번 역 자 료

캐나다 노동력조사 방법론

2001. 9

통 계 기 획 국
조 사 관 리 과

목 차

1. 서 론

1.1 배경 및 목적

캐나다 노동력조사(Labour Force Survey:LFS)는 대규모 노동시장의 변화 양상 및 시의성 있는 노동시장의 정보를 파악하기 위해 2차 세계대전 이후 도입되었고, 주로 주(Province) 지역 및 국가 단위의 고용 및 실업통계를 생산할 목적으로 설계되었다. LFS는 1945년 분기별 조사로 시작하여 1952년 월별 조사로 변경되었고, 1960년부터 캐나다 실업통계를 생산하기 위한 공식조사로 승인되었다. 그 후 LFS를 통해 노동시장의 다양한 통계를 작성할 수 있도록 표본개편 및 조사방법에 관한 연구가 지속적으로 진행되었고, 현재는 캐나다 노동시장의 세부적인 변화에 관한 정보를 제공할 수 있을 정도로 발전을 거듭하였다. 매월 고용인구와 실업인구 총계 및 실업률에 관한 추정치, 노동인구의 특성(연령, 결혼여부, 교육정도, 가족현황) 등에 관한 공식통계는 LFS를 통해 작성된다.

LFS는 주 지역 및 전국 단위의 고용 및 실업통계 작성 외에 고용보험 경제구역(EIER:Employment Insurance Economic Regions), 센서스 도시지역(CMA:Census Metropolitan Areas) 등과 같은 주 내의 특정 행정구에 대한 통계작성도 가능하도록 표본이 설계 되어있다. 최근에 들어서는 주 지역 내의 센서스 조사구(CD:Census Divisions)와 같은 소지역들에 대해서도 소지역 추정기법을 이용하여 관련 통계지표를 작성하고 있으며, 지방정부의 소지역에 대한 예산 배분 또는 정책 결정 등의 사안에 이러한 소지역 통계들이 이용되고 있다.

LFS 추정치들은 매월 "Labour Force Information"라는 책자를 통해 공표된다. 또한 노동시장의 좀 더 다양한 정보들은 캐나다 통계국의 전자정보 데이터베이스의 일종인 "CANSIM"을 통해 획득할 수 있으며, LFS의 결과로부터 매월 9000 항목 이상의 시계열 자료들이 정기적으로 수정 보완된다. 이외에 노동시장의 중심지표가 되는 다양한 주제에 대한 세부적인 고찰을 다룬 "Labour Force Update"가 1997년부터 계간지로써 출간되고 있으며, 1976년 이래로 최근까지의 방대한 시계열 자료(time series data) 및 횡단면 자료(cross-sectional data)를 포함하고 있는 "Labour Force Historical Review on CD-ROM"이 매년 제작되고 있다.

캐나다 노동력 조사에 의해서 매월 발표되는 통계수치는 자영업, 부업과 전업을 포함한 취업자 총수와 실업자 총수이다. 매월 발표하는 노동시장의 표준지표는 실업률, 취업률과 경제활동 참가율이고 노동력 조사의 주요 정보 요소로서 15세 이상 인구의 개인적 특성은 나이, 성별, 혼인상태, 교육정도와 가족사항이다.

취업통계의 추정값들에는 인구학적 특성, 산업과 업종, 정규직과 통상적인 근로시간 등이 포함되어 있으며 설문내용에는 비자발적 부업적 취업, 복수 직업 여부와 휴직 등에 대해서 분석할 수 있는 주제들이 포함되어 있다. 특히 1997년 이후에는 근로자들의 노조가입 여부와 임금수준에 대한 정보와 작업장의 근로자 수 및 직업의 정규직 또는 임시직 여부에 대한 정보를 제공하고 있다.

실업통계의 추정값은 인구학적 범주별, 실업기간, 구직활동 전의 활동 및

바로 이전 직장에서 이직한 이유 등에 대한 정보를 제공하고 있다. 노동력 조사에 의해서 발표되는 통계는 국가 단위와 주 단위 추정값이 핵심적인 내용이지만 경제구역(ER : Economic Region)과 센서스 도시지역(CMA ; Census Metropolitan Area)과 같은 소지역 단위에 대한 노동력 상태의 추정값을 제공하고 있다.

1.2 노동력 상태 결정과정

취업과 실업의 개념은 생산의 요소로서 노동력 공급이론을 근거로 정의하였으며 생산은 국민계정 체계(SNA : the System of National Acconts)에서 언급한 것과 같이 상품과 서비스로 정의되는 개념이므로 작업의 목적이나 성질에서는 보수를 받는 근로활동과 조금도 차이가 없는 무보수의 가사노동이나 자원봉사활동은 근로시간으로 계산하지 않고 있다.

노동력 공급의 측정단위는 개별적인 근로시간이지만 조사에서 모집단의 개별적인 구성원들의 구분은 취업, 실업, 비경제활동 인구로 분류되어야 한다. 조사기간 중에 보수 근로 중인 사람은 근로시간에 상관없이 취업자로 구분하고 근로시간과 무관하게 노동시장에서 구직 행위가 있을 경우에는 실업자로 구분한다. 나머지 인구는 현재 일을 하고 있지 않거나 또는 노동시장에서 구직 활동하지 않는 경우로 비경제활동 인구로 정의한다.

통계조사에서 적용하는 취업자와 실업자의 정의와 개념은 국제노동기구(ILO : International Labor Organization)의 기준에 준거하고 있다.

(1) 취업자(Employment)

조사대상 기간 중에 직업이 있거나 개인 사업에서 일을 하는 사람을 말하며 자기에게 직접적인 소득이 없을 지라도 가구단위로 운영되는 농장, 사업체 또는 전문적인 기관에서 일하는 경우도 포함된다. 또한 직업이나 사업체가 있을지라도 일시적인 병이나, 휴가, 노동쟁의 등의 이유로 일을 하지 못하는 일시적인 휴직자도 포함한다.

(2) 실업자(Unemployment)

이용되지 못하는 공급된 노동력으로 실업자를 정의하고 있으며 실업자의 구분은 구직행위와 근로행위의 준비여부를 기준으로 하고 있으나 구직행위는 가구조사에서 구체적이고 지속적인 의사 표명을 전제로 하고 있다. 실업자는 조사 대상 기간 중 다음 3가지 항목 중 하나에 해당하는 사람으로 정의될 수 있다.

1) 현재 휴직 중이지만 근로 활동이 가능하고 복직을 기대할 수 있는 경우

2) 일을 하지 않고 있으나 구직 활동을 하고 있으며 일이 주어지면 바로 시작할 수 있는 경우

3) 조사대상 기간으로부터 4주 이내에 새로운 일을 할 직업이 주어졌으며 바로 일을 할 수 있는 경우

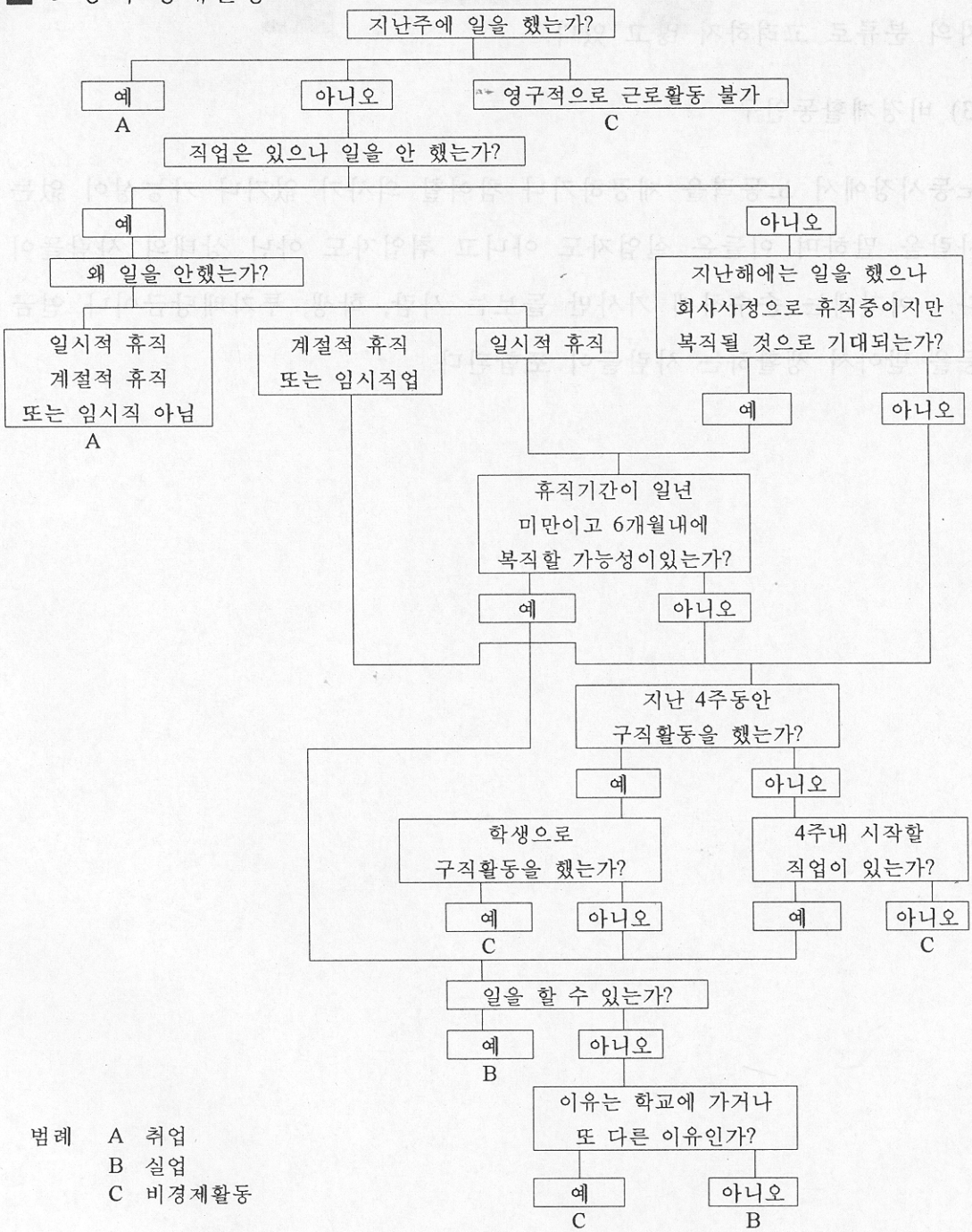
현재 학교에 다니는 학생이 전업적인 일을 찾고 있을 지라도 구직활동으

로 간주하지 않으며 여름방학동안의 근로 활동이나 구직 활동 역시 실업자의 분류로 고려하지 않고 있다.

