조사동향

2008 서울서베이 표본추출틀 구축 및 표본추출 사례 연구

A Case Study on the Construction of the Sampling Frame and Sampling Design for 2008 Seoul Survey

강현철* · 박승열** · 김지연*** · 김인수**** · 이동수***** · 황재일****** · 박민규*******
Kang, Hyuncheol · Park, Seungyeol · Kim, Jeeyoun · Kim, Insoo · Lee, Dongsu · Hwang, Jaeil · Park, Mingue

추출된 표본을 바탕으로 관심 모집단의 특성을 파악하는 조사연구에 있어서는 실제로 표본이 추출되는 표본추출틀의 모집단 대표성이 매우 중요하다. 표본추출틀이 관심 모집단을 적절한 수준으로 포함하지 못하는 경우 심각한 표본추출틀 편향이 발생하게 되고 이로 인하여 효율적인 추출법에 의하여 추출된 표본의 통계적 신뢰도 역시 손상된다. 그러나 대규모 조사를 위한 표본추출틀의 구축은 시간과 비용의 측면에서 비효율적이고 따라서 국가에서 제공하는 전수조사 기반의 표본추출틀이 흔히 사용된다. 대표적으로 국내의 가구조사를 위한 표본추출틀로는 매 5년마다 시행되는 인구주택총조사 기반의 자료가 사용된다. 그러나 인구주택총조사 기반 표본추출틀의 경우 인구주택총조사 시점과 실제 조사 시점과의 시간적 차이로 인한 표본추출틀의 모집단 대표성에 문제가 발생하게 된다. 특별히 인구 유동성이 심한 서울과 같은 대도시의 경우 시간의 경과에 따른 모집단 분포의 변화가 심하게 나타나리라 예측할 수 있다. 따라서 본 연구에서는 2008 서울서베이 가구 조사를 위해 새롭게 표본추출틀을 구축한 것과 새 표본추출틀을 기초로 하여 표본을 추출한 사례를 다룬다. 기존 인구주택총조사 기반 표본추출틀이 시간이 지남에 따라 대표성을 상실하는 문제점을 지적하고 주민등록 DB와 과세대장 DB를 기반으로 한 새로운 표본추출틀을 2008년 서울서베이 가구조사를 위한 표본추출틀로 제시하였다. 새롭게 작성된 표본추출틀로부터의 가구표본추출과정과 가중치 및 모평균 추정량 또한 제시되었다.

* 호서대학교 정보통계학과 부교수
** (주)월드리서치 대표이사
*** (주)밀워드브라운 미디어리서치 상무
**** 서울시 정보화기획담당관실 통계분석팀장
***** 서울시 정보화기획담당관실 통계분석팀 주임
****** 서울시 정보화기획담당관실 통계분석팀 주임
******* 교신저자(corresponding author) : 고려대학교 통계학과 부교수 박민규.
E－mai : mpark2@korea.ac.kr

주제어 : 표본추출틀, 인구주택총조사, 주민등록 DB, 서울서베이

For a survey research in which the characteristics of the population of interest are investigated from a sample, representativeness of the sampling frame is one of the most important part to be considered. If the sampling frame fails to represent the population properly, statistical procedures based on the even efficient sampling design result in significant nonsampling biases and thus the statistical validities of the results could be damaged. But the construction of the reliable sampling frame that covers the population properly costs money and time and thus the sampling frame based on a census or a large scale survey is often used in practice. For example, the sampling frame based on the population households census is used for many household surveys in Korea. But due to the time difference between the census and a survey of interest, the sampling frame constructed from the census is expected to fail to cover the population of interest. Especially, one could expect a large amount of population and household movement in a large city like Seoul. Thus in our research, we considered the construction of new sampling frame and the procedure of sample selection for 2008 Seoul survey. We analyzed the sampling frame based on 2005 population households census and found that it does not represent the population properly. Thus, we proposed a new sampling frame based on resident registration DB for 2008 Seoul survey. We also proposed the sampling weights and estimator of the population mean based on the sample selected from the newly constructed sampling frame.

Key words : sampling frame, population households census, residents registration DB, Seoul survey

Ⅰ. 서론

세계적으로 사회구조의 변화가 빠르게 나타나고 지역단위의 경쟁력이 주요 이슈로 등장하고 있는 시기를 맞아, 서울시에서는 빠른 사회변화의 흐름을 수용하는 동시에 지역별 특성을 반영하여 시정운영에 적절히 활용하기 위해 적절한 통계작성의 필요성을 인식하게 되었다. 또한 민선자치시대를 맞이하여 주기별로 달라지는 시정운영 현황을 파악하고, 정책방향 설정 및 운영 등에 다양하게 활용될 수 있는 통계를 생산할 필요를 느끼게 되었다. 이러한 필요성을 바탕으로 서울시는 행정자료 및 가구조사, 사업체 조사를 통하여 도시발전 지향을 지속적으로 모니터링 할 수 있는 정책지표의 개발 및 구축을 목적으로 하는 '서

울서베이' 사업을 2003년부터 매년 시행하고 있다.

