

연구노트

행정자료와 의존적 가구 명부를 이용한 가구조사 표집틀: 2023 한국종합사회조사*

김지범** · 유광중*** · 김규영**** · 김월화***** · 양동평***** · 김솔이*****

다단계 지역 확률 표집 일반 가구 조사에서 표집틀로 이용하는 지역의 가구 목록은 자료의 질을 결정하는 하나의 중요한 요인이다. 통계청의 승인 통계가 아닌 경우 집계구에서 가구를 선정하기 위해 가구 목록이 필요하지만 가구 목록 작성과 관련하여 알려진 것이 없다. 본 연구는 2023년 한국종합사회조사에서 적용한 행정자료와 의존적 가구 명부를 이용한 가구조사 표집틀 과정을 상세하게 기술하는 데 그 목적이 있다. 동시에, 행정자료 기반 가구 목록, 표집원 검토 가구 목록, 그리고 면접원 실사 가구 목록 등을 비교하여 행정자료의 완전성을 평가한다. 2023년 선정된 200개의 집계구는 117개의 아파트 집계구와 83개의 비아파트 집계구로 구성되었다. 아파트 지역은 행정자료에만 의존하였으며, 비아파트 지역은 조사 주관기관인 성균관대학교 서베이리서치센터에서 파악한 행정자료 가구 목록을 바탕으로 표집원이 지역을 방문하여 가구의 누락과 중복 등을 검토하여 수정한다. 이렇게 갱신된 가구 목록에서 선정된 12가구에 대한 가구 명부에 따라 면접원은 각 가구를 방문하여 실제로 조사를 수행한다. 결과를 살펴보면, 아파트 가구에서의 행정자료와 면접원 실사자료를 비교했을 때, 행정자료에 기반한 면접대상 가구의 99%는 실제로 사람이 거주하였다. 반면, 비아파트 가구에서는 행정자료와 표집원 자료 그리고 면접원 자료의 세 개 자료 간 비교가 가능했는데, 표집원 자료와 비교하여 행정자료가 놓치고 있는 실거주 가구 주소의 비율은 약 21%나 되었다(총 8,031가구 중 1,672가구). 표집원 자료와 면접원 자료 비교에서는 표집원 자료에서 약 4.8%(총 996가구 중 48가구)가 오류가 있었으며, 행정자료와 면접원 자료 비교에서는 행정자료가 놓치고 있는 실거주 가구 주소의 비율이 약 25% (총 996가구 중 245가구)였다. 결론적으로, 비아파트 지역 가구 조사의 정확한 표집틀을 위해서는 표집원의 의존적 가구 목록 작성이 필요하다.

주제어: 표집틀, 집계구, 의존적 가구 명부 작성, 가구면접조사, 행정자료, 한국종합사회조사(KGSS)

* 본 논문은 2020년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임.
(NRF-2020S1A3A2A03096777).

본 논문을 향상시키는 데 큰 도움을 주신 세 분의 심사위원, 성균관대 동아시아학술원 서베이리서치센터의 김우경, 민희수, 이지원과 표집원 학생들, 한국갤럽 박성준 연구원, 그리고 2023년 한국종합사회조사 면접원께 감사드린다.

** 성균관대학교 사회학과, 소셜이노베이션 융합전공 교수(jbk7000@skku.edu), 제1저자, 교신저자.

*** 성균관대학교 사회학과, 소셜이노베이션 융합전공 석사과정(nagara9712@gmail.com).

**** 성균관대학교 사회학과 학사과정(kimkyuyoungsrc@gmail.com).

***** 성균관대학교 사회학과 박사과정 수료·동아시아학술원 서베이리서치센터 연구원(kimworlwha@gmail.com).

***** 성균관대학교 사회학과 박사과정 수료(yangdongpingy@gmail.com).

***** 성균관대학교 서베이리서치센터 전임연구원(sorikim0313@gmail.com).

I. 서론

표집틀(sampling frame)은 표본이 추출되는 목록이며, 표집틀 오류는 (서베이)조사의 전반적 과정에서 나타날 수 있는 오류 중 하나이다(Weisberg 2005; Groves et al. 2009). 표집틀이 목표 모집단을 제대로 반영하지 못해서 나타나는 포함오류(coverage error)는 조사자료의 품질을 저하한다. 한국에서 일반 가구 성인 조사는 다단계 지역 확률 표집(multi-stage area probability sampling)에 따라 선정된 가구에서 가구원을 추출한다. 국가 승인 통계 가구 조사는 통계청에서 구축한 전수조사 기반의 DB 자료인 인구주택총조사 조사구 주소 자료를 사용할 수 있는 자격이 부여된다(박인호 2014). 반면, 비승인 통계 조사는 통계청으로부터 조사구 주소 자료를 제공받지 못하기 때문에 집계구의 가구 주소 명부를 구축해야 한다. 기관의 의뢰를 받아 실사를 담당하는 조사회사는 면접원에게 표본으로 선정된 가구 주소 명부를 제공하지 못하는 실정이다. 대신에 집계구의 중심점을 면접원에게 알려주면서 중심점으로부터 계통추출로 가구를 선정하게 하는 방법을 택하고 있다. 따라서, 면접원이 자의적으로 가구를 선정할 수 있는 개연성이 높아 조사자료의 품질을 담보하기 어렵다.

표집틀은 크게 세 가지 방식으로 작성된다. 첫째는 행정자료이다. 하나 혹은 여러 가지 행정자료 결합을 통해 만들 수 있다. 한국에서는 조사구 주소 자료, 미국에서는 우정공사의 우편배달 주소 파일(United States Postal Delivery Sequence File)이 대표적이다. 두 번째는 오랜 기간 활용된 표집틀 작성 방법의 표준인 전통적 가구 명부 작성(traditional listing)이다(O'Muircheartaigh et al. 2006). 표집원(lister)은 해당 표집 지역에 직접 방문하여 가구 명부를 작성하는 것이다. 셋째는 의존적(갱신된, 향상된) 가구 명부 작성(dependent, update or enhanced listing) 방식이다(Eckman & Kreuter 2011). 행정자료 가구 주소 목록의 오류를 보완하려는 시도로써 표집원은 주어진 행정자료의 가구 목록을 가지고 표집 지역을 방문한 후 가구 목록을 수정 및 보완하는 과정을 거쳐 가구 명부를 향상시킨다.