(3) 비경제활동인구

노동시장에서 노동력을 제공하거나 참여할 의사가 없거나 가능성이 없는 사람을 말하며 이들은 실업자도 아니고 취업자도 아닌 상태의 사람들이다. 여기에는 순수하게 가사만 돌보는 사람, 학생, 투자배당금이나 연금 등을 받아서 생활하는 사람들이 포함된다.

■ 노동력 상태결정



범례 A 취업
B 실업
C 비경제활동

LFS는 캐나다 인구의 약 98%에 해당하는 인구를 목표모집단(target population)으로 설정하였다. 캐나다 북서부지역, 인디언 보호구역, 왕실 소유지, 수감자 및 직업군인에 해당하는 약 2%는 LFS의 조사대상에서 제외된다. 1998년 12월 현재 LFS의 표본크기는 52,350가구이다. 1970년 표본개편을 통해 LFS의 표본크기는 이전의 35,000가구에서 50,000가구로 증편된 후, 1980년 47,000가구로 감소되었다가 1989년 주 지역 내의 행정구역인 EIER 지역의 통계작성에 신뢰성을 확보하기 위해 63,000조사가구로 표본설계가 대폭 개편되었으며, 1993년에 약 59,000 조사가구로 다시 감소된 후 1995년부터 52,350 조사가구(EIER 지역에 16,500 조사가구, 나머지 지역에 35,850 조사가구)로 조정되어 현재에 이르고 있다.

LFS의 표본교체 체계(sample rotation)는 일종의 연동교체표본설계(rotating panel sample design)를 따른다. 표본가구는 6개의 부차표본(sub-sample)로 분리되어 6개월간 관리되며, 매월 1/6의 표본이 새로운 표본으로 대체되는 형식이다. 캐나다의 LFS는 매월 15일이 포함되는 주중에 실시되며, 80명의 조사전문인력을 포함하여 약 850명의 조사인력이 투입되어 조사가 이루어지고, 조사결과는 5개의 지방사무소(RO: Regional Office)에서 각각 취합되어 중앙으로 이관된다. 첫 달의 조사는 방문면접 형식을 취하며, 이 후 연속되는 다섯달 은 전화조사를 통해 조사가 이루어진다. 조사자는 휴대용 컴퓨터를 이용하여 직접 설문 항목의 결과를 입력하는 일종의 컴퓨터 보조 면접(CAI: Computer Assisted Interviewing) 방식을 이용한다. 조사결과는 조사완료시점에서 정확히 13일후에 공표된다.

2. 층화 및 추출단위 구성

캐나다의 주(Province) 지역들은 지리적인 경계에 의해 여러 개의 경제구역(ER:Economic Regions)들로 분할되며, 현재 캐나다 ER 지역은 총 72개가 존재한다. LFS에서는 1960년대 이후로 이러한 ER 지역들을 캐나다 노동력 조사의 1차 층(primary strata)으로 이용해오고 있다. 주 지역 및 전국 단위의 통계 작성에 ER 지역들이 이용된다.

초기의 LFS에서는 ER 지역이 표본설계 시 반영되었던 주 지역 내의 유일한 행정구역이었고 노동력 조사의 관심은 이러한 ER 지역에 집중되었으나, 1989년부터 HRDC(Human Resources Development Canada)의 자금지원에 의해 16,500개의 표본조사가구가 LFS에 추가되면서 EIER 지역 에 대한 노동력 조사도 추가적으로 가능하게 되었다. 따라서 현재의 노동력 조사에서는 ER 지역과 EIER 지역 모두가 표본설계 시 층화에 반영되며, 추가표본은 주로 EIER 지역의 추정치의 신뢰도를 확보하기 위해 할당된다. 여기에서 ER 지역과 EIER 지역은 서로 조사 목적이 상이한 지역들이며 133개 지역이 서로 중복되어 조사된다. 한편 LFS 층화 시 반영되는 행정구역으로 CMA 지역을 들 수 있다. CMA 지역은 인구 100,000명 이상인 지역들로써 현재의 CMA 지역은 EIER 지역과 정확히 일치한다.

대영역 내에서의 세부적인 층화는 지리적인 구분에 관계없이 집락화 알고리즘에 의해 이루어진다. 그룹들 간의 가중 제곱합을 최소화하는 층화변수들을 이용하여 가능한한 동질적인 층으로 분할되며, 세부적인 알

고리즘은 Drew et al.(1985)과 Singh et al.(1990)에서 참조할 수 있다. 층화 알고리즘에 이용되는 층화변수들은 다음과 같다. 이 층화변수들은 1991년 센서스 자료를 이용하여 선정되었으며, 각각의 층화변수들은 전체인구의 2%이상을 설명할 수 변수들로 선정되었다.

- o 농업부문 종사자 수
- o 임업, 어업부문 종사자 수
- o 광업부문 종사자 수
- o 제조업(소비재분야) 종사자 수
- o 제조업(고무, 플라스틱, 가죽분야) 종사자 수
- o 제조업(섬유, 의류 분야) 종사자 수
- o 제조업(가구, 펄프, 제지, 인쇄, 목재분야) 종사자 수
- o 제조업(금속, 광업분야) 종사자 수
- o 제조업(석유화학, 화학분야) 종사자 수
- o 운수업 부문 종사자 수
- o 건설업 부문 종사자 수
- o 서비스업(상업분야) 종사자 수
- o 서비스업(금융분야) 종사자 수
- o 서비스업(개인/사업분야) 종사자 수
- o 서비스업(정부분야) 종사자 수
- o 종사인원 총계
- o 총 소득
- o 15세 이상 인구

o 15-24세 인구

o 55세 이상 인구

o 1인 거주 가구 수

o 2인 거주 가구 수

o 개인 소유 가구 수

o 총 임차료

o 고졸학력 인구

o 영어를 모국어로 하는 인구

o 프랑스어를 모국어로 하는 인구

o 영어/프랑스어 이외의 언어를 모국어로 하는 인구

LFS 추출틀은 농촌 지역, 인구 50,000명 이상의 대도시 지역과 소도시 지역의 3가지 유형의 지역들로 구분된다. 각 지역에 대한 층화는 다음과 같은 방법으로 이루어진다.

농촌지역에서 층화는 EI 지역과 EIER 지역의 교집합 내에 있는 2~3개의 CD(Census Division) 지역들을 함께 묶어 지리적 층으로 구성하였다.

캐나다 내의 대도시 지역인 17개의 CMA 지역들에 대해서는 충분한 수의 아파트들이 있기 때문에 각각의 CMA 지역들에 대해 독립적인 아파트 추출틀을 작성하며, 아파트 추출틀을 제외한 나머지 지역에 대해서는 일종의 지역 추출틀(area frame)을 형성하였다. 또한 $100,000 이상의 평균 소득을 갖는 고소득 지역들은 독립된 층으로 구성하여 고소득 가구들에

대한 대표성을 제고하였고, 부수적으로 소득관련 조사 및 소득관련 정보 수집이 용이하도록 하였다. 이러한 부수적인 정보는 고소득층에 대한 무응답의 경향을 분석하고자 하는 경우에도 유용하게 이용될 수 있다. 아파트 추출틀과 고소득 층을 제외한 나머지 지역들은 SNF(Street Network File) 지역으로 구분하였다. 최종적으로 마지막 층에는 적어도 48가구의 표본이 배정되도록 하였다.

LFS 표본설계에서 대규모 CMA 지역들에 대해서는 아파트 추출틀을 사용해오고 있다. 현재 18개 CMA 지역에서 표본 추출틀로써 이 목록이 이용되며, 각 CMA 지역에서 새로운 아파트가 건설되면 바로 표본목록에 추가된다. 또한 7개의 CMA 지역에 대해 평균소득이 $20,000 미만의 저소득 아파트 단지를 파악하여 저소득 아파트 층을 구성하여, 고소득 아파트 층과 더불어 소득관련 정보를 획득한다. 캐나다의 각 CMA 지역에 대한 아파트 추출틀 층화 현황은 다음 <표1>에 주어졌다.

<표1> 아파트 추출틀 층화

CMA	지리적 층	층의 총수	CMA	지리적 층	층의 총수
Halifax	2	2	London	1	2
Quebec	2	2	Windsor	1	2
Montreal*	4	9	Winnipeg*	1	6
Ottawa-Hull*	3	6	Saskatoon	1	1
Oshawa	1	2	Calgary*	1	3
Toronto*	6	16	Edmonton*	1	3
Hamilton	2	4	Vancouver*	4	6
St. Catharines	1	1	Victoria	1	1
Kitchener	2	2	Total	34	68

(주) ① "*" 는 저소득 아파트 층을 갖는 CMA 지역을 표시함
② "층의 총수"는 저소득 아파트 층을 포함한 수임

대부분의 소도시에서는 EA(enumeration area) 지역을 층화단위로 이용하였다. 조사 비용을 절약하기 위해 각 층에서 직접 표본가구를 추출하지 않고, 우선 각 층을 여러 개의 집락으로 구분하고 추출된 집락 내에서 표본가구들을 추출하였다. 통상적으로 농촌지역에서는 EA 지역이 집락으로 사용되었고, 도시지역에서는 좀 더 다양한 형태의 집락들이 이용되었다. 다음 <표2>는 LFS 표본설계에 이용된 일 단계 단위(first-stage unit)의 유형들을 요약한 것이다. 여기에서 "추출가구 수"는 LFS에서 조사되는 가구 수를 나타낸다. 집락과 표본가구에 대한 추출방법은 3장에서 논의하기로 한다.

<표2> 일단계 단위(first-stage unit), 단위당 가구 수와 추출가구 수

지 역	추출단위	단위당 가구수	추출가구수
Toronto, Montreal, Vancouver	cluster	200-250	6
Other cities	cluster	150-200	8
Apartment frame	apartment	varies	5
Most rural areas and non-SNF prarts of cities	EA	300	10

3. 표본배정, 추출, 순환

(Sample Allocation, Selection and Rotation)

LFS의 표본설계는 캐나다 전체, 주(Province) 지역, EIER 지역, CMA 지역, ER 지역에 대해 각각 다음과 같은 실업자 추정값의 목표 CV값이 만족되도록 설계되었다. 캐나다 전체의 실업자 추정치의 목표 CV 는 약 2%이내, 주(Province) 지역의 CV 값은 약 4~7%선에서 관리되도록 하였다. EIER 지역과 CMA 지역은 분기 실업자 추정치의 CV 값이 15% 이내가 되도록 하였고, 여기에서 하나의 EIER 지역에 배당되는 최소 표본크기는 매월 600가구가 배정된다. ER 지역에 대한 분기별 목표 CV 값은 25% 이내가 되기를 기대하고 있다. 캐나다의 각 주 내의 ER 지역, EIER 지역과 CMA 지역의 현황은 다음 <표3>과 같다.

