 모든 인구를 누락과 중복없이 조사해야 한다(UN, 2008).

경향이 있다. 일부 국가에서는 센서스 응답률이 낮아지고 특정 집단의 커버리지 오차가 커지자, 센서스 자료의 정확성에 대해서 문제를 제기하기도 한다. 그런데 커버리지 오차의 심각성은 센서스의 불완전성 그 자체 보다는 오차의 수준과 그 범위를 파악할 수 없을 때 발생하므로 커버리지 오차에 대해서 객관적으로 평가하는 것이 무엇보다도 중요하다. 커버리지 오차를 측정하여 이용자들에게 제공함으로써 센서스 결과를 해석하는데 도움을 줄 수 있기 때문이다. 그리고 그 결과는 최상의 인구를 추정하기 위한 기초 자료로 제공되거나 식별된 오차를 고려하여 센서스 결과를 보정하는데 활용할 수도 있다(UN, 2008).

커버리지 오차는 여러 가지 방법으로 측정되는데 인구분석과 사후조사 방법이 보편적으로 활용되고 있다. 인구분석[3]은 이전 센서스의 결과에 출생, 사망, 이동에 관한 행정자료를 결합하여 성, 연령, 인종 등에 대한 인구추정치를 산출하는 방법으로 전국수준(national level)에서 신뢰할 만한 수치를 제공해 줄 뿐만 아니라 시계열적으로도 그 패턴을 비교할 수 있다는 점에서 유용한 방법으로 평가받고 있다. 그러나 불법적인 출입국자들의 규모가 알려져 있지 않고, 내부 인구이동 자료가 불충분해서 하위지역수준(sub-national level)에서의 인구추정치를 제공하는 데에는 한계가 있다. 이에 따라 사후조사라는 표본조사를 활용하여 인구의 누락규모를 추정하는 방법이 함께 사용되어오고 있다. 사후조사는 센서스 결과를 자체적으로 평가하기 위해 실시하는 표본조사로 소지역에서 인구를 추정할 수 있다는 장점을 가진다(NRC, 1994). 우리나라를 비롯하여, 캐나다, 호주, 뉴질랜드, 영국 등은 센서스 커버리지 오차를 평가할 필요성을 인식하고 사후조사방법으로 커버리지 오차를 측정하고 있다.

영국은 커버리지 오차를 측정하는데서 머무르지 않고 2001년 센서스에서 누락규모를 보정하여 센서스 데이터베이스를 구축하였다. 그 결실을 맺지는 못하였으나, 미국의 경우에도 1990년과 2000년 센서스에서 이와 유사한 시도를 한 바 있다. 이와 같이 보정 후의 수치를 얻기 위한

3) 인구분석 방법은 인구균형방정식, 코호트성분분석방법, 센서스간 코호트생잔률법, 코호트회귀생잔계수 비교방법 등이 포함되나, 인구균형방정식이 가장 대표적인 방법으로 알려져 있다. 자세한 설명은 이지연(2004) 참고.

방향으로 개선된 센서스 방법론을 원넘버 센서스(One Number Census)라고 한다(전광희, 2008a). 누락 혹은 중복 등의 이유로 커버리지 오차를 측정하는 국가들은 센서스 본조사[4] 결과와 사후조사를 통해서 보정한 결과가 공존하게 됨으로써 공식적이든 비공식적이든간에 실제로 두 가지의 인구수가 존재하게 되는 반면에, 원넘버 센서스에서는 보정 후의 단일 인구 수치만 존재한다. 이런 두 가지의 인구수가 공존하는 전통적인 센서스와 구별하기 위해 원넘버 센서스라고 부르는 것이다. 원넘버 센서스를 좀더 확대하여 해석해 보면, 단일 인구를 작성하는 것 뿐 아니라 작성된 단일의 인구를 추정이나 추계인구의 기준으로 제공함으로써 센서스인구를 모든 인구통계의 중심에 두고자 하는 프로젝트라고 볼 수 있다.

센서스 커버리지 오차의 문제로 고심하고 있는 국가들은 영국의 원넘버 센서스 결과를 예의주시하고 있다. 우리나라에서는 아직까지 센서스 커버리지 오차에 대한 공공의 관심이 영국이나 미국처럼 높지는 않다. 센서스 커버리지에 대한 연구(이지연, 2004; 2006)가 극소수에 불과한 사실은 이러한 현실을 잘 말해준다. 그러나 우리나라에서도 원넘버 센서스를 주목할 필요가 있다. 왜냐하면, 우리나라 센서스 커버리지 오차의 특성이 영국이나 미국에서 원넘버 센서스에 대한 논의가 출발하게 된 배경과 유사한 특성을 보이고 있기 때문이다. 이 두 국가에서 모두 특정 인구 집단의 과소집계 문제가 지속적으로 제기되면서 원넘버 센서스에 대한 논의가 시작되었는데, 우리나라에서도 지리적 이동성이 높은 20대의 커버리지 오차가 지속적으로 높게 발생하고 있다(이지연, 2006). 또한 이러한 상황 하에서 센서스인구가 추계인구 및 주민등록인구와 비교되면서 국민들에게 혼란을 야기하고 있다. 세 가지 인구가 개념, 조사 범위 및 공표 시기 등에 엄연히 차이가 있음에도 불구하고 수치 그 자체가 직접적으로 비교되고 있다. 이에 따라 우리나라의 실제 인구에 대한 혼란이 예상되며, 이에 통계청 일각에서는 단일 인구의 필요성이 제기되고 있다(통계청, 2008a).

---

4) 사후조사와의 명확한 비교를 위해서 필요한 경우에는 전통적인 의미에서의 센서스를 센서스 본조사로 부르고자 한다.

2. 연구목적 및 내용

위와 같은 문제제기 하에 본 연구에서는 우리나라 총인구와 가구의 정확한 규모 및 분포를 파악할 수 있는 원넘버 센서스 방법론을 연구해 보고자 한다. 방법론 연구는 원넘버 센서스를 실시하고 있거나 시도한 경험이 있는 국가들의 사례연구를 통해서 진행할 것이다. 그리고, 우리나라에서 센서스 커버리지 오차 측정방법과 그에 기반하여 측정된 오차의 수준 및 특성 분석을 통해서 시사점을 도출해 볼 것이다.

사례연구의 대상 국가는 영국, 미국, 일본이다. 영국은 2001년에 원넘버 센서스를 처음으로 실시한 국가로 전광희(2007)에 의해서 관련 내용이 소개된 바 있다. 그러나 영국의 원넘버 센서스가 기존과 갖는 차별적인 특징인 보정 부분에 대한 검토가 충분하지 않으므로 본 연구에서는 이를 보완하여 다루고자 한다. 미국과 일본의 원넘버 센서스에 대한 내용은 거의 소개된 적이 없다. 두 국가 모두 명시적으로 원넘버 센서스를 실시하고 있지 않기 때문이다. 그러나, 미국의 경우 원넘버 센서스와 관련한 오랜 역사를 가지고 있다. 문헌연구(Wright, 1999)에 따르면 센서스 초기 부터 과소집계에 대한 관심을 갖고 있었으며, 1980년 이후 원넘버 센서스에 대한 논의가 본격적으로 시작되었다. 비록 좌절되기는 하였으나 2000년 센서스에서 원넘버 센서스를 계획한 적이 있기도 하다. 그리고 이러한 경험은 영국의 원넘버 센서스에 직간접적으로 영향을 끼쳤다. 일본은 영국과 미국처럼 과소집계의 규모를 측정한 이후 실제 인구를 추정하고 보정하는 방법의 원넘버 센서스를 실시하고 있지는 않다. 그러나, 일본 내에서 모든 인구통계가 센서스를 기반으로 작성되고 있다는 점에서 광의의 원넘버 센서스를 실시하고 있다고 판단되어 사례연구에 포함시켰다. 다음 절에서는 먼저 이들 세 국가의 원넘버 센서스를 등장배경, 개념, 방법론, 결과를 중심으로 설명하고자 한다.

제2절 원넘버 센서스 방법론 검토

1. 영국의 원넘버 센서스

가. 원넘버 센서스 등장배경

1801년에 시작된 영국의 센서스는 1941년의 센서스를 국민등록법(National Registration Act)이 대체한 것을 제외하고는 현장조사 방법에 의해서 매 10년마다 실시되고 있다. 센서스는 전국 인구에 대한 스냅샷을 제공하고 소지역 단위에서 양질의 자료산출을 목적으로 한다. 이와 함께 지방정부에 대한 예산 배분의 근거가 되는 추정인구의 기준이 되며 건강, 교육, 이동 및 주택 등에 대한 계획을 수립하는 데 사용된다. 따라서 정확한 인구수를 집계하는 것은 영국 통계청(Office of National Statistics)의 주요한 임무 중 하나이다(ONS, 2001a).

그러나 센서스를 완전하게 수행하기란 불가능하다. 이에 1966년부터 사후조사를 통해서 센서스를 평가하고 그 결과를 이용자들에게 제공함으로써 센서스의 커버리지에 대한 이해를 돕고 있다. 1991년 센서스 이전까지는 사후조사를 통해서 보정된 센서스인구와 인구분석에 의한 인구추정치가 거의 일치하는 것으로 나타났다. 또한 사후조사 결과 추정된 과소집계 수준은 1% 미만으로 다른 국가들과 비교해 볼 경우 상대적으로 작았다. 그런데, 1991년 센서스에서 다음과 같은 두 가지 문제가 발생하였다. 첫째, 과소집계 수준이 2.2% 가량으로 높아졌을 뿐 아니라 그 정도가 특정한 지역 및 성, 연령별 하위집단에서 차별적으로 높게 나타났다. 예컨데, 도시 내부의 젊은 남성이 20% 이상 누락된 것으로 밝혀졌다. 둘째, 1981년과 달리 사후조사를 통해 보정된 센서스인구와 인구분석을 통해서 추정된 인구 사이에 큰 차이가 발생해서 인구집계의 정확성에 대한 신뢰성의 위기 문제가 생겼다(Brown et al., 1999). 더구나, 센서스와 사후조사의 종속성으로 인해서 과소집계 규모와 그 분포를 정확히 밝히는데 실패함으로써 그에 대한 원인분석을 하는데 많은 시간이 소요되었다(ONS, 2001a).

이에 2001년 센서스에서는 커버리지를 최대로 향상시키고자 조사표

재설계, 조사대상으로서의 인구 개념에 대한 검토,[5] 조사방법 개선, 응답률이 낮은 지역에 대한 자원 집중 등의 전략을 시도하였다. 그러나, 이러한 노력에도 불구하고 현실적으로 일정 정도의 누락이 발생할 수 밖에 없으므로 원넘버 센서스를 계획하기에 이르렀다(ONS, 2001b).

원넘버 센서스는 사우스햄턴 대학(Southhamton University) 통계학과 교수들의 자문을 토대로 구상되었으며, 영국 통계청은 외부전문가, 지방자치단체대표, 통계학술기관 등을 포함하는 운영위원을 조직, 프로젝트의 방법론 개발과 운영관리 감독을 맡게 하였다. 또한 내부 국장급 직원과 여타 정부기관의 대표, 외부 전문가들로 구성된 원넘버 센서스 프로젝트 보드(Project Board)를 설치하여 진행상항을 수시로 점검하도록 하였다(전광희, 2007).

원넘버 센서스는 1997년에 450가구의 소규모 예비조사를 통해서 그 실행가능성을 검토하였으며, 1998년에는 2,000가구 정도의 대규모 시험조사를 통해서 사후조사가 정해진 시간 내에 완료될 수 있는지를 확인하였다. 1999년에는 대표적인 5개 지역을 선정하고 그 지역 내의 약 1만 8,000가구를 대상으로 시범예행조사(dress rehearsal)를 실시하였다. 이후에도 작은 규모의 조사를 수시로 진행하면서 원넘버 센서스의 방법론을 정교화하였다(Pereira, 2002).

나. 원넘버 센서스 개념

원넘버 센서스는 센서스 본조사에서 누락된 개인과 가구의 규모를 추정하고, 누락된 만큼을 센서스 본조사 결과에 보정하여 개인 수준의 데이터베이스를 구축하는 것을 일컫는다. 원넘버 센서스는 첫째 지방자치단체(Loacal Authority District) 수준에서 추정인구를 제공해 주며, 둘째 누락분이 보정된 데이터베이스를 구축함으로써 모든 통계표가 일관적으로 단일의 인구수치를 작성할 수 있게 해 주는 것을 목적으로 하였다

5) 영국 센서스 역사의 상당기간(1801~1971년) 동안 현주(de facto) 개념을, 1981년과 1991년에는 상주개념과 현주개념을 모두, 2001년에는 상주개념을 사용하였다. 2011년 센서스에서는 센서스 기준일의 자정에 발견되는 방문자 정보를 추가하여 사용하기로 했으며(NRC, 2006), 일상거주지 이외의 거주장소와 거주기간에 대해서도 추가로 질문할 것을 고려하고 있다.

(ONS, 2001a).

누락규모를 파악하기 위해서는 센서스커버리지조사(Census Coverage Survey)라고 불리는 사후조사[6]를 독립적으로 실시하였다. 센서스와 사후조사 자료를 매칭한 후 그 결과를 토대로 이원시스템 추정법(Dual System Estimation, DSE)을 사용하여 센서스에서 누락된 가구와 개인의 총수와 특성을 추정한 후에 이 개인과 가구를 센서스 데이터베이스에 보정하는 작업을 실시하였다. 그 과정을 도식화하면 [그림 1]과 같다.

다. 원넘버 센서스 방법론

1) 사후조사

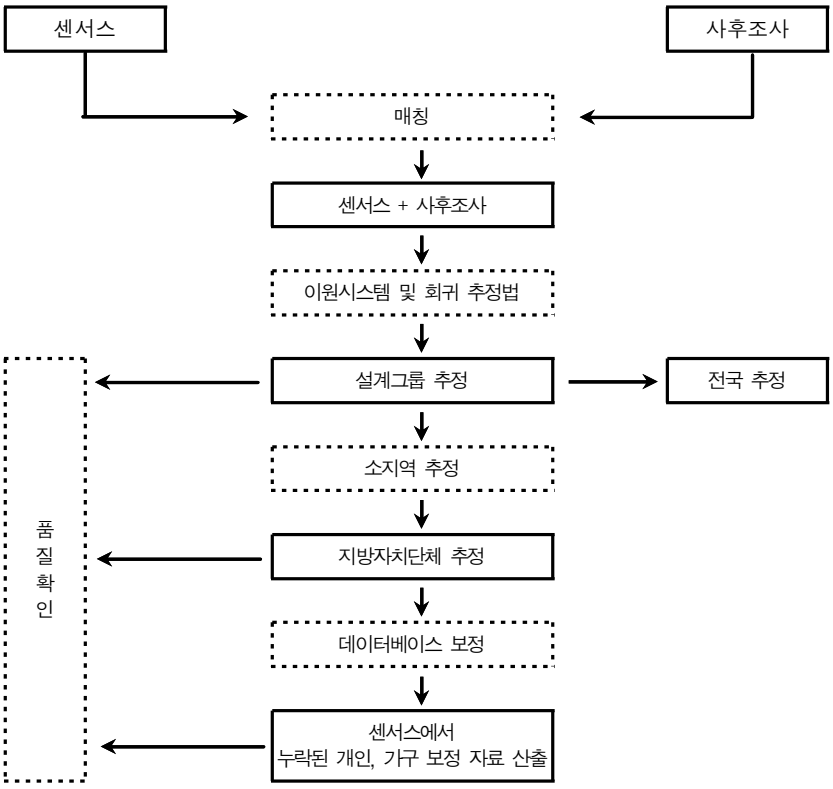
사후조사가 원넘버 센서스 목적에 부합하기 위해서 충족시켜야할 조건은 다음과 같았다(Pereira, 2002). 첫째, 센서스와는 전적으로 독립적이어야 한다. 이원시스템 추정법을 통해서 누락된 개인 및 가구의 규모를 추정하기 위해서 두 조사간의 독립성은 필수적인 전제조건이다. 둘째, 센서스의 일차적인 목적이 소지역 수준에서 양질의 자료를 제공하는 것인만큼 누락 규모의 보정은 지방자치단체 수준에서 이루어져야 한다. 셋째, 이와 연관지어 지방자치단체 내 우편번호는 지방자치단체를 대표할 수 있도록 뽑혀야 한다. 넷째, 센서스 예산 범위 내에서 효율적으로 운영해야 한다. 다섯째, 공공의 부담은 최대한 줄여야 한다.

사후조사는 표본설계와 자료수집단계로 구분하여 살펴볼 수 있다.

가) 표본설계

사후조사는 우리나라의 시군구에 해당하는 지방자치단체 수준에서 성 및 연령별 인구추정치를 얻는 것을 목적으로 하였다. 그런데 지방자치단체에서 직접적인 인구추정치를 얻기 위해서는 대규모의 표본이 요구되며 이는 많은 예산을 소요로 하였다. 이에 영국 통계청은 인구 50만 명을 기준으로 433개의 지방자치단체를 인접 지역끼리 묶어서 설계그룹

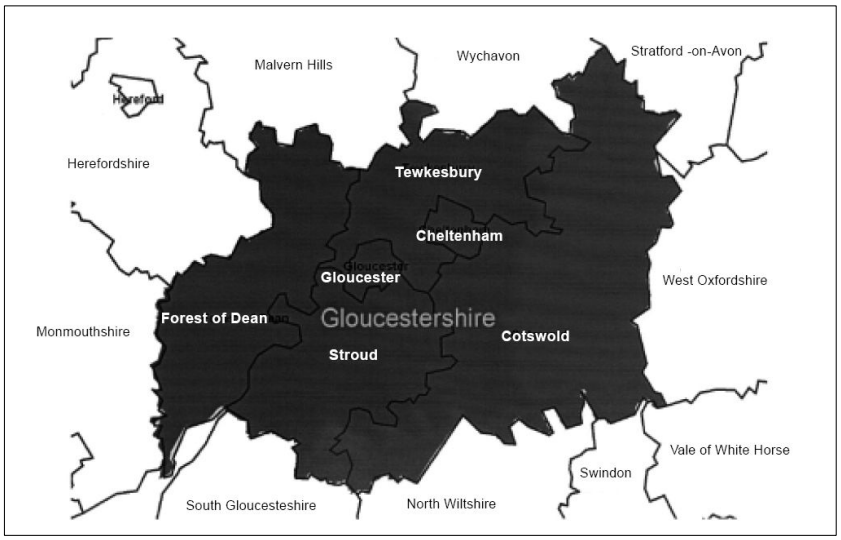
6) 2001년 원넘버 센서스에서 실시된 사후조사를 이전의 사후조사와 구별하기 위해 센서스커버리지조사라고 하나, 이후 미국이나 일본 등에서 실시된 사후조사와의 용이한 비교를 위해서 본 연구에서는 센서스커버지리조사를 사후조사로 부르고자 한다.



자료: ONS, 2001a.

[그림 1] 영국의 원넘버 센서스 흐름도

(Design Group)을 만든 후 이를 기반으로 개인과 가구를 추정하는 표집 전략을 채택하였다. 인구규모에 따라서 설계그룹은 하나의 지방자치단체 혹은 지방자치단체의 그룹으로 구성되었다([그림 2]). 표본설계에 따라서 설계그룹의 36개 성 및 연령별 그룹에서 누락 규모에 대한 직접추정이 가능하도록 하였다(ONS, 2001a).



자료: Pereira, 2002.

[그림 2] 단일지방자치단체로 구성된 설계그룹의 예

사후조사는 우편번호를 표집단위로 하는 지역기반(area-based) 표집을 하였다. 물론 기술적으로 볼 때, 영국우편주소파일(UK Postal Address File)의 배달지점(delivery points)을 추출함으로써 가구기반(household-based) 표집이 가능하였다. 그러나, 우편번호가 영국 전역을 완전하게 포함하고 있지 않아서 센서스 커버리지를 체크하는데에는 적당하지 않다는 판단하에 지역기반 표집전략을 택하였다.

각 지방자치단체에서 우편번호를 추출하기 위해서 2단계 표집방법이 적용되었다. 1단계에서 센서스 조사구를 추출한 후 2단계에서는 추출된 조사구 내에서 우편번호를 무작위로 선택하였다. 우편번호에는 가구 수 이외에 어떠한 정보도 포함되어 있지 않으므로 우편번호를 직접적으로 층화하는 것은 불가능하였다. 따라서 우편번호를 층화하기 위해서 각 가구 및 지역에 대한 마이크로 정보를 포함하고 있는 1991년의 센서스 조사구와 우편번호를 연결하였다.[7] 그리고 난 후에 추출된 조사구로부터 비용 및 실용성의 관점 하에 최대 5개의 우편번호를 무작위로

7) 영국의 우편번호는 평균 15개의 주소로 구성되어 있다.

뽑았다(ONS, 1998).

그런데, 센서스에서의 누락은 특정한 사회, 경제 및 인구학적 특성을 가진 지역 내에서 더 높을 것으로 예상되었다. 예컨데, 한 가구 이상이 함께 거주하는 주택에 있는 개인이 그렇지 않은 개인에 비해서 센서스에서 누락될 확률이 높다는 것이다. 이처럼 조사 확률에서의 이질성을 통제하기 위해서 각 설계그룹 내의 조사구는 집계난이도(hard to count)에 따라서 층화되었다(Brown et al., 1999).

집계난이도 지수는 1991년 센서스에서 누락에 영향을 미치는 요인으로 밝혀진 변수들로 구성되었다. 해당 변수는 실업률, 외국어(출생국가) 사용여부, 다가구주택, 세입자 동거여부, 보정된 가구 비율이며 이 변수들은 1999년도의 시범예행조사에서 다시 한번 검증하는 단계를 거쳤다. 집계난이도 지수에 따라서 영국 전역을 조사가 쉬운 지역부터 어려운 지역까지 40%, 40%, 20% 비율의 3개 집단으로 구분하였다. 최종적으로 집계가 쉬운 지역은 3.4%, 집계가 쉽지도 어렵지도 않은 중간지역은 3.7%, 집계가 어려운 지역은 4.5%의 표집율을 부여함으로써 대표성이 있는 표본추출이 이루어지도록 하였다(ONS, 2001b).

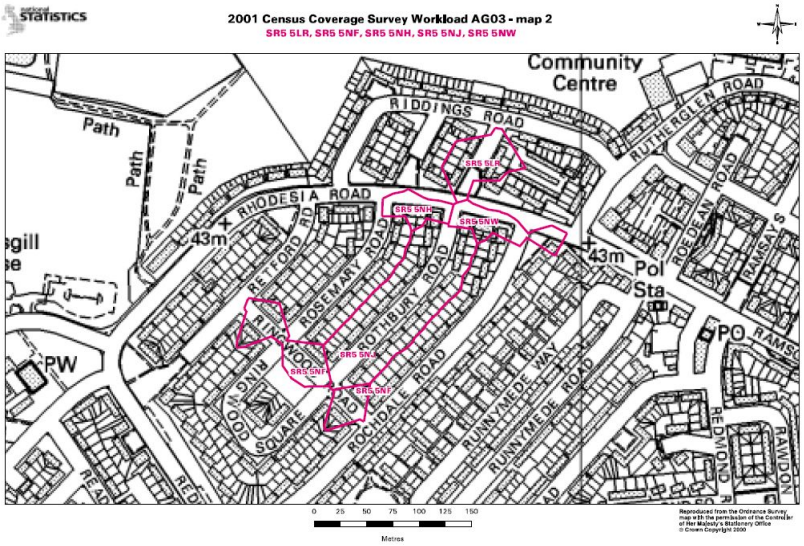
표본규모는 추정치의 정확성과 함께 예산과 관련하여 중요한 관심영역이다. 지방자치단체별로 직접적인 추정치를 얻기 위해서는 100만 가구 이상의 표본이 필요하나 인력 및 예산현황을 고려할 때 이러한 규모는 현실적으로 불가능한 것으로 판명되었다. 이에 따라서 영국 통계청에서는 비용 대비 정확성 분석을 통해서 표본규모를 산출하였다. 수차례의 시뮬레이션 결과 개별 설계그룹 인구에 대하여 1% 이하의 상대표준오차(relative standard error)를 허용하는 추정치를 얻기 위해서 잉글랜드와 웨일즈에서는 약 32만 가구, 스코틀랜드에서는 약 4만가구, 북아일랜드에서는 약 1만 가구를 표집하였다(ONS, 1998).

나) 자료수집

사후조사의 자료수집은 센서스 본조사 이후 3.5주 후에 약 3주간 실시되었다(Pereira, 2002). 조사기간은 센서스와 사후조사 사이 인구이동이 최소화되는 시점을 고려하여 설정되었다. 또한 응답자들이 센서스 시점에 해당 가구에 누가 거주했는지를 최대한 기억하도록 하는 것도

중요하게 고려되었다.

센서스와 사후조사의 독립성을 유지하기 위해 조사원에게는 센서스에 대한 어떠한 정보도 제공하지 않았으며, 조사를 시작한 처음 이틀 동안 표집된 우편번호 내에 있는 모든 가구를 방문하여 독자적으로 주소록을 작성하도록 하였다. 영국에서 우편번호는 지역적인 경계가 잘 정의되어 있지 않다. 따라서 조사원에게 [그림 3]과 같이 우편번호의 경계를 표시하도록 지시하였다.



자료: Pereira, 2002.

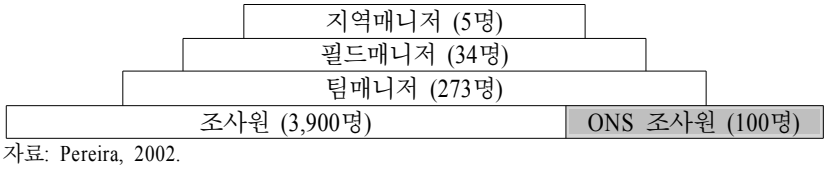
[그림 3] 영국의 사후조사 조사원 지도

조사원은 우편번호 지역에 방문하여 강도높은 면접조사를 하였다.[8] 면접에서는 응답자의 부담을 줄이기 위해서 매칭 작업에 필요한 핵심질문만을 반복 조사하였다. 조사항목은 해당가구에 거주하는 모든 개인에

8) 미국의 2000년 센서스 사후조사에서 채택하고 있는 컴퓨터를 이용한 조사 방법을 대안으로 고려하였으나 면접시간이 짧은 상황에서 컴퓨터의 사용이 오히려 면접에 방해요소로 작용할 가능성이 높다는 판단 하에 조사원 방문면접조사(Paper and Pencil Interviewing, PAPI)를 실시하였다(Pereira, 2002).

