행정간행물 등록번호
05400-02400-57-9609

표본개편 연구회 연구보고서

— 가구표본을 중심으로 —

1996. 8

통 계 청

머 리 말

우리청에서는 '95년도 기준 가구표본 개편을 앞두고 청내 표본 실무직원의 표본 전문지식을 보강하고, 현 가구표본 설계방법을 면밀히 검토하여, 미비점을 개선하고자 1995년 9월에 표본개편 연구회를 구성하였습니다.

이 보고서는, '95년 9월부터 '96년 2월까지 5회에 걸쳐 본 연구회에서 토의된 내용을 정리한 것으로, 향후의 가구표본 개편시 조금이라도 도움이 되길 바라며, 그밖에 표본에 관심이 있는 많은 분들께도 보탬이 되었으면 합니다. 끝으로 연구회를 위해 수고해주신 청내 직원여러분과 자문위원님들께 감사를 드립니다.

수록된 내용에 대해서 문의사항이 있으시거나 상세한 자료를 원하시는 분은 조사관리과(222-1852)로 연락하여 주시기 바랍니다.

1996년 8월

표본개편 연구회

통계기획국장 : 김 일현

조사관리과장 : 전 신애

담당사무관 : 이 상은

담 당 자 : 진 영

최 경아

자문위원:

계 훈방(한국보건사회연구원)

류 제복(청주대학교)

박 유성(고려대학교)

이 계오(공군사광학교)

한 근식(한신대학교)

목 차

제1장. 표본개편 연구회 개요

1. 표본개편 연구회의 필요성

표본조사는 사회적 또는 자연적 집단에 관한 정보를 얻는 적절한 수단의 하나로 대상집단 전부를 조사하는 것보다 경제적이고 신속성이 있으며, 효율적으로 정보를 얻을 수 있다는 장점이 있다. 통계청에서 나오는 대부분의 통계는 표본조사를 통해 작성되며 크게 가구표본과 사업체표본으로 나누어진다. 이 표본들은 급변하는 산업구조 및 사회현상을 반영하고 신뢰성있는 통계작성을 위해 매 5년마다 개편되고 있다. 이중 가구표본은 요구된 전제조건을 최대한으로 반영하고, 가구부문 경상조사(경제활동인구조사, 인구동태표본조사, 도시가계조사)를 동일 표본조사구에서 실시가능하도록 다목적 표본으로 설계되어 왔다.

우리청에서는 '95년 인구주택 총조사를 기초로 한 '95년 가구표본설계를 목전에 두고 있다. 이에 우리의 현실과 조사환경을 면밀히 파악하고, 현재 적용되고 있는 '90년 가구표본 설계방법을 검토하여, 개선방안을 모색할 필요성이 제기되었다. 이런 이유로 표본론 전문가들과 청내의 실무진을 중심으로 한 표본개편 연구회가 '95년 9월에 구성되었다.

2. 표본개편 연구회의 목적

표본개편 연구회는 현재의 가구표본을 표본론 전문가들의 이론적 관점과 청내의 표본 및 가구부문 담당직원의 실무적 관점에서 검토하여, 지역별 통계까지 세분화되고 정도(precision) 높은 통계를 작성할 수 있는 효율적인 표본설계방법 및 추정방법을 연구하는데 중점을 두었다. 또한 청내 표본 실무직원의 표본 전문지식을 보강하고, 자꾸만 열악해져가는 조사환경을 개선시킬 표본의 조사방법이나 관리방안을 모색하는 데에도 목적을 두고 있다.

3. 표본개편 연구회의 구성 및 운영

가. 연구회의 구성

연구회는 크게 청내의 표본 담당직원과 가구부문 담당직원, 학계의 표본론과 시계열 전문 교수들로 구성되어 있으며, 특히 '90년 가구표본 설계자인 보건사회연구소의 계훈방 연구위원을 초빙하여 자문을 구하였다. 연구회 구성은 다음 표와 같다.

	소 속	이 름	비 고
표본담당	조사관리과	임 명선 사무관	가구표본 I
	〃	이 상은 사무관	가구표본 II
	〃	김 용철 사무관	사업체표본
	통계정보과	김 규영 사무관	표본설계 전산
	조사관리과	진 영	가구표본 II
	〃	최 경아	〃
가구부문	인구통계과	김 동회 사무관	인구동태표본
	사회통계과	박 영진 사무관	경제활동인구
	〃	우 사임 사무관	도시가계
자문위원	한국보건사회연구원	계 훈방 연구원	'90년 가구표본 설계
	청주대학교	류 제복 교수	
	고려대학교	박 유성 교수	
	공군사관학교	이 계오 교수	
	한신대학교	한 근식 교수	

나. 연구회의 운영 및 주요 토의 내용

'95년 9월 15일에 첫모임을 가진 표본개편 연구회는 매월 1회(2시간)의 정기적인 모임을 6회 가질 계획이었으나, 11월의 인구주택 총조사의 실시로 5회의 모임을 가졌다. 가구표본설계의 소개로 시작된 연구회는 표본규모 결정, 추정치의 계산방법, 교체표본(Rotation Sampling System)과 소지역 추정방법(Small Area Estimation)과 같은 주제를 중심으로 토의되었다. 운영일정 및 토의 안건은 다음 표와 같다.

과 같은 주제를 중심으로 토의되었다. 운영일정 및 토의 안건은 다음 표와 같다.

	일 자	토 의 안 건
1 차	'95. 9. 15	가구부문 조사 개요 및 가구표본설계 소개
2 차	'95. 10. 17	가구표본의 표본규모 결정의 문제
3 차	'95. 12. 15	Rotation Sampling System 도입을 위한 미국의 사례(CPS) 소개
4 차	'96. 1. 26	· 지역통계 생산을 위한 Small Area Estimation의 사례 소개 · 비표본오차 관리 기법
5 차	'96. 2. 16	미국, 일본의 Rotation Sampling 사례를 통한 시계열 유지 방안

※ 5번에 걸친 연구회에서 토의된 내용을 Chapter별로 정리하였으며, 3차·5차(Rotation Sampling System)와 4차(Small Area Estimation)의 안건에 중점을 두고 이 보고서는 작성되었다.

제2장. 현재 다목적 표본설계의 검토

1. 가구표본설계

가. 가구부문 조사개요

현재 통계청 가구표본은 매월 지속적으로 실시하는 경제활동인구조사, 인구동태표본조사, 도시기계조사 등의 경상조사와 사회통계조사, 고용구조조사, 가구소비실태조사, 국부통계조사와 같은 연간 및 특별조사가 있다.

< 경상조사 >

1) 경제활동인구조사(경활조사)

가) 경제활동상태에 관한 통계작성을 위하여 매월 면접조사 실시

나) 취업 및 실업, 취업자의 직업·산업·종사상의 지위 등에 대해 전국 및 시·도별 통계자료 제공

2) 인구동태표본조사(인구동태조사)

가) 인구동태(출생·사망·인구이동·혼인 등)에 관한 통계작성을 위해 실시

나) 경제활동인구조사의 동일표본으로 조사실시

3) 도시가계조사

가) 도시지역의 가계수지에 관한 통계작성을 위하여 가계부 기장방법으로 실시

나) 가구당 수입과 지출에 관한 전도시의 분기별 통계자료를 제공

다) 지역별로는 서울 지역의 경우만 연간 통계를 제공

< 연간 및 특별조사 >

1) 사회통계조사

가) 국민생활의 질적인 측면과 복지정도를 측정하기 위해 실시

나) 경제활동인구조사 표본조사구

2) 고용구조조사

가) 경제활동인구조사에서 생산할 수 없는 노동력 유동실태 같은 취업 및 실업의 심층분석 자료를 생산하기 위해 실시

나) 전국의 약 2,260조사구, 약 147,000가구

3) 가구소비실태조사

가) 전국의 소득 및 소비구조를 파악하여 시·도별 가구소득 및 소비수준을 분석하기위해 실시

나) 경제활동인구조사 표본조사구

4) 국부통계조사

가) 경제적 국력을 파악하여 국가간 차이를 비교하고, 경제정책 기초자료 제공

나) 경제활동인구조사 표본조사구

나. 가구표본설계

1) 설계방법 : 다목적 표본설계

2) 표 본 틀 : 1990년 인구주택 총조사 자료 10% 실사표본 조사구

3) 표본규모 : 경제활동인구조사, 인구동태표본조사 - 약 1,100조사구, 약 35,000가구
도시가계조사 - 약 650조사구, 약 5,500적격가구

♣ 표본규모 결정 전제조건 ♣

① 3개 경상조사를 예산과 조사 담당자들의 관리측면에서 동일 표본조사구에서 실시할 수 있는 다목적 표본으로 설계하며, 경활조사와 인구동태조사는 동일가구에서 병행실시

② 경제활동인구조사 : 도 단위의 분기별 통계를 생산할 수 있도록 한다.

③ 도시가계조사 : 서울의 연간통계 작성과 기타 시도의 연간통계 작성이 가능토록 고려.

④ 조사원 업무량 : 한정된 조사능력 감안 ['95.3월 현재 경활 : 474명(2.3조사구, 71.8가구, 도시가계 : 269명(2.5조사구, 19.2가구)]

⑤ 지역 물가지수 가중치 산출을 위해 기초자료를 수집하는 대상도시에서는 적어도 가구수에 비례하는 수의 표본조사구를 추출

- 위의 전제조건을 만족하면서 도시가계조사와 경제활동인구조사의 주요 항목별 상대표준오차를 사용하여 각각의 표본규모 결정

4) 표본추출방법 : 층화계통추출

5) 기초자료 정리 : 표본틀에서 시설단위 조사구 및 섬지역 조사구를 제외한 18,524개 보통조사구를 1992년 7월 1일 현재 행정구역에 의해 1차 추출하고,

산업별 취업자수(농림어업, 광공업, 사회간접자본 및 기타 서비스업), 거처의 평균가구수 및 건평 등의 기초자료를 집계하여 분류지표를 산출하였으며, 인구주택 총조사의 가구수를 10으로 나누어 반올림한 결과를 크기의 측도로 부여

6) 추출단위 조사구명부 작성 : 6대도시, 도별 시부, 도별 군부별로 작성

7) 표본조사구 추출

· 추출단위 조사구 명부에서 크기의 측도에 비례하는 확률로 경제활동인구조사의 표본조사구를 계통추출한다. [경활조사의 시부 표본조사구 수와 도시가계조사의 표본조사구수가 같은 지역은 경활조사의 표본조사구에서 도시가계조사를 실시하며, 경활조사의 표본조사구수가 많은 지역은 경활조사의 표본조사구 중에서 도시가계조사 표본조사구를 계통추출한다.]

· 추출단위조사구 명부작성시 6대도시와 도별의 시부·군부의 분류지표를 다르게 적용

8) 표본조사구 명부 작성 : 각 표본조사구가 추출된 순서에 따라 작성

9) 표본구역 추출

· 조사구요도 및 가구명부 재작성 : '90년 인구주택 총조사 당시의 기본도를 가지고 현지 확인하여 수정 및 보완한다

· 조사구 분할 : 크기의 측도와 같은 수의 구역으로 조사구를 분할하되, 각 구역의 거처수는 달라도 가구수는 같도록 한다. 표본추출된 구역에서는 전수조사를 한다.

- 1구역을 임의로 추출하고 이 구역을 포함하여 북쪽의 시계바늘 방향으로 인접한 3개의 구역을 표본으로 추출하여 경제활동인구조사, 인구동태표본조사를 실시
- 도시가계조사는 경제활동인구조사의 표본구역 3개 중 임의로 1구역을 추출하여 실시

다. 표본의 유지 및 관리

1) 조사구역내 거처의 철거, 신축 및 가구변동

가) 철거거처는 조사를 중지하고, 신축거처는 즉시 조사

나) 표본조사구 확정 이후 신축거처의 증가 또는 지연증가로 인하여 조사대상 가구수가 '90년 인구주택 총조사 당시의 가구수보다 2배 이상 증가된 경우에는

조사구를 분할

다) 표본조사구의 전 거처가 철거되거나 일부 거처의 철거로 표본조사구 확정당시 가구수의 1/2미만이 남게된 경우, 표본조사구의 가구들이 윤락가·유흥가·시장 등으로 변동되어 일반 가구들이 1/2미만으로 감소된 경우에는 조사구의 특성과 같은 특성을 가진 비슷한 다른 조사구로 대체

라) 조사구내의 신축된 아파트의 분양가구수가 30가구 미만인 때는 즉시 조사실시, 60가구 이상일 때는 새로운 추출단위 조사구로 설정하여 조사실시

마) 전입가구는 전입이후부터 조사대상가구로 하여 계속하여 조사실시하고, 전출가구는 조사를 중지한다.

바) 도시가계조사에서 조사구역의 설정당시 적격가구의 2/3이상이 부적격가구로 변동되거나 적격가구수가 3가구이하로 된 경우에는 조사구내의 경제활동인구조사 구역중 1개의 구역으로 교체

사) 경제활동조사의 조사대상 가구수가 조사구내에서 2배 이상이 된 경우와 자연감소로 1/2미만이 남게 된 경우 구역 재설정

2) 모집단의 변화를 반영하기 위해 신축 아파트에 대한 추가조사구 설정 :
지역별로 총 가구수가 60~70가구가 되도록 1개의 동 또는 2개 이상의 인접한 동으로 조사구를 설정한 후, 각 지역별 표본추출율을 고려하여 새로 추출될 가구수와 조사구수를 정한다.

3) 가구표본조사에서 무응답의 경우 : 조사구내의 표본으로 추출되지 않은 다른 인접구역의 가구로 대체조사 하거나, 자료처리 과정에서 특성이 비슷한 가구의 조사결과를 복사하여 처리하며, 그대로 제외시키는 것은 바람직하지 못하다.

2. 표본규모 결정 및 추출방법

가. 표본규모 결정

제1장에서 제시되었던 토의안건 중에서 모두에게 관심사였던 표본규모 결정에 대해서 알아보자. 표본오차[1]의 크기를 나타내는 방법으로 표준오차(표준편차)[2]와

1) 표본조사의 추정치와 전수조사의 특성치(모수)와의 차이
2) 모수가 포함되는 신뢰구간 설정시 주로 사용

상대표준오차[3]가 있는데, 표본규모 결정시에는 여러 항목의 표본오차를 고려할 수 있는 상대표준오차를 주로 이용한다.

○ 도시가계조사의 표본조사구수

1) 기존표본에서의 표본오차

가) 기존설계에서의 기타 지역을 조사구 수를 비슷하게 3개 지역으로 구분

1 : 경기, 2 : 경북·경남·제주(지역A), 3 : 강원·충북·충남·전북·전남(지역B)

나) 조사구의 분류지표로 주택당 평균 가구수 및 평균 건평을 추가

다) 지역별 표본조사구수 결정은 주요항목[4]의 상대표준오차의 평균을 사용하고 있다.

라) 현 표본설계에서 지역별 표본오차의 특성은

① 6대도시 : 가구수가 많을수록 가계수지에 대한 조사구간의 분산이 큼

② 소도시일수록 조사구간 분산 작고, 대도시일수록 분산이 큼

③ 가구수가 많은 도 일수록 조사구간 분산이 큼

④ 가구수가 비슷할 때에는 시의 수가 많을수록 조사구간 분산이 큼

⑤ 각 시·도의 가구수에 관계없이 표본조사구 수는 일정수준 이상이 되어야 함

2) 시·도별 표본조사구수 고려사항

① 한정된 조사능력

② 시·도별 표본오차의 특성

③ 주택관련 분류지표의 사용효과는 6대도시에서 크며, 대도시일수록 더 커짐

④ 각 시·도별 투입 가능한 조사인력

⑤ 시·도별 표본오차가 전도시의 표본오차에 미치는 기여도

⑥ 각도의 시의 수

○ 경제활동인구조사의 표본조사구 수 결정

1) 기존표본에서의 표본오차

가) 7개 주요항목[5]에 대한 분기별 상대표준오차의 최소치와 최대치의 평균을 대표

---

3) 표준오차를 추정치로 나눈 수치로서 변동계수(Coefficient of variability, CV)라고도 하며, 표본오차의 크기를 상대적으로 비교할 때에 주로 사용된다. 값이 작을수록 좋은 추정량이 된다.

4) 식료품, 광열, 수도, 피복, 신발, 교육, 교양오락

5) 경제활동인구, 취업자, 농림어업, 광공업, SOC 및 기타 산업, 실업자, 피고용자

치로 함

나) 지역별 표본오차의 특성

① 6대도시 : 농림·어업 종사자의 조사구간의 분산이 큼

② 9개도 : 경기에서만 농림·어업 종사자의 조사구간의 분산이 큼

③ 경기, 경남을 제외한 7개 도에서는 실업자의 조사구간 분산이 큼

④ 시·도별 표본오차는 가구수 규모보다는 표본조사구 수에 의해 영향을 받음 즉, 각 시·도의 가구수에 관계없이 표본조사구수가 비슷한 수준이면 전체적인 표본오차도 비슷한 수준이 됨

2) 시·도별 표본조사구 수 고려사항

① 한정된 조사능력

② 시·도별 표본오차의 특성

③ 시·도별 표본오차가 전국 표본오차에 미치는 기여도

④ '89년 조사결과의 표본오차

나. 표본조사구의 추출

표본조사구수가 결정된 후 조사구를 시·도별 단위로 추출하게 되며 그 절차는 다음과 같다.

1) 기초자료 : '90년 인구주택 총조사의 10% 표본조사구를 기준

2) 표본조사구 추출을 위해 현재에 맞는 분류지표를 이용하여, 추출단위 조사구 명부를 작성한다.

3) '90년 표본설계당시 사용한 분류지표

가) 6대도시의 추출단위 조사구 명부

① 1차 분류 : 농림·어업 종사율

② 2차 분류 : 농림·어업, 광공업, 사회간접자본 및 기타 서비스업 종사율

③ 3차 분류 : 독립주택, 아파트, 연립 및 다세대주택, 기타 조사구

④ 4차 분류 : 거처당 평균 가구수

⑤ 5차 분류 : 거처당 평균 건평

⑥ 6차 분류 : 행정구역 번호와 조사구 번호순

나) 도별 시부의 추출단위 조사구 명부

① 1차 분류 : 농림·어업 종사율

② 2차 분류 : 광공업 종사율

③ 3차 분류 : 지역 물가지수 가중치산출

④ 4차 분류 : 거처당 평균 가구수

⑤ 5차 분류 : 독립 주택, 아파트, 연립 및 다세대 주택, 기타 조사구

⑥ 6차 분류 : 행정구역 번호와 조사구 번호순

다) 도별 군부의 추출단위 조사구 명부

① 1차 분류 : 아파트, 기타 조사구

② 2차 분류 : 광공업 종사율

③ 3차 분류 : 사회간접자본 및 서비스업 종사율

④ 4차 분류 : 농림·어업 종사율

⑤ 5차 분류 : 해안지역, 내륙지역 읍면 조사구

⑥ 6차 분류 : 행정구역 번호와 조사구 번호순

< 참고 > 지역별 인구 및 표본조사구

지 역	'90년 인구[1] (명, %)		'90년 일반가구 (가구, %)		보 통 조 사 구[2] (개, %)					
					10% 표본		계 획		시 행	
전 국	42,782,526	100.00	11,354,540	100.00	18,536	100.00	1,076	100.00	1,091	100.00
시 부	31,856,255	74.46	8,462,417	74.53	13,900	74.99	805	74.81	818	74.98
군 부	10,926,271	25.54	2,892,123	25.47	4,636	25.01	271	25.19	273	25.02
서 울	10,526,454	33.04	2,814,845	33.26	4,679	33.66	150	18.63	152	18.58
부 산	3,749,409	11.77	993,375	11.74	1,659	11.94	95	11.80	96	11.74
대 구	2,199,580	6.90	597,150	7.06	973	7.00	70	8.70	70	8.56
인 천	1,800,291	5.65	485,404	5.74	809	5.82	70	8.70	74	9.05
광 주	1,126,100	3.53	287,950	3.40	458	3.29	65	8.07	65	7.95
대 전	1,022,609	3.21	262,193	3.10	422	3.04	63	7.83	63	7.70
경 기	6,033,269		1,619,156		2,583		85		91	
시 부	4,025,066	12.64	1,098,678	12.98	1,785	12.84	55	6.83	59	7.21
군 부	2,008,203	18.38	520,478	18.00	798	17.21	30	11.07	32	11.72
강 원	1,556,587		412,918		689		60		60	
시 부	768,166	2.41	206,400	2.44	335	2.41	30	3.73	30	3.67
군 부	788,421	7.22	206,518	7.14	354	7.64	30	11.07	30	10.99
충 북	1,364,283		354,064		577		60		61	
시 부	690,843	2.17	174,147	2.06	277	1.99	29	3.60	30	3.67
군 부	673,440	6.16	179,917	6.22	300	6.47	31	11.44	31	11.36
충 남	1,972,064		478,579		772		62		62	
시 부	456,163	1.43	109,109	1.29	174	1.25	26	3.23	26	3.18
군 부	1,515,901	13.87	369,470	12.78	598	12.90	36	13.28	36	13.19
전 북	2,047,439		517,181		869		62		62	
시 부	1,126,579	3.54	271,416	3.21	449	3.23	33	4.10	33	4.03
군 부	920,860	8.43	245,765	8.50	420	9.06	29	10.70	29	10.62
전 남	2,475,708		619,767		929		64		64	
시 부	748,009	2.35	180,215	2.13	287	2.06	28	3.48	28	3.42
군 부	1,727,699	15.81	439,552	15.20	642	13.85	36	13.28	36	13.19
경 북	2,789,392		788,896		1,299		68		68	
시 부	1,105,572	3.47	310,681	3.67	504	3.63	30	3.73	30	3.67
군 부	1,683,820	15.41	478,215	16.54	795	17.15	38	14.02	38	13.92
경 남	3,608,818		991,695		1,599		68		69	
시 부	2,193,720	6.89	587,313	6.94	960	6.91	41	5.09	42	5.13
군 부	1,415,098	12.95	404,382	13.98	639	13.78	27	9.96	27	9.89
제 주	510,523		131,367		219		34		34	
시 부	317,694	1.00	83,541	0.99	129	0.93	20	2.48	20	2.44
군 부	192,829	1.76	47,826	1.65	90	1.94	14	5.17	14	5.13

1) '90년 인구주택 총조사 자료(집단가구 제외). 2) '92.10월

3. 추정치 산출

가구표본 조사에서 요구하는 항목 및 특성을 추정하면 다음과 같다.

가. 경제활동인구조사

경제활동인구조사는 표본조사구내에 상주하는 자로서, 현재 만15세 이상인 사람을 대상으로 조사하며, 매월 남·녀별로 6대 도시와 9개 도의 시부와 군부별로 승수를 구한다. 여기서 말하는 인구수는 모두 15세 이상 인구를 말한다.

1) 추계대상인구 구하기

추계대상인구는 남·녀별 전국 추계인구에서 남·녀별 시설단위 조사구 인구[6]를 뺀 값에 지역별 구성비를 적용하여 구한다.

D(전국 경활대상인구)

$= A$(추계된 만 15세이상 전국 인구*) $- E$(전국 제외자 인구**)

(* : 인구과 제공, ** : 사회통계과 제공)

$$J_i\ (\text{지역별 구성비}) = \frac{R_i\ (\text{추계된 만 15세이상의 지역별 인구})}{A}$$

(i = 15개 시·도의 시부·군부)

$$D_i^{\cdot}(\text{시부·군부별 경활대상인구}) = D \times J_i$$

2) 승수 구하기

$$M_i\ (\text{15개 시·도의 시부·군부별 승수}) = \frac{D_i^{\cdot}}{S_i\ (\text{지역 } i \text{에서 조사된 인구})}$$

3) 추정치 구하기

지역별로 구해진 승수(M)를 매월 조사되어진 지역별 취업자수, 실업자수, 비경제활동 인구수 등에 곱해 주면 지역별로 각각 취업자수, 실업자수, 비경제활동 인구수 등의 추정치가 구해진다.

---

6) 제외자 인구라고 부르며 군인, 교도소 수감자 등이 이에 속한다.

$$\hat{X}_{im}\text{(추정치)} = X_{im} \times M_{im}$$

[X : 조사된 수치(취업자수, 실업자수, 비경제활동 인구수, …)

$_i$: 15개 시·도의 시부 및 군부, $_m$: 1월 ~ 12월]

나. 도시가계조사

도시에 거주하고 있으며, 정상적인 가계수지 파악이 가능한 2인 이상의 가구(적격가구)를 조사대상으로 한다. 조사대상 이외의 가구 및 가계수지 파악이 곤란한 가구는 조사대상에서 제외된다. 여기에서 가구란 취사, 취침 및 생계를 같이하는 것을 의미한다.

1) 승수 구하기

전국 가구수와 표본가구수를 대비한 추출율의 역수에 의해 각 15개 시·도별로 승수가 계산되어진다.

$$M_i = \frac{1}{F_i} = \frac{T_i}{S_i}$$

[$_i$: 15개 시·도, M : 승수, F : 추출율,

S : 표본가구수, T : 총 가구수]

2) 가중치 산정 (분기별로 구해짐)

시·도별 가중치 : $W_{qi} = \dfrac{M_i G_{qi}}{\sum_{i=1}^{15} M_i G_{qi}}$ [G : 적격가구수, q : 분기]

3) 가중치의 적용

가) 15개 시·도별 근로자 가구의 년간 월평균 식료품비를 알고 싶다면:

15개 시·도별 각각의 분기별 적격가구의 식료품비의 합을 15개 시·도별 각각의 분기별 적격가구수로 나누면 15개 시·도별 분기별 월평균 식료품비가 되며, 이렇게 계산된 수치들의 평균은 15개 시·도별 근로자가구의 년간 월평균 식료품비가 된다. 지역별로 품목에 대한 추정을 할 때는 가중치가 필요하지 않으며, 이것을 수식으로 나타내면 다음과 같다.

· 15개 시·도별 근로자가구의 분기별 월평균 식료품비

$$= \frac{\text{15개 시·도별 분기별 적격가구의 식료품비 합}}{\text{15개 시·도별 분기별 적격가구수}}$$

$$\overline{X}_q = \frac{1}{3} \sum_{m=1}^{3} \frac{\sum_{j=1}^{n} X_{qmj}}{G_m}$$

[$\overline{X}_q$: 특정지역 식료품비의 분기별 월평균 추정치
X : 식료품비에 대해 조사된 값
q : 분기
m : 분기내의 월을 나타내는 첨자
j : 특정지역 표본조사구
G_m : m^{th}달의 특정지역의 총 적격가구수]

· 15개 시·도별 근로자가구의 년간 월평균 식료품비

$$= \frac{\text{15개 시·도별 분기별 월평균 식료품비 합}}{4}$$

$$\overline{X} = \frac{\sum_{q=1}^{4} X_q}{4}$$

나) 전국 근로자가구의 년간 월평균 식료품비를 알고 싶다면 :

15개 시·도별 각각의 분기별 적격가구들의 식료품비의 합에 15개 시·도별 각각의 가중치를 곱한 수치의 합계를 15개 시·도별 각각의 분기별 적격가구수에 15개 시·도별 각각의 가중치를 곱한 값의 합계로 나누면 전국 근로자가구의 분기별 월평균 식료품비가 된다. 전국 근로자가구의 년간 월평균 식료품비는 분기별로 계산된 4개 값들의 평균을 내면 된다.

· 전국 근로자가구의 분기별 월평균 식료품비

$$= \frac{\text{[15개 시·도별 분기별 적격가구의 식료품비 합} \times \text{15개 시·도별 가중치]의 합}}{\text{[15개 시·도별 분기별 적격가구수} \times \text{15개 시·도별 가중치]의 합}}$$

$$\overline{Y}_q = \frac{1}{3} \sum_{m=1}^{3} \sum_{i=1}^{15} \frac{\sum_{j=1}^{n} X_{qmj} W_i}{\sum_{j=1}^{n_{im}} W_i}$$

[W_i : 15개 시·도별 가중치, $\sum_{i=1}^{15} W_i = 1$
n_{km} : m^{-th} 달의 i^{-th} 지역의 적격가구수]

· 전국 근로자가구의 년간 월평균 식료품비

$$= \frac{\text{분기별 전국 근로자가구의 월평균 식료품비 합}}{4}$$

$$\overline{Y}_y = \frac{\sum_{q=1}^{4} \overline{Y}_q}{4}$$

4. 주요 토의내용(Q:질문사항, A:답변, D:공통 토의내용)

표본설계에 관한 사항

Q : 10% 표본추출시 전국 조사구의 1.7%인 기숙시설 조사구를 전부 표본으로 사용하였으나, 결과적으로 10%표본의 13%에 해당하는 표본이 추출된 것은 현 다목적 표본에 영향을 미칠 수 있지 않을까?

A : 그렇지 않다. 다목적 표본설계에서는 10%표본의 보통 조사구만을 모집단으로 하기 때문에 기숙시설 조사구는 전혀 문제가 되지 않는다. 다만 13%인 기숙시설 조사구가 10% 인구주택 총조사 표본조사구의 문제가 될 수 있다면 그것은 총조사의 문제일 뿐이다.

Q : 10% 표본조사구 추출시 제주도 면부의 추출율을 20%로 준 것은 어떤 이유인가?

A : 전국의 도별 추정치를 내기 위해 과거 자료를 집계해 본 결과 최소 20%는 뽑아주어야 도별집계가 의미가 있다. 그러나 표본설계시에는 제주도의 표본수를 정해놓은 후, 그 수만큼만 계통추출을 하였으므로 다목적 표본설계에서는 문제될 것이 없다.

Q : 그렇다면 추출간격을 5로 주지 말고 10으로 주어서 다시 한번 뽑으면 되지 않는지?

A : 전국단위의 표본추출이기 때문에 두번 추출시에는 예산문제와 인력문제가 있어 처음부터 간격을 조정했다.

Q : 표본규모의 결정은 어떻게 이루어졌나?

A : 도시가계조사는 주요 항목의 상대표준오차의 평균을 사용하여 구하였으며, 경제활동인구조사는 각 항목 상대표준오차의 최소치와 최대치의 평균을 사용하였다.

Q : 조사구를 인접가구로 묶어서 표본구역을 정한 이유는?
랜덤하게 추출할 수는 없는지? Fixed된 3개 구역이 현실을 잘 반영할 수 있는지?

A : 조사상의 편의를 위해 인접가구를 구역으로 정하였으며 랜덤으로 추출할 경우에는 관리가 힘들다.

D : 도시가계조사의 분산을 비교할 때 경기·기타1·기타2로 분류하여 비교하였으나, 그렇게 분류 비교하였을 경우 교통·통신비의 표본오차가 커진다. 경기, 강원, 충남·북, 경남·북으로 나누어서 비교하는 것이 적당하다고 생각된다.

표본추출틀을 작성시 분류지표 선정에 대해

Q : 분류지표를 사용한 근거는?

A : 6대도시와 15개 도의 시부·군부별로 각각 분류지표를 선정했는데 trial and error에 의해 분류지표와 그 순서를 정하고 희소성 있는 자료에 대한 조사가 가능하도록 지역별로 희소성 있는 분류지표를 제일 먼저 적용했다. 예를 들면 대도시일 경우 농림어업, 도의 군부일 경우에는 광공업 종사율 분류지표의 순서를 먼저 적용했다.

Q : 조사구 재편성이란 어떤 뜻인지?

A : 현재 방법에서는 도별 자료 발표를 위해서는 표본규모가 커질 수밖에 없다. 분산은 분류지표에 의해 변하므로 분산을 작게 하는 분류지표를 찾아 그 분류지표에 의해 조사구를 편성하는 것이다.

표본규모 결정에 관련된 사항

Q : 도시가계조사 표본설계시 7개 주요 항목에 대한 평균 상대표준오차를 사용하여 표본의 크기를 정한 것은 타당성이 있는지?
항목당 가중치가 다른 것을 똑같이 평균낸 것은 문제가 될 수 있지 않을까?

A : 도별로 주요 항목의 상대표준오차가 큰 항목이 매우 달라서 평균을 내서 사용했다.

Q : 표본추출 단위에 대하여 오차를 구해 표본규모를 구한다. 조사구의 평균 상대표준오차를 사용해 표본추출에 사용했는데 표본추출 단위는 무엇인가?

A : 한 조사구의 조사가구인 20가구를 1개의 unit로 해서 분산을 구했으므로 표본추출 단위는 조사구의 구역이라고 할 수 있다.

D : 대도시는 가구수가 많아서 조사구간 분산이 큰데, 서울의 경우에는 상대표준오차에 비해 조사구수가 작다.

D : 지역별 자료를 생산하기 위해서는 지역별 표본규모가 도시규모에 상관없이 일정하다. 도별로 허용오차를 정해주면 표본규모가 결정될 것이다.

D : 각 도의 특성을 고려해서 표본규모가 결정된다면 지역별 상대비교가 불가능하지 않는지의 문제점 검토

D : 전체 표본규모를 정해서 지역별로 minimum을 주고 전국에 할당한다면 정확성을 추구할 수 있다.

D : 현재의 방법으로는 오차를 줄이려면 조사구수가 늘 수밖에 없지만, 조사구 설정시 항목별 상대표준오차를 고려하여 조사구를 재편성한다면 조사구간 오차가 줄고 표본규모도 줄일 수 있을 것이다.

최종 표본추출 단위의 크기에 대해

Q : Final Sampling Unit를 Cluster로 한 이유는? Simple Random Sampling으로 하여 Final Sampling Unit를 가구로 하지 않은 이유는 무엇인가?

A : 조사를 한번으로 끝낸다면 S.R.S 방법도 가능하지만 연속 조사이므로 10개의 조사대상가구 중 8가구가 전출가고 그 집이 몇 달간 비워지는 경우도 생길 수 있다. 그러나 지금의 구역추출 방법에서는 그 구역 내의 10가구 중 전 가구가 한번에 이동하는 문제는 없을 것이다. 그리고 인접한 지역은 관리하기

가 쉬워 조사원들의 업무부담도 경감될 것이다.

인구동태표본조사

D : 인구동태표본조사의 결과와 현 신고통계와의 괴리가 많다. 그래서 발표는 않고 있지만 신고통계의 보완자료로 사용되고 있다. 인구동태표본조사는 신고통계의 보완과 인구동태 사항을 분석하기 위해 꼭 필요하다.

D : 현 표본은 인구동태표본조사를 위한 표본이 아니므로 맞지 않는 것은 당연한 일이다. 표본추출시에 인구동태표본조사를 고려해 보았으나 분산이 커서 다른 방법으로 층화를 하든지 아니면 표본을 아주 크게 하여야 했다.