LFS의 표본크기는 총 52,350가구로써 HRDC의 자금지원에 의해 추가된 16,500 표본가구를 제외한 35,850 표본가구는 전국 및 주(Province) 지역의 최적 추정을 위해 배정되며, 이 표본을 핵심표본(core sample)이라고 지칭한다. 핵심표본은 각 주의 실업률 추정의 목표정도를 만족하도록 배정된다. 주 내에서 각각의 ER 지역에 대한 핵심표본의 배정은 총 가구수에 비례하여 배정하며, 최소한 200~300 표본가구가 배정되도록 하였다. 또한 16,500 가구의 추가표본은 핵심표본을 보충하여 EIER 지역의 추정의 정확도를 높이기 위해 추가로 배정된다. 우선 핵심표본 만으로 EIER 지역에 대한 CV 값을 계산 한 후, CV 값이 큰 EIER 지역에 대해

추가 표본을 배정하는 형식을 취하며, EIER 지역에 대해서는 최소한 600 가구가 배정되도록 하였다.

<표3> ER, EIER, CMA 지역 현황

주(Province)	ERs	EIERs	CMAs
Newfoundland	4	3	1
Prince Edward I	1	1	0
Nova Scotia	5	5	1
New Brunswick	5	4	1
Quebec	16	13	6*
Ontario	11	18	10*
Manitoba	8	3	1
Saskatchewan	6	4	2
Alberta	8	4	2
British Columbia	8	6	2
Canada	72	61	25*

(주) Ottawa-Hull CMA 는 Ontario와 Quebec 양쪽에서 카운트 됨.

LFS에서 표본 추출은 대부분의 지역에서 이 단계 추출(two-stage sampling) 만을 이용한다. 일 단계 추출은 지역 추출(area sample)이고, 이 단계 추출은 일 단계 추출 지역들에 대해 거주가구 목록을 작성하여 이 목록으로부터 표본가구를 추출한다.

농촌지역의 경우 일 단계 추출단위는 EA 지역, 이 단계 추출단위는 가구 단위가 된다. 일 단계의 지역 추출은 확률비례계통추출법, 이 단계

의 가구 추출은 계통추출법이 이용된다. 아파트 추출틀을 이용하지 않는 주요 도시지역의 경우 일 단계 추출단위는 EA 지역, 이 단계 추출단위는 가구 단위가 된다. 일 단계의 지역 추출은 Rao-Hartley-Cochran(RHC)의 랜덤그룹법이 이용되며, 이 단계의 가구 추출은 계통추출법이 이용된다. 아파트 추출틀을 이용하지 않는 기타 도시 지역들도 몇몇 특별한 경우를 제외하고는 주요 도시지역과 유사한 방법을 취하고 있다. 랜덤그룹법은 인구유입 및 유출양상을 LFS의 표본크기에 반영시킨 방법으로써 시기에 따라 유동성이 큰 대도시 지역의 인구변화를 표본크기에 반영시킨 방법이다. 아파트 추출틀을 이용하는 주요 도시지역의 경우 일 단계 추출단위는 아파트 단지, 이 단계 추출단위는 가구 단위가 된다. 일 단계 아파트 단지에 대한 추출은 확률비례계통추출법, 이 단계 가구단위의 추출은 계통추출법이 이용된다.

LFS에서는 표본의 일부가 매달 새로운 표본으로 교체된다. 다 단계 표본설계의 매 단계에서 표본 단위의 교체가 이루어지며, 표본추출의 최종 단위인 가구는 6개월마다 교체된다. 조사자의 업무 부담과 응답가구가 오랜 기간 동안 표본으로 조사되는 동안 발생할 수 있는 무응답에 대한 가능성을 최소화하기 위해 매 달 1/6의 표본이 대체된다. 따라서 하나의 집락에 포함된 표본가구는 연속적으로 6개월 간 조사된 후 표본에서 완전히 삭제되고 새로운 표본으로 대체된다.

4. 특별조사와 보충조사

캐나다 노동력조사 외에 추가적인 많은 조사가 LFS 추출틀 또는 LFS 표본을 이용하여 실시되며, 이러한 조사들은 캐나다 정부 부처의 자금 지원에 의해 시행된다. 다음 <표4>는 LFS 연동교체표본 또는 LFS 추출틀을 이용한 특별조사 및 보충조사 현황을 요약한 것이다.

<표4> 특별조사 및 보충조사 현황(1998년 현재)

Survey	조사기간	Survey	조사기간
Canadian Travel Survey	1-12월(매월조사)	Homeowner Repair and Renovation Survey	3월
Employment Insurance Coverage	1월	Survey of Consumer Finances	4월
Survey of Household Spending	1월-3월	Cultural Capital Survey	4월
Survey of Labour and Income Dynamics	1월, 5월	National Population Health Survey	2, 6, 8, 11월
Adult Education and Training Survey	1월	National Longitudinal Survey of Children	11월
Resident Telephon Services Survey	2, 5, 8, 11월	Survey of Work Arrangements	11월
Survey of Household Energy Use	2월		

SCF(Survey of Consumer Finances)는 일년에 한번 시행되는 소비자 재정에 관한 조사로써 보통 4월에 실시된다. 모든 가구들이 4개의 순환 그룹에 배정되어 LFS 조사에 추가된다. 4월의 LFS 조사에 앞서 각 가정에 우편으로 설문지가 배달되고 LFS 조사 기간 동안 회수된다. SCF에서 조사하는 주요 정보는 소비자의 평균 소득과 세금 공제 전과 공제 후의 소

득관련 정보들로써, 이러한 결과들은 저소득 기준점 결정 등과 같은 소득 관련 측도들로 이용된다.

SHS(Survey of Household Spending)는 일년에 한번 실시되는 가계비 지출 및 식료품 지출 조사로써 보통 1월~3월에 걸쳐 시행되며, 소비자 가격 지표 산정의 정보로써 이용된다. SHS 조사는 LFS 표본을 포함하는 집락들에서 표본가구를 추출하나 표본가구들은 LFS 조사와는 별도로 조사된다.

SLID(Survey of Labour and Income Dynamics)는 노동력과 소득 변천과정 파악을 조사의 목적으로 일년에 2번, 1월과 5월에 조사가 이루어지며 LFS와 병행하여 시행된다. 조사 결과는 저소득 층의 유입, 유출 동향, 노동 시장의 변화, 가족변화와 경제적인 복지와의 상관관계 등을 분석하기 위해 이용된다.

NPHS(National Population Health Survey)는 일종의 국민 건강조사로써 분기별로 실시되며 국민 건강의 계절적 요인을 파악하기 위해 NPHS 표본을 2월, 6월, 8월과 11월조사에 각각 1/4씩 배분한다. NPHS 조사는 주정부의 자금지원에 의해 실시되며, LFS 조사와 병행하여 실시되지는 않는다. 초기에 선택된 가구의 구성원은 2년에 한번 심층 면접을 받으며 20년 동안 지속적으로 관리된다. 기초적인 건강정보는 거주 가구의 모든 구성원들에 대해서 취합되며 이러한 시계열 자료는 횡단면 추정 목적에 이용된다.

NLSCY(National Longitudinal Survey of Children and Youth)는 아동에

대해 유아기부터 성인기까지의 발달 과정을 모니터하는 조사로써 일년에 한번 실시되는 일종의 시계열적 장기조사에 해당한다. LFS 조사가구의 약 30%만이 대상연령에 있는 아동을 포함하는 관계로 NPHS의 추가표본이 조사에 이용되며, 조사방법은 NPHS 조사와 유사하다.

5. 가중치와 추정

LFS의 가중치는 다음의 세가지 요인들에 기인하여 작성된다. 표본설계를 반영하는 일종의 설계 가중치, 무응답 가구를 보정하는 일종의 무응답 보정 가중치와 모집단 총계에 표본 추정치를 일치시키는 일종의 사후층화 가중치(g-factor)의 3가지 요인들에 의해 LFS의 최종 가중치가 결정된다. LFS 조사 추정치는 LFS 표본이 확률표본이기 때문에 추정치의 표본오차를 추정하여 신뢰도를 판단할 수 있다. 표본설계에 계획되지 않은 관심지역들에 대한 추정문제는 소지역 추정기법을 도입하여 추정량의 신뢰도를 확보하였다. EIER 지역들이 여기에 해당된다.

LFS에서는 가구단위에 대해 계통추출이 이루어지는 마지막 단계를 제외하고는 모든 단계에서 확률비례추출법으로 표본이 추출되는 층화 다단계 추출법이 이용된다. 설계가중치는 이러한 복합표본설계에서 가장 기본적인 가중치로써 각 추출단위에 대해서 추출률의 역수로 결정된다. 예를 들어 층에 대한 설계가중치는 $R_h = N_h / n_h$와 같이 표현될 수 있다. 여기에서 n_h는 층 h에 대한 표본가구 수, N_h는 표본설계 시 층 h에 있는 총 가구 수를 나타낸다.

이 단계 추출의 경우는 다음과 같이 설명될 수 있다. n^*_{hj}를 층 h에 있는 j번째의 일 단계 추출단위(FSU: First Stage Unit)라고 하자. 추출해야할 FSU의 수는 $n_{1h} = n_h / n^*_{hj}$ 로 주어진다. h층에서 j번째 FSU에 있는 가구들의 수를 N_{hj}라 하면 j번째의 FSU에 대한 추출율은

$R_{hj}=N_{hj}/n^*_{hj}$로 주어지며, 이때 j번째 FSU에 대한 일 단계 산입확률(inclusion probability)은 $\pi_{1hj}=\frac{n_{1h}}{\sum_{j\in h}R_{hj}}R_{hj}$로 나타낼 수 있다. j번째 FSU가 주어진 상태에서 k번째 가구가 선택될 조건부 산입확률은 $\pi_{k/j}=\frac{n^*_{hj}}{N_{hj}}=\frac{1}{R_{hj}}$로 주어지며, h번째 층에서 k번째 가구에 대한 산입확률은 $\pi_{hk}=\pi_{1hj}\cdot\pi_{k/j}=\frac{n_{1h}}{\sum_{j\in h}R_{hj}}R_{hj}\frac{1}{R_{hj}}=\frac{n_{1h}}{\sum_{j\in h}R_{hj}}$로 계산될 수 있다. 여기에서 $\sum_{j\in h}R_{hj}=\sum_{j\in h}\frac{N_{hj}}{n^*_{hj}}=\frac{n_{1h}}{n_h}\sum_{j\in h}N_{hj}=n_{1h}R_h$ 인 관계가 성립하며, 결국 산입확률은 $1/R_h$과 같게된다. 참고적으로 LFS에서는 기본적으로 각 층내에서 동일한 설계 가중치(R_h)를 이용한다.