대한 기본적인 인구학적 특성과 거주유형, 그리고 관계정보에 대한 것이었다. 이외에도 우연히 누락되었을 개인을 찾기 위한 프로빙 질문을 포함하고 있었다. 특히 센서스 기준 시점에 일을 하러 나간 가구원이나 군대 혹은 병원에 있거나 휴가를 간 가구원이 있는지를 다시 물었다.

조사원 업무량을 고려하여 인접 우편번호를 하나로 모아서, 1인당 90～200가구의 면접이 가능하도록 설계하였다. 또한 사후조사의 성공적인 수행을 위해서 조사원간 팀웍을 중요시 여겼으며, 사후조사 조사원과는 별도로 영국 통계청 가구 담당 조사원 100여명을 상시로 두어서 조사가 어려운 가구 및 지역을 맡고 있는 조사원에게 도움을 주도록 하였다([그림 4]).



자료: Pereira, 2002.

[그림 4] 영국의 사후조사 조사원 조직도

2) 센서스와 사후조사 자료의 매칭작업

자료수집이 완료되면, 센서스와 사후조사 자료를 매칭하였다. 이 때 매칭을 정확하게 하는 것이 매우 중요하였다. 왜냐하면 매칭결과는 센서스에서 누락된 가구와 개인을 추정하는데 직접적인 영향을 미치기 때문이다. 예컨데, 0.1%의 매칭오류가 발생할 경우 0.1%의 정적인 편향(positive bias)이 발생하였다. 매칭과정은 크게 자동매칭과 수동매칭으로 구분되었다(ONS, 2001c; 2005).

가) 자동매칭

자동매칭은 다시 정밀매칭(exact matching)과 확률매칭(probability matching)으로 이루어졌다. 정밀매칭은 센서스의 개인과 가구가 사후조사의 그것들과 자동적으로 정확하게 매칭되는 것을 말한다. 정밀매칭에 사용되는 변수는 가구의 경우 우편번호, 주소, 주택유형, 가구대표자의 성, 가구원 수이며, 개인의 경우 성과 이름, 출생년월, 혼인상태, 가구대

표자와의 관계이다.

정밀매칭이 되지 않은 경우는 확률매칭의 단계로 넘어갔다. 이 단계에서의 매칭은 확률가중치(probability weight)에 따라서 이루어졌다. 확률가중치란 센서스와 사후조사의 두 레코드가 동일인(가구)일 가능성으로 확률가중치가 높을 수록 두 레코드 사이의 일치도가 높다는 것을 의미하였다. 예컨데, 레코드의 쌍이 하나의 사항만 예외로 하고 똑같다면 기재오류(recording error)일 가능성이 있다고 보고 높은 가중치를 부여하였다. 확률가중치가 높은 가구들을 연결하고 그 가구 안에 있는 개인들을 비교하여 매칭하였다.

나) 수동매칭

센서스와 사후조사의 두 레코드가 어느 정도는 일치하나 매칭되지 않을 경우에는 수동해법(clerical resolution)의 단계로 넘어갔다. 이 단계에서는 레코드의 쌍을 보고 매칭 쌍인지의 여부를 판단하였다. 마지막으로 매칭에 실패한 개인과 가구를 센서스 조사표에서 직접 찾는 수동매칭을 하였다. 수동매칭과정에서 매칭요원들은 컴퓨터 화면으로 사후조사와 센서스 조사표를 보면서 작업하였다(Computer Assisted Matching System).

다) 품질확인

일반매칭요원(matcher)의 작업결과는 철저한 품질확인작업을 통해서 검증된 경우에만 다음 단계로 넘어가는데 크게 세 단계를 거쳤다. 일반매칭요원이 작업한 결과는 전문매칭요원(expert matcher)에 의해서, 전문매칭요원이 작업한 결과는 품질확인요원(quality assurer)에 의해서 검토되었다. 일반매칭요원은 200건의 레코드를 정확히 매칭해야만 그 품질을 인정받으며 그 이후에는 작업결과 중 무작위로 선택된 10%에 대해서 품질검토를 하였다. 이 때 한 건의 매칭오류라도 발견될 경우에는 다시 200건의 기록을 검토하였다. 전문매칭요원에 의한 검토결과는 다시 품질확인요원에 의해서 일반매칭요원에 대한 방법과 동일한 방법으로 검토되었다.

3) 추정작업

다음 단계는 센서스와 사후조사의 매칭자료를 사용하여 모든 지방자치단체의 인구수를 추정하는 것이다. 추정은 설계그룹에 대한 직접추정과 지방자치단체에 대한 간접추정으로 나뉘었다.

가) 설계그룹 수준에서의 추정

설계그룹의 인구는 사후조사 표본인 우편번호에서의 총인구를 추정한 후 이 수치에 비 추정법(ratio estimation)을 적용함으로써 직접적으로 추정되었다. 우편번호 내의 인구를 추정하는데에는 과소집계를 추정하는 표준적인 방법인 이원시스템 추정법을 사용하여 다음과 같이 총인구를 추정하였다(<표 1>).

〈표 1〉 이원시스템 추정모델

구분		센서스		
		포함	누락	합계
사후조사	포함	N_{11}	N_{12}	N_{1+}
	누락	N_{21}	N_{22}	N_{2+}
	합계	N_{+1}	N_{+2}	N_{++}

자료: ONS, 2000a.

$$DSE = N_{++} = \frac{(N_{+1})(N_{1+})}{N_{11}}$$

이원시스템 추정법은 본래 동물의 모수를 추정하는데 사용되어온 방법[9]이나, 찬드라세카와 데밍(Chandraseka and Deming, 1949)에 의해서 인구추정법으로 발전되었다. 1970년 미국의 센서스국(U.S. Bureau of

9) 호수에 있는 물고기 총수를 추정하는 예로 이 방법을 설명해 보자. 호수에 있는 물고기를 가능한 모두 잡은 결과 1,000마리였다. 그 물고기에 빨간색 표시를 한 후 다시 호수로 던졌다. 그리고 나서 100마리의 물고기를 다시 잡았는데, 그 중 빨간색 표시가 되어있는 것은 90마리였다. 이것은 호수의 물고기 중 90%가 빨간색 표시가 되어 있음을 의미하는 것이다. 그러므로, 빨간색 표시가 된 것과 표시가 되지 않은 것을 고려하여 물고기의 총수를 추정해 보면 약 1,111마리(=1,000/0.90) 이다. 이 때 첫번째의 포획을 센서스로, 두번째의 재포획을 사후조사로 볼 수 있다(Freedman, 1991).

Census)에서 이 방법이 사용된 이래로 여러 국가에서 센서스 커버리지 오차를 측정하는데 사용하고 있다.

이원시스템 추정치의 편향을 최소화하기 위해서 다음의 두 가정이 필수적으로 전제되어야 한다. 개인이 센서스에서 조사될 여부와 사후조사에서 조사될 여부에 아무런 영향을 미치지 않는다는 독립성(independence) 가정과 센서스와 사후조사에서 개인이 조사될 확률이 대상 집단의 어떤 하위집단에서도 차이가 없다는 동질성(homogeneity) 가정이 그것이다. 그러나, 이 가정들은 두 조사 사이의 인과적 의존성(casual dependence) 혹은 같은 층 내에서 조사된 확률의 이질성(heterogeneity)등으로 인해 현실적으로 충족되기 어렵다는 문제가 있다. 대안적으로 다음과 같은 전략을 취함으로써 가정의 정당성을 확보하고 있다. 독립성 가정과 관련해서는, 센서스와 사후조사의 독립적인 운영을 통해서 두 조사간의 인과적인 의존성을 최소화하였다. 동질성 가정과 관련해서는, 조사될 확률에서의 이질성과 관련이 있다고 판단되는 특성에 기반하여 집계난이도 지수를 기준으로 층화된 우편번호에 대해서 추정을 함으로써 가정을 충족시키고자 하였다. 두 가정이 전제된 이후 이원시스템 추정법을 통해서 총인구를 추정하였다(ONS, 2005).

다음단계는 우편번호 수준에서의 이원시스템 추정치를 설계그룹 전체로 확대하는 것이었다. 제로절편 회귀모형(zero-intercept regression model)을 구축하여 각 우편번호에서의 이원시스템 추정치를 종속변수로 한 후 이것을 우편번호에 해당하는 센서스 집계결과와 연결시켰다. 이 비 모형(ratio model)은 센서스 집계결과와 각 우편번호 내의 이원시스템 추정치가 상호비례적이라는 가정에 기반한 것이다. 1991년의 센서스 결과에 따르면 지역적인 특성 뿐 아니라 성 및 연령별로 누락이 다르게 나타났다. 이에 따라 설계그룹 내의 집계난이도에 따라서 각 성 및 연령별 하위집단 내에서 비 모형을 독립적으로 적용하였다(ONS, 2001b).

나) 지방자치단체 수준에서의 추정

다음으로 433개 지방자치단체에 대한 추정을 해야 한다. 설계그룹은 하나 이상의 지방자치단체로 구성되었다. 규모가 커서 하나의 지방자치단체가 하나의 설계그룹을 구성한 경우를 제외하고, 규모가 크지 않은

대부분의 지방자치단체는 누락규모에 대한 직접적인 추정을 할 만큼 충분한 우편번호를 포함하고 있지 않으므로 간접적인 추정단계를 거쳐야 했다. 영국 통계청은 지방자치단체의 추정을 위해서 소지역이 대지역과 동일한 특성을 갖고 있다는 가정 하에 대지역 모델을 소지역에 적용하는 간접 추정기법인 합성추정법(synthetic estimation)을 적용하였다. 즉, 112개 설계그룹과 같이 상대적으로 규모가 큰 지역의 자료를 통해서 구축된 모형을 소지역에 적용함으로써 해당 지역의 총인구를 추정하는 것이다. 합성추정법의 가정에 따라서, 하위지역들의 추정치 합은 상위지역의 추정치와 동일해야 한다. 그러므로 설계그룹 아래에 있는 지방자치단체의 합성추정치 합은 설계그룹에 대한 이원시스템 추정치에 의해서 조정되었다(ONS, 2001a).

4) 보정작업

지방자치단체에 대한 추정이 완료된 이후에 센서스 집계치와 추정치의 차이만큼을 센서스 결과에 보정하여 데이터베이스를 구축하였다. 보정은 매칭된 센서스와 사후조사 자료를 사용하여 가구와 개인의 특성에 따라서 센서스에서 조사될 확률 모델을 만들고 이 확률을 커버리지 가중치(coverage weights)로 전환함으로써 시작되었다. 보정은 센서스에서 누락된 개인에 대한 두 가지 접근법에 따라서 독립적으로 이루어졌다. 첫째는 가구 무응답으로 인해서 모든 가구원이 누락된 경우이고, 다른 하나는 조사된 가구 내의 어떤 개인이 누락된 경우이다. 커버리지 가중치에 의해서 누락된 가구와 개인의 주요 특징이 결정되며, 주요 특징 이외의 특성은 기증자 보정방법(donor method)에 의해서 채워졌다. 마지막으로 보정작업의 대상이 된 개인과 가구의 총수 및 분포가 센서스에서 누락된 개인과 가구의 총수 및 분포와 동일한가를 확인하는 조정작업을 거쳤다(ONS, 1999; Steele, 2002).

가) 누락된 가구와 누락된 가구 내의 개인에 대한 보정

센서스와 사후조사 이후에 사후조사 지역 내에 있는 모든 가구는 아래의 세 범주 중 하나에 포함되었다. 이 가정에 따르면 두 조사 모두에서 누락된 가구는 없었다. 이것은 비록 비현실적인 가정이긴 하지만,

두 조사 모두에서 누락된 가구는 이원시스템 추정과정을 통해서 설명되었다. 그리고 보정된 데이터베이스는 이 총수를 만족시키기 위해서 제약되었다.

범주 1: 센서스에서는 조사되었으나, 사후조사에서는 누락된 경우
범주 2: 사후조사에서는 조사되었으나, 센서스에서는 누락된 경우
범주 3: 센서스와 사후조사 모두에서 조사된 경우

가구가 각 범주에 속할 확률은 다음과 같이 정의되었다. 지방자치단체 l의 조사구 d에서 우편번호가 i인 가구 j가 t범주에 속할 확률은 $\Theta^{(t)}_{jidl}$ 이며, 여기에서 t는 범주 1, 2, 3 이다. 이 중 범주 1과 범주 3에 속할 확률은 센서스에서 조사된 개인과 가구의 특성에 따라서 달라진다. 이 확률은 사후조사 표본 내에 있는 우편번호에 해당하는 센서스 자료에 적합하게 되고, 센서스 지역 전체로 외삽(extrapolate)됨으로써 조사된 모든 가구의 확률을 예측할 수 있게 해준다. 센서스에서 조사된 각각의 가구에 대해서 커버리지 가중치는 바로 이 예측확률의 역수로 다음과 같이 정의된다.

$$w^{hh}_{jidl} = \frac{1}{\Theta^{(1)}_{jidl} + \Theta^{(3)}_{jidl}}$$

그런데 지방자치단체 수준에서 조사된 가구에 대한 가중치의 합은 일반적으로 추정치와 일치하지 않았다. 그러므로 커버리지 가중치는 추정치에 맞게 반복비례척도법(iterative proportional scaling)을 통해서 조정되고, 조정된 가중치(calibrated weight)를 기준으로 다음과 같은 순서로 보정이 이루어졌다. 먼저, 센서스에서 조사된 가구 파일과 조정된 가중치를 매칭 한 후 지역과 가중치에 따라서 이 파일을 정렬하였다. 조정된 가중치와 가구수의 누적치를 순차적으로 계산하고 두 변수의 차이가 0.5이상 나게되면 이 사건이 발생한 가장 근접한 위치에 합성가구(synthetic household)를 보정하였다(<표 2>). 가구수의 누적치에는 보정가구가 포함되어 계산되었다.

〈표 2〉 영국 2001년 센서스에서의 누락가구 보정방법

센서스 가구	조사구	조정된 가구 가중치	누적 가중치	누적 가구수(보정포함)
1	1	1.3	1.3	1
2	1	1.3	2.6	2
‡	1	0.0	2.6	3
3	2	1.3	3.9	4
4	1	1.5	5.4	5
5	3	1.5	6.9	6
‡	3	0.0	6.9	7
6	2	1.8	8.7	8
‡	2	0.0	8.7	9
7	3	1.8	10.5	10
‡	3	0.0	10.5	11
8	1	1.9	12.4	12
9	2	2.2	14.6	13
‡	2	0.0	14.6	14
‡	2	0.0	14.6	15

자료: Steele, 2002: 500.
주: ‡ 는 보정된 가구가 놓여지는 센서스 과정의 한 지점을 뜻함.

합성가구의 특성은 기증자 보정방법을 통해서 부여되었다. 기증자는 같은 지역, 같은 가중치를 갖는 가구 중에서 무작위로 선택되었다. 예컨데 첫번째 보정되는 가구의 경우, 기증자는 센서스 가구 1과 2중에서 무작위로 선택되었다. 기증자가 선택되면, 가구의 특성과 그 가구의 가구원이 모두 복사되었으며, 이 가구는 해당 조사구에서 우편번호에 무작위로 할당되었다.

나) 누락된 개인에 대한 보정

가구보정이 끝난 이후에는, 비록 가구는 조사되었으나 그 가구 내에서 개인이 누락된 경우에 대한 보정을 하였다. 개인보정은 가구보정과 유사한 과정으로 진행되었다. 조사된 개인이 포함될 수 있는 가능한 범주는 가구의 경우와 마찬가지로 센서스에서는 조사되었으나 사후조사에서는 누락된 경우(범주 1), 사후조사에서는 조사되었으나 센서스에서

는 누락된 경우(범주 2), 센서스와 사후조사 모두에서 조사된 경우(범주 3)의 세 가지 중 하나에 속하였다. 이 때, $\pi^{(t)}_{kjidl}$은 지방자치단체 l의 조사구 d에서 우편번호가 i인 가구 j에서 가구원 k가 응답범주 t에 포함될 확률을 말한다. 가구모델에서와 마찬가지로 이 확률은 사후조사 표본 내에 있는 우편번호에 해당하는 센서스 자료에 적합하게 되고, 사후조사 표본이 아닌 센서스 전체 지역으로 외삽되어 어떠한 개인이 특정의 응답 범주에 포함될 확률을 예측해 주었다. 이 추정된 예측 확률을 통해서 다음과 같이 개인에 대한 커버리지 가중치가 계산되었다.

$$w^{ind}_{kjidl} = \frac{1}{\pi^{(1)}_{kjidl} + \pi^{(3)}_{kjidl}}$$

커버리지 가중치의 합 또한 반복비례척도법을 통해서 지방자치단체의 총인구와 일치하도록 조정되었다. 센서스에서 조사된 개인은 조정된 커버리지 가중치와 매칭이 된 이후에 이 가중치와 지역에 따라서 정렬되었다. 누적 가중치와 누적 개인수를 순차적으로 계산을 하면서 그 차이가 0.5이상 초과되는 지점에 합성개인(synthetic individual)을 보정하였다.

그런데, 누락된 개인에 대한 보정은 보정과정 중 가장 복잡한 단계이다. 왜냐하면 가구에 개인을 집어넣게 되면, 이 개인을 받게 되는 가구, 즉 수령가구(recipient)의 구조가 변하기 때문이다. 개인보정을 위한 기증자를 찾는 과정은 가구보정과 마찬가지로 같은 조사구 내 같은 가중치를 갖는 개인들 중에서 무작위로 선택되었다. 적정한 기증자가 찾아지면 보정될 개인이 놓이게 될 가구를 탐색하였다. 그 가구의 특성은 선택된 수령가구로부터 오게 된다. 수령가구의 유형은 기증자의 연령, 혼인상태, 가구구조 등에 따라서 좌우되는데, 왜냐하면 가구구조가 단독가구, 16세미만의 자녀를 둔 편부모가구, 부부가구, 16세 미만의 자녀를 둔 부부가구 등으로 정의되기 때문이다.

수령가구는 기증자와 잠재적인 수령가구에서 개인들의 성과 연령에 따라서 정렬한 후, 기증자 가구 내에서 k번째 가구원(비기증자)의 성 및 연령과 잠재적인 수령가구의 k번째 성원을 매칭한 후 비교를 통해서 선택되었다. 어떠한 기증자에 대해서는 적합한 수령가구가 매우 여럿이

될 경우도 있는데, 이러한 경우에는 부가적인 매칭기준을 적용하였다. 일차적으로는 기증자의 가구와 잠재적인 수령가구 사이에 통계적인 거리(statistical distance)를 계산하였으며, 만일 통계적인 매칭에서도 결정이 나지 않을 경우에는 지리적으로 인접한 곳에 위치한 가구를 선택하였다. 수령가구가 선택된 이후에 보정된 개인은 이 가구에 더해지고, 개인의 특성에 수령가구의 특성이 할당되었다. 그리고 보정 이후에는 센서스와 동일하게 내검과정을 거쳤다.

다) 추정인구 수치의 조정

최종적으로 보정된 가구와 개인의 분포가 센서스에서 누락된 것으로 조사된 가구와 개인의 분포와 동일한지를 검토해야 한다. 가구와 개인의 수는 추정 및 보정단계에서 원넘버 센서스의 추정치와 동일하게 조정되었으나, 개인의 보정과정을 통해서 너무 큰 규모의 가구가 생성되어 가구규모 분포는 정확하지 않을 수 있기 때문이다. 이에 따라서 사후보정 데이터베이스(post-imputation database)에서 다시 보정된 개인을 제거(pruning) 혹은 추가(grafting)하는 작업을 하였다. 제거 혹은 추가의 과정은 규모가 큰 가구에서부터 시작하여 규모가 1인 가구 순으로 진행하였다. 그 결과 개인수준의 데이터베이스는 최적의 추정치를 만들어내게 되며 이 데이터베이스로부터 나온 집계표는 자동적으로 모든 변수와 모든 지역수준에서 인구추정치와 일관적이게 되었다.

지방자치단체 수준에서 원넘버 센서스의 총수는 T이며, 사후보정된 총수는 $T^{(imp)}$라고 할 때, T_s 와 $T_s^{(imp)}$는 가구규모 s에서, T_a와 $T_a^{(imp)}$는 성 및 연령별 그룹 a 에서, T_{as}와 $T_{as}^{(imp)}$는 a라는 성 및 연령별 그룹과 가구규모 s 인 그룹에서 각각 원넘버 센서스의 총수와 사후보정된 총수를 뜻한다. 원넘버 센서스의 분포와 일치시키기 위해서 보정된 데이터베이스에서 원넘버 센서스 수치와의 차이 만큼을 제거해 주어야 한다($X = T^{(imp)} - T$). 최종적으로 보정될 수치는 $X_s = (T_s^{(imp)} - T_s)/s$로, 이것은 가구규모 s에서 사후보정 데이터베이스와 원넘버 센서스 추정치를 맞추기 위해서는 X_s 만큼 보정되어야 함을 뜻한다. 그런데 이 경우, 성 및 연령별 총수가 동시에 고려되어야하므로 성 및 연령별 그룹에서 원넘버 센서스 추정치와 맞추기 위해서 추가되어야 할 개인의 총

수는 $X_a = T_a^{(imp)} - T_a$ 가 된다. 결국, 해당 가구규모의 성 및 연령별 그룹에서 추가 혹은 제거되어야 할 총수는 $X_{as} = X_s \frac{X_a}{X}$ 이다.

예컨데, <표 3>과 같이 3개의 성 및 연령별 그룹이 있고 최대 가구규모가 4이며 $T^{(imp)}$=1,490이고 T=1,465인 데이터베이스가 구축되었다고 가정을 해보자. 이에 따르면, 원넘버 센서스의 목표치와 맞추기 위해서는 25명이 제거되어야 한다. 첫번째 단계의 제거와 추가 과정에서, 가구규모가 4인 경우가 보정되어야 하므로 X_4=(560-540)/4=5로 4명이 가구원으로 있는 5가구가 제거되어야 한다. 이 5가구를 성 및 연령별 분포를 고려하여 가구규모 4로부터 추가 혹은 제거해준다. 추가와 제거 이후에 <표 3>의 2단계와 같은 데이터베이스가 구축된다. 여기에서 $T^{(imp)}$=1,485로 추정치와 비교해 볼 경우 여전히 20명의 개인이 데이터베이스로부터 제거되어야 한다. 가구규모 3은 가구규모 4인 경우와 마찬가지의 조정단계를 각 성 및 연령별 그룹에서 개인을 추가 혹은 제거해주며 이러한 과정은 가구규모 2와 1에서도 반복적으로 시행됨으로써 인구 및 그 분포가 추정치와 동일하게 조정된다(Steel, 2002).

6) 설계그룹 및 지방자치단체에 대한 품질확인작업

영국 통계청은 원넘버 센서스 추정치를 행정자료 및 인구분석자료, 질적자료 등을 활용하여 지역, 성 및 연령별 비교를 통해서 진단하는 품질확인을 실시하였다(ONS, 2000b). 이 과정에서 센서스, 사후조사와 함께 제 3의 자료원으로써 행정자료를 이용한 삼원시스템 추정방법(triple system estimation)을 검토한 바 있다. 그런데, 행정자료가 개인 수준에서 정확하지 않으므로 편향된 추정을 할 가능성이 있기 때문에 이 추정방법은 고려대상에서 제외하였다. 그럼에도 불구하고 특정의 연령집단에 대해서는 행정자료가 완전한 커버리지를 제공할 것으로 평가되어 동일한 집단에 대해서 원넘버 센서스 수치와 비교함으로써 진단할 수 있었다. 예를 들어, 센서스에서 조사가 가장 어려운 그룹이었던 노인과 아동의 경우 각종 급여수령 목적의 자료 명부 등은 거의 완전한 커버리지를 제공해 주는 것으로 알려져 있다. 따라서 이러한 행정자료를 근거로 원

넘버 센서스 추정치를 평가하는 것이 가능하였다.

영국 통계청은 다양한 유형의 행정자료를 기반으로 하여 전국수준의 인구추정치에 대한 고위와 저위 변이를 추정하였다. 여기에서는 출생, 사망, 이동 수준의 변이뿐만 아니라 군인 및 망명자(refugees)와 같이 조사가 어려운 집단의 행정자료를 추가적으로 활용하였다. 이는 센서스 실시 이전에 획득가능한 최고 수준의 인구 지표로 원넘버 센서스의 추정치와 비교되었다. 비교과정에서는 행정자료와 인구분석치 뿐만 아니

〈표 3〉 보정된 데이터베이스에서 인구추정치 조정의 예

1단계

성 및 연령별 그룹(a)	가구규모별 $T_{as}^{(imp)}$				$T_a^{(imp)}$	T_a
	1	2	3	4		
1	60	100	90	200	450	455
2	40	80	200	150	470	460
3	50	120	190	210	570	550
$T_s^{(imp)}$	150	300	480	560	1490	-
T_s	161	296	468	540	-	1465

↓

a	X_a	X_{a4}	방향
1	-5	5(-5/25)=1	추가
2	10	5(10/25)=2	제거
3	20	5(20/25)=4	제거

2단계 ↓

성 및 연령별 그룹(a)	$T_{as}^{(imp)}$				$T_a^{(imp)}$	T_a
	1	2	3	4		
1	60	100	?	?	451	455
2	40	80	?	?	468	460
3	50	120	?	?	566	550
$T_s^{(imp)}$	150	300	495	540	1485	-
T_s	161	296	468	540	-	1465

자료: Steel, 2002: 504.

라 센서스와 사후조사의 현장작업에 대한 정보, 1991년 센서스 추정치의 보정에 대한 세부정보 등에 대한 기술적인 자료 또한 보조 정보로 함께 제공함으로써 질적인 분석도 가능하도록 하였다(ONS, 2001b).

