노동력조사결과의 패널자료화 해외사례연구

- 패널가중치와 응답오차의 문제 -

2006 12

: 년 월

연구자 : 사회통계국 이 지연
e-mail : jylee@nso.go.kr

主 要 內 容

□ 이 연구의 목적은 경활을 종단자료로 활용할 때 발생하는 방법론적인 문제점(패널소실과 응답오차)을 파악하고 해결방안 모색을 위해 해외 사례들을 살펴보고 시사점을 도출하는 것임

□ 미국의 CPS의 사례를 통해 패널소실의 유형과 추정결과의 편향정도를, 패널무응답 보정을 위한 가중치설계는 영국 ONS 분기별 LFS 종단자료 구축과정을, 응답오차로 인한 편향 정도는 미국, 캐나다, 스웨덴, 영국의 사례를 살펴보았음

□ 이상의 논의를 통해 얻어진 시사점은 패널무응답에 대한 보정절차 없이 경활자료를 패널로 구축할 경우, 특히 기간이 길어질수록 추정치의 편향 문제가 심각해질 것으로 판단됨

□ 경활패널 구축시 패널무응답을 보정할 수 있는 다양한 방법들과 그 실제적인 효과분석 및 응답오차로 인한 편향정도와 수준문제에 관한 경험적인 후속연구가 요구됨

Ⅰ. 서 론

최근에 들어 고용의 양과 질, 임금, 근로시간 등의 노동시장 주요 현안에 대한 종합적인 진단과 노동정책의 효과분석을 위한 기초자료로서 종단자료의 중요성이 크게 부각되고 있다. 종단자료는 노동시장에서 개인의 특성과 행위가 시간에 따라 어떻게 변화하는지를 보여줄 수 있는 자료원이다. 이 자료를 통해 분석자들은 한 시점에서 취업자와 실업자의 전체적인 규모(stock) 뿐만 아니라, 특정 시점들 사이에 발생하는 개개인의 노동력상태의 변화(flow)를 파악할 수 있다.

포괄적인 의미로 시간의 흐름에 따른 변화가 측정되는 모든 조사를 종단조사라고 부른다. 이는 크게 특정한 주제를 반복해서 측정하는 반복횡단 조사와 동일한 대상을 반복적으로 측정하는 패널조사의 두 가지 유형으로 구분할 수 있다. 통계청에서 매월 실시하는 경제활동인구조사(이하 경활), 미국의 Current Population Survey(CPS), 영국의 Labour Force Survey(LFS)는 반복횡단조사의 대표적인 사례들이다(이지연, 2006).

이러한 노동력조사에서 추출되는 표본은 각 조사 시기마다 완전히 다를 수도 있고, 일정부분 겹쳐지기도 한다. 만약 주기적으로 실시되는 노동력조사에서 공통으로 조사된 개인들이 있다면, 관련된 정보를 개인별로 매칭할 경우 패널자료로서도 충분히 활용될 수 있다. 실제로 CPS의 경우 오래전부터 월별로 중복되는 표본을 매칭해서 노동력 상태의 이행 추이를 측정해 왔다(U.S. Department of Labor, 2003). 또한 최근에 영국의 ONS에서는 LFS에서 분기별로 공통적으로 응답한 표본을 매칭한 자료와 주요 노동력상태간의 변화를 보여주는 flow변수를 사후적으로 추가한 데이터 셋을 일반에게 제공하고 있다(Office of National Statistics, 2002).

이렇게 반복횡단조사 결과를 패널자료로 구축할 경우 처음부터 패널조사로 설계된 자료들을 이용하는 것에 비해 여러 가지 장점이 있다. 먼저 표본규모가 일반적인 패널조사에 비해서 상대적으로 크고, 지리적인의 정보의 이

용가능성이 높다. 또한 패널들이 조사주제에 따라 코호트나 특정 계층으로 한정되어 있지 않고, 조사결과를 연구자들이 상대적으로 빨리 얻을 수 있다는 장점이 있다(Neumark and Kawaguchi, 2004).

그러나, 반복 횡단조사를 통해 얻어진 조사를 연계해서 종단자료로 분석할 경우 추정 결과가 왜곡될 수 있는 두 가지 방법론적인 문제점이 존재 한다. 무응답 및 표본소실로 인한 편향과 응답오차로 인한 편향이 그 것이다. 일반적으로 반복 횡단조사는 표본으로 추출된 가구가 이사를 가게 되면, 추적조사를 하지 않고 유사한 특성을 가진 다른 가구로 표본을 대체한다. 이렇게 조사된 반복 횡단자료를 패널자료로 사용할 때 패널에서 체계적인 소실이 일어나고 있을 경우 종단자료로서의 추정치에 편향이 생기게 된다[1]. 두 번째 문제는 응답자가 질문을 잘못 이해하거나 관련된 지식이 없어서 부정확한 응답을 해서 생기는 응답오차이다. 예를 들어 한 시점의 경활상태에 대한 정보가 잘못 기록된 자료가 인접한 다른 차수의 자료와 연계되면 유사흐름(spurious flow)이 만들어 지게 된다(Clarke and Tate, 1999). 이러한 문제점은 반복 횡단조사 형태의 노동력조사를 실시한 후 이를 종단자료로도 활용하고 있는 다른 나라에서도 발견되는 문제점들이다.

연구목적

이 연구의 목적은 경활과 같은 가구단위 반복횡단조사 결과를 종단자료로 활용할 때 발생하는 패널소실과 응답오차의 문제를 파악하고 해결방안 모색을 위해 미국, 영국, 캐나다의 사례들을 알아보고, 시사점을 도출하는 데 있다. 먼저 경활이 가진 종단자료적 특성을 간략히 소개할 것이다. 반복횡단조사를 패널자료로 활용할 경우 나타나는 패널소실 유형과 추정결과의 편향 정도는 미국의 CPS 사례를 통해 살펴보고자 한다. 패널무응답을 보정하기 위한 가중치 부여방법은 영국통계청 지난 1992년 부터 횡단조사로 설계된

1) 조사대상자가 거주지 이전 및 사망 등의 이유로 조사에서 탈락되거나 응답을 거절하거나 경우를 패널조사에서는 절단사례(censored cases)라고 지칭한다. 이 때 절단된 사례들이 무작위적으로 발생하는 것이 아니라, 일정한 소득이나 직업을 갖고 있지 않아서 지리적 이동성이 크고 추적하기 어려운 빈곤층과 같이 특정한 계층이나 집단에서 주로 발생할 경우 이를 '패널소실'이라고 한다 (이지연, 2006).

LFS에 연동표본체계를 도입한 이후부터 생산한 분기별 LFS 종단자료를 통해 알아볼 것이다. 마지막으로 반복횡단조사를 패널자료로 사용할 경우 응답오차로 인해 발생하는 종단자료의 편향 정도는 미국, 캐나다, 스웨덴의 사례 중, 특히 캐나다 통계청의 연구결과를 중심으로 알아보고자 한다.

경활의 종단적 성격

경활은 인구주택총조사 자료를 표본프레임으로 사용해서 지역, 거처, 가구주의 특성을 기준으로 층화추출한 표본가구에서 경제활동대상인구의 노동력상태를 조사해 왔다. 1963년 부터 실시된 경활은 그 후 시대의 변화에 따라 조사주기와 조사대상, 조사표와 조사항목등의 변화가 있었다. 조사주기는 1963년부터 1982년 6월까지는 매 분기(3,6,9,12월)별로 실시되었고, 이후 7월부터는 매월조사가 실시되었다. 조사대상자의 최저연령도 국민소득의 급격한 향상과 중학교 진학율이 99%를 상회하기 시작하자 1987년 1월부터는 14세에서 15세로 상향조정되었다 (통계청, 2006).