LFS와 같은 연속조사에서는 고정된 추출률을 유지할 경우, 인구 증가 및 인구 유입 등으로 시간이 지남에 따라 표본 수도 증가하게 된다. 급속한 인구 증가가 발생하는 특정지역의 표본 수는 급격히 증가되어 조사원의 업무부담을 가중시키고 조사의 질을 떨어뜨릴 가능성이 있다. 이러한 문제점을 해결하기 위해 LFS 표본 설계에서는 이러한 집락들에 대해서 표본을 부차추출(subsampling)하여 전체 표본크기를 일정하게 유지한 후 집락 부차가중치(cluster subweight)를 산출하여 LFS 설계가중치에 추가하여 사용하였다. 구체적으로 SC(Subclustering)방법, SRC(Self- Representing Cluster)방법, CSS(Cluster Subsampling)방법 등을 이용한다.

SC 방법은 인구유입 등으로 인해 원래의 집락의 크기가 약 3배를 초

과할 때 이러한 집락들을 여러개의 작은 집락으로 분할하여 표본 가구를 추출하는 방법이다. 분할된 작은 집락들의 표본 수를 n_{2hj}라 하자. N_{hji}를 새로운 집락의 크기, n_{hji}를 새로운 집락의 표본크기, R_{hji}를 새로운 집락들에 대한 추출률이라 할때, 새로운 부차집락들로부터 집락추출률 $R^*_{hj}=\sum_{i\in j}\frac{R_{hji}}{n_{2hj}}$ 를 얻을 수 있다. 이때 집락 부차가중치는 $K=R^*_{hj}/R_{hj}$로 주어지며, 설계가중치는 원래의 가중치에 이러한 부차가중치를 곱하여 계산된다.

SRC 방법은 임의의 층에서 증가하는 거주단위들의 특성이 나머지 단위들의 특성과 구별되거나 집락의 규모가 층 규모의 20%를 초과할 경우, 해당 집락을 하나의 층으로 재분류하여 가중치를 결정하고, 나머지 집락들에 대해서는 가중치를 재 보정해 주는 방법이다. 이렇게 생성된 새로운 층을 (hj) 라 하자. 이러한 층내에서 새로운 집락들이 구성되어 표본들이 추출된다. $N^N_{(hj)}$을 새로운 층의 크기, $n^N_{(hj)}$을 새로운 층의 표본크기라 할 때, 층 추출률은 $R^N_{(hj)}=N^N_{(hj)}/n^N_{(hj)}$으로 주어지며, 이러한 새로운 층으로부터 추출된 가구들은 가중치 $K=R^N_{(hj)}/R_h$가 배정된다. 나머지 집락들의 가구들에 대한 가중치는 다음과 같은 방법으로 보정된다. 하나의 층에서 6개의 집락들이 추출되었다고 하자. 층 내에서 1개의 성장집락이 발생하여 새로운 층으로 재 분류 되었다면 나머지 5개의 집락의 가구들에 대한 가중치는 보정되어야만 한다. $N^R_h=N_h-N^N_{(hj)}$를 새로운 층을 제외한 층의 크기, $n^R_h=n_h-n_{hj}$을 표본의 크기라 할때, 층에 대한 추출률은

$R_h^R = N_h^R / n_h^R$로 계산되며, 이때 집락 부차가중치는 $K = R_h^R / R_h$로 결정된다.

CSS 방법은 집락단위가 부차표본으로 추출되는 경우에 이용된다. 임의의 집락이 추출률 R_{hj}로 추출되고, 부차추출이 추출률 R_{hj}^*로 추출되었다면, 이때 집락 부차가중치로써 $K = R_{hj}^* / R_{hj}$를 이용하는 방법이다.

표본추출의 마지막 단계는 계통추출이 이용된다. 가구에 대한 추출률은 일관되게 적용되기 때문에 인구 증가로 인한 표본가구수의 증가는 조사규모 및 조사비용의 증가를 초래한다. 이러한 조사비용을 통제하기 위해 표본 수 안정화 작업이 수행된다. 표본 수 안정화 작업은 전체 표본가구 수를 적절한 수준에서 유지하기 위해 초과 표본들을 랜덤하게 제거하는 작업을 말한다. 이러한 과정에서 가구단위의 산입확률(inclusion probability)은 당연히 바뀌게 된다. 현 표본설계에서는 같은 EIER 지역에 속해 있고 같은 순환그룹에 포함되어 있는 모든 가구단위들을 안정화 지역(stabilization area)으로 정의하고, 각 안정화 지역(a)에 대해서 표본크기가 결정된다. 안정화 지역 a의 표본크기를 b_a라고 나타내자. 안정화 지역에서 안정화작업을 거치지 않은 표본크기를 n_a라 할 때, n_a가 b_a를 초과한다면 $n_a - b_a$만큼의 표본가구가 랜덤하게 제거된다. LFS에서는 임의의 집락이 부차추출 되었을 경우에는 이 집락을 안정화 작업에서 제외시키기 때문에 이러한 집락에 포함된 가구단위들은 안정화 가중치(stabilization weight)의 영향을 받지 않는다. 이러한 집락을 제외시킨 안정화 지역 a의 가구단위의 총계를 c_a라 할때, 지역 a에 있는 조사가구의

안정화 가중치는 $s_a=(n_a-c_a)/(b_a-c_a)$ 이 이용된다.

통계조사에서 무응답은 항목 무응답(item nonresponse)과 단위 무응답(unit nonresponse)의 두 가지 유형으로 분류된다. LFS에서는 항목 무응답은 대체법(imputation)으로, 단위 무응답은 전체적인 가중치 조정을 통해 처리하고 있다. 항목 무응답은 지리적으로 또는 인구 통계적으로 유사한 특성을 갖는 응답자의 응답패턴을 이용하여 대체된다. LFS에서는 단위 무응답을 처리하기 위한 무응답 층을 " 같은 EIER 지역에 속하고, 같은 유형의 지역적 특성을 가지며, 같은 표본순환그룹 내에 있는 가구들"로써 정의한다. 무응답 층을 올 바르게 구성하였을 경우 동일한 무응답 층에 속한 응답 가구와 무응답 가구는 서로 비슷한 속성을 갖기 때문에 응답 가구가 무응답 가구를 대표한다고 가정할 수 있게 된다. 단위 무응답에 대한 보정은 설계 가중치에 보정요인 $f_b=\sum_{k=1}^{n}\pi_k^{-1}/\sum_{k=1}^{r}\pi_k^{-1}$ 을 곱하여 계산한다. 여기에서 π_k^{-1} 은 각 표본가구에 부여된 설계 가중치, n 은 무응답 층 b 에 있는 표본가구 수, r 은 응답가구 수를 나타낸다.

LFS에서의 최종 가중치는 특별한 경우를 제외하고는 설계 가중치와 보조정보로부터 얻게되는 사후 가중치(g-factor)의 곱으로 계산되며, 여기에서 사후 가중치 계산은 일반적인 회귀추정방법이 이용된다. 자세한 추정절차는 Lemaitre and Dufour (1987)에 소개되어 있다. 먼저 다음과 같은 기호를 정의하자.

p = 1, 2, ···, 10 : 주 지역을 나타내는 기호,

u= 1, 2, ···, U : p번째 주 내에 있는 EIER 지역,

f = 1, 2, ···, F : u번째 EIER 지역 내에 있는 추출틀의 형태,

h = 1, 2, ···, H : 추출틀 f 내에 있는 층,

r = 1, 2, ···, 6 : h번째 층 내에 있는 순환그룹,

j = 1, 2, ···, J : 순환그룹 r 의 집락,

k = 1, 2, ···, K : 집락 j에서의 가구,

i = 1, 2, ···, c_k : k번째 가구 내에 있는 구성원,

SC(subclustering) 방법은 우선 해당 집락을 여러개의 작은 집락들로 재구성한 후 표본 추출을 위한 집락들을 선정하고 전체 표본 수를 참고하여 선정된 집락 내에서 표본가구를 추출한다. 원래의 집락 추출율이 $R_{pufh \cdot j}$, 해당 집락 추출율이 $R^*_{pufh \cdot j}$라 하면 해당 집락의 부차 가중치는 $c_{pufh \cdot j} = R^*_{pufh \cdot j} / R_{pufh \cdot j}$ 이다.

SRC(self-representing cluster) 방법은 층 내에서 성장 집락을 분리하여 새로운 층(h)으로 구성하고 h층 내에서 여러개의 집락들을 재 구성한 후 표본가구를 추출하는 방법이다. 원래의 층 추출율이 R_{pufh}, 새로 형성된 층의 추출율이 R^*_{pufh}라 하면 이때 새로 형성된 층에서 가구단위들에 배정되는 집락 부차가중치(cluster subweight)는 $c_{pufh} = R^*_{pufh} / R_{pufh}$ 이다. 성장 집락을 제외한 나머지 집락의 가구단위에 대해서는 보정된 집락 부차가중치 $c_{pufh} = R^R_{pufh} / R_{pufh}$를 적용한다. 여기에서 R^R_{pufh}은 나머지 층의 추출율을 나타낸다.

CSS(cluster Subsampling) 방법은 추출된 가구단위들이 부차추출되고

이러한 부차추출된 가구단위들만 조사하는 방법이다. 원래의 집락 추출율이 $R_{pufh\cdot j}$이고 부차추출하기 위한 해당 집락 추출률이 $R^*_{pufh\cdot j}$라면 이때 집락 부차가중치는 $c_{pufh\cdot j}=R^*_{pufh\cdot j}/R_{pufh\cdot j}$이다.

안정화 가중치(stabilization weight)는 앞서 언급되었던 안정화지역(stabilization area) 내에서만 적용된다. 각 안정화 지역 내에서의 표본크기를 $b_{pu\cdot\cdot r}$, 실제 추출된 표본크기를 $n_{pu\cdot\cdot r}$, 안정화 지역에서 CSS 방법으로 추출된 표본의 크기를 $c_{pu\cdot\cdot r}$라 할때, 안정화 가중치는 $s_{pu\cdot\cdot r}=\frac{n_{pu\cdot\cdot r}-c_{pu\cdot\cdot r}}{b_{pu\cdot\cdot r}-c_{pu\cdot\cdot r}}$ 로 계산된다.