라. 원넘버 센서스 결과

영국 통계청은 2002년 9월, 원넘버 센서스 결과를 공표하였다(ONS, 2005). 그 내용을 보면 센서스에서 총인구의 약 6.1%가 누락된 것으로 추정되었다. 구체적으로는 약 130만 가구(추정된 가구의 5.9%에 해당하는 규모)가 보정되었으며, 가구가 보정됨으로써 보정된 개인은 추정인구의 4.9%에 해당하는 250만 명이었다. 또한, 센서스에서 가구는 집계되었으나 가구원이 누락된 경우는 약 60만 명으로 이는 총인구 추정치의 1.2%에 해당하였다.

이러한 결과는 최단 시일에 달성할 수 있었던 최고 품질의 추정수치라고 자체 평가되었으나, 1991년 센서스의 인구집계와 출생・사망・이동 등 인구동태변수를 인구방정식에 적용하여 작성한 추정인구와 비교해 본 결과 약 110만명이라는 차이가 발생하였다. 이는 어떤 지방자치단체에서는 추정치의 향상이 있었으나, 다른 지방자치단체에 대해서는 추정치의 신뢰도가 떨어졌음을 의미하는 것이다. 이에 통계청에서는 그 차이에 대한 원인분석에 착수하였다. 처음에는 국제이동통계의 불완전성에 의한 것으로 평가내려 2003년 9월 국제인구이동통계의 수정을 통해서 추정인구를 약 30만명 가량 상향조정하였다. 이후에도 여러 차례 추가적인 작업을 하였으나, 21만명 가량이 여전히 설명이 불가능한 것으로 나타나 센서스 방법론에 대한 개선의 여지가 있음을 보여주었다

2. 미국의 원넘버 센서스

가. 원넘버 센서스 등장배경

미국 헌법은 센서스를 통해서 미국 내 거주하는 모든 인구수를 집계하도록 의무화하고 있다. 센서스인구는 각 주(state)의 의석수 및 연방정부

의 예산을 배분하는데 중요한 근거로 사용되기 때문이다. 따라서 센서스의 완전하고도 정확한 실시가 무엇보다 중요하다. 하지만, 미국 내에서 센서스를 통해 집계된 인구수가 한 번도 완전한 적은 없다(NRC, 1999).

센서스의 불완전성 문제는 센서스 초기 시점까지 거슬러 올라가며[10], 최근 몇 십년 동안 이에 대한 공공의 관심이 급격하게 증가하였다. 1980년 센서스 이후, 뉴욕시를 비롯한 몇 몇 시민단체는 과소집계된 센서스 결과를 보정하라는 소송을 제기하기도 하였다. 센서스국은 센서스를 개선할 수 있을 정도로 정확하게 보완할 방법이 없으므로 보정이 불가능하다는 판단을 내렸고 법정에서도 센서스국의 이러한 결정을 지지해주었다. 이후 센서스국은 1990년 센서스에서 과소집계를 측정하고 보정할 수 있는 방안에 대한 연구를 착수하였다. 사후조사는 자료의 보정을 위해서 중요한 방법론으로 대두되었으며, 1990년 센서스에서는 사후조사를 통해서 블록수준까지 자료를 보정하는 방법이 제안되었다. 그러나, 미국 상무국(U.S. Department of Commerce)은 보정된 센서스 수치가 의석수 배분을 목적으로 사용되어서는 안된다는 결정을 내린바 있다(Hogan, 1993). 원넘버 센서스에 대한 논의는 센서스 커버리지의 향상을 위한 투자에도 불구하고 특정 인종의 과소집계 문제가 지속적으로 나타난 1990년 센서스 이후 본격적으로 이루어졌다.

〈표 4〉 미국의 연도별 센서스 실시 비용과 순과소집계 비율

연도	비용		순과소집계		
	전체	가구	전체	흑인	비흑인
1980	22억달러	24달러	1.2%	4.5%	0.8%
1990	33억달러	32달러	1.9%	5.7%	1.3%

자료: NRC, 2004: 94를 재구성하였음.

10) 토마스 제퍼슨은 1790년의 센서스 결과를 대통령에게 보고하면서 “직접 회수된 것은 검은 색 잉크로, 추정된 것은 붉은 색 잉크로 표시하였으며, 이것은 실제에 매우 근접한 수치”라고 설명한 바 있다. 1830년 센서스에서는 이미 원래의 센서스 집계치(original count)와 수정된 집계치(corrected count)가, 1840년에는 수정된 집계치만 공표하였다. 1850년 센서스에서는 센서스 집계표를 작성함에 있어서 캘리포니아에서 누락된 부분을 센서스 이외의 자료를 통해서 추정하였다(Wright, 1999).

<표 4>는 1990년 센서스에서 소요된 비용이 1980년에 비해 총비용의 경우에는 11억달러, 가구당비용의 경우에는 8달러가 증가하였음에도 불구하고,[11] 과소집계 비율 또한 1.2%에서 1.9%로 증가하였음을 보여준다.[12] 그런데 특정 인종의 과소집계 비율의 증가폭은 더욱 컸다. 흑인집단의 과소집계는 1980년의 4.5%에서 1990년의 5.7%로 1.2%가 증가한 반면에 흑인이 아닌 집단의 커버리지 오차는 0.8%에서 1.3%로 그 증가폭이 0.5%에 그쳐서 두 집단 사이의 과소집계 차이가 훨씬 크게 벌어졌다. 센서스국은 이를 두고 1990년 센서스에서 추가적으로 들인 비용과 노력이 센서스 커버리지의 향상을 담보해 주지 못한 것으로 받아들였다(NRC, 2004).

이에 센서스국은 통계적 기법을 활용하여 누락 혹은 중복인구를 보정하여 단일의 인구를 작성하는 원넘버 센서스 개념을 채택하였다. 센서스 디자인 계획을 검토할 전문가 패널을 구성하여, 사후조사의 표본설계와 추정절차 및 현장조사, 행정자료 사용 등을 검토하였으며, 1995·1996년에 두 차례의 시험조사를 거치면서 원넘버 센서스의 단계를 차근히 밟아나갔다.

나. 원넘버 센서스 개념

1993년 초에 센서스국에서 채택한 원넘버 센서스의 개념은 “집계와 할당, 통계적 추정의 적절한 조합에 기초하여 법적인 마감시일까지 최고의 단일 인구수를 만들어 내는 것(NRC, 1994: 21)”이었다. 여기에서 집계라 함은 응답자를 직접 접촉하여 인구수를 세는 방법으로 우편조사, 방문조사, 전화조사 등을 포함한다. 할당은 현장에 대한 직접적인 확인 없이 특정한 지역적 위치에 일정 수의 사람을 배분하기 위해 행정자료를 활용하는 것을 뜻하며, 통계적 추정은 집계나 할당을 통해서도 파악되지 않은 누락인구를 보정하기 위해 사용하는 기법을 말한다.

집계, 할당, 통계적 추정의 상호의존성은 원넘버 센서스의 핵심이었

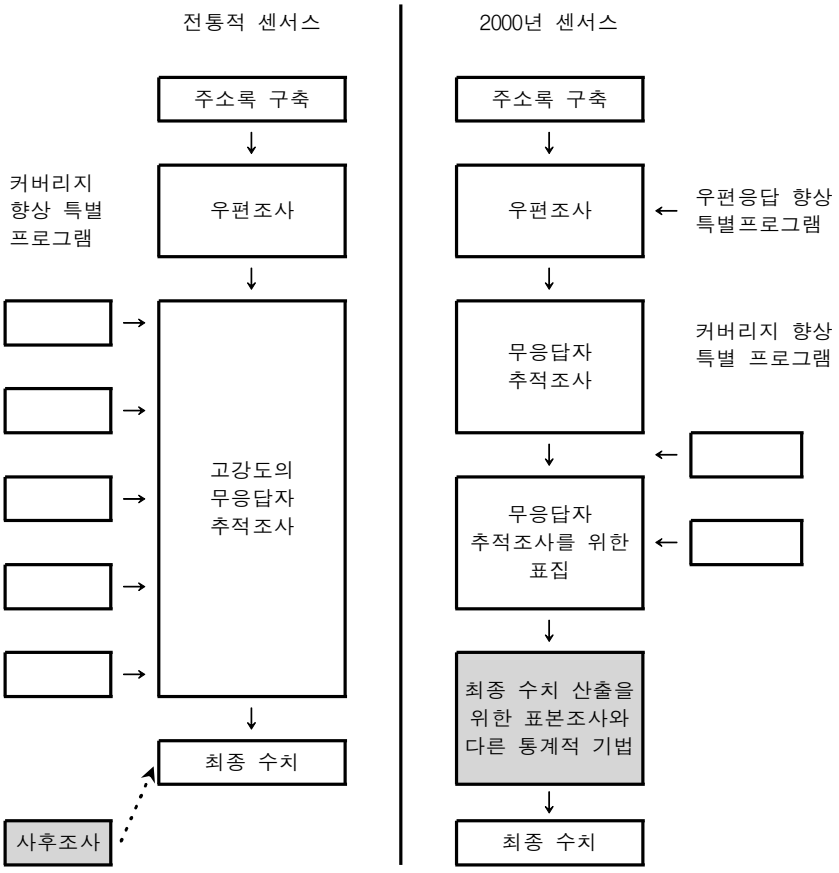
11) 2000년 센서스의 경우, 전체 비용은 66억달러, 가구당 비용은 56달러가 소요되었다. 이 시기의 커버리지 오차는 미국의 원넘버 센서스 결과 부분을 참고할 것.

12) 1980년의 사후조사 결과 분석에 의한 커버리지 오차 정보가 없으므로 인구분석에 의한 커버리지 오차를 비교하였다.

다. 원넘버 센서스와 전통적인 센서스의 가장 큰 차이점은 센서스 커버리지 측정을 위한 사후조사가 기존처럼 센서스와 독립적으로 이루어지는 것이 아니라 센서스 과정 내에서 중심적인 부분이 된다는 점이었다. 일반적으로 센서스는 ① 주소파일의 구축, ② 구축된 주소로부터 우편조사를 통한 응답얻기, ③ 초기 단계에서 응답하지 않은 가구에 대한 추적조사, ④ 커버리지 측정의 네 단계로 구분된다([그림 5]). 전통적인 센서스는 앞의 세 단계로 이루어진 반면에, 새롭게 디자인되는 원넘버 센서스에서는 마지막 단계가 센서스 과정에서 필수적인 부분으로 포함되었다. 이러한 특성을 강조하기 위해서 원넘버 센서스에서의 사후조사를 통합적인 커버리지 측정(Integrated Coverage Measurement, ICM)이라고 하였다. ICM 결과에 의해 누락 혹은 중복된 것으로 추정된 인구는 센서스 본조사에 보정되어 총인구를 작성하는데 기여하게 된다(NRC, 1994).

그런데, 원넘버 센서스 실시를 위한 준비 중에도 의회와 센서스국 간의 논쟁은 지속되었다(Brunell, 2002). 의회는 표집과 추정 등의 통계적 기법의 사용이 정치적으로 이용될 가능성과 소지역에서의 표본오차의 크기로 인해서 원넘버 센서스를 반대하는 입장에 선 반면, 센서스국은 통계적 기법을 통해서 인구를 정확하게 측정하고 비용을 통제할 수 있다는 근거를 들어서 원넘버 센서스를 지지하였다. 1997년, 백악관과 의회는 1998년의 시범예행조사에서 원넘버 센서스와 함께 센서스 본조사 이후에 사후조사를 실시하는 전통적인 센서스를 모두 테스트할 것을 지시하였다(Wright, 1999). 그러나, 그 결과와는 무관하게 1999년 1월 25일, 대법원이 표본조사를 토대로 센서스의 집계결과를 보정한 추정인구를 기반으로 의석수를 배분하는 것은 연방 법률의 정신에 위배된다고 판단함에 따라서 원넘버 센서스 계획의 변경이 불가피하였다.

이러한 결정에 따라서 센서스국은 사후조사를 재설계하였다. 의석수 배분이라는 본래의 목적을 달성하기 위해 당초 주 수준에서의 인구 추정을 목적으로 설계되었던 것은 전국 및 주요 하위 지역 수준에서 인구영역에 대한 신뢰할 만한 추정치의 생산이라는 목적으로 변경되었다. 그리고 이 수치는 의석수 배분 이외의 목적에 활용하고자 하였다. 표본규모도 75만 가구에서 30만 가구로 축소되었으며, 사후조사의 명칭 또한



자료: NRC, 1995:77

[그림 5] 미국의 전통적 센서스와 2000년 센서스 계획 비교

정확성 및 커버리지 평가조사(Accuracy and Coverage Evaluation Survey; ACE)로 바뀌었다(Anderson and Fienberg, 2001). 그러나 목적이나 표본규모, 명칭 등이 바뀌었음에도 불구하고, 소지역 수준에서 인구를 추정하여 데이터베이스를 보정한다는 원넘버 센서스의 기본 골격은 유지되었다.[13] 그러므로 보정수치의 공표여부와는 무관하게, 보정수치가 나오기까지의 과정을 검토해 보는 것은 영국의 원넘버 센서스 방법론과 비교

13) 물론 [그림 5]에서 보여주고 있는 2000년 센서스 디자인이 2000년도에 그대로 시행되지는 않았다. 예컨데, 무응답자 추적조사를 위한 표집 등은 제외되었다.

해 본다는 점에서 의의가 있다.

ACE를 도입하기 전 센서스국은 실행가능성(feasibility)을 평가했다. 실행가능성이란 운영적인 부분(operational feasibility)과 기술적인 부분(technical feasibility)으로 나뉘었다. 전자는 블록수준에서 통계적으로 보정된 수치를 마감기한까지 제공할 수 있느냐의 여부이며, 후자는 통계적으로 보정된 수치가 과연 센서스 자료의 전반적인 정확성을 향상시켜 줄 수 있느냐이다. 두 가지 실행가능성은 시범예행조사를 통해서 검증되었다(U.S. Census Bureau, 1999).

다. 원넘버 센서스[14] 방법론

[그림 6]에서 보듯이 원넘버 센서스 방법론은 표본설계, 표본가구의 주소록 구축, 센서스 주소록과의 가구매칭, 표본가구 내 개인면접, 센서스와 사후조사의 개인매칭, 추정 및 보정 단계로 구분되었다(U.S. Census Bureau, 2004).

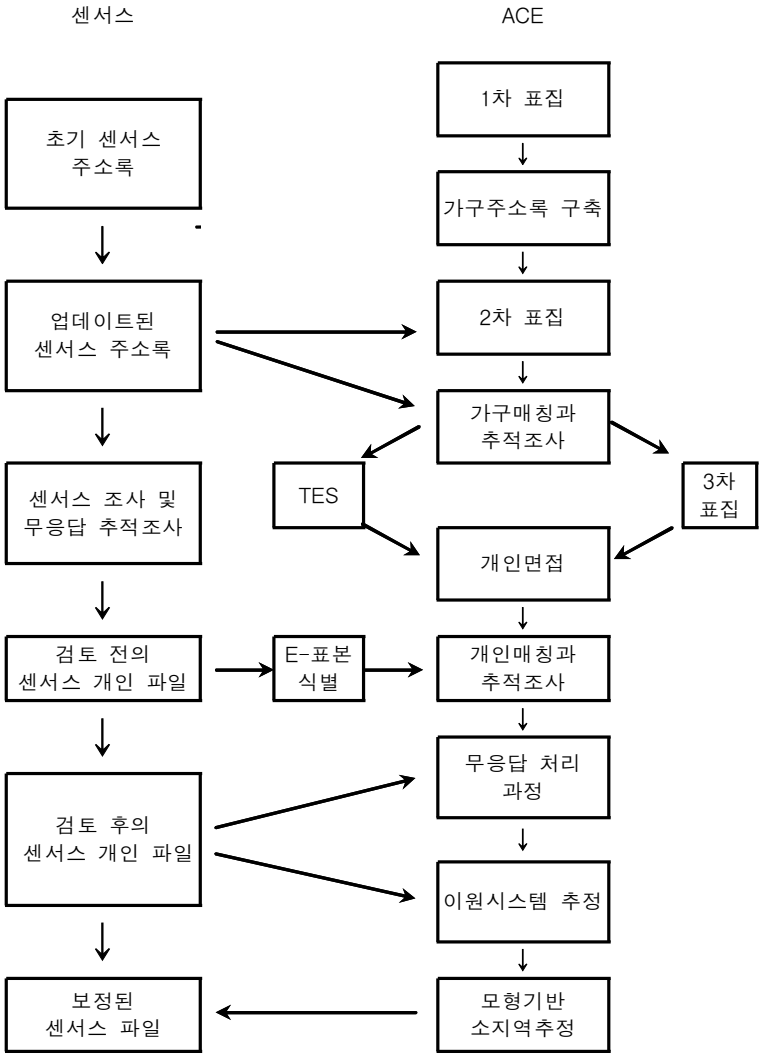
1) 표본설계

사후조사는 P-표본과 E-표본의 두 부분으로 구성된다. 개념적으로 P-표본과 E-표본은 동일하다. 다만, P-표본은 센서스와 독립적으로 수행된 사후조사에서 응답한 표본을, E-표본은 P-표본과 같은 지역에서 응답한 센서스를 뜻한다. 센서스와 사후조사의 매칭과 조정을 거친 이후에 P-표본은 센서스에서 누락된 개인에 대한 비율을, E-표본은 중복 혹은 잘못 포함되어 조사된 비율을 추정하며, 두 오차율을 합하여 커버리지 오차를 산출한다.

표본설계는 기본적으로 ICM의 영향을 받았다. 왜냐하면 ICM의 진행이 중단된 1999년에 이미 ICM 표집단계에 들어갔을 뿐 아니라 시간적으로도 ACE 표본설계로 변경하는 데에는 무리가 있었기 때문이다. 따라서, ACE 표집은 다단계 설계로 발전되었다.

---

14) 2000년 센서스가 원넘버 센서스는 아니나 그 방법론이 원넘버 센서스의 골격을 유지하고 있다는 점에서 본 연구에 한해서만 원넘버 센서스로 부르고자 한다. 또한 영국과 마찬가지로 누락과 중복을 측정하기 위한 표본조사를 ACE가 아닌 사후조사로 부르고자 한다.



자료: U.S. Census Bureau, 2004.

[그림 6] 미국의 2000년 센서스 흐름도

1단계에서는 ICM 표본을 추출한 후 이들에 대한 가구리스트를 독립적으로 구축하였다. 여기에서 중요한 것은 일차표집단위(primary sampling unit)에 대한 개념정의와 일차표집단위의 추출이었다. 일차표집단위는

지리적으로 인접한 하나 이상의 센서스 블록을 묶은 블록클러스터(block cluster)이다. 블록클러스터는 정확성과 함께 비용 및 조사원의 효율적인 현장업무 관리를 위해 평균적으로 30가구가 되도록 하였으며, 물리적인 경계가 명확한 지역을 중심으로 형성하였다.[15] 일차표집단위는 각 주 내에서 이전 센서스 주소록으로부터 나온 가구 수에 따라 0-2가구로 구성된 소규모 클러스터, 3-79가구로 구성된 중간규모 클러스터, 80가구 이상으로 구성된 대규모 클러스터, 3가구 이상으로 구성된 어메리칸 인디언 구역(American Indian Reservation)으로 층화되었다. 각 층 내에서 표본들을 정렬한 이후에 계통추출을 하고 표본들의 주소록을 독립적으로 구축하였다.

2단계에서는 1단계에서 추출된 중간 및 대규모 클러스터, 그리고 소규모 클러스터가 독립적으로 재층화되었다. 중간 및 대규모 클러스터는 블록클러스터 내의 인구학적 특성, 사후조사 표본 가구주소록과 업데이트된 2000년 센서스 주소록 사이의 가구 수 등에 따라서 각 주 별로 다시 층화하였다. 이 결과 발생하게 되는 하위층을 가리켜 축약된 층(reduction strata)이라고 하고, 이 층들을 정렬한 후 계통추출 하였다. 소규모 클러스터 또한 사후조사 가구주소록과 업데이트된 2000년 센서스 주소록 사이의 가구수를 사용하여 층화한 후 계통추출하였다. 면접 및 추적조사의 가구 당 비용을 고려하여 중간 및 대규모 클러스터 보다 소규모 클러스터에서 낮은 표집 비율을 적용하였다. 한편, 1단계의 어메리칸 인디언 구역은 2단계에서 모두 추출하였다. 이것은 어메리칸 인디언 구역에 대한 충분한 표본을 확보하여 커버리지를 추정하기 위한 것이다.

3단계에서는 대규모 블록클러스터를 하위표집하였다. 추출된 블록클러스터가 79가구 이하일 경우에는 클러스터 전체가 사후조사 표본으로 구성되는 반면에, 80가구 이상의 대규모 클러스터일 경우에는 하위표집을 실시하였다. 자료수집 비용을 줄이고, 조사원 업무량을 효율적으로 관

15) 일차표집단위로써 블록을 더 작은 단위로 나눈 단위(블록의 반 정도에 해당하는 크기)가 대안으로 고려되기도 하였다. 그러나 이 안은 배제되었다. 블록을 더 작게 나눈 단위가 표본오차는 줄여줄 수 있으나 매칭에 대한 업무 부담과 조사원의 이동거리를 증가시킴으로써 비용을 증가시킬 뿐 아니라 일차표집단위 간 경계 식별의 어려움으로 인해서 매칭오류를 증가시켰기 때문이다.

리하기 위해서 인접가구끼리 묶어서 구역(segment)을 형성하였다. 이는 정확성에 영향을 미치지 않고 높은 급내 상관을 유지하는 범위 내에서 이루어졌다. 구역이 나뉘어진 이후에는 각 클러스터 내 계통추출을 통해서 최종적으로 11,303 블록클러스터를 추출하였는데 이는 약 301,000가구에 해당하는 규모이다.

2) 현장조사와 매칭

자료수집단계는 가구주소록 구축, 가구매칭, 개인면접, 개인매칭으로 나뉘었다.

가) 가구주소록 구축

조사원은 1차 표집단계에서 추출된 블록클러스터 표본에 방문하여 모든 가구의 주소록을 독립적으로 작성하였다. 이 주소록을 기반으로 2, 3차 표집을 하며, 최종적으로 P-표본을 구축하였다.[16] 가구주소록은 기본거리주소(basic street address)에 따라서 작성하고, 이 주소는 거점번호(map spot number)로 할당되어 사후조사 지도에 기록하였다. 가구원 혹은 아파트 관리자와 같은 대리인으로부터 가구정보를 수집한 후 해당 주소에 위치하고 있는 가구 수 및 주택구조를 기록하였다. 구축된 주소록은 자료 매칭을 위해서 국립자료센터(National Processing Center)에 전달되어 정리되었다.

나) 가구매칭

가구주소록이 구축된 이후에는 센서스 매스터주소파일의 가구와 매칭하였다. 이것은 다음 단계인 개인면접의 대상이 되는 가구를 식별하기 위한 것인 동시에 지오코딩(geo coding) 에러를 찾아 주소록을 보완하기 위한 목적에서 수행되었다. 가구매칭은 컴퓨터매칭, 수동매칭, 매칭되지 않은 가구에 대한 추적조사, 추적조사 이후의 코딩작업으로 구분되었다.

컴퓨터매칭은 2차 표집이 완료된 이후에 수행되었으며, 매칭이후 어

16) 센서스 기준 시점에 존재할 것 같은 모든 가구에 대한 주소록을 만들기 위해서 잠재적인 가구, 예컨데 공사중이거나 앞으로 거주 예정인 가구 등도 모두 포함하였다.

느 한 쪽의 가구 주소라도 빈 칸으로 되어 있거나 주소가 표준화되어 있지 않을 경우에는 수동매칭 단계로 넘어갔다. 수동매칭은 일반요원(clerks, 115명), 기술요원(technicians, 46명), 전문요원(analysts, 10명) 순으로 작업이 이루어졌다. 일반요원에 의해서 매칭된 모든 작업은 인정할 만한 수준의 품질이 지속적으로 유지될 때까지 기술요원의 검토를 받고, 그 이후에는 계통추출을 통해서 뽑힌 주소를 대상으로만 자료의 품질을 검토하였다. 이 과정은 컴퓨터 화면을 통해서 보여지며, 만일 수동매칭작업의 결과가 일정 품질 이하의 수준으로 떨어질 경우에는 기술요원에 의해서 재검토되었다. 기술요원의 수동매칭결과는 전문요원에 의해서 같은 방식으로 품질검토가 이루어졌다.[17] 이러한 품질검토과정은 앞서 살펴본 영국과 매우 유사하다. 이는 영국이 미국의 사례를 벤치마킹하였기 때문이다.

수동매칭 이후에 매칭되지 않은 가구나 센서스와 사후조사에서 동일한 주소로 추정되어 향후 매칭 가능성이 있는 가구, 센서스 내 혹은 사후조사 내에서 주소가 중복된 것으로 추정되는 가구 등에 대해서도 현장추적 조사를 하였다.[18] 현장조사를 통해서 획득된 정보를 활용하여 다시 수동매칭하였다. 최종적으로 가구가 아닌 것, 중복된 것, 지오코딩 에러를 제거하였으며, 가구는 매칭된 것과 매칭되지 않은 것, 그리고 지위가 확정되지 않은 것(unresolved)으로 분류되었다.

다) 개인면접

가구매칭과 현장추적조사, 그리고 3차 표집단계가 끝나면 P-표본이 형성되고, 이 표본들에 대한 면접을 실시하였다. 면접의 목적은 센서스 시점과 사후조사 시점에 가구에 살고 있는 모든 개인에 대한 가구 명부(household roster)를 구축하기 위한 것이다.

개인면접을 통해서 거주지위(residency status)와 이동자 지위(mover status) 등을 확정함으로써 P-표본을 식별할 수 있다. 면접을 통해서 센

17) 2000년 센서스 사후조사의 목적 중 하나는 수동매칭과정에서 종이를 사용하지 않는 것이었다. 이에 모든 자료를 컴퓨터에서 볼 수 있게 하였으며, 이는 상당한 시간을 단축시켜 준 것으로 평가되었다.

18) 주소록 구축과정에서 공사중, 건축예정 등으로 표시한 가구의 경우, 가구정의에 적합한지 현장추적을 통해서 재차 확인하는 절차를 거쳤다.

서스 시점과 사후조사 시점에 누가 거주했는지를 묻고 이를 기반으로 해서 이동자 지위를 결정하였다. 이것은 이동자를 추정하기 위한 것이다.[19] 이동자 지위는 센서스 시점과 사후조사 시점에 모두 거주하고 있는 비이동자(nonmover), 센서스 시점에는 거주하였으나 사후조사 시점에는 이사를 간 전출자(outmover), 센서스 시점에는 없었으나 사후조사 시점에 이사를 온 전입자(inmover)로 구분되었다.