Q : 신고통계와 조사통계중에서 참값은 무엇인가?

A : 신고통계가 신뢰성이 더 높다고 본다. 인구동태표본조사는 발생빈도가 낮아서 월별로 추정한다는 것은 어렵다.

Q : 서울역 시계탑의 인구증가 조정은 어떻게 하는가?

A : 인구주택 총조사결과 출생률을 가정하여 자동으로 시시각각 변하게 조정되어 있다. 일종의 회귀식으로 추정된다.

Q : 원래 출생과 사망에는 어떤 규칙성이 있는가?

A : 구조식으로 추정하고는 있지만 고정표본의 적용에서 문제가 생겨 맞지 않는 것 같다.

Q : 인구동태에서 한 조사구의 30가구만을 조사하는 것은 문제가 있다.

A : 출생 사망률을 제대로 조사하려면 조사구 수를 줄이면서 조사가구수 규모는 커져야 한다. 그러나 업무량 증가와 조사의 어려움 등이 수반된다.

D : 인구동태표본조사의 정도가 높아지려면 발생빈도가 높아야 하는데 현 표본은 비표본오차가 크고 빈도가 낮아 신뢰성이 없다. 그리고 이 표본은 경제활동인구조사 주요항목의 상대표준오차를 이용한 표본이므로 발표하기에는 무리가 있다. 결국 인구동태를 위해서 층화를 다르게 해야만 할 것이다. 이 표본은 인구동태에는 부적절하다.

인구추계에 관해서

Q : 지역별 연령별 추계인구 산정에 대한 인구과의 의견 ?

A : 5년마다 Census에 의해 인구변동 요인인 인구이동을 구하여 전국인구를 추계하고 다시 전국인구 추계에서 연령별 인구를 산정하는데 인구이동 자료의 변동폭이 커서 5년 간격의 평균에는 적용할 수 있으나 매년 적용에는 약간의 문제점이 있다. 지역별 구성비는 Census자료를 바탕으로 5년에 한번 정해지고 있지만 서울과 경기 지역은 변동이 심한 실정이다.

Q : 표본지역을 조사해서 전국인구를 추계할 수 있지 않나?

A : 표본구역을 정하는 데 있어서 일부 시·도는 인구이동량이 1%이고, 서울·경기와 부산·경남은 인구이동에 미치는 영향이 크며, 대전·충남의 경우도 영향이 커지고 있는 등의 문제가 많다. 매년 주민등록 인구이동도 고려하고 있다.

도시가계조사

Q : 도시가계조사의 추출율의 역수를 구할 때 모집단에서 농가가 들어간 건지?

A : 도시가계조사는 표본추출 후에 농가를 제외했다. 농림수산부에서 인구주택 총조사에 의한 조사구를 받아다가 4~5개를 다시 한 조사구로 하여 농가 총조사, 어가 총조사를 실시하고 있는데, 다음 표본설계에서는 이 자료를 이용하여 농가수와 어가수를 구한다면 보다 합리적으로 표본추출을 할 수 있을 것이다.

Q : 도시가계의 가계부를 category 방식으로 체크하는 형식은 어떤가.

A : 400가지의 품목을 모두 매일 기입하는 것은 어렵다.
현 가계부도 매우 괜찮은 방법이다.

Q : 도시가계조사는 지역단위 자료를 생산하기 위해 표본추출을 하는데 전국단위의 정확한 조사를 위해서는 표본규모를 줄여 지역단위 통계를 포기하고 전국단위 추정을 정확하게 하는 것이 좋지 않을까?

A : 지역통계 작성의 중요성은 지방자치 시대를 맞아 더욱 증가되고 있다.
지역별 통계작성시에는 세세한 항목별로 집계는 어려우므로 중요한 기본 항목만을 작성하고 전국단위 추정 때만 세분화한 항목을 작성하자.

Q : 통계청에서는 전국단위 추정을 위해 전국표본을 사용하고, 지역별 통계작성시에는 현 표본에 더해 지방자치 단체가 새로 표본을 추출하여 사용하는 것은 어떤가?

A : 지방에서 통계수요가 확대될 것이므로 고려해야 할 사항이다.

도시가계조사의 근로자외 가구에 대한 발표여부

Q : 도시가계조사의 근로자가구에 대해서만 소득에 관한 자료를 생산하는데 근로자외 가구에 대해서는 왜 조사하지 않나?

A : 조사는 하고 있지만 발표는 하고 있지 않다.

Q : 현재 근로자 가구의 소득만 발표되고 있는데 근로자외 가구에 대한 발표는 왜 이루어지지 않는지?

A : 근로자외 가구는 가계전입 소득만이 조사되며, 그것은 사업소득의 순소득 중 일부일뿐 소득의 개념도 아니고 신뢰도도 떨어진다.

Q : 외국의 경우에는 어떤지?

A : 일본의 경우에는 근로자외 가구는 전혀 조사를 하지 않고 있다.

Fixed표본과 Rotation표본에 관하여

Q : 조사를 특정 구역에서만 Fixed시키지 말고 조사구내 구역을 Rotation하는 것은 어떤지.

A : 도시가계에서는 회수율이 낮아지므로 Rotation은 불가한 면이 있다.

A : 미국의 경우 Chain식 Rotation 방법을 사용하고 있다.

Q : 조사구역이 고정되면 조사가 느슨해지고, 조사원의 자질이 부족한 면이 있지 않을까?

A : 조사관리과에서 조사원들의 자질을 담당하고 있으므로 문제될 것은 없다.

D : Rotation 비율을 20%로 하면 고정표본을 보완한 체감통계가 작성될 것이며, 업무부담도 경감될 것이다.

D : 표본은 모집단을 반영하여야 하므로 Rotation하는 것도 바람직하다.

D : 그러나 소비자 물가 작성을 위해서는 고정되어져야 한다.

D : 그렇다면 도시가계는 Rotation이 불가하겠지만 경활조사는 가능할 것이다.

D : 인구동태표본조사와 경제활동인구조사 등은 Fixed된 표본을 사용할 경우 횡단면적인 성격이 강하므로 위험하다. 예를 들어 출생률을 조사할 경우 50세 이상인 가구만 조사하게 된다면 아주 이상한 결과가 나올 수 있다.

D : 표본의 교체기간을 2.5년으로 하는 것은 어떤가?

D : 총조사의 10%조사구를 모집단으로 하는 표본을 5년간 고정시키는 것은 문제가 있다.

Rotation Sampling에 관해

D : Fixed 된 가구를 조사할 때는 그 Segment의 변화만을 볼 수 있다. 75%를 Fixed시키고 25%를 Rotate하면서 변화율을 조사하면 전체 모집단의 추세도 볼 수 있다. 선진국의 경우는 Rotate방법을 사용하는 경우가 많은데 우리는 표본관리의 효율성을 이유로 왜 Fixed방법만 고집하는가?

D : 매월 1/6조사구가 Rotate되고 조사기간은 조사구당 연속 6개월이 되며, 조사객체는 2/3가 동원되어 3개월 연속으로 조사를 실시하는 방법은 어떤지.

D : 조사인력, 예산, 표본규모와 같은 전제 조건을 파악해야만 표본설계에 관한 개선점을 얘기할 수 있다.

무응답 가구에 관한 처리

Q : 무응답이나 불응가구에 대한 조치로 지역이나 조사구의 평균값을 대체하는 것은 어떤가? 또는 시계열 추정에 의해 대체하는 것은 어떤가?

A : 현재는 무응답 가구와 생활수준 등 특성이 같은 가구를 찾아 대체하는 방법을 사용하고 있으며 고려해 볼 사항이다.

A : 응답율이 80%를 넘는다면 평균대체가 좋다.

비표본오차에 대한 의견

D : 무응답을 줄이기 위해 재방문을 할 경우에도 응답거부에 기인한 무응답 줄일 수가 없다.

D : 무응답에 대한 대체를 하기 위해 유사한 통계자료를 이용하는 콜덱과 현재 조사의 유사한 대체하는 핫덱이 있다. 평균으로 대체할 것인가 혹은 중간값으로 대체할 것인기는 모집단의 분포에 따라 결정될 문제이다.

D : 현재 도시가계조사의 무응답률이 10%일 경우 이의 대체를 위해서는 품목별로 평균값이나 중앙값을 대체해야 한다.

D : Census에서 undercount 문제가 발생했을 때 원래 모집단과 조사된 모집단에 관한 비교가 있어야 한다.

D : Census에 의한 모집단 list와 외부자료에 의한 list를 비교하여 누락이나 중복 여부를 따지는 비표본오차 관리가 있을 수 있다.

5. '95년 기준 가구표본설계시 보완 및 개선 사항

가. 현 표본설계에서 제시된 문제점

1) 한 가구에 대해 5년간 계속 조사하기 때문에 응답자의 응답거부 또는 형식적인 답변의 증가.

2) 신축아파트 조사구 추가시 기존 조사구와 특성(취업률, 실업률 등)이 달라 기존의 결과와 괴리가 발생할 우려가 있음.

3) 전입·전출의 대량 발생시 기존 조사구의 특성과 달라질 수 있음.

4) 경활의 승수계산시 남녀별 추계인구만을 적용함으로 인해 연령별에 따른 경활인구의 분포와 추계인구 분포의 경향이 다름
   → 승수계산시 연령에 따른 추계인구도 감안할 필요성이 있음

나. 표본설계시 고려되어야 할 점

1) 지역표본의 randomness를 분석하여 지역통계 생산가능의 여부를 연구
: 지역적으로 세분화된 통계자료 제공

※현재 가구표본조사 자료의 발표현황
· 경제활동인구조사 : 전국, 15개 시·도(서울, 부산, 대구, 인천, 광주, 대전, 경기, 강원, 충북, 충남, 전북, 전남, 경북, 경남, 제주)의 결과자료 발표
· 도시가계조사 : 전국, 서울의 결과자료 발표

2) 전입, 전출 및 신축 APT의 표본관리를 위한 분석이 필요하다.
(예 : 전입, 전출자의 가구특성 검토)

3) 비표본오차 점검
· 조사 목적의 명시는 명확한가?
- 주요 목적의 명시와 기타 목적의 상반여부
· 표본조사 방법으로 명시된 목적을 달성할 수 있는가?
- 표본조사 이론의 실현성 여부
· 목표모집단과 표본모집단이 일치하는가?
- 표본모집단이 부분모집단인가 아니면 목표모집단이 부분모집단인가?
· 모집단의 정의는 조사목적에 적합하고 명확한가?
· 모수 추정량의 정도를 수량적으로 표현가능한가?
- 통계학자와 이용자간의 협의
· 인원과 비용으로 요구되는 정도는 달성가능한가?
· 요구되는 정도가 실제활용에 적합한가?
· 자료수집 방법은 타당한가?
- 면접조사, 우편조사, 자계식기입, 타계식기입 또는 직접조사
· 조사원에 대한 충분한 교육과 훈련은 실시하였는가?
· 무응답이나 응답거부에 대책이 수립되었는가?
· 표본틀이 완벽한가?
- 누락, 중복, 시기적으로 낡은 것은 아닌지?

· 표본추출단위의 선정은 적절한가?

  - 선정방법의 적절성, 추출단위의 크기의 적합성 등

· 표본오차의 계산가능한가?

· 조사원의 통제업무는 잘되는가?

  - 감독원이나 조사원이 조사업무나 목적을 정확하게 이해하는가?

  - 조사원의 보고체계는 확립되었는가?

  - 조사원의 업무의 질을 평가할 수 있는 방안은?

· 자료입력 및 분석과정 중 확인 절차는 수립되었는가?

· 통계적 분석방법이나 발표방식은 적절한가?

제3장. Rotation Sampling에 대한 외국의 사례 소개 및 연구

1. 서 론

가. 표본조사 방법에 대하여

표본설계는 표본조사 방법에 의해 전반적인 설계가 결정되는데, 표본조사 방법으로는 조사객체의 결정방법에 의해 크게 세 가지로 나누어진다. 최초에 대표성이 높은 표본(표본조사구 및 조사객체)을 선정하고 그것을 장기간 고정하여 조사하는 방법(고정표본)이 기본적 표본조사 방법이며, 매회 전면적으로 교체하는 방법(완전교체표본)과 표본에 순서를 붙여 매회 부분적으로 교체하는 방법(부분교체표본)이 있다.

이 방법들의 일반적 특징을 비교하면 다음과 같다. 고정표본에서 조사구를 고정하는 경우에는 매월 조사구를 교체(부분 또는 전부)하는 경우보다 조사객체의 균질성이 보전되며, 시계열의 정도가 높아 표본간의 상관관계를 이용하여 표본오차를 최대한 줄일 수 있는 장점이 있다. 반면에, 연평균 같은 매월의 표본을 누적하여 산출하는 결과치에 대해서는 조사구를 매월 교체하는 경우보다 정도가 낮고, 추출단위 명부의 정리가 부실할 수 있다. 또한, 조사객체를 고정하면 조사원의 교체가 적어 경비를 줄일 수 있으며, 오랜 기간의 만남으로 인한 친밀감 증대로 조사의 정도(precision)가 증가할 수도 있으나, 장기간 동일한 조사대상자들에게 조사를 반복하여 실시하기 때문에 응답자의 응답기피 증가와 조사담당자의 매너리즘에 의한 형식적 조사의 증가로 비표본오차가 커지고 표본효율이 점차 떨어져 조사의 정확성이 감소될 위험도 있다.

완전교체표본의 경우, 매월 표본을 완전교체하기 때문에 비용이 많이 들며, 월별 표본들의 상관관계가 낮아 표본오차가 증가할 수 있으며, 시계열의 안정성에도 문제가 생길 수 있다. 또한 매번 새로운 독립표본을 조사할 경우, 조사때마다 새로운 명부를 작성해야 하는 번거러움이 있으며, 짧은 조사기간으로 인해 조사담당자와 응답자간의 안면부족으로 조사의 신뢰성에 불신 우려가 있다.

부분교체표본은 위의 2가지 방법의 장점을 살리고 단점을 감소하려는 중간정도의 표본설계방식으로 매월 부분적으로 표본을 교체하여 고정표본의 단점중 큰 비중을 차지하고 있는 응답자의 부담감을 줄일 수 있다. 또한 월간(년간) 중복부분의 상관관계를 이용하여 분산을 줄일 수 있어 표본오차를 줄이는 효과가 나타난다. 조사구를 교체하는데 소요되는 약간의 추가적 비용으로 중복되는 표본에 대해 특정한 분석들을 가능하게 해준다.

나. Rotation Sampling System 정의와 목적

Rotation Sampling Method는 몇 달간 계속 표본이었던 마지막 추출단위(조사구)들의 일부를 새로운 것으로 교체하여 조사를 실시하는 방법이다.

1) 기본가정

가) **급격한 표본의 교체를 방지하기 위해서** 각 층(Homogeneous stratum)마다 가능한 독립이며 같은 정보를 가지는 표본을 그룹화한다. 즉 부표본(sub-sample)이 만들어 져야 한다.

나) **표본의 대표성을 유지하기 위해서** 매달 모든 부표본이 조사되어야 한다.

다) **시계열 유지를 위해서** 동일한 표본의 조사가 일정기간 이루어져야 한다.

2) 장점

가) 월간의 높은 상관관계를 이용하여 각 항목 추정치들의 표본오차를 줄일 수 있는 복합추정방법(composite estimation method)이 가능하다.

나) 표본의 월간(년간) 중복비율은 특성치들 변화에 대한 추정치들의 표본오차에 강한 영향을 미친다. 그러므로, 약간의 추가비용으로 마지막 추출단위(집락:cluster)의 대표성을 증가시켜 월별의 표본오차를 줄일 수 있다.

3) 추정

가) 추정의 관심대상은 월간변동(Monthly change)에 있으며, 월별 추정치의 신뢰성 증가를 위하여 월별중복이 있어야 하며, 년간 추정치의 신뢰성 증가를 위하여 년간중복이 있어야 한다. 이 방법을 사용함으로써 전월대비변화(month to month change)와 전년대비변화(year to year change)에 대한 분석이 가능하다.

나) month to month(year to year) change의 추정 : 표본을 추정할 때의 주

목적은 표본간의 상관관계를 이용하여 표본오차를 최대한으로 줄이는 것이며, 부분교체표본의 추정량은 $\Delta = X_t - X_{t-1}$ 이 되며, 분산은 다음의 식과 같다.

$$Var(\Delta) = Var(X_t) + Var(X_{t-1}) - 2Cov(X_t, X_{t-1})$$

부분교체표본은 표본의 월간(년간)중복이 있으므로 월간(년간)의 상관관계가 비교적 높다. 그러므로 $2Cov(X_t, X_{t-1})$는 양수의 값을 갖게 되며, 그만큼 분산이 감소하게 된다. 표본간에 상관관계가 있을 경우에는 비추정 방식(Ratio Estimates Method)보다는 복합추정방식(Composite Estimates Method)을 사용할 때에 큰 효과를 얻을 수 있다.(ref. The Effects of Rotation Group Bias on Estimates from Panel Survey)

여기서, 현재 부분교체표본을 실시하고 있는 미국의 CPS(Current Population Survey)의 표본설계와 일본 노동력조사의 표본설계를 통해 양국의 부분교체 표본을 비교해 보자.

2. 미국의 CPS 표본설계

CPS(Current Population Survey)는 노동력에 관한 추정치를 제공하기 위해서 Bureau of the Census에서 매월 실시하는 가구표본 조사로서, 미국의 매달 노동력자료(실업률, 임금, 성별, 교육수준, 가족관계, 혼인 등)에 대한 시계열 추정치를 제공해준다.

가. CPS 표본 및 역사

1) CPS 표본

CPS는 두개의 표본집단인 전국표본과 보조표본으로 구성되어 있다.

- 전국표본(National Sample) : 461개의 1차 추출단위인 PSU(Primary Sampling Units)로 이루어져 있으며, PSU는 923개의 군(County)과 시(City)로 구성되어 있다. '75년 전국 표본은 월평균 약 58,000개의 가구

(Housing units) 또는 거처하는 곳(Living Quarter)이 선정되었으며, 그 중에서 13,000가구는 조사에 활용할 수 없는 가구(3000: 비거주·철거 가구, 2000: 조사거부 가구, 8000: 빈가구, 기타)로 실제조사에는 45,000가구가 사용되었다.

- 보조표본(State Supplementary Sample) : 1975년 7월에 추정치의 확실성을 증진시키기 위해 처음으로 14,000개의 보조표본이 사용되었는데, '76년말부터는 매월 11,000개의 보조표본이 사용되었다.

2) CPS의 역사

CPS의 기원은 1930년 경제공황기간 중 실업률에 대한 문제제기에 따라 표본을 근거로 매달 실업자의 수를 파악하기 위해 시작되었다. 처음에는 조사표본으로 추출된 가구를 연속 6개월을 조사하고 교체하는 방식이었으나, 1953년 7월부터 4-8-4 Rotation Sampling System을 사용하였다. CPS의 역사에 대해 살펴보면 <표 3.1>과 같다.

<표 3.1> CPS의 역사

년 월	내 용	비 고
1937	실업자 등록원부로 실시	The Enumerative Check Census에서 실시
1942. 8	실업자 조사(Sample Survey of Unemployment)	
1943. 10	표본개정(68PSU, 125개주와 시)	· The Bureau of the Census에서 실시
1945	68PSU, 25,000표본가구	
1953. 7	4-8-4 Rotation System Method 실시	· 표본크기 동일
1954. 2	230PSU	
1956. 5	330PSU, 40,000표본가구	· Alaska, Hawaii추가
1960. 1	333PSU	· 표본크기 동일
1963. 3	357PSU	· 예산증가로 인해 50% 표본수증가
1967. 1	449PSU, 60,000표본가구	
'71. 12~ '73. 3	461PSU, 58,000표본가구 (특징 : PSU수는 증가한 반면 표본가구수는 감소하였으며 조사구의 크기를 6에서 4로 줄여 보다 더 효율적인 표본을 제공함)	· '70년 CENSUS를 바탕으로 인구수와 분포를 감안하여 표본가구가 감소됨
1975. 9	보조표본(PSU 165개, 14,000표본가구) 도입	
1976. 8	보조표본(PSU 155개, 11,000표본가구) 감소	

나. CPS 전국 표본의 설계

30년동안 표본설계의 신뢰성에 대한 관심이 계속되어 왔다. 정확성에 대한 영향과 비용을 최소화하기 위해 표본설계의 변화가 있었으며, 보다 작은 주에 대해 전국표본으로부터 얻은 추정치의 신뢰성을 증진시키기 위해 보조표본에 대한 설계가 사용되었다.

1) 조사와 표본설계의 특징

가) 조사의 특징

① CPS는 확률표본이다. 즉 표본에서 생성된 조사오차(Survey Error)에 대한 대부분의 추정이 가능하다.

② 표본은 주로 노동력에 관한 주요 사항들에 대한 추정치를 생산하기 위해서 설계되었으며, 일반적으로 주어진 비용 하에서 분산이 최소가 되도록 함.

③ CPS에 의해서 생산되는 통계량은 미국의 노동력 특성치이며, 월별조사는 관공서 지역(the civilian noninstitution population)을 제외한 모든 지역을 포함한다.

④ 표본이 확률적으로 관리되도록 표본지역을 선정한다.

나) 표본설계

CPS표본설계는 다단계 층화표본으로, A-설계와 C-설계로 불리는 두개의 독립적인 전국표본으로 구성되어 있다.

- A-설계 : 군(County) 또는 몇 개의 군으로 된 PSU를 동질의 그룹으로 묶어 층을 만든 후, 각 층으로부터 하나의 PSU를 추출하여 여기에서 표본을 추출한다. 대부분의 표본가구는 Census 목록에서 얻는다.

- C-설계 : 우선 A-설계로부터 표본층들을 선정하고 이들 층내에서 A-설계에서와 같은 원리를 이용하여 PSU나 표본을 선정한다.

위의 두 가지 방법의 다단계 추출에서 마지막 추출단위인 USU(Ultimate Sampling Unit)는 4개의 이웃하고 있는 가구(housing units)들로 구성되어 있다. CPS는 매달 조사되는데 월간 변동(month-to-month change)에 대한 추정치들의 신뢰성을 증진하기 위해서 지난 달(pre-month)의 USU중 3/4을 다음달의 표본으로 계속해서 사용한다. 또한 연간변동(year-to-year change)

의 신뢰성 증진을 위해 일년전 같은 달의 USU중 1/2을 그 기간의 표본에 포함시킨다. 두 표본설계에 관련된 사항들은 다음과 같다.

① A-설계의 USU는 C-설계의 두배이다.

② 이들 각 표본은 자체가중치(self-weighting)를 가진다.
악간의 예외는 있지만 A-설계에 있는 모든 USU들은 전체적으로 같은 확률을 가지므로 각각의 추정치는 표본 USU 각각에 일정한 가중치를 적용할 수 있다. USU는 편의없이(without bias) 추출되므로 비록 어떤 추정치가 악간의 편의를 갖고 있다고 해도 이들은 불편추정치(Unbiased Estimates)로 보며 이는 두가지 설계방식에 모두 적용된다.

③ 전체 CPS의 불편추정치는

$$X = (2/3)X_a + (1/3)X_c$$

[X_a : A-설계의 불편추정치. X_c: C-설계의 불편추정치.
2/3, 1/3 : 가중치]

④ A-설계와 C-설계는 통계적 의미에서 거의 독립이다.

$$E(X_aX_c) = E(X_a)\,E(X_c)$$
$$Var(X) = (2/3)^2\,Var(X_a) + (1/3)^2\,Var(X_c)$$

2) 표본 PSU들의 추출

최종 USU크기를 변화시키지 않은 상태에서 PSU의 수를 증가시키면 표본 USU들을 보다 많은 군으로 분산시키는 효과가 있기 때문에, 추정치 대부분의 표본오차(Sampling Error)를 줄일 수 있게 된다. 하지만 PSU수를 증가시키면 조사비용(PSU내에서 USU간의 이동경비) 또한 증가하게 된다. 이런 단점을 보완하기 위해 **집락의 크기를 6에서 4로 줄였는데**, 이는 표본조사구 수를 줄이고 집락의 크기를 작게 하면 보다 효율적인 표본을 제공한다는 Census Bureau의 연구결과를 반영한 것이다. PSU의 층(strata)수를 결정시에는 분산과 비용을 고려한다.

많은 인구를 가진 PSU는 SR(self-representing)이라는 자체가중치(self-weighting)를 가진 별도의 층으로 다루며, 반드시 표본에 포함시킨다.

3) 표본 PSU의 선정

A-설계에서는 156개의 SR PSU들과 220개의 NSR(nonself-representing) 층에서 각각 하나의 PSU를 추출하여 총 376개의 PSU들을 추출한다. 그리고 C-설계에서는 A-설계의 156개의 SR PSU들과 A-설계의 NSR의 1/2인 110개

의 NSR층에서 각각 하나씩 추출하여 266개의 PSU를 추출한다. 이렇게 결합된 두 설계에는 서로 다른 461개의 표본 PSU들이 있다. NSR층의 PSU 추출은 1970년 census의 총인구수를 사용하여 수정된 층내에서 실시되었으며, NSR은 다음 3가지 목적을 위해 고안되었다.

① census인구에 비례한 확률을 가진 각 층에서 하나의 PSU를 추출하기 위해,
② 구 표본설계의 PSU를 최대한 사용하기 위해,
③ 표본PSU의 수가 주(State)의 인구수를 반영하도록 하기 위해서이다.

4) PSU내에서의 표본추출

1970년 census후의 CPS설계에서는 매달 60,000가구가 표본으로 사용되는데, 그 중 2/3인 40,000가구가 A-설계로부터 추출된다. 70.6백만 가구(또는 거주지) 중에서 40,000가구가 추출되어 표본 추출율은 1:1,800(계획단계에서 변하여 실제 비율은 1:1,968)이 되며, 표본이 SR일 때 전체적인 추출확률은 A-설계의 표본에 있는 모든 단위에 대해서 같다.

$$P_{h_i} \times \frac{1}{W_{h_i}} = \frac{40,000}{70.6\text{백만}} \fallingdotseq \frac{1}{1,800} \Rightarrow \frac{1}{1,968}$$

[P_{h_i} : h번째 층에 있는 i번째 PSU추출확률
$1/W_{h_i}$: PSU내에 있는 표본 USU들의 추출확률]

5) USU 추출

USU추출에 앞선 중간추출 단계에서 추출되는 구획(segment : small area, place)은 크기가 같지 않지만 USU들의 수로 표시될 수 있다. 각 USU들은 특정구획에 속하며 한 구획에 있는 단지 하나의 USU가 주어진 시기에 표본으로 사용된다. 또한, census에서 바뀐된 ED(enumeration districts)는 비교적 큰 지역을 의미이며, 약 350가구를 포함한다. PSU내에서의 부차표본 추출은 1단계로 추출된 PSU내에서 PPS(probability proportional to size)로 계통추출(systematic sampling)하여 ED를 추출한 후, 그 안에서 USU들을 추출한다. USU는 4개의 가구로 구성되어 있으며, 이는 CPS을 하기 위한 최적의 크기이다.

다. ED추출에 대해서

PSU내 ED의 추출과정은 다음과 같이 한다. 각 ED 크기를 정하여 지리적 연관성에 의해 배열한 다음 ED크기에 비례하는 확률로 계통추출한다. census내에서 ED는 PSU내의 SMSA[7]와 non-SMSA그룹 내에서 4개 categories로 나누어진 C, B, U, R[8] 각각에 대하여 **독립적으로 추출**된다. 각각의 8개 분류에서 주어지는 PSU내 ED의 무작위 시작점에서 출발하여 표본을 계통추출한다.

1) ED의 추출

PSU412를 예로 들어서 A29에서의 ED와 USU의 추출에 대하여 알아보자. 이 PSU는 non-SMSA이고 B category가 없기 때문에 단지 U, R부류만 가진다. 표본추출율(sampling rate)은 PSU가 가진 자체가중치(PSU412의 경우는 677.68)로 만들어지며, 시작점은 1에서 677.68사이의 값을 가진다.

가) ED크기의 측정 [표 3.2의 (6)난]

$$M_e = [H_e + P_e / 3] / 4$$

[M_e : ED에서 할당된 USU숫자로서 e번째 선택될 확률로 결정됨.
크기는 정수(반올림), 항상 1보다 크거나 같은 값을 가짐
H_e : ED의 총 가구수(빈집을 포함)
P_e : 집단거주지역 인구수(공공기관의 거주인 제외)]

나) ED추출 [표 3.2의 (8)난]

① 각 ED의 크기인 Me를 합한다. [(7)난]

② 표본선정번호(sample designation numbers)의 선택 [(8)난] :

㉮ 추출간격(sampling interval)인 1에서 677.68사이에서 시작점(RS:random start)을 임의로 선택한다. RS는 각 category마다 선택되며 <표 3.2>에서는 222.14와 538.89가 각각 U와 R의 RS로 선택되며, 거기에 추출간격(677.68)을 더하여 계속하여 다음의 Number들을 선택한다.

---

7) Standard Metropolitan Statistical Area

8) C : SMSA의 central city
B : 인접지역을 포함하여 50,000이상의 인구가 사는 하나이상의 city로 구성된 도시화한 지역(Urbanized area)
U : 도시지역(Urban place: C, B제외)
R : 기타(All other ED's)

㉯ RS를 반올림한 값은 (7)난의 Me보다 적거나 같아야 한다.

③ ED안의 USU결정 [표 3.2의 (9)난]

㉮ RS를 반올림한 다음 (6)난의 Me를 더한다.

㉯ '㉮'에서 (7)난의 ΣMe를 뺀 값이 USU가 된다.

즉, '222+166-328=60'과 같이 계산되며, U category의 A29 USU중 60번째 USU가 A29의 첫 번째 표본이 된다.

<표 3.2> PSU412내 A-설계(A29 USU)의 ED표본추출

(Sampling Interval : 1 in 677.68)

ED categories	Identification codes		Total housing units	Pe	Total USUs		Sample designation numbers	USU in sample A29	Hit number
	county	ED		(GQ)	Me	ΣMe			
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
U	091	512	595	5	149	149			
		512B	43	—	11	160			
		513	8	—	2	162			
		514	665	—	166	328	222.14[F]	60	0001
		514B	101	—	25	353			
		:	:	:	:	:			
		520	742	28	188	934	899.82	154	0002
		:	:	:	:	:			
		526	646	—	162	1738	1577.50	2	0003
					중 간 생 략				
		562	***288***	—	***17***	***5023***	***4965.90***	***15***	***0008***
		563	173	—	43	5066			
		:	:	:	:	:	:	:	
	149	009	294	—	74	5662	5643.58	56	0009
		:	:	:	:	:			
		12B[L]	72		18	5910			
R	091	576	346	1	87	565	538.89[F]	61	0001
		:	:	:	:	:	:	:	:
		506	680	—	170	1339	1216.57	48	0002
		:	:	:	:	:	:	:	:

주) F : RS(random start point) L : last ED

④ 추출과정에서의 임의의 시작점(RS) 결정

㉮ RS는 8개 분류(SMSA, non-SMSA내의 C,B,U,R)의 각 ED내에서 추출되며, 추출간격내의 범위에서 선택된다. 만약 표본추출 간격이

543.21이면 RS는 0.01에서 543.21사이에 포함되어야 한다. 8개 부류의 RS는 아래의 수식과 같이 계산되어지며, PSU_1의 RS가 선택되어졌다는 가정 하에서 PSU_2의 RS를 선택해 보자.

> RS_1 : PSU_1의 ED category의 RS
> W_1 : PSU_1의 표본추출 간격(sampling interval)
> m_1 : PSU_1에서 선택된 USU의 총계
> D_1 : $RS_1 + (m_1-1) \times W_1$이며 이는 마지막 ED의 표본 선정번호(sample designation number)와 같게 된다.
> $M1 : \sum_{e}^{last} M_e$ (PSU_1의 모든 ED크기의 합)
> W_2 : PSU_2의 표본추출 간격
> RS_2 : PSU_2의 ED category의 RS

$$RS_2 = (D_1 + W_1 - M_1) \times (W_2 / W_1)$$

㉯ 만약, PSU_1이 PSU412의 U category라 하면 RS_1은 222.14, W_1은 677.68, m_1은 9(last Hit Number)이고, D_1= 222.14+(9-1)×677.68 = 5643.580이며, M_1은 5910 이 된다. 여기서 W_2가 642.06으로 주어진다면 U category의 PSU_2의 시작점(RS_2)은 다음과 같다.

$$RS_2 = (5643.58 + 677.68 - 5910) \times (642.06 / 677.68) = 389.64$$

⑤ Rotation Sample에서의 USU정의

㉮ <표 3.2>에서 (9)난의 A29USU에 '1'을 더하면 A30USU가 되며, A48USU까지 같은 방법으로 USU를 선택한다. USU는 (6)난의 Me까지 선택할 수 있는데, 만약 (6)난의 Me가 적으면 다음 ED에서 USU를 추출한다.

㉯ <표 3.2>의 ***562ED***에 대하여 예를 들어보자.

(10)난에서 추출되는 8번째 그룹(Hit number)의 A29USU는 15번째, A30USU는 16이고, A31USU는 17이 된다. 그런데, (6)난의 Me는 17밖에 없으므로 8번째 그룹의 A32USU는 다음 ED인 563ED의 1번째가 되며 계속해서 '1'씩 더하여 A33USU~A48USU까지를 추출한다.

라. 4-8-4 ROTATION SYSTEM 표본설계

1) 기본개념(Basic Idea)

가) **월별 추정치의 신뢰성 증가를 위하여 전월표본의 3/4(75%)을 중복**하여 다음달 표본으로 사용한다. 즉 <그림 3.1>의 1974년 3월을 보면 A31의 5,6,7표본과 A33의 1,2,3표본은 전달('74년 2월)과 같은 표본이고, A31의 8표본과 A33의 4표본만이 새로운 것이다.