이상의 가중치들을 이용하여 조사가구에 대한 설계가중치를 산출하며 LFS에서는 다음과 같은 설계가중치를 고려한다.

$$\pi^{-1}_{pufhrjk}=w_{pufh}\times c_{pufh\cdot j}\times s_{pu\cdot\cdot r},$$

여기에서 w_{pufh}는 같은 층 내에 있는 모든 가구단위들에 대해서 동일한 가중치를 배정했던 표본설계 당시의 가중치를 나타낸다. 표기상의 편의를 위해 앞으로 $\pi^{-1}_{pufhrjk}$를 π^{-1}_k로 나타내기로 한다.

LFS에서는 무응답 층에 대해서 무응답 보정을 실시하며, 보정 가중치로써 $f_{puf\cdot r}=\sum_{k\in s}\pi_k^{-1}/\sum_{k\in r}\pi_k^{-1}$를 이용한다. 여기에서 분자의 s에 대한 합은 무응답 층에 있는 모든 가구들에 대한 합을 나타내며, 분모의 r에 대한 합은 무응답 층에 있는 모든 응답가구들에 대한 합을 나타낸다. 같은 무응답 층에 속해 있는 모든 가구단위들은 동일한 무응답 가중치를 갖는

다.

무응답 보정요소가 추가될 경우 부차 가중치는 $a_k = f_{puf..r} \times \pi_k^{-1}$ 와 같이 설계가중치와 무응답 보정가중치의 곱으로 표현된다. 즉 같은 조사가구 내의 모든 구성원들은 동일한 부차가중치를 갖는다.

위에서 언급한 부차가중치를 이용하여 고용 인구 Y에 대한 총계 추정값을 산출해 보자. 모집단에서 고용인구의 총계를 $t_y = \sum_U y_i$ 라고 하자. 여기에서 U에 대한 합은 모집단에서 관심영역 내에 있는 모든 구성원들의 합을 나타내며, y_i는 조사대상자가 고용일 경우 1, 아닐 경우 0의 값을 갖는다. 이때 표본조사에 의한 총계 추정치는 부차가중치에 의존하며 $\hat{t}_{ya} = \sum_s y_i a_i$ 와 같이 표현될 수 있다. 여기에서 s에 대한 합은 표본으로 추출된 조사 대상자들에 대한 합을 나타내고, a_i는 부차가중치를 의미한다. 위의 t_y와 $\hat{t}_{ya}$는 각각 다음과 같이 다시 표현할 수 있다.

$$t_y = \sum_{k=1}^{N} \sum_{i=1}^{c_k} y_i = \sum_{k=1}^{N} y_k \quad , \quad \hat{t}_{ya} = \sum_{k=1}^{n} a_k \sum_{i=1}^{c_k} y_i = \sum_{k=1}^{n} y_k a_k \ ,$$

여기에서 c_k는 k번째 조사가구의 구성원의 수, N은 모집단의 가구 수, n은 표본 가구 수, y_k는 $y_k = \sum_{i \in k} y_i$를 의미하며, k는 가구 총 수, i는 구성원을 나타낸다.

LFS에서 적용하고 있는 마지막 단계의 가중치로써 사후층화 가중치(g-factor)를 들 수 있다. 사후층화 가중치는 사후층화를 통한 보조정보로부터 획득하며 회귀추정방법을 이용하여 산출한다. 각 주 단위의 성별-연

령대별 그룹, ER 지역과 CMA 지역에 대한 인구 총계, 센서스 결과에 의한 인구 추계 정보 등이 보조정보로 활용되었다. 추가적인 논의를 위해 다음과 같은 기호를 정의하기로 하자.

y_i : i번째 조사자에 대한 특성치,

y_k : k번째 가구에 대한 특성치 총계,

Q : 추정에 이용된 보조변수의 수, $q=1,2,\cdots,Q$,

x_{qi} : 조사자 i에 대한 q번째 지표변수의 값, 지표변수는 조사자 i가 j번째 범주에 속할 경우 1, 기타 0의 값을 갖는다.

x_{qk} : k번째 가구단위에 속하는 조사자들에 대한 q번째 지표변수 값의 총계,

x_k : q번째 원소가 x_{qk}인 $Q\times 1$ 벡터,

c_k : k번째 가구의 크기,

$\hat{t}_{ya}$: 위에서 언급한 부차가중치에 근거한 추정치,

$\hat{t}_{x_qa}$: q번째 보조변수에 대한 부차가중치에 근거한 추정치.

사후층화 가중치를 산출하기 위해 $\hat{t}_{yr}=\hat{t}_{ya}+\sum_{q=1}^{Q}\hat{B}_q(t_{x_q}-\hat{t}_{x_qa})$와 같은 회귀추정량을 이용한다. 여기에서

$$\hat{t}_{x_qa}=\sum_s x_{qi}a_i,$$

$$\hat{B}=(\hat{B}_1,\cdots,\hat{B}_Q)^T=\left(\sum_{k=1}^{n}\frac{x_kx_k^Ta_k}{c_k}\right)^{-1}\sum_{k=1}^{n}\frac{x_ky_ka_k}{c_k}$$ 이며,

$(\sum_{k=1}^{n}\frac{x_k x_k^T a_k}{c_k})^{-1}$은 $Q\times Q$ 행렬, $\sum_{k=1}^{n}\frac{x_k y_k a_k}{c_k}$는 $Q\times 1$ 벡터를 나타낸다.

회귀추정량 $\hat{t}_{yr}$은 $\hat{t}_{yr}=\sum_{k\in s}y_k a_k g_k$와 같이 사후층화 가중치를 포함하는 식으로 재표현 할 수 있다. 여기서,

$$g_k=1+(t_x-\hat{t}_{xa})^T(\sum_{k\in s}\frac{x_k x_k^T a_k}{c_k})^{-1}\frac{x_k}{c_k}$$

로써 일명 g-factor로 불리우는 사후층화 가중치이다. 가구 구성원에 대한 사후층화 가중치는 $g_i=1+(t_x-\hat{t}_{xa})^T(\sum_{i\in s}z_i z_i^T a_i)^{-1}z_i$이며, 여기에서 $z_i=\frac{1}{c_k}\sum_{i=1}^{c_k}x_i$이다. 즉 사후층화 가중치의 특징은 가구에 대한 가중치와 가구 내의 구성원에 대한 가중치가 일치하여 모든 구성원들이 동일한 가중치를 갖는다는 점이다.

LFS의 최종가중치는 설계 가중치, 무응답 조정 가중치와 사후층화 가중치의 곱으로 표현되며, 이러한 최종가중치를 이용하여 주 지역 및 전국 단위의 경제활동인구 총계, 취업자 총계, 실업자 총계, 취업 및 실업률 등이 추정된다.

LFS에서 추정량의 분산계산은 잭나이프 방법을 이용한다. 좀 더 일반적인 경우의 잭나이프 방법에 대한 기술은 Wolter(1985)를 참조할 수 있으며, 여기에서는 LFS에서 적용하는 잭나이프 알고리즘을 소개하기로 한다.

(ⅰ) 잭나이프 방법을 적용하기 위해 h번째 층은 J_h개의 반복표본을 갖는다고 가정한다($a=1,2,\cdots,J_h$). 우선 특정 반복표본에 해당하는 모든 가구들을 제거한다. 여기에서 해당 표본에서 반복 총계는 $J=\sum_{h=1}^{H} J_h$이고, H는 해당 표본에서 층의 총계를 나타낸다.

(ⅱ) 주어진 층에서 나머지 J_h-1 개의 반복표본의 모든 가구들에 대해 부차가중치에 대한 보정이 이루어진다. 보정된 가중치의 값은 $a_k^{adj}=\frac{J_h}{(J_h-1)}\,a_k$ 이다.

(ⅲ) 보정된 부차가중치와 남아있는 해당표본을 이용하여 관심 추정치 $\hat{t}_{yr(ha)}$를 계산하기 위한 최종가중치를 계산한다. 여기에서 (ha)는 h번째 층으로부터 a번째 반복이 제거되었다는 것을 나타낸다.

해당표본의 모든 반복에 대해서 위의 (ⅰ)~(ⅲ)의 절차가 반복되며, 결과로써 관심 추정치에 대한 J개의 서로 다른 추정값을 얻게된다. 이러한 추정값을 이용하여 추정값의 분산을 계산하며 다음과 같은 분산 추정 공식을 이용한다.

$$\hat{V}(\hat{t}_{yr})=\sum_{h=1}^{H}\frac{(J_h-1)}{J_h}\sum_{a=1}^{J_h}(\hat{t}_{yr(ha)}-\hat{t}_{yr})^2 .$$

실업률에 대한 분산 추정은 다음의 추정공식을 이용할 수 있다.

$$\hat{V}(100\frac{\hat{t}_{yr}}{\hat{t}_{zr}})=100^2\sum_{h=1}^{H}\frac{(J_h-1)}{J_h}\sum_{a=1}^{J_h}(\frac{\hat{t}_{yr(ha)}}{\hat{t}_{zr(ha)}}-\frac{\hat{t}_{yr}}{\hat{t}_{zr}})^2 ,$$

여기에서 y는 실업인구 총계, z는 경제활동인구 총계를 나타내며, 위의

결과는 실업률 $100(y/z)\%$ 에 대한 잭나이프 분산 추정공식이다.

월 변화량의 추정치에 대한 잭나이프 분산추정을 다음과 같이 산출할 수 있다. 연속되는 두 달의 월 추정치로부터 다음의 차분추정치 $\hat{D}_{yr} = \hat{t}_{yr}^{\,2} - \hat{t}_{yr}^{\,1}$ 를 고려하자. 여기에서 윗 첨자는 연속되는 월을 나타낸다. 대응되는 잭나이프 추정치는 $\hat{D}_{yr(ha)} = \hat{t}_{yr(ha)}^{\,2} - \hat{t}_{yr(ha)}^{\,1}$ 으로 표현할 수 있다. 이때 분산 추정치는 다음과 같이 주어질 수 있다.

$$\hat{V}(\hat{D}_{yr}) = \sum_{h=1}^{H} \frac{(J_h - 1)}{J_h} \sum_{a=1}^{J_h} (\hat{D}_{yr(ha)} - \hat{D}_{yr})^2 .$$

LFS의 연속된 두 달의 조사에서는 5/6의 표본이 일치한다. 공통표본을 이용하여 두 달 간의 월 변화량의 차를 추정하는 것이 위의 추정방법에 비해 훨씬 효율적일 수 있다. Singh et al.(1997)은 이러한 공통표본을 이용하여 다음과 같은 합성추정량을 제안하였다.

$$est^{C}_{(t+1)} = K \times est_{(t+1)} + (1-K) \times [est^{C}_{(t)} + change_{common}],$$

여기에서 윗첨자 C는 복합추정법을 의미하며, $change_{common}$ 은 공통표본을 이용한 변화량을 나타낸다. 이 추정량은 1998년 현재 캐나다 통계청에서 채택하고 있지는 않지만 새로운 표본설계에서는 사용될 전망이다.