자료수집은 컴퓨터를 이용한 전화면접(Computer Assisted Telephone Interviewing, CATI)과 방문면접(Computer Assisted Personal Interviewing, CAPI) 방식으로 이루어졌다. 가구매칭을 통해서 센서스 조사 당시 전화번호가 수집된 것으로 확인된 가구에 대해서는 CATI를 먼저 시도하였고, 그렇지 않은 가구에 대해서는 CAPI를 수행하였다. 방문면접의 경우, 최대한 가구원에게 응답을 얻도록 하였으며 대리응답은 3주 이후에 허용하였다.[20] 총 업무량 중 약 29.4%가 CATI에 의해서 수행되었다. 해당 조사방법에 고용된 조사원 현황은 <표 5>와 같다.

〈표 5〉 미국의 2000년 사후조사 조사원 현황

(단위: 명)

	CATI	CAPI	추적면접
조사원	450	4,502	4,470
크류리더	794	836	714
슈퍼바이저	189	186	184

자료: U.S. Census Bureau, 2004.
주: 전화면접의 경우, 크류리더와 슈퍼바이저도 조사를 함께 수행하였음.

조사표는 가구의 이동지위와 응답자의 유형(가구원 vs. 대리응답)에 따라서 세가지로 구분되었다. 먼저, 가구원 전체가 이사를 간 전출가구와 그 외의 모든 경우로 조사표 양식이 분리되었다. 전출가구인 경우에

19) 사후조사는 전입자를 이용하여 사후층 내에서 P-표본 이동자 수를 추정하는 반면에 전출자를 이용하여 이동자의 매칭율을 추정하였다.

20) 무응답을 제외하고 최종적으로 면접한 결과(264,103가구)중 가구원에 의한 면접은 94.6%(249,854가구)이며 대리인에 의한 면접은 4.7%(12,317가구), 나머지는 부분면접(sufficient partial interview)이 이루어진 것이었다.

는 현 표본지역에 살고 있는 대리인으로부터 전출자에 대한 정보를 얻어내기 위해 별도로 설계된 조사표를 활용하기 때문이다. 다음으로, 전출가구 이외의 모든 경우에 해당하는 가구는 응답자가 가구원인 경우와 대리인인 경우로 조사표가 구분되었다. 여기에는 가구원 전체가 비이동자인 경우와 가구원 전체가 전입자인 경우, 한 가구 내에 비이동자・전입자・전출자가 혼합된 경우가 포함되었다.

면접이 완료된 이후에는 품질확인을 위해서 리인터뷰(reinterview)를 실시하였다. 조사원 면접량 중에서 5%를 무작위로 뽑는 한편, 각 지역사무소에서 슈퍼바이저가 허위작성 등의 징후를 보이는 표본을 선정하여 면접을 하였다.

라) 개인매칭

면접이 완료된 이후에 사후조사와 센서스 조사 시점에 해당 가구에 거주했던 모든 가구원 명부가 구축되면 E-표본의 식별이 가능해졌다. 이에 따라서 P-표본과 E-표본 개인매칭을 하였다. 매칭의 결과는 최종 인구추정치에 직접적인 영향을 미치므로 정확하게 이루어져야 한다.

매칭은 컴퓨터매칭과 수동매칭으로 이루어졌다. 컴퓨터매칭으로 P-표본이 E-표본과 매칭되고, 매칭되지 않은 P-표본은 E-표본이 아닌 센서스 지역 전체로 확대되어 매칭이 시도되었다. 이 때 다음의 두 단계를 거쳤다. 첫번째는 레코드의 순위 매기기(record pair ranking), 두번째는 매칭 절사점(match cutoffs)의 결정이다. P-표본과 E-표본을 문자열 비교 방법(string comparison)을 사용해서 표준화된 이름을 비교, 각 쌍별로 순위를 매긴 후 최적의 쌍을 선정하였다. 클러스터 내 최적의 쌍은 매칭과 비매칭에 대한 절사점을 결정하기 위해서 사용되었다. 매칭오류를 방지하기 엄격하게 매칭 절사점이 선정되었으며, 매칭 절사점 이상에 있는 모든 쌍은 매칭된 것으로, 비매칭 절사점 이하에 있는 모든 쌍은 매칭되지 않은 것으로, 둘 사이에 있는 것은 매칭 가능한 것으로 분류하였다.

컴퓨터매칭 이후에 매칭되지 않은 P-표본은 수동검토 되었는데 그 절차는 가구매칭과정과 유사하였다. 235명의 일반요원과 46명의 기술요

원, 10명의 전문요원이 수동매칭에 가담하였으며, 일반요원 이후에 기술요원에 의해서 품질확인이 이루어지고, 기술요원이 검토한 것은 다시 전문요원이 검토하였다.

3) 무응답 자료 처리

사후조사에서는 세 가지 유형의 무응답 보정이 실시되었다. 접촉이 어렵거나 응답을 거절하여 가구전체가 면접을 하지 못한 경우에 대한 무응답 보정(noninterview adjustment), 특정 항목이 누락되었을 경우에 대한 무응답 보정(characteristic imputation), 매칭결과 및 거주지위 등이 확정되지 않은 경우에 대한 무응답 보정이다.

첫째, 가구수준에서의 무응답 보정은 이동지위에 따라서 다른 보정가중치가 적용되었다. 한 가구에 대해서 센서스 시점의 가구명부와 사후조사 시점의 가구명부가 구축되었는데, 이동지위에 따라서 가중치가 다르게 적용되었다. 비이동자와 전출자에게는 센서스 시점의 가구지위에 기반한 가중치가, 전입자에게는 사후조사 시점의 가구지위에 기반한 가중치가 적용되었다(Cantwell and Ikeda, 2003).

둘째는 항목무응답 보정에 관한 것이다. 연령, 성, 가구소유여부, 인종은 이원시스템 추정에서 중요한 변수이므로 해당 항목에 무응답이 있을 경우에는 보정을 해야 한다. 그런데 항목의 유형에 따라서 보정방법에 차이가 있다. 주택소유여부나 인종은 지리적으로 인접한 사람들 간에 상관관계가 높으므로 최근방 핫덱 절차(nearest-neighbor hot-deck)를 사용하여 보정을 하는 한편, 성 및 연령의 경우에는 지리적으로 덜 군집화되어 있으며, 배우자 유무 등과 같은 변수와 관계가 있으므로 해당 지역의 인구학적 분포에 기반하여 보정을 하였다.

셋째, 매칭작업의 결과 역시 추정단계에서 사용되므로 매칭여부, 정확한 조사여부, 거주지위 등이 결정되어야 한다. 매칭여부, 정확한 조사여부, 거주지위 등이 컴퓨터 및 수동매칭 혹은 현장 추적을 통해서도 확정되지 않을 경우 추적면접을 통해서 파악된 산출물의 분포에 기반하여 확률을 추정하였다.

4) 추정

가) 이원시스템 추정모델

사후조사로부터 2000년도 센서스의 커버리지를 추정하는데 이원시스템 추정법이 사용되었다. 영국의 원넘버 센서스에서 이미 설명하였듯이, 이 추정법은 센서스와 사후조사 사이의 독립성과 동질성 가정을 전제로 한다. 센서스와 사후조사의 독립적인 운영을 통해서 두 조사간의 인과적인 의존성을 최소화하였으며, 조사될 확률에서의 이질성과 관련이 있다고 판단되는 특성에 기반하여 사후층화를 하고 이 층 내에서 이원시스템 추정치를 계산함으로써 동질성을 확보하고자 하였다.

사후층화 변수는 성, 연령, 인종(race/hispanic origin), 가구소유여부, 지역, 메트로폴리탄 통계구의 규모, 트랙 수준에서의 센서스 회수율 등 9개이다. 위 변수들 중에 성과 연령을 제외, 각 변수에서 기대되는 표본크기를 설정하고, 이에 미치지 못할 경우에는 사전에 범주를 합하였다. 이에 따라서 64개의 그룹이 발생하였다. 이 그룹을 다시 7개의 성 및 연령으로 나누면 448개의 사후층화가 발생하며, 이 층 내에서 이원시스템 추정치가 계산되었다.

미국 센서스국은 이원시스템 추정법의 기본모델(<표 1>)을 좀 더 정교화하였다. E-표본에서 잘못 조사된 것으로 파악된 레코드나 센서스와 매칭하여 불충분한 정보를 가진 것으로 파악된 P-표본의 레코드가 수정되었다. 이 기준을 충족시키는 센서스 레코드를 일컬어 자료에 의해서 정의된 개인(data-defined person)이라고 하였으며,[21] 이원시스템 추정법의 기본모델은 다음과 같이 재정의되었다.

$$DSE = DD \times \frac{CE}{N_e} \times \frac{N_p}{M}$$

여기에서,

DD 는 사후조사 매칭을 위해 자료에 의해서 정의된 인구수

CE 는 E-표본에서 정확하게 조사된 것으로 추정된 수치

N_e 는 추정된 E-표본수

21) 센서스 항목 중 100% 모두 응답해야 하는 항목에 2개 이상 응답한 개인을 뜻한다.

N_p 는 추정된 P-표본수
M 은 센서스와 매칭된 P-표본으로부터 추정된 인구수

이원시스템 추정치를 센서스 집계결과(C)로 나누어 주면, 센서스 집계치가 실제인구를 얼마나 적게 혹은 많게 집계하였는지를 알 수 있다. 이를 커버리지 수정요인(Coverage Correction Factor, CCF)이라고 하며, 그 값이 1에 가까울수록 센서스 커버리지가 완전하다는 것을 의미한다.

$$CCF = \frac{DSE}{C}$$

커버리지 수정요인은 사후층 내에서 모두 동일하다는 가정을 전제하고 있다. 즉 사후층 내의 모든 블록 간에 커버리지 수정 요인은 똑같다. 그러므로 커버리지 수정요인을 이용하여 블록 수준에서 인구수를 추정하는 것이 가능하게 된다.

나) 소지역에 대한 모델 기반 추정

지금까지의 과정을 통해서 산출된 추정치는 두 가지 목적에 기여한다. 첫째, 센서스 품질에 대한 정보를 제공함으로써 자료 분석 시 도움을 주거나 향후 센서스 절차를 개선하는데 사용된다. 둘째, 만일 그 결과치가 적정하다면, 센서스 집계치를 보정할 수 있다. 그런데, 두 번째 목적을 수행하기 위해서는 부가적인 추정절차가 더 요구된다. 30만 정도의 표본규모로는 대부분의 주, 카운티, 시티 등에 대해서 직접추정을 하기가 어렵기 때문이다.

(1) 합성추정법

블록, 트랙, 의회구역(congressional districts)과 같은 소지역에 대한 커버리지 보정을 위해서는 합성추정법이라는 간접적인 추정기법이 사용되었다. 영국에서도 이 방법을 적용하여 지방자치수준에서의 인구를 추정하였다. 합성추정법에서는 소지역이 대지역과 동일한 특성을 갖고 있다는 가정을 하므로, 하위지역의 인구추정치 합은 상위지역의 인구추정치와 동일하다. 그러므로 블록수준의 인구추정치는 어떠한 지역수준에서도 합역될 수 있다.

$$\widehat{N}_{ig}^{S} = C_{ig} \times CCF_i \quad \cdots\cdots \text{식 1)} \qquad \widehat{N}_{g}^{S} = (\sum_i C_{ig} \times CCF_i) \quad \cdots\cdots \text{식 2)}$$

C_{ig}는 지역 g에서 사후층 i에 대한 센서스 집계치, CCF_i는 사후층 i의 커버리지 수정요인이다. 그러므로 지역 g에서 사후층 i에 대한 합성추정치는 첫 번째 식과 같으며, 지역 g에서 모든 사후층의 합성추정치 합은 지역 g에서 총인구에 대한 합성추정치를 만들어낸다(식 2).

그런데, 어느 지역에서의 합성추정치는 정수가 아닐 수도 있다. 제어라운딩(controlled rounding) 방법은 블록수준의 집계표에서 보정된 정수값(inter-valued adjusted)을 산출하는 목적으로 사용되었다. 제어라운딩 프로그램은 한 지역수준에서의 사후층화와 더 낮은 지역 수준에서의 총수로 구성된 두 차원의 매트릭스를 사용하여 정수값으로 라운딩하는 방법이다.

라운딩과 보정은 지역수준에 따라서 단계별로 진행되었다. 트랙수준에서 합성추정치가 라운딩・보정되고, 카운티와 주수준으로 올라가면서 이 절차를 반복하였다. 그 결과, 블록, 트랙, 카운티에서 라운딩・보정된 합성추정치는 서로 일관적이게 되며, 최종적으로 주수준에서 합성추정치의 모든 합은 전국수준에서의 총인구추정치와 같게 된다.

(2) 개인 레코드의 복사

블록수준에서의 합성추정치는 센서스 집계치와 비교되었다. 만일, 합성추정치가 센서스 집계치보다 크면, 이것은 센서스에서 과소집계된 것을 의미하였다. 그리고 그 차이만큼 개인 레코드를 복사하여 해당 블록에 집어 넣었다. 이 때 과소집계된 개인의 레코드는 사후층 i의 블록 b에 있는 개인의 레코드 중에서 복원(replacement)없이 무작위로 선택된 것이다. 그 레코드 각각은 집계표에서 +1의 유효가중치(effective weight)를 가짐으로써 상향보정(upward adjustment)된 결과를 얻는다.

이와는 반대로, 센서스 집계치가 합성추정치 보다 크면 이것은 과대집계된 것으로 과대집계된 만큼의 개인 레코드를 블록 b에 있는 레코드들 중에서 복원없이 무작위로 선택하여 복사하였다. 복사된 레코드 각각은 집계표에서 -1의 유효가중치(effective weight)를 가지면서 하향보정(downward adjustment)된 효과를 갖는다.

그런데 보정된 개인은 가구가 아닌 특별범주(special category)에 넣어졌다.[22] 센서스 개인 레코드에는 개인 속성과 가구 속성이 모두 있으나, 보정된 개인에는 가구 속성이 부여되지 않았다. 전수조사의 항목에서 다루어지고 있는 개인의 속성은 갖고 있지만 가구주와의 관계 정보가 제거되어 가구와는 연결되지 않았다. 따라서 보정된 개인은 집단시설(group quarters)과 같은 범주에 속하는 개인과 같게 되었다. 그리고 이 개인은 [그림 7]과 같이 비가구(Non-Household)의 범주에 있는 개인으로 분류되었다(Ikeda and Tsay, 2003).[23]

각 블록별:

가구 범주	비가구 범주
가구주와의 관계, 개인 속성 정보를 포함하고 있는 개인과 가구 자료	개인 속성 정보만 있는 자료 집단시설 자료

[그림 7] 미국의 2000년 센서스 전수조사 파일의 예

(3) 소지역에 대한 분산추정

최종적으로는 합성추정치에서 발생하는 오차를 추정해야 한다. 그런데, 소지역에서의 수많은 합성추정치 각각에 대해 표준오차를 나열해 주는 것은 가능하지 않으므로 대신에 분산에 대한 일반화된 계수(generalized coefficient of variation)를 제공하는 전략을 취하였다. 사용자들은 추정치와 추정치의 표준오차에 대한 비율인 일반화된 계수를 통해서 어떤 추정치에 대한 표준오차를 짐작할 수 있다.

라. 원넘버 센서스 결과

위의 과정을 통해서 사후조사 추정치는 기한 내인 2001년 2월에 공표 되었다. 그런데, 보정된 수치가 정확하지 않다는 평가에 따라 당초의

22) 그러므로 가구당 개인의 수는 커버리지 수정요인에 따라서 변하지 않는다.

23) 자료 이용자들은 이러한 인공적인 범주에 익숙치 않았다. 보정된 개인은 가구에 있는 실제 개인인데, 보정된 개인을 계산하기 위해서 사용된 추정치는 가구에 있는 개인에 대한 추정치이기 때문이다. 많은 이용자들은 이러한 보정된 개인의 가구 특성을 제공해 줄 것을 요구하였다. 2001년 3월에 ACE 추정치가 기각된 이유 중의 하나는 바로 거주지위(residency status)의 부적절성이었다(Ikeda and Tsay, 2003).

목적과 달리 인구추정을 위한 기준인구로는 사용되지 못하였다. 사후조사 결과로는 약 1.18%(약 330만 명) 과소집계된 것으로 추정되었으나, 인구분석 결과에 따르면 센서스 집계치는 오히려 0.65% 과대집계된 것으로 추정되었다. 이에 센서스국은 사후조사 결과 센서스에서 중복으로 조사된 부분을 결과에 반영시키고, 상관편향(correlation bias)를 보정해서 센서스인구가 0.48% 과대집계 되었다는 2차 수정안을 발표하였다. 그럼에도 불구하고 여전히 상관편향, 합성추정치편향(syntehtic estimation bias), 어린이 집단의 커버리지에 대한 센서스와 인구분석 자료의 차이 등의 문제가 남았다. 최종적으로는 보정한 결과를 인구추정의 기준인구로 사용하지 않는다는 결정을 내렸다(Robinson and Adlakha, 2002). 이러한 결정이 우리에게 시사하는 바는 사후조사가 많은 오차에 노출되어 있기 때문에 사후조사 자료를 기반으로 한 인구추정치에 대한 신중한 접근이 필요하다는 것이다.

오차는 크게 표본오차와 비표본오차로 구분된다. 표본오차는 전수에서 표본을 추출하여 추정치를 만듦으로써 발생하는 표집분산에 의한 체계적인 오차로 계량화가 가능하다. 반면에 비표본오차는 표본오차 이외의 모든 오차로 자료수집 및 조사과정에서의 오차, 무응답 자료의 처리 과정에서의 오차, 충분하지 않은 정보를 활용하여 매칭을 할 경우 발생하는 오차, 오염오차(contamination error), 상관편향, 합성추정치편향 등이 있다([그림 8]). 비표본오차의 경우 계량화에 어려움이 있기 때문에 오차를 알 수 없다는 점에서 문제가 더욱 심각하다고 하겠다(U.S. Census Bureau, 2004).

일부에서는 센서스 보정결과가 분포에서의 정확성, 즉 하위집단간 차별적인 과소집계의 문제는 상당부분 완화시키고 있다고 평가하고 있다. 추정인구의 기준인구로 활용되지는 못할 지라도 많은 표본조사에서 통계 모집단 수치로 활용될 가능성이 있음을 의미하는 것이었다. 이는 1990년 센서스의 예를 통해서 알 수 있다. 1990년 역시 보정된 결과를 인구추정치의 기준으로 삼지 않기로 결정하였으나 센서스 자료의 특정한 부분에서는 정확성이 향상되었음을 인정하고, 연방표본조사(federal sample survey)의 스폰서에게 보정된 센서스 결과를 이용한 인구추정치를

P-표본 매칭오류와 E-표본 절차(processing)오류: P-표본에서 발생하는 매칭오류는 P-표본과 E-표본을 매칭하기 위해 센서스 기록을 찾는 과정에서 발생한다. P-표본은 잘못 매칭될 수도 있고 혹은 잘못 매칭된 것 그 자체가 잘못된 것일 수도 있다. E-표본 절차오류는 E-표본이 정확하게 조사되었는지 혹은 잘못 조사되었는지의 여부를 분류하는 과정에서 발생한다. 이 오류는 정확히 조사된 것을 잘못 조사된 것으로 혹은 그 반대의 경우로 인해서 발생한다.

P-표본과 E-표본 자료수집오차: 자료수집과정에서 또한 오차가 발생할 수 있다. 조사원이 질문을 하는 과정에서 혹은 응답자가 답을 하는 과정에서, 또는 자료의 코딩 과정에서도 오차가 발생한다. 이로 인해서 이동자의 지위나 거주지위 혹은 매칭지위 등이 정확하지 않게 할당되며, 추정치의 왜곡이 발생한다. 또 다른 유형의 오차는 지오코딩에 관한 것으로, P-표본 혹은 센서스 가구주소록을 구축하는 동안에 특정 가구를 다른 지역에 코딩함으로써 오차가 발생하기도 한다.

무응답 자료: 조사과정에서는 여러가지 유형의 무응답이 발생하기 마련이다. 어떤 가구는 아예 면접을 하지 못할 수도 있으며, 또 다른 가구는 특정 항목에 대해서 응답을 하지 않기도 한다. 또한 매칭여부와 같은 분류 자체가 불명확하기도 하다. 이에 따라 각각의 무응답에 대해서 보정을 하게 되는데 이 무응답 보정으로 인해 오차가 생길 수 있다. 예컨대 매칭여부에 대한 무응답 보정은 추적조사를 통해서 파악된 산출물의 분포에 기반해서 추정된 확률에 따라서 이루어지는데 이 때 추적조사의 비율이 너무 낮기 때문에 오차가 발생할 수 있다.

자료: U.S. Census Bureau, 2004.

[그림 8] 사후조사의 비표본 오차 종류

상관편향: 같은 사후층 내에서 개인이 집계될 확률이 다르며 그들 사이에 의존성이 있다면, 상관편향이 발생한다. 상관편향은 추정치를 과대 혹은 과소추정하는 것으로 알려져 있으나, 일반적으로 이원시스템 추정치를 과소추정한다. 왜냐하면, 두 조사 모두에서 포함되지 않는 개인들이 있기 때문이다. 상관편향은 사후조사 결과 보정된 자료에서의 성비(sex ratio)와 인구분석 결과에서의 성비 비교를 통해서 파악할 수 있다. 그러나 이러한 비교는 전국적인 수준에서는 유효하나 소지역 수준에서는 적절하지 않다.

오염오차: 주어진 블록클러스터 내에서 사후조사 표본으로 선택된 것이 센서스 본조사의 수행과정이나 조사결과에 영향을 미칠 수도 있다. 사후조사 표본 주소록을 구축하기 위해서 해당 가구의 거주자와 접촉하는 것이 센서스에 응답하지 않도록 하는 원인이 될 수도 있다. 왜냐하면 이 접촉 자체를 센서스에 대한 응답으로 생각할 수 있기 때문이다.

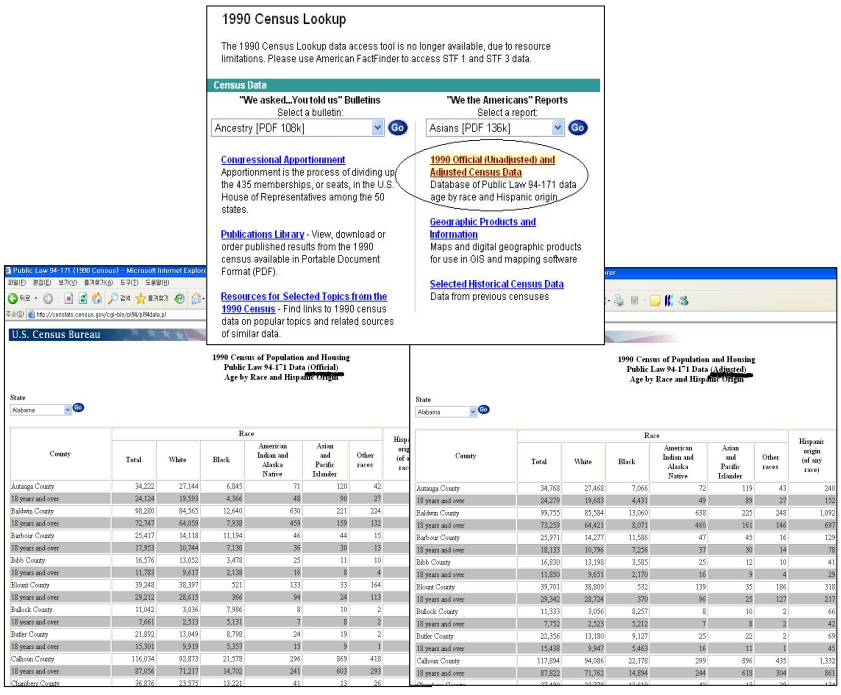
합성추정치 편향: 사후층 내에서의 커버리지 수정요인이 서로 다른 블록에서도 동일하다는 가정 하에 합성추정법을 적용하였다. 실제로 조사지역이 다름에도 불구하고 동일한 사후층에 속함으로써 동일한 수정요인을 갖게 되는 것이 편향의 원인이 되기도 한다.

균형오차(balancing error): 균형오차는 P-표본 매칭과 E-표본의 정확한 조사 여부를 찾기 위해서 탐색지역을 설계하는 과정에서 발생하는 오차이다. 매칭과정에서 전체 센서스 지역을 모두 탐색할 수 없기 때문에 탐색지역을 제한했는데, 그것이 충분한 매칭 정보를 수집하는데 방해요소로 작용할 수 있다.

자료: U.S. Census Bureau, 2004.

[그림 8] 사후조사의 비표본 오차 종류 (계속)

옵션으로 제공해 주기로 결정하였다. 노동통계국(the Bureau of Labour Statistics)의 입장은 과소집계가 보정된 추정치가 센서스 원자료에 비해서 전체 인구 분포에 대해 보다 정확한 정보를 제공하고 있다고 보고 경상인구조사(Current Population Survey)나 소비자지출조사(Consumer Expenditure Survey)등에서 이 결과를 사용하도록 하였다. 다른 주요 가구조사(national household surveys)에서도 보정된 인구를 기반으로 연방통계시스템을 변경해 주도록 요구하는 등 1990년 센서스의 보정자료가 연방통계시스템과 연동되어 활용되었다(Prewitt, 2000). 뿐만 아니라 일반인들 사이에서 1990년 센서스 보정자료의 요구가 증가하자, 센서스국은 [그림 9]와 같이 이 자료를 보정되지 않은 결과(공식수치)와 함께 홈페이지에 공표하는 전략을 채택하였다.



자료: 미국 센서스국 홈페이지(www.census.gov).

[그림 9] 미국 1990년 센서스의 보정전과 보정후 결과 공표

3. 일본의 원넘버 센서스

지금까지 영국과 미국의 원넘버 센서스를 등장배경과 개념, 방법론, 결과를 중심으로 검토해 보았다. 두 국가 모두 센서스에서 과소집계 문제가 심각하게 제기되면서 원넘버 센서스 프로젝트에 착수하였으며, 원넘버 센서스는 센서스 본조사에서 누락 혹은 중복된 것으로 추정된 인구를 센서스 본조사 결과에 통합하는 것을 의미하였다. 누락인구를 추정하는 데에는 사후조사방법이 이용되었다. 그런데 일본의 원넘버 센서스는 이 두 국가와 개념 및 방법론에서 명확한 차이를 보이고 있다. 여기에서는 일본의 원넘버 센서스에 대해서 살펴보도록 하겠다.