<그림 3.1> A-설계와 C-설계 표본의 Rotation Chart

Year & Month		Sample & Rotation																																							
		A29 C13								A30 C14								A31 C15								A32 C16								A33 C17							
1972	NOV					5	6	7	8									1	2	3	4																				
	DEC						6	7	8	1									2	3	4	5																			
1973	JAN							7	8	1	2									3	4	5	6																		
	FEB								8	1	2	3									4	5	6	7																	
	MAR									1	2	3	4									5	6	7	8																
	APR										2	3	4	5									6	7	8	1															
	MAY											3	4	5	6									7	8	1	2														
	JUNE												4	5	6	7									8	1	2	3													
	JULY													5	6	7	8									1	2	3	4												
	AUG														6	7	8	1									2	3	4	5											
	SEPT															7	8	1	2									3	4	5	6										
	OCT																8	1	2	3									4	5	6	7									
	NOV																	1	2	3	4									5	6	7	8								
	DEC																		2	3	4	5									6	7	8	1							
1974	JAN																			3	4	5	6									7	8	1	2						
	FEB																				4	5	6	7									8	1	2	3					
	MAR																					5	6	7	8									1	2	3	4				
	APR																						6	7	8	1									2	3	4	5			
	MAY																							7	8	1	2									3	4	5	6		
	JUNE																								8	1	2	3									4	5	6	7	
	JULY																									1	2	3	4									5	6	7	8

나) **년간 추정치의 신뢰성 증가를 위하여 전년동월 표본의 1/2(50%)을 중복**하여 다음해 같은 달의 표본으로 사용한다. <그림 3.1>의 1974년 3월을 보면, A31의 5,6,7,8표본은 '73년 3월의 표본과 같은 것이고 A33의 1,2,3,4표본만이 새로운 것이다.

2) 표본설계

가) 표본(sample) : 추출된 USU들의 집합으로 추출확률은 A-설계에서는 1/1968이고, C-설계에서는 1/(2×1968)이다.

나) Rotation Group의 정의 : 추출된 표본은 거의 같은 크기의 8개의 부차표본으로 다시 나누어지며, A-설계와 C-설계는 각각 8개의 부차표본들의 특정조합으로 매월의 CPS 표본을 형성한다.

다) 4-8-4 System : 이 부차표본에서 4개의 USU를 추출하여 4개월 동안은 조사하고 다음 8개월 동안은 조사에서 제외하였다가, 이듬해에 전년의 같은 달부터 4개월 동안 다시 표본에 포함하여 조사를 실시하며, 그 후에는 CPS표본에서 완전히 제거된다. 이와 같이 **표본의 일부가 매년 중복되므로 시계열 자료를 제공**하게 되며, 예시는 앞의 기본개념(Basic Idea)에 잘 나타나 있다.

3) 기타 사항

가) 추가표본의 생성

① 일정량의 표본을 대체시키는 것은 월간변동에 대한 추출오차를 증진시키지만, 이런 오차는 특성이 유사한 USU들을 대체시킴으로써 최소화할 수 있다. 즉 **이웃해 있거나 가까이 있는 USU들로 대체시킨다.**

② Bureau of Census로 하여금 USU들을 대체할 때 표본들에서 얻은 표본을 다른 조사에 같이 사용하지 않을 수 있다. 추가표본을 얻을 때 USU는 계통추출한다.

③ 작은 수의 USU로 구성된 PSU에서 순환시스템을 사용할 때, 필요한 USU들의 수가 PSU내에 있는 USU들의 수보다 큰 경우 보완하는 방법은 다음과 같다.

- 새로운 표본에는 반드시 사용하지 않은 표본이 포함되어야 한다는 요구조건을 삭제하고 이미 사용했던 USU들을 사용한다.
- NSR층에 있는 규모가 작은 PSU들을 묶은 집락(cluster)을 만들어 사용한다.

나) PSU 대체효과

PSU를 대체함으로써 새로운 조사원을 훈련시켜야 하며, PSU추출과 관련하여 추가비용이 들게 되지만, 대체하는 PSU의 수가 작기 때문에 다른 비용에 관련해서는 매우 적은 비용이 든다. 10년동안 적은 수의 PSU들을 대체함으로써 대체에 따른 효과를 최소로 하여 월간 추정치를 구할 수 있다.

4) 추정방식

월간 표본오차를 감소시키기 위해 이번 달과 다음 달의 추정치들의 가중평균으로 계산하는 복합추정량을 사용한다. CPS의 복합추정량은 다음과 같다.

$$\hat{X}_t = (1-K)\, X_t' + K[\, \hat{X}_{t-1} + d_{t,t-1}\,]$$

여기서,

$\hat{X}_t$: 이번달 t의 추정량

$\hat{X}_{t-1}$: 전달$(t-1)$의 복합추정량

K : 보통 0.5로 지정한다.

$d_{t,t-1}$: 이번달(t)과 전달$(t-1)$의 복합추정량 계산시 각각 그 전달에 공통으로 쓰였던 USU 규모에 대한 차이의 추정량

[$d_{t,t-1} = X_t'' - X_{t-1}'''$

X_t'' : 이번달(t)과 전달$(t-1)$의 공통부분 단순추정량

X_{t-1}''': 전달$(t-1)$과 전전달$(t-2)$의 공통부분 단순추정량]

3. 일본의 노동력조사 표본설계

가. 개요

일본의 노동력조사는 국민생활에서 독신세대 가계수지의 실태파악과 경제 및 사회문제등을 처리하는 기초자료의 제공을 목적으로 매월 실시하고 있다. 1946년 9월에 시험조사 실시후 1947년 7월부터 본격적 조사를 실시했으며, 1961년 10월 이후부터는 현재의 4-8-4 부분표본 교체방식을 채택하고 있다.

나. 표본설계

대상은 국세조사구내 상주하는 사람으로 하며, 전국을 10개의 block(북해도, 동

북, 남관동, 북관동, 갑신, 북륙, 동해, 근기중국, 서국, 구주)으로 나누어 각 block 별로 조사구를 층화한다. 층화의 목적은 결과치의 표본오차를 줄이는 것으로 1980년 조사구의 층화에 있어서는 산업, 종사상의 지위, 주거형태를 기준의 중심으로 하였다. 최종 조사대상자는 한 사람(1인)이며, 인구비례의 확률표본을 기본으로 하고 있다.

1) 표본조사구 추출

표본추출은 층화 2단계 추출법을 사용하였다. 제1차 추출단위는 조사구이며, 제2차 추출단위는 조사구내의 주소로 하였다. 표본조사구는 각각의 조사구 크기에 비례한 추출확률을 부여하여 추출후 조정을 하며, 또한 표본오차의 추정과 rotation sampling을 위해 부표본법(8의 배수로 된 표본을 1/8의 비율로 구분하여 부표본 형성)을 채택하고 있다. 즉, block내 조사구의 층별 배당은 각 층내에 있는 조사구들의 weight합계에 의해 비례배분한 후, 부표본수(8)의 정수배가 되도록 조정한다.

표본조사구의 추출방법을 자세히 살펴보자.

① 각 층마다 추출용 일련 누적번호(A_i)를 아래와 같은 방법으로 부여한다.

$$A_i = A_{i-1} + W_i \qquad i = 1, \cdots, N \quad (N\text{은 조사구 총수})$$

$$[\, A_0 = 0, \quad W_i = i^{th}\text{조사구의 } weight \,]$$

② i층의 추출간격은 층별로 다음과 같이 산출하고, 추출시작번호는 층별 부표본별로 I_i를 초과하지 않는 값을 무작위로 선택한다.

$$I_i = \frac{i\text{층 조사구들의 } weight\text{의 합}}{(i\text{층의 표본조사구수})/8}$$

③ 추출시작번호에 추출간격을 순차적으로 더하여 추출번호를 산출하는데, 산출한 추출번호에 대응한 추출층내 누적번호를 가진 조사구가 표본조사구가 된다. 제2차 추출단위는 한 조사구당 2쌍의 주소를 15개씩 각각 추출한 후, 조사원이 직접 작성한 추출용 리스트를 이용한다. 조사구마다의 weight의 역수를 추출율로 산출하고 추출간격은 weight(환산가구수:15)로

한다. 추출시작번호는 추출간격을 넘지않는 범위에서 무작위로 추출한다.

2) 4-8-4 부분표본 교체방식

고정표본의 장점인 시계열 유지와 조사원과 조사대상자와의 안면부족으로 야기될 신뢰성의 회복을 위해 다음과 같은 방식을 채택했다.

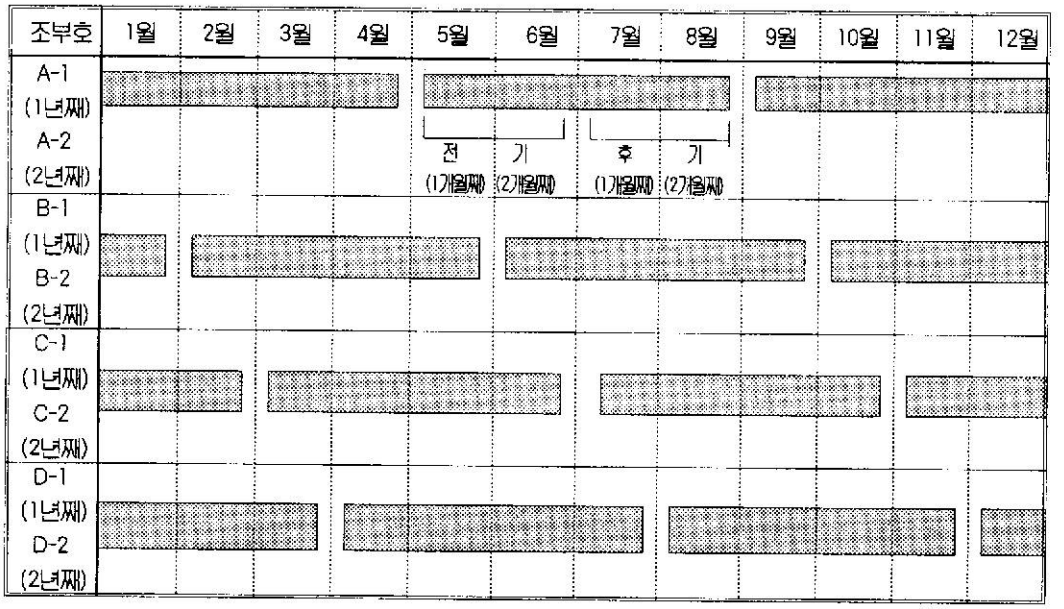
① **동일조사객체(주소)는 2개월간 계속 조사한다.**
이것은 집계결과에 대한 전월분과의 비교의 정도를 향상시키기 위함이다.

② **2개월간 조사를 마친 조사객체(주소)는 익년의 동일시기에 한번 더 조사한다.** 이것은 집계결과에 대한 전년동월분과의 비교의 정도를 향상시키기 위함이다.

③ 하나의 표본조사구를 4개월간 계속 조사한다.

①,②에 의해 시계열의 정도향상이라는 조건은 일단 만족되지만, 이와 같은 구분을 조사구와 같은 지역단위로 하면, 하나의 표본조사구는 2개월만 조사할 수밖에 없으므로 3개월 째부터는 다른 지역으로 교체해야 하며, 대부분의 경우 조사원도 교체하지 않을 수 없다. 이것을 보완하여 **당초에 추출한 조사구가(기준조사구)가 소정의 조사기간을 마친 후에 교체하는 조사구(교체조사구)는 기준조사구와 동일층에서 기준조사구 추출용 일련번호를 시작번호로 하여 그 층의 교체조사구 추출간격을 더하여 계통추출한다.** 추출간격은 기준조사구 추출간격의 1/12로 한다. 또 교체조사구가 합병전의 층의 범위를 초과하여 변한 때에는 합병전 층의 처음으로 되돌아가서 새로 고쳤다. 1980년 국세조사 조사구에서의 개편은 '82년 10월부터 단계적으로 실시하여 '83년 1월에 완료하였다.

<그림 3.2> 조사구 및 조사객체(주소)의 연결상황



<그림 3.2 의 설명>

① 조별부호 : A, B, C 및 D는 각각 조사구에 대한 조사개시월에 의한 구분을 나타내며, '-1' 조사구는 금년에 새로 표본조사구로 추출되어 익년 동일 시기에 재조사하는 조사구(1년째 조사구)이고, '-2'조사구는 작년의 표본조사구로서 금년에 재조사하는 조사구(2년째 조사구)를 나타낸다.

조 별 부 호		조사를 시작하는 달
1 년째 조사구	2년째 조사구	
A - 1	A - 2	1월, 5월, 9월
B - 1	B - 2	2월, 6월, 10월
C - 1	C - 2	3월, 7월, 11월
D - 1	D - 2	4월, 8월, 12월

② 조사구의 조사기간 : 1월에 시작한 조사구는 4월에 조사를 완료한다

③ 조사객체(주소)의 조사기간 : 5월에 시작한 조사구 중 5, 6월의 전반 2개월은 전기 조사분이고, 7,8월의 후반 2개월은 후기 조사분으로 분리한다.

위와 같이 실시하면 매월 1/4의 조사구가 교체되는 동시에 1/2의 조사객체가 교체된다.

다. 추정방법

추계인구를 기준으로 하여 비추정 방식(Ratio Estimator method)[9]으로 추정하며, 그 원리는 다음과 같다.

$$\text{목적항목의 비추정치} = \text{목적항목의 선형추정치} \times \frac{\text{보조항목의 참값}(bench\ mark)}{\text{보조항목의 선형추정치}}$$

노동력조사에서는 남녀별, 지역별(7대도시, 7대도시 이외의 구별), 연령계급별 인구를 bench mark인구로 하고 있으며, 목적항목 및 보조항목의 선형 추정치 X는 다음과 같은 1차식으로 구해진다.

$$\hat{X} = \sum_{h}^{10} \hat{X}_h = \sum_{h}^{10} \sum_{i}^{L_h} \hat{X}_{h_i}$$

$$\left[\hat{X}_{h_i} = \frac{1}{m_{h_i}} \sum_{j}^{m_{h_i}} \frac{1}{P_{h_{ij}}} r_{h_{ij}} \frac{1}{1/f_{h_{ij}}} X_{h_{ij}} \right.$$

$$= \frac{1}{m_{h_i}} \sum_{j}^{m_{h_i}} \frac{W_{h_i}}{W_{h_{ij}}} r_{h_{ij}} f_{h_{ij}} X_{h_{ij}}$$

$$\left. = F_{h_i} \sum_{j}^{m_{h_i}} r_{h_{ij}} X_{h_{ij}} \right]$$

여기서,

$\hat{X}$: 속성X를 가진 인구의 선형추정치

$\hat{X}_h$: h지역의 속성X를 가진 인구의 선형추정치

$\hat{X}_{h_i}$: h지역, i층의 속성X를 가진 인구의 선형추정치

L_h : h지역의 층수

$F_{h_i} = \dfrac{W_{h_i}(h\text{지역, } i\text{층에 있어서 전체조사구의 } weight \text{ 합계})}{m_{h_i}(h\text{지역, } i\text{층의 표본조사구수})}$

$P_{h_{ij}}$: h지역 i층의 표본조사구의 추출확률

$r_{h_{ij}}$: h지역 i층의 j표본조사구의 보정률

$f_{h_{ij}}$: h지역 i층의 j표본조사구의 조사구내 추출율의 역수

h : $block$번호

i : 층번호

j : 표본조사구번호

9) 분자, 분모가 Random Variable이고, 보조정보의 활용이 가능할 때 사용하는 방식으로 추정치의 정도를 향상시키는데 목적이 있다.

4. 미국과 일본의 표본설계 비교

미국의 CPS와 일본의 노동력조사의 근본적인 공통점은 표본설계에서 2단계추출법을 사용하고 있으며, 부분표본 교체방식은 4-8-4시스템을 쓰고 있다는 것이다. 두 표본설계를 비교해보자.

가. 표본설계

1) 미국의 모집단은 4개이고 일본은 10개의 그룹으로 되어있다.

2) 층화의 목적은 어떤 통계치의 분산을 최소화시키기 위한 것으로 미국의 경우에는 376개 층으로 되어 있고, 일본은 44개 층으로 구성되어 있다.

3) 이들 각 조사의 근본적인 차이는 교체단위와 교체방식에서 나타난다.

마지막 추출단위(final sampling unit)를 살펴보면, 미국은 평균 4가구로 구성된 집락(USU)이지만 일본은 한 가구이다. 미국의 교체단위는 마지막 추출단위인 집락이지만, 일본은 평균 50가구로 이루어진 조사구를 교체단위로 하고있다. 표본의 교체시 미국은 가장 인접한 USU로 교체하고 있으나, 일본은 계통추출에 의해 표본조사구를 교체한다. 그러므로 미국의 경우가 일본보다 좀 더 유사한 표본으로 교체하여 추출오차를 최소화하고 있음을 알 수 있다.

나. 4-8-4 시스템

1) 4개월동안 조사하고 8개월은 조사에서 제외한 후 다음 4개월동안 다시 조사한 후 표본에서 제거하는 4-8-4시스템의 장점은 조사비용의 감소와 좀 더 정확한 년간변동 추정을 할 수 있다는 것이다. 또한 표본의 순환으로 항목 대부분의 수준과 변화에 대한 추정치의 추출오차를 줄이는데 유효한 복합추정방법의 적용이 가능하며, 약간의 추가비용으로 많은 가구를 가진 USU들의 대표성을 증대시킴으로써 표본오차를 줄일 수 있다.

2) 각 조사의 중복비율을 보면 연간중복은 미국, 일본 모두 50%이지만, 월별중복은 미국이 75%, 일본 50%로 일본의 중복비율이 미국보다 낮았다. 이로 보아 일본의 방법이 미국보다 조사비용이 더 든다고 할 수 있다.

3) Rotation할 때, 미국은 PSU를 고정시키고 USU만 rotation시키며, ED 크기가 작으면 옆의 ED로 옮겨 조사한다. 일본은 조사구와 가구 모두 rotation시킨다.

5. 우리나라 현실에 맞는 Rotation Sampling 연구

지금까지는 외국 사례를 검토하였는데, 그것을 토대로 우리의 조사환경에 맞는 몇 가지 부분교체표본 조사방식이 제안 되었다.

가. Rotation방법을 도입할 경우 유의 사항

1) 부분교체표본추출을 실시할 경우 추가되는 비용(사용방법에 따라 달라질 수 있음)을 고려한다.

2) 추가 조사원의 확보

3) 6대 도시와 9개 지역별로 유사한 성질을 갖고 있는 PSU를 구성하여 지역별 통계를 얻을 수 있다. 기존의 표본설계에서는 특정지역의 표본조사구가 적어 대표성에 문제가 생기고, 추정치의 분산이 증가하는 문제점이 있었다. 표본조사구수를 늘리면 이를 해결할 수 있지만 비용의 증가와 표본관리의 어려움이 있으므로 현재의 표본에서 가구수를 유지하면서 조사구만을 늘리는 방안을 검토해야 한다. 현재의 구역별 가구수(10가구)를 줄이는 방법도 검토될 수 있다.

4) 충분한 PSU내 조사구수의 확보 : 부분교체표본 시스템을 사용할 경우 PSU당 조사구수가 적으면 최소 5년간(60개월)의 부분교체 표본조사에 필요한 새로운 표본의 제공이 불가능할 수 있다. 이러한 경우의 대비책이 필요하다.

위의 전제조건하에서 우리의 실정에 알맞으며 이론상으로 실시가 가능한 부분교체표본 방식은 **2-10-2, 3-9-3, 4-8-4, 6-6-6** 등의 4가지 방식을 들 수 있다.

나. 표본설계 예시

1) 2-10-2 system

이 방식은 4개의 부표본을 가진다. 매달 4개의 부표본이 전부 조사되며, 중복비율(부분교체비율)은 매월 50%, 매년 50%이다. 여기에서는 2달간만이 조사되고 10달 후에야 재조사를 해야 하므로 조사에 어려움이 많으며, 부분교체비율 역시 낮은 편이다.

<그림 3.3> 2-10-2 Rotation Chart

조사구 / 월	A1	A2	A3	A4	A5	A6	A7
	1 2 3 4	3 4 1 2	1 2 3 4	3 4 1 2	1 2 3 4	3 4 1 2	1 2 3 4
1	1 2			3 4			
2	2 3			4 1			
3	3 4			1 2			
4	4	3		2	1		
5		3 4			1 2		
6		4 1			2 3		
7		1 2			3 4		
8		2	1		4	3	
9			1 2			3 4	
10			2 3			4 1	
11			3 4			1 2	
12			4	3		2	1

2) 3-9-3 system

3-9-3방식은 6개의 부표본이 필요하며, 중복비율은 매월 66.6%, 매년 50%이다. 2-10-2방식보다 교체비율도 높고, 부표본의 수도 많아 표본의 대표성을 나타내는데 유리한 조건을 가지고 있다. 하지만, 9개월이 지난 후에야 재조사를 실시해야 하므로 조사에 어려움이 따른다.

<그림 3.4> 3-9-3 Rotation Chart

조사구 / 월	A1	A2	A3	A4
	1 2 3 4 5 6	1 2 3 4 5 6	4 5 6 1 2 3	4 5 6 1 2 3
1	1 2 3		4 5 6	
2	2 3 4		5 6 1	
3	3 4 5		6 1 2	
4	4 5 6		1 2 3	
5	5 6	1	2 3	4
6	6	1 2	3	4 5
7		1 2 3		4 5 6
8		2 3 4		5 6 1
9		3 4 5		6 1 2
10		4 5 6		1 2 3
11		5 6	4	2 3
12		6	4 5	3

3) 4-8-4 system

현재 미국과 일본에서 쓰고 있는 교체방식으로 8개의 부표본이 필요하며, 중복비율은 비율은 매월 75%, 매년 50%이다. 이는 미국, 일본 사례에서 충분히

검토되었다.

<그림 3.5> 4-8-4 Rotation Chart

조사구 / 월	A1	A2	A3	A4
	1 2 3 4 5 6 7 8	1 2 3 4 5 6 7 8	1 2 3 4 5 6 7 8	1 2 3 4 5 6 7 8
1	1 2 3 4	5 6 7 8		
2	2 3 4 5	6 7 8	1	
3	3 4 5 6	7 8	1 2	
4	4 5 6 7	8	1 2 3	
5	5 6 7 8		1 2 3 4	
6	6 7 8	1	2 3 4 5	
7	7 8	1 2	3 4 5 6	
8	8	1 2 3	4 5 6 7	
9		1 2 3 4	5 6 7 8	
10		2 3 4 5	6 7 8	1
11		3 4 5 6	7 8	1 2
12		4 5 6 7	8	1 2 3

4) 6-6-6 system

6-6-6방식은 12개의 부표본을 필요로 하며, 4가지 방법 중 연속 조사기간이 6개월로 가장 길며, 쉬는 기간은 6개월로 가장 짧다. 중복비율은 매월 83.3%, 매년 50%로 가장 높다.

< 그림 3.6 > 6-6-6 Rotation Chart

조사구 / 월	A1	A2	A3
	1 2 3 4 5 6 7 8 9 10 11 12	7 8 9 10 11 12 1 2 3 4 5 6	1 2 3 4 5 6 7 8 9 10 11 12
1	1 2 3 4 5 6	7 8 9 10 11 12	
2	2 3 4 5 6 7	8 9 10 11 12 1	
3	3 4 5 6 7 8	9 10 11 12 1 2	
4	4 5 6 7 8 9	10 11 12 1 2 3	
5	5 6 7 8 9 10	11 12 1 2 3 4	
6	6 7 8 9 10 11	12 1 2 3 4 5	
7	7 8 9 10 11 12	1 2 3 4 5 6	
8	8 9 10 11 12	7 2 3 4 5 6	1
9	9 10 11 12	7 8 3 4 5 6	1 2
10	10 11 12	7 8 9 4 5 6	1 2 3
11	11 12	7 8 9 10 5 6	1 2 3 4
12	12	7 8 9 10 11 6	1 2 3 4 5

앞에서 제시한 4가지 방법의 공통점은 전년대비의 정도를 높이기 위해서 1년이란 일정기간만을 주기로 하여 동일 조사구를 조사대상으로 하였다. 또한 표본조사구들의 월간 상관관계가 조사의 정도(precision)에 영향을 미침은 너무도 당연한 일이므로 각 부표본내의 조사구는 가능한 동질성을 유지하였으며, 표본간의 상관관계는 표본관리 차원에서 항상 보완·관리된다는 가정을 주었다.

위의 방법 중에서는 6-6-6방법으로 부분교체표본을 설계하는 것이 적절하다고 보여지지만, 이것은 이론상의 결론일 뿐 6-6-6방법이 우리 실정에 실제로 적합한지에 대해서는 확정적으로 말할 수 없다. 왜냐하면 기존표본은 고정표본으로 설계되어 표본간의 상관관계가 비교적 높기 때문에, 6-6-6 Rotation Sampling 방법으로 모의실행을 했을 때의 추정치들과 고정표본으로 추정했을 때의 수치와 별 차이가 없었다. 또한 부분교체표본을 실시했을 경우 조사시에 나타날 문제점과 제반 사항에 대하여 충분히 검토되어야 한다.

다. 과거의 표본연동교체

과거 1983년부터 약 3년간 가구표본조사시 부분교체표본을 실시한 적이 있었지만, 참고자료가 많지 않은 탓에 깊은 검토 및 분석을 하지 못했다. 보관되어 있던 자료를 가지고 나름대로 정리해 보았다.

1) 필요성 : 고정표본 사용의 문제점 배제하기 위해서
   가) 응답가구의 응답부담 경감
   나) 응답가구의 형식적 답변 가능성 방지
   다) 실사상의 조사원 의사반영 가능성 방지

2) 고려사항
   가) 전월, 전분기, 전년 사용표본과 부분적인 중복(비교상의 정도제고를 위해) 고려
   나) 조사원의 업무량 증가 고려
   다) 동일표본의 최소한 6개월 이상 사용한다.(인구동태표본조사를 고려)

3) 현재 표본
   가) 조사구수 : 546조사구(시부 453, 군부 93)

나) 조사구별 조사구역 수

· 시부 - 경제활동인구조사 및 인구동태표본조사 : 2개 구역
도시기계조사 : 2개 구역 중 1개 구역

· 군부 - 경제활동인구조사 및 인구동태표본조사 : 6개 구역

4) 교체방법

가) 경제활동인구조사 및 인구동태표본조사

· 매월 1/6 교체하며, 1/3조사구에서는 1/2구역 교체

· 중복비율 : 전월과 5/6 중복 , 전분기와 1/2 중복, 전년과 1/2 중복

· 동일가구에 대한 조사기간 : 6-6-6 Rotation Sampling System 사용
(6개월 조사, 6개월 조사중지, 6개월 조사 후에 조사 중지)

나) 도시가계조사

· 매분기별 1/6 교체하며, 1/6조사구에서는 조사구역 교체

· 중복비율 : 전분기와 5/6 중복, 전년과 1/3 중복

· 동일가구에 대한 조사기간 : 1년 6개월 조사후 조사중지

다) 실시기간 : '83. 10월 ~ '87. 10월

5) 교체방법 수정

가) 필요성

① 조사의 정확성 제고 : 조사자료의 신뢰도를 가지는 단계에서 표본이 교체되므로 새로운 가구원 설득과정(1~3개월)에서 오는 조사착오를 감소시킬 방책이 필요함

② 조사상 애로부문 조정필요 : 현재의 표본연동교체는 공백기간(6개월) 때문에 불응가구 증대, 대상가구 설득에 애로가 있음.

나) 기본방향

① 도시가계조사 : 현행유지

② 경제활동인구조사 및 인구동태표본조사 : 교체비율과 교체방법 조정

· 교체비율을 1/6 → 1/12로 조정

다) 개선내용

	현 행	개 선
교 체 비 율	1/6	1/12
월별 교체 조사구수 (총 549 조사구)	182 조사구 (2,750 가구)	90~91 조사구 (1,380 가구)
전년동기와 중복비율	1/2	1/2
구역당 교체주기	6개월 조사, 6개월 중지, 다시 6개월 조사후 조사중지	1그룹: 12개월 조사, 6개월 중지, 12개월 조사 2그룹: 12개월 조사, 6개월 중지, 6개월 조사 3그룹: 6개월 조사, 6개월 중지, 6개월 조사

· 장점 : 대상가구 설득용이, 공백기간의 문제 해결, 조사자료의 신뢰성 제고

· 실시시기 : '84. 8월부터 실시, 표본설계방식은 부록 1에 수록

라) 중단이유 : 시계열 유지의 어려움으로 중단

마) 과거 Rotation Sampling System 검토

존재하는 자료만을 중심으로 과거의 rotation sampling을 다음과 같이 검토해 봤다.

① 우선 표본설계시 부표본 그룹(sub-group)의 idea가 미흡한 듯 하며, rotation sampling의 기본 가정인 모든 부표본의 사용 원칙 미비로 시계열 유지에 문제를 나타낸 듯 하다.

② 표본교체시 규칙적인 교체보다 불규칙적이며, 6-6-6 system을 유지하기 위해 유의적인 표본대체를 볼 수 있었다.

③ 추출된 표본의 대표성의 충분한 고려는 다목적 표본설계로 인해 미흡한 듯 하다.

라. 결 론

위에서 살펴본 미국과 일본의 부분교체표본의 경우, 오랜기간의 연구 끝에 '교체하는 마지막 조사단위의 정의', '부표본의 분류기준', '추정공식', '교체표본의 중복비율',… 등에서 각국의 조사환경에 맞는 표본설계를 했음을 알 수 있었다. 우리나라는 다목적 표본설계를 기본 전제조건으로 하기 때문에, 각 조사의 특성에 따

라 설계하는 부분교체표본을 적용하려면 상당기간의 심층적이며 지속적인 연구가 필요하다. 앞에서 제시한 부분교체 표본방법들을 토대로 과거 우리 청에서 실시했던 부분교체표본의 방법 및 내용 등을 깊이있게 연구하여 우리나라의 조사환경과 실정에 맞는 가구표본설계 방법이 정립되었으면 한다.

Reference

1. 일본의 노동력조사(표본설계의 해설, 1984), 총리부 조사국 통계과

2. The Current Population Survey : Design and Methodology(1985) Bureau of the Census, Technical Paper 40

3. Eckler, Albert Ross(1955), *"Rotation Sampling"*, Annal of Mathematical Statistics 26, 664-685

4. Leslie Kish, *"Survey Sampling"*, John Wiley & Sons, Inc., New York · London · Sydney

제4장. Small Area Estimation의 사례 소개 및 연구

현재 **소지역 통계추정(Small Area Estimation)**은 한국뿐만이 아니라 세계 각국에서 관심의 대상이 되고 있다. 왜냐하면 소지역 즉, 행정적으로 규모가 작거나 조사가 어려운 지역(빈민지역 등)통계의 필요성이 국가뿐만 아니라 개인차원에서도 커졌다. 이런 요구에 따라 소지역까지 통계를 산출하려고 노력하고 있지만 무조건 표본규모만을 증가시키는 것이 조사환경 악화와 같은 문제들을 해결하려는 데에는 한계가 있다. 이렇게 판단한 많은 국가들은 적은 표본만으로도 소지역을 추정할 수 있는 기법(Small Area Estimation Method)에 대해 연구하고 있으며, 몇몇 나라들은 이 기법으로 이미 통계를 작성하고 있다. 우리나라 역시 지방자치제 실시 이후로 지역통계 생산이 요구되고 있으며, 우선 Small Area Estimation기법의 소개와 필요성을 시작으로 우리나라의 지역통계 생산의 가능성을 검토해 보기로 하자.

1. Small Area Estimation에 대해서

표본조사 자료를 이용하여 특정집단(domain)의 총계나 평균을 추정하는 문제는 조사연구자들에게 일상적인 것이다. Area(domain)이 크고 표본크기가 충분할 때 조사자료 자체에 의한 추정치(direct estimator)나 조사설계법에 따른 추정치(design-based estimator)는 상당히 정도(precision)가 높다. 그러나, Area(domain)이 작고 표본의 크기가 적을 때 조사자료 자체에 의한 추정량들은 변동량(variability)이 커서 추정치로서 이용하기엔 부적당하다. 이러한 문제의 해결을 위해 small area estimation기법의 연구가 시작되었다.

여기서 "small area estimator"란 용어는 주로 지리적으로 구분된 특정지역들에 대한 추정치들을 생산하는데서 유래된 것으로서, "small"이란 특정지역(area, or domain)의 모집단이나 지역(area, domain)자체의 크기가 작다는 것이 아니라, 그 지역에서 추출된 표본의 크기가 작음을 의미한다. 따라서 관심의 초점은 크기가 적은 표본으로부터 추정된 추정치의 분산을 어떻게 줄이느냐 하는 것이다. "area"란

임의의 특정 소지역만을 의미하는 것이 아니라, 임의의 지역과 domain(arbitrary area and arbitrary domain)에 모두 적용되는 용어이다. (예: 특정지역의 추정치, 4인 가구소득의 중앙값 등).

소지역(small area, small domain)에 대한 추정은 Royal(1976)이후로 급속히 발전된 model based estimator나 synthetic estimator의 특수한 방법 혹은 대체 방법이라고 할 수 있다. 이러한 소지역 추정방법의 특징은, **추정치를 산출하려는 소지역과 유사한 다른 소지역(area or domain)으로부터 정보를 빌려서 특정 모수들에 대한 추정치의 정도를 높이는 것은 물론 소지역으로부터 보조정보와 결합하여 small area(domain)의 추정치의 정도를 높이려는데 있다.**

비록 대규모 표본조사가 전국적 혹은 각 도별 통계량을 산출하기 위해서 설계되고 있으며, 필요시에는 이런 자료들을 이용하여 각 domain별(예: 3인가구, 4인가구, 5인가구 소득 등) 혹은 소지역별(예 : 중소도시별) 통계량을 산출하여 '표:(Cross-classified table)'를 만들고 있다. 그러나 이렇게 산출된 Table내의 통계량은 조사 설계 당시 혹은 추정단계에서 일정한 정도(precision)를 감안함이 없이 추정하게 되는 경우가 많기 때문에 추정의 정도(precision)가 낮아 소지역 통계로 사용할 수 없는 경우가 많다.

이를 보완하기 위해서 Canada에서는 Economic Regions, Unemployment Regions, Health Planning Regions등 특별한 층(strata) 혹은 집락(Clusters)을 설정하여, 대규모 표본조사시 소지역(small area or domain)에 대한 통계량을 추정하기 위한 작업을 병행한다. 그러나 소지역의 통계량 산출을 위해 기획과 설계당시에 많은 주의를 기울였다 하더라도 특정지역에 대한 small area(domain) 통계량의 필요성은 항상 존재하게 된다. 결국 수요자들 모두를 만족시킬 수 있도록 많은 특별층(strata) 혹은 집락(cluster)을 구성할 수 없는 것이 현실이므로, 표본설계당시 감안되지 않은 domain이나 소지역에 대한 통계량을 추정하기 위해서는 기술적인 문제를 고려하지 않을 수 없었다. 그래서 최소한의 자료를 가지고 제 문제점들을 감안하여 우리에게 필요한 통계를 작성하기 위해 소지역 추정방법에 대한 연구의 필요성이 부각됐다.