미국에서는 비용과 시간이 많이 소요되는 전통적 가구 명부 작성 방법의 대안으로 우정공사의 우편배달 주소 파일이 이용되고 있는데, 이 파일의 정확성을 확인하는 연구가 지속적으로 이루어지고 있다(Iannacchione 2011; Dohrmann et al. 2014;

Jones, Brick, & Van De Kerckhove 2023). 우편배달 주소 파일은 전통적 가구 명부를 대체하거나 혹은 우편배달 주소 파일의 결함을 표집원이 보정(추가 혹은 삭제)하는 의존적 가구 명부 작성에 이용된다(Kalton, Kali, & Sigman 2014). Iannacchione, Staab, & Redden(2003)은 반개구간(half-open interval) 방식을 활용하여 주소 자료의 미포함(undercoverage) 정도를 확인하여 우편배달 주소 파일의 완전성을 평가하였으며, 주소 목록과 반개구간 방법의 연계를 통해 주소 목록의 완전성을 증가시킬 수 있다는 점을 강조했다. 하지만 의존적 가구 명부 작성에서 표집원들이 우편배달 주소 파일을 그대로 믿는 확증편향(confirmation bias)이 나타난다는 점도 밝혔다(Eckman & Kreuter 2011). 또한 전통적 가구 명부 작성을 우편배달 주소 파일로 완전히 대체 혹은 보완하기보다는 도시 지역에서는 우편배달 주소 파일을 활용하고 농촌에서는 전통적 가구 명부 작성을 제안하기도 하였다(O'Muircheartaigh et al. 2003; O'Muircheartaigh et al. 2006). 실제로, 미국의 종합사회조사(General Social Survey, GSS)는 2004년부터 지역의 우편배달 주소 목록의 완벽성에 따라 그대로 이용하거나, 우편배달 주소 파일을 표집원이 갱신하는 의존적 가구 명부 작성을 취하거나, 아니면 전통적 가구 명부 작성을 통하여 표집틀을 구축하였다(Harter et al. 2010).

반면 한국에서는 가구조사 표집틀 연구가 간과되었다고 해도 무리가 아니다. 극히 소수의 선행연구가 있다. 주로 인구주택총조사에 기초한 표집틀이 특정 조사의 조사 시점과의 차이로 인하여 생기는 표집틀 오류를 행정자료 결합으로 해결하려고 했다. 강현철 외(2009)는 서울 서베이 조사에 주민등록DB와 과세대장DB를 통합해서 만든 표집틀을 제시하였고, 박현아 외(2010)는 인구주택총조사와 KB 국민은행 아파트 시세조사 자료를 혼용한 표집틀을 제안하였으며 박진우·변종석·박민규(2010)는 주민등록 통/반/리 자료와 KB 국민은행 아파트 시세조사 자료를 혼용한 표집틀을 제안했다. 집계구를 활용한 표집틀 연구도 있었지만, 11장의 파워포인트로만 남아있다(GDS CONSULTING 2016). 다른 한편, 전통적 가구 명부 작성으로는 2006년부터 2018년까지 실시된 한국종합사회조사의 표집틀 연구(구묘정 외 2014)가 대표적 사례이다. 의존적 가구 명부 작성 관련 연구는 찾을 수 없었다.

한국에서 행정자료 기반 표집틀 작성 선행연구는 표집틀의 정확성을 높여 자료의 질을 향상시키려는 의미있는 연구였지만 몇 가지 한계점이 있다. 비승인 통계조사를 수행하는 연구자들이 기존 연구에서 사용한 몇몇 행정자료는 쉽게 접할 수 없으며, 행정자료의 정확성도 파악할 수 없었다. 더불어, 최근에 많이 활용되는 통

계청 집계구를 표집지역으로 한 표집틀 작성 과정이 거의 알려지지 않았다. 즉, 표집틀 작성 및 표집틀 오류에 대한 연구 공백이 크다.

본 연구는 2023년 한국종합사회조사에서 적용한 행정자료와 의존적 가구 명부를 이용한 가구조사 표집틀 과정을 상세하게 기술하려는 목적이 있다. 이와 함께, 행정자료 기반 가구 목록, 표집원이 행정자료를 가지고 가구에 직접 방문하여 작성한 가구 명부, 그리고 면접원 실사 자료 가구 명부를 비교하고자 했다. 상세한 가구 명부 작성 과정과 자료 간 비교는 향후 정확하고 효과적인 표집틀 구축 방법을 만들기 위한 돌파구이다.

이러한 목표로 본 연구는 다음과 같이 구성된다. 2장에서는 의존적 가구 명부 작성에 이용한 자료를 설명하고 분석방법을 소개한다. 3장에서는 작성된 가구 명부 및 검증 관련 과정을 상세히 기술한다. 4장에서는 먼저 비아파트 지역 가구의 행정자료와 표집원 자료를 대조하여 행정자료의 정확성을 파악한다. 다음으로, 아파트 지역 가구의 행정자료와 면접자료를 검토하고, 비아파트지역 가구의 표집원 자료와 면접원 자료를 비교한 후 행정자료와 면접자료를 비교한다. 마지막으로, 분석 결과를 정리하고 그 결과의 함의를 논한다.