가. 원넘버 센서스 개념

일본은 10년마다 대규모 조사(large scale census)를, 중간년도에는 간이 조사(simplified census)를 실시한다.[24] 일본에서 센서스 결과는 인구, 가구, 산업구조 등의 실태 파악을 통해 기초행정자료로 사용된다. 특히 지방자치법, 도시계획법, 지방교부세법 등의 법적 근거에 따라서 도도부현 및 시구읍면 의회의 의원수 상한 설정이나, 시(인구 5만이상)・지정시(인구 50만이상)・핵심시(인구 30만 이상)등의 설치 규정 요건, 지방 교부세의 배분 근거가 된다(www.stat.go.jp). 영국이나 미국과 마찬가지로 일본 또한 센서스 결과의 활용에 대한 법적인 근거가 매우 명확하다는 것을 알 수 있다. 이는 원넘버 센서스의 출발과 밀접한 관련을 맺는 것으로 판단된다. 센서스를 통해서 집계된 인구수가 법적인 근거에 따라서 활용됨에 따라서, 그 만큼 완전하고도 정확한 집계가 필요하게 된 것이다. 물론 일본 통계국(Statistics Bureau of Japan)에서 원넘버 센서스라는 용어를 사용하지는 않고 있다. 그러나, 센서스는 누락과 중복이 없다고 가정되며, 센서스를 통해서 단일의 인구가 작성된다는 점에서 원넘버 센서스라고 할 수 있다. 나아가 이렇게 작성된 인구가 모든 인구통계의 중

24) 두 조사는 항목에서 차이를 보인다. 대규모 조사의 경우, 성 및 연령과 같은 인구학적 특성과 직업, 산업과 같은 경제학적 특성, 그리고 주택, 교육, 내부인구이동을 조사하는 한편, 간이조사의 경우에는 인구학적 특성과 경제학적 특성 및 주택에 대해서 조사한다.

심에 있다. 센서스인구는 현재인구(current population)[25]와 장래추계인구의 기준인구로 사용되고 있다는 점에서 명실상부한 원넘버 센서스를 실시하고 있다고 할 수 있을 것이다. 물론 누락과 중복이 없다는 가정의 현실성은 떨어지지만, 이것은 일본 내에서 센서스의 높은 위상을 보여준다는 점에서 의의가 있다. Takami(2003)는 2000년 센서스 결과의 정확성에 대한 평가의 글에서 사후조사 결과 센서스에서 1.1%의 누락을 보고하는 동시에 “일본 통계국은 누락과 중복없이 모든 개인을 조사하는 센서스 과정에 대한 자부심이 매우 높다”라고 언급하고 있기도 하다.

나. 원넘버 센서스 방법론

일본의 원넘버 센서스는 크게 두 단계로 구분된다고 볼 수 있다. 첫 번째는 센서스 기간에 장기부재로 인해 누락될 가능성이 높은 가구의 규모를 파악하는 것이고, 두번째는 센서스인구 확정 이후에 현재인구와 추계인구의 기준인구로 사용하는 것이다(통계개발원, 2008).

센서스 본조사 기간의 누락 가구에 대한 파악은 영국이나 미국과 다르게 이웃청취조사를 통해서 진행되고 있다. 이 조사는 센서스 기간에 응답을 얻어내지 못한 가구에 한해 조사원이 해당부재가구의 가구원수 및 이름과 성별 정보를 이웃집에 물어서 파악하는 방법이다. 이웃청취조사를 통해서 파악된 항목은 총인구, 가구수, 성별 인구수에 반영되어, 단일 인구를 작성하는데 활용된다. 그러나, 해당 항목 이외의 항목별 집계표에는 포함되지 않는다. 2005년의 경우, 조사표가 회수되지 않아서 세 가지 항목만 조사한 비율은 전체 인구의 0.4% 수준이었다.

다음으로 센서스에서 집계된 인구는 현재인구의 기준인구로 활용된다. 일본은 두 센서스 시점 사이의 인구를 추정하여 발표하고 있다. 인구추정의 대상은 외국인을 포함해 일본에 상주하는 인구로 센서스 조사대상과 동일한 정의를 따르고 있다. 인구추정의 기본식은 가장 최근의 센서스인구를 기준으로 그 후의 인구동태(출생, 사망, 이동)를 다른 인구관련 자료로부터 얻어서 산출하는 것이다.[26]

---

25) 일본에서는 추정인구를 현재인구로 부르기도 한다.

26) 센서스에서 연령 미상의 경우에는 모든 연령 그룹에 평균배분한 후 인구추정을 시작한다.

현재인구는 전국적으로는 매년 10월 1일과 매월 1일에, 도도부현별로는 매년 10월 1일에 공표된다. 영국이나 미국의 경우, 인구추정의 기준시점은 연 중앙인 7월 1일이며 이 기준은 국제적인 비교를 위해서 많은 국가들에서 받아들여지고 있다. 그런데 일본은 10월 1일을 기준으로 한다. 이는 센서스 기준 시점과 일치하는 것으로 일본 내에서 센서스인구가 모든 인구통계에서 중심적인 위치에 있음을 단적으로 보여주는 것이라고 할 수 있다.

직전 센서스를 기준으로 하여 추정된 현재인구는 다음 센서스 실시연도가 되면 센서스인구로 대체되고 두 센서스 사이에서 추정된 인구는 센서스인구를 기준으로 보정된다. 센서스간 보정은 센서스인구가 100% 정확하다는 전제 아래 이전 센서스를 기준으로 추정된 현재인구와 다음 센서스 동안의 인구에 대해서 보정을 실시한다. 보정에서는 총인구 및 일본인 인구, 도도부현 인구에 대한 총수 보정이 이루어진다. 이전 센서스 기준에서 산출된 인구와 현재 센서스인구의 차이를 평균배분하여 각 연도(월)에 증감분을 더해준다.

2005년의 예를 들어 월간보정분과 연간보정분 계산법을 비교해 보자([그림 10]). 2005년 센서스인구와 2000년 센서스인구에 기반하여 추정된 2005년 10월 1일 인구 사이의 차이를 60개월로 평균 배분하면 월간보정분이 나오게 된다. 연간보정분도 월간보정분과 마찬가지로 계산되며, 다만 차이점은 연간보정분은 5년으로 평균배분된다는 것이다. 이렇게 하여 추정된 현재인구는 [그림 11]에서 보듯이 센서스가 실시된 시점에는 센서스인구와 동일하게 된다.

센서스간 월간보정= $\frac{1}{60} \times (p^{(2005)} - p^{(2000)}_{2005})$

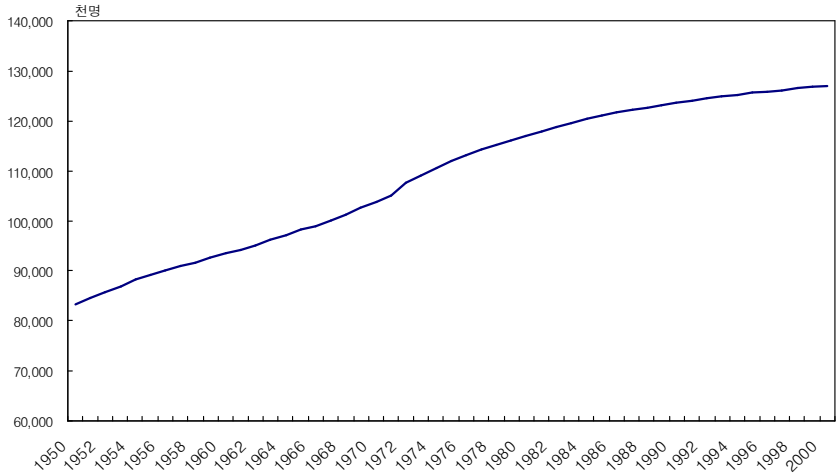
센서스간 연간보정= $\frac{1}{5} \times (p^{(2005)} - p^{(2000)}_{2005})$

$p^{(2005)}$= 2005년 센서스 확정인구

$p^{(2000)}_{2005}$= 2000년 센서스인구 기준으로 추정된 2005년 10월 1일 인구

자료: 일본 통계청 홈페이지(www.srat.go.jp)

[그림 10] 일본의 현재인구 보정방법



자료: 일본 통계청 홈페이지(www.stat.go.jp)
주: 센서스 실시연도(0, 5)는 센서스인구이며 센서스 실시 연도 사이는 현재인구임.

[그림 11] 일본의 인구: 센서스인구와 현재인구(10월 1일 기준)

영국의 경우, 사후조사를 통해서 과소집계를 보정한 결과를 추정인구의 기준인구로 사용하고 있으나, 직전 센서스 결과를 바탕으로 추정한 인구와 센서스 실시연도의 센서스 보정결과와는 큰 차이를 보이고 있음이 1991년과 2001년도 센서스 결과에서 나타났다. 미국은 과소집계의 문제가 있음에도, 추정인구의 기준인구로 보정 전의 센서스 결과를 사용하고 있다. 이들 국가와 비교해 볼 경우, 센서스인구와 현재인구가 동일하게 나타나고 있는 일본은 가장 완벽한 의미에서 원넘버 센서스를 실시하고 있다고 해석될 수 있을 것이다.

다. 원넘버 센서스 결과

일본의 원넘버 센서스는 별도의 평가 과정이 있지는 않으나, 사후조사나 주민기본대장 인구와 같은 행정자료를 통해서 그 결과가 평가되고 있다고 볼 수 있다(Takami, 2003).

1) 사후조사

영국과 미국에서 커버리지 오차 평가를 목적으로 사후조사를 실시

하고 있는 것과 달리 일본에서 사후조사는 내용 오차에 대한 평가를 주목적으로 하고 있다. 사후조사가 커버리지 오차의 평가에 활용되지 않는 것은 다음과 같은 두 가지 이유 때문이다. 첫째, 센서스에 대한 응답은 의무인 반면 사후조사에 대한 응답은 의무가 아니므로 사후조사의 응답률이나 정확성이 낮을 가능성이 있다. 둘째, 표본조사인 사후조사는 다양한 종류의 비표본 오차에 노출되어 있다. 이는 미국의 예에서도 잘 드러난다. 이러한 이유로 인해서 사후조사는 커버리지를 추정하는데 사용되지 않고 내용 오차 확인과 항목무응답 보정을 위한 기본자료로만 활용되고 있다. 그렇다고 해서 커버리지에 대한 평가를 전혀 하지 않는 것은 아니다. 다만 그 비중이 센서스 전체에 비추어 볼 때 낮은 수준이다.

2005년 센서스 사후조사는 센서스 시점으로 2개월 후인 12월에 실시되었다. 일반조사구와 특별조사구[27]로 구분하여 표본을 추출하였으며, 표본규모는 약 6만 가구이다. 사후조사와 센서스 조사표를 매칭하여 가구와 개인이 얼마나 정확하게 조사되었는지를 검토하였다. 사후조사에서는 이름, 성별, 출생년월일 항목에 대해서 물었으며 이것 이외에도 센서스 시점의 거주지, 센서스 시점에 머물렀던 장소, 센서스에서 조사된 장소를 물어서 이 세 가지 항목과 사후조사에서의 조사장소를 비교하였다. 비교는 사후조사에 기입된 모든 사람에 대해서 도도부현 및 시구정촌번호, 조사구번호, 가구번호, 성별, 가구주와의 관계, 출생년월일을 확인하는 방법으로 이루어졌다(SBJ, 2008a).

사후조사에서 조사된 모든 개인은 다음의 네가지 범주중 하나로 분류되었다. ① 한 장소에서 확인된 경우(센서스에서 정확하게 조사됨), ② 확인되지 않은 경우(센서스에서 누락되었을 가능성이 높음)[28], ③ 2곳 이상의 장소에서 확인된 경우(센서스에서 두 번 중복 조사될 가능성이 있음), ④ 완전하지 않은 입력으로 인해 확인이 불가능한 경우. 이에 따라서 2005년 센서스 커버리지 오차 결과를 보면 누락(0.93%)과 중복

27) 특별조사구라 함은 사회시설, 200명 이상의 수용시설을 보유한 병원 및 50명 이상의 단신자가 거주하고 있는 기숙사 구역을 말한다.

28) 이 범주의 모든 사례가 센서스에서 누락된 것은 아니다. 예컨데, 해당 장소에서 조사는 되었으나 개인이 부정확하게 기록되거나 연령을 알 수 없는 경우도 여기에 포함되었다.

(0.16%)을 통한 총오차율 1.09%로 우리나라의 2005년 센서스 총오차율인 3.9%에 비해서 훨씬 낮음을 알 수 있다.[29)]

2) 주민기본대장인구

일본은 우리나라와 유사하게 이름, 생년월일, 성별, 가구주와의 관계 등에 대해 신고하는 주민등록제도를 실시하여 의료보험, 선거인 명부, 학령부의 자료로 활용하고 있다. 일본인이라면 누구나 이사 및 가구주와의 관계가 변경되었을 경우, 관할 시구읍면에 14일 이내에 신고를 해야 하며, 정당한 사유없이 신고를 지연할 경우 5만엔 이하의 과태료를

〈표 6〉 일본의 센서스인구와 주민기본대장인구 비교

(단위: 명)

지역	센서스	주민기본대장	차이(c)	비율(%)
	(a)	(b)	(a)−(b)	(c)/(a)×100
(센서스인구가 주민기본대장인구보다 많은 도도부현)				
쿄토부	181,823	158,569	23,254	12.79
도쿄도	832,740	792,934	39,806	4.78
아이치현	425,292	424,050	1,242	0.29
미야기현	152,491	151,429	1,062	0.70
카나가와현	540,965	519,468	21,497	3.97
(주민기본대장인구가 센서스인구보다 많은 도도부현)				
와카야마현	46,544	56,725	-10,181	-21.87
삼현	70,860	76,609	-5,749	-8.11
카가와현	46,193	54,079	-7,886	-17.07
미야자키현	54,891	61,983	-7,092	-12.92
에히메현	67,592	77,167	-9,575	-14.17

자료: SBJ, 2008c.
주1: 20~24세 연령을 대상으로 한 것임.
주2: 센서스인구는 2005년을, 주민기본대장인구는 2005년과 2006년 3월 31일 현재인구를 평균하여 산출한 것임.

29) 일본 통계국의 Ishihara과장보좌와의 2008년 7월 17일 면접내용임.

내야 한다. 이외에도 일년 이상 해외에 체류하게 될 경우에도 신고를 하게 되어 있다(SBJ, 2008b).[30]

주민기본대장은 일본에 거주하는 일본인을 대상으로 하므로 센서스와는 일본 국적을 가진 사람에 한해서 비교가 가능하다. [표 6]을 보면, 일반적으로 도쿄와 같은 대도시의 경우에는 센서스인구가 주민기본대장인구보다 많고, 시골의 경우에는 센서스인구가 주민기본대장인구보다 적은 경향이 나타났음을 알 수 있다.

이는 두 자료의 차이에서 비롯된 것이다. 센서스는 조사원이 센서스 시점에 해당 가구에 방문해서 응답자들의 실제 거주지를 조사하는 반면에 주민기본대장은 신고에 의한 거주지를 대상으로 하고 있기 때문이다. 이러한 개념적인 차이 이외에도 어떤 이동자들은 법에 의해서 규정되어 있음에도 불구하고, 자신들의 이동을 보고하지 않는 경우도 있다. 예컨대, 취학 및 취업 등을 이유로 이동을 할 경우 신고를 하지 않기도 한다(Takami, 2003).

이러한 객관적인 차이에도 불구하고, 센서스인구와 주민기본대장인구의 활용도에 있어서는 단연 센서스인구가 높다. 지정통계인 센서스는 광범위하게 사용되는 반면에 등록된 인구를 집계한 주민기본대장은 업무 등에 참고자료로 활용되고 있다. 예를 들어, 센서스 조사기간에 장기부재가구의 가구원 이름을 알 경우에는 주민기본대장의 정보를 이용하여 센서스 조사항목을 채운다.

센서스의 높은 위상에도 불구하고, 최근에 센서스인구와 장래추계인구와의 비교를 통해서 센서스인구의 정확성에 대한 문제가 제기되고 있다. 추계인구와 센서스인구의 차이가 벌어짐에 따라서 센서스 자료의 품질에 대한 문제가 지적된 것이다. 2005년 센서스 결과 역시 이러한 점을 반영하고 있다. 조사대상가구와 조사원과의 거리가 멀어서 접촉이 어려운 사례가 증가하는가 하면 어떠한 가구는 조사표 작성에 비협조적이기도 하였다. 일본 통계국은 이를 두고 맞벌이 가구 및 독신가구가 증가하고, 개인의 프라이버시 의식이 높아짐에 따른 것으로 평가하고, 이에 대한 대응책 마련에 부심하고 있다(Ishihara, 2008).

---

30) 주민기본대장이 주민세 부과의 기준이 되므로, 이러한 사유로 전출신고서를 제출하는 경우가 많다고 평가하나 정확한 규모는 파악되지 않고 있는 것으로 알려져 있다.

4. 영국 · 미국 · 일본의 원넘버 센서스 비교

지금까지 살펴본 영국과 미국, 일본의 원넘버 센서스의 주요 특징을 <표 7>을 통해서 다시 한 번 정리해 보았다.

영국과 미국 모두 특정집단의 커버리지 오차가 지속적으로 높게 나타남에 따라서 원넘버 센서스 프로젝트에 착수하였다. 이에 따라서, 원넘버 센서스는 누락 혹은 중복 조사된 것으로 추정된 부분을 센서스 결과에 보정하는 것으로 그 개념을 정의하였다. 반면에 일본은 명시적으로 원넘버 센서스라는 용어를 사용하지는 않고 있으나 센서스의 완전한 커버리지를 가정함에 따라서 센서스에서 조사된 인구를 추정 및 추계인구의 기준으로 하며, 이것을 원넘버 센서스라고 할 수 있다.

방법론을 보면, 영국과 미국에서는 센서스와는 독립적인 사후조사를 실시하고 그 결과를 센서스 자료와 매칭하여 누락 혹은 중복조사된 것을 추정하고, 센서스에 보정하는 방법을 채택하고 있다.[31]

사후조사는 먼저 표본설계와 자료수집 측면에서 비교가능하다. 두 국가의 사후조사는 커버리지 오차를 측정하는 것을 목적으로 하고 있다. 커버리지 오차는 어느 지역단위에서 측정할 것인가에 따라서 표본규모가 결정된다. 영국은 우리나라의 시군구 수준인 지방자치단체에서의 성 및 연령별 인구 추정을 목적으로 함에 따라서 약 37만 가구의 표본이 필요했으며, 이는 지방자치단체에 대한 간접추정을 할 수 있는 규모였다. 미국은 전국 및 주요 하위지역에서의 인구 추정을 목적으로 하며 이는 약 30만가구로도 충분한 것으로 나타났다. 표집단위는 두 국가 모두 지역이다. 가구단위 표집도 고려했으나, 비용 및 정확성을 고려할 경우에 지역기반 표집이 효율적인 것으로 나타났다. 자료수집방법은 영국은 조사원 방문면접을, 미국은 컴퓨터를 보조적으로 활용한 전화와

31) 영국의 경우, 보정 이후에 품질확인절차가 추가되는 반면에 미국의 경우는 이 절차가 포함되지 않는다는 차이가 있다. 그렇다고 해서 미국의 원넘버 센서스에서 품질확인절차가 과소평가되는 것은 아니다. 영국의 품질확인절차와 유사한 단계가 센서스 보정 결과에 대한 평가 단계에 마련되어 있다. 미국에서는 센서스 보정 결과를 인구추정의 기준인구로 사용하기 위해서 매 단계에 대한 평가를 철저히 하였다. 특히 통계적 기법을 사용하여 인구수를 작성하는 것에 대한 논란의 여지가 많았던 만큼 보정 결과에 대한 평가는 공표에 앞서서 반드시 진행되어야 하는 필수조건이었다.

〈표 7〉 영국, 미국, 일본의 원넘버 센서스 비교

구분		영국	미국	일본
개념		센서스 본조사에서 과소집계된 것으로 추정된 것을 센서스 본조사 결과와 통합하여 최고 품질의 단일 인구를 작성	집계, 할당, 통계적 추정의 조합을 해서 법적인 마감시일까지 최고의 단일 인구를 작성	센서스 본조사에서 과소 및 과대집계 없이 모든 개인을 조사한 후 이 인구를 추정인구의 기준으로 사용
방법론		센서스와 사후조사 자료를 매칭하여 과소 및 과대집계를 추정하고 그 결과를 센서스 본조사에 보정하는 방법		이웃청취조사를 통해서 무응답 가구 정보를 파악하여 센서스인구를 확정하고, 이 인구를 기준으로 현재인구를 보정하는 방법
사후조사	목표추정집단	지방자치단체	전국 및 주요지역	-
	표본규모	약 37만 가구	약 30만 가구	-
	표집단위	지역		-
	조사방법	PAPI	CATI와 CAPI	-
매칭	시점	면접 → 가구・개인매칭	가구매칭 → 면접 → 개인매칭	-
	방법	자동 및 수동 매칭		-
	품질확인절차	3단계		-
추정	직접추정지역	설계그룹	사후층	-
	기본추정법	이원시스템+회귀	이원시스템	-
	간접추정지역	지방자치단체	블록	-
	간접추정법	합성추정법		-
보정	대상	가구・개인	개인	-
	방법	기증자 보정방법		-
	보정위치	우편번호	블록	-
공표	소요기간	약 17개월	약 11개월	약 16개월
	공표내용	보정 후 결과	보정 전 결과	센서스 조사 결과

방문면접을 혼합하여 사용하였다.

사후조사가 완료되면 센서스 자료와의 매칭이 시도되었다. 매칭시점을 제외하고는 매칭방법이나 주요 절차는 영국과 미국에서 매우 유사하였다. 영국은 모든 면접이 완료된 이후에 센서스 조사표와 사후조사 조사표를 매칭하여 가구와 개인을 같은 시점에 매칭하는 반면, 미국은 표본설계 단계에서 주소록을 구축한 이후에 센서스 매스터주소록의 가구와 사후조사를 위한 주소록의 가구를 매칭한 후 그 결과를 바탕으로 표집을 하였다. 그리고 표집된 가구에 방문하여 개인면접을 하고 난 이후에 개인매칭을 하였다. 두 국가 모두 자동매칭과 수동매칭을 하고 있으며, 수동매칭은 크게 3단계를 거쳐서 품질확인을 하였다.

매칭된 자료를 바탕으로 추정을 한다. 두 국가 모두 추정은 직접추정단계와 간접추정단계로 구분되었다. 표본지역에 대해서는 이원시스템 추정법을 통해서 직접적인 추정치를 작성하였다. 영국의 경우에는 우편번호 수준에서 이원시스템 추정법으로 인구를 추정하고 회귀식을 이용하여 표본이 아닌 우편번호지역의 인구를 추정하여 설계그룹 전체에 확대하였다. 미국은 자료수집 이후 성, 연령, 인종, 주택소유여부 등 9개 변수를 이용하여 사후층화를 한 후 이 층 내에서 이원시스템 추정을 하였다. 설계그룹 및 사후층보다 하위지역에 대해서는 합성추정이라는 간접추정법을 이용하였다. 이 방법은 대지역과 소지역이 동일하다는 비교적 간단한 가정으로 인해서 많이 사용되고 있다.

다음으로 누락 혹은 중복 조사된 것으로 추정된 수 만큼 센서스 본조사 결과에 보정하였다. 보정은 유사한 특성을 가진 개인의 정보를 복사하는 기증자 보정방법을 사용한다는 점에서는 동일하나, 세부 내용에서는 차이를 보였다. 보정의 대상이 영국의 경우에는 과소집계된 가구와 개인이었던 반면에, 미국은 과소집계 뿐 아니라 과대집계된 개인을 보정대상으로 하였다. 또한 보정시 영국은 개인 보정시 가구 속성을 부여한 반면에 미국은 개인 보정시 가구 속성은 보정되지 않았다. 기증자 보정방법을 통해서 복사된 개인은 영국은 우편번호 내의 가구에, 미국은 블록수준에서의 특별범주에 넣어졌다. 미국의 경우, 보정되는 개인에게 가구 속성이 부여되지 않으므로 가구에 넣지 않았다.

일본에서는 센서스 이후에 이웃청취조사를 통해서 무응답 가구에 대

한 최소한의 정보를 대리인을 통해서 파악하고 있다. 무응답 가구에 대한 가구원수, 이름과 성별을 파악하여 총인구 및 가구, 성별 인구에 반영하여 센서스인구를 확정한다. 이 인구는 현재인구의 기준으로 활용될 뿐 아니라 두 센서스 사이의 현재인구를 보정하는 데에도 사용된다.

센서스 본조사 이후 보정까지의 과정을 모두 완료하는데, 영국은 약 17개월, 미국은 약 11개월이 소요되었다. 인구추정치를 두고, 영국은 최단 시일 내에 작성된 최고 품질이라고 평가한 반면, 미국은 정확성 향상에 대한 근거를 발견하지 못해 추정인구의 기준으로 활용하지 않았다. 일본은 보정하지 않은 센서스 결과가 그대로 확정되며, 2005년 센서스의 경우 총인구를 공표하는데 약 16개월 정도의 시간이 소요되었다.

지금까지 세 국가의 원넘버 센서스 주요 특징에 대해서 비교해 보았다. 다음 절에서는 각 국가의 원넘버 센서스 방법론의 우리나라에 대한 시사점을 도출해 보고자 한다. 우리나라의 센서스 커버리지 오차 현황을 분석해 보고 그 결과를 바탕으로 시사점을 찾아보고자 한다.