서울시의 각 시기별 상태에 대한 객관적이고 심층적인 분석과 도시발전 수준을 모니터링 하기 위한 도시정책지표의 지속적인 개발을 목적으로 수행 중인 서울서베이는 크게 가구조사, 외국인 조사 그리고 사업체 조사의 3개 부분으로 구성되어 있다. 가구조사의 조사내용은 인구, 경제, 도시발전과 주거, 문화, 관광, 복지, 여성과 가족, 환경, 교통 정보와 참여, 안전과 재난, 가치와 의식 등의 총 12개 분야별 지표와 관련된 문항으로 구성되어 있다. 외국인 조사는 동거 유형, 삶의 질 만족도 그리고 서울의 대표적 관광지를 묻는 단순한 문항으로 구성되어 있으며, 사업체 조사는 조세 부담, 사업상의 규제 그리고 행정 편의성과 관련된 내용으로 구성되어 있다.

서울 도시정책지표조사로 명명되며 서울서베이의 가장 큰 부분을 차지하는 가구조사는 서울시 거주 가구(또는 세대)와 15세 이상의 모든 가구원(또는 세대원)을 목표 모집단으로 정의하고 있으나, 인구와 가구유동성이 높은 서울시의 특성상 목표 모집단과 정확히 부합하는 적절한 표본추출틀이 존재하지 않으며 목표 모집단과 부합하는 표본추출틀의 설계 또한 용이하지 않다. 본 연구에서는 먼저 2005년 인구주택총조사 기반 표본추출틀을 2008년 조사를 위해 사용할 때 발생할 수 있는 대표성 문제를, 통계청에서 제공하고 있는 2005년 인구주택총조사 기반 통계와 2008년 추계 통계를 바탕으로 살펴본다. 또한 2005년 인구주택총조사 기반 표본추출틀의 시의성 문제를 해결하기 위한 방안으로 주민등록 DB와 과세대장 DB를 이용한 시의성 있는 표본추출틀의 구축방법을 소개하고 이로부터 2008 서울서베이 가구조사를 위한 표본추출과정을 소개한다.

Ⅱ. 표본추출틀의 구축

2008년 서울서베이 가구조사의 목표 모집단은 2008년 9월 1일 기준 서울시 전체 거주 가구(또는 세대)와 15세 이상의 모든 가구원(또는 세대원)으로 정의된다. 대부분의 전국단위 또는 대형 가구조사의 경우 통계청에서 제공하는 인구주택총조사의 90% 자료를 표본추출틀로 사용하게 된다. 참고로 인구주택총조사 10% 자료는 통계청 특수목적에 의하여 조사 시 long form을 이용하여 구축된 자료로 일반에게 표본추출틀로 제공되지 않는다. 인구주택총조사의 90% 자료를 표본추출틀로 사용하는 경우 기본 관측단위는 가구가 되며 1차 추출단위로는 근접 가구들의 집합인 조사구가 주로 사용된다. 통계청에서 정의한 가구의 정의는 '1인 또는 2인 이상이 모여 취사, 취침 등 생계를 같이 하는 생활단위'로서 일반가구

와 집단가구로 구분된다. 일반가구와 집단가구의 정의는 아래와 같다.

1. 일반가구
   ① 통상 가족단위로 이루어져서 생활을 같이 하고 있는 가구(혈연가구)
   ② 친구 또는 혈연관계가 없는 사람들끼리 모여 생활을 같이 하고 있는 가구(비혈연 5인 이하 가구)
   ③ 혼자서 살림하는 가구(1인 가구)

2. 집단가구
   ① 집단시설가구: 기숙사, 고아원, 양로원, 모자원, 특수병원 등의 사회시설 내에서 생활하는 가구
   ② 비혈연 6인 이상 가구 : 혈연관계가 없는 6인 이상의 사람들이 모여 동일한 거처 내에서 생활을 같이 하고 있는 가구

표본추출을 위해 인구주택총조사 기반의 표본추출틀을 이용하는 많은 조사들은 통상 일반가구만을 대상으로 하는데, 여기에는 비혈연가구가 포함되며 가정부, 기타 가사사용인, 동거인(점원, 견습인, 하숙인) 등이 가구원으로 간주된다. 이와는 달리 세대는 주로 가족 몇 사람이 동거하여 소득을 통합하고, 공동으로 지출·구입·소비를 하는 하나의 조그만 경제집단을 말한다. 따라서 동일가족이라 해도 별거하여 수입과 지출을 달리하는 경우에는 그 세대에 포함될 수 없으며, 가족이 아닌 타인이 동거하여 수입과 지출을 같이하는 경우에는 세대원으로 간주된다. 경제적 의미가 부여된 세대는 가구보다 조금 더 세밀하게 생활단위를 구분하는 경향이 있음이 알려져 있으며 세대의 이동 및 주소와 세대주 같은 기본적인 정보는 주민등록의 내용을 통해 상시적으로 갱신된다. 실제 많은 경우 세대와 가구는 거의 유사한 개념으로 사용되고 있으며 통계청 주관 조사의 경우 가구를 기준으로 한 가구통계를 제공하며, 기타 행정자료의 경우는 주민등록상에 등록된 가구 즉 세대를 기준으로 한 통계를 제공하고 있다. 최근 제 5차 국민건강조사와 같은 1년 또는 단기 주기의 조사의 경우, 인구주택총조사 기반의 가구조사 대신 주민등록 자료 기반의 세대조사를 이용하는 사례가 늘고 있다.