II. 분석자료 및 분석방법

1. 분석자료

한국종합사회조사: 한국종합사회조사는 2003년에 태동되어 한국 사회의 변화와 안정성을 파악할 수 있는 정치, 경제, 사회, 문화 등 다양한 질문과 국제비교를 위한 문항으로 구성된다(김지범 외 2022). 일반 가구에 살며 한국어를 할 수 있는 대한민국 18세 이상 성인을 대상으로 한 면접조사이다. 2023년 한국종합사회조사의 최종 최소 표본 크기는 1,200명으로 50%의 응답률을 목표로 했다. 다단계 지역 확률 표집을 활용하였는데, 시·도를 층화변수로 이용하여, 네 단계의 과정을 거쳐 조사 대상자를 선정했다. 시·도의 가구 수를 크기비례확률표집 방법을 이용하여 1단계는 200개의 1차 표집단위인 동/읍/면을 선정했다. 1차 표집단위 200개는 표본의 다양성, 효율적인 통계 추정, 조사 예산, 운영 비용등을 종합적으로 고려하여 결정

했다. 2단계는 2차 표집단위인 통/반/리를 선정했으나 2021년부터는 집계구를 추출하였다. 집계구는 평균 200가구를 포함하는데 통계청의 SGIS(통계지리정보서비스)에 기반한 표본 조사구 2~3(한 조사구에 60가구 정도)개를 묶은 것이다(박인호 2016). 3단계는 선정된 200개 집계구를 표집원이 방문하여 가구 주소목록을 작성했다. 그리고, 이 주소목록에서 집계구의 각 가구가 동일한 추출 확률을 갖도록 집계구당 12개 가구를 선정했다. 마지막 단계는 추출된 각 가구에서 적절한 조사 대상자를 생일법 혹은 키시테이블(Kish table)(2018만 해당)을 이용하여 선정하였다. 2023년 한국종합사회조사 응답률은 51%이며, 응답자는 1,230명이었다.

2023년 선정된 200개 집계구 중 117개는 아파트 지역이고 83개는 비아파트 지역(아파트와 비아파트 혼재 지역 6개 포함)이었다. 아파트와 비아파트가 공존하는 지역에서는 아파트와 비아파트를 모두 포함하는 가구 명부를 작성하였다. 아파트 지역 가구 명부는 서베이리서치센터 구성원 5인이 작업을 담당했으며, 비아파트 지역 83곳은 서베이리서치센터 연구원 및 학생으로 이루어진 총 24명의 인원이 표집원으로 참여하였다. 행정자료 기반 가구 명부 작성은 3월 6일부터 시작되어 4월 26일까지 진행되었다. 표집원 실사 작업은 4월 27일부터 시작되어 5월 19일에 마무리되었고, 면접원의 실제 면접 조사는 5월 30일부터 9월 26일까지 진행되었다.

통계지리정보시스템(SGIS): 통계청에서 제공하는 통계지역 서비스 시스템으로, 본 연구에서는 해당 서비스에서 제공하는 지역정보들을 바탕으로 QGIS 프로그램을 활용하여 200개 집계구를 살펴보았다. 지도 화면에서 모든 집계구에 대해 중심점과 집계구 경계를 제시하고 있다.

행정안전부 도로명주소 시스템(<https://www.juso.go.kr>): 행정안전부에서 도로명주소 사용의 독려를 위해 제공하는 주소 검색 시스템으로, 현재 약 637만 개의 도로명주소와 약 39만 개의 상세주소가 기재되어 있다. 도로명을 검색하면 해당하는 주소의 목록을 열람할 수 있고, 상세주소가 기재된 건물의 경우에는 층수와 호수를 열람할 수 있다.

네이버 지도-거리뷰 시스템: 포털 사이트 네이버에서 제공하는 스트리트뷰 시스템으로, 실제 현장에 방문한 것과 같은 360도 이미지를 전국 상당수 지역에서 확인할 수 있다. 주소를 입력하면 네이버 측에서 촬영한 360도 사진 열람이 가능하며, 촬영

시기별 사진도 다양하게 열람할 수 있어서 건물의 신축 여부는 물론 실제 거주 흔적이 있는지도 어느 정도 확인할 수 있다.

네이버 부동산 시스템: 포털 사이트 네이버에서 제공하는 부동산 정보 시스템으로 아파트 지역의 동·호수 정보를 파악할 수 있다. 주상복합 아파트의 경우에는 상가/오피스텔로 사용되는 동·호수는 별도로 표시되어 있고, 동별 층수와 층별 호수가 다른 경우도 보여주고 있다.

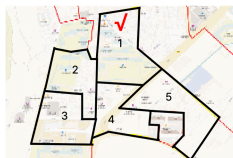
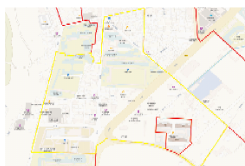
본 연구에서는 위에서 언급한 자료들을 행정자료로 통칭한다. 그리고 표집원에 의해서 작성된 주소 자료를 표집원 자료로 명명하고, 전문 면접원들이 면접을 위해 실제로 가구를 방문하여 확인한 주소 자료를 면접원 자료로 지칭한다.

2. 분석방법

우선 KGSS의 의존적 가구 명부 작성 방법을 상세히 기술한다. 그리고, 본 연구에서 이용한 자료를 비교한다. 첫째, 비아파트 지역 가구의 행정자료와 표집원 자료로 행정자료가 실제 거주 현황을 얼마나 정확히 반영하고 있는지를 살펴보았다. 둘째, 아파트 지역 가구의 경우에는 표집원의 현장 방문이 이뤄지지 않았기 때문에, 행정자료와 면접자료와의 비교를 진행했다. 셋째, 비아파트 지역 가구에 대해, 표집원 자료와 면접원 자료와의 비교를 진행했다. 마지막으로, 비아파트 지역 가구에서 행정자료와 면접원 자료를 비교했다.

Ⅲ. KGSS의 의존적 가구 명부 작성 및 검증 과정

한국종합사회조사의 의존적 가구 명부 작성 및 검증은 크게 총 세 단계로 구성되어 있다. 1단계는 행정자료 등을 활용한 가구 명부 작성이고, 2단계는 표집원의 표집 지역 현장 방문을 통한 가구 명부 확인 및 수정이며, 3단계는 조사 면접원이 2단계에서 작성된 가구 명부를 가지고 실제로 가구를 방문한 이후의 확인 작업이다.