2. 추정방법(Estimation Methods)

시대적으로 요구되는 small area estimation의 방법론은 크게 2가지로 나눌 수 있다. 그 첫번째는 관심있는 area(domain)에서 추출된 표본의 관측치들을 이용하여 추정량을 구하는 직접추정치(Direct estimator)와 관심있는 area(domain)에서 추출된 표본의 관측치들과 이와 유사한 다른 area(domain)의 관측치를 이용하여 추정하는 간접추정치(Indirect estimator)를 들 수 있다. 이때 보조변수(auxiliary variable)의 사용여부가 Direct estimator와 Indirect estimator를 구분하는 잣대는 아니다.

가. Direct Estimator

Direct Small Area estimators는 오직 small area에서 얻은 조사자료를 이용하여 추정되는데, 보조변수로서 인구센서스 자료나 행정자료(administrative records)를 이용하는 경우도 있다.

1) 총계(Total)추정을 위한 가장 단순한 추정치로 (1)의 식을 사용한다.

$$\hat{Y}_a = \sum_{i=s_a} w_i y_i \qquad \text{---------- (1)}$$

[s_a : small area인 a지역에서 추출된 표본
w_i : i번째 관측치의 weight]

$\hat{Y}_a$는 불편추정치이지만 small area내의 표본크기에 따라 변동량(variability)이 심할 수 있다는 문제점이 있다.

2) small area인 a의 크기(N_a)가 알려진 경우, 사후층화 추정량[10](post stratified estimator)을 이용하며, 이 추정치의 수식은 아래와 같다.

$$\hat{Y}_{post,a} = N_a \frac{\sum_{i=s_a} w_i y_i}{\sum_{i=s_a} w_i} = N_a \frac{\hat{Y}_a}{\hat{N}_a} = N_a \bar{y}_a \qquad \text{---------- (2)}$$

10) 각 부분집단의 크기를 알고 있으나 부분집단별 list가 없는 경우에 사용된다. 우선 임의의 표본을 뽑은 후, 추정단계에서 층별 자료를 이용하여 추정치의 정도를 향상시키는 방법이다.

이 추정치는 식(1)보다 안정된 추정량을 제공하나, 조사설계가 복잡한 경우 비추정 편의(ratio estimation bias)가 발생할 수 있다는 단점을 가지고 있다. 만약 층화추출이 가능하며 어떤 층의 소지역 a의 크기($N_{h,a}$)가 알려진 경우에는 다음과 같은 사후층화 추정량(post stratified estimator)으로 계산할 수 있다. 여기서 층은 조사설계당시의 층이 아니라 사후층화된 것을 의미한다.

$$\hat{Y}_{st,post,a} = \sum_h \left\{ N_{h,a} \frac{\sum_{i=s_{h,a}} w_i y_i}{\sum_{i=s_{h,a}} w_i} \right\} = \sum_h N_{h,a} \frac{\hat{Y}_{h,a}}{\hat{N}_{h,a}} = \sum_h N_{h,a} \bar{y}_{h,a} \quad \text{———(3)}$$

[$N_{h,a}$: h번째 층내의 a 지역의 모집단크기]

단순추정치외에 비추정치의 경우는 사후추정량과 유사하나 모집단의 크기 N_a와 $N_{h,a}$대신에 보조변수를 이용하는 것이 다르다. 여기서 보조변수 X 는 추정하려는 변수 Y 와 상관관계가 클수록 정도 높은 추정량을 얻을 수 있으며, 비추정량은 (4)와 (4-1)로 정의된다.

$$\hat{Y}_{r,a} = \frac{\hat{Y}_a}{\hat{X}_a} X_a \quad \text{———(4)}$$

$$\hat{Y}_{st,r,a} = \sum_h \frac{\hat{Y}_{h,a}}{\hat{X}_{h,a}} X_{h,a} \quad \text{———(4-1)} .$$

[X_a : 보조변수 X의 small area총계
$X_{h,a}$: h번째 층의 a지역 총계]

이외에도 보조변수로 small area의 모집단과 표본과의 차이를 보정하기 위해 회귀선을 이용하는 추정치도 있다.

$$\hat{Y}_{reg,a} = \hat{Y}_a + \hat{\beta}_a (X_a - \hat{X}_a) \quad \text{———(5)}$$

$$\hat{\beta}_a = \sum_{i=s_a} \nu^{-1} w_i y_i x_i' \left\{ \sum_{i=s_a} \nu^{-1} w_i y_i \right\}^{-1} \quad \text{———(5-1)}$$

여기서 $\hat{Y}_a$ 는 (1), (2), (3), (4), (4-1)식 중의 어느 추정량을 사용하여도 무방하며, $\hat{X}_a$ 는 $\hat{Y}_a$ 와 같은 방법으로 추정된다. (5-1)식은 회귀식에서 일반

적으로 얻어지는 추정치이며, 이를 일반화된 회귀추정량(generalized regression estimator)이라고도 한다.

1) Direct estimator의 보정

이는 domain밖에서 얻어진 조사자료를 이용하여 direct estimator를 얻기 위한 모형설정에 따르는 편의(bias)를 종합적으로 보정한(synthetic adjustment) 추정치를 말한다.

$$\hat{Y}_{M,reg,a} = \hat{Y}_a + \hat{\beta}(X_a - \hat{X}_a) \quad \text{——————————(6)}$$

이 수식은 회귀선을 이용해서 일반 회귀선 추정식에 의해 다음 같이 추정된다.

$$\hat{\beta} = \sum_{i=s} \nu^{-1} w_i y_i x_i' \left\{ \sum_{i=s} \nu^{-1} w_i y_i x_i' \right\}^{-1} \quad \text{————(6-1)}$$

(6-1)의 $\hat{\beta}$는 (5-1)의 $\hat{\beta}_a$보다 안정되어 있다. 다시 말해서 a지역에 대한 $\hat{\beta}$의 분산이 $\hat{\beta}_a$보다 적은 경우가 많다. 흔히 악속된 β의 추정량을 이용하는데, 이는 $\lambda_a \hat{\beta}_a + (1-\lambda_a)\hat{\beta}$ 형태를 취하며 λ_a는 설정된 모형에 따라서 적절히 선택된다.

2) Indirect estimator

추정하고자 하는 domain과 유사한 특성을 지닌 다른 domain의 정보를 이용하여 추정하는 것을 말하며, 대표적인 방법으로 **Synthetic estimator Method**가 있다.

Synthetic estimator는 Gonzalez(1973)와 Ericsen(1974)에 의해 시작되었으며, 모집단을 구성하는 집락들의 특성이 유사할 때, 표본조사를 통해 얻어진 불편추정치는 모집단에 대한 불편추정치일뿐 아니라, 각각의 소집락에 대한 불편추정치이기도 하다. 즉, 큰 지역에 속하는 소지역들은 대규모 지역과 같은 특징을 가지고 있을 것이라는 가정 하에 대규모 지역의 추정치가 소지역에 대한 추정치를 유도하는데 사용되는데, 이렇게 얻어진 추정치를 synthetic estimator라고 정의하였다. 결국 Synthetic estimator를 구할 때 small area

의 a지역은 a지역을 포함하는 지역의 다른 small area a'지역과 유사하다는 가정 하에서 시작된 방법이므로 가정이 위배되면 bias가 크다.

모집단이 g개의 domain으로 분할되었고 모집단의 총계 $Y_{.g}$의 추정치는 $\hat{Y}_{.g}$(direct estimator)라 하자. 이때 small domain을 a라 하고, a는 g개의 domain 모두에 걸쳐있다고 하면, g domain의 총계는 $Y_{.g}=\sum_{a} Y_{ag}$가 된다. Y_{ag}는 (a,g)cell의 총계를 나타내며, (a,g)cell의 보조정보인 $X_{a,g}$가 이용가능할 때 synthetic estimator는 다음과 같이 주어진다.

$$\hat{Y}_{a,syn}=\sum_{a}\frac{X_{a,g}}{X_{.g}}\ \hat{Y}_{.g}$$

이때 direct estimator인 $\hat{Y}_{.g}$(ratio estimator)는 다음과 같이 정의된다.

$$\hat{Y}_{.g}=\frac{y_{.g}}{x_{.g}}X_{.g} \qquad \text{——————————(7)}$$

[$y_{.g}$: g domain에서 추출된 y의 표본총계
$x_{.g}$: g domain에서 추출된 x의 표본총계]

3) Composite estimator

Direct estimator에서 변동량의 문제와 synthetic estimator의 모형설정에 따른 bias 문제를 해결하기 위한 것으로 사용되는 composite estimator는 다음과 같이 정의된다.

$$\hat{Y}_{a}^{c}=w_i\ \hat{Y}_{1,a}+(1-w_i)\ \hat{Y}_{2,a}$$

[$\hat{Y}_{1,a}$: direct estimator
$\hat{Y}_{2,a}$: indirect estimator]

Composite estimator의 요점은 w_i의 결정이며, w_i의 적정(optimal)한 값은 small area추정치의 **MSE**[11]를 최소화하는 것이다. 여기서 random area

11) 평균평방오차(mean squared error)를 의미하며, $E(\hat{\theta}-\theta)^2$로 정의된다. MSE가 작을수록 좋은 추정량이다.

specific effect를 고려한 모형에 따른 composite estimator의 예를 들어보자. 우선 $x_a=(x_{a1},\ldots,x_{ap})'$가 이용가능하고 y_a와 x_a와의 상관관계가 있다고 가정하면 다음과 같은 모형이 된다.

$$Y_a = x_a\beta + \nu_a z_a + e_a$$

[a : 1, 2 , ... A
z_a : known positive constants
β : 회귀계수
$\nu_a \sim (0,\ \sigma_\nu^2)$ or $\nu_a \sim N(0,\ \sigma_\nu^2)$]

이때, 이 모형은 general mixed linear model의 특수한 형태이며, design induced random variable e_a와 model based random variable ν_a를 포함하게 된다. 1986년 Freedman과 Navidi가 이 모형에 대해서 사후조사(post enumeration survey)를 실시한 결과 추정치의 bias가 크게 나타났다. 두 사람은 표본의 크기가 작을 때 bias가 커지는 것을 이 모형의 문제점으로 제시를 하였다. 이 모형을 응용한 사람들은 Ericsen & Kadane(1985,1987), Cressie(1992), Fay & Herriot(1979)들이 있다.

다른 모형으로는 element specific auxiliary data와 $X_{aj}=(x_{aj1}, x_{aj2}, \ldots, x_{ajp})^p$이 이용가능한 경우에 사용하는 nested error regression모형을 들 수 있다. 이에 따른 추정치는 아래와 같으며, 응용한 사람들은 Battese, Harter, Fuller(1988)들이 있다.

$$Y_{aj} = X_{aj}' + \nu_a + e_{aj}$$

[$j=1,2,\cdots,N_a$, N_a : a번째 $area$의 요소수
$a=1,2,\cdots,A$
$e_{aj} = \tilde{e}_{aj}\ k_{aj}$, $\tilde{e}_{aj} \sim N(0,\sigma^2)$, k_{aj} : 알려진 상수]

3. 미국의 이용사례

< 표 4.1 > Small Estimation Area Method 이용사례

시 행 처	추 정 량	추정하려는 변수	Area(Domain)	비 고
Bureau of the Census	indirect regression	Census후의 인구	Counties	매 년
Bureau of the Census	indirect composite	4인 가구의 소득의 중앙값	States	매 년
Bureau of Economic Analysis	indirect regression (Ratio)	개인소득, 연간소득, 연간 생산량	States Counties	매 년 분기별
Bureau of Labor Statistics	indirect regression	고용률, 실업률	States	매 달
National Agricultural Statistics Service	indirect regression	목화, 쌀, 콩의 재배면적	Counties	매 년
National Agricultural Statistics Service	indirect composite	가축수,과일생산량, 과일재배면적	Counties	매 년
National Center for Health Statistics	indirect synthetic	유아와 산모의 건강상태	States	정기적
National Center for Health Statistics	indirect composite	지체부자유자, 치과의사 방문수	States	정기적

가. 사례연구

여기서 미국의 small area estimator method의 실질적 예를 검토한 후, 우리나라의 1994년 경제활동인구조사 자료에 이 방법을 적용하여 소지역 통계의 생산 가능성을 검토·분석해 보았다.

사 례 1 **작물의 재배면적 조사**

작물의 재배면적 조사는 미국 농업부에서 정책입안에 중요한 기본자료가 되고

있으며, 대규모 작물재배단지, 각 주별 작물의 경작면적, 미국 전체의 작물 경작면적을 추정할 수 있도록 표본설계 되었다. 그러나, County 같은 소지역에 대한 경작면적의 추정은 이들 지역의 표본크기가 작기 때문에 시도되지 않았다고 한다.(Battese, Harter, Fuller)

1978년 이전의 조사에서는 각 County별 옥수수와 콩의 재배면적의 추정은 자료의 부족으로 인하여 시도되지 않았으나, 위성사진의 활용으로 인해 소지역(Small Area) 추정치에 대한 연구가 시작되면서, Rashid and Battese, Harter, and Fuller(1987), Hanuschak(1979), Hung and Fuller(1987), Fuller and Battese(1981,1987,1988), Datta and Ghosh(1991), Nandran (1995) 등에 의해서 각 County별 경작면적 추정을 위한 연구가 활발해지기 시작하였다.

가) 조사의 목적 : Iowa주의 12개 County에서 옥수수와 콩의 평균 경작면적을 조사하는 것이다.

나) 자료의 구성

12개 Iowa County를 대상으로 미 농업부에서 1978년 6월 옥수수와 콩의 경작면적을 표본조사 하였다. 작물의 성장시기에 위성(LANDSAT)사진을 촬영하여 특정작물 재배면적을 위성사진의 분석으로 조사하였다. 각 지역별 조사자료의 구성은 뒤에 나오는 <표 4.2>에서 보는 바와 같이 Cerro Gordo County의 경우 545개의 Segments(PSU)로 구성되어져 있으며, Hamilton County의 경우 566개의 segments로 구성되어져 있다. Hardin County의 경우는 556개의 segments로 구성되어 있다. (참고로 1개의 Segment는 250hectares이다)

표본조사는 Cerro Gordo County에서 545개의 Segments중 1개만을 조사하였고, Hamilton County 역시 1개의 Segment만을 조사하였다. 12번째 County인 Hardin에서는 6개의 Segments를 조사하였다.

다) 위성사진의 분석

1978년 8월과 9월에 LANDSAT을 통하여 얻은 위성사진을 12개 County

에서 추출된 Segment에 대해 옥수수와 콩의 재배지역을 Pixel(0.45 hectares)로 분류하였다. 즉 Cerro Gordo County에서는 545개의 Segments 중에서 1개의 Segment를 추출하여 조사한 결과 옥수수의 재배면적은 165.76hectares였고, 위성사진 분석결과 위 Segment에서의 옥수수 재배면적은 374 Pixels(374*0.45=168.3hectares)이었다. 12번째 County인 Hardin에서는 556개의 Segments중에서 6개의 Segments를 표본추출하여 조사한 결과 88.59(99), 165.35(159.75), 104.00(117.45), 88.63(84.15), 153.70(157.5)이며, 2번째 segments 자료는 버렸다.

참고로 조사를 통한 옥수수의 재배면적과 위성사진분석자료를 통한 옥수수의 재배면적과의 상관관계는 0.80이었으며, 각 County별 특성을 고려한 random effect가 있다고 가정하고 작물의 경작면적의 추정시 다음과 같은 nested error regression 모형을 생각하게 된다.

라) 각 County별 평균추정

○ Components-of-variance 모형

모형: $y_{aj} = \beta_0 + \beta_1 x_{1aj} + \beta_2 x_{2aj} + u_{aj}$ [$u_{aj} = \nu_a + e_{aj}$]

$$E(u_{aj}\ u_{pq}) = \begin{cases} \sigma_\nu^2 + \sigma_e^2 & a = p,\quad j = q \\ \sigma_\nu^2 & a = p,\quad j \neq q \end{cases}$$

$Y = X\beta + Z\nu + e$ 형태로 mixed linear model의 특수형태인 nested error regression 모형을 사용하였다. 이 모형은 County 내의 각 농가에서 보고한 경작면적간의 상관관계를 감안할 수 있다.

일반적으로 super population approach에서는 a번째 County의 segment당 작물의 평균 경작면적을 추정하기 위해

$$N_a^{-1}\left\{\sum_{j=1}^{n_a} y_{aj} + \sum_{j=n_a}^{N_a} (x_{aj}\tilde{\beta} + \tilde{\nu}_a)\right\}$$

을 이용한다. 이 때의 모형은 다음과 같다.

$$\bar{y}_{i.} = \beta_0 + \beta_1 \bar{x}_{1i.} + \beta_2 \bar{x}_{2i.} + v_i + \bar{e}_{i.}$$

[$\bar{x}_{1i}$, $\bar{x}_{2i}$: 각 *county*내의 표본 *segment*에 대응하는 위성사진 자료]

그러나, Small area 추정에서는 추정하려는 변수가 속한 area의 표본추출비가 작기 때문에 다음과 같은 추정식을 이용한다.

$$\bar{x}_{a(p)}\hat{\beta} + (\bar{y}_{a.} - \bar{x}_{a.}\hat{\beta})\delta_a$$

$$\left[\begin{array}{l} \bar{x}_{a(p)} = N_a^{-1} \sum_{j=1}^{N_a} x_{1aj} \\ \delta_a = \dfrac{\sigma_v^2}{m_a} \\ m_a = (\sigma_v^2 + n_a^{-1}\sigma_e^2) \end{array} \right]$$

이 모형에서 $\hat{\beta} = \beta$일 때 위 추정치는 **BLUE**[12]가 된다. 추정결과는 <표 4.2>와 같으며, <표 4.2>에서 보는 바와 같이 Cerro Gordo County의 경우 545개의 Segments의 옥수수 평균 재배면적은 122.2hectares이고, Hardin County의 경우 556개 segments의 옥수수 평균 재배면적은 143.6hectares이었다.

< 표 4.2 > 작물의 재배면적 조사 DATA

(12 Counties in North-Central Iowa: Farm Interviews and LANDSAT)

county	sample	popul-ation	corn	corn soybean	soybean BHF(se)	corn BHF(se)	soybean R-est	R-est
Cerro Gordo	1	545	65.76	8.09	122.2(9.6)	77.8(12.0)	122.5	66.8
Hamilton	1	566	96.32	106.03	126.3(9.5)	94.8(11.8)	126.0	104.9
Worth	1	394	76.08	103.60	106.2(9.3)	86.9(11.5)	93.5	85.2
Humboldt	2	424	185.35 116.43	6.47 63.82	108.0(8.1)	79.7(9.7)	106.8	65.1
Franklin	3	564	162.08 152.04 161.75	43.50 71.43 42.49	145.0(6.5)	65.2(7.6)	149.7	59.6

12) 최량선형불편추정량(best linear unbiased estimator)

Pocahontas	3	570	92.88 149.94 64.75	105.26 76.49 174.34	112.6(6.6)	113.8(7.7)	114.4	116.4
Winnebago	3	402	127.07 133.55 77.70	95.67 76.57 93.48	112.4(6.6)	98.5(7.7)	109.1	101.4
Write	3	567	206.39 108.33 118.17	37.84 131.12 124.44	122.1(6.7)	112.8(7.8)	123.9	111.1
Webster	4	687	99.96 140.43 98.95 131.04	144.15 103.60 88.59 115.58	115.8(5.8)	109.6(6.7)	118.5	108.9
Hancock	5	569	114.12 100.60 127.88 116.90 87.41	99.15 124.56 110.88 109.14 143.66	124.3(5.3)	101.0(6.2)	123.1	104.4
Kossuth	5	965	93.48 121.00 109.91 122.66 104.21	91.05 132.33 143.14 104.13 118.57	106.3(5.2)	119.9(6.1)	104.2	121.7
Hardin	6	556	88.59 88.59 165.35 104.00 88.63 153.70	102.59 29.46 69.28 99.15 143.66 94.49	143.6(5.7)	74.9(6.6)	144.6	79.4

사 례 II

주(state)별 4인가구 소득의 중앙값(Median family incomes by states) 추정

1974년도 소득(income)부터 미국 Census Bureau 에서 census 자료, CPS 자료, BEA(Bureau of Economic Analysis)의 PCI(Per Capita Income)의 추정치들을 이용하여 4인가구에 대한 각 주별 소득의 중앙값을 추정하고 있다. 이때의 추정치는 Model based estimate이며, Schaible, W.Gonzalez, M. 등이 주축이 된 Small area Estimation Committee에서 연구한 것이다.

가) 자료의 구성

CPS자료, Census 자료(매 10년 실시), BEA(Bureau of Economic Analysis)의 PCI(Per Capita Income) 추정치들

나) 접근방법

① 10년마다 실시되는 Census로부터 얻는 주(state)별 가구소득의 중앙값은 추정표본의 크기가 크기 때문에 각 주별로 매 10년마다의 가구소득의 중앙값에 대한 표본오차가 작다.

② CPS자료에 의한 CPS추정치는 가구크기(3인, 4인, 5인 가구 등)별 소득의 중앙값에 의해 산출된 표본추정치로 매년 유용하게 이용되지만, CPS표본의 크기 때문에 변동량이 커서 추정치의 활용은 제한되어 있다.

③ BEA로부터의 PCI의 매년 추정치(PCI 추정치는 각 개인의 평균소득을 추정한 것을 이용한다.

다) 추정방법

1984년 소득분부터 시작된 추정방법은 다음과 같으며, 이 방법에 의해 매년 4인가구 소득의 중앙값이 추정되고 있다.

① CPS를 통해 각 주별로 4인가구 소득의 중앙값에 의해 산출된 표본추정치(direct estimate)인 $\hat{Y}_{s4}$가 추정된다. 3인가구와 5인가구의 소득의 중앙값 역시 같은 방법으로 추정한다.

② 주별로 3인가구와 5인가구 소득의 중앙값에 대한 가중결합(weighted combination) 추정치는 다음과 같다.

$$\hat{Y}_{sc} = 0.75\ \hat{Y}_{s3} + 0.25\ \hat{Y}_{s5}$$

여기서 0.75와 0.25는 표본크기에 비례하여 얻은 값으로서, 3인 가구의 표본수가 5인 가구의 표본수보다 3배 많음을 의미이다.

③ $\hat{Y}_{s4}$과 $\hat{Y}_{sc}$의 회귀추정치 $\hat{Y}_{reg,s4}$ 과 $\hat{Y}_{reg,sc}$을 추정하기 위해 이용되는 보조변수들은 다음과 같다.

♣ 4인가구 소득의 중앙값을 추정하기 위해 회귀모형에 이용되는 보조변수 ♣

· $X_{s41} = 1$ -------> 회귀모형의 상수항

· $X_{s42} = \frac{BEA_{st}}{BEA_{sb}} Y_{cen,s4}$

여기서 BEA_{st}는 CPS로부터 4인가구 소득의 중앙값을 추정($\hat{Y}_{s4}$)한 년도와 같은 년도 t에서의 *BEA PCI*를 나타낸다. BEA_{sb}는 previous census의 4인가구 소득에 대해 기준년도 b의 *BEA PCI*를 나타낸다. 또한, $Y_{cen,s4}$는 직전 census 소득의 기준년도 b에서의 4인가구 소득의 census 중앙값이다. 따라서 X_{s42}는 직전 census 이래로 BEA PCI의 비례적인 증가에 의해서 보정된 census median을 나타내고, $X_{s43} = Y_{cen,s4}$는 직전 census의 소득의 중앙값을 나타낸다.

가중평균 $\hat{Y}_{sc}$에 대한 회귀모형은 위와 유사하게 다음 변수를 이용한다.

$$X_{sc1} = 1$$

$$X_{sc2} = \frac{BEA_{st}}{BEA_{sb}} Y_{cen,c}$$

$$X_{sc3} = Y_{cen,sc}$$

라) 복합추정방법(Composite Estimate Method)

Composite estimate $\hat{Y}_{comp,s4}$는 $\hat{Y}_{s4}$, $\hat{Y}_{sc}$, $\hat{Y}_{reg,s4}$, $\hat{Y}_{reg,sc}$를 이용하여 산출된다. $\hat{Y}_{s4}$와 $\hat{Y}_{reg,s4}$의 결합은 small area estimation에서 흔히 사용되는 방법인데, 여기서는 $\hat{Y}_{sc}$ 과 $\hat{Y}_{reg,sc}$을 모두 이용한다는 점이 다르다. 참고로 X_{s42}와 X_{s42}의 관계, 즉 1979년부터 1989년까지 census 중앙값과 *BEA PCI*의 증가분(percent increase)과의 관계는 다음과 같다. 추정치를 구하기 위한 일반식을 행렬형식으로 표현하면 다음과 같다.

$$\hat{Y}=(\ \hat{Y}_{1,4}\quad \hat{Y}_{1,c}\quad \hat{Y}_{2,4}\quad \hat{Y}_{2,c}\ \cdots\cdots\ \hat{Y}_{51,4}\quad \hat{Y}_{51,c})'$$

다음은 각 주별로 CPS표본으로부터 추정한 중앙값들이다.

$X_{s4}=(X_{s41},\ X_{s42},\ X_{s43})$각 주의 4인 가구소득 중앙값에 대한 3predictor변수들,

$X_{sc}=(X_{sc1},\ X_{sc2},\ X_{sc3})$은 가중평균 변수 $\hat{Y}_{sc}$의 값이다.

$$X=\begin{bmatrix} X_{14} & 0 \\ 0 & X_{1c} \\ X_{24} & 0 \\ 0 & X_{2c} \\ \vdots & \vdots \\ \vdots & \vdots \end{bmatrix}$$

○ small domain 추정모형

$$\hat{Y}=X\beta+b+e$$

[β : 회귀계수,
b : individual true median과 regression predictor의 차이를 나타내는 random effects
e : sampling error]

$$A^*=\sum(b)=\begin{bmatrix} A & 0 & 0 & \cdot & \cdot & \cdot \\ 0 & A & 0 & \cdot & \cdot & \cdot \\ 0 & 0 & A & \cdot & \cdot & \cdot \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

여기서 A는 2 × 2 공분산 행렬이다.

$$D^*=\sum(e)=\begin{bmatrix} D_1 & 0 & 0 & \cdots \\ 0 & D_2 & 0 & \cdots \\ 0 & 0 & D_3 & \cdots \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

$$\hat{\beta}=[X'(D^*+A^*)^{-1}X]^{-1}X'(D^*+A^*)^{-1}\hat{Y}$$

따라서 4인가구 소득 중앙값의 **BLUE**는 $Y=X\beta+b$이고, 그 추정치는 다음과

같다.

$$\hat{Y}_{comp} = X\hat{\beta} + A^*(D^* + A^*)^{-1}(\hat{Y} - X\hat{\beta})$$

여기서 $\hat{Y}_{comp}$은 4인 가구소득의 중앙값과 3인가구와 5인가구의 소득에 대한 중앙값의 가중평균치 $\hat{Y}_{sc}$에 대해 각 Count내에서의 표본추정치와 회귀추정치의 가중평균이 된다. 4인가구 소득의 중앙값을 추정하기 위한 모형에서 b와 e의 분산이 서로 독립이라고 가정하였으나, $V = J\sigma_b + I\sigma_e$를 이용하면 보다 정도가 높은 추정치를 구할 수 있을 것으로 생각된다. 참고로 소득과 상관관계가 높은 세금관계 자료 혹은 또 다른 변수의 이용이 가능하다면 추정치의 정도를 더 높일 수 있다고 생각된다.

4. '94년 경활자료를 이용한 Small Area Estimation 연구

경제활동인구조사는 취업, 실업, 노동력 등과 같은 인구의 경제적 특성을 조사하여 거시경제분석과 인력자원의 개발정책수립에 필요한 기초자료인 노동조합, 고용구조, 가용노동시간 및 인력자원의 활용정도를 측정하여 고용창출, 취업훈련, 소득증진 등을 위한 정부정책 입안 및 평가에 필요한 자료를 제공하고있다.

본격적인 지방자치제의 시작으로 지역의 정책수립을 위한 지역통계자료의 요구가 날로 늘어나고 있는 현실을 반영하여, 1994년 경제활동인구조사 자료중 일부를 가지고 소지역 표본추정 방법에 적용, 지역통계를 생산 가능성을 검토·분석하였다.

여기서는 1994년 C지역의 경제활동인구조사 자료를 이용하여 A시와 B군의 실업자, 취업자, 비경제활동인구를 중점으로 소지역 표본추정방법에 따른 추정치를 CV관점에서 비교를 하였다.

1) 자료의 구성

가) 주민등록인구를 이용한 A, B의 추계 경제활동인구

나) 보조자료로써 A, B보다 행정적으로 큰 지역(C)의 추계 경제활동인구

다) 경제활동인구조사에서

· A, B의 경제활동인구수

· A, B의 취업자수

· A, B의 실업자수

· A, B의 비경제활동 인구수

2) 접근 방법

가) A, B의 실업자와 취업자. 비경제활동인구에 대한 추정치를 direct, synthetic, composite estimation 등 3가지 추정방법을 이용하여 계산한다.

나) 각각의 추정치에 따른 CV를 계산한다.

다) 각각의 추정방법의 안정성을 검정한다

3) 추정방법

가) Direct Estimator

$$\hat{Y}_a = \sum_{i=s_a} w_i y_i$$

$$SE(\hat{Y}_a) = \sqrt{var(\hat{Y}_a)}$$

$$CV(\hat{Y}_a) = \frac{SE(\hat{Y}_a)}{\hat{Y}_a} \times 100$$

[$\hat{Y}_a$ = a지역의 어떤 특성의 추정치

y_i = i 조사구의 어떤 특성을 가진 인구수

$var(\hat{Y}_a)$ = y_i의 표본분산

(자세한 수식은 "다목적 표본설계"의 55Page 참고)

s_a = 표본조사구수 (A지역의 s_a = 19개 조사구수

B지역의 s_a = 5개 조사구수)

w_i = a 지역의 인구추계에 의한 가중치]

나) Synthetic Estimator

$$\hat{Y}_{a,syn} = \sum_a \left[\frac{X_{ag}}{X_{.g}} \right] \hat{Y}_{.g}$$

$$Var(\hat{Y}_{a,syn}) = \sum_{a}\left[\frac{X_{ag}}{X_{.g}}\right]^2 \sigma_{\hat{Y}_{.g}}$$

$$CV(\hat{Y}_{a.syn}) = \frac{SE(\hat{Y}_{a.syn})}{\hat{Y}_{a,syn}} \times 100$$

[$\hat{Y}_{.g}$ = g 지역의 비추정(ratio estimator: 수식 (7) 참조)
X_{ag} = (a,g) 지역의 추계 층수
$X_{.g} = \sum_{a} X_{ag}$, g 지역의 추계 층수
$\sigma_{\hat{Y}_{.g}}$ = $Var(\hat{Y}_{.g})$]

다) Composite Estimator

$$\hat{Y}_{a.com} = \alpha\ \hat{Y}_a + (1-\alpha)\ \hat{Y}_{a,syn}$$

$$Var(\hat{Y}_{a.com})$$
$$= \alpha^2 Var(\hat{Y}_a) + (1-\alpha)^2 Var(\hat{Y}_{a,syn}) + 2\alpha(1-\alpha)Cov(\hat{Y}_a,\ \hat{Y}_{a,syn})$$

$$CV(\hat{Y}_{a.com}) = \frac{SE(\hat{Y}_a)}{\hat{Y}_{a,com}} \times 100$$

α는 $Var(\hat{Y}_{a.com})$를 최소화하는 값으로 정의된다.

$$\alpha = \frac{MSE(\hat{Y}_{a,syn}) - E(\hat{Y}_{a,syn} - Y)(\hat{Y}_a - Y)}{MSE(\hat{y}_{a,syn}) + MSE(\hat{Y}_a) - 2E(\hat{Y}_{a,syn} - Y)(\hat{Y}_a - Y)}$$

다음 도표들은 이 3가지 방법에 의한 계산된 A, B지역의 CV값들이다.