2008년 서울서베이 가구조사를 위한 표본추출틀은 세대를 기본단위로 하는 주민등록 자료를 기본적으로 사용하였다. 이는 경제적 생활단위인 세대가 가구에 비해 본 조사의 주 내용인 경제 및 문화활동의 기반이 된다는 점과 더불어 상시 갱신되는 세대 단위의 주민등록자료가 5년 단위로 갱신되는 인구주택총조사 자료보다 목표 모집단을 잘 반영한다고 여겨지기 때문이다. 실제로 2008년 서울서베이를 위한 표본을 인구주택총조사 기반 표본추출틀에서

추출할 경우 이용 가능한 최근 자료는 2005년에 조사된 자료이며, 따라서 2005년 자료를 바탕으로 작성된 표본추출틀을 이용할 경우 매해 시행되는 서울서베이 조사 결과의 시의성에 대한 문제가 있으리라 판단된다. 이는 특히 가구와 인구의 유동성이 많은 서울과 같은 대도시의 경우 3년 간 발생하는 모집단의 분포 변화는 상당하리라 예측할 수 있기 때문이다.

〈표 1〉에서 〈표 4〉까지는 통계청(KOSIS: www.kosis.kr)에서 2005년 인구주택총조사 결과를 바탕으로 작성한 연령별, 성별 모집단 전체 인구분포와 2008년 추계 인구분포를 나타낸다. 2008년 추계인구자료의 경우 각 광역시·도별 통계가 제공되지 않아 2005년과 2008년 인구분포의 비교를 위해서는 전국 기준 분포를 고려하였다. 〈표 1〉을 통해 2005년 기준 20대의 전체 인구 대비 비율은 감소하였으며, 60대 이상의 노인층의 비율은 증가했음을 알 수 있다. 〈표 3〉은 2005년 기준 각 연령대별 전국 대비 서울시의 인구비율을 나타내고 있다. 전국 대비 20~30대의 구성비가 서울에서 높게 나타나며 60대 이상의 노인층의 비율은 낮게 나타나고 있다. 이는 인구의 증감 폭이 크게 나타나는 연령대의 서울 인구비율이 상대적으로 크게 나타나고 있음을 의미하며, 따라서 2005년 인구주택총조사 기반의 표본추출틀을 표본추출을 위하여 사용할 경우 추출된 표본이 서울시의 연령별 분포를 왜곡할 가능성이 있음을 시사한다. 또한 〈표 2〉는 2008년 남·여의 성비가 2005년과 반대로 나타남을 보여주고 있다. 〈표 4〉에 나타난 전체 인구 대비 서울시 인구의 성별 비율이 전국 성별 분포와 유사함을 볼 때, 2005년 인구주택총조사 기반 표본추출틀을 2008년 서울서베이를 위해 사용할 경우 성별 분포의 왜곡이 발생할 수 있음을 짐작할 수 있다.

〈표 1〉 2005년 모집단 및 2008 추계 연령별 전국 인구분포

(단위: 명)

연령별	2005 인구주택총조사		2008 추계인구	
0~9세	5,551,237	11.80%	5,092,743	10.48%
10~19세	6,535,414	13.89%	6,642,016	13.66%
20~29세	7,333,970	15.59%	7,181,464	14.77%
30~39세	8,209,067	17.45%	8,283,010	17.04%
40~49세	8,023,940	17.06%	8,360,030	17.20%
50~59세	5,133,735	10.91%	6,018,796	12.38%
60~69세	3,568,920	7.59%	3,829,324	7.88%
70~79세	2,019,604	4.29%	2,372,507	4.88%
80세 이상	665,547	1.41%	826,897	1.70%
합 계	47,041,434	100.00%	48,606,787	100.00%

〈표 2〉 2005년 모집단 및 2008 추계 성별 전국 인구분포

(단위: 명)

성 별	2005 인구주택총조사		2008 추계인구	
남	23,465,650	49.88%	24,415,883	50.23%
여	23,575,784	50.12%	24,190,904	49.77%

〈표 3〉 2005년 모집단 전국 및 서울시 연령별 인구분포

(단위: 명)

연령별	2005 인구주택총조사			
	전 국		서울특별시	
0~9세	5,551,237	11.80%	991,679	10.16%
10~19세	6,535,414	13.89%	1,243,130	12.73%
20~29세	7,333,970	15.59%	1,835,235	18.80%
30~39세	8,209,067	17.45%	1,783,293	18.27%
40~49세	8,023,940	17.06%	1,633,559	16.73%
50~59세	5,133,735	10.91%	1,163,035	11.91%
60~69세	3,568,920	7.59%	701,502	7.19%
70~79세	2,019,604	4.29%	307,405	3.15%
80세 이상	665,547	1.41%	103,708	1.06%
합 계	47,041,434	100.00%	9,762,546	100.00%

〈표 4〉 2005년 모집단 전국 및 서울시 연령별 인구분포

(단위: 명)

성 별	2005 인구주택총조사		
	전 국	서울특별시	
남자(명)	23,465,650	4,837,112	20.61%
여자(명)	23,575,784	4,925,434	20.89%