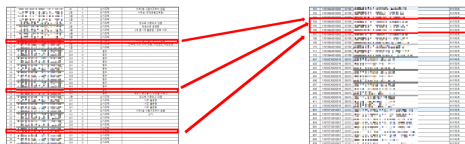
[illegible]

1단계
집계구 가구명부 작성



2단계

표집원의 현장 방문



(표집된 가구명부를 바탕으로 각 집계구별 12가구 추출)

3단계

최종 조사 대상 가구 목록
작성 및 실제 조사 진행

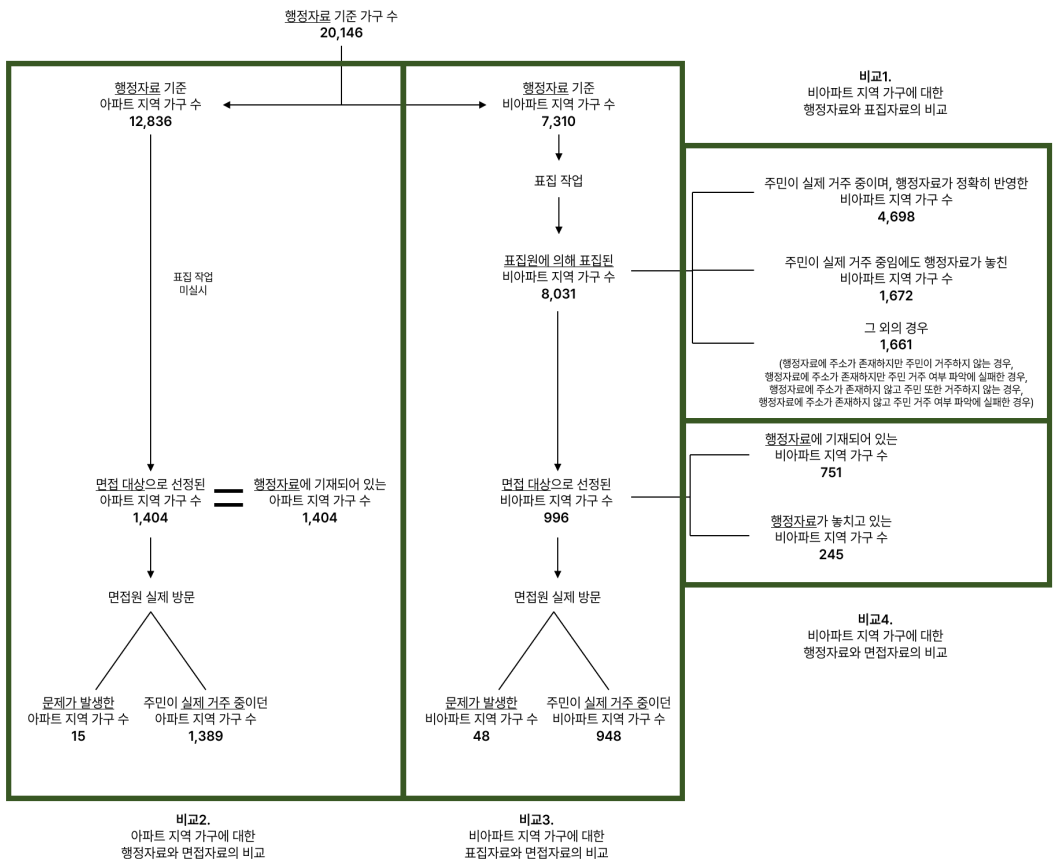
1단계 작업(그림 1-1단계, (1)~(6))은 성균관대학교 서베이리서치센터 연구원들이 표집지역으로 선정된 200개 집계구 번호를 통계지리정보시스템(QGIS)에 입력하여 집계 지역의 지도를 확인함으로써 시작된다(그림 1-1단계, (1)). 한 집계구당 평균 가구 수는 120가구 정도로 분할되어 있는데, 이 과정에서 비용과 시간을 고려하여, 연구원들은 각 지역 모든 가구의 가구 명부를 작성하는 대신 50개 가구가 추출될 수 있도록 우선 지역을 여러 구역으로 분할한다(그림 1-1단계, (2)). 분할의 기준은 대형 도로와 자연기물 등의 물리적 기준으로, 분할 작업이 종료되면 분할된 구역들에 번호를 할당한 뒤 무작위 난수 추출을 통해 가구 명부를 작성할 구역을 선정한다. 이때 해당 구역들은 아파트 지역과 비아파트 지역(아파트, 비아파트 혼합 지역도 비아파트 지역으로 분류)으로 분류되며, 지역 특성과 관계없이 선정된 구역에 대해 연구원들은 주소 목록을 작성한다(그림1-1단계, (3)). 주소 목록은 행정안전부가 제공하는 도로명 주소 사이트와 네이버 포털 사이트의 지도 로드뷰 시스템을 활용하여 작성된다. 연구원은 해당 수단들을 통해 선정된 구역 속 건물들의 주소 목록을 작성하고, 건물의 외관을 로드뷰 시스템을 통해 육안으로 확인함으로써 표집원에게 전달될 가구 목록에 특이사항 및 확인점을 기재한다(그림1-1단계, (4)~(6)).

2단계 작업(그림1-2단계, (1)~(2))은 표집원 교육과 표집원의 표집 지역 방문 가구 명부 작성이다. 먼저 서베이리서치센터 연구원은 표집원에게 표집틀 작성 교육을 실시한다(그림1-2단계, (1)). 교육은 실제 현장에서의 실사 요령과 주의사항 등을 다루며, 특히 행정자료 기반의 가구 명부에서 누락된 상세주소 등을 기재해야 한다는 사실과 실제 거주 여부 파악 활동이 중요함을 강조한다. 표집원은 서베이리서치센터에서 제공한 가구 명부를 가지고 비아파트 지역에 방문하여 누락된 가구를 추가하거나 잘못 기재된 가구를 수정하는 등의 임무를 맡는다(그림1-2단계, (2)). 예를 들어 표집 구역에 존재하는 3층 다세대 주택에 대해 표집틀에 오직 한 가구의 주소만 기재되어 있다면, 표집원은 층수와 호수 등을 확인하여 해당 건물에 대한 상세주소를 반영하여 가구 목록을 추가한다. 이때 표집원은 택배 존재 여부, 우편물의 존재 여부, 쓰레기 봉투의 존재 여부, 계량기 수치 범위, 야간 점등 여부 등의 다양한 방법을 통해 비대면으로 가구의 거주 여부를 확인한다. 만약 표집원이 배정받은 구역에 실제로 거주하는 가구가 50가구에 미달하면, 해당 표집원은 선정된 구역의 시계 방향 구역으로 이동하여 추가 조사를 한다.