위와 유사한 방법으로 n 개의 월 추정치의 평균에 대한 분산추정을 고려할 수 있다. n 개의 월 추정치의 평균은 $\hat{A}_{yr} = \sum_{i=1}^{n} \frac{\hat{t}_{yr}^{\,i}}{n}$ 이고, 이에 대응하는 잭나이프 추정값은 $\hat{A}_{yr(ha)} = \sum_{i=1}^{n} \frac{\hat{t}_{yr(ha)}^{\,i}}{n}$ 로 계산할 수 있으며, 추정치의 분산은 다음의 추정공식을 이용할 수 있다.

$$\hat{V}(\hat{A}_{yr}) = \sum_{h=1}^{H} \frac{(J_h - 1)}{J_h} \sum_{a=1}^{J_h} (\hat{A}_{yr(ha)} - \hat{A}_{yr})^2.$$

6. 데이터 품질관리

모든 표본조사에서와 마찬가지로 LFS 추정값들도 표본 오차와 비표본 오차를 수반한다. 따라서 조사 추정값들이 올바르게 해석되기 위해서는 추정값들의 정도를 나타내는 측도에 관한 점검이 요구된다.

표본오차에 직접적으로 영향을 미치는 것은 표본크기라 할 수 있다. 일반적으로 표본크기가 증가함에 따라 표본오차는 감소한다. 표본의 크기 뿐만 아니라 모집단의 변동, 추정 및 표본설계의 방법과 같은 요인들이 표본오차에 관련된 요인들이라 할 수 있다. 표본오차는 층화방법, 표본할당, 추출단위의 선택에서 뿐 만 아니라 다 단계 표본설계에서 매 단계에서 선택된 표본 추출방법 및 추정방법 등과 같은 요인들에 크게 의존한다.

표본설계 및 추정방법에 대한 효율성을 점검하기 위한 측도로는 평균제곱오차(MSE)가 이용된다. MSE는 추정값과 모집단의 실제값과의 편차제곱합의 평균으로 보통 정의된다. 표본오차와 관련된 또 다른 중요한 측도로써 변동계수(CV)가 이용된다. Y가 어떤 특성치에 대한 추정치이고, d가 추정치의 표준오차라 할때, CV값은 $(d/Y)\times 100$ 으로 정의된다. 또한 추정치의 신뢰구간은 표준오차 d로부터 추론될 수 있다. 한편 시간에 따

른 표본설계의 낙후성을 평가하기 위한 지표로써 설계효과(design effect)가 이용된다. 설계효과는 기존 표본설계로부터 조사된 추정값의 분산과 단순임의표본으로부터 계산된 추정값의 분산의 비로써 정의되며, LFS에서는 표본설계효과(sample design effect)와 전체설계효과(overall design effect)와 같은 두 가지 형태의 설계효과를 계산한다. 표본설계효과는 모총계에 대한 가중치의 조정없이 부차가중치만을 이용하여 계산되며, 전체설계효과는 앞서 언급되었던 최종가중치를 반영하여 계산된다. 따라서 표본설계효과는 표본설계의 효율성만이 반영되며, 전체설계효과는 층화, 다단계추출, 사후층화 및 추정 등의 표본설계의 전반적인 사항들이 반영된다고 보면 된다.

1997년 1월부터 7월까지의 LFS 조사자료에 대한 주 단위 및 전국 단위의 고용 및 실업관련 추정치들의 CV 값이 다음 <표5>에 주어졌다. 고용과 실업에 대한 월 변화 추정값의 표준오차는 <표6>에, 고용과 실업에 관한 설계효과는 <표7>에 주어졌다

<표5> LFS 고용 및 실업 추정치들의 CV 값 (1997)

Province	Employed CV(%)	Unemployed CV(%)	Province	Employed CV(%)	Unemployed CV(%)
Newfoundland	2.2	6.1	Manitoba	0.91	6.5
Prince Edward Island	1.7	6.5	Saskatchewan	1.1	7.4
Nova Scotia	1.2	5.3	Alberta	0.76	5.9
New Brunswick	1.2	5.5	British Columbia	0.90	5.1
Quebec	0.79	3.5	Canada	0.32	1.72
Ontario	0.54	3.0			

<표6> LFS 고용 및 실업에 대한 월변화 추정값의 표준오차(1997)

Unit: thousands

Province	SE (employed)	SE (Unemployed)	Province	SE (employed)	SE (Unemployed)
Newfoundland	3	2	Manitoba	4	3
Prince Edward Island	1	1	Saskatchewan	3	2
Nova Scotia	4	3	Alberta	9	6
New Brunswick	3	2	British Columbia	12	9
Quebec	18	14	Canada	32	24
Ontario	20	15			

<표7> LFS 고용 및 실업에 대한 설계효과 (1997)

Province	Employed		Unemployed	
	Sample	Overall	Sample	Overall
Newfoundland	2.7	0.83	1.4	1.3
Prince Edward Island	2.0	0.53	1.1	1.1
Nova Scotia	2.2	0.51	1.2	1.1
New Brunswick	2.0	0.56	1.4	1.4
Quebec	2.1	0.55	1.1	1.0
Ontario	3.3	0.50	1.2	1.1
Manitoba	2.2	0.41	1.1	1.1
Saskatchewan	2.4	0.63	1.2	1.2
Alberta	4.1	0.40	1.1	1.1
British Columbia	2.1	0.50	1.2	1.1
Canada	2.8	0.51	1.2	1.1

비표본오차는 표본조사의 매 단계에서 발생할 수 있으며 주로 조사원의 무관심, 오해 및 잘못된 해석 등의 사유에 기인하며, 추정값의 편향 및 변동에 직접적인 영향을 미친다. 관측값의 수가 많거나 혹은 대영역의 조사에서는 비표본오차에 기인한 효과는 무시될 수도 있는 양이나, 소지역 추정의 문제에서는 민감한 문제로 인식되어진다. 비표본 편향 및 분산은 조사원에 대한 교육 및 조사원의 태도, 설문지 설계 상의 문제 또는 무응답을 처리하기 위해 이용되는 대체방법 등에서 발생될 수 있으며, 여기에서는 LFS에서의 적용범위 오차(coverage error), 무응답 오차, 결측 오차(vacancy errror), 응답 오차(response error), 처리과정 오차(processing error)등에 관해서 설명하기로 한다.

적용범위 오차는 표본추출틀의 조사단위들이 목표모집단을 제대로 반영하지 못할 경우 발생할 수 있다. 조사단위들이 표본추출틀에서 누락되어 있는 경우, 목표모집단에 속하지 않은 단위들이 표본추출틀에 포함되어 있는 경우 또는 조사단위들이 표본추출틀에 중복되어 있는 경우 등이 적용범위 오차를 유발할 수 있는 일반적인 유형들이며, 이 중 조사단위들이 표본추출틀에서 누락되어 있는 경우가 LFS에서 가장 빈번히 발생하는 유형이라 할 수 있다. 나머지 유형의 문제는 LFS에서는 거의 무시된다. LFS에서는 적용범위 오차를 측정하기 위한 지표로써 손실률(slippage rate)을 이용한다. 손실률은 LFS 인구 추정치와 최근의 센서스 인구 추정치와의 차이에 대한 센서스 인구 추정치의 비율로써 정의된다. LFS에서는 CMA 지역, ER 지역, 주 및 전국 단위와 캐나다 지역의 성별(남, 녀)-연령대별(15-19, 20-24, 25-29, 30-39, 40-54, 55+) 범주에 대한 손실률을 매월 정기적으로 작성하고 있다. 손실률로부터 발생하는 비표본오차에 대한 보정은 추정과정에서 처리하고 있다. LFS 조사에서 평균적인 손실률의 양은 다음 <표8>과 같이 주어진다.

<표8> LFS 평균 손실률(Average Slippage Rate(%))

Provinces		Average	Provinces	Average
Canada	all	9.3	Nova Scotia	8.6
	15-19세	6.1	New Brunswick	10.4
	20-24세	15.6	Quebec	8.0
	25-29세	16.1	Ontario	9.7
	30-39세	9.8	Manitoba	6.1
	40-54세	8.0	Saskatchewan	10.7
	55이상	6.9	Alberta	7.4
Newfoundland		9.8	British Coumbia	12.4
Prince Edward Island		11.6		

조사가구에 대한 무응답 발생요인으로써 가구 구성원의 부재, 가구 구성원의 비 정상적인 거주 환경, 인터뷰 거절 등의 요인을 들 수 있다. 여기에서 인터뷰 거절의 비율은 매월 조사에서 1~2% 정도로 매우 낮게 나타나며, 주 지역에 대해서도 월 조사와 비슷한 비율을 보이나 높게는 약3% 정도까지 나타난다. 단위 무응답에 대해서는 바로 전 달의 정보를 이용할 수 있다면 이를 이용하여 대체되며, 항목 무응답에 대해서는 표본 대체법이 이용된다. 가구 구성원이 거주하고 있지 않는 결측 가구 및 건물 철거 등으로 인한 비 존재 가구들에 대해서 발생하는 무응답은 편향에 영향을 미치지는 않지만 표본 분산에는 영향을 미치게 되므로 LFS에서는 이러한 유형의 오차 정보를 파악하기 위한 VC(vacancy check) 프로

그램을 운영하고 있다. 다음 <표9>는 1997년 LFS 조사에서 발생한 평균 무응답률을 나타낸다.