〈표 5〉 2005년 모집단 및 2008 추계 가구구성별 서울시 가구분포

(단위: 가구)

	부 부	1세대 기타	부부+ 자녀	한 부모+ 자녀	부부+ 부모	2세대 기타	3세대 이상	1인 가구	비친족 가구
2005	364,596	107,657	1,442,057	317,607	19,639	122,488	213,458	675,739	46,649
	11.02%	3.25%	43.57%	9.60%	0.59%	3.70%	6.45%	20.42%	1.41%
2008	400,135	110,254	1,505,043	336,130	21,987	122,336	219,117	718,940	43,873
	11.51%	3.17%	43.28%	9.66%	0.63%	3.52%	6.30%	20.67%	1.26%

〈표 6〉 2005년 모집단 및 2008 추계 가구주의 성별, 연령별 서울시 가구분포

(단위: 가구)

가구주 연령	2005		2008	
0~14세	226	0.01%	219	0.01%
15~19세	10,842	0.33%	10,924	0.31%
20~24세	102,641	3.10%	81,547	2.34%
25~29세	242,698	7.33%	256,062	7.36%
30~34세	375,410	11.34%	351,629	10.11%
35~39세	396,766	11.99%	438,990	12.62%
40~44세	421,471	12.73%	413,732	11.90%
45~49세	452,314	13.67%	456,051	13.11%
50~54세	364,694	11.02%	417,558	12.01%
55~59세	304,723	9.21%	315,353	9.07%
60~64세	243,982	7.37%	259,164	7.45%
65~69세	181,899	5.50%	211,025	6.07%
70~74세	112,999	3.41%	139,131	4.00%
75~79세	60,304	1.82%	76,863	2.21%
80~84세	28,189	0.85%	34,650	1.00%
85세 이상	10,732	0.32%	14,917	0.43%
가구주 성	2005		2008	
남 자	2,529,317	76.42%	2,634,836	75.76%
여 자	780,573	23.58%	842,979	24.24%

〈표 5〉와 〈표 6〉은 2005년 인구주택총조사 결과와 통계청이 제공한 2008년 서울시 추계 가구분포를 나타내고 있다. 가구구성별 분포는 비친족 가구를 제외하고 그 차이가 크지 않음을 알 수 있다. 가구주의 연령별 두 해의 분포 차이를 살펴보면 두 해의 분포의 차이가 유의함을 알 수 있다. 2008년 추계 가구의 분포를 보면 가구주의 연령이 60세 이상인 가구의 비율이 높아졌으며 가구주의 연령이 30세 이하인 가구의 비율은 줄어든 것을 확인할 수 있다. 2008년 추계 가구주의 연령별 자료가 제한적인 이유로 시간의 경과에 따른 연령별 분포를 직접적으로 비교할 수 없으나 가구주의 연령이 20~50세인 가구의 비율에도 2005년과 2008년 사이에 변화가 있음을 〈표 6〉을 통해 알 수 있다. 즉 통계청에서 제공하고 있는 2005년 인구 및 가구 모집단 통계와 이를 바탕으로 계산된 2008년 추계 결과를 비교해 볼 때 2005년 인구주택총조사 기반 표본추출틀을 이용하여 2008 서울서베이 표본을 추출할 경우, 기본적인 인구학적 변수 기준 표본의 대표성에 문제가 발생할 수 있음을 〈표 1〉에서 〈표 6〉까지를 통해서 알 수 있다. 특별히 인구나 가구의 이동이 단시간 내에 빈번하게 발생하는 서울의 경우 3년이 경과된 표본추출틀로부터 추출된 표본의 대표성은 심각하게 훼손될 것으로 예측된다.

2005년 인구주택총조사 자료를 표본추출틀로 사용할 경우에 발생할 수 있는 표본추출틀의 대표성과 시의성 문제를 해결하기 위하여, 본 연구에서는 행정자료인 주민등록 DB와 과세대장 DB를 연계시켜서 2008년 9월 1일 기준 세대주가 서울시에 거주하는 세대들의 리스트를 표본추출틀로 작성하여 사용하였다. 주민등록 DB에는 개인 ID, 세대주·세대원 여부, 성, 연령, 거주지 주소 코드 등의 변수가 있으며, 과세대장 DB에는 주민등록 DB의 주소코드와 대응되는 주소코드와 주택유형을 포함하는 건물용도의 변수가 포함되어 있다. 주민등록 DB와 과세대장 DB의 연계는 주소코드를 기준으로 주민등록 DB에 과세대장 DB의 건물용도 변수를 추가하여 이루어졌다. 이후 건물용도 내에서 정의된 주택유형을 각 세대에 부여하였다. 자료의 특성상 주민등록 DB와 과세대장 DB는 상시 갱신되며, 따라서 이 두 DB의 연계를 통해서 얻어지는 표본추출틀은 2008년 서울시 거주자 모집단을 대표하는 것으로 간주될 수 있다. 또한 주택유형으로 일반가구와 아파트 가구만을 고려하여 약 60개의 근접가구로 정의된 조사구를 1차 추출단위로 사용하는 인구주택총조사 기반 표본추출틀과는 달리, 본 연구에서는 과세대장 DB로부터 얻을 수 있는 보다 구체적인 주택유형 정보를 바탕으로 각 통·반 내의 동일한 주택유형을 갖는 세대들을 1차 추출단위로 사용하였다.