3단계 작업(그림1-3단계)은 센터 연구원이 표집원들이 보내온 수정된 가구 명부를 바탕으로 실거주 가구 기반의 가구 명부를 재작성한다. 이후, 집계구별 12가구씩

을 무작위 추출하여 총 2,400가구의 주소를 조사회사에 송부한다. 조사회사는 선정된 가구에 사전 안내 우편과 공문을 발송한다. 면접원들은 이 가구 주소를 가지고 해당 주소에 직접 방문하는데 실제 거주하지 않는 가구의 주소이거나 주소 목록 자체에 오류가 있는 경우, 면접원은 면접 팀장에게 보고하고 면접 팀장은 서베이리서치센터에 그 사실을 통보한다. 이후 센터 연구원들은 확인 작업을 거친 후 해당 사실을 기록하고, 필요한 경우 사전에 준비된 예비 주소 목록의 동일 번호에 해당하는 주소를 해당 면접원에게 제공한다.

IV. 연구결과



<그림 2> 행정자료, 표집원 자료, 면접원 자료 가구 명부 비교

<그림 2>는 행정자료, 표집원 자료, 그리고 면접원 자료와의 비교를 전반적으로 보여준다. 먼저 비아파트 지역 가구에 대한 행정자료와 표집원들이 작성한 표집자료에 대한 비교를 진행했다. 200개 집계구 중 비아파트 지역 집계구는 총 83개로, 이들 집계구에서 표집된 행정자료 기반 가구 주소 개수는 총 7,310곳이었다. 하지만 현장에 방문해서 실제로 주소를 표집해 온 표집원 작성 가구 명부의 가구 주소 개수는 총 8,031곳에 달했다. 즉, 행정자료에서 주소가 상당히 누락되었다.

<표 1> 비아파트 지역 행정자료와 표집원 자료 비교

행정자료 반영 여부	표집원 자료: 주민 거주 여부	행정 자료	표집원 추가자료	전체	(단위: %)
행정자료 반영	(1) 주민 실제 거주			4,698	58
	(2) 주민 거주하지 않음			892	11
	(3) 주민 거주 파악 실패			222	3
	(1)+(2)+(3)			(5,812)	(72)
행정자료 미반영	(4) 주민 실제 거주			1,672	21
	(5) 주민 거주하지 않음			440	5
	(6) 주민 거주 파악 실패			107	1
합계	(1)+(2)+(3)+(4)+(5)+(6)	7,310	721	8,031	100

<그림 2>의 비교1과 비교3 과정에서 드러난 결과를 <표 1>로 정리했다. 이렇게 표집된 8,031곳의 가구 중 실제 주민이 거주하고 있으며 행정자료가 해당 주소를 정확하게 반영하고 있는 가구는 총 4,698가구였으며, 반대로 주민이 실제 거주하고 있음에도 행정자료에서 그 주소를 놓치고 있는 가구는 총 1,672가구에 달했다. 이외에 892가구는 행정자료에 존재했으나 주민이 거주하지 않는 경우, 222가구는 행정자료에 존재하지만 거주 여부 파악에 실패한 경우, 440가구는 그 주소가 행정자료에 존재하지 않고 주민이 거주하지 않는 경우, 107가구는 그 주소가 행정자료에 존재하지 않고 거주 여부 파악에 실패한 경우에 해당했다.

결과적으로 약 21% 정도의 실거주 가구 주소를 행정자료가 놓치고 있음이 드러났다. 누락 사례들은 다세대 주거형태에서의 층수나 호수 등의 상세주소가 누락된

경우가 98%로 절대적인 비중을 차지하였다. 이외에도 지하층과 같은 특정 층만 누락되어있거나 상세주소가 한 층씩 밀려서 기록된 사례도 존재했다.

다음으로는 아파트 지역 가구에 대한 행정자료와 면접자료의 비교를 진행했다. 먼저 행정자료의 가구 목록을 면접원 자료와 비교해 본 결과, 비교적 행정자료가 실제 거주 상황을 잘 반영하고 있음이 드러났다. 117개 집계구에서 표집된 12,836개의 아파트 가구들은 표집원 확인 과정 없이 1,404곳이 면접 대상가구로 선정되었는데, 면접원이 방문해본 결과 15곳을 제외한 1,389곳에 실제 주민이 거주하고 있음이 드러났다.

다만 상기된 1,389가구 중에는 공동현관 출입 불가 등의 이유로 면접원 방문이 어려웠던 사례나 방문 시 무응답이 반복된 사례도 포함되어 있으므로 해당 가구들에 대해 완전무결한 실거주 여부 확인이 이뤄졌다고는 볼 수 없다. 그러나 한국종합사회조사의 조사과정에서 면접원들은 여러 번 방문하는 원칙을 고수하고, 부재 또는 미거주 추정 상황을 비교적 분명히 구분하여 서베이리서치 센터에 보고해야 하기 때문에, 상당한 정도의 정확성이 있다고 볼 수 있다.

면접원 방문 당시 문제가 발생한 15곳은 면접원에게 제공한 가구 명부상의 주소가 실제로 존재하지 않는 ‘주소 오류’가 6건, 주상복합 아파트 또는 아파트 내 어린이집과 같은 비주거용 가구 5건, 그리고 빈 집으로 추정되는 4건이다.

요약하자면, 행정자료를 기반으로 하여 아파트 지역 가구를 방문했을 때 행정자료의 오류 비율은 약 1%이다. 즉, 아파트 지역 가구는 비아파트 지역 가구와 비교해서 행정자료가 실제 거주 상황을 잘 반영했다. 부적격 아파트 가구(예, 어린이집 또는 주상복합의 상가)에 대해서는 다른 가구로 교체했다.