<표9> LFS 평균 무응답률(1997)

Provinces	Average(%)	Maximum(%)	Minimum(%)
Newfoundland	4.2	5.4	3.0
Prince Edward Island	3.5	4.8	2.4
Nova Scotia	6.3	7.3	4.6
New Brunswick	4.6	5.4	3.1
Quebec	5.4	6.6	3.7
Ontario	4.8	5.7	3.7
Manitoba	3.6	5.4	2.1
Saskatchewan	3.6	4.6	2.4
Alberta	4.9	6.3	3.1
British Columbia	5.7	6.7	4.5
Canada	4.9	5.5	3.8

LFS에서 결측가구(dwelling vacant)는 사람이 거주하고 있지 않는 가구, 계절 가구 또는 공사 중인 가구로 정의되어 분류된다. 철거 또는 조사 가구가 상점 등으로 용도가 변경된 경우는 비 존재 가구(dwelling non-existence)로 분류된다. 결측가구로 확인된 가구들은 LFS 추정치의 편향에 영향을 미치지는 않지만 표본 조사단위가 줄어들므로 추정 분산은 커지게 된다. 결측가구들은 새로운 입주자들이 상주할 가능성이 항상 존재하므로 매월 조사 대상에 포함된다. 그러나 LFS 조사에서 비 존재 가

구로 확인된 조사가구들은 표본추출틀에서 일제히 삭제된다. 1997년 LFS 조사에서 발생한 평균적인 결측가구에 대한 비율이 다음 <표10>에 주어졌다.

<표10> LFS 평균 결측률(vacant rate)(1997)

Provinces	Average(%)	Maximum(%)	Minimum(%)
Newfoundland	15.4	14.9	16.4
Prince Edward Island	20.5	18.6	23.0
Nova Scotia	16.8	15.2	18.7
New Brunswick	14.1	13.5	15.2
Quebec	14.0	11.9	15.8
Ontario	10.8	10.0	11.3
Manitoba	17.1	16.4	17.7
Saskatchewan	14.7	12.5	15.5
Alberta	8.7	8.1	9.8
British Columbia	9.5	8.7	9.8
Canada	13.0	12.2	13.5

응답 오차(response error)는 설문지 설계, 문항 구성, 응답자의 인지력, 인터뷰 방식, 조사가 수행되는 상황 및 조사정보가 수집되고 집계되는 과정 등에 기인할 수 있다. 조사정보가 수집되고 집계되는 과정에서 발생하는 응답오차는 CAI 시스템에 의해 어느 정도 보완되었다고 볼 수 있다.

처리과정 오차(processing error)는 자료 획득, 편집, 코딩, 가중치를 산

출하는 과정 및 목록화 작업 등의 매 단계에서 발생할 수 있다. LFS 조사에서는 이러한 매 단계의 처리과정을 전산화 작업으로 통합하여 자료처리과정에서 발생하는 오차를 최소화하고 있다. 전산화 통합 모드는 1993년부터 채택되어 시행되고 있으며 앞서 언급되었던 CAI 모드는 조사행정에서 조사원들의 조사과정을 보조하는 일종의 컴퓨터 보조관리 시스템을 말한다.

7. 소지역 추정법

캐나다 노동력 조사에서는 표본설계 단계에서 EIER 지역과 CMA 지역과 같은 소지역 단위에 대한 추정을 고려하여 층화, 표본추출, 표본 배정 등이 이루어진다. 소지역 통계 추정을 위한 추정량은 크게 설계 기반 추정량(design-based estimator), 간접추정량(indirect estimator), 모형 기반 추정량(model-based estimator)이 이용된다. 소지역 통계 작성 시 설계 기반 추정량이 목표 요구정도를 만족한다면 우선적으로 설계 기반 추정량을 이용하며 그렇지 못할 경우에는 추정량의 신뢰도를 확보할 수 있는 다른 추정방법을 이용한다.

가. 설계 기반 추정량(Design-Based Estimator)

일반적으로 설계 기반 추정량은 직접추정량(direct estimator)과 수정된 직접추정량(modified direct estimator)으로 구분된다. 관심변수와 밀접한 관련이 있는 보조정보가 있는 경우에 이를 이용하는 사후층화추정량(post stratified estimator), 비추정량(ratio estimator), 회귀추정량(regression estimator) 등은 직접추정량의 일종이다. 직접추정량은 편향이 없는 추정량이지만 해당 소지역에 배정된 표본의 크기가 작은 경우에는 추정량의 분산이 커져서 신뢰성이 떨어지게 된다. 한편 수정된 직접추정량(modified direct estimator)은 해당 소지역 이외의 다른 지역의 조사결과를 추정과정에 추가적으로 이용하며 추정량의 불편성은 근사적으로 유지된다.

직접추정량(direct estimator)은 보통 해당 소지역에서 조사된 자료만을

이용하여 추정되며, 간혹 센서스나 행정자료로부터 획득된 보조정보를 조사자료에 추가하여 추정되기도 한다. 가장 간단한 총계추정에 대한 직접추정량으로써 다음과 같은 단순추정량(expansion estimator)을 들 수 있다.

$$\hat{Y}_{e,a} = \sum_{i \in s_a} \omega_i y_i \quad , \tag{7.1}$$

여기에서 s_a는 소지역 a의 표본들의 집합, ω_i는 조사단위 i에대한 가중치를 나타낸다. 위의 직접추정량은 불편추정량이나 소지역 a의 표본크기가 작을 경우에는 분산이 커지기 때문에 신뢰성에 문제가 있을 수 있다.

소지역 a의 모집단의 크기 N_a를 알고있을 경우에는 다음과 같은 사후층화추정량이 이용될 수 있다.

$$\begin{aligned} \hat{Y}_{pst,a} &= N_a \frac{\sum_{i \in s_a} \omega_i y_i}{\sum_{i \in s_a} \omega_i} \\ &= N_a \frac{\hat{Y}_{e,a}}{\hat{N}_{e,a}} \\ &= N_a \, \bar{y}_{e,a} \end{aligned} \tag{7.2}$$

위의 사후층화추정량은 먼저 언급된 단순추정량보다는 안정적이나 보다 복잡한 조사에서는 비추정편향(ratio estimation bias)이 발생할 가능성이 있다.

표본이 층화추출되고 층 h에서 소지역 a의 모집단의 크기 $N_{h,a}$가 알려져 있을 경우에는 다음과 같은 유형의 사후층화추정량이 소지역 추

정에 이용될 수 있다.

$$\hat{Y}_{st,pst,a} = \sum_h \left(N_{h,a} \frac{\sum_{i \in s_{h,a}} \omega_i y_i}{\sum_{i \in s_{h,a}} \omega_i} \right)$$

$$= \sum_h N_{h,a} \frac{\hat{Y}_{h,e,a}}{\hat{N}_{h,e,a}}$$

$$= \sum_h N_{h,a} \bar{y}_{h,a} \quad . \qquad (7.3)$$

여기에서 층 h는 표본설계 시 반영된 층이라기 보다는 사후층화에 의해 형성된 층을 말한다.

비추정법(ratio estimation)은 사후층화추정법과 유사하나 모집단 총계 N_a와 $N_{h,a}$대신에 보조정보에 의해 획득된 소지역 총계 X_a와 $X_{h,a}$를 이용하며, 이 값들을 알고 있을 경우 비추정량은 다음과 같이 정의된다.

$$\hat{Y}_{r,a} = X_a \hat{R}_a \ , \quad \hat{Y}_{st,r,a} = \sum_h X_{h,a} \hat{R}_{h,a} \ , \qquad (7.4)$$

여기에서 $\hat{R}_a = \hat{Y}_{e,a} / \hat{X}_{e,a}$는 Y_a / X_a의 추정값, $\hat{R}_{h,a} = \hat{Y}_{h,e,a} / \hat{X}_{h,e,a}$를 나타낸다.

회귀추정법(regression estimation)이 소지역 총계 추정에 이용되기도 한다. 이 방법은 관심변수 y와 공변량 x사이의 관계에서 회귀모수를 추정하여 소지역 총계 추정에 이용하는 방법으로써 추정량은 다음과 같은 형태로 주어진다.

$$\hat{Y}_{reg,a} = \hat{Y}_a + \hat{\beta}_a (X_a - \hat{X}_a) \ , \qquad (7.5)$$

여기에서 $\hat{Y}_a$는 직접추정 또는 사후층화추정법에 의해 추정된 소지역 a에 대한 총계 추정값이며 $\hat{X}_a$는 보조정보를 통해 $\hat{Y}_a$과 유사한 방법으로 추정된다. 추정모수 $\hat{\beta}_a$은 관심변수 y와 공변량 x의 관계로부터 추정되며 $\hat{\beta}_a = \sum_{i \in s_a} \nu_i^{-1} \omega_i y_i x_i^T (\sum_{i \in s_a} \nu_i^{-1} \omega_i x_i x_i^T)^{-1}$와 같이 주어진다. 여기에서 ν_i는 회귀가중치로써 주어지는 값이며, x가 상수이고 $\nu_i = x_i$일 경우에는 $\hat{\beta}_a = \hat{R}_a$인 관계가 성립한다. 회귀추정량의 불편성은 $\hat{Y}_a$와 $\hat{X}_a$의 불편성에 의존한다.

한편, 회귀추정량을 변형한 일종의 수정된 직접추정량(modified direct estimator)이 소지역 특성치 추정에 이용되기도 한다. 수정된 직접추정량은 해당 지역 외의 조사자료를 특성치 추정에 이용하며, 추정량의 불편성은 회귀추정량과 마찬가지로 근사적으로 만족된다. 예를 들면 식(7.5)에서 추정 회귀모수 $\hat{\beta}_a$ 대신에 회귀모수에 대한 합성추정량의 일종인 $\hat{\beta} = \sum_{i \in s} \nu_i^{-1} \omega_i y_i x_i^T (\sum_{i \in s} \nu_i^{-1} \omega_i x_i x_i^T)^{-1}$이 이용되었다면 이러한 추정량을 수정된 직접추정량(modified direct estimator)이라 부른다. 일반적으로 소지역 추정 시 $\hat{\beta}$이 $\hat{\beta}_a$보다 안정적인 것으로 알려져 있으며, $\hat{\beta}$과 $\hat{\beta}_a$의 가중평균 $\lambda_a \hat{\beta}_a + (1-\lambda_a)\hat{\beta}$이 추정 회귀모수로 이용되기도 한다. 여기에서 λ_a는 적절히 선택되는 값이다. x가 상수이고 $\nu_i = x_i$인 경우에는 $\hat{\beta}$대신 $\hat{R} = \hat{Y}_e / \hat{X}_e$이 이용될 수도 있다.

나. 간접추정량(Indirect Estimator)

간접추정량은 합성추정량(synthetic estimator), 복합추정량(composite estimator), 표본수 의존 복합추정량(sample size dependent estimator) 등의 유형으로 구분되며, 해당 지역의 조사자료뿐만 아니라 해당 지역을 포함하고 있는 더 큰 지역의 조사자료를 소지역 추정과정에 이용하여 소지역 추정의 신뢰성을 확보하는 방법이다.