〈표 7〉 2005년 모집단 및 2008 서울서베이 표본추출틀의 성별, 연령별 서울시 가구주 분포

(단위: 가구)

가구주 연령	2005 인구주택총조사		2008 서울서베이 표본추출틀	
0~14세	226	0.01%	1,203	0.03%
15~19세	10,842	0.33%	4,282	0.12%
20~24세	102,641	3.10%	76,242	2.10%
25~29세	242,698	7.33%	305,519	8.41%
30~34세	375,410	11.34%	388,370	10.69%
35~39세	396,766	11.99%	469,500	12.93%
40~44세	421,471	12.73%	440,077	12.12%
45~49세	452,314	13.67%	471,401	12.98%
50~54세	364,694	11.02%	426,341	11.74%
55~59세	304,723	9.21%	308,150	8.48%
60~64세	243,982	7.37%	255,695	7.04%
65~69세	181,899	5.50%	218,314	6.01%
70~74세	112,999	3.41%	138,139	3.80%
75~79세	60,304	1.82%	74,575	2.05%
80~84세	28,189	0.85%	35,815	0.99%
85세 이상	10,732	0.32%	18,164	0.50%
가구주 성	2005 인구주택총조사		2008 서울서베이 표본추출틀	
남 자	2,529,317	76.42%	2,554,825	70.35%
여 자	780,573	23.58%	1,076,962	29.65%

〈표 7〉은 2008년 주민등록 DB와 과세대장 DB를 연계하여 얻은 표본추출틀의 가구주의 연령별, 성별 분포를 나타내고 있다. 표본추출틀의 구성에 있어서 인구주택총조사 기반 표본추출틀과의 비교를 위하여 근린생활시설, 교육연구시설, 종교시설 그리고 사무용 오피스텔은 제외시켰다. 편의상 주민등록 DB와 과세대장 DB를 연계하여 작성된 표본추출틀을 2008 서울서베이 표본추출틀로 명명하였다. 두 표본추출틀의 비교를 위해서는 인구주택총조사 자료의 2008년 기준 가구주의 연령분포를 사용하여야 하나, 2008년 기준 각 광역시도별 그리고 연령별 통계량이 제공되지 않아 2005년 모집단을 직접 비교하였다. 3년의 시간 및 가구와 세대의 정의가 다름에 기인하여 발생하는 표본추출틀 규모의 변화를 고려하더라도 두 표본추출틀의 가구주 또는 세대주의 연령 및 성별 분포가 매우 다름을 알 수 있다.

가구주의 성별 분포의 경우 여성 가구주의 비율이 2008년 서울서베이 표본추출틀에서 높게 나타나고 있으며, 20세 미만의 가구주와 고령 가구주의 분포에 있어서도 두 표본추출틀이 큰 차이를 나타내고 있다. 행정자료를 이용한 2008 서울서베이 표본추출틀의 작성 시 발생하는 연계과정상의 오류와 행정자료 자체가 갖는 오류를 인구주택총조사 시 발생하는 비표본 오류와 동일한 수준으로 간주하거나 또는 이를 무시하는 경우, 〈표 7〉을 통해서 우리는 2005년 인구주택총조사 기반 가구 모집단이 가구주의 성별, 연령별 분포를 올바르게 반영하지 못하고 있다는 사실을 확인할 수 있다.

〈표 8〉은 2005 인구주택총조사 서울시 가구 수와 2008 서울서베이 표본추출틀의 세대수 분포를 각각 나타낸다. 연립주택 그리고 기타 부분을 제외한 모든 주택유형에서 2008 서울서베이 표본추출틀의 세대수가 2005 인구주택총조사의 가구 수보다 많게 나타나나 각 주택유형별 분포는 상대적으로 유사한 것으로 파악된다. 각 주택유형별 가구 또는 세대수의 차이가 크게 나타나므로, 1차 추출단위를 정의함에 있어서 주택유형정보를 이용하여 1차 추출단위 내의 모든 가구 또는 세대가 동일한 주택유형을 갖도록 하는 것이 바람직하다. 인구주택총조사 자료의 경우 가구들로 구성된 조사구를 크게 아파트와 일반 조사구로 구분하여 정의하고 있으나, 본 연구에서는 과세대장 DB로부터 얻어지는 각 세대별 구체적인 주택유형정보를 활용하여 동일 통·반 내의 동일 주택유형을 갖는 세대들의 집합을 1차 추출단위로 사용하였다. 결론적으로 2008 서울서베이를 위한 표본추출틀로는 3년 간의 인구 및 가구 변동을 고려할 수 없는 2005 인구주택총조사 자료 대신 상시 갱신되는 주민등록 DB와 과세대장 DB를 연계한 새로운 표본추출틀을 사용했다. 조사구를 1차 추출단위로 사용하는 대부분의 인구주택총조사 기반 가구조사와는 달리 이용 가능하며 보다 상세한 각 세대별 주택유형을 고려한 1차 추출단위를 정의하여 2008 서울서베이 표본을 추출하였다.