그러나 표본체계는 1983년부터 4년간의 연동표본 기간 제외하면, 1963년부터 2003년까지는 인구주택총조사를 통해 표본프레임이 갱신되기 전까지 5년간은 반복적으로 조사되는 고정표본체계가 유지되었다. 2003년 부터는 동일한 응답자로 부터 최대 3년간 조사가 실시되고, 표본이 매월 1/36씩 교체되는 연동표본 체계를 도입되면서, 실제 2005년 부터는 매월 일정량의 표본교체가 이루어 지고 있다. 경활은 반복 횡단조사로 설계되었지만 인접한 달의 조사결과들을 개인별로 연계해서 사용한다면 패널자료로서도 충분히 활용될 수 있다는 견해가 제기되었다. 이러한 인식은 비단 최근의 일은 아니다. 한국에서 본격적인 패널조사는 90년대 중반에 최초로 실시되었지만, 이보다 10여년 앞선 1984년에 서로 다른 두 시점상의 경활자료를 매칭 시켜서 사후적으로 종단자료를 구축한 연구가 있었다(류재우・배무기, 1984). 이후 남재량(1997)은 1982년부터 1994년까지의 경활의 하반기 자료들을 연계해서, 남재량・류근관(1999)은 1985년부터 1997년까지 경활의 월별 자료를 연결해서 노동력 상태별 유동율을 측정하는 연구들을 수행했다. 또한 최근에는 경활의 표본가구를 관리하기 위해 만들어진 가구관리명부 자료를 이용해서 경

활을 종단자료화 시킬 때 나타나는 표본 소실의 정도와 특성을 개인과 가구 차원에서 분석한 연구도 있었다(이지연, 2005).

Ⅱ. 패널소실 유형과 추정결과 편향 - 미국의 CPS 패널

1. CPS 연동표본체계와 패널자료

미 센서스국과 노동통계국이 주관하는 CPS는 매월 가구단위의 노동력조사이다. CPS가 가장 최근에 개편된 시기는 1995년 7월인데, 1990년 센서스결과를 표본추출틀로 사용하였다. CPS는 먼저 약 754개의 기초표본단위(primary sampling unit: PSU)를 추출하고, 이 PSU내에서 약 59,000개의 주거단위(housing unit)를 다시 추출하는 다단계 층화표본이다. 미 전역의 3,141개의 카운티와 시가 2,007개의 기초표본단위로 구분되는데, 주로 메트로폴리탄 지역이나 대규모 카운티, 또는 소규모 카운티 그룹으로 이루어 진다. 최종표본단위는(ultimate sampling unit: USU)는 4개의 주거단위로 이루어진다. 매월 약 72,000개의 주거단위를 조사대상으로 선정하는데, 이중 실제 조사가 가능한 주거단위는 약 60,000개 정도이다(U.S. Department of Labor, 2003; 강석훈, 2001).

CPS 조사는 1953년에 동일표본을 6개월간 연속 조사하는 체계에서 4-8-4 연동체계로 전환했다. 이 방법은 한 가구가 4개월간은 연속해서 조사되고, 8개월간은 조사되지 않다가, 다시 4개월간 조사되고 나면 표본으로 부터 완전히 빠지게 되는 방법이다. 예를 들어 한 가구에 대한 조사가 올해 처음으로 1월에서 4월까지 매달 진행된 후, 중단되었다가 다음해 1-4월에는 다시 동일가구에 대해 반복 조사가 실시되는 것이다. 한 달 동안 조사되는 가구의 구성을 살펴보면 1/8은 첫 번째 조사되는 가구, 나머지 1/8은 2회차 가구 식으로, 최종 1/8은 8회차 조사가구인 셈이다. 이 달에 조사된 응답가구의 6/8은 지난달에도 조사된 가구이기 때문에, 인접한 두 달 동안은 항상 75%의 표본이 중복된다. 이 방법의 장점은 표본의 중복을 통해 추정치의 변이를 줄

여서 시계열 자료의 비연속성을 감소시키고, 특정가구에 대한 장기간의 응답 부담을 분산시키는데 있다.

이렇게 월간 중복되는 75%의 표본가구는 오래전부터 월별 “grows flow” 통계를 추정하는데 이용되어 왔다. 연속한 두 달 동안의 stock의 변화나 차이는 노동력 상태간 유동성의 순 변화분이 반영된 것으로 추정되어 왔다. 예를 들어, 월별 취업수준의 변화는 첫 달에 비취업자에서 두 번째 달에 취업된 사람의 수에서 정반대 경로를 보인 사람의 수를 뺀 값이다. 월별 “grows flow”의 추정은 분석기간 동안 공통으로 조사된 표본가구의 마이크로 데이터를 종단적으로 매칭시켜서 얻어진다.

CPS자료의 종단적 매칭 절차는 가구와 개인 ID를 매칭키로 사용한다. 그러나 일정한 달에 공통으로 조사된 사람이라도 모두 다 매칭되지는 않는다. CPS는 이사가는 사람을 추적하지 않고, 또한 조사대상자들 중에 무응답이 발생하기 때문에 연속한 두 달간의 매칭 성공율은 잠재적으로 매칭 가능한 표본의 90-95% (전체 표본으로 보면 약 67-71%에 해당) 정도이다.

2. CPS의 패널소실 유형

Harris-Kojetin and Tucker (1998)는 2년에 걸쳐 진행된 총 8개월간의 CPS에 모두 응답한 집단과 무응답이 한번이라도 발생한 집단의 사회인구학적인 특성 차이를 분석했다. 이들은 1994년에서 1995년까지 7개의 코호트(동일한 달에 동일한 차수의 조사에 응답한 사람들)를 대상으로 종단자료 셋을 구축했는데, 가구수준에서는 전체 표본의 약 60%를 매칭할 수 있었다. 비매칭된 사례의 대부분은 총 8개월간의 조사동안 주택소실이 발생했거나, 빈집이거나, 주거용이 아니거나, 이사한 경우들로 분석대상에서는 제외되었다. 45,395가구가 총 8개월간 매칭이 가능했는데, 이중 8개월간 모두 응답한 가구는 63.3%인 반면, 8개월간 모두 무응답한 가구는 1.5%정도였다. 개인수준의 매칭은 총 8개월 동안 적어도 한 번 이상의 응답과 한번이상의 무응답이 있었던 가구에 거주하는 99,639명이 매칭 가능했는데, 이들 중 약 85%는 8개

월 모두 응답했다고 한다.

가구수준에서 8회차 모두 응답한 경우, 부분적으로 무응답이 발생한 경우, 모두 무응답한 경우를 비교한 결과, 모두 무응답한 가구들은 일반적으로 도심지역에 거주하는 경우가 많았다. 모두 응답한 가구와 부분적으로 무응답이 발생한 가구의 특성을 비교해 보면, 후자는 주로 세 들어 사는 사람이고, 가구원수가 적고, 가구소득 항목에 대해 응답하지 않는 경우가 많았다.