제3절 우리나라의 센서스 커버리지 오차 현황 및 시사점

1. 센서스 커버리지 오차 현황

우리나라에서 센서스 커버리지는 인구분석과 사후조사를 통해 평가되고 있다. 그러나 앞서 보았듯이 인구분석에 의한 추정치는 전국수준에서는 신뢰할 만하나 지역수준에서는 자료의 불충분으로 인해서 커버리지를 평가하는데 한계가 있다. 반면에 사후조사는 소지역 인구를 추정할 수 있다는 점에서 많이 사용되어 오고 있다. 실제로 영국에서 원넘버 센서스가 실시될 수 있었던 것도 정교하게 설계된 사후조사가 뒷받침되었기 때문이다. 이에 우리나라의 사후조사방법과 이를 근거로 측정된 센서스 커버리지 오차현황을 살펴본 후 우리나라에서의 원넘버 센서스 시사점에 대해 검토해 보고자 한다. 물론, 사후조사 자체가 완전하지

않기 때문에 이를 기준으로 평가한 센서스 커버리지 오차에 절대적으로 의존하는 것은 위험할 수 있다. 그러나, 현재 주요 지역별로 커버리지 오차를 측정할 수 있는 대안이 사후조사 이외에는 없다. 게다가 어느 한 시점의 사후조사 결과가 아닌 다양한 시점에서 조사된 결과를 비교해 봄으로써 커버리지 오차의 일반적인 패턴은 파악해 볼 수 있을 것으로 기대된다.

가. 2005년 센서스 사후조사 방법론

1) 사후조사의 역사

우리나라의 사후조사는 커버리지 오차 측정과 함께 내용 오차 파악을 목적으로 1960년 센서스에서부터 시작되었다(<표 8>). 영국이나 미국에서 독립적인 방법의 사후조사를 실시하고 있는 것과는 달리, 우리나라에서는 1970년과 1975년 2회를 제외하고는 종속적인 방법의 사후조사를 실시하고 있다. 독립적인 방법은 표집틀이나 조사원 및 관리자 등 사후조사의 운영이 센서스와는 독자적으로 진행되는 것인 반면에 종속적인 방법은 센서스의 프레임과 운영이 사후조사에서도 그대로 활용되는 것을 말한다(김민경, 2002). 1975년 이후의 사후조사에서 독립적인 방법을 검토한 적은 있다. 그러나 이 방법은 센서스 조사표와 사후조사 조사표의 매칭 작업이 쉽지 않아서 신속한 분석이 어려우며, 정밀한 매칭을 위해서는 많은 인력의 투입과 시간이 소요된다는 단점 때문에 지속적으로 시행되지 못했다(경제기획원, 1982). 반면에 종속적인 방법은 조사 및 자료 처리가 용이하고, 응답자와 조사원의 부담이 경감되는 동시에 조사비용이 저렴하다는 장점으로 1980년 이후 시행되어 오고 있다.

사후조사의 시점은 센서스 기준일 이후 1개월 후에 진행되고 있다. 1960년에는 6개월이었던 것이 회차를 거듭하면서 그 기간이 점점 단축되어 1975년 이후 현재까지는 센서스 1개월 이후에 사후조사가 진행되었다. 이는 센서스 시점의 정보에 대한 응답자들의 회상오차(recall error)를 감안한 것이었다.

표본규모는 전체 조사구를 모집단으로 하였을 경우의 비율로 표시해

보았는데, 1960년에는 0.60%이었던 것이 점차 감소하여 2005년에는 약 0.27% 정도인 것으로 나타났다.

〈표 8〉 우리나라 사후조사 실시현황

구분	1960	1966	1970	1975	1980	1985	1990	1995	2000	2005
조사방법	종속	종속	독립	독립	종속	종속	종속	종속	종속	종속
본조사후기간	6개월	1.5개월	2개월	1개월	1개월	1개월	1개월	1개월	1개월	1개월
조사구수	404	-	293	260	309	425	550	600	600	730
표본크기	0.60%	-	0.40%	0.30%	0.30%	0.29%	0.30%	0.28%	0.24%	0.27%

자료: 경제기획원(1982), 김민경(2002), 통계청(2006a).
주: 표본크기는 전체 조사구수에 대한 표본 조사구수의 비율임.

2) 2005년 센서스 사후조사 방법론

가) 표본설계

2005년 사후조사의 모집단은 보통 조사구 및 아파트 조사구로 구성된 일반 조사구와 기숙시설 조사구 및 특수사회시설 조사구로 구성된 집단시설 조사구이며, 표본추출은 두 조사구에서 독립적으로 실시되었다. 집단시설 조사구에 대한 사후조사는 2005년도에 처음으로 이루어졌다. 이 조사구의 경우, 조사 자체가 어렵기 때문에 누락규모가 과소집계될 가능성이 높아서 모집단에서 제외되어 왔다. 영국이나 미국의 경우에도 이 집단에 대한 누락규모 파악의 어려움으로 고심하고 있다.

표본규모는 커버리지 오차의 정확성 및 인력, 예산을 감안하여 선정하였다. 일반 조사구의 경우, 시도별로 과소집계 및 과대집계를 분석할 수 있을 정도의 표본규모 추출을 목적으로 하였다. 2000년 센서스 사후조사의 표준오차와 표본규모를 참고로 하여 목표 표준오차를 0.025 ~ 0.050 로 한 후 시도별로 표본규모를 산출하였는데, 그 결과 693개의 조사구가 선정되었다. 그런데 이 방법은 2000년 자료를 사용함으로써 그 동안의 모집단 변화가 반영되지 않는다는 한계가 있다. 이에 따라서 가설정된 2005년 조사구의 크기를 반영, 최종적으로 700개 조사구가 선정되었다. 집단시설 조사구의 표본은 전국 수준에서 분석이 가능한 규모

를 추출하고자 하였다. 분석을 위한 최소 표본규모는 30개 조사구이며, 해당 범위 내에 있는 조사구를 정렬한 후 계통추출 하였다(통계청, 2005a).

나) 자료수집

자료수집은 준비조사, 본조사, 전입가구원에 대한 대조작업 단계로 이루어졌다(통계청, 2005b). 준비조사에서 조사원은 크게 두 가지 업무를 수행하였다. 하나는 해당 조사구에 대한 센서스 정보를 확보하는 것이며, 다른 하나는 조사구 경계 및 거처를 확인하는 것이다. 조사원은 표본으로 추출된 조사구의 2005년 센서스 가구명부, 조사표, 조사구 요도를 수집하여 그 내용을 사후조사의 가구명부와 조사표의 항목에 그대로 기록하였다. 사후조사의 가구명부에는 거처번호, 가구번호, 주소, 가구주 성명, 센서스 시점의 가구원 수, 센서스 조사원의 인적사항을 기록하고, 사후조사의 조사표에는 성명, 성별, 가구주와의 관계, 연령, 혼인상태, 센서스 응답자를 기록하였다. 이는 종속적인 방법이 센서스의 틀을 그대로 활용하는 것에 따름이다. 이후에는 센서스 정보를 바탕으로 조사구 경계를 확인하여 조사구 내에 거처 누락이 있는지를 확인하였으며, 특히 인접 조사구의 거처가 잘못 포함되어 있지는 않은지 등을 확인하도록 하였다. 조사구 경계확인이 끝나면 조사구 요도를 가지고 모든 거처를 확인하면서 요도에 빠진 주요 지형물을 보완하였다.

다음으로 조사원은 센서스 내용이 기록된 사후조사 가구명부와 사후조사 조사표를 바탕으로 해당 가구를 방문하여 면접조사를 실시하였다. 조사내용은 가구원에 관한 사항, 가구에 관한 사항, 조사표 작성에 관한 사항으로 구분되었다. 연령, 혼인상태 등의 개별항목을 확인하여 센서스의 정확성을 파악하고, 중복 및 누락, 전출입 사항을 확인하여 센서스의 커버리지를 파악하였다.

조사원 업무량은 1인당 1조사구이며, 이에 따라서 총 730명의 조사원이 고용되었다. 실사지도원은 조사원 9명당 1명꼴인 77명이며, 중앙실사지도원은 20명으로 구성되었다. 조사원으로는 지방통계청 조사직원이 동원되었으며, 가능한 센서스와의 독립성을 유지하고 허위조사가능성을 줄이기 위해서 시군구간 교체조사를 하였다. 즉, 조사원으로 하여금 센

서스에서 담당하지 않았던 지역을 조사하도록 하였다. 그러나, 경상조사 업무를 병행함에 따라서 업무가 과중되었다는 평가가 2000년 사후조사 이후 제기되고 있다(통계청, 2001; 2006a).

커버리지 오차를 측정하기 위해서는 누락과 중복 뿐 아니라 가구의 이동지위를 확인해야 한다. 센서스 시점과 사후조사 시점의 차이로 인해서 전입과 전출과 같은 변동이 발생할 수 있으므로 이들 이동자에 대한 누락과 중복을 파악해야 한다. 사후조사에서 이동자의 누락정도를 파악하는 방법으로는 ① 전출자를 조사하는 방법, ② 전입자를 조사하는 방법, ③ 전출자 및 전입자를 모두 조사하되 전출자의 조사결과를 전입자에 환산 적용하는 방법이 채택될 수 있다. 우리나라는 전입자를 조사하는 방법을 채택하였다. 전출자 조사방법은 전출지역에 대한 조사의 어려움으로 인해서 이웃주민 또는 대리응답자를 통해서 얻게 되므로 응답이 확실하지 않다는 결정적인 단점으로 인해서 채택되지 않았다(경제기획원, 1982). 미국에서는 1990년 사후조사에서는 전입자를 조사하는 방법을 채택하였으나, 2000년에는 사후조사에서 최대한의 커버리지를 확보하고자 세번째 방법으로 이동자를 추정하였다.[32)]

이동자는 센서스 시점을 기준으로 센서스에서 조사되었지만, 사후조사일 현재 조사구 밖으로 전체 가구원이 전출하였는지와 센서스에서 조사여부와 관계없이 센서스 시점 이후에 조사구 내로 전체 가구원이 전입하였는지에 따라서 전출가구와 전입가구를 구분하였다. 전입가구에 대해서는 사후조사를 통해서 얻은 센서스 당시 거주지 정보를 활용하여 전입자 대조 확인작업을 거쳐서 조사여부 결과를 입력하였다. 이 때 대조가능한 경우와 대조가 불가능한 경우로 구분하도록 했다. 대조가능자는 전입자 중 이전의 거주지 주소가 확실한 자로 조사, 중복, 누락으로 구분하였다. 조사는 전입지, 전출지의 센서스 결과를 비교하여 센서스 당시 한 곳에서 조사가 된 경우, 중복은 전입지와 전출지 센서스 결과를

32) 미국의 2000년 센서스 시범예행조사 평가 결과에서는 예상대로, 전입자와 전출자를 모두 파악하는 방법이 전입자를 파악하는 방법에 비해서 더 많은 추정치가 나오는 것으로 파악되었다. 전출자에 대한 정보파악은 두 가지로 이루어질 수 있다. 하나는 해당 가구의 전입자 및 아파트 관리인 등 대리인으로부터 정보를 확보하는 방법이며, 다른 하나는 전출자를 추적하여 정보를 파악하는 방법이다. 시범예행조사에서 두 방법을 모두 평가한 결과 추정치에서 유의미한 차이가 발견되지 않았다. 따라서 전출자를 직접적으로 추적하는 방안은 대안에서 제외되었다(Shindler, 1999).

비교한 결과 전입지와 전출지 모두 조사가 된 경우, 누락은 전입지, 전출지 센서스 결과를 비교한 결과 전입지와 전출지 모두 조사가 되지 않은 경우이다. 대조불능자는 전입자 중 이전의 거주지 주소가 불확실한 자이다(통계청, 2006a).

다) 커버리지 오차의 계산

종속적인 조사방법에서 커버리지 오차는 단수시스템 추정법(Single System Estimation)으로 측정되었다. 이원시스템 추정법에서는 센서스와 사후조사의 독립성 및 동질성 가정 하에서 총인구를 추정하고 이 인구를 센서스인구와 비교하여 오차율을 계산하는 것과 달리, 단수시스템 추정법에서는 사후조사를 통해서 센서스인구를 재구성한 후에 이 인구와 센서스인구의 비율을 고려하여 오차율을 계산하였다. 이는 사후조사가 완전한 커버리지를 제공한다는 가정에 의한 것으로, 본래의 센서스 집계치에서 누락된 것으로 파악된 부분은 더해주고, 잘못 조사된 부분은 제거해 줌으로써 센서스인구를 재구성하였다(Schindler and Navarro, 1994). 이에 따라서 2005년 사후조사에서 커버리지 오차율을 계산한 방법은 [그림 12]와 같다.[33] 이 공식에 따라서 전국 및 시도별 커버리지 오차를 계산하였으며, 그 결과는 인구 및 가구 추계시 보완자료로 활용되고 있다.

중복률 = (중복인구) / (비이동인구+누락인구-중복인구)
누락률 = (누락인구) / (비이동인구+누락인구-중복인구)
순누락률 = 누락률-중복률
총오차율 = 누락률+중복률

* 누락인구에는 전입중 누락인구가 포함, 중복인구에는 전입중 중복인구가 포함된 것임

자료: 통계청, 2006a.

[그림 12] 우리나라의 2005년 센서스 커버리지 오차 계산방법

33) 사후조사에서 오차율을 계산하는 방법은 사후조사 실시 시점마다 차이를 보여주고 있다. 오차율 계산방법에 대한 재검토가 필요할 것으로 보인다.

나. 센서스 커버리지 오차 현황

사후조사를 통해서 측정된 커버리지 오차는 총수의 정확성(numeric accuracy)과 분포의 정확성(distributive accuracy)의 두 가지 측면에서 평가해 볼 수 있다. 전자는 해당 집단에서의 총 인구수가 실제에 얼마나 근접해 있는가를 말하며, 후자는 특정의 지역 혹은 인구학적 그룹의 오차가 다른 지역 혹은 그룹의 오차와 얼마나 비슷한 분포를 보이는가를 뜻한다(Prewitt, 2000).

먼저, 총수의 정확성은 전국수준에서의 커버리지 오차를 살펴봄으로써 짐작해 볼 수 있다(<표 9>). 누락률에서 중복률을 뺀 값인 순누락률이 2005년의 경우 0.9%로 1995년(1.3%)과 2000년(1.6%)보다 감소하였다. 그러나 순누락률이 낮다고 해서 전국의 모든 인구가 완전에 근접하게 집계되었다는 뜻은 아니다. 센서스에서 조사되기 어려운 인구는 종속방식을 취하고 있는 사후조사에서도 마찬가지로 조사되기 힘들기 때문에 누락인구가 과소추정될 가능성이 있다. 또한, 총오차율 지표와 비교할 경우 오차정도를 축소해서 보여주기도 한다. 왜냐하면 누락률과 중복률이 둘 다 높을 경우 서로의 영향이 상쇄되기 때문이다. 1990년의 사후조사 결과 중복률과 누락률을 합한 총오차율은 6.9%로 사후조사가 시작된 이래로 가장 높았음에도 불구하고 순누락률은 -0.1%로 가장 낮았다. 이에 따라서 본 연구에서는 중복률과 누락률을 구분해서 살펴보았다. 일반적으로 거의 모든 시점에 누락률이 중복률에 비해서 더 높은 것으로 나타났으며, 많게는 2배정도 차이가 남을 알 수 있었다.

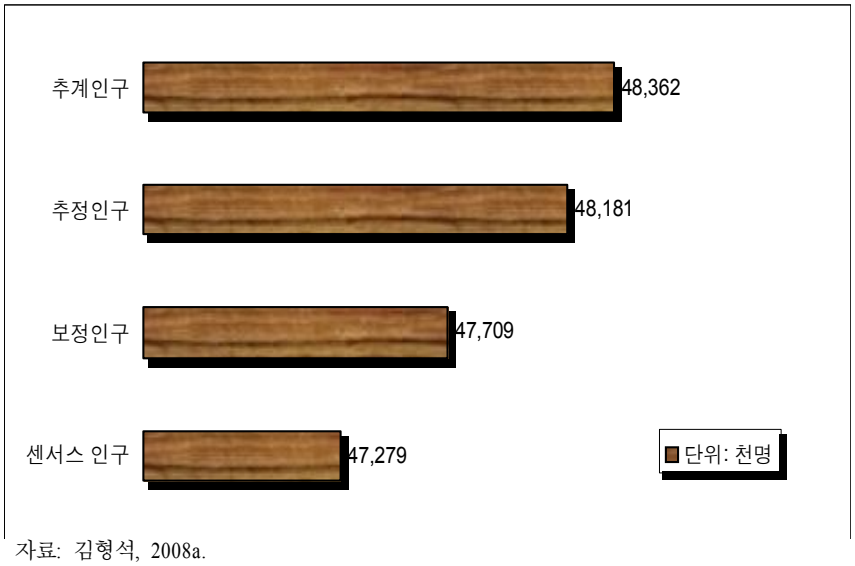
〈표 9〉 우리나라 센서스 커버리지 오차의 연도별 비교

(단위: %)

구분	1985년	1990년	1995년	2000년	2005년
중복률(A)	1.3	3.5	1.8	1.7	1.5
누락률(B)	2.1	3.4	3.1	3.3	2.4
총오차율(A+B)	3.4	6.9	4.9	5.0	3.9
순누락률(B-A)	0.8	-0.1	1.3	1.6	0.9

자료: 통계청, 2006a.

총수의 정확성에 대한 평가는 동일한 기준 시점에 서로 다른 방법으로 산출된 인구수에 대한 비교를 통해서도 평가해 볼 수 있다([그림 13]). 만일, 모집단이 동일하다면 각각의 자료에서 측정된 인구수가 동일해야 한다. 2005년 11월 1일을 기준으로, 센서스에서 집계된 인구는 47,279천명이며, 센서스의 커버리지 오차를 보정한 인구는 47,709천명, 가장 최근의 센서스 보정인구를 기준으로 한 추정인구는 48,181천명, 추계인구는 48,362천명이다. 센서스인구와 추정인구 사이에는 약 90만명이, 추계인구 사이에는 약 100만 명 이상의 차이가 발생하고 있다.



자료: 김형석, 2008a.

[그림 13] 인구자료별 인구수 현황(2005년 11월 1일 기준)

사후조사를 통해서 살펴본 결과 센서스인구에서 총수의 정확성은 완전하지 않음을 알 수 있다. 순누락률은 점차 감소하여 1%미만으로 나타나고 있지만, 누락률과 중복률을 모두 반영하고 있는 총오차율은 여전히 약 4% 정도로 높았다. 또한 추정인구 및 추계인구와 비교해 본 결과 100만명 가까운 차이가 발생하고 있다. 물론, 사후조사의 한계로 커버리지 오차 수준에 대한 정확성에 의문이 제기될 수도 있다. 또한 행정자료의 불완전으로 인해서 추정 및 추계인구의 정확성에 대해서도 논란의

여지가 있을 수 있다.[34] 그러나 이러한 상황 자체가 우리나라에서 단일 인구의 중요성을 더욱 부각시켜주는 것으로 볼 수 있을 것이다.

다음으로 분포의 정확성을 성, 연령 및 지역별로 살펴보도록 하자. 자료 이용이 가능한 최근 3개 시점(1995~2005년)의 사후조사 결과를 분석해 본 결과 주요 특성별 하위집단에서 커버리지 오차의 분포에 차이가 있으며 이러한 경향은 시계열적으로도 지속되고 있음을 알 수 있었다. 분석결과는 누락률과 중복률을 구분하여 제시하였으며, 오차율 계산 공식이 시기마다 차이가 있어서 2005년 공식을 기준으로 다시 계산하였다. 지역별 비교의 경우에 한해 행정구역 변경 등의 문제가 있어서 2000년과 2005년만을 비교하였다.

먼저, [그림 14]를 통해서 남자와 여자 사이의 커버리지 오차 분포를 비교해 보자. 누락률은 2000년을 제외하고는 남자의 누락률이 여자에 비해서 높게 나타났다. 2000년의 경우 비록 여자의 누락이 높게 나타나기는 하나, 그 차이는 미미한 수준이다. 반면에 중복률은 남자의 중복률이 여자의 중복률에 비해서 일관적으로 높으며 그 차이 또한 유지되고 있다.

[그림 15]는 5세 단위의 연령 집단별 누락률과 중복률을 나타낸 것이다. 누락률을 보면, 1995, 2000, 2005년도의 그래프 모양과 그 값이 차이를 보이기는 하나, 누락률이 높은 집단과 누락률이 낮은 집단이 세 시점에서 유사하게 나타나고 있음을 알 수 있다. 5세이하, 15-30세, 75세 이상의 노년층에서 누락률이 전체 평균보다 높은 반면에 5-14세 그리고, 35-64세 연령층에서는 누락률이 평균보다 낮게 나타났다. 중복률의 그래프 모양은 1995, 2000, 2005년도가 매우 유사하다. 15-35세 연령층의 중복률이 매우 높으며 이 집단을 제외한 거의 모든 연령층에서의 중복률은 전체 평균에 비해서 낮게 나타난다. 이 두 결과를 종합해 볼 때 20대 전후 연령집단에서 누락률과 중복률이 모두 높음을 알 수 있다.

[그림 16]과 [그림 17]은 성 및 연령별 누락률과 중복률의 분포를 보여주고 있다. 누락률은 남자나 여자 모두에서 연령에 따라 등락을 거듭하며 복잡한 양상을 보인다. 그러나, 시계열적으로 보면 평균을 기준으

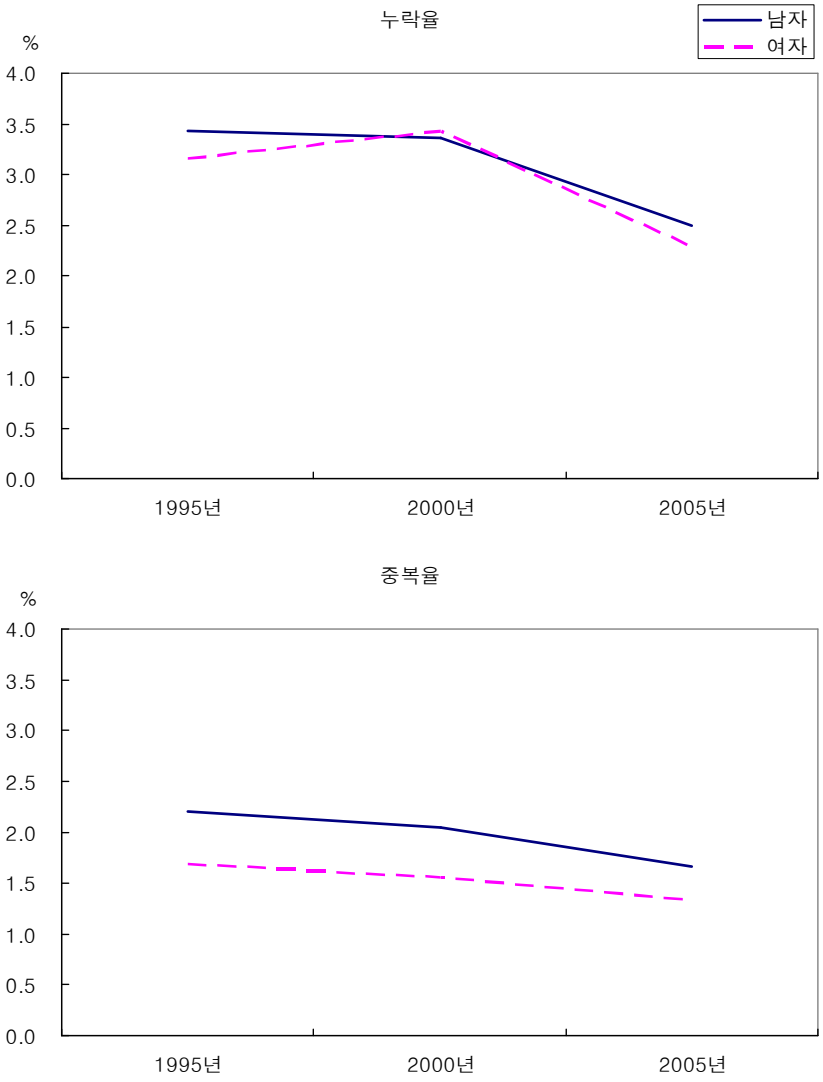
34) 주민등록의 이중 부여 및 말소 문제로 인해서 추정 및 추계인구 그 자체에도 과대 및 과소집계의 문제가 있을 수 있다.

로 15-35세 집단의 누락률이 남자와 여자 모두에서 높게 나타나고 있다는 점에서 유사하다고 할 수 있다. 다만 차이점은 남자와 달리 여자에서 70대 이상 노년층의 누락률이 높게 나타나고 있다는 것이다. 중복률의 경우, 두 집단 모두 20대 지점에서 뾰족한 모양을 보이고 있다는 점에서 매우 유사하다. 그러나 그 값을 비교해 보면, 남자의 경우는 여자에 비해서 20대의 누락률이 2배 가량 더 높게 나타남을 알 수 있다.

시도별로 누락률과 중복률을 보면([그림 18]), 주요 광역시의 누락률과 중복률이 도 단위에 비해서 낮게 나타나고 있음을 알 수 있다. 2005년도의 누락률을 보면 대전을 제외한 서울, 인천, 광주, 대구, 부산, 울산 등 특·광역시에서의 그 비율이 낮게 나타났다. 대전의 경우에도 다른 도 단위 지역에 비해서는 누락률이 낮은 편에 속한다. 중복률은 2000년에 비해서 2005년도에는 경북 지역을 제외하고 거의 모든 지역에서 커버리지 오차가 낮아졌다. 특히 서울과 대구의 중복률이 낮아진 모습이다.