〈표 8〉 서울시 가구수와 세대수 비교

(단위: 가구, 세대)

	2005 인구주택총조사 서울시 가구수		2008 서울서베이표본추출틀 서울시 세대수	
단독, 공동 주택	1,404,272	43.36%	1,611,355	44.37%
아파트	1,218,779	37.63%	1,283,591	35.34%
다세대주택	414,983	12.81%	556,836	15.33%
연립주택, 기타	200,434	6.19%	180,005	4.96%
총 합	3,238,468	100.00%	3,631,787	100.00%

〈표 5〉에서 〈표 8〉을 통한 두 표본추출틀의 비교에 있어 두 표본추출틀 모두 유한모집단 전수에 대한 분포이므로 통계적 유의성을 직접 논의하기 어렵지만, 유한모집단이 가상의 무한모집단으로부터의 표본임을 가정하여도 각 표의 범주별 관측치의 수가 매우 크기 때문에 모든 범주별 두 표본추출틀의 차이는 통계적으로 유의하게 나타난다.

Ⅲ. 표본추출

2008 서울서베이를 위한 표본추출을 위해서는 2단계 층화집락추출이 사용되었다. 세대의 집합이자 1차 추출단위인 집락은 기본적으로 동일 통·반 내의 세대들로 정의되나 동일 통/반 내 여러 형태의 주택유형이 존재하는 경우 동일한 주택유형을 갖는 세대들의 집합으로 정의되었다. 2차 추출단위로는 집락 내의 세대가 정의되었다. 새롭게 정의된 집락을 편의상 서울서베이 집락으로 명명한다. 따라서 구성된 서울서베이 집락 내의 세대들은 층화변수로 사용된 구, 동 그리고 주택유형에 대하여 동일한 값을 갖게 된다. 2008 서울서베이 표본추출틀을 이용하여 구성된 서울서베이 집락의 수는 109,173개이며 각 집락의 평균 세대수는 약 33세대이다.

모집단의 층화를 위하여 사용된 변수로는 각 구의 행정동(460)과 주택유형(4)을 고려하였다. 주택유형의 경우 단독, 공동주택/아파트/다세대주택/연립주택, 기타의 범주를 고려하였다. 층화변수로 지역과 주택유형이 사용된 이유는 각 주택유형과 지역별로 생활패턴이 다르다는 기존 조사결과를 바탕으로 한 경험적 지식과, 또한 각 구별로 통계적으로 정도가 높은 통계작성을 위한 최소표본을 확보하기 위함이다. 서울서베이 조사 초기인 2003년부터 지역과 주택유형 변수가 층화 변수로 사용되어 왔다. 각 층으로부터 추출될 서울서베이 집락수의 결정은 적절한 수준의 정도를 갖는 각 구별 통계량의 산출에 필요한 세대수를 바탕으로 이루어졌다. 즉 서울서베이 집락의 층별 분포를 이용한 층별 집락의 수를 결정하는 배분 방법이 아닌 각 층별 세대수 분포를 이용하여 층별 표본 세대수를 정의하고 이를 바탕으로 필요한 집락의 수를 결정하였다. 이때 각 서울서베이 집락 내에서의 추출 세대수를 5세대 이하로 제한하여 충분한 수의 서울서베이 집락이 추출되도록 하였다. 이는 집락 내의 세대들이 동일한 주택유형을 가지며 지리적으로 근접함으로써 본 조사의 주 내용인 문화, 경제적 행태가 유사할 것을 예측할 수 있기 때문이다. 효율적인 집락추출 방법과 관련해서는 Cochran(1977)과 Sarndal et al.(1992)를 참조하면 된다.

각 구별 세대의 표본배분을 위해서는 구별 최소 표본 세대수인 400 세대를 만족하며 층별 표본 세대수의 변동이 단순비례배분보다 적은 제곱근 비례배분을 사용하였다. 이는 단순비례배분의 경우 모집단 층 규모에 따라 각 층별로 배분되는 표본의 크기가 매우 다르게 나타나며 따라서 적은 수의 표본이 배분된 층 또는 관심 모집단의 통계량의 신뢰도가 매우 떨어지는 것을 막기 위함이다. 즉 비례배분의 장점 및 관심 부모집단 통계의 통계적 정도를 유지하기 위하여 제곱근 비례배분이 사용되었다. 각 구에서 동별, 주택유형별 표본배분은 단순비례배분을 통하여 이루어졌다. 배분결과 얻어진 서울서베이 표본 집락 수는 4,940ro 이며 표본 세대수는 20,000 세대이다. 추출된 집락과 세대수는 전체 집락과 세대의 각 각 4.5%와 0.6%에 해당한다. 이는 집락내의 세대 간 높은 유사성을 고려하여 충분한 수의 서울서베이 집락을 추출함으로 조사결과의 정도를 높이기 위함이다. 〈표 9〉와 〈표 10〉은 구별, 주택유형별 표본 세대수의 분포를 나타낸다. 각 동별 표본배분 결과를 위해서는 서울특별시(2009)를 참조하면 된다. 각 층에서 1차 추출단위인 서울서베이 집락을 추출하기 위해서는 집락 내의 세대수를 이용한 확률비례 추출법을 이용하였고 추출된 서울서베이 집락 내의 2차 추출단위인 세대 추출을 위해서는 단순임의 추출법을 사용하였다.