가구원 차원에서도 모두 응답한 사람과 부분적으로 무응답이 발생한 사람간에 유의미한 차이가 있었다. 유색인이거나, 대학졸업자이며, 25-34세의 젊은 연령인 경우 부분적으로 무응답이 발생할 확률이 높았다. 응답자의 노동력 상태의 변화가 월별로 특히 후반기 회차에 변동이 많을수록 무응답이 많이 발생하는 경향을 보였다고 한다.

3. CPS 패널의 추정치 편향 및 패널자료 활용 분야

CPS패널은 이사가는 사람들을 추적하지 않기 때문에 매칭가능한 표본이 12개월 동안 전체표본의 70-80% 수준으로 축소된다는 문제점이 있다. 그러나 보다 근본적인 문제는 지리적 이동이 개인의 연령과 성, 경제적인 상태에 따라 매우 선택적으로 발생하는 사건이라는 점이다. 횡단조사에서는 표본으로 선정된 사람이 이사를 가는 경우 유사한 특성을 가진 다른 사람으로 대체하기 때문에 해당시기의 추정치에 별다른 영향을 미치지 않는다. 그러나, 동일인을 반복적으로 측정하는 종단조사에서는 이동자가 한번 표본에서 탈락하게 되면 사례가 절단되기 때문에 추정치에 편향이 발생할 수 있다.

Neumark and Kawaguchi(2004)는 CPS가 추적조사를 하지 않음으로써 발생할 수 있는 추정치의 편향을 동일한 표본프레임과 설계방식을 사용하지만 이사가는 사람을 추적하는 Survey of Income and Program Participation(SIPP)의 결과를 통해 간접적으로 추정하고 있다. 이들은 SIPP를 이주자를 추적조사한 결과가 포함된 원자료 셋과 CPS 처럼 이주자를 표본에서 제외시킨 제2의 자료 셋을 만들었다. 두 개의 자료 셋에서 노동조합이 임금에 미치는 효과와 남자의

결혼이 임금에 미치는 효과를 추정한 후 상호 비교를 통해서 추적조사를 하지 않음으로써 발생하는 추정치의 편향 정도를 평가해 보았다. 연구결과 전자는 유의미한 차이가 없었지만, 후자는 패널소실이 추정값에 유의미한 차이를 미치고 있었다. 이것은 결혼이나 이혼이 거주지 이동과 밀접히 관련되기 때문이다. 그러나, CPS를 종단자료로 사용할 때 발생하는 추정치의 편향 여부는 아직 경험적으로 확고한 결론에 도달한 것은 아니다.

최근에 종단적으로 매칭된 CPS 자료를 이용해서 정책 분석을 실시한 연구들은 다음과 같다: 환율변화와 취업 안정성(Goldberg, Tracy and Aaronson., 1999); 취업 및 빈곤가족의 청소년 취학에 대한 최저임금 효과(Neumark and Wascher, 1996); 빈곤탈출에 근로소득세액공제의 효과(Neumark and Wascher, 2001) 등이 있다. 행동과학 분석은은 이민자와 비이민자의 임금성장 (Duleep and Regets, 1997) 보상차이 효과와 소득에서 성별 임금격차와 성별 분리(Macpherson and Hirsch, 1995), 소득이동성(Gittleman and Joyce, 1996)등이 있다.

Ⅲ. 패널무응답 보정절차 -영국 LFS 종단자료

1. LFS 종단자료 구축과 방법론적 문제점

LFS는 1992년 부터 분기별로 약 6만 가구를 조사하고 있는데, 이 때부터 각 표본가구는 5분기 연속으로 조사되고, 각 분기마다 표본의 20%가 교체되는 연동표본을 도입했다. 이 조사는 횡단자료 생산을 목적으로 만들어 졌으나, 분기별로 개인들의 자료를 연결할 경우 종단 자료로 활용가능하다는 점을 최근에 인식하기 시작했다.

그러나, 종단적으로 자료를 연계할 경우 1) 무응답과 표본소실로 인한 편향, 2) 응답오차로 인한 편향 (특히 경제활동상태의 변화 플로우 생산시 발생하는 문제점)으로 인해 자료가 왜곡될 수 있다는 문제점들이 제기되었다. 이

를 해결하기 위해서 ONS는 Southhampton 연구소와 공동으로 방법론적인 문제를 해결하기 위한 방법을 연구해 왔다. ONS는 현재 분기별 종단자료를 일반에게 제공하고 있다. 이 자료 셋에는 무응답으로 인한 편향을 보정하는 절차에 대한 안내와 함께 이용자가 경제활동상태간의 변화를 보여주는 flow 변수들을 직접 만들어서 사용할 수 있도록 SPSS syntax 코드도 함께 제공하고 있다. 또한 2003년에는 2001년 센서스의 총인구에 맞춰서 가중치를 새롭게 조정한 분기별 자료 셋도 제공했다.

현재 ONS에서 제공하고 있는 분기별 종단자료셋을 이용한 연구들로는 여성과 남성의 노동력 상태 유동성의 특징을 비교하거나, 이직자들의 특성, 비경활상태에서 전환된 사람들의 특성에 관한 분석등이 있다(Young, 2001; McIntyre, 2002.)

2. 패널무응답으로 인한 편향을 보정한 종단자료 구축절차

1) 패널기간

LFS의 종단자료는 2분기와 5분기 패널로 구분된다. 2분기 패널은 92/93 겨울부터 인접한 2개의 분기를 연결한 자료 셋이다. 예를 들어 92/93년 겨울과 93년 봄, 93년 봄과 여름, 93년 여름과 가을, 93년 가을과 겨울자료가 연계되어 있다. 2분기 패널의 표본규모는 횡단자료 셋의 약 80% 정도 수준인데, 실제로는 분기별로 공통으로 조사된 개인자료들만이 연계되기 때문에 이보다 더 작다. LFS의 5분기 패널은 93년 봄 이후부터 연속해서 5분기 동안 응답한 응답자로 구성된다. 연동표본 체계와 표본소실로 인해서 5분기 패널자료는 표본규모가 횡단자료의 20% 수준에 못 미친다는 문제점이 있다.

2) Coverage

LFS 패널자료는 각 분기별로 주요 경제활동 연령에 속하는 계층을 주로 포함시키기 위해서 성과 연령에 따라 커버리지에 약간의 차이를 두고 있다. 여자는 분기패널의 첫 분기에 15-59세인 경우, 남자는 첫 분기에 15-64세로 한정하고 있다.

3) 연계절차

분기별 자료에 공통적으로 응답한 사람들을 매칭시킨 방법은 조사일자, 주소, 가구번호, 가구원번호를 연계해서 새로운 ID를 만들었다. 이때 비매칭된 사례들과 한 분기라도 경제활동상태가 없는 자료, imputation 된 자료는 패널자료 셋에서 제외시켰다.

4) 패널변수생성 방법

패널화된 데이터는 일종의 LFS sub-data set이라고 할 수 있다. ID를 제외한 1분기 패널에는 모든 변수명에 1을 추가하고, 2분기 패널에는 모든 변수명에 2를 추가하는 방식으로 패널변수를 만들었다. 이때 연속해서 조사되지 않은 변수는 해당 분기만 표기했다.

5) 노동력 Flow 범주

LFS 분기패널 자료가 가진 주요한 특징은 세가지 노동력 상태별 유동성을 측정할 수 있는 변수가 포함되어 있다는 점이다. 다음은 2분기 패널자료에서 노동력상태 이행 변수에 속한 하위항목들이다. 실제로 2분기 동안의 취업, 실업, 비경활간의 상태 변화는 9가지, 마지막분기에 경제활동연령에 진입하거나 은퇴연령에 도달한 경우 2가지를 포함해서 총 11가지의 유형이 있다.