다음으로는 비아파트 지역 가구에 대해 표집원 자료와 면접자료의 비교를 진행했다. 8,031가구의 표집된 가구 중 996가구(아파트와 비아파트 혼재 가구 27가구 포함)가 면접 대상으로 선정되어 면접원이 방문한 결과, 이 중 948가구에 주민이 실제 거주하고 있었으며 48가구는 부적격 가구였다. 48가구 중 17개 가구는 비거주 목적으로 이용되었고, 16개 가구는 주기적으로 비어있어 빈집으로 추정되는 경우였고, 11개 가구는 주소 오류의 사례이다. 특기할 만한 점은, 문제 발생 사례의 상당수가 표집원의 비대면 표집 방식에서 비롯되었다는 것이다. 실제로 ‘비거주 목적으로 사용’ 사례는 건물이 홈오피스나 사찰로 사용되고 있는 경우로 직접 면접하지 않는 이상 파악될 수 없는 경우였다.

마지막으로 비아파트 지역 가구에 대한 행정자료와 면접자료를 비교했다. 문제 발생 가구에 대한 조치가 이루어진 실거주 가구 996가구 중 행정자료가 누락한 가구 수는 무려 245곳이었다. 행정자료에서 약 25%의 실거주 가구가 누락되었다는 점에서, 비아파트 지역에서 표집원 검토를 통한 가구 명부 향상이 이루어지지 않는다면 표본의 대표성 문제가 충분히 제기될 수 있다.

전체적으로 행정자료 기반의 가구 명부는 층 표기 또는 호수 등 상세주소를 많이 누락시키는 면이 있다. 특히 다세대주택 및 빌라가 밀집된 지역에서 두드러졌다. 한편 재개발이 최근 진행된 지역, 혹은 발전하고 있는 신도시 권역의 지역 또는 농촌과 공업 단지가 혼재된 지역의 경우에도 행정자료가 실제 거주현황을 잘 반영하지 못하는 경우가 잦았다. 또한 상가주택의 경우, 행정자료상에 상업공간과 주거공간의 구분 없이 혼재되어 기록되어 있고, 저층부 상업공간의 존재로 인하여 전체적인 층수가 오기재되는 등 상세주소의 정확성이 떨어졌다.

이러한 사례들을 종합하자면, 전반적인 조사 과정에서 발생한 주소 불일치 및 오류 수정 사례들은 다음과 같다. 먼저 (1) QGIS 자료가 거주 현실을 제대로 반영하지 못하여 센터 연구자들이 이를 표집원 파견 이전 단계에서 수정한 유형, (2) 행정자료 기반의 주소 목록이 거주 현실을 제대로 반영하지 못하여 표집원이 표집 실시 단계에서 이를 수정한 유형, (3) 비대면 표집의 한계로 인해 발생한 표집원 오류를 면접원이 면접원 단계에서 파악한 유형으로 분류할 수 있다.

(1)의 경우에는 QGIS 자료가 과거 지도 기반으로 작성되어 있어 일부 신축 아파트 단지 또는 재건축 진행 상황이 반영되지 않은 지역들이 존재했다. 센터 연구자들이 네이버 지도 등의 최신 지도 및 행정자료와 비교·대조해가며 집계구 내 조사대상 구역을 변경하는 조치를 취했다. 이렇게 작성된 가구목록은 표집원에게 전해졌는데, (그림 1의 1단계 부분 참조), (2)의 경우에는 표집원의 현장 방문 과정에서 발생하였으며, 다수의 주소 수정이 해당 유형에 해당된다. 센터 연구자 및 표집원들이 해당 표집 과정에서 실거주 여부 파악을 최대한 정확하게 진행하려 했으나, 비대면 표집 방식의 한계로 일부 가구에서는 (3)유형의 오류가 발생했다. 계량기 작동 여부나 택배 존재 여부 등의 방법을 통해 표집원들이 실거주 가구로 판단했으나, 실제 면접원의 대면 확인 결과 농촌 지역과 공단 지역에서 공가(空家)로 판명되는 가구가 일부 존재했다.

V. 결론

본 연구는 조사 과정을 투명하게 공개하려는 노력의 일환으로 한국에서 등한시 되었던 가구조사 표집틀 작성 과정에 관한 상세한 정보를 제공했다. 동시에 행정자료, 표집원 자료, 면접원 자료를 비교하여 향후 표집틀 개선 방향을 모색하였다. 아파트 지역과 비아파트 지역으로 나누어 결과를 요약하면 다음과 같다. 표집원의 의존적 가구 명부 작성을 거치지 않은 아파트 지역의 경우, 행정자료 중 면접원이 방문했을 때 비거주 가구로 밝혀진 아파트 가구의 비율은 1%였다. 비아파트 지역은 3가지를 비교했는데, (1) 행정자료와 비교하여 표집원은 21%나 더 많은 가구를 추가할 수 있었다; (2) 표집원 자료 중 5%는 면접이 가능하지 않은 가구였다; (3) 최종 면접 대상 가구 중 25%는 행정자료에서 누락되었다. 아파트 지역에서는 표집원이 의존적 가구 주소 명부를 작성하지 않았고, 비아파트 지역은 의존적 가구 주소 명부를 작성한 표집원이 오류를 범할 수 있는 한계점이 있지만 다음과 같이 조심스럽게 결론을 내린다.

1. 가구조사 표집틀에서 아파트 지역은 행정자료를 이용할 수 있다.
2. 가구조사 표집틀에서 비아파트 지역은 행정자료의 가구 주소 누락이 심각하기 때문에 표집원의 검토 작업을 거치는 의존적 가구 명부 작성이 필요하다.