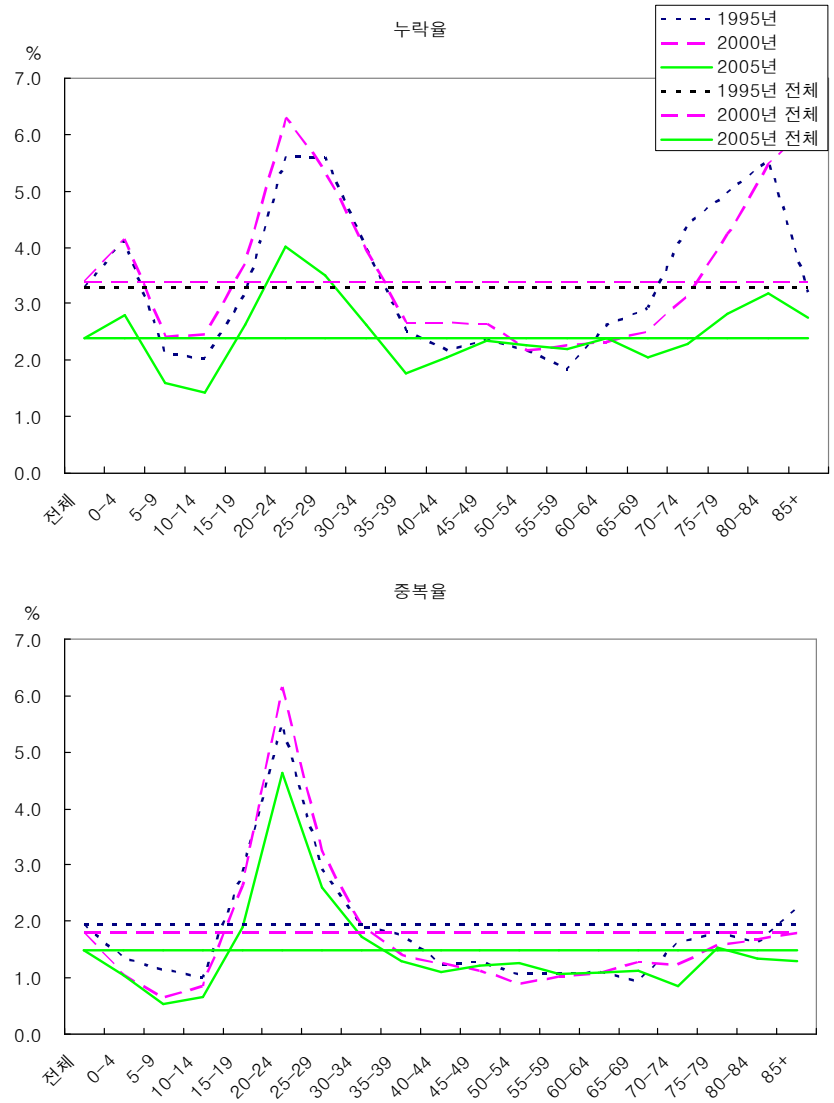
요약해보면, 1995~2005년 센서스 커버리지 오차의 모습을 통해서 주요 하위 집단간 누락률과 중복률의 패턴이 매우 유사한 모습으로 지속되고 있음을 알 수 있었다. 남자의 누락률과 중복률이 여자에 비해서 높으며, 다른 어느 연령집단보다도 20대 남자의 누락률과 중복률이 높게 나타났다. 지역별로는 광역시의 누락률과 중복률이 도 단위 보다는 낮게 나타나고 있다. 이러한 결과는 다음과 같은 의미를 갖는 것으로 해석될 수 있다.

첫째, 사후조사의 본래 목적이 센서스 자료의 질을 평가하고 향후의 센서스를 개선하는데 활용하고자 하는 목적이라는 점을 볼 때, 지난 15년간 누락률이나 중복률의 패턴이 유지되었다는 것은 사후조사 결과를 센서스 개선에 충분히 활용하고 있는지에 대해 다시 한 번 생각해 볼 필요가 있다. 둘째, 그럼에도 불구하고 센서스 커버리지 오차에서 안정적인 패턴이 발견되었다는 것은 센서스 결과에서 누락 혹은 중복 조사된 인구의 특성을 보여주는 것으로 보정을 위한 근거를 제공해 준다고도 볼 수 있다. 20대 남자의 누락률과 중복률이 높게 나타나는데 이 집단에 대한 보정없는 센서스 결과는 왜곡된 해석을 낳을 수도 있기 때문이다.



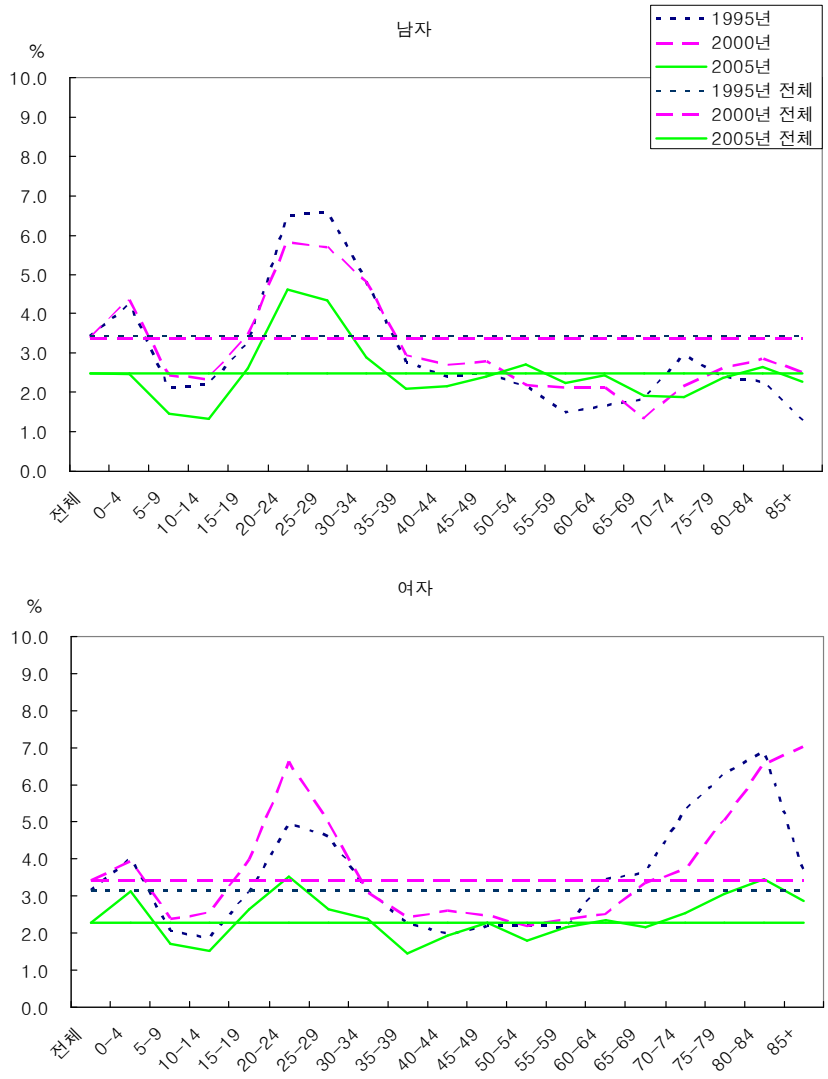
자료: 통계청, 1998, 2001, 2006a.

[그림 14] 성별 누락률과 중복률 (1995 · 2000 · 2005년)



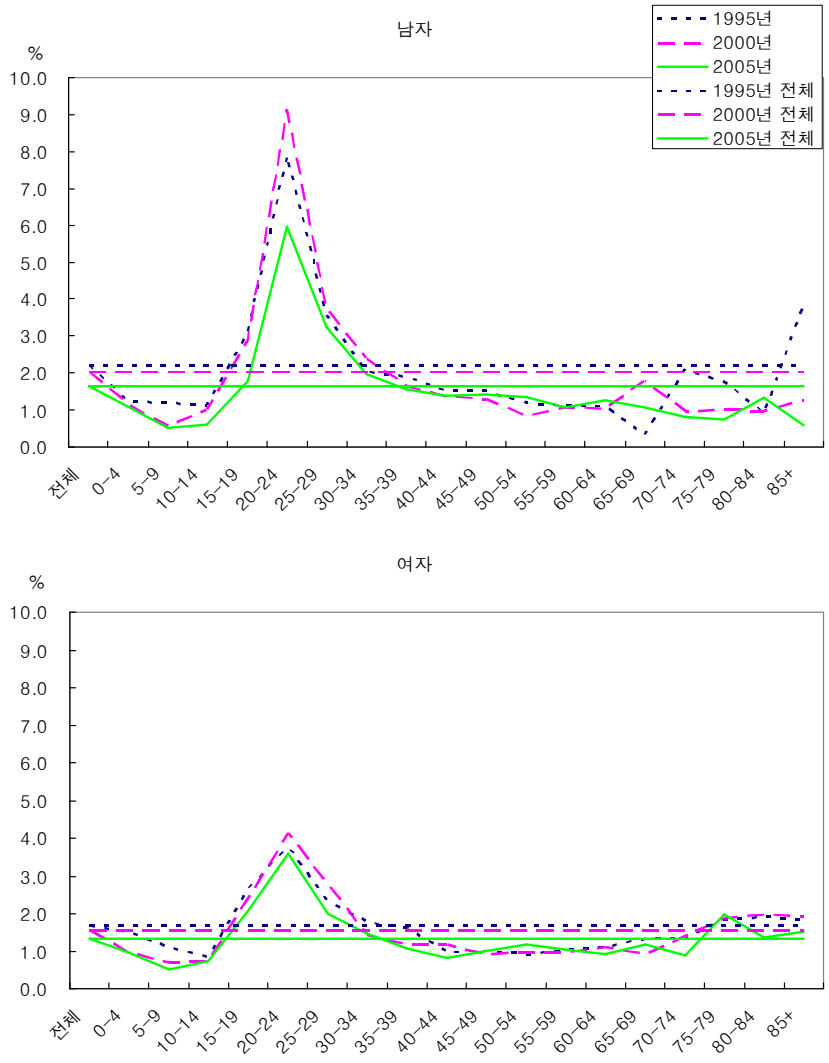
자료: 통계청, 1998, 2001, 2006a .

[그림 15] 연령별 누락률과 중복률 (1995 · 2000 · 2005년)



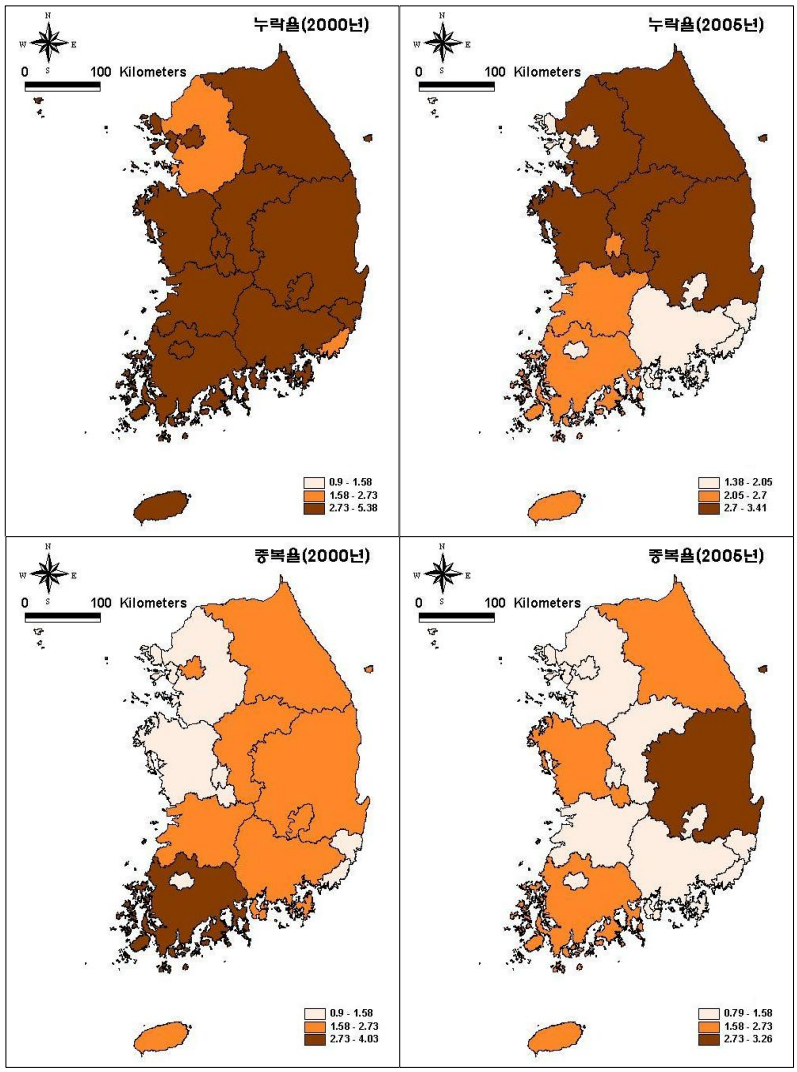
자료: 통계청, 1998, 2001, 2006a.

[그림 16] 성 및 연령별 누락률 (1995 · 2000 · 2005년)



자료: 통계청, 1998, 2001, 2006a.

[그림 17] 성 및 연령별 중복률 (1995 · 2000 · 2005년)



자료: 통계청, 2001, 2006a.

[그림 18] 시도별 누락률과 중복률 (2000 · 2005년)

2. 우리나라에 대한 시사점

원넘버 센서스 방법론과 우리나라 센서스 커버리지 오차의 현황 검토를 통해서 원넘버 센서스가 갖는 시사점을 다음의 세 가지 측면에서

제안해 보고자 한다. 첫째, 과연 우리나라에서 원넘버 센서스는 필요한가? 둘째, 우리나라에서 원넘버 센서스는 어떠한 방법으로 수행될 수 있는가? 셋째, 우리나라에서 원넘버 센서스는 실행가능한가?

가. 원넘버 센서스 필요성

아직까지 우리나라에서 원넘버 센서스에 대한 관심이 표면 위로 나타나고 있지는 않다. 우리나라 센서스 커버리지 오차를 검토해 본 결과 외국에서 원넘버 센서스가 등장하게 된 배경과 그 양상이 유사함을 볼 때 우리나라에서 원넘버 센서스에 대한 논의를 하는 것은 일면 타당하다고 하겠다.

2005년 센서스 결과, 순누락률은 0.9%, 총오차율은 3.9%로 나타났으며, 센서스인구와 이전 센서스를 기초로 하여 추정된 인구 사이에는 약 90만명, 추계인구와는 약 100만명 가량의 차이가 발생하고 있음을 알 수 있다. 물론, 행정자료의 한계로 인해서 추정 및 추계인구의 정확성에 대한 의문이 제기될 수는 있다. 반대로 센서스 혹은 사후조사의 한계가 있을 수도 있다. 그렇다고 해서 원넘버 센서스의 필요성에 대한 논의가 무위로 돌아가는 것은 아니다. 실제로 외국의 사례연구를 보면, 원넘버 센서스의 필요성이 제기된 것은 주요 하위 집단별로 커버리지 오차의 차이가 인식되면서부터이다. 영국에서는 도시 내부의 20대 남자가, 미국에서는 흑인의 누락 문제가 끊임없이 제기된 것이 원넘버 센서스의 직접적인 원인이 되었다. 우리나라의 1995 ~ 2005년간의 사후조사 결과를 비교분석해 본 결과, 20대 남자의 커버리지 오차는 다른 어느 집단에 비해서도 높은 것으로 나타났으며 이는 시계열적으로도 지속되고 있었다. 특히 20대 여자와 비교해 볼 경우, 그 차이가 2배 이상 높게 나타나므로 이 부분에 대한 보정없이 자료를 분석할 경우, 결과에 대한 왜곡된 해석이 나올 수 있다. 센서스의 완전하고도 정확한 집계를 위한 통계청의 노력에도 불구하고 이 집단의 커버리지 오차 수준이 크게 향상되지 않고 있음은 조사방법 이외의 다른 통계적인 기법을 통한 대안이 마련되어야 함을 의미한다고 볼 여지가 있다.

정확한 추정과 보정이 결합된 원넘버 센서스의 실시를 통해서 우리는 다음과 같은 잇점을 얻을 수 있다. 총수의 정확성 향상 뿐 아니라 성

및 연령과 같은 주요 하위 집단의 커버리지 오차를 제거 혹은 감소시킴으로써 분포의 정확성도 향상시킬 수 있다. 이를 통해서 최고 품질의 단일 인구(혹은 가구) 수치를 얻음으로써 우리나라 전체의 인구 규모 및 특성의 분포에 대해서 비교적 정확히 파악할 수 있다. 궁극적으로는 센서스의 활용가능성을 증대시킬 수 있다. 미국이나 영국 등에서 센서스의 정확한 집계는 의석 및 예산 배분 등과 같은 목적과 연동되어 매우 중요하게 다루어지고 있는 반면에, 우리나라에서 센서스는 법적인 근거를 가지고 활용되는 부분이 많지 않다. 센서스 결과의 정확성 향상을 통해서 많은 정책적 자료로 활용될 수 있을 것을 기대해 볼 수 있다.

나. 원넘버 센서스 방법론

사후조사 및 인구분석을 통해서 센서스의 누락 혹은 중복이 일반적으로 받아들여지고 있는 현 상황에서 일본의 원넘버 센서스 방법론 보다는 영국과 미국의 원넘버 센서스 방법론이 우리에게 적합한 것으로 판단된다. 이 두 국가에서는 센서스의 과소 혹은 과대집계를 파악하는 방법으로 사후조사를 활용하고 있다.

이와 함께 영국이나 미국에서 원넘버 센서스의 방법론 중 하나로 행정자료를 검토한 바 있다(NRC, 1994). 개인수준에서의 행정자료가 정확하다면, 센서스에서 과소 혹은 과대집계된 인구를 추정하는데 행정자료 또한 대안적인 방법이 될 수 있기 때문이다. 삼원시스템 추정법은 센서스, 사후조사, 행정자료의 세 가지 자료를 비교하여 인구를 추정하는 방법이다. 센서스에서 행정자료의 활용을 검토하고 있는 우리나라의 현 상황에서 과도기적인 단계로 세 가지 자료를 활용한 인구 추정법의 개발을 고려해 볼 수도 있을 것으로 생각된다.

1) 사후조사

영국, 미국과 마찬가지로 우리나라도 1960년대 이래 사후조사를 실시하여 센서스 커버리지 오차를 평가하고 있다. 그러나, 우리나라에서 현재 실시하고 있는 사후조사 방법이 과연 원넘버 센서스를 수행하기에 적절한지는 다시 검토해 보아야 한다.

〈표 10〉 영국, 미국, 우리나라 및 일본의 사후조사 비교

구 분	영 국	미 국	우리나라	일 본
조사방법	독립	독립	종속	독립
측정오차유형	커버리지	커버리지	커버리지 +내용	내용+ 커버리지
표본규모	약 37만가구	약 30만가구	약 4만가구	약 6만가구
표집단위	우편번호 (15가구)	블록클러스터 (30가구)	조사구 (60가구)	조사구 (50가구)
본조사후 기간	약 1개월 후	CATI: 약 3주후 CAPI:약 3.5개월후	1개월 후	2개월 후
자료수집방법	PAPI	CATI, CAPI	PAPI	PAPI
조사원 업무량	90-200가구	CATI:약 62가구 CAPI:약 47가구	약 60가구	약 100가구

주 1: 영국 2001년, 미국은 2000년, 우리나라와 일본은 2005년 사후조사 결과를 비교한 것임.
주 2: 측정오차유형은 해당 국가에서 비중있게 측정하고자 하는 오차의 유형을 먼저 기술하였음.
주 3: 미국의 경우 전체 사후조사의 면접 중 CATI를 이용한 경우가 29.4%임.

<표 10>은 영국과 미국, 우리나라의 사후조사 방법을 비교한 것이다. 일본은 다만 참고하기 위해서 표에 함께 제시하였다. 일본은 미국이나 영국과 달리 원넘버 센서스에서 사후조사가 차지하는 비중이 크지 않으며 사후조사의 목적도 커버리지 오차 보다는 내용 오차의 측정에 중점을 두고 있기 때문이다. 사후조사 비교는 조사방법 및 목적, 표본규모, 표집단위, 자료수집시기 및 방법, 조사원 체계 등의 차원에서 이루어졌다.

먼저, 자료수집방법을 비교해 보면, 영국과 미국은 독립적인 방법의, 우리나라는 종속적인 방법의 사후조사를 실시하고 있다. 종속적인 방법은 센서스 본조사와 사후조사간의 인과적인 의존성으로 인해서 커버리지 오차가 과소추정될 가능성이 높으므로 커버리지 오차에 대한 정확한 파악을 위해 독립적인 방법이 권고되고 있다(UN, 2008). 독립적인 조사

라고 해서 인과적인 의존성의 문제가 전혀 없는 것은 아니다. 영국에서 '센서스 타당성 조사(Census Validation Survey)'라는 이름으로 수행된 1991년의 사후조사는 이전의 조사에 비해서 누락률을 제대로 측정하지 못했다는 평가를 받았는데 그 가장 큰 이유가 바로 사후조사가 센서스로부터 독자적인 조사가 되지 못했기 때문이다. 이는 미국의 경우에도 마찬가지이다. 사후조사 결과가 센서스 보정에 활용되지 못한 중요한 요인 중의 하나가 바로 센서스와 사후조사간의 의존성으로 인해 발생하는 상관편향때문이다. 독립적인 조사방법에서조차도 인과적인 의존성의 문제가 완전히 해소되지 않음은 역설적으로, 종속적인 조사방법에서의 인과적인 의존성 문제의 심각성을 보여준다. 우리나라에서도 과거 두 차례에 걸쳐서 독립조사방법으로 사후조사가 수행된 적이 있다. 그러나 센서스와 사후조사간 매칭의 어려움, 인력 및 예산상의 문제로 인해서 1980년 이후에는 계속해서 종속적인 방법으로 사후조사를 실시해 오고 있다. 정확성을 목적으로 하는 원넘버 센서스를 수행하고자 한다면, 인력 및 예산상의 문제 해결을 통해서 조사방법을 독립적인 방법으로 변경할 필요가 있다.

사후조사의 표본규모를 비교해 보자. 표본규모는 원넘버 센서스에서 성공을 좌우하는 요인이라고 해도 무리가 아니다. 소지역 수준에서 인구를 추정하기 위해서 영국과 미국 모두 이전의 사후조사 보다 표본규모를 대폭 늘렸다. 영국은 1991년 2만 가구에서 2001년에 37만 가구로, 미국은 1990년 16만 5천가구에서 2000년 30만 가구로 표본규모를 확대했다. 우리나라는 현재 4만 가구를 대상으로 사후조사를 하고 있는데 이는 시도 단위로 오차율을 계산할 수 있는 수준이다. 보정된 센서스인구를 시군구 혹은 읍면동 단위로 공표하기 위해서는 더 큰 표본규모가 요구된다.

유사한 규모의 인력 및 예산을 투입하여 더 낮은 수준에서의 인구를 추정하기 위해서는 간접추정기법을 적용할 수도 있다. 소지역을 직접추정하는 것에 비해서 간접추정할 경우에 상당한 규모로 표본이 줄어들 수 있기 때문에 예산을 줄일 수 있다. 영국과 미국은 소지역 인구를 추정하기 위해서 간접추정법의 일종인 합성추정법을 사용하였다. 이는 대지역 정보를 사용하여 소지역 인구를 간접적으로 추정하는 것으로 그

방법이 비교적 간단해서 많이 사용되고 있다. 우리나라에서도 사후조사를 실시할 경우에 어느 수준까지 인구 및 가구에 대한 추정치를 산출할 것인지를 결정하고, 그 추정치를 직접 혹은 간접적으로 추정하기에 적합한 표본설계 전략을 마련해야 한다. 표본설계는 추정치의 정확성을 좌우하는 매우 중요한 일이므로 충분히 연구된 후 시행되어야 한다.

다음으로 표집단위에 대해 비교해 보자. 표집단위는 세 국가 모두 동일하게 지역에 기반하고 있다. 지역기반 표본추출은 추정치의 정확성과 함께 조사원 업무량 관리 등에서 야기될 수 있는 비용의 문제를 고려하여 채택된 것이다. 영국은 15가구 정도의 주소로 구성된 우편번호, 미국은 30가구 정도로 구성된 블록클러스터를 표집단위로 하고 있다. 우리나라의 경우에도 조사구를 표집단위로 하고 있다는 점에서는 영국이나 미국과 같다. 그러나 그 규모는 영국이나 미국에 비해 각각 4배와 2배 정도 크다. 만일 지역기반 표본추출을 유지할 경우에는 표집단위를 조사구보다 작은 규모로 하는 방안을 고려해 볼 수 있다. 미국의 경우에는 블록의 반 정도에 해당하는 규모를 표집단위로 검토하기도 하였다.

사후조사의 면접시작일은 영국이나 미국, 우리나라 모두 센서스 기준일 이후 3~4주 후이다. 이 시점은 응답자들이 센서스 시점에 해당 가구에 누가 거주했는지를 기억해내는가가 고려된 것이다. 향후 사후조사 시점은 이 기간을 유지하는 것이 바람직할 것이다.

자료수집은 영국과 우리나라는 PAPI를, 미국은 CATI와 CAPI를 혼합한 방법을 이용하고 있다. 영국도 미국과 마찬가지로 CAPI를 검토한 적이 있으나, 5~10분 가량 소요되는 면접에서 컴퓨터를 이용할 경우에 오히려 조사의 효율성을 떨어뜨릴 수 있다고 보았다. 우리나라에서도 아직까지 사후조사에서 컴퓨터를 이용하는 것은 어려운 것으로 보여진다. 장비 뿐 아니라 조사원의 컴퓨터 활용 교육 등 선결해야 할 문제들이 산적해 있다. 반면에 CATI에서 시사점을 찾아보면, 사후조사에서 전화조사방법의 도입을 검토해 볼 여지가 있다. 센서스 조사 당시에 응답자의 성명과 전화번호의 기입을 원칙으로 하고 있으므로 사후조사 표본으로 선정된 표본에 대해서는 일차적으로 전화면접을 시도하고 나머지 가구에 한해서 방문면접을 할 수 있을 것이다. 혼합방법조사(mixed-mode survey)는 현장조사의 어려움 및 비용의 증가 등에 대응하여 최근에 여

러 조사에서 채택되고 있다.

표본가구간의 거리 및 자료수집 방법 등의 차이로 인해서 조사원 업무량의 직접적인 비교는 어렵겠으나 대략적으로 보면 조사원의 업무량은 전화조사를 별도로 할 경우, 영국, 우리나라, 미국 순으로 많은 것으로 나타났다. 그런데, 우리나라 조사원의 경우 경상조사 업무를 병행함에 따라 실제로는 더 많은 업무를 하고 있는 셈이다. 영국의 경우에도 통계청 가구 담당 조사원을 고용했는데 이들은 면접 수행 그 자체가 일차적인 업무라기보다는 특별기동대(SWAT)와 같은 역할을 한다는 점에서 우리나라와는 대조적이다.