<표 1> 2분기 패널의 노동력 Flow 범주

범주 유형	첫분기	마지막 분기
1	16세이상	경제활동연령 진입
2	취업	취업
3		실업
4		비경활
5	실업	취업
6		실업
7		비경활
8	비경활	취업
9		실업
10		비경활
11		은퇴연령 도달

출처 : Office for National Statistics. 2002. "User Guide for Two-quarter and Five-quarter Longitudinal Datasets," *Labour Force Survey Two-Quarter Longitudinal Dataset.*

<표 2> 5분기 패널의 노동력 Flow 범주

	첫 분기	2,3,4 번째 분기	마지막 분기
1	취업(E)		
2	실업(U)		
3	비경활(N)		
4	취업		실업
5	취업		비경활
6	실업		비경활
7	실업		취업
8	비경활		취업
9	비경활		실업
10	취업	실업	취업
11	취업	비경활	취업
12	실업	비경활	실업
13	실업	취업	실업
14	비경활	취업	비경활
15	비경활	실업	비경활
16	취업	취업	비경활
17	취업	비경활	취업
18	실업	취업	비경활
19	실업	비경활	취업
20	비경활	취업	실업
21	비경활	실업	취업
22	3-4번 변화		

출처 : Office for National Statistics. 2002. "User Guide for Two-quarter and Five-quarter Longitudinal Datasets," *Labour Force Survey Two-Quarter Longitudinal Dataset.*

5분기 패널자료에서 노동력 상태의 이행변수가 가진 하위항목들은 최대 243개의 시퀀스가 가능하다. 그러나, 대부분의 하위항목들이 실제로 해당되는 사례수가 매우 적었다. 특히 5분기 동안 3-4번 이상 노동력 상태가 변화하는 경우는 그렇게 많지 않았다. 따라서, 노동력 상태 이행형태별로 유의미한 사례수를 확보하기 위해 분기중 몇 번째 차수에 상태변화가 있었는지에 관계없이 첫 분기와 마지막 분기의 노동력 상태를 중심으로 하위항목의 수를 22개로 간략하게 축소했다.

6) 가중치 부여 방법

LFS의 분기별 패널자료에는 패널가중치 변수가 포함되어 있다. 이 가중치 변수는 무응답으로 인한 편향을 보정하고, 전체 인구 수준에서 추정치를 산출하기 위한 것이다. 먼저 초기 가중치는 첫 분기 횡단자료에서 설계가중치로 이용되었던 주택소유 유형(소유, 개인 렌트, 주택조합 및 지방정부로 부터 렌트로 구분)을 그대로 사용하였다. 이 초기 가중치들에 a single grossing factor를 곱해서 얻은 가중된 표본의 합은 전체 population의 추정치가 된다. 이렇게 초기가중치를 통해서 얻어진 결과는 마지막 분기 가중치 산출하는데도 사용되었다.

CALMAR 소프트웨어를 이용해서 calibration 가중 과정을 2분기 표본에 적용했다. 이 방법은 첫 분기 가중치와 마지막 분기 가중치간의 차이가 적도록 만들면서도, 마지막 가중치가 아래의 4개의 control total에 제한받도록 했다:

a) 2분기 횡단 LFS 자료 셋의 가중을 위해 사용된 인구추정치를 성과 연령(24세까지는 1세 간격, 이후부터는 5세 간격으로)에 따라서 분류
b) 2분기 횡단 LFS 자료 셋의 가중을 위해 사용된 인구추정치를 특정 연령내에서 지역에 따라 분류
c) 연계된 2분기의 특정연령으로 부터 나온 가중된 횡단 추정치를 3가지 경제활동상태에 따라 분류
d) 연계된 1분기의 특정연령으로 부터 나온 가중된 횡단 추정치를 3가지 경제활동 상태에 따라서 분류하고, 비경제활동인구 범주를 줄여서 a)와 c)의 total과 유사하도록 조정

이중 a)와 b)는 population 수준의 추정치를 생산하고, 무응답 편향을 어느 정도 보정하는 효과를 가져오게 된다.

<표3>은 위와 같은 절차를 거쳐서 만들어진 2004년도 2분기 종단자료에서 노동력상태 이행변수의 항목별 사례수 분포이다. 5분기 종단자료의 가중치는 2분기 자료에 사용된 방법을 각 분기별 횡단적 경제활동분포를 고려하여 5분기로 확장한 것이다. 따라서, calibration 방법에 사용된 constraint들을 1분기 뿐만 아니라 2, 3, 4분기에도 그대로 적용해서 5분기까지 일관적인 결과를 유도하도록 만들었다.

<표 3> 2004년 2분기 종단 flow 변수 범주 및 추정치

유형	첫분기	마지막 분기	Unweighted	Weighted	Flow %
1		경제활동연령 진입	269	180,327	
2		취업	34,549	26,324,784	72.7
3	취업	실업	471	357,047	1.0
4		비경활	620	478,703	1.3
5		취업	411	391,109	1.
6	실업	실업	784	681,758	1.9
7		비경활	404	305,351	0.8
8		취업	554	459,783	1.3
9	비경활	실업	336	313,667	0.9
10		비경활	8,944	6,905,055	19.1
11		퇴직연령 도달	210	140,777	
합계			47,552	36,538,361	100.0

출처 : Brook, Keith and Catherine Barham, 2005, "Reliability of the Two-quarter Longitudinal LFS flows data"

7) 패널 표본규모와 공표분계점

표본의 변동성으로 인해서 규모가 적은 집단일수록 추정치의 정확성이 떨어지는 문제가 발생한다. 공표 분계점이란 신뢰성 있는 추정치를 얻을 수 있는 최소한의 표본규모로서 공표 가능한 규모를 의미한다. 일반적으로 횡단 LFS의 분기별 공표분계점은 10,000명이었고, 표준오차(SE)가 약 20%수준이며, 95% 신뢰구간에서 +/- 4,000 정도였다.

2분기 종단자료 셋에도 동일한 원칙이 적용되지만, 연계가 가능한 표본수가 원자료(일상적으로 60,000명)보다는 적기 때문에 분계점을 17,000명 수준으로 상향조정하였다. 5분기 종단자료 셋에서는 공표분계점이 더 높아질 수밖엔 없다. 5분기 모두 연계가 가능한 사례수 자체가 원표본에 비해 적고, 높은 패널소실율로 인해서, 2분기 데이터 셋은 60,000개의 사례가 남아 있었다면 5분기를 데이터 셋은 약 11,000 사례 정도만 존재하기 때문이다. 따라서 표본의 변동성이 증가했기 때문에 분계점은 권장된 17,000명 수준 이상이어야만 신뢰성있는 추정치 사용이 가능해진다. 그러나, 공표분계점이 너무 높아지게 되면 자료의 이용도가 낮아지기 때문에, 몇개의 자료셋에서 나온 결과치를 결합하게 되면 분계점을 낮출 수도 있다.