< 표 4.3 > A, B지역의 월별 Direct Estimator의 CV값

지 역	월	실 업 자		취 업 자		비경제활동인구	
		남 자	여 자	남 자	여 자	남 자	여 자
A	1	3.52	5.83	24.90	31.38	6.36	3.88
	2	4.54	4.71	17.49	16.61	8.32	3.34
	3	3.22	5.91	24.65	23.75	6.99	3.70
	4	3.44	5.80	27.53	22.79	7.75	4.02
	5	4.17	6.36	29.71	39.74	8.55	4.46
	6	4.28	6.50	40.88	36.53	8.30	4.66
	7	4.36	6.67	36.37	27.60	8.54	4.63
	8	4.67	6.79	46.16	32.19	8.82	4.63
	9	4.54	6.22	34.99	33.42	8.33	4.31
	10	4.76	6.23	34.80	33.77	8.79	4.15
	11	4.74	5.66	31.62	45.28	8.17	3.88
	12	4.79	5.60	26.87	49.16	7.49	3.79
B	1	3.36	14.91	53.02	*0.00*	7.85	11.87
	2	3.41	16.50	109.74	77.41	7.58	12.50
	3	2.76	14.64	*0.00*	44.93	7.67	13.42
	4	3.15	6.44	77.65	48.67	9.54	8.71
	5	2.43	4.74	*0.00*	49.17	8.81	6.23
	6	2.61	4.87	*0.00*	60.71	8.91	6.39
	7	2.66	5.33	76.83	61.45	7.07	7.17
	8	2.43	4.65	115.14	61.71	8.31	6.00
	9	2.95	4.82	56.03	63.32	9.79	7.30
	10	4.56	7.00	63.33	58.29	11.45	9.69
	11	4.35	5.62	63.95	58.46	11.56	7.79
	12	4.60	9.39	63.87	59.99	10.81	9.65

< 표 4.4 > A, B지역의 월별 Synthetic Estimator의 CV값

지 역	월	실 업 자		취 업 자		비경제활동인구	
		남 자	여 자	남 자	여 자	남 자	여 자
A	1	1.29	2.20	8.65	20.02	2.91	1.59
	2	1.64	1.70	7.46	12.83	3.87	1.40
	3	1.15	1.92	10.22	16.35	3.51	1.71
	4	1.19	1.84	11.36	12.77	4.24	1.96
	5	1.45	2.00	15.39	18.21	4.73	2.21
	6	1.47	2.03	20.03	19.94	4.71	2.35
	7	1.52	2.08	14.55	15.75	4.73	2.33
	8	1.63	2.11	14.52	20.75	4.95	2.37
	9	1.58	1.92	14.23	26.22	4.65	2.21
	10	1.65	1.92	19.19	22.40	4.98	2.14
	11	1.64	1.81	13.07	25.68	4.68	1.95
	12	1.71	2.10	9.21	40.80	3.99	1.65
B	1	0.25	1.16	1.57	*0.00*	0.55	0.80
	2	0.25	1.18	1.49	1.86	0.53	0.89
	3	0.20	1.01	*0.00*	4.88	0.57	0.95
	4	0.24	0.51	1.20	5.75	0.63	0.52
	5	0.18	0.39	*0.00*	8.52	0.58	0.34
	6	0.19	0.39	*0.00*	8.51	0.59	0.35
	7	0.19	0.43	2.31	10.10	0.48	0.39
	8	0.17	0.37	3.62	11.54	0.56	0.33
	9	0.21	0.39	2.74	5.19	0.67	0.39
	10	0.32	0.56	3.46	7.09	0.79	0.53
	11	0.31	0.45	3.17	9.35	0.75	0.42
	12	0.32	0.73	2.58	5.42	0.70	0.57

< 표 4.5 > A, B지역의 월별 Composite Estimator의 CV값

(Alpha = 0.5)

지 역	월	실 업 자		취 업 자		비경제활동인구	
		남 자	여 자	남 자	여 자	남 자	여 자
A	1	3.21	4.17	18.86	34.22	5.68	2.71
	2	3.61	3.65	15.22	19.18	6.62	2.51
	3	2.96	3.79	20.07	25.18	6.31	2.68
	4	2.22	3.67	24.68	22.38	5.69	3.08
	5	2.47	3.86	34.17	35.49	6.12	3.28
	6	2.54	2.92	45.97	33.50	6.01	3.45
	7	2.50	3.85	27.05	22.97	6.12	3.33
	8	2.62	3.79	31.19	26.46	6.25	3.28
	9	2.58	3.55	23.70	26.68	5.98	3.12
	10	2.99	2.20	21.81	27.66	7.25	3.15
	11	3.02	1.90	18.20	34.64	6.91	1.93
	12	3.15	2.54	16.75	38.06	6.33	2.04
B	1	2.83	8.86	24.49	58.32	9.50	7.08
	2	2.98	9.28	27.07	33.77	10.02	7.50
	3	2.53	8.18	30.95	32.43	9.66	8.06
	4	2.04	4.12	30.02	32.90	11.83	4.98
	5	1.77	3.43	53.37	37.34	11.44	4.15
	6	1.87	3.49	69.28	43.30	11.46	4.32
	7	2.14	3.54	36.31	44.01	10.09	4.31
	8	2.07	3.18	46.23	45.95	10.87	3.88
	9	2.26	3.23	29.90	37.86	11.74	4.27
	10	3.34	7.25	31.87	39.80	13.33	8.95
	11	3.30	7.00	30.27	43.02	13.66	8.41
	12	3.51	8.96	28.42	40.64	12.62	8.03

4) 결 론

결과표인 <표 4.3>, <표 4.4>, <표 4.5>을 요약한 <표 4.6>에서 보면, 세 개의 추정방법중 Synthetic estimation method을 사용한 CV값이 가장 낮다는 것을 쉽게 알 수 있으며, Synthetic estimator가 Direct estimator보다 정도면에서 많이 향상된 것을 알 수 있다. 하지만 P.D.Falorsi, S.Falorsi와 A.Russo(1994)에 따르면, Synthetic estimation method를 사용한 추정치는 높은 편의(bias)를 갖는다고 하였다. 그러므로 작은 값의 CV으로만 추정방법을 선택하는 것은 무리가 있음을 참작해야 한다. 여기에서는 각 추정치의 적절한 편의를 구하는 것은 차후로 미루고, 각 추정치가 편의성이 있다고 가정했으며, 단지 CV값만을 가지고 안정성을 비교하여 적당한 추정방법을 검토해 보았다.

<표 4.6> A, B지역의 추정치(CV)의 평균

지 역	성별 / 추정방법	실 업 자		취 업 자		비경제활동인구	
		남	여	남	여	남	여
A	Direct	4.25	6.02	31.33	32.77	8.03	4.12
	Synthetic	1.49	1.97	13.16	20.98	4.33	1.99
	Composite	2.84	3.41	24.81	28.87	6.27	2.80
B	Direct	3.27	8.24	56.63	53.68	9.11	8.89
	Synthetic	0.24	0.63	1.85	6.52	0.62	0.54
	Composite	2.55	5.88	36.52	40.78	11.35	6.16

<표 4.3>과 <표 4.4>의 Direct와 Indirect estimator를 보면 표본수가 작은 B의 경우에 취업자의 CV에서 '0'이 나타난 경우를 볼 수 있는데, 이는 표본에서 조사된 자료가 없음을 나타낸다. 또한 조사 정도에 대한 추정을 불가능하게 하며, 그에 따른 추정치의 의미 또한 잃어버리게 된다. 그러므로 비록 소지역 추정방법일지라도 적절한 표본 크기가 요구됨을 알 수 있다. 하지만 Composite 추정법을 사용한 <표 4.5>를 보면, 적은 표본수에도 불구하고 추정치의 변동량이 안정적으로 나타났다. 그러므로 우리의 미래과제는 "우리가 원하는 추정치의 정도(precision)를 유지하기 위한 최소한의 표본수는 얼마인가?" 하는 문제이다. 이와 더불어 추정치가 '0'의 값을 갖는 경우를 대비하여 소지역 통계생산

이 가능한 적정 표본수와 그에 따른 보조변수의 활용법 연구 또한 미래과제 중 하나이다.

< 표 4.7 > 실업자 추정치에 따른 연간 상대표준오차(1994년)

지역 / 성별	실업자 추정치에 따른 상대표준오차 (CV)						
	C[1]	A			B		
	EAPS	Direct	Synthetic	Composite	Direct	Synthetic	Composite
남	4.41	4.25	1.49	2.83	3.27	1.49	2.55
여	6.87	6.02	1.97	3.40	8.24	1,.91	5.87

1) 1994년 C지역(64개 조사구) 경제활동인구조사 자료

<표 4.7>은 경제활동인구연보에 있는 C지역의 실업자의 연간 상대표준오차(CV)와 A와 B지역에서 small estimation method에 따라 산출된 실업자의 연간 상대표준오차(CV)이다.

EAPS의 결과를 3가지 방법에 의한 추정치와 비교해보자. 우선 A의 Direct 추정치는 C와 거의 비슷하지만, A의 Synthetic 추정치는 C보다 상당히 작음을 알 수 있다. 또한, A의 Composite 추정치는 Direct 추정치보다 작고 Synthetic 추정치보다는 큰 값을 가진다. 이것만을 보면 Synthetic 추정치가 상당히 안정되어 보인다.

하지만 Falorsi and Russo(1994)가 말한대로 Synthetic의 추정량이 높은 편의(bias)를 갖는다고 가정하면, 비록 Composite의 CV값이 Synthetic보다 크지만 Composite 추정치를 사용하는 데에는 큰 무리가 없을 것이다. 이는 조사의 정도(precision)면에서 Composite 방법에 따른 추정치에 의해 지역 통계의 생산이 가능함을 보여 주었다.

결론적으로 단기간의 실험적 자료에 의한 연구결과지만 표본수를 늘리지 않고 지역 통계를 생산할 수 있는 가능성을 볼 수 있었다. 앞으로 장기간의 연구와 실제적 적용·검토를 통해 우리의 현실에 맞는 지역통계를 생산할 수 있었으면 좋겠다.

Reference

1. 다목적 가구표본설계. 통우회

2. Battese, G.E. Harter, R.H. & Fuller, W.A.(1988), "An Error Components Model for Prediction of County crop Area using survey and Satellite Data" , *JASA*, Vol 83, 28-36

3. Falorsi, P.D., Falorsi S. & Russo A.(1994), "Empirical Comparison of Small Area Estimation Methods for the Italian Labour Force Survey", *Survey Methodology*, Vol. 20, 171-176.

4. Fay, R.E. & Nelson, C.T.(1994), "Estimation of Median Income for 4 Person Families by State" , *Survey Methodology*

5. Ghosh, M. and Rao, J.N.K.(1994), "Small Area Estimation", *Statistical Science*, Vol 9, No.1, 55-93

6. Schaible,W.L.," Indirect Estimation"

7. Singh, M.P., Gambino, J. and Mantel, H.J(1994), " Issues and Strategies for Small Area Data", *Survey Methodology*, Vol 20, 3-22

< 부 록 >

[부 록 1] 과 거 표 본 연 동 교 체 개 황

[부 록 2] Small Area Estimation 참 고 문 헌

[부 록 1] 과 거 표 본 연 동 교 체 개 황

(1983.10 ~ 1987.10)

과거 가구조사 표본연동교체 개황

Ⅰ. 개 요

1. 실시기간 : 1983. 10 ~ 1987.10

2. 필 요 성
   - 고정 표본사용의 문제점 배제
     · 응답가구의 응답부담 경감
     · 응답가구의 형식적 답변 가능성 방지
     · 실사상의 조사원 의사반영 가능성 방지

3. 고려사항
   - 전월, 전분기, 전년사용 표본과 부분적 중복
     · 비교상의 정도제고

   - 조사원의 업무량 증가
     · 무리한 업무량 증가는 조사의 질적 저하 초래

   - 동일 표본 최소한 6개월 이상 사용
     · 인구동태표본조사를 고려

4. '80년 기준 표본
   - 조사구수
     · 546조사구(시부 453, 군부 93)

   - 조사구별 조사구역수
     · 시부 2개 구역 : 경제활동인구조사 및 인구동태표본조사
       2개 구역 중 1개 구역 - 도시가계조사
     · 군부 6개 구역 : 경제활동인구조사 및 인구동태표본조사

5. 교체방법

가. 경제활동인구조사 및 인구동태표본조사
   - 매월 1/6 교체
     · 1/3 조사구에서 1/2구역 교체

- 중복비율
  · 전월과 5/6 중복
  · 전분기와 1/2 중복
  · 전년과 1/2 중복

- 동일가구에 대한 조사기간
  · 6개월 조사, 6개월 중지, 6개월 조사, 이후 중지

나. 도시가계
- 매월 1/12 교체
  · 1/12 조사구에서 조사구역 교체

- 중복비율
  · 전분기와 3/4 중복

- 동일가구에 대한 조사기간
  · 1년 조사후 조사중지

6. 실시시기
- 1983년 10월분 조사부터 실시

Ⅱ. 표본연동교체에 대한 문제점 발생 및 개선사항

1. 5개월 조사결과 1차 검토내용('84. 1)

가. 개황

주 요 사 항	도시가계조사	인구동태표본조사 · 경제활동인구조사
가. 월별 교체비율	1/12	1/6
나. 전년동기와의 표본중복 비율	·	1/2
다. 조사구당 교체주기	12개월	3개월
라. 구역당 교체주기	12개월	6개월
마. 구역당 조사기간	12개월	6개월
바. 구역당 조사회수	1회(동일 구역에 대해서는 다시 조사를 하지 않음)	2회 (6개월 조사, 6개월 중지, 6개월 조사)

나. 실사과 요구사항 및 문제점

1) 도시가계조사 : 표본교체에 대한 문제점 없음

2) 경제활동인구조사 및 인구동태표본조사
- 실사과(인구과, 사회과)의 요청 :
  자료의 분석을 위해서는 전년동기와의 표본중복 비율이 1/20이상이 되어야 함

가) 조사원의 업무량 증가
- 가구표 작성
  · 매월 평균 10가구 정도의 가구표본을 신규로 추가된 구역에 대하여 작성하여야 함

- 조사구역의 분리로 인한 업무량 증가
  · 인구동태 및 경활조사의 조사구역과 도시가계조사의 조사구역이 분리됨으로 인하여 실질적으로 1개 구역이 증가됨

- 가구표 수정, 보완 업무의 증가
  · 매월 182개 조사구의 가구표가 새로 작성되어, 조사관리과 및 자료처리과의 업무량이 증가

나) 조사상 애로부분의 증가
- 6개월 동안 조사한 후 일단 중지하였다가 재조사하게 되므로 오히려 불응가구가 증대되는 경향도 있음

- 교체빈도가 높아 대상가구의 설득에 애로가 있음

다. 검토내용

1) 기본방향 : 경제활동인구조사 및 인구동태표본조사의 대표도를 유지하되 조사상의 애로부분 조정

2) 검토내용
- 1안 : 교체비율 고정하에 조사구역을 조정하여 중복비율을 1/2로 유지
- 2안 : 교체비율과 조사구역을 조정하여 중복비율을 1/2로 유지

3) 교체비율에 따른 비교 검토 결과
- 1안 : 교체비율 고정하에 조사구역 조정만으로 중복율을 1/2로 유지하는 것은 불능인 것으로 판정

- 2안 : 교체비율 조정하에 중복율 유지 방안

교체방법	2 안	
	(제 1 안)	(제 2 안)
교 체 비 율	1/18	1/24
월별 교체조사구수 (총 546개 조사구)	60~61개	45~46개
월별 교체구역수 (총 1,092개 구역)	60~61개	45~46개
전년동기와의 중복비율	1/2	1/2
구역당 교체주기	18개월	24개월
구역당 조사회수	1회	1회
장 점	1. 대상가구 설득용이 2. 교체비율을 1/6으로 하였을 때 공백기간의 문제해결 (대상가구:18개월간 조사)	좌 동 (대상가구 24개월 조사)
단 점	1. 외국과 비교하여 상대적으로 교체비율이 낮음 2. 현재 대상가구에 조사기간 연장으로 설득상 애로	좌 동

4) 결론

- 제1안과 제2안 중에서 교체비율이 상대적으로 큰 제1안을 채택하는 것이 좋을 것으로 사료됨

- 교체비율을 1/6으로 하여 구역당 6개월 조사후 6개월동안 중지하였다가 다시 6개월동안 조사하는 체계를 18개월동안 계속 조사한 후 재조사하지 않는 방법으로 변경

라. 실시시기

- 현재 시행중인 표본연동교체의 완료시점인 '84. 3월 이후가 적당함

2. 최종 표본연동교체의 비율 변경 및 시행('84. 6)

가. 필요성

1) 조사의 정확도 제고
조사자료의 신뢰도를 가지는 단계에서 표본이 교체되므로 새로운 가구원 설득과정(1~3개월)에서 조사착오를 감소시킬 방책 필요

2) 조사상 애로부분 조정 필요
가) 현재의 표본연동교체제는 공백기간(6개월) 때문에 불응가구증대, 대상가구 설득에 애로가 있음

나) 경활조사 및 인구동태 표본조사 구역에서 도시가계조사 구역이 분리되어 1개 구역이 증가함에 따라 대상가구 설득에 애로점이 증대되었음.

나. 기본방향

1) 도시가계조사 : 당시 현행대로 유지

2) 경제활동인구조사 및 인구동태표본조사
- 교체비율과 교체방법 조정(교체비율을 1/6에서 1/12로 조정)
- 대표도 및 중복율은 1/2 유지

다. 개선내용

1) 내용

	현 행	개 선
교 체 비 율	1/6	1/12
월별 교체 조사구수 (총 549 조사구)	182 조사구 (2,750 가구)	90~91 조사구 (1,380 가구)
전년동기와 중복비율	1/2	1/2
구역당 교체주기	6개월 조사, 6개월 중지, 다시 6개월 조사후 조사중지	1그룹: 12개월 조사, 6개월 중지, 12개월 조사 2그룹: 12개월 조사, 6개월 중지, 6개월 조사 3그룹: 6개월 조사, 6개월 중지, 6개월 조사

2) 장점
- 대상가구 설득 용이
- 교체비율을 1/6로 하였을 때 공백기간의 문제해결
- 조사자료의 신뢰성 제고

라. 실시시기 : 1984. 8월부터 실시

마. 해당조사명
- 경제활동인구조사, 인구동태표본조사
- 도시가계조사는 종전과 동일

사. 실시방법

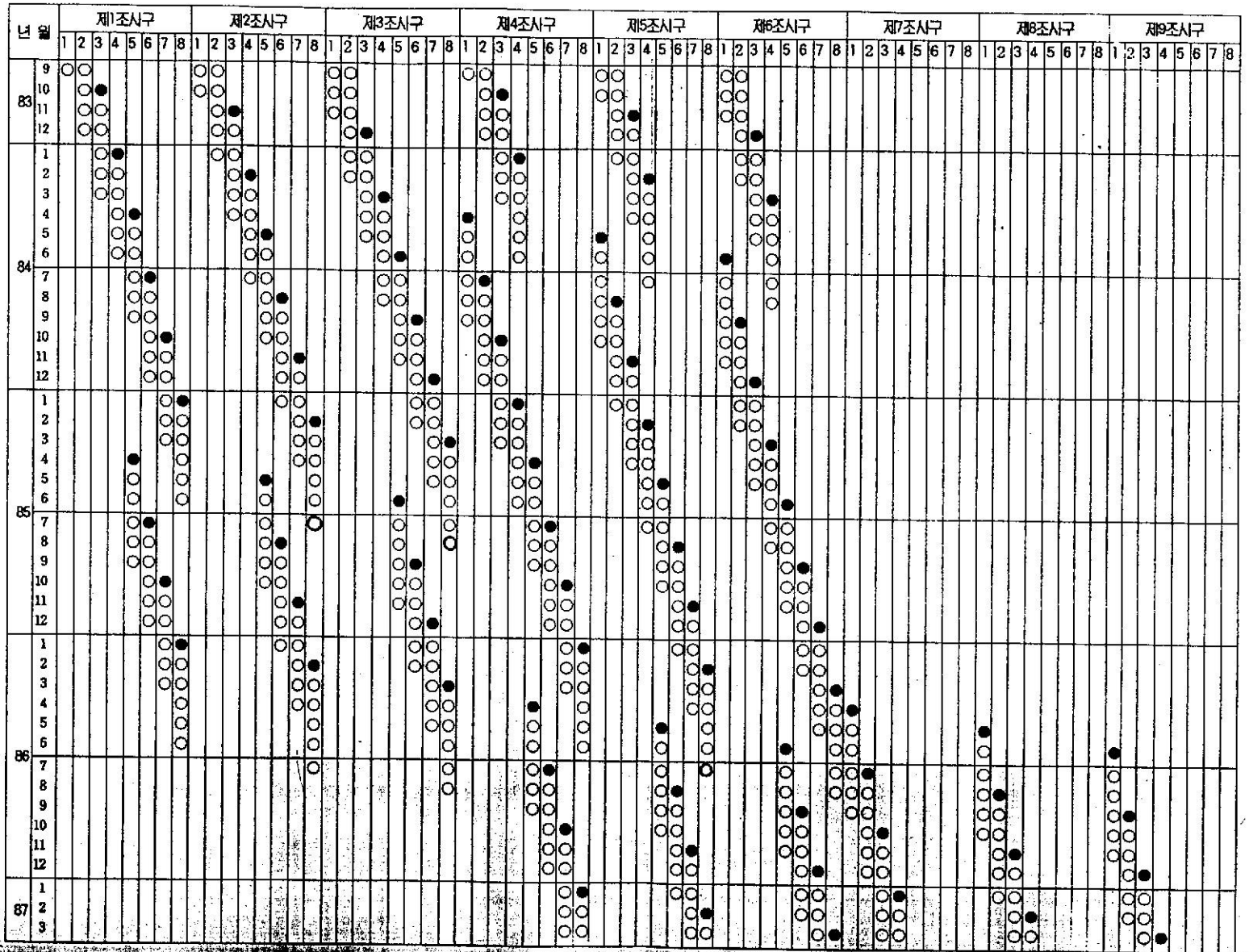
1) 조의 편성 및 교체방법 : 549개 조사구에 대하여 6개조로 분류되어 있는 것을 사용하여 매월 1조부터 차례로 1개조씩 교체

2) 표본조사구의 사용기간 : 다음 센서스에 의한 조사구의 추출까지 2년 6개월 정도이며, 신 표본설계시 새로운 연동교체안을 작성하여 실시

< 표본연동교체 변경 내역 예시도 >

조 및 구역번호 / 교체월	A조	B조	C조	D조	E조	F조
'84. 2	1. 2					
3	1. 2	1. 2				
4	1. 2	1. 2	1. 2			
5	1. 2	1. 2	1. 2	1. 2		
6	1. 2	1. 2	1. 2	1. 2	1. 2	
7	1. 2	1. 2	1. 2	1. 2	1. 2	1. 2
8	3. 2	1. 2	1. 2	1. 2	1. 2	1. 2
9	3. 2	3. 2	1. 2	1. 2	1. 2	1. 2
10	3. 2	3. 2	3. 2	1. 2	1. 2	1. 2
11	3. 2	3. 2	3. 2	3. 2	1. 2	1. 2
12	3. 2	3. 2	3. 2	3. 2	3. 2	1. 2
'85. 1	3. 2	3. 2	3. 2	3. 2	3. 2	3. 2
2	3. 1	3. 2	3. 2	3. 2	3. 2	3. 2
3	3. 1	3. 1	3. 2	3. 2	3. 2	3. 2
4	3. 1	3. 1	3. 1	3. 2	3. 2	3. 2
5	3. 1	3. 1	3. 1	3. 1	3. 2	3. 2
6	3. 1	3. 1	3. 1	3. 1	3. 1	3. 2
7	3. 1	3. 1	3. 1	3. 1	3. 1	3. 1
8	2. 1	3. 1	3. 1	3. 1	3. 1	3. 1
9	2. 1	2. 1	3. 1	3. 1	3. 1	3. 1
10	2. 1	2. 1	2. 1	3. 1	3. 1	3. 1
11	2. 1	2. 1	2. 1	2. 1	3. 1	3. 1
12	2. 1	2. 1	2. 1	2. 1	2. 1	3. 1
'86. 1	2. 1	2. 1	2. 1	2. 1	2. 1	2. 1
2	2. 3	2. 1	2. 1	2. 1	2. 1	2. 1
3	2. 3	2. 3	2. 1	2. 1	2. 1	2. 1
4	2. 3	2. 3	2. 3	2. 1	2. 1	2. 1
5	2. 3	2. 3	2. 3	2. 3	2. 1	2. 1
6	2. 3	2. 3	2. 3	2. 3	2. 3	2. 1
7	2. 3	2. 3	2. 3	2. 3	2. 3	2. 3
8		2. 3	2. 3	2. 3	2. 3	2. 3
9			2. 3	2. 3	2. 3	2. 3
10				2. 3	2. 3	2. 3
11					2. 3	2. 3
12						2. 3

< 표 본 교 체 내 역 표 >


년 월
제1조사구
제2조사구
제3조사구
제4조사구
제5조사구
제6조사구
제7조사구
제8조사구
제9조사구
83
84
85
86
87

[부 록 2] Small Area Estimation 참 고 문 헌

Small Area Estimation: An Appraisal

M. Ghosh and J. N. K. Rao

Abstract. Small area estimation is becoming important in survey sampling due to a growing demand for reliable small area statistics from both public and private sectors. It is now widely recognized that direct survey estimates for small areas are likely to yield unacceptably large standard errors due to the smallness of sample sizes in the areas. This makes it necessary to "borrow strength" from related areas to find more accurate estimates for a given area or, simultaneously, for several areas. This has led to the development of alternative methods such as synthetic, sample size dependent, empirical best linear unbiased prediction, empirical Bayes and hierarchical Bayes estimation. The present article is largely an appraisal of some of these methods. The performance of these methods is also evaluated using some synthetic data resembling a business population. Empirical best linear unbiased prediction as well as empirical and hierarchical Bayes, for most purposes, seem to have a distinct advantage over other methods.

Key words and phrases: Borrowing strength, demographic methods, empirical Bayes, empirical best linear unbiased prediction, hierarchical Bayes, synthetic estimation

1. INTRODUCTION

The terms "small area" and "local area" are commonly used to denote a small geographical area, such as a county, a municipality or a census division. They may also describe a "small domain," i.e., a small subpopulation such as a specific age-sex-race group of people within a large geographical area. In this paper, we use these terms interchangeably.

The use of small area statistics originated several centuries ago. Brackstone (1987) mentions the existence of such statistics in 11th century England and 17th century Canada. Many other countries may well have similar early histories. However, these early small area statistics were all based either on a census or on administrative records aiming at complete enumeration.

For the past few decades, sample surveys, for most purposes, have taken the place of complete enumeration or census as a more cost-effective means of obtaining information on wide-ranging topics of interest at frequent intervals over time. Sample survey data certainly can be used to derive reliable estimators of totals and means for large areas or domains. However, the usual direct survey estimators for a small area, based on data only from the sample units in the area, are likely to yield unacceptably large standard errors due to the unduly small size of the sample in the area. Sample sizes for small areas are typically small because the overall sample size in a survey is usually determined to provide specific accuracy at a much higher level of aggregation than that of small areas. Thus, until recently, the use of survey data in developing reliable small area statistics, possibly in conjunction with the census and administrative data, has received very little attention.

Things have changed significantly during the last few years, largely due to a growing demand for reliable small area statistics from both the public and private sectors. These days, in many countries including the United States and Canada, there is "increasing government concern with issues of distribution, equity and disparity" (Brackstone, 1987). For example, there may exist geographical subgroups within a given population that are far below the average in certain respects, and need definite upgrading. Before taking remedial action, there is a need to identify such regions, and accordingly, one must have statistical data at the relevant geographical levels. Small area statistics are also needed

M. Ghosh is Professor, Department of Statistics, University of Florida, Gainesville, Florida 32611-2049. J. N. K. Rao is Professor of Statistics, Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada K1S 5B6.

in the apportionment of government funds, and in regional and city planning. In addition, there are demands from the private sector since the policy-making of many businesses and industries relies on local socio-economic conditions. Thus, the need for small area statistics can arise from diverse sources.

Demands of the type described above could not have been met without significant advances in statistical data processing. Fortunately, with the advent of high-speed computers, fast processing of large data sets made feasible the provision of timely data for small areas. In addition, several powerful statistical methods with sound theoretical foundation have emerged for the analysis of local area data. Such methods "borrow strength" from related or similar small areas through explicit or implicit models that connect the small areas via supplementary data (e.g., census and administrative records). However, these methods are not readily available in a package to the user, and a unified presentation which compares and contrasts the competing methods has not been attempted before.

Earlier reviews on the topic of small area estimation focussed on demographic methods for population estimation in post-censual years. Morrison (1971) covers the pre-1970 period very well, including a bibliography. National Research Council (1980) provides detailed information as well as a critical evaluation of the Census Bureau's procedures for making post-censual estimates of the population and per capita income for local areas. Their document was the report of a panel on small-area estimates of population and income set up by the Committee on National Statistics at the request of the Census Bureau and the Office of Revenue Sharing of the U.S. Department of Treasury. This document also assessed the "levels of accuracy of current estimates in light of the uses made of them and of the effect of potential errors on these uses." Purcell and Kish (1979) review demographic methods as well as statistical methods of estimation for small domains. An excellent review provided by Zidek (1982) introduces a criterion that can be used to evaluate the relative performance of different methods for estimating the populations of local areas. McCullagh and Zidek (1987) elaborate this criterion more fully. Statistics Canada (1987) provides an overview and evaluation of the population estimation methods used in Canada.

Prompted by the growing demand for reliable small area statistics, several symposia and workshops were also organized in recent years, and some of the proceedings have also been published: National Institute on Drug Abuse, Princeton Conference (see National Institute on Drug Abuse, 1979), International Symposium on Small Area Statistics, Ottawa [see Platek et al. (1987) for the invited papers and Platek and Singh (1986) for the contributed papers presented at the symposium]; International Symposium on Small Area Statistics, New Orleans, 1988, organized by the National Center for Health Statistics; Workshop on Small Area Estimates for Military Personnel Planning, Washington, D.C., 1989, organized by the Committee on National Statistics; International Scientific Conference on Small Area Statistics and Survey Designs, Warsaw, Poland, 1992, (see Kalton, Kordos and Platek, 1993). The published proceedings listed above provide an excellent collection of both theoretical and application papers.

Reviews by Rao (1986) and Chaudhuri (1992) cover more recent techniques as well as traditional methods of small area estimation. Schaible (1992) provides an excellent account of small area estimators used in U.S. Federal programs (see NTIS, 1993, for a full report prepared by the Subcommittee on Small Area Estimation of the Federal Committee on Statistical Methodology, Office of Management and Budget).

The present article considerably updates earlier reviews by introducing several recent techniques and evaluating them in the light of practical considerations. Particularly noteworthy among the newer methods are the empirical Bayes (EB), hierarchical Bayes (HB) and empirical best linear unbiased prediction (EBLUP) procedures which have made significant impact on small area estimation during the past decade. Before discussing these methods in the sequel, it might be useful to mention a few important applications of small area estimation methods as motivating examples.

As our first example, we cite the Federal-State Cooperative Program (FSCP) initiated by the U.S. Bureau of the Census in 1967 (see National Research Council, 1980). A basic goal of this program was to provide high-quality, consistent series of county population estimates with comparability from area to area. Forty-nine states (with the exception of Massachusetts) currently participate in this program, and their designated agencies work together with the Census Bureau under this program. In addition to county estimates, several members of the FSCP now produce subcounty estimates as well. The FSCP plays a key role in the Census Bureau's post censual estimation program as the FSCP contacts provide the bureau a variety of data that can be used in making post censual population estimates. Considerable methodological research on small area population estimation is being conducted in the Census Bureau.

Our second example is taken from Fay and Herriot (1979) whose objective was to estimate the per

capita income (PCI) for several small places. The U.S. Census Bureau was required to provide the Treasury Department with the PCI estimates and other statistics for state and local governments receiving funds under the General Revenue Sharing Program. These statistics were then used by the Treasury Department to determine allocations to the local governments within the different states by dividing the corresponding state allocations. Initially, the Census Bureau determined the current estimates of PCI by multiplying the 1970 census estimates of PCI in 1969 (based on a 20 percent sample) by ratios of an administrative estimate of PCI in the current year and a similarly derived estimate for 1969. The bureau then confronted the problem that among the approximately 39,000 local government units about 15,000 were for places having fewer than 500 persons in 1970. The sampling errors in the PCI estimates for such small places were large: for a place of 500 persons the coefficient of variation was about 13 percent while it increased to about 30 percent for a place of 100 persons. Consequently, the Bureau initially decided to set aside the census estimates for these small areas and use the corresponding county PCI estimates in their place. This solution proved unsatisfactory, however, in that the census estimates of PCI for a large number of small places differed significantly from the corresponding county estimates, after taking account of the sampling errors. Fay and Herriot (1979) suggest better estimates based on the EB method and present empirical evidence that these have average error smaller than either the census sample estimates or the county averages. The proposed estimate for a small place is a weighted average of the census sample estimate and a "synthetic" estimate obtained by fitting a linear regression equation to the sample estimates of PCI using as independent variables the corresponding county averages, tax-return data for 1969 and data on housing from the 1970 census. The Fay-Herriot method was adopted by the Census Bureau in 1974 to form updated estimates of PCI for small places. Section 4 discusses the Fay-Herriot model and similar models for other purposes, all involving linear regression models with random small area effects.

Our third example refers to the highly debated and controversial issue of adjusting for population undercount in the 1980 U.S. Census. Every tenth year since 1790 a census has been taken to count the U.S. population. The census provides the population count for the whole country as well as for each of the 50 states, 3000 counties and 39,000 civil divisions. These counts are used by the Congress for apportioning funds, amounting to about 100 billion dollars a year during the early 1980s, to the different state and local governments.

It is now widely recognized that complete coverage is impossible. In 1980, vast sums of money and intellectual resources were expended by the U.S. Census Bureau on the reduction of non-coverage. Despite this, there were complaints of undercounts by several major cities and states for their respective areas, and indeed New York State filed a lawsuit against the Census Bureau in 1980 demanding the Bureau to revise its count for that state.

An undercount is the difference between omissions and erroneous inclusions in the census, and it is typically positive. In New York State's law suit against the Census Bureau, E.P. Ericksen and J.B. Kadane, among other statisticians, appeared as the plaintiff's expert witnesses. They proposed using weighted averages of sample estimates and synthetic regression estimates of the 1980 Census undercount, similar to those of Fay and Herriot (1979) for PCI, to arrive at the adjusted population counts of the 50 states and the 16 large cities, including the State of New York and New York City. The sample estimates are obtained from a Post Enumeration Survey. Their general philosophy on the role of adjustment as well as the explicit regression models used for obtaining the regression estimates are documented in Ericksen and Kadane (1985) and Ericksen, Kadane and Tukey (1989). These authors also suggest using the regression equation for areas where no sample data are available. As a historical aside, we may point out here that the regression method for improving local area estimates was first used by Hansen, Hurwitz and Madow (1953, pages 483–486), but its recent popularity owes much to Ericksen (1974).

While the Ericksen-Kadane proposal was applauded by many as the first serious attempt towards adjustment of Census undercount, it has also been vigorously criticized by others (see, e.g., the discussion of Ericksen and Kadane, 1985). In particular, Freedman and Navidi (1986, 1992) criticized them for not validating their model and for not making their assumptions explicit. They also raise several other technical issues, including the effect of large biases and large sampling errors in the sample estimates. Ericksen and Kadane (1987, 1992), Cressie (1989, 1992), Isaki et al. (1987) and others address these difficulties, but clearly further research is needed. Researchers within and outside the U.S. Census Bureau are currently studying various models for census undercount and the properties of the resulting estimators and associated measures of uncertainty using the EBLUP, EB, HB and related approaches.

Our fourth example, taken from Battese, Harter

and Fuller (1988), concerns the estimation of areas under corn and soybeans for each of 12 counties in North-Central Iowa using farm-interview data in conjunction with LANDSAT satellite data. Each county was divided into area segments, and the areas under corn and soybeans were ascertained for a sample of segments by interviewing farm operators; the number of sample segments in a county ranged from 1 to 6. Auxiliary data in the form of numbers of pixels (a term used for "picture elements" of about 0.45 hectares) classified as corn and soybeans were also obtained for all the area segments, including the sample segments, in each county using the LANDSAT satellite readings. Battese, Harter and Fuller (1988) employ a "nested error regression" model involving random small area effects and the segment-level data and then obtain the EBLUP estimates of county areas under corn and soybeans using the classical components of variance approach (see Section 5). They also obtain estimates of mean squared error (MSE) of their estimates by taking into account the uncertainty involved in estimating the variance components. Datta and Ghosh (1991) apply the HB approach to these data and show that the two approaches give similar results.

Our final example concerns the estimation of mean wages and salaries of units in a given industry for each census division in a province using gross business income as the only auxiliary variable with known population means (see Särndal and Hidiroglou, 1989). This example will be used in Section 6 to compare and evaluate, under simple random sampling, several competing small area estimators discussed in this paper, treating the census divisions as small areas. We were able to compare the actual errors of the different small area estimators since the true mean wages and salaries for each small area are known.