지금까지 영국과 미국, 그리고 우리나라의 사후조사 방법을 비교해 보았다. 그 결과 원넘버 센서스 실시를 위해서는 표본설계 및 자료수집 방법에서 개선의 여지가 있으며, 무엇보다도 조사방법과 표본규모에 있어서의 변경은 불가피한 것으로 보인다.

2) 주민등록자료와의 비교

사후조사를 통해서 과소집계를 파악한 후 이를 기반으로 인구를 추정하더라도 원넘버 센서스에 대한 품질평가는 반드시 이루어져야 한다. 품질평가는 행정자료와 비교되는 경우가 많다. 행정자료가 통계적인 목적으로 작성되지 않았기 때문에 개념이나 포괄범위, 정확성 등에 대해 평가가 필요하기는 하나, 일부의 행정자료는 특정 하위 영역에서는 그 포괄범위가 정확한 경우도 있기 때문이다. 예컨데, 영국의 경우 지금까지의 센서스에서 가장 집계가 어려운 그룹이었던 노인과 아동의 커버리지를 일부의 행정자료에서 거의 완전하게 재현하고 있는 것으로 평가되고 있다.

우리나라도 최근에는 국민들의 인식도 많이 바뀌고 등록제도도 선진화되면서 국민들의 연령이 출생과 동시에 등록되기 때문에 실제 연령과 등록에 의한 연령의 차이가 줄어들고 있다(김형석, 2008b). 또한 통계청에서 현재 추진 중인 현재인구는 일정한 지역수준에서는 인구 추정을 정확하게 해 줄 것으로 기대되고 있다. 현재인구는 주민등록인구에서 주민등록은 되어 있으나 해외에 거주하고 있는 인구를 추정하여 제외하

고, 동일 시점 현재 주민등록에는 포함되지 않은 상주 외국인과 주민등록 말소자를 포함하여 추정함으로써 센서스와 비교가 가능하다.[35] 만일, 현재인구가 정확하다면 센서스인구를 대체할 수 있겠으나 현재로서는 시도 수준에서 인구통계로서의 가능성을 확인한 정도라고 평가된다. 제한적이기는 하지만 시도수준에서 센서스, 사후조사, 현재인구의 세 인구를 비교하는 삼원시스템 추정방법이 검토될 수도 있다.

다. 원넘버 센서스 실행가능성

원넘버 센서스가 필요하다고 판단될 경우, 그 실행가능성을 검토해야 한다. 실행가능성은 기술적인 측면과 운영상의 측면으로 나누어 볼 수 있다. 전자는 원넘버 센서스의 추정치가 센서스 자료의 정확성을 향상시켜 줄 수 있는가를, 후자는 현재의 자원을 이용하여 공표일정에 맞추어 원넘버 센서스 수치를 제공할 수 있는가를 평가하는 것이다. 시험조사나 시범예행조사 등을 통한 경험자료가 축적되어 있지 않은 상황에서 경험에 근거한 판단을 할 수는 없지만, 원넘버 센서스의 실행가능성을 판단하는 기준을 제시할 수 있을 것이다.

1) 기술적인 실행가능성

원넘버 센서스를 통해 추정된 수치가 센서스 자료의 정확성을 향상시켜 줄 수 있는가? 원넘버 센서스 수치의 정확성은 앞서 보았듯이 총수의 정확성과 분포의 정확성으로 나누어 볼 수 있다. 해당 집단 내의 총수의 정확성이 향상되었는지, 다른 집단과 비교해 볼 때 그 오차 분포

35) 김형석(2008b)은 행정자료를 활용하여 주민등록인구를 기준인구로 사용하기 위해 4단계에 걸쳐 주민등록 인구자료를 평가하고 보정했다. 첫째, 주민등록인구 중 실제 외국에 장기간 거주하는 인구를 추정하여 주민등록인구에서 제외시켰다. 둘째, 생잔율이 1보다 큰 7세이하 주민등록 인구에 대한 보정의 필요성을 발견했다. 이 연령층의 생잔율이 1보다 큰 이유는 혼인 외의 출생이나 해외에서 출생해 귀국한 아동 때문인 것으로 파악되었다. 과거 8년간 출생 지연신고 패턴을 이용하여 2006년 1월 1일 기준으로 0세부터 7세까지 주민등록인구를 보완할 수 있었다. 셋째, 연령 및 지역 보정계수를 적용해 주민등록인구의 연령 및 거주지를 시도 수준에서 조정하였다. 마지막 단계에서는 국내에 거주하는 외국인을 추정하는 것인데, 2005년말 기준으로 법무부의 성 및 연령별 등록외국인(48만5천명)과 단기체류자 중 불법으로 전환된 외국인 총수를 성 및 연령별로 추정하여 합산하였다.

가 유사하여 분포의 정확성이 향상되었는지를 평가해 보아야 한다. 전체적으로는 커버리지 오차가 줄어들어 총수의 정확성은 향상되었을지 모르나, 이것이 어느 한 집단의 조사는 완전하지 못한 상태에서 다른 한 집단의 완전한 조사에 기반한 것이라면 분포의 정확성은 낮아질 것이다. 반대의 경우도 있다. 두 집단의 오차율이 비슷할 지라도 이 수준이 높은 수준에서라면 총수의 정확성은 낮을 수밖에 없다.

총수와 분포의 정확성을 모두 향상시키기 위해서는 정교한 사후조사의 설계에 기반한 원넘버 센서스가 수행되어야 한다. 사후조사의 중요성은 사후조사가 센서스 이후의 부가적인 조사가 아니라 센서스 과정의 일부로 포함되어야 함을 강조하며 그 이름을 통합적인 커버리지 측정조사인 ICM이라고 한 미국의 예에서 잘 알 수 있다. 즉, 원넘버 센서스에서 사후조사가 센서스 과정의 일환임을 다시 한 번 주지할 필요가 있다. 이러한 인식 하에 지금까지의 우리나라 사후조사 현황에 대한 철저한 평가가 필요하다. 사후조사에서 사용되고 있는 주요 개념과 표본추출 및 자료수집과정, 센서스 자료와의 대조작업과 오차율 계산방법 등에 대한 검토를 통해서 과연 우리나라 사후조사가 원넘버 센서스를 뒷받침할 수 있는지 냉정하게 평가해야 한다.

또한, 원넘버 센서스의 추정치에 영향을 미칠 수 있는 오차의 가능성에 대해서도 염두해 두어야 한다. 사후조사가 아무리 정교할 지라도 항상 수많은 오차에 노출되어 있다. 미국이 센서스 보정을 두고 여전히 논쟁하고 있는 이유 중의 하나가 바로 알려지지 않은 비표본오차의 가능성 때문이라는 점은 우리에게 시사하는 바가 크다. 일본에서 사후조사가 커버리지 오차 보다는 내용 오차에 대한 평가를 중심으로 하는 것도 유사한 이유에서이다.

2) 운영상의 실행가능성

다음으로 원넘버 센서스의 운영이 실행가능한지에 대한 평가를 해야 한다. 이 평가는 인력과 예산의 동원이 가능한지, 실제 운영과정이 계획대로 진행될 수 있는지, 자료 공표는 적정한 시간에 가능한지를 중심으로 이루어진다.

가) 인력과 예산

원넘버 센서스 추정치의 정확성 못지않게 중요하게 고려되어야 할 부분이 인력과 예산이다. 사후조사 방법을 통해서 소지역 인구를 직접적으로 추정하기 위해서는 대규모의 표본이 필요한데 이는 많은 인력과 예산을 필요로 하는 일이다. 정확성 향상을 위해서 일정 정도의 예산 증가는 불가피하나, 예산의 증가분이 과연 정확성 향상을 담보할 수 있는지에 대해서는 분석을 해보아야 한다.

인력과 예산이 표본규모의 영향만을 받는 것은 아니다. 정확한 추정치를 얻기 위해서는 센서스 본조사와 사후조사의 가구 및 개인에 대한 정확한 매칭이 중요하다. 영국이나 미국의 경우, 자동매칭과 함께 수동매칭을 하며, 수동매칭은 전문적으로 훈련된 매칭요원이 3단계를 거쳐서 진행할 정도로 노동집약적인 작업이다. 매칭 이외에도 많은 부분에서 새로운 인력을 필요로 하며 이는 예산과 직결된다.

여기에서는 표본규모의 증가에 따라서 필요한 조사원과 조사비용의 규모를 가늠해보고자 한다. 정교한 수준의 표본규모를 산출하기 위해서는 별도의 연구과정이 필요하다. 본 연구에서는 영국과 유사한 방법으로 표본을 추출하여 그 비용을 대략적으로 산출해 보고자 한다. 전광희(2008b)는 영국과 마찬가지로 50만 규모의 설계 그룹을 구성하여 이 그룹의 직접추정치를 이용, 시군구별로 간접추정치를 얻는 전략을 제안하였다. 이에 따르면 표본규모는 약 30만 가구가 산출된다.

지역간의 경계를 고려하지 않고 표본규모 30만가구에 대한 조사원 비용을 2010년 센서스 2차 시험조사의 사후조사 예산배정안(통계청, 2008b)에 근거하여 산정해 보면 다음과 같다. 조사원 1인당 채용단가는 41,620원이며, 1일 업무량은 전수조사의 보통조사구 업무량과 동일한 12가구이다. 2005년과 동일하게 조사기간을 5일로 할 경우에 조사원 1인당 업무량은 60가구가 된다. 계산을 해 보면, 조사원은 약 5,000명 정도가 필요하며, 이들에 대한 비용은 약 10억원으로 계산된다(5,000명 × 5일 × 41,620원=1,040,500,000원). 이는 2005년 사후조사 비용 (약 5억원)보다도 많은 것이다. 물론 이는 대략적으로 계산한 조사원 규모와 비용이며 실제로 원넘버 센서스 수행에 필요한 비용은 이를 포함하여 더 많이 소요될 것은 자명한 일이다. 그러므로 원넘버 센서스를 수행하기 위

해서 이에 준하는 인력과 예산이 동원가능한지에 대한 검토가 사전에 이루어져야 할 것이다.

나) 자료의 공표

원넘버 센서스는 언제 완료되어야 하는가? 센서스 자료의 공표를 신속히 하기 위해 노력하고 있는 상황에서 원넘버 센서스의 공표시기에 대한 검토는 원넘버 센서스 실행가능성을 평가하는데 중요한 기준이 될 것이다. <표 13>은 2005년 센서스의 전수자료 공표시기를 나타낸 것이다. 인구, 가구 및 주택의 규모에 대한 잠정집계는 센서스 시점 이후 약 1개월 후인 12월에, 인구부분에 대한 전수집계결과는 그 이듬해 5월에, 가구 및 주택부분에 대한 결과는 7월에 공표되었다.

먼저, 2005년도 기준과 동일하게 공표 시점을 맞추어야 하는지 혹은 공표시점 조정이 가능한지를 결정해야 한다. 사후조사 결과에 대한 잠정집계는 약 1개월 후인 2005년 12월 말에, 최종 결과표 작성 및 분석은 2006년 4월에 완료되었는데, 사후조사 완료 이후에도 매칭, 추정 및 보정을 해야 하므로 실질적으로 2005년 센서스 전수집계 공표시기를 맞추기는 수월하지 않다. 그러므로 원넘버 센서스의 실시가 결정된다면 언제까지 공표할 것인지 공표시기에 대한 사전 합의가 있어야 한다.

〈표 11〉 우리나라의 2005년 센서스 자료 공표 시기

시 기	구 분	내 용
2005년 12월	잠정집계결과	읍면동별로 조사결과를 잠정집계하여 제출한 집계표를 기초로 전국 및 지역별(시도, 시군구, 읍면동) 인구, 가구 및 주택규모 공표
2006년 5월	전수집계결과 (인구부문)	연령별, 교육정도별, 혼인상태별, 가구주와의 관계별 인구 등의 내용을 분석하여 공표
2006년 7월	전수집계 결과 (가구 및 주택부문)	가구규모, 가구구성, 가족형태, 가구주 특성과 가구의 주거형태, 주택수, 주택유형, 주택규모, 건축연도별 주택 등에 관한 내용을 분석하여 공표

자료의 공표와 관련하여 공표 방법에 대한 논의도 필요하다. 이는 자료의 시계열 문제와 연결된다. 지금까지는 보정 전의 결과가 공식통계였다면, 원넘버 센서스를 수행할 경우 보정 후의 결과가 공식통계가 된다. 이는 시계열의 단절을 의미할 수도 있으며 이에 따라서 사용자들에게 많은 혼란을 줄 수 있다. 그러므로 보정후의 자료 공표를 어떻게 할 것인가에 대한 논의가 필요하다. 미국의 경우, 1990년 센서스 보정 전후의 자료를 홈페이지에서 동시에 제공하고 있다. 본래는 보정전의 자료만 공표하였으나 보정 후 자료에 대한 요구가 증가하면서 이러한 결정을 내린 것이다. 만일 우리나라에서 원넘버 센서스를 실시한다면 이와 반대의 상황이 발생할 수도 있다.

다) 시험조사 및 시범예행조사를 통한 평가

원넘버 센서스 추정치의 정확성을 향상시킬 수 있는 방법론이 완성되고 인력 및 예산이 확보되면, 원넘버 센서스가 정해진 기간에 성공적으로 수행될 수 있는가에 대한 전반적인 평가가 필요하다. 주요 일정에 맞게 진행될 수 있는지, 운영과정 상에 위험 변수는 발생하지 않는지를 시험조사 및 시범예행조사 과정에서 확인해야 할 것이다.

제4절 결 론

본 연구는 조사환경의 악화와 응답자의 부담 증가, 해마다 늘어나고 있는 비용, 그리고 자료의 정확성에 대한 신뢰의 문제 등 현재 센서스가 직면하고 있는 어려움에 대한 대안을 마련하고자 하는 일환으로 계획되었다. 원넘버 센서스는 이 중에서도 정확한 자료의 작성이라는 목적을 갖고 출발하였다.

센서스는 일정한 지역에 거주하고 있는 모든 인구를 빠짐없이 조사해야 함에도 불구하고 조사의 전 과정에 개입하는 여러 가지 오차로 인해 항상 불완전하게 마련이다. 이에 일부 국가에서는 현장조사 이외에 추정이나 보정과 같은 통계적 기법을 접목시킴으로써 센서스의 완전성을 지향하고자 한다. 우리나라는 아직까지 센서스 커버리지에 대한 관

심이 높은 편은 아니나, 우리나라의 센서스 자료 역시 커버리지가 완전하지 않으며 특히 특정 집단은 다른 집단에 비해서 지속적으로 높은 커버리지 오차를 보이고 있는 것으로 나타나 이에 대한 대안마련이 요구되는 것으로 분석되었다. 이에 본 연구에서는 원넘버 센서스에 대한 주요 방법론을 검토해 보고, 우리나라에서 원넘버 센서스의 필요성과 실행가능성을 평가해 보았다.

방법론 검토는 영국, 미국, 일본의 통계기관에서 수행하고 있는 원넘버 센서스를 살펴보는 식으로 진행되었다. 원넘버 센서스는 아직까지 일반적인 형태의 센서스 방법론이 아니므로 용어 조차 생소할 수 있다. 그러나, 점차 센서스의 커버리지 오차 문제가 심각해지면서 센서스 자료의 정확성에 대한 의문이 제기되자 원넘버 센서스에 주목하고 있는 국가들이 많아지고 있다. 일본은 명시적으로 원넘버 센서스를 실시하고 있지는 않으나 원넘버 센서스의 개념을 좀 더 확대해 보면 누락과 중복이 없다고 전제되는 센서스인구가 모든 인구통계의 중심에 있다는 점에서 광의의 원넘버 센서스 실시 국가로 포함시킬 수 있다. 그러나, 우리나라가 시사점을 얻을 수 있는 원넘버 센서스의 사례는 영국과 미국에서 시도한 것이다. 이 두 국가는 센서스가 조사대상을 완전히 집계하지 못하고 있다고 보고, 누락 혹은 중복된 부분을 사후조사를 통해서 파악한 후 그 결과를 센서스 본조사 결과와 통합하는 방법론을 취하고 있다.

사후조사 방법에 근거한 원넘버 센서스가 과연 우리나라에서 실행가능한가를 검토하기 위해서 우리나라의 사후조사를 살펴보았다. 분석결과, 사후조사가 원넘버 센서스의 목적에 부합하기 위해서는 여러 가지 측면에서 개선의 여지가 있는 것으로 나타났다. 특히 소지역 수준에서 정확한 인구를 추정하기 위해서는 현재의 종속적인 조사방식을 독립적인 조사방식으로 변경하고, 표본규모를 대폭 증가하는 등의 전환이 필요하다. 그리고 표본설계 및 오차율의 추정 등과 관련하여 통계적인 기법에 대한 연구가 뒷받침되어야 한다.

원넘버 센서스는 방법론 뿐 아니라 운영상의 실행가능성이 함께 고려되어야 한다. 센서스 자체가 워낙 큰 조사이므로 작은 것처럼 보이는 변화 조차도 예상치 않은 많은 인력과 예산을 필요로 하는 경우가 있기 때문이다. 하물며 센서스의 방법론을 개선하는 일에 있어서 인력과 예

산 등과 같은 여러가지 행정적인 문제를 사전에 검토해야 함은 두말할 나위가 없다. 영국이나 미국 등의 예에서 보듯이 원넘버 센서스를 계획하면서 시도한 가장 큰 변화 중의 하나는 사후조사의 표본규모 증대이다. 영국에서는 이전 사후조사에 비해서 약 18배 정도로 표본규모를 증가시켰다. 이는 필연적으로 인력 및 예산의 증가와 맞물린다. 공표시점 또한 중요하다. 원넘버 센서스가 사후조사 결과를 바탕으로 본조사 결과를 보정하는 것이므로 2005년과 비교할 경우 자료의 공표시점이 늦어질 가능성이 높다. 그러므로, 센서스 결과의 신속한 집계로 높은 평가를 받고 있는 현 상황에서 원넘버 센서스의 공표시점에 대한 합의가 이루어져야 할 것이다.

결론적으로, 본 연구를 통해서 사후조사 방법에 의한 원넘버 센서스를 당장에 실시하는 것이 쉽지 않음을 알 수 있었다. 영국이나 미국의 원넘버 센서스는 관계기관 간의 수차례 검토 끝에 결정된 것이다. 그 진행과정을 보면 학계의 자문을 통해서 이론적인 모형을 구축하고 이는 여러 차례의 시험조사를 통해서 검증되었다. 또한 실무적인 지침 하나하나 시험조사 및 시범예행조사를 통해 검토하는 등 철저한 준비를 하였다. 원넘버 센서스는 전통적인 현장조사 방법에 추정이나 보정과 같은 통계적인 기법을 결합하여 최종 수치를 작성한다는 점에서 기존의 센서스와는 차별적이다. 또한 대규모의 인력과 예산을 필요로 하는 일이므로 사후조사 방법에 의한 원넘버 센서스에 대한 결정은 신중히 내려져야 할 것이다. 무엇보다도 주지해야 할 점은 원넘버 센서스 자체도 완전하지 않을 수 있다는 점이다. 원넘버 센서스를 통해서 작성된 추정치는 다른 자료에 의해서 항상 비교 검토되어야 한다. 특히 측정되지 않은 비표본오차의 가능성으로 인해서 원넘버 센서스 추정치의 공식화를 보류한 미국의 결정은 우리에게 시사하는 바가 크다.

참고문헌

경제기획원(1982),「1980년 인구 및 주택 센서스 사후조사 결과 보고」. 내부자료.

김민경(2002),「인구센서스의 이해」. 글로벌.

김형석(2008a), "Register-based 센서스, Rolling 센서스, One-number 센서스 통합방안." 인구주택총조사 개선 T/F 발표자료.

_____(2008b), "주민등록인구 자료의 평가 및 보정 방안." 인구학회 발표자료.

이지연(2004), "범위오차의 평가와 원인분석." 2004년도 연구결과 모음집. 통계청.

_____(2006), "2005년 인구주택총조사의 범위오차평가와 지역별 차이 연구." 2006년도 연구결과 모음집. 통계청.

전광희(2007), "영국의 2001년 One Number Census 프로젝트: 방법론의 골자와 2011년의 준비과정을 중심으로." 2005년 인구주택총조사 종합분석. pp259-314. 통계청.

_____(2008a), "한국판 One Number Census: 도전과 전망." 경제통계 및 인구주택총조사의 행정자료 활용을 위한 국제세미나 자료집: 297-310. 통계청.

_____(2008b), "One Number Census 방법론과 도입전략." 인구주택총조사 개선 T/F 발표자료.

통계개발원(2008), 고용통계 소지역 추정 및 사후조사와 인구추계. 내부자료.

통계청(1998),「1995 인구주택총조사 사후조사 결과분석」. 내부자료.

_____(2001),「2000 인구주택총조사 사후조사 결과분석」. 내부자료.

_____(2005a), "2005 인구주택총조사 사후조사 표본규모 및 표본추출(안)." 내부자료.

_____(2005b),「2005 인구주택총조사 사후조사 지침서」. 내부자료.

_____(2006a),「2005 인구주택총조사 사후조사 결과분석. 내부자료.

_____(2008a), "2008 현재인구 작성계획(안)." 내부자료.

_____(2008b), "2010 인구주택총조사 제2차 시험조사 사후조사 실시계획(안)." 내부자료.

Anderson, M. and S. E. Fienberg(2002), "Why Is There Still a Controversy About Adjusting the Census for Undercount?" *www.apsanet.org*.

Brown, J. J., I. D. Diamond, and R. L. Chambers(1999), "A Methodological Strategy for a One-Number Census." *Journal of Statistical Society Association* 162(2): 247-267.

Brunell, T. L.(2002), "Why There Is Still a Controversy About Adjusting the Census." *www.apsanet.org*.

Cantwell, P. J. and M. Ikeda(2003), "Handling Missing Data in the 2000 Accuracy and Coverage Evaluation Survey." *Survey Methodology* 29(2): 139-153.

Chandraseka. C. and W. Deming(1949), "On a Method of Estimating Birth and Death Rates and the Extent of Registration." *Journal of the American Statistical Associations* 44: 101-115.

Freedman, D. A.(1991), "Adjusting the 1990 Census." *Policy Forum:* 1233-1236.

Hogan, H.(1993), "The 1990 Post-Enumeration Survey: Operations and Results." *Journal of the American Statistical Association* 88(423): 1047-1060.

Ikeda, M. and J. Tsay(2003), "Transparent File Construction for the State of New Jersey in Census 2000." *Research Report Series(Statistics #2003-03)* U.S. Census Bureau.

Ishihara. H.(2008), “Stragey for the 2010 Population Census, and the Statistical Use of Administrative Data and Its Issues." 경제통계 및 인구주택총조사의 행정자료 활용을 위한 국제세미나 자료집: 287-294. 통계청.

NRC(1994), *Counting People in the Information Age*. National Academy Press.

____(1995), *Modernizing the U.S. Census*. National Academy Press.

____(1999), *Measuring A Changing Nation: Modern Methods for the 2000*

Census, Washington, D.C.: National Academy Press.

____(2004), *The 2000 Census: Counting Under Adversity*, Washington, D.C.: National Academy Press.

____(2006), *Once, Only Once, and in the Right Place,* Washington, D.C.: National Academy Press

ONS(1998), "Census Coverage Survey Design." *One Number Census Steering Committee Paper 98/12*. Office for National Statistics, Britain.

____(1999), "A Donor Imputation System a Create a Census Database Fully Adjusted for Underenumeration." *One Number Census Steering Committee Paper 99/08*. Office for National Statistics, Britain.

____(2000a), "Estimation Strategy for Design Group Estimates by Age and Sex from the Census Coverage Survey." *One Number Census Steering Committee Paper 00/03A*. Office for National Statistics, Britain.

____(2000b), "A Quality Assurance and Contingency Strategy for the One Number Census." *One Number Census Steering Committee Paper 00/04*. Office for National Statistics, Britain.

____(2001a), "One Number Census Methodology." *One Number Census Steering Committee Paper 01/01*. Office for National Statistics, Britain.

____(2001b), "Census 2001: A Guide to the One Number Census." Office for National Statistics, Britain.

____(2001c), "One Number Census Matching." *One Number Census Steering Committee Paper 01/05*. Office for National Statistics, Britain.

____(2005), *Census 2001 Review and Evaluation*. Office for National Statistics, Britain.

Pereira, R.(2002), *The Census Coverage Survey-the key elements of a One Number Census*. Office for National Statistics, Britain.

Prewitt, K.(2000), A*ccuracy and Coverage Evaluation: Statement on the Feasibility of Using Statistical Methods to Improve the Accuracy of Census 2000*. U.S. Census Bureau.

Robinson, G. J. and A. Adlakha(2002), "Comparioson of A.C.E. RevisionⅡ

Results with Demographic Aanlysis." *DSSD A.C.E. Revision Memorandum Ⅱ Series # PP-41*. U.S. Census Bureau.

SBJ(2008a), 2005년 국세조사 사후조사. 내부자료.

___(2008b), 주민기본대장인구요람. 내부자료.

___(2008c), 센서스인구와 주민기본대장인구의 비교. 내부자료.

Shindler, E.(1999), Comparison of Dual system Estimation A and C. *Census 2000 Dress Rehearsal Evaluation Memorandum C8a.*

Schindler, E. and A. Navarro(1993), "Census Plus: An Alternative Coverage Methodology." *Proceedings of the Survey Research Methods*, 248-253.

Steele, F.(2002), "A Controlled Donor Imputation System for a One-Number Census." *Journal of Statistical Society Association* 165(3): 495-522.

Takami(2003), "Eavaluation of the Accuracy of the 2000 Population of the Census." Paper to be presented at the 21st Population Census Conference November 19-21, 2003 Tokyo, Japan.

UN(2008), *Principles and Recommendations for Population and Housing Censuses, Rev. 2*. Statistical Papers Series M No. 67/Rev.2.

U.S. Census Bureau(1999), *Census 2000 Dress Rehearsal Evaluation Summary*.

____(2004), *Accuracy and Coverage Evaluation of Census 2000: Design and Methodology*.

Wright, T.(1999), "A One-Number Census: Some Related History." *Science* 283: 491-492.

Watcher, K.(2008), "The Future of Census Coverage Surveys." *IMS Collections Probability and Statistics: Essays in Honor of David A. Freedman*. Vol.(2): 234-245.

http://www.census.gov

http://www.stat.go.jp