<표 4> 연계된 패널 웨이브 숫자별 노동력 상태에 따른 공표분계점

연계된 패널 웨이브 숫자	1	2	3	4
취업, 또는 비경활	68,000	34,000	23,000	17,000
실업	130,000	65,000	44,000	33,000
기타	100,000	50,000	33,000	25,000

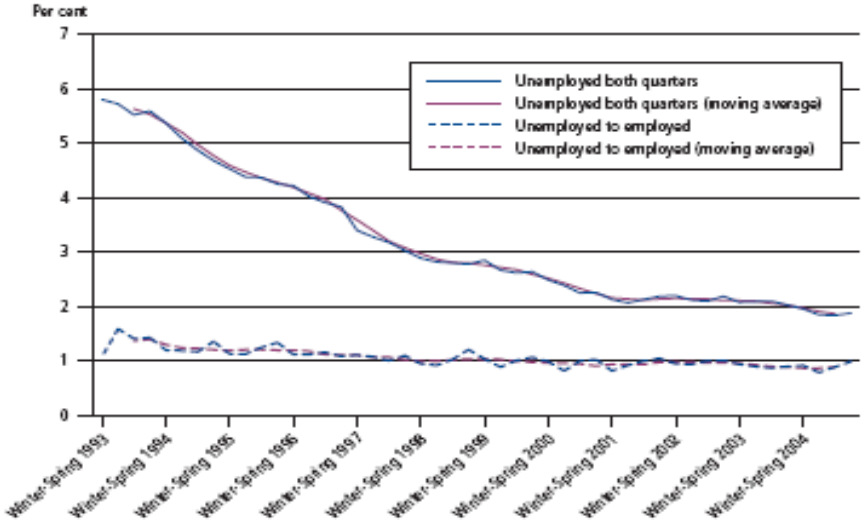
출처 : Brook, Keith and Catherine Barham, 2005, "Reliability of the Two-quarter Longitudinal LFS flows data"

위의 도표상의 숫자보다 낮은 추정치는 표준오차가 20% 보다 높아지기 때문에 추정치로서는 사용되지 않는다. 실업의 공표분계점 숫자가 높은 이유는 이 집단자체의 높은 소실율과 초기 설계 효과 때문이다. 그 외의 분류들은 실제로 그 숫자가 각 데이터 셋에서 많지 않았다.

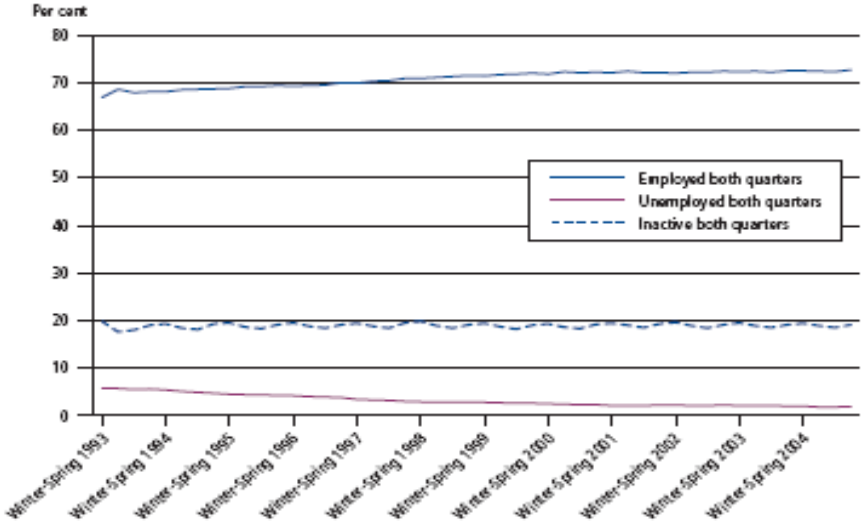
패널 가중치를 조정했음에도, 연구자가 분석을 특정 하위그룹을 선택해 진행하게 되면 무응답 편향에 의해서 왜곡된 결과가 산출될 수 있다. 예를 들어 양분기에 편부모의 성과 가장 나이가 어린 자녀의 연령별 경제활동상태별 교차표를 만든다든가, 18-24세의 청년실업자의 첫분기 교육정도 및 마지막 분기 경제활동상태에 따른 교차표, 마지막 분기에 은퇴연령에 도달한 사람의 양분기의 경제활동상태와 비경활인구라면 그 사유별 교차표를 작성하는 것과 같은 경우이다. 이러한 교차표를 만들게 되면 특정 셀의 사례수가 분계점 이하로 떨어지게 되는 경우가 발생할 수 있다.

이상의 과정을 거쳐서 패널무응답이 조정된 2분기 종단자료의 경제활동상태별 결과는 다음과 같다. 그림 1은 1993년에서 2004년 까지 2분기 패널자료를 이용하여 양분기 동안 실업상태에 머무는 비율 및 실업에서 취업으로 전환한 패널추정치와 분기별 자료를 4분기 이동평균한 결과치를 각각 비교한 것이다. 패널추정치와 이동평균결과치의 추이가 상당히 유사하지만, 실업에서 취업으로 전환한 비율이 실업으로 머무는 비율에 비해 분기별로 기복이 더 심하다는 것을 알 수 있다. 그림2는 동일기간동안 양분기간 취업, 실업, 비경활의 상태가 동일한 비율, 그림3은 실업에서 취업으로, 비경활에서 취업으로, 비경활에서 실업으로 이동한 비율이다

<그림 1> 분기별 flow와 4분기 이동평균 비교: 실업, 실업 → 취업

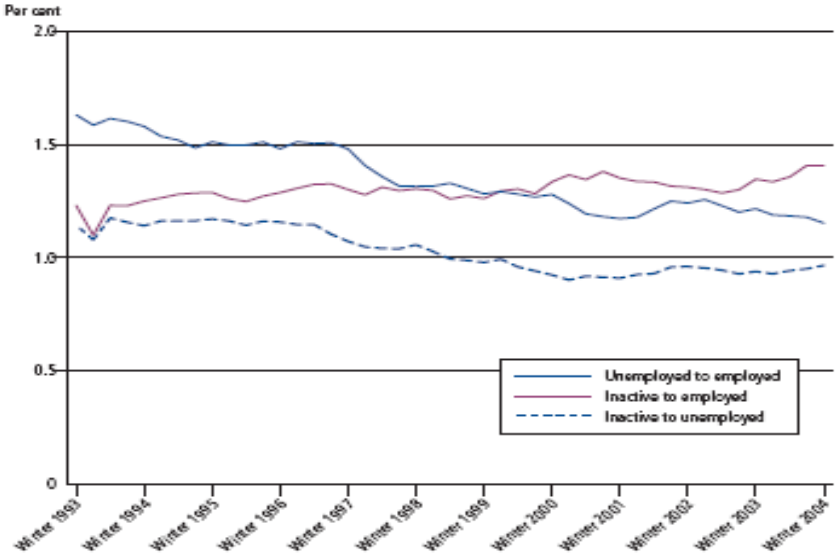


<그림 2> 양분기간 경활상태 변동없는 비율: 취업, 실업, 비경활



출처 : Labour Force Survey, 1993-2004. Brook and Barham (2006)에서 재인용

<그림 3> 양분기간 경활상태 변동비율:실업→취업, 비경활→취업, 비경활→실업



출처 : Labour Force Survey, 1993-2004. Brook and Barham (2006)에서 재인용

Ⅳ. 응답오차로 인한 종단자료의 편향 - 캐나다의 LFS

1. 노동력상태 응답오차

반복횡단조사 결과를 종단자료로 사용하기 위해 패널무응답으로 인한 편향을 보정한다 하더라도 응답오차의 문제는 여전히 존재한다. 응답오차는 연속된 조사에서 응답자가 1회차 이상 부정확한 응답을 하게 될 경우 응답자의 정보가 잘못 분류 될 때 발생한다. 이러한 문제가 발생하는 데는 많은 이유가 있다. 조사지침이 불완전하거나, 질문이나 질문지의 설계가 부실하거나, 응답자와 면접자 모두 조사의 세세한 사항들을 잘 모르거나 주의하지 않는 경우 등이 있다. 또 다른 이유는 조사당시에 응답자가 전환기에 있다면 정확한 상태를 대답하기 어려운데, 이런 경우 조사기준시점을 자의적으로 해석해서 잘못 응답하는 경우가 많다(Brook and Barham, 2005).