The outline of the paper is as follows. Section 2 gives a brief account of classical demographic methods for local estimation of population and other characteristics of interest in post-censual years. These methods use current data from administrative registers in conjunction with related data from the latest census. Section 3 provides a discussion of traditional synthetic estimation and related methods under the design-based framework. Two types of small area models that include random area-specific effects are introduced in Section 4. In the first type, only area specific auxiliary data, related to parameters of interest, are available. In the second type of models, element-specific auxiliary data are available for the population elements; and the variable of interest is assumed to be related to these variables through a nested error regression model. We present the EBLUP, EB and HB approaches to small area estimation in Section 5 in the context of basic models given in Section 4. Both point estimation and measurement of uncertainty associated with the estimators are studied. Section 6 compares the performances of several competing small area estimators using sample data drawn from a synthetic population resembling the business population studied by Särndal and Hidiroglou (1989). In Section 7, we focus on special problems that may be encountered in implementing model-based methods for small area estimation. In particular, we give a brief account of model diagnostics for the basic models of Section 4 and of constrained estimation. Various extensions of the basic models are also mentioned in this section. Finally, some concluding remarks are made in Section 8.

The scope of our paper is limited to methods of estimation for small areas; but the development and provision of small area statistics involves many other issues, including those related to sample design and data development, organization and dissemination. Brackstone (1987) gives an excellent account of these issues in the context of Statistics Canada's Small Area Data Program. Singh, Gambino and Mantel (1992) highlight the need for developing an overall strategy that includes planning, designing and estimation stages in the survey process.

2. DEMOGRAPHIC METHODS

As pointed out earlier, demographers have long been using a variety of methods for local estimation of population and other characteristics of interest in post-censual years. Purcell and Kish (1980) categorize these methods under the general heading of Symptomatic Accounting Techniques (SAT). Such techniques utilize current data from administrative registers in conjunction with related data from the latest census. The diverse registration data used in the U.S. include "symptomatic" variables, such as the numbers of births and deaths, of existing and new housing units and of school enrollments whose variations are strongly related to changes in population totals or in its components. The SAT methods studied in the literature include the Vital Rates (VR) method (Bogue, 1950), the composite method (Bogue and Duncan, 1959), the Census Component Method II (CM-II) (U.S. Bureau of the Census, 1966), and the Administrative Records (AR) method (Starsinic, 1974), and the Housing Unit (HU) method (Smith and Lewis, 1980).

The VR method uses only birth and death data, and these are used as symptomatic variables rather than as components of population change. First, in a given year, say t, the annual number of births,

b_t, and deaths, d_t, are determined for a local area. Next the crude birth and death rates, r_{1t} and r_{2t}, for that local area are estimated by

$$r_{1t} = r_{10}(R_{1t}/R_{10}), \qquad r_{2t} = r_{20}(R_{2t}/R_{20}),$$

where r_{10} and r_{20} respectively denote the crude birth and death rates for the local area in the latest census year ($t = 0$) while $R_{1t}(R_{2t})$ and $R_{10}(R_{20})$ respectively denote the crude birth (death) rates in the current and census years for a larger area containing the local area. The population P_t for the local area at year t is then estimated by

$$P_t = \tfrac{1}{2}(b_t/r_{1t} + d_t/r_{2t}).$$

As pointed out by Marker (1983), the success of the VR method depends heavily on the validity of the assumption that the ratios r_{1t}/r_{10} and r_{2t}/r_{20} for the local area are approximately equal to the corresponding ratios, R_{1t}/R_{10} and R_{2t}/R_{20}, for the larger area. Such an assumption is often questionable, however.

The composite method is an extension of the VR method that sums independently computed age-sex-race specific estimates based on births, deaths and school enrollments (see Zidek, 1982, for details).

The CM-II method takes account of net migration unlike the previous methods. Denoting the net migration in the local area during the period since the last census as m_t, an estimate of P_t is given by

$$P_t = P_0 + b_t - d_t + m_t,$$

where P_0 is the population of the local area in the census year $t = 0$. In the U.S., the net migration is further subdivided into military and civilian migration. The former is readily obtainable from administrative records while the CM-II estimates civilian migration from school enrollments. The AR method, on the other hand, estimates the net migration from records for individuals as opposed to collect units like schools (see Zidek, 1982, for details).

The HU method expresses P_t as

$$P_t = (H_t)(PPH_t) + GQ_t,$$

where H_t is the number of occupied housing units at time t, PPH_t is the average number of persons per housing unit at time t and GQ_t is the number of persons in group quarters at time t. The quantities H_t, PPH_t and GQ_t all need to be estimated. Smith and Lewis (1980) report different methods of estimating these quantities.

As pointed out by Marker (1983), most of the estimation methods mentioned above can be identified as special cases of multiple linear regression.

Regression-symptomatic procedures also use multiple linear regression for estimating local area populations utilizing symptomatic variables as independent variables in the regression equation. Two such procedures are the ratio-correlation method and the difference-correlation method. Briefly, the former method is as follows: Let 0, 1 and $t(> 1)$ denote two consecutive census years and the current year, respectively. Also, let $P_{i\alpha}$ and $S_{ij\alpha}$ be the population and the value of the jth symptomatic variable for the ith local area ($i = 1, \ldots, m$) in the year $\alpha(= 0, 1, t)$. Further, let $p_{i\alpha} = P_{i\alpha}/\Sigma_i P_{i\alpha}$ and $s_{ij\alpha} = S_{ij\alpha}/\Sigma_i S_{ij\alpha}$ be the corresponding proportions, and write $R'_i = p_{i1}/p_{i0}$, $R_i = p_{it}/p_{i1}$, $r'_{ij} = s_{ij1}/s_{ij0}$ and $r_{ij} = s_{ijt}/s_{ij1}$. Using the data $(R'_i, r'_{i1}, \ldots, r'_{ip}; i = 1, \ldots, m)$ and multiple regression, we first fit

$$(2.1) \qquad R'_i = \hat{\beta}'_0 + \hat{\beta}'_1 r'_{i1} + \ldots + \hat{\beta}'_p r'_{ip},$$

where $\hat{\beta}$s are the estimated regression coefficients that link the change, R'_i, in the population proportions between the two census years to the corresponding changes, r'_{ij}, in the proportions for the symptotmatic variables. Next the changes, R_i, in the post censual period are predicted as

$$\tilde{R}_i = \hat{\beta}'_0 + \hat{\beta}'_1 r_{i1} + \ldots + \hat{\beta}'_p r_{ip},$$

using the known changes, r_{ij}, in the symptomatic proportions in the post censual period and the estimated regression coefficients. Finally, the current population counts, P_{it}, are estimated as

$$\tilde{P}_{it} = \tilde{R}_i p_{i1} \left(\sum_i P_{it} \right),$$

where the total current count, $\Sigma_i P_{it}$, is ascertained from other sources. In the difference-correlation method, differences between the proportions at the two pairs of time points, $(0, 1)$ and $(1, t)$, are used rather than their ratios.

The regression-symptomatic procedures described above use the regression coefficients, $\hat{\beta}'_j$, in the last intercensual period, but significant changes in the statistical relationship can lead to errors in the current postcensal estimates. The sample-regression method (Ericksen, 1974) avoids this problem by using sample estimates of R_i to establish the current regression equation. Suppose sample estimates of R_i are available for k out of m local areas, say $\hat{R}_1, \ldots, \hat{R}_k$. Then one fits the regression equation

$$\hat{R}_i = \hat{\beta}_0 + \hat{\beta}_1 r_{i1} + \ldots + \hat{\beta}_p r_{ip}$$

to the data $(\hat{R}_i, r_{i1}, \ldots, r_{ip})$ from the k sampled areas, instead of (2.1); and then obtains the sample-regression estimators, $\hat{R}_{i(\text{reg})}$, for all the areas using the known symptomatic ratios r_{ij} $(i = 1, \ldots, m)$:

$$\hat{R}_{i(\text{reg})} = \hat{\beta}_0 + \hat{\beta}_1 r_{i1} + \ldots + \hat{\beta}_p r_{ip}.$$

Using 1970 census data and sample data from the Current Population Survey (CPS), Ericksen (1974) has shown that the reduction of mean error is slight compared to the ratio-correlation method but that of large errors (10% or greater) is more substantial. The success of Ericksen's method depends largely on the size and quality of the samples, the dynamics of the regression relationships and the nature of the variables.

3. SYNTHETIC AND RELATED ESTIMATORS

Gonzalez (1973) describes synthetic estimates as follows: "An unbiased estimate is obtained from a sample survey for a large area; when this estimate is used to derive estimates for subareas under the assumption that the small areas have the same characteristics as the large area, we identify these estimates as synthetic estimates." The National Center for Health Statistics (1968) first used synthetic estimation to calculate state estimates of long and short term physical disabilities from the National Health Interview Survey data. This method is traditionally used for small area estimation, mainly because of its simplicity, applicability to general sampling designs and potential of increased accuracy in estimation by borrowing information from similar small areas. We now give a brief account of synthetic estimation and related methods, under the design-based framework.

3.1 Synthetic Estimation

Suppose the population is partitioned into large domains g for which reliable direct estimators, $\hat{Y}'_g$, of the totals, $Y_{\cdot g}$, can be calculated from the survey data; the small areas, i, may cut across g so that $Y_{\cdot g} = \Sigma_i Y_{ig}$, where Y_{ig} is the total for cell (i,g). We assume that auxiliary information in the form of totals, X_{ig}, is also available. A synthetic estimator of small area total $Y_i = \Sigma_g Y_{ig}$ is then given by

$$\hat{Y}_i^S = \sum_g (X_{ig}/X_{\cdot g})\hat{Y}'_{\cdot g}, \tag{3.1}$$

where $X_{\cdot g} = \Sigma_i X_{ig}$ (Purcell and Linacre, 1976; Ghangurde and Singh, 1977). The estimator (3.1) has the desirable consistency property that $\Sigma_i \hat{Y}_i^S$ equals the reliable direct estimator $\hat{Y}' = \Sigma_g \hat{Y}'_{\cdot g}$ of the population total Y, unlike the original estimator proposed by the National Center for Health Statistics (1968) which uses the ratio $X_{ig}/\Sigma_g X_{ig}$ instead of $X_{ig}/X_{\cdot g}$.

The direct estimator $\hat{Y}'_g$ used in (3.1) is typically a ratio estimator of the form

$$\hat{Y}'_{\cdot g} = \left[\left(\sum_{\ell \in s_{\cdot g}} w_\ell y_\ell\right) \Big/ \left(\sum_{\ell \in s_{\cdot g}} w_\ell x_\ell\right)\right] X_{\cdot g} = (\hat{Y}_{\cdot g}/\hat{X}_{\cdot g}) X_{\cdot g},$$

where $s_{\cdot g}$ denotes the sample in the large domain g and w_ℓ is the sampling weight attached to the ℓth element. For this choice, the synthetic estimator (3.1) reduces to $\hat{Y}_i^S = \Sigma_i X_{ig}(\hat{Y}_{\cdot g}/\hat{X}_{\cdot g})$.

If $\hat{Y}'_{\cdot g}$ is approximately design-unbiased, the design-bias of $\hat{Y}_i^S$ is given by

$$E(\hat{Y}_i^S) - Y_i \doteq \sum_g X_{ig}(Y_{\cdot g}/X_{\cdot g} - Y_{ig}/X_{ig}),$$

which is not zero unless $Y_{ig}/X_{ig} = Y_{\cdot g}/X_{\cdot g}$ for all g. In the special case where the auxiliary information X_{ig} equals the population count N_{ig}, the latter condition is equivalent to assuming that the small area means $\bar{Y}_{ig}$ in each group g equal the overall group mean, $\bar{Y}_{\cdot g}$. Such an assumption is quite strong, and in fact synthetic estimators for some of the areas can be heavily biased in the design-based framework.

It follows from (3.1) that the design-variance of $\hat{Y}_i^S$ will be small since it depends only on the variances and covariances of the reliable estimators $\hat{Y}'_{\cdot g}$. The variance of $\hat{Y}_i^S$ is readily estimated, but it is more difficult to estimate the MSE of $\hat{Y}_i^S$. Under the assumption $\text{cov}(\hat{Y}_i, \hat{Y}_i^S) \doteq 0$, where $\hat{Y}_i$ is a direct, unbiased estimator of Y_i, an approximately unbiased estimator of MSE is given by

$$\text{mse}(\hat{Y}_i^S) = (\hat{Y}_i^S - \hat{Y}_i)^2 - v(\hat{Y}_i). \tag{3.2}$$

Here $v(\hat{Y}_i)$ is a design-unbiased estimator of variance of $\hat{Y}_i$. The estimators (3.2), however, are very unstable. Consequently, it is customary to average these estimators over i to get a stable estimator of MSE (Gonzalez, 1973), but such a global measure of uncertainty can be misleading. Note that the assumption $\text{cov}(\hat{Y}_i, \hat{Y}_i^S) \doteq 0$ may be realistic in practice since $\hat{Y}_i^S$ is much less variable than $\hat{Y}_i$.

Nichol (1977) proposes to add the synthetic estimate, $\hat{Y}_i^S$, as an additional independent variable in the sample-regression method. This method, called the combined synthetic-regression method, showed improvement, in empirical studies, over both the synthetic and sample-regression estimates.

Chambers and Feeney (1977) and Purcell and Kish (1980) propose structure preserving estimation (SPREE) as a generalization of synthetic estimation in the sense it makes a fuller use of reliable direct estimates. SPREE uses the well-known method of iterative proportional fitting of margins in a multi-way table, where the margins are direct estimates.

3.2 Composite Estimation

A natural way to balance the potential bias of a synthetic estimator against the instability of a direct estimator is to take a weighted average of the two estimators. Such composite estimators may be written as

$$(3.3) \qquad \hat{Y}_i^C = w_i \hat{Y}_{1i} + (1 - w_i)\hat{Y}_{2i},$$

where $\hat{Y}_{1i}$ is a direct estimator, $\hat{Y}_{2i}$ is an indirect estimator and w_i is a suitably chosen weight ($0 \le w_i \le 1$). For example, the unbiased estimator $\hat{Y}_i$ may be chosen as $\hat{Y}_{1i}$, and the synthetic estimator $\hat{Y}_i^S$ as $\hat{Y}_{2i}$. Many of the estimators proposed in the literature, both design-based and model-based, have the form (3.3). Section 5 gives such estimators under realistic small area models that account for area-specific effects. In this subsection, we mainly focus on the determination of weights, w_i, in the design-based framework using $\hat{Y}_{1i} = \hat{Y}_i$ and $\hat{Y}_{2i} = \hat{Y}_i^S$.

Optimal weights, $w_i(\text{opt})$, may be obtained by minimising the MSE of $\hat{Y}_i^C$ with respect to w_i assuming $\text{cov}(\hat{Y}_i, \hat{Y}_i^S) \doteq 0$:

$$(3.4) \quad w_i(\text{opt}) = \text{MSE}\,(\hat{Y}_i^S)/[\text{MSE}\,(\hat{Y}_i^S) + V(\hat{Y}_i)].$$

The optimal weight (3.4) may be estimated by substituting the estimator mse $(\hat{Y}_i^S)$ given in (3.2) for the numerator and $(\hat{Y}_i^S - \hat{Y}_i)^2$ for the denominator, but the resulting weights can be very unstable. Schaible (1978) proposes an "average" weighting scheme based on several variables to overcome this difficulty, noting that the composite estimator is quite robust to deviations from $w_i(\text{opt})$. Another approach (Purcell and Kish, 1979) uses a common weight, w, and then minimizes the average MSE, i.e., $m^{-1}\Sigma_i$ MSE $(\hat{Y}_i^C)$, with respect to w. This leads to estimated weight of the form

$$(3.5) \quad \hat{w}(\text{opt}) = 1 - \sum_i v(\hat{Y}_i) \Big/ \sum_i (\hat{Y}_i^S - \hat{Y}_i)^2.$$

If the variances of $\hat{Y}_i$'s are approximately equal, then we can replace $v(\hat{Y}_i)$ by the average $\bar{v} = \Sigma_i v(\hat{Y}_i)/m$ in which case (3.5) reduces to James-Stein type weight:

$$\hat{w}(\text{opt}) = 1 - m\bar{v} \Big/ \sum_i (\hat{Y}_i^S - \hat{Y}_i)^2.$$

The choice of a common weight, however, is not reasonable if the individual variances, $V(\hat{Y}_i)$, vary considerably. Also, the James-Stein estimator can be less efficient than the direct estimator, $\hat{Y}_i$, for some individual areas if the small areas that are pooled are not "similar" (C.R. Rao and Shinozaki, 1978).

Simple weights, w_i, that depend only on the domain counts or the domain totals of a covariate x have also been proposed in the literature. For example, Drew, Singh and Choudhry (1982) propose the sample size dependent estimator which uses the weight

$$(3.6) \qquad w_i(D) = \begin{cases} 1, & \text{if } \hat{N}_i \ge \delta N_i, \\ \hat{N}_i/(\delta N_i), & \text{otherwise,} \end{cases}$$

where $\hat{N}_i$ is the direct, unbiased estimator of the known domain population size N_i and δ is subjectively chosen to control the contribution of the synthetic estimator. This estimator with $\delta = 2/3$ and a generalized regression synthetic estimator replacing the ratio synthetic estimator $\hat{Y}_i^S$ is currently being used in the Canadian Labour Force Survey to produce domain estimates. Särndal and Hidiroglou (1989) propose an alternative estimator which uses the weight

$$(3.7) \qquad w_i(S) = \begin{cases} 1, & \text{if } \hat{N}_i \ge N_i \\ (\hat{N}_i/N_i)^{h-1}, & \text{otherwise,} \end{cases}$$

where h is subjectively chosen. They, however, suggest $h = 2$ as a general-purpose value. Note that the weights (3.6) and (3.7) are identical if one chooses $\delta = 1$ and $h = 2$.

To study the nature of the weights $w_i(D)$ or $w_i(S)$, let us consider the special case of simple random sampling of n elements from a population of N elements. In this case, $\hat{N}_i = N(n_i/n)$, where the random variable n_i is the sample size in ith domain. Taking $\delta = 1$ in (3.6), it now follows that $w_i(D) = w_i(S) = 1$ if n_i is at least as large as the expected sample size $E(n_i) = n(N_i/N)$, that is, the sample size dependent estimators can fail to borrow strength from related domains even when $E(n_i)$ is not large enough to make the direct estimator $\hat{Y}_i$ reliable. On the other hand, when $\hat{N}_i < N_i$ the weight $w_i(D)$, which equals $w_i(S)$ when $h = 2$, decreases as n_i decreases. As a

result, more weight is given to the synthetic component as n_i decreases. Thus, the weights behave well unlike in the case $\hat{N}_i \geq N_i$. Another disadvantage is that the weights do not take account of the size of between area variation relative to within area variation for the characteristic of interest, that is, all characteristics get the same weight irrespective of their differences with respect to between area homogeneity.

Holt, Smith and Tomberlin (1979) obtain a best linear unbiased prediction (BLUP) estimator of Y_i under the following model for the finite population:

$$(3.8) \qquad \begin{aligned} y_{ig\ell} &= \mu_g + e_{ig\ell}, \\ \ell &= 1,\ldots,N_{ig}; \qquad g = 1,\ldots,G; \qquad i = 1,\ldots,m \end{aligned}$$

where $y_{ig\ell}$ is the y-value of the ℓth unit in the cell (i,g), μ_g's are fixed effects and the errors $e_{ig\ell}$ are uncorrelated with zero means and variances σ_g^2. Further, N_{ig} denotes the number of population elements in the large domain g that belong to the small area i. Suppose n_{ig} elements in a sample of size n fall in cell (i,g), and let $\bar{y}_{ig}$ and $\bar{y}_{\cdot g}$ denote the sample means for (i,g) and g, respectively.

The best linear unbiased estimator of μ_g under (3.8) is $\hat{\mu}_g = \bar{y}_{\cdot g}$ which in turn leads to the BLUP estimator of Y_i given by

$$\hat{Y}_i^B = \sum_g \hat{Y}_{ig}^C,$$

where $\hat{Y}_{ig}^C$ is a composite estimator of the total Y_{ig} giving the weight $w_{ig} = n_{ig}/N_{ig}$ to the direct estimator $\hat{Y}_{ig} = N_{ig}\bar{y}_{ig}$, and the weight $1 - w_{ig}$ to the synthetic estimator $\hat{Y}_{ig}^S = N_{ig}\bar{y}_{ig}$. It therefore follows that the BLUP estimator of Y_i tends to the synthetic estimator $\hat{Y}_{\cdot g}^S = \Sigma_g N_{ig}\bar{y}_{\cdot g}$ if the sampling fraction n_{ig}/N_{ig} is negligible for all g, irrespective of the size of between area variation relative to within area variation. This limitation of model (3.8) can be avoided by using more realistic models that include random area-specific effects. We consider such models in Section 4, and we obtain small area estimators under these models in Section 5 using a general EB or a variance components approach as well as a HB procedure.

4. SMALL AREA MODELS

We now consider small area models that include random area-specific effects. Two types of models have been proposed in the literature. In the first type, only area-specific auxiliary data $\mathbf{x}_i = (x_{i1},\ldots,x_{ip})^T$ are available and the parameters of interest, θ_i, are assumed to be related to $\mathbf{x}_i$. In particular, we assume that

$$(4.1) \qquad \theta_i = \mathbf{x}_i^T\beta + v_i z_i, \qquad i = 1,\ldots,m,$$

where the z_i's are known positive constants, β is the vector of regression parameters and the v_i's are independent and identically distributed (iid) random variables with

$$E(v_i) = 0, \qquad V(v_i) = \sigma_v^2.$$

In addition, normality of the random effects v_i is often assumed. In the second type of models, element-specific auxiliary data $\mathbf{x}_{ij} = (x_{ij1},\ldots,x_{ijp})^T$ are available for the population elements, and the variable of interest, y_{ij}, is assumed to be related to $\mathbf{x}_{ij}$ through a nested error regression model:

$$(4.2) \qquad \begin{aligned} y_{ij} &= \mathbf{x}_{ij}^T\beta + v_i + e_{ij}, \\ j &= 1,\ldots,N_i; \qquad i = 1,\ldots,m. \end{aligned}$$

Here $e_{ij} = \bar{e}_{ij}k_{ij}$ and the $\bar{e}_{ij}$'s are iid random variables, independent of the v_i's, with

$$E(\bar{e}_{ij}) = 0, \qquad V(\bar{e}_{ij}) = \sigma^2,$$

the k_{ij}'s being known constants and N_i the number of elements in the ith area. In addition, normality of the v_i's and $\bar{e}_{ij}$'s is often assumed. The parameters of inferential interest here are the small area totals Y_i or the means $\bar{Y}_i = Y_i/N_i$.

For making inferences about the θ_i's under model (4.1), we assume that direct estimators, $\hat{\theta}_i$, are available and that

$$(4.3) \qquad \hat{\theta}_i = \theta_i + e_i, \qquad i = 1,\ldots,m$$

where the e_i's are sampling errors, $E(e_i|\theta_i) = 0$ and $V(e_i|\theta_i) = \psi_i$, that is, the estimators $\hat{\theta}_i$ are design-unbiased. It is also customary to assume that the sampling variances, ψ_i, are known. These assumptions may be quite restrictive in some applications. For example, in the case of adjustment for census underenumeration, the estimates $\hat{\theta}_i$ obtained from a post-enumeration survey (PES) could be seriously biased, as noted by Freedman and Navidi (1986). Simlarly, if θ_i is a nonlinear function of the small area total Y_i and the sample size, n_i is small, then $\hat{\theta}_i$ may be seriously biased even if the direct estimator of Y_i is unbiased. We also assume normality of the $\hat{\theta}_i$'s, but this may not be as restrictive as the normality of the random effects v_i, due to the central limit theorem's effect on the $\hat{\theta}_i$'s.

Combining (4.3) and (4.1), we obtain the model

$$\hat{\theta}_i = \mathbf{x}_i^T\beta + v_i z_i + e_i, \qquad i = 1, \ldots, m \tag{4.4}$$

which is a special case of the general mixed linear model. Note that (4.4) involves design-induced random variables, e_i, as well as model-based random variables v_i.

Turning to the nested error regression model (4.2), we assume that a sample of size n_i is taken from the ith area and that selection bias is absent; that is, the sample values also obey the assumed model. The latter is satisfied under simple random sampling. It may also be noted that model (4.2) may not be appropriate under more complex sampling designs, such as stratified multistage sampling, since the design features are not incorporated. However, it is possible to extend this model to account for such features (see Section 7).

Writing model (4.2) in matrix form as

$$\mathbf{y}_i^P = \mathbf{X}_i^P\beta + v_i\mathbf{1}_i^P + \mathbf{e}_i^P, \tag{4.5}$$

where $\mathbf{X}_i^P$ is $N_i \times p$, $\mathbf{y}_i^P, \mathbf{e}_i^P$ and $\mathbf{1}_i^P$ are $N_i \times 1$ and $\mathbf{1}_i^P = (1, \ldots, 1)^T$, we can partition (4.5) as

$$\mathbf{y}_i^P = \begin{bmatrix} \mathbf{y}_i \\ \mathbf{y}_i^* \end{bmatrix} = \begin{bmatrix} \mathbf{X}_i \\ \mathbf{X}_i^* \end{bmatrix}\beta + v_i \begin{bmatrix} \mathbf{1}_i \\ \mathbf{1}_i^* \end{bmatrix} + \begin{bmatrix} \mathbf{e}_i \\ \mathbf{e}_i^* \end{bmatrix}, \tag{4.6}$$

where the superscript $*$ denotes the nonsampled elements. Now, writing the mean $\bar{Y}_i$ as

$$\bar{Y}_i = f_i\bar{y}_i + (1 - f_i)\bar{y}_i^*, \tag{4.7}$$

with $f_i = n_i/N_i$ and $\bar{y}_i$, $\bar{y}_i^*$ denoting the means for sampled and nonsampled elements respectively, we may view estimation of $\bar{Y}_i$ as equivalent to prediction of $\bar{y}_i^*$ given the data $\{\mathbf{y}_i\}$ and $\{\mathbf{X}_i\}$.

Various extensions of models (4.4) and (4.6), as well as models for binary and Poisson data, have been proposed in the literature. Some of these extensions will be briefly discussed in Section 7.

In the examples given in the Introduction, the models considered are special cases of (4.4) or (4.6). In Example 3, Ericksen and Kadane (1985, 1987) use model (4.4) with $z_i = 1$ and assume σ_v^2 to be known. Here $\hat{\theta}_i$ is a PES estimate of census undercount $\theta_i = \{(T_i - C_i)/T_i\}100$, where T_i is the true (unknown) count and C_i is the census count in the ith area. Cressie (1992) uses (4.4) with $z_i = C_i^{-1/2}$, where $\hat{\theta}_i$ is a PES estimate of the adjustment factor $\theta_i = T_i/C_i$. In Example 2, Fay and Herriot (1979) use (4.4) with $z_i = 1$, where $\hat{\theta}_i$ is a direct estimator of $\theta_i = \log P_i$ and P_i is the average percapita income (PCI) in the ith area. Further, $\mathbf{x}_i^T\beta = \beta_0 + \beta_1 x_i$ with x_i denoting the associated county value of log (PCI) from the 1970 census. In Example 4, Battese, Harter and Fuller (1988) use model (4.6) with $k_{ij} = 1$ and $\mathbf{x}_{ij}^T\beta = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij}$, where y_{ij}, x_{1ij} and x_{2ij} respectively denote the number of hectares of corn (or soybeans), the number of pixels classified as corn and the number of pixels classified as soybeans in the jth area segment of the ith county. A suitable model for our final example is also a special case of (4.6) with $\mathbf{x}_i^T\beta = \beta_0 + \beta_1 x_{ij}$ and $k_{ij} = x_{ij}^{1/2}$, where y_{ij} and x_{ij} respectively denote the total wages and salaries and gross business income for the jth firm in the ith area (census division).

5. EBLUP, EB AND HB APPROACHES

We now present the EBLUP, EB and HB approaches to small area estimation in the context of models (4.4) and (4.6). Both point estimation and measurement of uncertainty associated with the estimators will be studied.

5.1 EBLUP (Variance Components) Approach

As noted in Section 4, most small area models are special cases of a general mixed linear model involving fixed and random effects, and small area parameters can be expressed as linear combinations of these effects. Henderson (1950) derives BLUP estimators of such parameters in the classical frequentist framework. These estimators minimize the mean squared error among the class of linear unbiased estimators and do not depend on normality, similar to the best linear unbiased estimators (BLUEs) of fixed parameters. Robinson (1991) gives an excellent account of BLUP theory and examples of its application.

Under model (4.4), the BLUP estimator of $\theta_i = \mathbf{x}_i^T\beta + v_i z_i$ simplifies to a weighted average of the direct estimator $\hat{\theta}_i$ and the regression-synthetic estimator $\mathbf{x}_i^T\tilde{\beta}$:

$$\tilde{\theta}_i^H = \gamma_i\hat{\theta}_i + (1 - \gamma_i)\mathbf{x}_i^T\tilde{\beta}, \tag{5.1}$$

where the superscript H stands for Henderson,

$$\tilde{\beta} = \left[\sum_i \mathbf{x}_i\mathbf{x}_i^T/(\sigma_v^2 z_i^2 + \psi_i)\right]^{-1} \cdot \left[\sum_{i=1}^m \mathbf{x}_i\hat{\theta}_i/(\sigma_v^2 z_i^2 + \psi_i)\right] \tag{5.2}$$

is the BLUE estimator of β and

$$\gamma_i = \sigma_v^2 z_i^2/(\sigma_v^2 z_i^2 + \psi_i).$$

The weight γ_i measures the uncertainty in modelling the θ_is, namely, $\sigma_v^2 z_i^2$ relative to the total variance $\sigma_v^2 z_i^2 + \psi_i$. Thus, the BLUP estimator takes proper account of between area variation relative to the precision of the direct estimator. It is valid for general sampling designs since we are modelling only the θ_i's and not the individual elements in the population. It is also design consistent since $\gamma_i \to 1$ as the sampling variance $\psi_i \to 0$.

The mean squared error (MSE) of $\tilde{\theta}_i^H$ under model (4.4) may be written as

$$M_{1i}(\sigma_v^2) = E(\tilde{\theta}_i^H - \theta_i)^2 = g_{1i}(\sigma_v^2) + g_{2i}(\sigma_v^2),$$

where

$$g_{1i}(\sigma_v^2) = \sigma_v^2 z_i^2 \psi_i (\sigma_v^2 z_i^2 + \psi_i)^{-1} = \gamma_i \psi_i$$

and

$$g_{2i}(\sigma_v^2) = (1-\gamma_i)^2 \mathbf{x}_i^T \left[\sum_i \mathbf{x}_i \mathbf{x}_i^T / (\sigma_v^2 z_i^2 + \psi_i) \right]^{-1} \mathbf{x}_i.$$

The first term $g_{1i}(\sigma_v^2)$ is of order $O(1)$ while the second term $g_{2i}(\sigma_v^2)$, due to estimating β, is of order $O(m^{-1})$ for large m.

The BLUP estimator (5.1) depends on the variance component σ_v^2 which is unknown in practical applications. However, various methods of estimating variance components in a general mixed linear model are available, including the method of fitting constants or moments, maximum likelihood (ML) and restricted maximum likelihood (REML). Cressie (1992) gives a succinct account of these methods in the context of model (4.4). All these methods yield asymptotically consistent estimators under realistic regularity conditions.

Replacing σ_v^2 with an asymptotically consistent estimator $\hat{\sigma}_v^2$, we obtain a two-stage estimator, $\hat{\theta}_i^H$, which is referred to as the empirical BLUP or EBLUP estimator (Harville, 1991), in analogy with the EB estimator. It remains unbiased provided (i) the distributions of v_i and e_i are both symmetric (not necessarily normal); (ii) $\hat{\sigma}_v^2$ is an even function of $\hat{\theta}_i$'s and remains invariant when $\hat{\theta}_i$ is changed to $\hat{\theta}_i - \mathbf{x}_i^T \mathbf{a}$ for all $\mathbf{a}$ (Kackar and Harville, 1984). Standard methods of estimating variance components all satisfy (ii). We may also point out that the MSE of the EBLUP estimator appears to be insensitive to the choice of the estimator $\hat{\sigma}_v^2$.

If normality of the errors v_i also holds, then we can write the MSE of $\hat{\theta}_i^H$ as

$$M_{2i}(\sigma_v^2) = M_{1i}(\sigma_v^2) + E(\hat{\theta}_i^H - \tilde{\theta}_i^H)^2, \tag{5.3}$$

see Kackar and Harville (1984). It follows from (5.3) that the MSE of $\hat{\theta}_i^H$ is always larger than that of the BLUP estimator $\tilde{\theta}_i^H$. The second term of (5.3) is not tractable, unlike the first term $M_{1i}(\sigma_v^2)$; but it can be approximated for large m (Kackar and Harville, 1984; Prasad and Rao, 1990; Cressie, 1992). We have, for large m,

$$E(\hat{\theta}_i^H - \tilde{\theta}_i^H)^2 \doteq g_{3i}(\sigma_v^2) \tag{5.4}$$

where

$$g_{3i}(\sigma_v^2) = \psi_i^2 z_i^4 (\sigma_v^2 z_i^2 + \psi_i)^{-3} \overline{V}(\hat{\sigma}_v^2),$$

and the neglected terms in the approximation (5.4) are of lower order than $O(m^{-1})$. Here $\overline{V}(\hat{\sigma}_v^2)$ denotes the asymptotic variance of $\hat{\sigma}_v^2$; Cressie (1992) gives the asymptotic variance formulae for ML and REML estimators. It is customary to ignore the uncertainty in $\hat{\sigma}_v^2$ and use $M_{1i}(\hat{\sigma}_v^2) = g_{1i}(\hat{\sigma}_v^2) + g_{2i}(\hat{\sigma}_v^2)$ as an estimator of MSE of $\hat{\theta}_i^H$, but this procedure could lead to severe underestimation of the true MSE. A correct, approximately unbiased estimator of MSE $(\hat{\theta}_i^H)$ is given by

$$\text{mse}(\hat{\theta}_i^H) = g_{1i}(\hat{\sigma}_v^2) + g_{2i}(\hat{\sigma}_v^2) + 2g_{3i}(\hat{\sigma}_v^2), \tag{5.5}$$

(see Prasad and Rao, 1990). The bias of (5.5) is of lower order than m^{-1}.