우리의 결과는 가구조사 표집틀 작성과 관련하여 조사회사뿐만 아니라 연구자들이 가구조사 표집틀을 집계구에서 작성하려고 할 때 유용하고 실제 구현할 수 있는 기초적인 내용을 제공했다는 기여점이 있다. 3개의 다른 방식의 표집틀 자료를 비교함으로써, 비아파트 지역에서 행정자료 기반 표집틀의 부정확성과 더불어 비용과 시간이 많이 드는 전통적 가구 명부 작성을 대신할 수 있는 방법을 제시했다는 점에서도 의의가 있다. 지역에 따라 다른 방식의 표집틀이 작성되어 면접원에게 제공된다면 면접원의 집계구 내 가구 추출 업무 부담과 자의적 가구 대체 유혹은 줄여 줄 수 있으리라 예상된다.

정확한 표집틀 작성이 단지 전반적 조사 과정의 한 단계에서 나타나는 오류를 줄

이는 것만을 의미하지는 않는다. 면접원에게 배포된 정확한 가구 주소는 무응답 오류와 심지어 측정오류를 줄일 수 있어 조사의 질을 높일 수 있다. 가구 주소가 배포되지 않으면 가구 대체가 용이해져 가구조사 응답률을 제대로 계산하기가 쉽지 않고 무응답 이유도 파악하기 어렵다. 또한, 면접원에게 가구 주소가 정확하게 배포되었기 때문에 면접원은 응답자와의 면접에만 전념하여 측정오류를 줄일 수 있다.

본 연구는 여러 가지 제한점이 있다. 첫째, 비록 아파트 가구는 표집틀로 행정자료를 그대로 이용할 수 있다고 제안했지만 실제로 본 연구에서 아파트 가구 목록을 표집원이 직접 방문하여 검토하지 않았다. 면접 자료와의 비교에 국한되었기 때문에 행정자료에 신축 혹은 누락 주소가 존재한다. 그렇지만 행정자료를 그대로 표집틀로 이용했을 때, 적어도 비거주 가구가 거의 없었다는 점은 아파트 행정자료의 완전성이 높다는 것을 시사한다. 둘째, 표집원이 작성한 가구 명부의 정확성이다. 표집원 자료의 정확성은 실제 면접 자료만을 가지고 확인하였다. 표집원 검토 후 향상된 가구 명부 주소 거의 대부분이 면접원 실사에서 문제가 없었다. 하지만, 실제 표집 과정에서 표집원이 누락된 주소를 그대로 유지하거나 부적절한 주소에 대한 삭제 작업을 제대로 진행하지 않을 수도 있다는 점을 배제할 수 없다(Eckman 2011). 마지막으로, 우리가 작성한 의존적 가구 명부 작성 방법이 가장 효율적인 방법이 아닐 수 있다. Geocoding을 활용하여 집계구에 전체 행정 주소 자료를 연결하여 더 쉽게 처리하는 방법도 있고, 더 혁신적인 방법도 있으리라 추측한다. 물론, 이러한 방법도 행정자료 자체의 정확성은 이용 전에 충분한 검토 과정을 거쳐야 한다.

본 조사 과정에서 학생 표집원들은 중요한 역할을 수행했다. 표집원들의 비전문성과 성실성에 대한 우려에도 불구하고, 전반적인 표집 활동은 상당히 원활하게 진행되었다. 일원화된 사전 교육과 현장 방문 검증 과정을 통해 학생들의 적극적인 업무 수행을 유도했다. 일회성 참여자보다는 과거 조사에서의 표집원 활동 유경험자와 다수 지역의 표집원 활동 지원자를 적극 기용하는 방법이 유효했던 것으로 보인다. 그러나 극소수 학생의 경우, 교육에도 불구하고 조사 목적을 제대로 이해하지 못해 수차례의 재조사를 진행한 사례도 있으며, 표집원의 가구 주소 확인 작업이 비대면 방식으로 이루어져 일부 오류도 발생했다. 향후 표집틀 작성에서 체계적이고 깊이 있는 표집원 교육이 필요하다고 볼 수 있다.

한 가지 특기할 만한 점은 현행 한국종합사회조사의 표집틀 작성 방식이 단순히 표집원에게만 의존하지 않았다는 점이다. 앞서 언급했듯이, 서베이리서치센터 연구

원들이 미리 네이버 지도와 로드뷰 시스템을 활용하여 검토된 표집 주소를 배포하는 방식이었다. 단순히 표집 구역을 지정해 주기만 하는 것이 아니라, 표집원이 현장에서 확인해 보아야 할 구역의 범위와 구역 내의 건물들의 주소 및 특성들(다세대주택 추정 여부, 누락 여부 확인 요망 등)까지 표집원들에게 제공하여 가구명부 작성의 정확성을 높이는 데 이바지했다. 물론, 표집원들에게 실제로 얼마나 큰 도움이 되었는지는 분명하지 않다. 하지만 적어도 실제 면접원이 가구를 방문했을 때 문제가 비교적 적었던 것으로 보아, 표집원에게 상세한 집계구 구역 정보를 제공하는 방식은 상당히 유효했다고 보인다.

본 연구는 비아파트 지역에서 행정자료가 표집틀을 좌우해서는 안 된다는 점을 밝혔는데, 국가 승인 통계인 가구조사도 유사한 문제에 노출되었을 수 있다. 따라서 향후에 통계청에서 제공하는 조사구의 가구 명부를 표집원 자료 및 면접원 방문 결과 자료와 비교하는 연구가 수행된다면 행정자료의 오류 정도나 활용성에 대해 시사점을 제시하리라 생각한다.

양질의 조사자료는 조사과정의 모든 단계에 걸쳐 오류를 줄이려는 노력 없이는 이루어질 수 없다. 전화조사의 경우 선거여론조사에 국한되어 통신사의 전화번호 표집틀에 기초한 안심번호를 이용하여 전화조사의 정확성을 높이게 되었다. 그러나 비승인 통계 가구조사 표집틀은 방대한 양의 행정자료가 공개되어 있음에도 불구하고, 이에 대한 품질 검증은 물론 적절하게 활용되지 못하고 있으며, 더불어 면접원에게 가구 추출 업무를 부여하는 상황을 벗어나지 못하고 있다. 본 연구에서 제안하는 표집틀 구축 방법이 가구 조사자료의 질을 개선하려는 조사기관에게 도움이 되기를 바란다.