응답오차 정도를 측정하는 가장 좋은 방법은 사후조사를 실시하는 것이다. 보다 숙련된 면접자를 이용해서 동일인에게 동일한 질문을 원조사 실시 후 가급적 빠른 시일내에 실시하게 되면 정확한 응답오차의 정도와 범위를 평가할 수 있다. 하지만 연동표본체계를 채택하고 있는 대부분의 노동력 조사의 경우 응답자가 이미 적어도 수 개월 이상 같은 조사에 반복적으로 응답해왔다. 이들을 대상으로 다시 사후조사를 실시하는 것은 그다지 실익이 크지 않아서 실제로 사후조사를 실시한 나라들은 그리 많지 않다.

그러나, 미국, 캐나다, 스웨덴에서는 노동력 조사에 대한 사후조사를 실시한 경험이 있다. 아래의 <표 5>는 각 나라의 노동력조사 원조사와 사후조사 결과중 실업, 취업, 비경활의 세 가지 경활상태 응답을 각각 비교한 오분류 매트릭스이다. 노동력 상태가 일치하지 않는 경우는 응답오차가 있었음을 의미한다. 가장 최근에 원조사 직후에 사후조사를 실시한 스웨덴의 결과를 살펴보자. 스웨덴은 원표본의 7%에 해당하는 표본을 대상으로 사후조사를 실시했다. 그 결과 원조사에서 취업자라고 응답한 사람의 98.1% 만이 사후조사에서도 동일한 응답을 했고, 비경활은 95.7%, 실업자는 오직 90.7% 만이 원조사와 경활상태응답이 같았다.

이 매트릭스를 통해서 두 가지 사실을 알 수 있다. 첫째, 각 나라마다 응답오차에 정도가 차이가 있지만, 취업, 비경활, 실업자의 순으로 응답오차가 커지는 경향은 세 나라에서 공통적으로 발견된다. 두 번째로 노동력 상태중 비경활 에 대한 경계인식이 가장 모호하다는 점이다. 취업과 실업은 비경활에 비해 상대적으로 기준이 명확하기 때문에 원조사와 사후조사간에 취업이 실업으로, 실업이 취업으로 오분류 되는 경우는 많지 않다. 반면에 실업과 비경활이 서로 오분류되는 경우가 가장 많았고, 그 다음으로는 취업과 비경활간의 오분류가 많이 발생한다.

<표 5> 해외 노동력 조사와 사후조사의 노동력상태 결과 비교

미국 CPS 사후조사(1981): 조정표본[2)]

원조사 \ 사후조사	취업	실업	비경활
취업	0.981	0.035	0.020
실업	0.003	0.830	0.010
비경활	0.016	0.135	0.970
계	1.000	1.000	1.000

출처 : Poterba & Summers (1986). 계가 1이 되도록 저자 재계산.

캐나다 LFS 사후조사(1989): 조정표본

	취업	실업	비경활
취업	0.993	0.024	0.007
실업	0.002	0.900	0.008
비경활	0.005	0.076	0.985
계	1.000	1.000	1.000

출처 : Singh A. C. and Rao J. N. K (1995). 계가 1이 되도록 저자 재계산.

스웨덴 LFS 사후조사(1994)

	취업	실업	비경활
취업	0.981	0.023	0.035
실업	0.004	0.907	0.007
비경활	0.015	0.070	0.985
계	1.000	1.000	1.000

출처 : Tzavidis(2004), Brook and Barham(2005)에서 재인용.

2) 이 결과는 조정표본 결과이다.

미국 CPS의 사후조사는 전체 표본의 80%는 조정(reconciled)표본으로 나머지 20%는 비조정(unreconciled) 표본으로 구성된다. 전자는 사후조사조사원이 응답자의 원조사 응답정보를 가지고 있는 상태에서 동일한 질문을 반복한 후 두 응답결과를 상호비교하게 된다. 이때 응답이 차이가 날 경우 어느 쪽이 정확한지 확인해서 응답결과가 최종적으로 조정된다. 후자는 사후조사원이 응답자의 원조사 결과를 모르는 상태에서 동일한 질문을 하고 응답결과를 그대로 기록하여 응답결과를 조정하지 않는다 (Poterba & Summers, 1996).

영국 통계청은 LFS 패널자료의 제공을 검토하면서, Southampton 대학과 함께 패널화된 자료에 포함된 노동력 상태 flow 변수에서 나타날 수 있는 편향의 정도를 최대우도추정법 (Maximum Likelihood Estimation: MLE)을 이용해서 측정해 보았다. 영국에서는 LFS의 사후조사가 실시된 적이 없기 때문에 미국, 캐나다, 스웨덴의 오분류 매트릭스 유형을 그대로 사용했다. 먼저 2분기 패널자료의 flow 변수에서 관측된 노동력 상태 사례수와 오분류 매트릭스에 따라 응답오차가 보정된 자료를 비교해 보았다.

어느 나라의 매트릭스를 사용하느냐에 따라서 편차가 크긴 했지만, 응답오차를 보정한 결과 공통된 결론은 첫분기와 마지막 분기간에 노동력상태가 변하는 사례수가 감소한다는 점이다. 다음으로 이들은 계절적인 요인이나 시기적인 요인으로 인한 응답오차의 변동폭을 측정해 보기 위해 각분기별로 실제 관측값과 응답오차를 보정한 값을, 4분기 평균값을 써서 연도별 관측값과 보정값을 각각 비교해 보았다. 그 결과 분기나 연도와 같은 계절적 요인에 의해서 관측값과 보정값이 크게 차이나기 보다는 어느 나라의 매트릭스를 쓰느냐에 따라서 더 큰 차이가 난다는 점을 발견했다.

2. 캐나다 LFS 종단자료의 응답오차로 인한 편향

캐나다 통계청의 LFS는 연동표본 체계를 사용하는데 표본은 6개월간 연속해서 조사된다. 한 달간 조사되는 표본은 총 6개의 연동그룹으로 구성된다. Stasny(1986)과 Lemaitre(1988)는 LFS의 응답자별로 종단자료 셋을 구축해서 만들어진 노동력 상태간의 gross flow 추정치의 오차를 연구했다. Lemaitre는 추정치에 오차가 발생하는 원인은 응답자의 오류와 함께 노동력 조사의 경제활동상태별 조사기준 때문에 유사흐름이 생성되게 된다는 점을 지적했다. 예를 들어서 사용자의 필요에 의해서 일시적인 요구가 있을 때만 일하는 호출노동자와 사업체없는 자영자의 경활상태를 구분하면서 유사흐름이 생성된다.

Kinack(1991)는 경활과 비경활을 구분하는 구직여부에 관한 질문에 대한

응답의 일관성을 종단적으로 평가해보았다. 이 연구에 따르면 구직여부에 관한 응답이 실제적으로 비일관적인 경우가 많기 때문에, 종단적으로 자료를 사용할 경우 실업과 비경활을 나누기 보다는 취업과 비취업으로 경활상태를 구분해야 응답오차를 줄일 수 있다고 보고했다.