2008 서울서베이 가구조사를 위하여 설명된 표본추출 과정을 통해 얻어진 표본자료를 분석하기 위해서는 통계적 이론에 근거하여 산출된 가중치를 통계처리 과정에 적용하여야 한다. 일반적으로 조사자료에 부여되는 가중치는 표본추출 과정에서 부여되는 표본가중치, 무응답에 대한 조정 그리고 사후층화 또는 레이킹에 의한 조정 등의 세 가지 요인을 통합하여 산출된다. 설명된 표본추출 과정을 통하여 얻어진 h번째 층의 i번째 집락 내의 j번째 세대에 부여되는 표본가중치는 표본추출확률의 역수로 아래와 같이 정의된다.

$$\pi_{hij}^{-1} = \left[\pi_{hi}\pi_{j|hi}\right]^{-1} \tag{1}$$

여기서 h번째 층의 i번째 집락의 추출확률은 $\pi_{hi} \approx n_h\left[M_h^{-1}M_{hi}\right]$ 로서 M_{hi}는 h번째 층의 i번째 집락 내의 총 세대수를 나타내며, n_h는 h번째 층에서 추출된 표본 서울서베이 집락의 수이며 $M_h = \sum_i M_{hi}$ 이다. h번째 층의 i번째 집락이 1 단계에서 추출되었다는 조건하에서의 집락 내의 세대가 표본에 추출될 조건부 확률은 $\pi_{j|hi} = M_{hi}^{-1}m_{hi}$로 m_h는 집락에 할당된 표본 세대수이다.

〈표 9〉 각 구별 표본 세대수 배분

(단위: 세대)

구	서울서베이 표본추출틀		표 본	
종로구	61,926	1.71%	530	2.65%
중 구	42,059	1.16%	437	2.19%
용산구	88,958	2.45%	634	3.17%
성동구	115,173	3.17%	722	3.61%
광진구	147,604	4.06%	817	4.09%
동대문구	140,814	3.88%	798	3.99%
중랑구	163,145	4.49%	859	4.30%
성북구	164,021	4.52%	861	4.31%
강북구	129,641	3.57%	766	3.83%
도봉구	128,082	3.53%	761	3.81%
노원구	203,486	5.60%	959	4.80%
은평구	166,883	4.60%	869	4.35%
서대문구	125,286	3.45%	753	3.77%
마포구	143,167	3.94%	805	4.03%
양천구	162,504	4.47%	857	4.29%
강서구	192,391	5.30%	933	4.67%
구로구	147,194	4.05%	816	4.08%
금천구	87,495	2.41%	629	3.15%
영등포구	145,601	4.01%	812	4.06%
동작구	148,867	4.10%	821	4.11%
관악구	214,200	5.90%	984	4.92%
서초구	142,306	3.92%	802	4.01%
강남구	202,798	5.58%	958	4.79%
송파구	218,963	6.03%	995	4.98%
강동구	149,223	4.11%	822	4.11%
합 계	3,631,787	100.00%	20,000	100.00%

〈표 10〉 주택유형별 표본 세대수 배분

(단위: 세대)

	단독주택 (공동주택)	아파트	다세대주택	연립주택 기타
모집단	1,611,355 (44.4%)	1,283,591 (35.3%)	556,836 (15.3%)	180,005 (5.0%)
표 본	9,049 (45.2%)	6,900 (34.5%)	3,043 (15.2%)	1,008 (5.0%)

추출된 가구를 대상으로 조사를 진행하게 될 때 응답거절이나 이사 등으로 인한 단위 무응답의 발생은 필연적이다. 조사과정에서 발생하는 이러한 단위무응답을 보정하기 위하여 식 (1)에서 정의된 표본가중치를 조정하게 된다. 무응답 처리를 위한 가중치 보정 방법으로는 무응답 패턴이 유사한 셀을 구성하여 각 셀 내 응답률을 바탕으로 무응답으로 인한 가중치 보정을 실시하는 셀 무응답 보정 또는 무응답 여부와 이용 가능한 보조 변수들 간의 관계에 대한 모형을 설정하고 응답 확률을 예측하여 이를 바탕으로 가중치를 보정하는 성향점수(propensity score) 방법 등을 고려할 수 있다. 표본추출틀의 구축 시기와 조사 시점 간의 시간적인 차이가 있거나 표본설계 시 반영할 수 없었던 모집단 분포와 표본 분포를 일치시키기 위하여 무응답 보정이 이루어진 후에 사후층화 혹은 레이킹이 흔히 이루어진다. 2008 서울서베이 가구조사의 경우, 표본추출틀의 구성 시기와 조사 시점이 거의 일치하며 표본설계 시 중요한 보조 변수들이 모두 고려되었으므로 무응답 보정 후 추가적인 가중치 보정은 요구되지 않을 것으로 예측된다. 무응답 처리와 사후층화 그리고 레이킹에 대하여서는 Deville et al.(1993), Fuller(2002) 및 Kott(2006)를 참조하면 된다.