참고문헌

- 강현철·박승열·김지연·김인수·이동수·황재일·박민규. 2009. “2008 서울서베이 표본추출틀 구축 및 표본추출 사례 연구.” 《조사연구》 10(3): 157-172.
- 구묘정·김솔이·김하영·김지범. 2014. “가구조사 표집틀 작성: 한국종합사회조사.” 《조사연구》 15(4): 153-174.
- 김지범·강정환·김석호·김창환·박원호·이윤석·최슬기·김솔이. 2022. 《한국종합사회조

- 사 2003-2021.》 서울: 성균관대학교 출판부.
- 박인호. 2014. “조사구를 이용한 가구조사 조사수행 결과분석: 2013 식품소비행태조사 사례.” 《조사연구》 15(2): 1-22.
- 박인호. 2016. “가구조사를 위한 이단추출 표본설계에서의 집락선택.” 《한국데이터정보 과학회지》 27(2): 363-372.
- 박진우·변종석·박민규. 2010. “국민건강영양조사 표본설계를 위한 추출틀 구축.”《응용 통계연구》 23(5): 923-932.
- 박현아·박승환·전종우·박진우. 2010. “계속조사 표본설계에서 추출틀 변경에 따른 층화 변수 선정: 국민여행실태조사 사례연구.”《조사연구》 11(3): 103-114.
- GDS Consulting. 2016. “집계구를 이용한 샘플링 틀 요건 정의”
- Dohrmann, S., T.D. Buskirk, A. Hyon, and J. Montaquila. 2014. “Address Based Sampling Frames for Beginners.” in *American Statistical Association Proceedings of the Survey Research Methods Section* pp. 1009-1018.
- Eckman, S. and F. Kreuter. 2011. “Confirmation Bias in Housing Unit Listing.” *Public Opinion Quarterly* 75(1): 1-12.
- Groves, R.M., Fowler F.J. Jr, M.P. Couper, J.M. Lepkowski, E. Singer, and R. Tourangeau. 2009. *Survey Methodology*. Hoboken, NJ: John Wiley & Sons.
- Harter, R., E. Stephanie, N. English, and C. O’Muircheartaigh. 2010. “Applied Sampling for Large-Scale Multistage Area Probability Designs,” in P.V. Marsden & J.D. Wright (Eds.). *Handbook of Survey Research* (2nd ed., pp. 169-197). Bingley, UK: Emerald Group Publishing.
- Iannacchione, V.G. 2011. “The Changing Role of Address-Based Sampling in Survey Research.” *Public Opinion Quarterly* 75(3): 556-575.
- Iannacchione, V.G., J.M. Staab, and D.T. Redden. 2003. “Evaluating the Use of Residential Mailing Addresses in A Metropolitan Household Survey.” *Public Opinion Quarterly* 67(2): 202-210.
- Jones, M., J.M. Brick, and Van De W. Kerckhove. 2023. “Effects of Address Coverage Enhancement on Estimates from Address-Based Sampling Studies.” *Journal of Survey Statistics and Methodology* 11(2): 340-366.
- Kalton, G., J. Kali, and R. Sigman. 2014. “Handling Frame Problems When Address-based Sampling Is Used for In-person Household Surveys.” *Journal of Survey Statistics and Methodology* 2(3): 283-304.
- O’Muircheartaigh, C., S. Eckman, and C. Weiss. 2003. “Traditional and Enhanced Field Listing for Probability Sampling.” *Proceedings of the Section on Survey Research*

Methods Section, American Statistical Association.

- O'Muircheartaigh, C., S. Eckman, N. English, and J. Lepkowski. 2006. "Validating a Sampling Revolution: Benchmarking Address Lists against Traditional Listing." *American Statistical Association Proceedings of the Survey Research Methods Section* pp. 4189-4196.
- Weisberg, H.F. 2005. *The Total Survey Error Approach: A Guide to the New Science of Survey Research*. Chicago: University of Chicago Press.

<접수 2023.11.03; 수정 2023.11.11; 게재확정 2023.11.22>

Using Administrative Data and Dependent Household Listing for the Household Sampling Frame: 2023 Korean General Social Survey

Jibum Kim
(Sungkyunkwan University)
Gwangjong Yu
(Sungkyunkwan University)
Kyuyoung Kim
(Sungkyunkwan University)
Worlhwa Kim
(Sungkyunkwan University)
Dongping Yang
(Sungkyunkwan University)
Sori Kim
(Sungkyunkwan University)

In a household survey employing a multistage area probability sample design, the household addresses list serves as the sampling frame. Despite the crucial role played by the completeness of this sampling frame in ensuring data quality, there is limited knowledge regarding the methods employed in creating the list of household addresses in Korea. In this study, using data from the 2023 Korean General Social Survey (KGSS), we endeavor to shed light on the process of creating the household address list. Furthermore, we aim to evaluate the completeness of three distinct data sources for generating this list, namely administrative data, lister data, and interviewer data. In 2023, we carefully selected 200 Primary Sampling Units (PSUs), which consisted of 117 apartment areas and 83 non-apartment areas. For the apartment areas, we primarily relied on administrative data, while for the non-apartment areas, we opted for dependent household listing. Subsequently, we randomly selected 12 addresses from each area for conducting face-to-face surveys. Our findings reveal a noteworthy contrast between apartment and non-apartment areas. In the case of apartment areas, we observed an impressive 99% match between administrative data and interviewer data. Conversely, in non-apartment households, listers identified approximately 21% more addresses compared to what was available in administrative data. Furthermore, when comparing administrative data with interview data, we noted that administrative data failed to capture approximately 25% of the selected households. In conclusion, our study strongly recommends the involvement of listers in creating household addresses in non-apartment areas. This practice is essential for enhancing the quality of the sampling frame and, consequently, the accuracy of the survey data.

Key words: sampling frame, dependent listing, household face-to-face survey, administrative data, Korean General Social Survey(KGSS)