캐나다 통계청의 Rowe and Nguyen(2004)은 LFS를 월별로 패널화시킬 경우 발생할 수 있는 응답오차의 정도를 평가하기 위해 LFS 결과와 세가지 상이한 종단자료들을 비교했다. 먼저 고용보험지급 관련해서 고용주가 직원이 이직한 시점에 보고하도록 되어 있는 행정자료인 고용기록(Records of Employment)자료와 LFS를 20년간 비교해 보았다. 이 결과 두 자료가 대부분 유사했으나, LFS의 표본설계 및 질문지 설계가 변한 시점에서는 차이가 있었다. 또 다른 차이의 원인은 LFS에서는 복수취업자가 한 직업은 이직했을지라도 다른 직업은 여전히 가지고 있다면 실업자로 분류되지 않는다는 차이가 있다.

두 번째로 LFS와 1년 주기로 조사되는 Survey of Income and Labour Dynamics의 취업생존확률을 성별로 비교해본 결과 거의 일치했다. 다만, 6개월 이내의 생존확률은 차이가 있었는데 그 이유는 LFS가 월별 측정인데 비해 SILD는 1년 동안의 회상기간을 갖기 때문인 것으로 추측했다.

세 번째는 LFS에서 응답한 15에서 50세 이하의 여성들이 조사기간중에 낳은 자녀수를 출생동태자료와 비교해 본 것이다. 두 지표간의 출생아수는 완전히 일치하지는 않았지만, 연도별 출산율의 변화 패턴은 유사했다. 이상의 세 가지 실험을 통해서 Rowe and Nguyen(2004)은 LFS 패널이 분절적(fragmentary)인 특성을 가지고 있긴 하지만 유용한 종단자료라고 결론지었다.

Ⅴ. 결론 및 시사점

패널조사는 조사설계에서 공표까지의 전 과정에 시간이라는 차원이 개입되기 때문에 횡단조사에 비해 더 복잡하고 관리가 어렵다는 단점이 있다. 패널의 마모를 최소화시키기 위한 비용이 크기 때문에 최근에는 동일인을 추적조사 하기보다는 집합적으로 동일한 성격을 가진 개인들을 반복해서 조사하는 유사패널 (pseudo panel)조사도 실시되고 있다. 오랜 기간 결과가 축적된 반복횡단조사는 Rowe and Nguyen의 표현처럼 분절적인 사건사 (fragmentary event histories) 자료이다. 반복횡단조사결과를 종단자료화하게 되면, 상대적으로 적은 비용으로 시간에 따라 개인의 특성과 행위가 어떻게 변해왔는지를 보여주는 새로운 통계의 생산을 의미한다.

반복횡단조사 결과를 패널자료로 구축할 경우 처음부터 패널조사로 설계된 자료들을 이용하는 것에 비해 표본규모가 크기 때문에 이와 관련된 지리적인의 정보의 이용가능성이 높아진다. 또한 패널들이 조사주제에 따라 특정 집단으로 국한되지 않고, 조사결과의 업데이트가 상대적으로 빠르다는 장점이 있다.

그러나, 반복횡단조사로 설계된 조사 결과를 패널자료로 활용하기 위해서는 먼저 해결되어야 할 방법론적인 문제들이 있다. 이 연구는 미국, 영국, 캐나다의 사례를 통해서 패널무응답과 응답오차라는 두 가지 방법론적 문제와 해결방안을 살펴보았다. 미국의 CPS를 통해 월별로 중복되는 표본가구를 이용한 flow 통계에서 나타나는 패널소실의 유형과 추정결과의 편향정도를 알 수 있었다. 이러한 패널무응답과 표본소실을 보정하기 위한 대안은 영국통계청의 LFS 2분기 및 5분기 종단자료의 생산과정을 통해 살펴보았다. 마지막으로 반복횡단조사를 패널자료로 사용할 경우 응답오차로 인해 발생하는 추정치 편향의 문제는 영국과 캐나다의 연구결과를 중심으로 알아보았다.

반복횡단조사로 설계되었지만 경활이 가진 패널적인 특성은 오래전부터 인식되었었고, 실제로 경활자료를 이용해서 패널분석을 실시한 연구들도 있다.

그러나, 국내학계에서 반복횡단적으로 설계된 자료를 종단적으로 생산 및 분석할 때 발생하는 문제점들은 구체적으로 지적되지 않았었다. 이 연구를 통해 경활자료를 패널화 시키고 분석하는데 있어서 세 가지 시사점을 얻을 수 있었다.

첫 번째로, 패널무응답에 대한 보정절차 없이 경활자료를 패널적으로 사용할 경우 특히 자료의 기간이 길어질수록 추정치의 편향의 문제가 심각해진다는 점이다. 실제로 1998년에서 2002년까지 경활자료를 이용해서 60회차의 월별 패널을 구축하여 패널소실율을 측정한 연구결과에 따르면 경활패널의 최종 표본소실율은 가구차원 53.5%, 가구원차원 63.3%였다. 경활패널의 문제는 표본소실율의 수준보다도 패널무응답이 무작위적으로 발생하는 것이 아니라 개인차원에서는 주로 연령에 따라, 가구차원에서는 주택소유형태에 따라 체계적으로 발생하고 있다는데 있었다(이지연, 2005).

두 번째로, 경활을 패널자료로 생산하고자 할 경우 여러 가지 보정방법의 효과에 대한 실험과 다양한 측면에서의 결과비교가 반드시 선행되어야 한다는 점이다. 일반적으로 패널가중치를 통해 무응답을 보정하는 방법은 (개인차원에서 또는 가구차원에서 가중치를 부여하는 가에 따라 방법상의 차이가 있긴 하지만) 대부분 표본설계 초기의 추출확률과 각 조사 차수의 무응답률을 반영해서 산출한다(강석훈, 2000; 김영원・김재광・이기재・조유미, 2005). 패널가중치를 사용할 때 기본적인 전제는 응답가구를 표집단위로 해서 반복적으로 추적조사하기 때문에 무응답이 최소화되었다는 가정이다. 그러나, 경활은 거처를 표본선정의 단위로 하고 추적조사를 실시하지는 않는 '거처패널'이기 때문에 패널소실에 일정한 계절성이 나타나고 있다(이지연, 2005). 이러한 특성의 차이가 패널무응답을 보정할 때 충분히 고려되어야 할 것으로 보인다.

세 번째로 패널무응답 보정을 위해 영국과 같이 노동력상태별 flow 변수를 만들어서 패널자료 셋에 추가로 제공하는 방법은 현재로서는 경활에 그대로 적용하기는 어려울 것으로 보인다. 영국의 분기패널자료는 횡단자료로 설계된 조사결과를 패널자료화 할 때 발생하는 추정치 편향의 문제에 포커

스를 맞춰서 연구하고 실천적인 해결방안을 모색해서 실제 패널자료를 생산하고 있기 때문에 경활자료를 패널화하는 데 있어서 가장 모범적인 벤치마크이다. 그러나, 영국의 경우 최대 5분기 패널임에도 노동력 상태 이행의 범주가 22가지를 넘는다. 이에 반해 경활은 2003년 전까지는 월별로 최대 60회, 연동표본이후 36회 이상 월별로 사건사 기록이 생성된다. 이 기간 동안 발생 가능한 모든 노동력 상태의 이행 범주들을 구해서 각범주별로 패널가중치를 조정하는 것은 현재로서는 실익이 크지 않을 것으로 보인다. 물론 사례수가 일정수준을 넘어서는 주요 이행상태 중심으로 범주를 구성할 수도 있지만, 이로 인한 문제점과 효과정도가 구체적으로 검토된 후 결론지을 수 있을 것으로 보인다.