표본 가중치에 무응답 보정과 사후층화 또는 레이킹 과정을 통해 정의된 가중치를 w_{hij}라 할 때 이를 이용하여 정의되는 관심변수 y의 모집단 총합에 대한 불편 추정량으로 Horvitz−Thompson(1952) 추정량

$$\widehat{t_{HT}} = \sum_{hij \in S} w_{hij} y_{hij} \tag{2}$$

을 사용할 수 있다. 여기서 S는 추출된 세대 표본을 나타낸다. 모집단의 평균에 대한 추정량으로는 다음의 두 추정량으로 Horvitz−Thompson 추정량

$$\overline{y_{HT}} = \frac{\sum_{hij \in S} w_{hij} y_{hij}}{M} \quad (3)$$

과 비추정량

$$\overline{y_{\pi}} = \frac{\sum_{hij \in S} w_{hij} y_{hij}}{\sum_{hij \in S} w_{hij}} = \frac{\sum_{hij \in S} w_{hij} y_{hij}}{w_{...}} \quad (4)$$

을 고려할 수 있다. 여기서 M은 모집단 전체 세대수를 의미한다. 추정량 (3)은 모집단 평균에 대한 불편 추정량이며 (4)는 근사 불편 추정량이다. 본 조사의 경우 표본 세대의 수가 충분히 크기 때문에 추정량 (4)의 편의는 무시할 수 있을 것으로 고려된다. 각 집락별 세대의 크기가 다른 경우 비추정량 형식의 추정량 (4)가 더 효율적임이 알려져 있고, 따라서 모집단 평균의 추정량으로는 (4)를 사용한다. 비추정량의 효율성에 관해서는 Cochrane (1977)과 Sarndal 외(1992)를 참조하면 된다. 한편 제시한 추정량 (4)에 대한 분산추정량으로서는

$$var(\overline{y_{\pi}}) = \sum_{h=1} \frac{n_h (1 - f_h)}{n_h - 1} \sum_{i=1}^{n_h} (e_{hi.} - \overline{e_{h..}})^2$$

를 사용할 수 있다. 여기서 $e_{hi.} = \left(\sum_{j=1} w_{hij} (y_{hij} - \overline{y_{\pi}}) \right) / w_{...}$, $\overline{e_{h..}} = \left(\sum_{i=1} e_{hi.} \right) / n_h$ 이다. 추정량 (4)에 대한 상대표준오차로는

$$\widehat{CV}(\overline{y_{\pi}}) = \frac{\sqrt{var(\overline{y_{\pi}})}}{\overline{y_{\pi}}} \times 100$$

를 사용할 수 있다. 분산추정량과 상대표준오차에 대한 보다 자세한 내용은 Lohr(1999)와 Scheaffer(2006)를 참조하면 된다.

Ⅳ. 토의 및 결론

2005년 서울서베이 조사의 한 부분인 가구조사를 위한 표본추출틀의 작성과 이로써 표본추출방법을 본 연구에서는 소개하였다. 조사가 매해 수행되는 즉 조사주기가 1년인 서울서베이의 경우, 유동성이 심한 서울시의 인구와 가구의 변동을 감안한 표본추출틀의 구축이 매우 중요하다. 5년 주기로 갱신되는 인구주택총조사 90%자료를 이용한 표본추출틀의 경우 인구주택총조사 시점으로부터 실제 표본조사 시점까지의 시차가 커짐에 따라 표본추출틀의 모집단 대표성이 심각하게 훼손될 수 있음을 살펴보았다. 본 연구에서는 3년이 지난 2005년 인구주택총조사 기반 표본추출틀의 시의성 문제를 해결하기 위해 주민등록 DB와 과세대장 DB를 연계한 새로운 표본추출틀의 구성을 제안하였고, 이로부터 각 세대별 주택유형 정보를 활용한 서울서베이 집락을 구성하고 충분한 수의 집락을 추출하는 표본추출법을 제안하였다. 제안된 표본추출틀과 표본추출법을 통해 얻어진 표본은, 새롭게 구성된 표본추출틀의 적절한 모집단 대표성과 충분한 표본 집락수, 그리고 집락 내의 세대수를 크기 변수로 활용한 확률비례 표본추출로써 효율적인 추정량을 제공하리라 기대된다.

참고문헌

서울특별시 2009.《2008년 서울서베이 보고서》.

Cochran, W. G. 1977. *Sampling Technique*. New York: Wiley.

Deville, J. C., Samdal, C. E. and Sautory, O. 1993. "Generalized Raking Procedures in Survey Sampling." *Journal of the American Statistical Association* 88: 013−1020.

Suller, W. A. 2002. "Regression Estimation for Survey Samples." *Survey Methodology* 28: 5−23.

Horvitz, D. G. and Thompson, D. J. 1952. "A Generalization of Sampling Without Replacement from a Finite Universe." *Journal of the American Statistical Association* 47: 663−685.

Kott, P. S. 2006. "Using Calibration Weighting to Adjust for Nonresponse and Coverage Errors." *Survey Methodology* 32: 133−142.

Lohr, L. L. 1999. *Sampling: Design and Analysis*. Duxbury Press.

Samdal, C. E., Swensson, B. and Wretman, J. 1992. *Model Assisted Survey Sampling*. Springer.

Scheaffer, R. L., Mendenhall, W. and Ott, R. L. 2006. *Elementary Survey Sampling*. Thomson.

[접수 2009/8/20, 1차수정 2009/10/9, 2차수정 2009/10/25,
게재확정 2009/10/29]