Noting that $E[\Sigma_i (y_i - \mathbf{x}_i^T \tilde{\beta})^2 / (\sigma_v^2 z_i + \psi_i)] = m - p$, a method of moments estimator $\tilde{\sigma}_v^2$ can be obtained by solving iteratively

$$\sum_{i=1}^{m} (y_i - x_i^T \tilde{\beta})^2 / (\sigma_v^2 z_i + \psi_i) = m - p$$

in conjunction with (5.2) and letting $\tilde{\sigma}_v^2 = 0$ when no positive solution exists (Fay and Herriot, 1979). This method does not require normality, unlike the ML and REML. Alternatively, a simple moment estimator is given by $\tilde{\sigma}_v^2 = \max(\tilde{\sigma}_v^2, 0)$, where

$$\tilde{\sigma}_v^2 = (t-p)^{-1} \left[\sum_i \frac{1}{z_i^2} (y_i - \mathbf{x}_i^T \beta^*)^2 - \sum_i \frac{\psi_i}{z_i^2} \left\{ 1 - \mathbf{x}_i^T \left(\sum_i \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \mathbf{x}_i \right\} \right] \tag{5.6}$$

and $\beta^* = (\Sigma_i \mathbf{x}_i \mathbf{x}_i^T)^{-1} (\Sigma_i \mathbf{x}_i \hat{\theta}_i)$ is the ordinary least squares estimator of β. The estimator $\tilde{\sigma}_v^2$ is unbiased for σ_v^2 and under normality,

$$\overline{V}(\hat{\sigma}_v^2) = \overline{V}(\tilde{\sigma}_v^2) = 2t^{-2} \sum_i (\sigma_v^2 + \psi_i / z_i^2)^2$$

(see Prasad and Rao, 1990 for the case $z_i = 1$).

Lahiri and Rao (1992) show that the estimator of MSE, (5.5), using the moment estimator (5.6), is also valid under moderate nonnormality of the random effects, v_i. Thus, inference based on $\hat{\theta}_i^H$ and $\text{mse}(\hat{\theta}_i^H)$ is robust to nonnormality of the random effects.

We next turn to the nested error regression model (4.6). The BLUP estimator of $\overline{Y}_i$ in this case is obtained as follows: (i) using the model $\mathbf{y}_i = \mathbf{X}_i + v_i \mathbf{1}_{n_i} + \mathbf{e}_i$ for the sampled elements, obtain the BLUP estimator of $\overline{\mathbf{X}}_i^{*T}\beta + v_i$, where $\overline{\mathbf{X}}_i^*$ is the mean for nonsampled elements; (ii) substitute this estimator for $\bar{y}_i^*$ in (4.7). Thus the BLUP estimator of $\overline{Y}_i$ is given by

$$(5.7)\quad \tilde{\overline{Y}}_i^H = f_i \bar{y}_i + (1 - f_i)\left[\overline{X}_i^{*T}\tilde{\beta} + \gamma_i(\bar{y}_{iw} - \bar{\mathbf{x}}_{iw}^T\tilde{\beta})\right],$$

where $\tilde{\beta}$ is the BLUE of β,

$$\gamma_i = \sigma_v^2(\sigma_v^2 + \sigma^2/w_{i\cdot})^{-1}$$

with $w_{i\cdot} = \sum_{j=1}^{n_i} w_{ij}$ and $w_{ij} = k_{ij}^{-2}$, and $\bar{y}_{iw}$ and $\bar{\mathbf{x}}_{iw}$ are the weighted means with weights w_{ij} (see Prasad and Rao, 1990, and Stukel, 1991). The BLUE $\tilde{\beta}$ is readily obtained by applying ordinary least squares to the transformed data $\{(y_{ij} - \gamma_i \bar{y}_{iw})/k_{ij}, (\mathbf{x}_{ij} - \gamma_i \bar{\mathbf{x}}_{iw})/k_{ij}\}$ (see Stukel, 1991, and Fuller and Battese, 1973). If the sample fraction f_i is negligible, we can write $\tilde{\overline{Y}}_i^H$ as a composite estimator of the form

$$(5.8)\quad \tilde{\overline{Y}}_i^H \doteq \gamma_i[\bar{y}_{iw} + (\overline{\mathbf{X}}_i - \bar{\mathbf{x}}_{iw})^T\tilde{\beta}] + (1 - \gamma_i)\overline{X}_i^T\tilde{\beta},$$

where $\overline{\mathbf{X}}_i$ is the ith area population mean of $\mathbf{x}_{ij}$'s. It follows from (5.8) that the BLUP estimator is a weighted average of the "survey regression" estimator $\bar{y}_{iw} + (\overline{\mathbf{X}}_i - \bar{\mathbf{x}}_{iw})^T\tilde{\beta}$ and the regression synthetic estimator $\overline{\mathbf{X}}_i^T\tilde{\beta}$. If $k_{ij} = 1$ for all (ij), then the survey regression estimator is approximately design-unbiased for $\overline{Y}_i$ under simple random sampling even if n_i is small. In the case of general k_{ij}'s, it is model-unbiased conditional on the realized local effect v_i, unlike the BLUP estimator which is conditionally biased.

An empirical BLUP estimator, $\hat{\overline{Y}}_i^H$, is obtained from (5.7) by replacing (σ_v^2, σ^2) with asymptotically consistent estimators $(\hat{\sigma}_v^2, \hat{\sigma}^2)$. Further, assuming normality of the errors an approximately unbiased estimator of MSE $(\hat{\overline{Y}}_i^H)$, similar to (5.5) under model (4.4), is given by

$$(5.9)\quad \text{mse}(\hat{\overline{Y}}_i^H) = (1 - f_i)^2\left[g_{1i}(\hat{\sigma}_v^2, \hat{\sigma}^2) + g_{2i}(\hat{\sigma}_v^2, \hat{\sigma}^2) + 2g_{3i}(\hat{\sigma}_v^2, \hat{\sigma}^2)\right].$$

Here

$$g_{1i}(\sigma_v^2, \sigma^2) = \gamma_i(\sigma^2/w_{i\cdot}) + (1 - f_i)^2 N_i^{-2}\mathbf{k}_i^{*T}\mathbf{k}_i^*$$

with $\mathbf{k}_i^*$ denoting the vector of k_{ij}'s for nonsampled units in ith area, and

$$g_{2i}(\sigma_v^2, \sigma^2) = (\bar{\mathbf{x}}_i^* - \gamma_i \bar{\mathbf{x}}_{iw})^T\mathbf{A}^{-1}(\bar{\mathbf{x}}_i^* - \gamma_i \bar{\mathbf{x}}_{iw})\sigma^2$$

with

$$\mathbf{A} = \sum_{i=1}^{m}\left[\sum_{j=1}^{n_i} w_{ij}\mathbf{x}_{ij}\mathbf{x}_{ij}^T - \gamma_i w_{i\cdot}\bar{\mathbf{x}}_{iw}\bar{\mathbf{x}}_{iw}^T\right].$$

Further,

$$g_{3i}(\sigma_v^2, \sigma^2) = w_{i\cdot}^{-2}(\sigma_v^2 + \sigma^2/w_{i\cdot})^{-3}\left[\sigma^2\overline{V}(\hat{\sigma}_v^2) + \sigma_v^2\overline{V}(\hat{\sigma}_e^2) - 2\sigma^2\sigma_v^2\overline{\text{cov}}(\hat{\sigma}_v^2, \hat{\sigma}_e^2)\right].$$

where $\overline{\text{cov}}$ denotes the asymptotic covariance (see Stukel, 1991 and Prasad and Rao, 1990).

For the ML and REML methods, the asymptotic covariance matrix of $(\hat{\sigma}_v^2, \hat{\sigma}^2)$ can be obtained from general theory (see, e.g., Cressie, 1992). Stukel (1991) and Fuller and Battese (1973) use the method of fitting constants which involves two ordinary least square fittings: first, we calculate the residual sum of squares, SSE(1), with ν_1 degrees of freedom by regressing through the origin the y-deviations $k_{ij}^{-1}(y_{ij} - \bar{y}_{iw})$ on the nonzero x-deviations $k_{ij}^{-1}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_{iw})$ for these areas with $n_i > 1$. Second, we calculate the residual sum of squares SSE(2) by regressing y_{ij}/k_{ij} on $\mathbf{x}_{ij}/k_{ij}$. Then $\hat{\sigma}^2 = \nu_1^{-1}$ SSE(1) and $\hat{\sigma}_v^2 = \max(\tilde{\sigma}_v^2, 0)$ with

$$\tilde{\sigma}_v^2 = \eta_{**}^{-1}[SSE(2) - (n - p)\hat{\sigma}^2],$$

where

$$\eta^* = \sum_i w_{i\cdot}(1 - w_{i\cdot}\bar{\mathbf{x}}_{iw}^T\mathbf{A}_1^{-1}\bar{\mathbf{x}}_{iw})$$

with

$$\mathbf{A}_1 = \sum_i \sum_j w_{ij}\mathbf{x}_{ij}\mathbf{x}_{ij}^T.$$

The Appendix gives the variances and covariance of $\hat{\sigma}^2$ and $\hat{\sigma}_v^2$.

Again, ignoring the uncertainty in $\hat{\sigma}_v^2$ and $\hat{\sigma}^2$ and using $M_{1i}(\hat{\sigma}_v^2, \hat{\sigma}^2) = g_{1i}(\hat{\sigma}_v^2, \hat{\sigma}^2)$ as an estimator of MSE $(\hat{\overline{Y}}_i^H)$ could lead to severe underestimation of the true MSE.

Limited simulation results (Prasad and Rao, 1990; Datta and Ghosh, 1991 and Hulting and

Harville, 1991) indicate that the estimator of MSE, mse $(\widehat{\bar{Y}}_i^H)$, given by (5.9), performs well even for moderate m (as small as 15), provided σ_v^2/σ^2 is not close to zero.

5.2 EB Approach

In the EB approach, the posterior distribution of the parameters of interest given the data is first obtained, assuming that the model parameters are known. The model parameters are estimated from the marginal distribution of the data, and inferences are then based on the estimated posterior distribution. Morris (1983) gives an excellent account of the EB approach and significant applications.

Under model (4.4) with normal errors, the posterior distribution of θ_i given $\hat{\theta}_i, \beta$ and σ_v^2 is normal with mean θ_i^B and variance $g_{1i}(\sigma_v^2) = \gamma_i\psi_i$, where

$$\theta_i^B = E(\theta_i|\hat{\theta}_i, \beta, \sigma_v^2) = \gamma_i\hat{\theta}_i + (1-\gamma_i)\mathbf{x}_i^T\beta.$$

Under quadratic loss, θ_i^B is the Bayes estimator of θ_i. Noting that the $\hat{\theta}_i \sim N(\mathbf{x}_i^T\beta,\ \sigma_v^2 z_i^2 + \psi_i)$ are marginally independent, we can obtain the estimators $\hat{\sigma}_v^2$ and $\hat{\beta}$ as before using ML, REML or the method of moments. The estimated posterior distribution is $N(\hat{\theta}_i^{EB}, g_{1i}(\hat{\sigma}_v^2))$, where $\hat{\theta}_i^{EB}$ is identical to the EBLUP estimator $\hat{\theta}_i^H$. A naive EB approach uses $\hat{\theta}_i^{EB}$ as the estimator of θ_i and measures its uncertainty by the estimated posterior variance

$$V(\theta_i|\hat{\theta}_i, \hat{\beta}, \hat{\sigma}_v^2) = g_{1i}(\hat{\sigma}_v^2). \tag{5.10}$$

This can lead to severe underestimation of the true posterior variance $V(\theta_i|\hat{\theta})$ (under a prior distribution on β and σ_v^2), although $\hat{\theta}_i^{EB} = E(\theta_i|\hat{\theta}_i, \hat{\beta}, \hat{\sigma}_v^2)$ is approximately equal to the true posterior mean $E(\theta_i|\hat{\theta})$, where $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_m)^T$.

The above point is better understood when one writes

$$E(\theta_i|\hat{\theta}) = E_{\beta,\sigma_v^2}[E(\theta_i|\hat{\theta}_i, \beta, \sigma_v^2)]$$

and

$$\begin{aligned} V(\theta_i|\hat{\theta}) &= E_{\beta,\sigma_v^2}[V(\theta_i|\hat{\theta}_i, \beta, \sigma_v^2)] \\ &\quad + V_{\beta,\sigma_v^2}[E(\theta_i|\hat{\theta}_i, \beta, \sigma_v^2)], \end{aligned} \tag{5.11}$$

where E_{β,σ_v^2} and V_{β,σ_v^2} respectively denote the expectation and variance with respect to the posterior distribution of β and σ_v^2 given the data $\hat{\theta}$. It follows from (5.11) that (5.10) is a good approximation only to the first variance term on the right side of (5.11), but the second variance term is ignored in the naive EB approach, that is, it fails to take account of the uncertainty about the parameters β and σ_v^2. Note that the form of the prior distribution on β and σ^2 is not specified in the EB approach, unlike in the HB approach (Section 5.3).

Two methods of accounting for the underestimation of true posterior variance have been proposed in the literature. The first method is based on the bootstrap (Laird and Louis, 1987), while the second method uses an asymptotic approximation to the posterior variance $V(\theta_i|\hat{\theta})$ irrespective of the form of the prior on β and σ_v^2 (Kass and Steffey, 1989). In the bootstrap method, a large number, B, of independent bootstrap samples $\{\theta_1^*(b), \ldots, \theta_m^*(b);\ b = 1, \ldots, B\}$ are first drawn, where $\theta_i^*(b)$ is drawn from the estimated marginal distribution $N(\mathbf{x}_i^T\hat{\beta}, \hat{\sigma}_v^2 z_i^2 + \psi_i)$. Estimates $\beta^*(b)$ and $\sigma_v^{*2}(b)$ are then computed from the bootstrap data $\{\theta_i^*(b), \mathbf{x}_i,\ i = 1, \ldots, m\}$ for each b. The EB bootstrap estimator of θ_i is given by

$$\begin{aligned} \theta_i^*(\cdot) &= \frac{1}{B}\sum_{b=1}^{B} E[\theta_i|\theta^*(b), \beta^*(b), \sigma_v^{*2}(b)] \\ &= \frac{1}{B}\sum_{b=1}^{B} \theta_i^{*EB}(b), \end{aligned}$$

and its uncertainty is measured by

$$\begin{aligned} V_i^* &= \frac{1}{B}\sum_{b=1}^{B} V[\theta_i|\theta_i^*(b), \beta^*(b), \sigma_v^{*2}(b)] \\ &\quad + \frac{1}{B-1}\sum_{b=1}^{B}[\theta_i^{*EB}(b) - \theta_i^{*EB}(\cdot)]^2. \end{aligned} \tag{5.12}$$

The second term on the right side of (5.12) accounts for the underestimation. The EB bootstrap method looks promising, but further studies on its frequentist performance are needed.

In the Kass-Steffey method, $\hat{\theta}_i^{EB}$ is taken as the estimator of θ_i, but a positive correction term is added to the estimated posterior variance $V(\theta_i|\hat{\theta}_i, \hat{\beta}, \hat{\sigma}_v^2)$ to account for the underestimation. This term depends on the observed information matrix and the partial derivatives of θ_i^B, evaluated at the ML estimates $\hat{\beta}$ and $\hat{\sigma}_v^2$. This method also looks promising, but its frequentist properties remain to be investigated. (Steffey and Kass, 1991 conjecture that the MSE of EB estimator is approximately equal to their approximation to the posterior variance.) Kass and Steffey (1989) also give an improved second-order approximation to the true posterior variance, $V(\theta_i|\hat{\theta})$.

Turning to the nested error regression model (4.6), the estimated posterior distribution of $\bar{Y}_i$ given the data $\mathbf{y}$ is normal with mean equal to the EBLUP $\widehat{\bar{Y}}_i^H$ and variance equal to $(1-f_i)^2 g_{1i}(\hat{\sigma}_v^2, \hat{\sigma}^2)$ which

is a severe underestimate of the true posterior variance $V(\bar{Y}_i|\mathbf{y})$. Again, the bootstrap and Kass-Steffey methods can be applied to account for the underestimation.

If one wishes to view the EB approach in the frequentist framework, a prior distribution on β and σ_v^2 cannot be entertained. In this case, MSE is a natural measure of uncertainty and any differences between the EB and EBLUP approaches disappear under the normality assumption. It may also be noted that the EB estimator can be justified without the normality assumption, similar to the EBLUP, using the "posterior linearity" property (Ghosh and Lahiri, 1987; Ericson, 1969).

5.3 HB Approach

In the HB approach, a prior distribution on the model parameters is specified and the posterior distribution of the parameters of interest is then obtained. Inferences are based on the posterior distribution; in particular, a parameter of interest is estimated by its posterior mean and its precision is measured by its posterior variance. The HB approach is straightforward and clear–cut but computationally intensive, often involving high dimensional integration. Recent advances in computational aspects of the HB approach, such as Gibbs sampling (cf. Gelfand and Smith, 1990) and importance sampling, however, seem to overcome the computational difficulties to a large extent. If the solution involves only one or two dimensional integration, it is often easier to perform direct numerical integration than to use Gibbs sampling or any other Monte Carlo numerical integration method. Datta and Ghosh (1991) apply the HB approach to estimation of small area means, $\bar{Y}_i$, under general mixed linear models, and also discuss the computational aspects.

We now illustrate the HB approach under our models (4.4) and (4.6), assuming noninformative priors on β and the variance components σ_v^2 and σ^2. The HB approach, however, can incorporate prior information on these parameters through informative priors.

Under model (4.4), we first obtain the posterior distribution of θ_i given $\hat{\theta}$ and σ_v^2, by assuming that β has a uniform distribution over R^p to reflect absence of prior information on β. Straightforward calculations show that it is normal with mean equal to the BLUP estimator $\tilde{\theta}_i^H$ and variance equal to $M_{1i}(\sigma_v^2)$, the MSE of $\tilde{\theta}_i^H$, that is, $E(\theta_i|\hat{\theta},\sigma_v^2) = \tilde{\theta}_i^H$ and $V(\theta_i|\hat{\theta},\sigma_v^2) = \text{MSE}\,(\tilde{\theta}_i^H)$. Hence, when σ_v^2 is assumed to be known, the HB and BLUP approaches lead to identical inferences.

To take account of the uncertainty about σ_v^2, we need to calculate the posterior distribution of σ_v^2 given $\hat{\theta}$ under a suitable prior on σ_v^2. The posterior mean and variance of θ_i are then given by

$$E(\theta_i|\hat{\theta}) \equiv E_{\sigma_v^2}(\tilde{\theta}_i^H) \tag{5.13}$$

and

$$V(\theta_i|\hat{\theta}) = E_{\sigma_v^2}[M_{1i}(\sigma_v^2)] + V_{\sigma_v^2}(\tilde{\theta}_i^H), \tag{5.14}$$

where $E_{\sigma_v^2}$ and $V_{\sigma_v^2}$ respectively denote the expectation and variance with respect to the posterior distribution of σ_v^2 given $\hat{\theta}$. Numerical evaluation of (5.13) and (5.14) involves one dimensional integration. Ghosh (1992) obtains the posterior distribution, $f(\sigma_v^2|\hat{\theta})$, assuming that σ_v^2 has a uniform distribution over $(0,\infty)$ to reflect the absence of prior information about σ_v^2, and that σ_v^2 and β are independently distributed. It is given by

$$f(\sigma_v^2|\hat{\theta}) = (\sigma_v^2)^{-\frac{(m-p)}{2}} \left\{\prod_1^m \gamma_i^{1/2}\right\} \left|\sum_i \gamma_i \mathbf{x}_i \mathbf{x}_i^T\right|^{-\frac{1}{2}} \cdot \exp\left[-\frac{1}{2} Q_a(\hat{\theta})\right],$$

where

$$Q_a(\hat{\theta}) = (\sigma_v^2)^{-1}\left[\sum_i \gamma_i \hat{\theta}_i^2 - \left(\sum_i \gamma_i \hat{\theta}_i \mathbf{x}_i\right)^T \cdot \left(\sum_i \gamma_i \mathbf{x}_i \mathbf{x}_i^T\right)^{-1} \left(\sum_i \gamma_i \hat{\theta}_i \mathbf{x}_i\right)\right].$$

We next turn to the nested error regression model (4.6). We first obtain the posterior distribution of $\bar{Y}_i$ given $\mathbf{y}$, σ_v^2 and σ^2, by assuming that β has uniform distribution over R^p. Straightforward calculations show that it is normal with mean equal to the BLUP estimator $\tilde{\bar{Y}}_i^H$ and variance equal to MSE $(\tilde{\bar{Y}}_i^H) = M_{1i}(\sigma_v^2, \sigma^2)$, that is, $E(\bar{Y}_i|\mathbf{y}, \sigma_v^2, \sigma^2) = \tilde{\bar{Y}}_i^H$ and $V(\bar{Y}_i|\mathbf{y}, \sigma_v^2, \sigma^2) = \text{MSE}(\tilde{\bar{Y}}_i^H)$. Hence, when both σ_v^2 and σ^2 are assumed to be known, the HB and BLUP approaches lead to identical inferences.

To take account of the uncertainty about σ_v^2 and σ^2, Datta and Ghosh (1991) further assume $\beta, (\sigma^2)^{-1}$ and $(\sigma_v^2)^{-1} = (\sigma^2)^{-1}\lambda$ to be independently distributed with $(\sigma^2)^{-1} \sim$ gamma $((1/2)a_0, (1/2)g_0)$ and $(\sigma^2)^{-1}\lambda \sim$ gamma $((1/2)a_1, (1/2)g_1)$, where $a_0 \geq 0$, $g_0 \geq 0$, $a_1 > 0$, $g_1 \geq 0$ and $\lambda = \sigma^2/\sigma_v^2$. Here gamma (α, β) denotes the gamma random variable with pdf $f(z) = \exp(-\alpha z)\alpha^\beta z^{\beta-1}/\Gamma(\beta)$, $z > 0$. Datta and Ghosh (1991) obtain closed form expressions for $E(\bar{Y}_i|\mathbf{y}, \lambda)$

and $V(\overline{Y}_i|\mathbf{y}, \lambda)$ by showing that $f(\mathbf{y}^*|\mathbf{y}, \lambda)$ is a multivariate t-distribution. They also derive the posterior distribution of λ given $\mathbf{y}$, but it has a complex structure making it necessary to perform one-dimensional numerical integration to get $E(\overline{Y}_i|\mathbf{y})$ and $V(\overline{Y}_i|\mathbf{y})$ using the following relationships:

$$E(\overline{Y}_i|\mathbf{y}) = E_\lambda[E(\overline{Y}_i|\mathbf{y}, \lambda)]$$

and

$$V(\overline{Y}_i|\mathbf{y}) = E_\lambda[V(\overline{Y}_i|\mathbf{y}, \lambda)] + V_\lambda[E(\overline{Y}_i|\mathbf{y}, \lambda)],$$

where E_λ and V_λ respectively denote the expectation and variance under the posterior distribution of λ given the data $\mathbf{y}$.

Datta and Ghosh (1991) compare the HB, EB and EBLUP approaches using the data for our example 4 and letting $a_0 = a_1 = 0.005$ and $g_0 = g_1 = 0$ to reflect the absence of prior information on σ_v^2 and σ^2. As one might expect, the three estimates were close to each other as point predictors of small area (county) means; the EB estimate was obtained by replacing λ with the method-of-fitting constants estimate $\hat{\lambda}$ in $E(\overline{Y}_i|\mathbf{y}, \lambda)$. The naive variance estimate, $V(\overline{Y}_i|\mathbf{y}, \hat{\lambda}) = (s_i^{EB})^2$ associated with the EB estimate $E(\overline{Y}_i|\mathbf{y}, \hat{\lambda})$, was always found to be smaller than the true posterior variance, $V(\overline{Y}_i|\mathbf{y}) = (s_i^{HB})^2$, associated with the HB estimate $\widehat{\overline{Y}}_i^{HB} = E(\overline{Y}_i|\mathbf{y})$; for one county, s_i^{EB} was about 10% smaller than s_i^{HB}. Note that the customary naive EB variance estimate, $V(\overline{Y}_i|\mathbf{y}, \hat{\beta}, \hat{\sigma}_v^2, \hat{\sigma}^2)$, will lead to much more severe underestimation than $V(\overline{Y}_i|\mathbf{y}, \hat{\lambda})$ since the latter takes account of the uncertainty about β and σ^2. The estimated MSE, $\text{mse}(\widehat{\overline{Y}}_i^H) = (s_i^H)^2$, associated with the EBLUP estimate, $\widehat{\overline{Y}}_i^H$, was found to be similar to the HB variance estimate. Our example in Section 6 also gives similar results. Datta and Ghosh (1991) have also conducted a simulation study on the frequentist properties of the HB and EBLUP methods using the Battese, Harter and Fuller (1988) model. Their findings indicate that the simulated MSEs for the HB estimator are very close to those for the EBLUP estimator while the coverage probabilities based on $\widehat{\overline{Y}}_i^{HB} \pm (1.96)s_i^{HB}$ turn out to be slightly bigger than those based on $\widehat{\overline{Y}}_i^H \pm (1.96)s^H$, both being close to nominal confidence level of 95%. Hulting and Harville (1991) obtain similar results in another simulation study using the Battese, Harter and Fuller (1988) model and varying the variance ratio σ_v^2/σ^2. However, they find the HB method produces different and more sensible answers than the EBLUP procedure if the estimate for σ_v^2/σ^2 is zero or close to zero.

The HB approach looks promising, but we need to study its robustness to choice of prior distributions on the model parameters.

6. EXAMPLE

Several of the proposed small area estimators are now compared on the basis of their squared errors and relative errors from the true small area means $\overline{Y}_i$. For this purpose, we first constructed a synthetic population of pairs (y_{ij}, x_{ij}) resembling the business population studied by Särndal and Hidiroglou (1989) where the census divisions are small areas, y_{ij} denotes wages and salaries of jth firm in the ith census division and x_{ij} the corresponding gross business income. To generate the synthetic population, we fitted the nested error regression model (4.6) with $\mathbf{x}_{ij}^T\beta = \beta_0 + \beta_1 x_{ij}$ and $k_{ij} = x_{ij}^{1/2}$ to a real population to estimate β_0 and β_1 and the variance components σ_v^2 and σ^2. The resulting synthetic model is given by

$$\begin{aligned} y_{ij} &= -2.47 + 0.20x_{ij} + v_i + e_{ij}, \\ j &= 1, \ldots, N_i,\ i = 1, \ldots, m, \\ v_i &\overset{\text{i.i.d}}{\sim} N(0, 22.14), \\ e_{ij} &\overset{\text{i.i.d}}{\sim} N(0, 0.47x_{ij}). \end{aligned} \tag{6.1}$$

We then used model (6.1) in conjunction with the population x_{ij}-values to generate a synthetic population of pairs (y_{ij}, x_{ij}) with $m = 16$ small areas. Table 1 reports the small area population sizes, N_i, and the small area means $(\overline{Y}_i, \overline{X}_i)$ for this synthetic population of size $N = 114$. A simple random sample of size $n = 38$ was drawn from the synthetic population. The resulting small area sample sizes, n_i, and sample data (y_{ij}, x_{ij}) are reported in Table 2. Note that direct estimators cannot be implemented for areas 1, 4 and 13 since $n_i = 0$ for these areas. We have, therefore, confined ourselves to the following indirect estimators valid for all $n_i \geq 0$:

(i) Ratio-synthetic estimator: $\widehat{\overline{Y}}_i^{RS} = (\bar{y}/\bar{x})\overline{X}_i$, where $(\bar{y}, \bar{x})$ are the overall sample means.

(ii) Sample-size dependent estimator:

$$\widehat{\overline{Y}}_i^{SD} = \begin{cases} \widehat{\overline{Y}}_i^{\text{REG}} = \bar{y}_i + (\bar{y}/\bar{x})(\overline{X}_i - \bar{x}_i), & \text{if } w_i \geq W_i, \\ \frac{w_i}{W_i}(\widehat{\overline{Y}}_i^{\text{REG}}) + \left(1 - \frac{w_i}{W_i}\right)\widehat{\overline{Y}}_i^{RS}, & \text{if } w_i < W_i, \end{cases}$$

where $\widehat{\overline{Y}}_i^{\text{REG}}$ is a "survey regression" estimator, $(\bar{y}_i, \bar{x}_i)$ are the sample means, $w_i = n_i/n$ and $W_i = N_i/N$. This estimator corresponds to the weight (3.6) with $\delta = 1$ or the weight

TABLE 1

Small area sizes, N_i, and means $(\bar{Y}_i, \bar{X}_i)$ for a synthetic population ($N = 114$)

Area No.	N_i	$\bar{X}_i$	$\bar{Y}_i$	Area No.	N_i	$\bar{X}_i$	$\bar{Y}_i$
1	1	137.70	24.22	9	27	97.58	15.56
2	6	100.84	20.43	10	5	76.04	5.88
3	4	47.72	5.48	11	12	90.15	15.20
4	1	45.64	6.55	12	7	86.24	13.40
5	8	108.53	20.55	13	4	164.28	26.06
6	6	65.68	14.85	14	6	164.70	22.44
7	6	116.34	21.46	15	13	83.86	9.40
8	6	92.74	13.40	16	2	134.49	29.49

TABLE 2

Data from a simple random sample drawn from a synthetic population ($n = 38, N = 114$)

Area No.	n_i	x_{ij}	y_{ij}	Area No.	n_i	x_{ij}	x_{ij}
1	0	—	—	9	10	333.24	47.62
						80.91	5.27
2	3	33.70	5.90			43.65	6.97
		47.19	13.22			29.29	−0.19
		75.21	17.44			102.66	15.94
						109.34	19.84
3	1	36.43	2.54			30.56	2.57
						127.96	24.61
4	0	—	—			190.34	35.41
						52.16	2.54
5	1	28.82	3.61	10	1	45.91	−6.34
				11	2	43.03	8.83
6	2	30.60	11.48			190.12	27.31
		129.69	21.45	12	1	47.39	1.70
				13	0	—	—
7	4	200.60	46.96	14	3	35.66	−0.80
		113.92	15.57			40.23	2.75
		74.33	8.66			111.23	10.87
		53.00	11.90	15	6	51.61	−3.20
						67.46	12.47
8	3	95.43	11.76			190.97	21.77
		35.75	−0.69			35.11	2.92
		39.08	21.46			25.09	−5.46
						73.51	7.35
				16	1	229.32	53.83

(3.7) with $h = 2$. We have not included the optimal composite estimator due to difficulties in estimating the optimal weight (3.4).

(iii) EBLUP (or EB) estimator $\hat{\bar{Y}}_i^H$ under model (4.6) with $\mathbf{x}_{ij}^T\beta = \beta_0 + \beta_1 x_{ij}$ and $k_{ij} = x_{ij}^{1/2}$, where σ_v^2 and σ^2 are estimated by the method of fitting constants.

(iv) HB estimator $\hat{\bar{Y}}_i^{HB}$ under model (4.6) as in (iii), using Datta-Ghosh's diffuse priors with $a_0 = 0, g_0 = 0, a_1 = 0.05$ and $g_1 = 0$.

Using the sample data (y_{ij}, x_{ij}) and the known small area population means $\bar{X}_i$ we computed the above four estimates along with their average relative errors

$$\text{ARE} = \frac{1}{m}\sum_{i=1}^{m} |\text{est.} - \bar{Y}_i|/\bar{Y}_i$$

and average squared errors

$$\text{ASE} = \frac{1}{m}\sum_{i=1}^{m} (\text{est.} - \bar{Y}_i)^2.$$

These values are reported in Table 3. We also calculated the standard error, s_i^H, of EBLUP estimator using (5.9) and the posterior standard deviation

TABLE 3

Small area estimates and their (%) average relative errors and average squared roots; standard error (S.E.) of EBLUP and HB estimators

Area No.	n_i	$\bar{Y}_i$	RS	SD	EBLUP	HB	S.E. EBLUP	S.E. HB
1	0	24.22	19.79	19.79	22.16	22.16	7.40	8.29
2	3	20.43	14.90	19.20	20.47	20.18	2.20	2.47
3	1	5.48	6.86	5.34	4.85	4.87	2.62	2.60
4	0	6.55	6.56	6.56	4.97	4.94	5.40	5.99
5	1	20.55	15.60	15.52	17.98	17.81	3.10	3.17
6	2	14.85	9.44	14.39	13.99	13.47	2.07	2.40
7	4	21.46	16.72	21.62	21.31	21.22	1.59	1.74
8	3	13.40	13.33	11.22	11.44	11.58	1.86	2.00
9	10	15.56	14.02	14.27	13.95	13.98	1.14	1.22
10	1	5.88	10.93	6.27	3.30	3.96	3.06	3.63
11	2	15.20	12.96	13.29	14.66	14.44	2.61	2.57
12	1	13.40	12.11	11.17	9.97	10.17	3.14	3.14
13	0	26.06	23.61	23.61	27.13	27.13	5.52	6.13
14	3	22.44	23.67	18.98	24.05	24.22	3.10	3.48
15	6	9.40	12.05	7.40	8.24	8.43	1.32	1.50
16	1	29.49	19.33	40.20	30.31	30.24	2.58	2.87
Av. Rel. Error%:			17.85	12.40	11.74	11.23		
Av. Sq. Error:			22.10	12.38	2.84	2.69		

RS=ratio synthetic estimator; SD=sample-size dependent estimator; EBLUP=EBLUP or EB estimator; HB=HB estimator.

(standard error), s_i^{HB}, of HB estimator using one-dimensional numerical integration. These values are also reported in Table 3.

The following observations on the relative performances of small area estimates may be drawn from Table 3: (1) EBLUP and HB estimators give similar values over small areas, and their average relative errors (%) are 11.74 and 11.23 and squared errors are 2.84 and 2.69 respectively. Asymptotically (as $m \to \infty$), the two estimators are identical, and the observed differences are due to moderate $m(= 16)$ and the method of estimating σ_v^2 and σ^2 (REML or ML would give slightly different EBLUP values). (2) Standard error values for EBLUP and HB estimators are also similar. This is in agreement with the empirical results of Datta and Ghosh (1991) and Hulting and Harville (1991). (3) Under the criterion of average squared error, EBLUP and HB estimators perform much better than the ratio-synthetic and sample-size dependent estimators: 2.84 for EBLUP vs. 12.38 for sample-size dependent (SD) and 22.10 for ratio-synthetic (RS). (4) Under the criterion of average relative error (%), however, EBLUP and HB estimates are not much better than the sample-size dependent estimator: 11.74 for EBLUP versus 12.40 for SD. However, both perform much better than the ratio-synthetic estimator with % ARE = 17.85.

It may be noted that EBLUP, EB and HB estimators are optimal under squared error loss and cease to be so under relative error loss. This is due to the fact that the Bayes estimators under relative error loss can often differ quite significantly from those under squared error loss. This nonoptimality carries over to EBLUP estimator which usually mimics closely the Bayes estimators. The above observations could perhaps explain why in our example the Bayes and EBLUP estimator did not improve significantly over the SD estimator under relative error.