마지막으로, 경활을 패널자료화 할 때 발생하는 응답오차로 인한 편향의 정도와 수준에 대해 보다 심도있는 후속연구가 필요하다는 점이다. 지금까지는 경활을 이용해서 종단분석을 실시한 연구들은 연구자가 직접 경활패널자료를 구축해서 사용했는데, 이때 응답오차의 문제는 고려되지 않았다. 그러나, 미국, 캐나다, 스웨덴, 그리고 영국의 사례들을 참고해 볼 때, 노동력 상태에 관한한 경활도 응답오차의 문제로 부터 자유롭지 못할 것으로 보인다. 특히 노동력 상태 중 실업의 응답오차가 가장 크고, 응답자가 비경활 상태의 기준을 모호하게 해석하는 경향은 세 나라 모두 공통적으로 나타나고 있는 현상이다. 이러한 면에서 노동력 조사 결과를 종단적으로 사용할 경우 실업과 비경활을 나누기 보다는 취업과 비취업으로 구분하는 것이 응답오차를 줄일 수 있다고 한 캐나다의 연구결과 또한 주의를 기울일 필요가 있을 것으로 보인다. 앞으로의 과제는 실제 경활자료를 이용하여 패널가중치와 응답오차의 문제에 대한 해결방안을 제시하는 경험적인 연구에 촛점이 맞춰줘야 할 것으로 보인다.

참 고 문 헌

[1] 강석훈. 2003. "KLIPS의 가중치 부여방안 연구“, 한국노동패널연구 Working Paper Series 2003-04

[2] --------. 2001. "상시인구조사(CPS)의 이해," 노동정책연구 11: 127-146.

[3] 김영원・김재광・이기재・조유미, 2005 “한국노동패널 표본의 대표성과 가중치 보정 방법,” 제6회 한국노동패널 학술대회 논문집.

[4] 남재량. 1997. “우리나라 실업률 추세변화에 관한 연구,” 서울대학교 경제학 박사 학위논문.

[5] 남재량・류근관. 1999. “우리나라 여성 노동력 상태의 동태적 특성연구,” 「사회과학과 정책연구」 21:115-159.

[6] 류재우・배무기. 1984. 한국의 노동시장 플로우와 실업“ 「노동경제논집」 10:55-75.

[7] 이지연. 2006. “종단조사”, 『인구학사전』통계청.

[8] -------. 2005. “가구조사 자료의 종단화 방안” 통계연구보고서 2005-01. 통계청.

[9] 통계청. 2006. “경제활동인구조사 지침서”(내부자료). 통계청.

[10] Brook, Keith and Catherine Barham. 2006. “Labour Market Gross Flows Data from the Labour Force Survey”, *Labour Market Trends,* July. Office of National Statistics.

[11] -------. 2005. “Reliability of the Two-quarter Longitudinal LFS Flows data”, presented at the Tenth GSS Methodology Conference - Changes and flows: design and analysis of repeating and longitudinal surveys. Office for National Statistics, available at:

http://www.statistics.gov.uk/events/gss2005/agenda.asp

[12] Clarke, Paul S. and Pam F. Tate. 1999. "Methodological Issues in the Production and Analysis of Longitudinal Data from the Labour Force Survey," GSS Methodology Series. Office for National Statistics.

[13] Duleep, Harriet O. and Mark .C. Regets, 1997. "Measuring Immigrant Wage Growth Using Matched CPS Files," *Demography* 34:239-249.

[14] Gittleman, Maury, and Mary Joyce. 1996. "Earnings Mobility and Long-run Inequality: an Analysis Using Matched CPS Data," *Industrial Relations* 35: 180-196.

[15] Goldberg, Linda, Joseph Tracy and Stephanie Aaronson. 1999. "Evidence from Matched CPS Data," *American Economic Review* 89: 204-210.

[16] Harris-Kojetin, Brian A. and Clyde Tucker. 1998. "Longitudinal Nonresponse in the Current Population Survey," ZUMA Nachtrichten Spezial, 4: 263-272, available at: http://www.gesis.org/Publikationen/Zeitschriften/ZUMA_Nachrichten_spezial/zn-sp-4-inhalt.htm.

[17] Kinack, M. 1991. "Measuring Data Quality with Longitudinal Data." *1991 Proceedings of American Statistical Association*, section on Survey Research Methods: 514-519.

[18] Lemaitre, Georges. 1988. "The Measurement and Analysis of Gross Flows." Working Paper, Labour and Household Surveys Analysis Division, Statistics Canada.

[19] Macpherson, David A. and Barry T. Hirsch. 1995. "Wages and Gender Composition: Why Do Women's Jobs Pay Less?" *Journal of Labor Economics* 13: 426-471.

[20] McIntyre Andrew. 2002. "People Leaving Economic Inactivity: Characteristics and Flows," *Labour Market Trends,* April. Office of National Statistics.

[21] Neumark, David and Daiji Kawaguchi. 2004. "Attrition Bias in Labor Economics Research Using Matched CPS Files," *Journal of Economic and Social Measurement* 29: 445-472.

[22] Neumark, David and William Wascher. 1996. "The Effects of Minimum Wages on Teenage Employment and Enrollment: Estimates from Matched CPS Data," *Research in Labor Economics* 15: 25-1363.

[23] -------. 2001. "Using the EITC to Increase Family Earnings: New Evidence and a Comparison with the Minimum Wage," *National Tax Journal* 54: 281-317.

[24] Office for National Statistics. 2002. "User Guide for Two-quarter and Five-quarter Longitudinal Datasets," *Labour Force Survey Two-Quarter Longitudinal Dataset,* available at:

[25] http://www.data-archive.ac.uk/doc/4679/mrdoc/pdf/longitudinal.pdf

[26] Poterba, James M., and Lawrence H. Summers (1986). "Reporting Errors and Labor Market Dynamics." Econometrica 54: 1319-1338.

[27] Singh A. C. and Rao J. N. K.1995. "On the Adjustment of Gross Flow Estimation for Classification Error with Application to Data from the Canadian Labour Force Survey," *Journal of the American Statistical Association,* 90:478-488.

Tzavidis Nikos, 2004. "Correcting for Misclassification Error in Gross Flows Using Double Sampling: Moment-based Inference vs. Likelihood-based Inference." Methodology Series Working Papers, 1-33.

[28] Rowe, Geoff and Huan Nguyen. 2004. "Longitudinal Analysis of Labour Force Survey Data." *Survey Methodology* 30: 105-114.

[29] Stasny, Elizabeth A. 1986. "Estimating Gross flows Using Panel Data with Nonresponse: An Example from the Canadian Labour Force Survey," *Journal of the American Statistical Association* 81: 42-47.

[30] U.S. Department of Labor. 2003. "Labor Force Data Derived from the Current Population Survey," Chapter 1 of *BLS Handbook of Methods* (updated version of 1997), available at:

[31] http://www.bls.gov/opub/hom/homch1_a.htm

Young, Mike. 2001. "Time Series Analysis of the Labour Force Survey Two-quarter Longitudinal Datasets," *Labour Market Trends,* August. Office of National Statistics.