All in all, our results in Table 3 clearly demonstrate the advantages of using the EBLUP or HB estimator and associated standard error when the assumed random effects model fits the data well. (Note that we simulated the data from an assumed model.) It is important, therefore, to examine the aptness of the assumed model using suitable diagnostic tools; Section 7.1 gives a brief account of diagnostics for models (4.4) and (4.6).

7. SPECIAL PROBLEMS

In this section we focus on special problems that may be encountered in implementing model-based methods for small area estimation. We also discuss some extensions of our basic models (4.4) and (4.6).

7.1 Model Diagnostics

Model-based methods rely on careful checking of the assumed models in order to find suitable models

that fit the data well. Model diagnostics, therefore, play an important role. However, the literature on diagnostics for mixed linear models involving random effects is not extensive, unlike standard regression diagnostics. Only recently have some useful diagnostic tools been proposed. See, for example, Battese, Harter and Fuller (1988); Beckman, Nachtsheim and Cook (1987); Calvin and Sedransk (1991); Christensen, Pearson and Johnson (1992); Cressie (1992); Dempster and Ryan (1985) and Lange and Ryan (1989).

We first consider the Fay-Herriot type model (4.4), where only area-specific covariates are used. When the model is correct, the standardized residuals $r_i = (\hat{\sigma}_v^2 z_i^2 + \psi_i)^{-(1/2)} (\hat{\theta}_i - \mathbf{x}_i^T \hat{\beta})$, $i = 1, \ldots, m$ are approximately iid $N(0,1)$ for large m where $\hat{\beta}$ is the BLUE estimator (5.2) with σ_v^2 replaced by $\hat{\sigma}_v^2$. We can, therefore, use a $q - q$ plot of r_i against $\Phi^{-1}[F_m(r_i)]$, where $\Phi(r)$ and $F_m(r)$ are the standard normal and empirical cdfs, respectively. A primary goal of this plot is to check the normality of the random effects v_i since the sampling errors e_i are approximately normal due to the central limit theorem effect. Dempster and Ryan (1985) note that the above $q - q$ plot may be inefficient for this purpose since it gives equal weight to each observation, even though the $\hat{\theta}_i$s differ in the amount of information contained about the v_is. They propose a weighted $q - q$ plot which uses a weighted empirical cdf $F_m^*(r) = \Sigma_i I(r - r_i) W_i / \Sigma_i W_i$ in place of $F_m(r)$, where $I(t) = 1$ for $t \geq 0$ and 0 otherwise, and $W_i = (\hat{\sigma}_v^2 + z_i^{-2} \psi_i)^{-1}$ in our case. This plot is more sensitive to departures from normality than the unweighed plot since it assigns greater weight to those observations for which $\hat{\sigma}_v^2$ account for a larger part of the total variance $\hat{\sigma}_v^2 + z_i^{-2} \psi_i$.

We next turn to the nested errors regression model (4.6), where the y_{ij}'s are correlated for each i. In this case, the transformed residuals $r_{ij} = k_{ij}^{-1} (y_{ij} - \hat{\gamma}_i \bar{y}_{iw}) - k_{ij}^{-1} (\mathbf{x}_{ij} - \hat{\gamma}_i \bar{\mathbf{x}}_{iw})^T \hat{\beta}$ are approximately uncorrelated with equal variances σ^2. Therefore, traditional regression diagnostics may be applied to the r_{ij}s, but the transformation can mask the effect of individual errors e_{ij}. On the other hand, standardized BLUP residuals $k_{ij}^{-1} (y_{ij} - \mathbf{x}_{ij}^T \hat{\beta} - \hat{v}_i) / \hat{\sigma}$ may be used to study the effect of individual units (ij) on the model, provided they are not strongly correlated. Lange and Ryan (1989) propose methods for checking the normality assumption on the random effects v_i using the BLUP estimates $\hat{v}_i$.

Christensen, Pearson and Johnson (1992) develop case-deletion diagnostics for detecting influential observations in mixed linear models. Their methods can be applied to model (4.6) as well as to more complex small area models.

7.2 Constrained Estimation

Direct survey estimates are often adequate at an aggregate (or large area) level in terms of precision. For example, Battese, Harter and Fuller (1988), in their application, find that the direct regression estimator of the mean crop area for the 12 counties together has adequate precision. It is, therefore, sometimes desirable to modify the individual small area estimators so that a properly weighted sum of these estimators equals the model-free, direct estimator at the aggregate level. The modified estimators will be somewhat less efficient than the original, optimal estimators, but they avoid possible aggregation bias by ensuring consistency with the direct estimator. One simple way to achieve consistency is to make a ratio adjustment, for example, the EBLUP estimator $\hat{Y}_i^H$ of a total Y_i is modified to

$$(7.1) \qquad \hat{Y}_i^H \text{ (mod)} = \left(\hat{Y}_i^H / \sum_i \hat{Y}_i^H \right) \hat{Y},$$

where $\hat{Y}$ is a direct estimator of the aggregate population total $Y = \Sigma_i Y_i$. Battese, Harter and Fuller (1988) and Pfeffermann and Barnard (1991) propose alternative estimators involving estimated variances and covariances of the optimal estimators $\hat{Y}_i^H$.

The previous sections focused on simultaneous estimation of small area means or totals, but in some applications the main objective is to produce an ensemble of parameter estimates whose histogram is in some sense close to the histogram of small area parameters. Spjøtvoll and Thomsen (1987), for example, were interested in finding how 100 municipalities in Norway were distributed according to proportion of persons not in the labor force. They propose constrained *EB* estimators whose variation matched the variation of the small area population means. By comparing with the actual distribution in their example, they show that the *EB* estimators are biased toward the prior mean compared to the constrained *EB* estimators. Constrained estimators reduce shrinking towards the synthetic component; for example, in (5.1) the weight $1 - \gamma_i$, attached to the synthetic component, is reduced to $1 - \gamma_i^{1/2}$. Following Louis (1984), Ghosh (1992) develops a general theory of constrained HB estimation. Ghosh obtains constrained HB estimates by matching the first two moments of the histogram of the estimates, and the posterior expectations of the first two moments of the histogram of the parameters and minimizing, subject to these conditions, the posterior expectation of the Euclidean distance between the estimates and the parameters. Lahiri (1990) obtains

similar results in the context of small area estimation, assuming "posterior linearity," thus avoiding distributional assumptions. Constrained Bayes estimates are suitable for subgroup analysis where the problem is not only to estimate the different components of a parameter vector but also to identify the parameters that are above or below a specified cut-off point. It should be noted that synthetic estimates are inappropriate for this purpose.

The optimal estimators (i.e., EBLUP, EB and HB estimators) may perform well overall but poorly for particular small areas that are not consistent with the assumed model on small area effects. To avoid this problem, Efron and Morris (1972) and Fay and Harriot (1979) suggest a straightforward compromise that consists of restricting the amount by which the optimal estimator differs from the direct estimator by some multiple of the standard error of the direct estimator. For example, a compromise estimator corresponding to the HB estimator $\hat{\theta}_i^{HB}$, under a normal prior on the θ_i's, is given by

$$\bar{\theta}_i^{HB} = \begin{cases} \hat{\theta}_i^{HB}, & \text{if } \hat{\theta}_i - c\psi_i^{1/2} \le \hat{\theta}_i^{HB} \le \hat{\theta}_i + c\psi_i^{1/2} \\ \hat{\theta}_i - c\psi_i^{1/2}, & \text{if } \hat{\theta}_i^{HB} < \hat{\theta}_i - c\psi_i^{1/2} \\ \hat{\theta}_i + c\psi_i^{1/2}, & \text{if } \hat{\theta}_i^{HB} > \hat{\theta}_i + c\psi_i^{1/2}, \end{cases}$$

where $c > 0$ is a suitable chosen constant, say $c = 1$. A limitation of the compromise estimators is that no reliable measures of their precision are available.

7.3 Extensions

Various extensions of the basic models (4.4) and (4.6) have been studied in the literature. Due to space limitation, we can only mention some of these extensions.

Datta et al. (1992) extend the aggregate-level model (4.4) to the case of correlated sampling errors with a known covariance matrix and develop HB and EB estimators and associated measures of precision. In their application to adjustment of census undercount, the sampling covariance matrix is block diagonal. Cressie (1990a) introduces spatial dependence among the random effects v_i, in the context of adjustment for census undercount. Fay (1987) and Ghosh, Datta and Fay (1991) extend model (4.4) to multiple characteristics and perform hierarchical and empirical multivariate Bayes analysis, assuming that the sampling covariance matrix of $\hat{\boldsymbol{\theta}}_i$, the vector of direct estimators for ith area, is known for each i. In their application to estimation of median income for four-person families by state, $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2})^T$ with θ_{i1} = population median income of four-person families in state i and $\theta_{i2} = \frac{3}{4}$ (population median income of three-person families in state i) $+\frac{1}{4}$ (median income of four-person families in state i). By taking advantage of the strong correlation between the direct estimators $\hat{\theta}_{i1}$ and $\hat{\theta}_{i2}$, they were able to obtain improved estimators of θ_{i1}.

Many surveys are repeated in time with partial replacement of the sample elements, for example, the monthly U.S. Current Population Survey and the Canadian Labor Force Survey. For such repeated surveys considerable gain in efficiency can be achieved by borrowing strength across both small areas and time. Cronkite (1987) developed regression synthetic estimators using pooled cross-sectional time series data and applied them to estimate substate area employment and unemployment using the Current Population Survey monthly survey estimates as dependent variable and counts from the Unemployment Insurance System and Census variables as independent variables. Rao and Yu (1992) propose an extension of model (4.4) to time series and cross-sectional data. Their model is of the form

$$\hat{\theta}_{it} = \theta_{it} + e_{it}, \qquad t = 1, \ldots, T, \tag{7.2}$$

$$\theta_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + v_i + u_{it}, \qquad i = 1, \ldots, m, \tag{7.3}$$

where $\hat{\theta}_{it}$ is the direct estimator for small area i at time t, the e_{it}'s are sampling errors with a known block diagonal covariance matrix Ψ = block diag $(\Psi_1, \ldots, \Psi_m)$, $\mathbf{x}_{it}$ is a vector of covariates and $v_i \overset{\text{iid}}{\sim} N(0, \sigma_v^2)$. Further, the u_{it}'s are assumed to follow a first order autoregressive process for each i, i.e., $u_{it} = \rho u_{i,t-1} + \epsilon_{it}, |\rho| < 1$ with $\epsilon_{it} \overset{\text{iid}}{\sim} N(0, \sigma^2)$. They obtain the EBLUP and HB estimators and their standard errors under (7.2) and (7.3).

Models of the form (7.3) have been extensively used in the econometric literature, ignoring sampling errors (see, e.g., Anderson and Hsiao, 1981; Judge, 1985, Chapter 13). Choudhry and Rao (1989) treat the composite error $w_{it} = e_{it} + u_{it}$ as a first order autoregressive process and obtain the EBLUP estimator of $\mathbf{x}_{it}^T\boldsymbol{\beta} + v_i$. A drawback of their method is that the area by time specific effect u_{it} is ignored in modelling the θ_{it}'s.

Pfeffermann and Burck (1990) investigate more general models on the θ_{it}'s, but they assume modeling of sampling errors across time. They obtain EBLUP estimators of small area means using the Kalman filter. Singh and Mantel (1991) consider arbitrary covariance structures on sampling errors and propose recursive composite estimators using the Kalman filter. These estimators are not optimal but appear to be quite efficient relative to the corresponding EBLUP estimators.

Turning to extension of the nested error regression model (4.6), Fuller and Harter (1987) propose a multivariate nested error regression model and obtain EBLUP estimators and associated standard errors. Stukel (1991) studies two-fold nested error regression models, and obtains EBLUP estimators and associated standard errors. Such models are appropriate for two-stage sampling within small areas. Kleffe and Rao (1992) extend model (4.6) to the case of random error variances, σ_i^2, and obtain EBLUP estimator and associated standard errors in the special case of $\mathbf{x}_{ij} = 1$.

MacGibbon and Tomberlin (1989) and Malec, Sedransk and Tompkins (1991) study logistic regression models with random area-specific effects. Such models are appropriate for binary response variables when element-specific covariates are available. MacGibbon and Tomberlin (1989) obtain EB estimators of small area proportions and associated standard errors, but they ignore the uncertainty about the prior parameters. Farrell, MacGibbon and Tomberlin (1992) apply the bootstrap method of Laird and Louis (1987) to account for the underestimation of true posterior variance. Malec, Sedransk and Tompkins (1991) obtain HB estimators and associated standard errors using Gibbs sampling and apply their method to data from the U.S. National Health Interview Survey to produce estimates of proportions for individual states.

EB and HB methods have also been used for estimating regional mortality and disease rates (see, e.g., Marshall, 1991). In these applications, the observed small area counts, y_i, are assumed to be independent Poisson with conditional mean $E(y_i|\theta_i) = n_i\theta_i$, where θ_i and n_i respectively denote the true rate and number exposed in the ith area. Further, the θ_is are assumed to be random with a specified distribution (e.g., a gamma distribution with unknown scale and shape parameters). The EB or HB estimators are shrinkage estimators in the sense that the crude rate y_i/n_i is shrunk towards an overall regional rate, ignoring the spatial aspect of the problem. Marshall (1991) proposes "local" shrinkage estimators obtained by shrinking the crude rate towards a local neighbourhood rate. Such estimators are practically appealing and further work on their statistical properties is desirable.

De Souza (1992) studies joint mortality rates of two cancer sites over several geographical areas and obtains asymptotic approximations to posterior means and variances using the general first order approximations given by Kass and Steffey (1989). The bivariate model leads to improved estimators for each site compared to the estimators based on univariate models.

8. CONCLUSION

In this article, we have used the term "small area" to denote any local geographical area that is small or to describe any small subgroup of a population such as a specific age-sex-race group of people within a large geographical area. Sample sizes for small areas are typically small because the overall sample size in a survey is usually determined to provide desired accuracy at a much higher level of aggregation. As a result, the usual direct estimators of a small area mean are unlikely to give acceptable reliability; and it becomes necessary to "borrow strength" from related areas to find more accurate estimators for a given area or, simultaneously, for several areas. Considerable attention has been given to such indirect estimators in recent years.

We have attempted to provide an appraisal of indirect estimation covering both traditional design-based methods and newer model-based approaches to small area estimation. Traditional methods covered here include demographic techniques for local estimation of population and other characteristics of interest in post-censal years, and synthetic and sample size dependent estimation. Model-based methods studied here include EBLUP, EB and HB estimation. Two types of basic small area models that include random area-specific effects are used to describe these methods. In the first type of models, only area-specific auxiliary data are available for the population elements while in the second type element-specific auxiliary data are available for the population elements.

We have emphasized the importance of obtaining accurate measures of uncertainty associated with the model-based estimators. To this end, an approximately unbiased estimator of MSE of the EBLUP estimator is given as well as two methods of approximating the true posterior variance, irrespective of the form of the prior distribution on the model parameters. The latter approximations may be used as measures of uncertainty associated with the EB estimator. In the HB approach, a prior distribution on the model parameters is specified and the resulting posterior variance is used as a measure of uncertainty associated with the HB estimator (posterior mean). We have also mentioned several applications of the model-based methods.

We have also considered special problems that may be encountered in implementing model-based methods for small area estimation; in particular, model diagnostics for small area models, constrained estimation, "local" shrinkage, spatial modelling and borrowing strength across both small areas and time. We anticipate quite a bit of future research on these topics.

Caution should be exercised in using or recom-

mending indirect estimators since they are based on implicit or explicit models that connect the small areas, unlike the direct estimators. As noted by Schaible (1992): "Indirect estimators should be considered when better alternatives are not available, but only with appropriate caution and in conjunction with substantial research and evaluation efforts. Both producers and users must not forget that, even after such efforts, indirect estimates may not be adequate for the intended purpose." (Also see Kalton, 1987.)

Finally, we should emphasize the need for developing an overall program that covers issues relating to sample design and data evolvement, organization and dissemination, in addition to those pertaining to methods of estimation for small areas.

APPENDIX

Variances and Covariance of $\hat{\sigma}_v^2$ and $\hat{\sigma}^2$

Let $\hat{\sigma}_v^2$ and $\hat{\sigma}^2$ be the estimators of σ_v^2 and σ^2 obtained from the method of fitting constants. Then

$$V(\hat{\sigma}^2) = 2\nu_1^{-1}\sigma^4$$

$$V(\hat{\sigma}_v^2) \doteq 2\eta_*^{-2}\Big[\nu_1^{-1}(n-p-\nu_1)(n-p)\sigma^4 + \eta_{**}\sigma_v^4 + 2\eta_*\sigma^2\sigma_v^2\Big]$$

with

$$\eta_{**} = \sum w_{i\cdot}^2\left(1 - w_{i\cdot}\bar{\mathbf{x}}_{iw}^T\mathbf{A}_1^{-1}\bar{\mathbf{x}}_{iw}\right) + \mathrm{tr}\left(\mathbf{A}_1^{-1}\sum w_{i\cdot}^2\bar{\mathbf{x}}_{iw}\bar{\mathbf{x}}_{iw}^T\right)^2$$

and

$$\mathrm{cov}\,(\hat{\sigma}_v^2, \hat{\sigma}_e^2) \doteq -2\eta_*^{-1}\nu_1^{-1}(n-p-\nu_1)\sigma^4.$$

(See Stukel, 1991.)

ACKNOWLEDGMENTS

The authors would like to thank the former editor, J. V. Zidek, and a referee for several constructive suggestions. M. Ghosh was supported by NSF Grants DMS-89-01334 and SES-92-01210. J. N. K. Rao's research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

ANDERSON, T. W. and HSIAO, C. (1981). Formulation and estimation of dynamic models using panel data. *J. Econometrics* 18 67–82.

BATTESE, G. E., HARTER, R. M. and FULLER, W. A. (1988). An error-component model for prediction of county crop areas using survey and satellite data. *J. Amer. Statist. Assoc.* 83 28–36.

BECKMAN, R. J., NACHTSHEIM, C. J. and COOK, R. D. (1987). Diagnostics for mixed-model analysis of variance. *Technometrics* 29 413–426.

BOGUE, D. J. (1950). A technique for making extensive postcensal estimates. *J. Amer. Statist. Assoc.* 45 149–163.

BOGUE, D. J. and DUNCAN, B. D. (1959). A composite method of estimating post censal population of small areas by age, sex and colour. Vital Statistics–Special Report 47, No. 6, National Office of Vital Statistics, Washington, DC.

BRACKSTONE, G. J. (1987). Small area data: policy issues and technical challenges. In *Small Area Statistics* (R. Platek, J. N. K. Rao, C. E. Särndal and M. P. Singh eds.) 3–20. Wiley, New York.

CALVIN, J. A. and SEDRANSK, J. (1991). Bayesian and frequentist predictive inference for the patterns of care studies. *J. Amer. Statist. Assoc.* 86 36–48.

CHAMBERS, R. L. and FEENEY, G. A. (1977). Log linear models for small area estimation. Unpublished paper, Australian Bureau of Statistics.

CHAUDHURI, A. (1992). Small domain statistics: a review. Technical report ASC/92/2, Indian Statistical Institute, Calcutta.

CHOUDHRY, G. H. and RAO, J. N. K. (1989). Small area estimation using models that combine time series and cross-sectional data. In *Symposium 89—Analysis of Data in Time—Proceedings* (A. C. Singh and P. Whitridge, eds.) 67–74. Statistics Canada, Ottawa.

CHRISTENSEN, R., PEARSON, L. M. and JOHNSON, W. (1992). Case deletion diagnostics for mixed models. *Technometrics* 34 38–45.

CRESSIE, N. (1989). Empirical Bayes estimation of undercount in the decennial census. *J. Amer. Statist. Assoc.* 84 1033–1044.

CRESSIE, N. (1990a). Small area prediction of undercount using the general linear model. In *Symposium 90—Measurement and Improvement of Data Quality—Proceedings* 93–105. Statistics Canada, Ottawa.

CRESSIE, N. (1992). REML estimation in empirical Bayes smoothing of census undercount. *Survey Methodology* 18 75–94.

CRONKITE, F. R. (1987). Use of regression techniques for developing state and area employment and unemployment estimates. In *Small Area Statistics* (R. Platek, J. N. K. Rao, C. E. Särndal and M. P. Singh, eds.) 160–174. Wiley, New York.

DATTA, G. S. and GHOSH, M. (1991). Bayesian prediction in linear models: Applications to small area estimation. *Ann. Statist.* 19 1748–1770.

DATTA, G. S., GHOSH, M., HUANG, E. T., ISAKI, C. T., SCHULTZ, L. K. and TSAY, J. H. (1992). Hierarchical and empirical Bayes methods for adjustment of census undercount: the 1980 Missouri dress rehearsal data. *Survey Methodology* 18 95–108.

DEMPSTER, A. P. and RYAN, L. M. (1985). Weighted normal plots. *J. Amer. Statist. Assoc.* 80 845–850.

DESOUZA, C. M. (1992). An approximate bivariate Bayesian method for analysing small frequencies. *Biometrics* 48 1113–1130.

DREW, D., SINGH, M. P. and CHOUDHRY, G. H. (1982). Evaluation of small area estimation techniques for the Canadian Labour Force Survey. *Survey Methodology* 8 17–47.

EFRON, B. and MORRIS, C. (1972). Limiting the risk of Bayes and empirical Bayes estimates—Part II: the empirical Bayes case. *J. Amer. Statist. Assoc.* 67 130–139.

ERICKSEN, E. P. (1974). A regression method for estimating populations of local areas. *J. Amer. Statist. Assoc.* 69 867–875.

ERICKSEN, E. P. and KADANE, J. B. (1985). Estimating the population in a census year (with discussion). *J. Amer. Statist. Assoc.* 80 98–131.

ERICKSEN, E. P. and KADANE, J. B. (1987). Sensitivity analysis of local estimates of undercount in the 1980 U.S. Census. In *Small Area Statistics* (R. Platek, J. N. K. Rao, C .E. Särndal and M. P. Singh, eds.) 23–45. Wiley, New York.

ERICKSEN, E. P. and KADANE, J. B. (1992). Comment on "Should we have adjusted the U.S. Census of 1980?," by D. A. Freedman and W. C. Navidi. *Survey Methodology* 18 52–58.

ERICKSEN, E. P., KADANE, J. B. and TUKEY, J. W. (1989). Adjusting the 1981 census of population and housing. *J. Amer. Statist. Assoc.* 84 927–944.

ERICSON, W. A. (1969). A note on the posterior mean. *J. Roy. Statist. Soc. Ser. B* 31 332–334.

FARRELL, P. J., MACGIBBON, B. and TOMBERLIN, T.J. (1992). An evaluation of bootstrap techniques for correcting empirical Bayes interval estimates. Unpublished manuscript, Dept. Statistics and Actuarial Science, Univ. Waterloo.

FAY, R. E. (1987). Application of multivariate regression to small domain estimation. In *Small Area Statistics* (R. Platek, J. N. K. Rao, C. E. Särndal and M. P. Singh, eds.) 91–102. Wiley, New York.

FAY, R. E. and HERRIOT, R. A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *J. Amer. Statist. Assoc.* 74 269–277.

FREEDMAN, D. A. and NAVIDI, W. C. (1986). Regression models for adjusting the 1980 Census (with discussion). *Statist. Sci.* 1 1–39.

FREEDMAN, D. A. and NAVIDI, W. C. (1992). Should we have adjusted the U.S. Census of 1980? (with discussion). *Survey Methodology* 18 3–74.

FULLER, W. A. and BATTESE, G. E. (1973). Transformations for estimation of linear models with nested error structure. *J. Amer. Statist. Assoc.* 68 626–632.

FULLER, W. A. and HARTER, R. M. (1987). The multivariate components of variance model for small area estimation. In *Small Area Statistics* (R. Platek, J. N. K. Rao, C. E. Särndal and M. P. Singh, eds.) 103–123. Wiley, New York.

GELFAND, A. and SMITH, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* 85 398–409.

GHANGURDE, D. D. and SINGH, M. P. (1977). Synthetic estimators in periodic households surveys. *Survey Methodology* 3 152–181.

GHOSH, M. (1992a). Hierarchical and empirical Bayes multivariate estimation. In *Current Issues in Statistical Inference: Essays in Honor of D. Basu* (M. Ghosh and P. K. Pathak, eds.) 151–177. IMS, Hayward, CA.

GHOSH, M. (1992b). Constrained Bayes estimation with applications. *J. Amer. Statist. Assoc.* 87 533–540.

GHOSH, M., DATTA, G. S. and FAY, R. E. (1991). Hierarchical and empirical multivariate Bayes analysis in small area estimation. In *Proceeding of the Bureau of the Census Annual Research Conference* 63–79. Bureau of the Census, Washington, DC.

GHOSH, M. and LAHIRI, P. (1987). Robust empirical Bayes estimation of means from stratified samples. *J. Amer. Statist. Assoc.* 82 1153–1162.

GONZALEZ, M. E. (1973). Use and evaluation of synthetic estimators. In *Proceedings of the Social Statistics Section* 33–36. Amer. Statist. Assoc., Washington, DC.

HANSEN, M., HURWITZ, W. N. and MADOW, W. G. (1953). *Sample Survey Methods and Theory*, 1. Wiley, New York.

HARVILLE, D. A. (1991). Comment on, "That BLUP is a good thing: The estimation of random effects," by G. K. Robinson. *Statist. Sci.* 6 35–39.

HENDERSON, C. R. (1950). Estimation of genetic parameters (abstract). *Ann. Math. Statist.* 21 309–310.

HOLT, D., SMITH, T. M. F. and TOMBERLIN, T. J. (1979). A model-based approach to estimation for small subgroups of a population. *J. Amer. Statist. Assoc.* 74 405–410.

HULTING, F. L. and HARVILLE, D. A. (1991). Some Bayesian and non-Bayesian procedures for the analysis of comparative experiments and for small area estimation: computational aspects, frequentist properties, and relationships. *J. Amer. Statist. Assoc.* 86 557–568.

ISAKI, C. T., SCHULTZ, L. K., SMITH, P. J. and DIFFENDAL, D. J. (1987). Small area estimation research for census undercount—progress report. In *Small Area Statistics* (R. Platek, J. N. K. Rao, C.E. Särndal and M.P. Singh, eds.) 219–238. Wiley, New York.

JUDGE, G. G. (1985). *The Theory and Practice of Econometrics*. 2nd ed. Wiley, New York.

KACKAR, R. N. and HARVILLE, D. A. (1984). Approximations for standard errors of estimators for fixed and random effects in mixed models. *J. Amer. Statist. Assoc.* 79 853–862.

KALTON, G. (1987). Discussion. In *Small Area Statistics* (R. Platek, J. N. K. Rao, C. E. Särndal and M. P. Singh, eds.) 264-266. Wiley, New York.

KALTON, G., KORDOS, J. and PLATEK, R. (1993). *Small Area Statistics and Survey Designs Vol. I: Invited Papers; Vol. II: Contributed Papers and Panel Discussion*. Central Statistical Office, Warsaw.

KASS, R. E. and STEFFEY, D. (1989). Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *J. Amer. Statist. Assoc.* 84 717–726.

KLEFFE, J. and RAO, J. N. K. (1992). Estimation of mean square error of empirical best linear unbiased predictors under a random error variance linear model. *J. Multivariate Anal.* 43 1–15.

LAHIRI, P. (1990). "Adjusted" Bayes and empirical Bayes estimation in finite population sampling. *Sankhyā Ser. B* 52 50–66.

LAHIRI, P. and RAO, J. N. K. (1992). Robust estimation of mean square error of small area estimators. Unpublished manuscript.

LAIRD, N. M. and LOUIS, T. A. (1987). Empirical Bayes confidence intervals based on bootstrap samples. *J. Amer. Statist. Assoc.* 82 739–750.

LANGE, N. and RYAN, L. (1989). Assessing normality in random effects models. *Ann. Statist.* 17 624–642.

LOUIS T. (1984). Estimating a population of parameter values using Bayes and empirical Bayes methods. *J. Amer. Statist. Assoc.* 79 393–398.

MACGIBBON, B. and TOMBERLIN, T.J. (1989). Small area estimation of proportions via empirical Bayes techniques. *Survey Methodology* 15 237–252.

MALEC, D., SEDRANSK, J. and TOMPKINS, L. (1991). Bayesian predictive inference for small areas for binary variables in the National Health Interview Survey. Unpublished manuscript.

MARKER, D. A. (1983). Organization of small area estimators. *Proceedings of Survey Research Methods Section* 409–414. Amer. Statist. Assoc., Washington, DC.

MARSHALL, R. J. (1991). Mapping disease and mortality rates using empirical Bayes estimators. *J. Roy. Statist. Soc. Ser. C* 40 283–294.

MCCULLAGH, P. and ZIDEK, J. (1987). Regression methods and performance criteria for small area population estimation. In *Small Area Statistics* (R. Platek, J. N. K. Rao, C. E. Särndal and M. P. Singh, eds.) 62–74. Wiley, New York.

MORRIS, C. (1983). Parametric empirical Bayes inference: Theory and applications (with discussions). *J. Amer. Statist. Assoc.* 78 47–65.

MORRISON, P. (1971). Demographic information for cities: a manual for estimating and projecting local population characteristics. RAND report R-618-HUD.

NATIONAL CENTER FOR HEALTH STATISTICS (1968). *Synthetic State Estimates of Disability*. P.H.S. Publication 1759. U.S. Government Printing Office, Washington, DC.

NATIONAL INSTITUTE ON DRUG ABUSE (1979). *Synthetic Estimates*

for Small Areas (research monograph 24). U.S. Government Printing Office, Washington, DC.

NATIONAL RESEARCH COUNCIL (1980). *Panel on Small-Area Estimates of Population and Income. Estimating Population and Income of Small Areas*. National Academy Press, Washington, DC.

NICHOL, S. (1977). A regression approach to small area estimation. Unpublished manuscript, Australian Bureau of Statistics, Canberra, Australia.

NTIS (1963). Indirect estimators in federal programs. Statistical policy working paper 21, prepared by the Subcommittee on Small Area Estimation, Federal Committee on Statistical Methodology, Office of Management and Budget, Washington, DC.

PFEFFERMANN, D. and BURCK, L. (1990). Robust small area estimation combining time series and cross-sectional data. *Survey Methodology* **16** 217–237.

PFEFFERMANN, D. and BARNARD, C. (1991). Some new estimators for small area means with applications to the assessment of farmland values. *Journal of Business and Economic Statistics* **9** 73–84.

PLATEK, R. and SINGH, M. P. (1986). *Small Area Statistics: Contributed Papers*. Laboratory for Research in Statistics and Probability, Carleton Univ.

PLATEK, R., RAO, J. N. K., SÄRNDAL, C. E. and SINGH, M. P. (1987). *Small Area Statistics*. Wiley, New York.

PRASAD, N. G. N. and RAO, J. N. K. (1990). The estimation of mean squared errors of small-area estimators. *J. Amer. Statist. Assoc.* **85** 163–171.

PURCELL, N. J. and LINACRE, S. (1976). Techniques for the estimation of small area characteristics. Unpublished manuscript.

PURCELL, N. J. and KISH, L. (1979). Estimation for small domain. *Biometrics* **35** 365–384.

PURCELL, N. J. and KISH, L. (1980). Postcensal estimates for local areas (or domains). *Internat. Statist. Rev.* **48** 3–18.

RAO, C. R. and SHINOZAKI, N. (1978). Precision of individual estimates in simultaneous estimation of parameters. *Biometrika* **65** 23–30.

RAO, J. N. K. (1986). Synthetic estimators, SPREE and best model based predictors. In *Proceedings of the Conference on Survey Research Methods in Agriculture* 1–16. U.S. Dept. Agriculture, Washington, DC.

RAO, J. N. K. and YU, M. (1992). Small area estimation by combining time series and cross-sectional data. In *Proceedings of the Survey Research Methods Section* 1–19. Amer. Statist. Assoc., Alexandria, VA.

ROBINSON, G. K. (1991). That BLUP is a good thing: the estimation of random effects (with discussion). *Statist. Sci.* **6** 15–51.

SÄRNDAL, C. E. and HIDIROGLOU, M. A. (1989). Small domain estimation: a conditional analysis. *J. Amer. Statist. Assoc.* **84** 266–275.

SCHAIBLE, W. L. (1978). Choosing weights for composite estimators for small area statistics. In *Proceedings of the Survey Research Methods Section* 741–746. Amer. Statist. Assoc., Washington, DC.

SCHAIBLE, W. L. (1992). Use of small area statistics in U.S. Federal Programs. In *Small Area Statistics and Survey Designs* (G. Kalton, J. Kordos and R. Platek, eds.) **1** 95–114. Central Statistical Office, Warsaw.

SINGH, A. C. and MANTEL, H. J. (1991). State space composite estimation for small areas. In *Symposium 91—Spatial Issues in Statistics—Proceedings* 17–25. Statistics Canada, Ottawa.

SINGH, M. P., GAMBINO, J. and MANTEL, H. (1992). Issues and options in the provision of small area data. In *Small Area Statistics and Survey Designs* (G. Kalton, J. Kordos and R. Platek, eds.) **1** 37–75. Central Statistical Office, Warsaw.

SMITH, S. K. and LEWIS, B. B. (1980). Some new techniques for applying the housing unit method of local population estimation. *Demography* **17** 323–340.

SPJØTVOLL, E. and THOMSEN, I. (1987). Application of some empirical Bayes methods to small area statistics. *Bulletin of the International Statistical Institute* **2** 435–449.

STARSINIC, D. E. (1974). Development of population estimates for revenue sharing areas. Census Tract Papers, Ser. GE40, No. 10 U.S. Government Printing Office, Washington, DC.

STATISTICS CANADA (1987). *Population Estimation Methods, Canada*, Catalogue 91-528E. Statistics Canada, Ottawa.

STEFFEY, D. and KASS, R. E. (1991). Comment on "That BLUP is a good thing: The estimation of random effects," by G. K. Robinson. *Statist. Sci.* **6** 45–47.

STUKEL, D. (1991). *Small Area Estimation Under One and Two-Fold Nested Error Regression Model*. Ph.D. Thesis, Carleton Univ.

U.S. BUREAU OF THE CENSUS (1966). Methods of population estimation: Part I, Illustrative procedure of the Bureau's component method II. *Current Population Reports, Series P-25, No. 339*. U.S. Government Printing Office, Washington, DC.

ZIDEK, J. V. (1982). A review of methods for estimating the populations of local areas. Technical Report 82–4, Univ. British Columbia, Vancouver.