합성추정법(synthetic estimation)은 소지역 추정 시 소지역을 포함하는 대영역의 정보를 함께 이용하는 방법으로써 소지역과 대영역의 특성 구조가 유사하다는 가정 하에서 이용된다. 합성추정량의 분산은 직접추정량의 분산에 비해 작으나 전제한 가정이 성립하지 않을 경우에는 심각한 편향이 발생할 수 있다.

소지역의 특성치 평균이 전체 지역의 특성치 평균과 같다는 가정 하에서 만들어진 가장 간단한 형태의 합성추정량은 다음과 같다.

$$\hat{Y}_{syn,m,a} = N_a \frac{\sum_{i \in s} \omega_i y_i}{\sum_{i \in s} \omega_i} = N_a \bar{y} \qquad (7.6)$$

층화 또는 사후층화에 근거한 합성추정량은 보통 다음과 같은 형태로 주어진다.

$$\hat{Y}_{syn,st,m,a} = \sum_h N_{h,a} \frac{\sum_{i \in s_h} \omega_i y_i}{\sum_{i \in s_h} \omega_i} = \sum_h N_{h,a} \bar{y}_h \qquad (7.7)$$

직접추정법에서와 마찬가지로 합성추정법에서도 비합성추정법(ration synthetic estimation)이 고려될 수 있다. 비합성추정법은 모집단의 크기

N_a 또는 $N_{h,a}$ 외에 소지역 추정을 위한 보조정보로써 공변량 x를 이용하며 추정량(ratio synthetic estimator)은 다음과 같은 형태로 정의된다.

$$\hat{Y}_{syn,r,a}=X_a\frac{\hat{Y}_e}{\hat{X}_e}\ ,\quad \hat{Y}_{syn,st,r,a}=\sum_h X_{h,a}\frac{\hat{Y}_{h,e}}{\hat{X}_{h,e}}\ , \qquad (7.8)$$

여기에서 $\hat{Y}_e=\sum_{i\in s}\omega_i y_i$는 y에 대한 모집단 총계 추정량, $\hat{Y}_{h,e}=\sum_{i\in s_h}\omega_i y_i$는 층의 총계 추정치를 나타낸다. 비합성추정량들은 Gonzalez(1973), Gonzalez and Waksberg(1973), Ghangurde and Singh(1977, 1978)에 자세히 소개되어있다. 한편, Singh and Tessier(1976)는 (7.8)식의 $\hat{Y}_{syn,r,a}$에서 $\hat{X}_e$ 대신에 X를 이용한 비합성추정량의 대체식 $\hat{Y}_{syn,r,a}=X_a\ \hat{Y}_e/X$을 제안하였다. 여기에서 $\hat{Y}_{syn,r,a}$와 $\hat{Y}_{syn,r,a}$는 모두 같은 양의 편향을 가지며, $\hat{Y}_{syn,r,a}$의 편향은 표본의 크기가 클 경우에는 무시될 수 있다. 두 추정량 중 하나의 추정량을 선택하는 문제에서는 $\hat{Y}_e$와 $\hat{X}_e$의 상관계수 ρ를 참조하도록 하였다. 일반적으로 표본의 크기가 클 경우, 두 추정량의 분산은 상관계수의 값이 $\rho\geq 0.5c_x/c_y$일때 $V(\hat{Y}_{syn,r,a})\leq V(\hat{Y}_{syn,r,a})$인 관계가 성립한다. 여기에서 c_x와 c_y는 각각 $\hat{X}_e$와 $\hat{Y}_e$의 변동계수(coefficient of variation)를 나타낸다. 상관계수 ρ의 값이 크거나 모집단의 분포가 한쪽으로 치우쳐져 있을 경우에는 $\hat{Y}_{syn,r,a}$가 선호되며, 변동계수 c_x의 값이 크거나 상관계수 ρ의 값이 적당할 경우에는 보통 $\hat{Y}_{syn,r,a}$를 선택한다.

소지역의 보조정보로써 이용된 공변량 x외에 추가적인 보조변수 z를 도입하여 소지역 특성치를 추정하는 다음과 같은 이변량 비합성추정량이 소지역 추정에 이용될 수 있다.

$$\hat{Y}_{syn,r,a}^{(2)} = \gamma_a X_a \frac{\hat{Y}_e}{\hat{X}_e} + (1-\gamma_a) Z_a \frac{\hat{Y}_e}{\hat{Z}_e} , \qquad (7.9)$$

여기에서 γ_a는 적절히 선택되는 값이다. 보다 일반적인 다변량 비합성추정량은 Olkin(1958)에서 참조할 수 있다.

회귀합성추정법은 비합성추정법과 유사하며 추정량은 다음과 같이 주어진다.

$$\hat{Y}_{syn,reg,a} = \hat{\beta} X_a , \quad \hat{\beta} = \sum_{i \in s} \nu_i^{-1} \omega_i y_i x_i^T \left(\sum_{i \in s} \nu_i^{-1} \omega_i x_i x_i^T \right)^{-1} \qquad (7.10)$$

회귀합성추정법은 표본설계의 층 내에서 또는 사후층화에 의해 형성된 층 내에서도 응용이 가능하며, Royall(1979)은 이러한 내용을 반영한 다음과 같은 회귀합성추정량을 제안하였다.

$$\hat{Y}_{syn,Roy,a} = \sum_{i \in s_a} y_i + \hat{\beta} (X_a - \sum_{i \in s_a} x_i), \qquad (7.11)$$

복합추정량(composite estimator)은 직접추정량의 불안정성과 합성추정량의 잠재적 편향 가능성을 보완하기 위해 두 추정량의 가중평균을 취하며 일반적인 형태는 다음과 같이 주어진다.

$$\hat{Y}_{com,a} = \lambda_a \hat{Y}_{dir,a} + (1-\lambda_a) \hat{Y}_{syn,a} , \qquad (7.12)$$

여기에서 가중치 λ_a는 적절히 선택되는 값이다.

가중치 λ_a는 결정하는 방법은 크게 세가지 정도로 구분될 수 있다. 첫 번째 방법은 가장 간단한 방법으로써 가중치 λ_a를 고정계수로 두는 방법인데 추정량의 신뢰성에 문제가 있어 많이 사용되지는 않는다. 두 번째 방법은 추정하고자하는 소지역의 표본크기를 반영하는 방법이다. 이 경우 가중치 λ_a는 $\hat{N}_{e,a}/N_a$의 함수로 표현된다. Drew et al.(1982)은 표본크기에 의존하는 복합추정량으로써 다음과 같은 추정량을 제안하였다.

$$\hat{Y}_{ssd,r,a} = \lambda_a \hat{Y}_{r,a} + (1-\lambda_a) \hat{Y}_{syn,r,a} , \qquad (7.13)$$

여기에서 $\lambda_a = \begin{cases} 1 & , \text{ if } \hat{N}_{e,a} \geq \delta N_a \\ \hat{N}_{e,a}/\delta N_a & , \textit{otherwise} \end{cases}$ 이며, δ는 합성추정량 부분의 편향을 보정하기 위해 주관적으로 결정되는 값이다. 캐나다 노동력조사에서는 $\delta = 2/3$를 이용한다. 식 (7.13)의 복합추정량은 직접추정량의 신뢰도가 완전히 확보된 지역에 대해서는 합성추정량의 가중치가 0이 되기 때문에 이러한 지역의 경우 직접추정값이 곧 복합추정값으로 선택된다고 볼 수 있다. 그렇지 않은 기타 지역에 대해서는 직접추정값과 합성추정값의 가중평균값으로 복합추정값이 계산된다. 캐나다 노동력조사에서 이러한 기타 지역들에 대한 합성추정량의 평균 가중치는 약 10% 정도이며 많아야 20%를 초과하지는 않는다. 이때 δ의 값은 [2/3 , 3/2]의 범위에 있는 것으로 알려져 있다. 이외의 표본크기 의존 복합추정량으로써 Sandal(1984)의 추정량 $\hat{Y}_{ssd,reg,a} = \lambda_a \hat{Y}_{sreg,a} + (1-\lambda_a) \hat{Y}_{syn,reg,a}$ 을 들 수 있

다. 사용된 가중치는 $\lambda_a = \hat{N}_{e,a}/N_a$ 이다. Rao(1986)는 위와 동일한 추정량에 대해 가중치를 약간 달리 적용할 것을 제안하였다. Rao의 가중치는 $\hat{N}_{e,a} \geq N_a$ 인 지역에 대해서는 $\lambda_a = 1$, 기타 지역에 대해서는 Sandal의 가중치와 동일하다. Sandal and Hidiroglou(1989)는 Rao의 가중치에서 $\hat{N}_{e,a} < N_a$ 일때 $\lambda_a = (\hat{N}_{e,a}/N_a)^{h-1}$ 를 사용할 것을 제안하였다. 여기에서 h는 합성추정량 편향을 감안하여 적절히 선택되는 값이다. 가중치를 결정하는 세 번째 방법은 직접추정량과 합성추정량의 평균제곱오차와 두 추정량의 공분산을 자료로부터 추정하여 최적가중치를 산정하는 방법이다. 복합추정량의 평균제곱오차는 다음 식과 같이 나타낼 수 있다.

$$MSE(\hat{Y}_{com,a}) = \lambda_a^2 MSE(\hat{Y}_{dir,a}) + (1-\lambda_a)^2 MSE(\hat{Y}_{syn,a}) + 2\lambda_a(1-\lambda_a)E(\hat{Y}_{dir,a} - Y_a)(\hat{Y}_{syn,a} - Y_a) \quad (7.14)$$

$\hat{Y}_{com,a}$ 의 MSE를 최소화하는 가중치 λ_a 다음 식과 같이 주어질 수 있다.

$$\hat{\lambda}_a = \frac{\widehat{MSE}(\hat{Y}_{syn,a}) - E(\hat{Y}_{syn,a} - Y_a)(\hat{Y}_{dir,a} - Y_a)}{\widehat{MSE}(\hat{Y}_{syn,a}) + \widehat{MSE}(\hat{Y}_{dir,a}) - 2E(\hat{Y}_{syn,a} - Y_a)(\hat{Y}_{dir,a} - Y_a)} \quad (7.15)$$

식 (7.15)에서 $\hat{Y}_{dir,a}$ 와 $\hat{Y}_{syn,a}$ 의 공분산의 항이 $\widehat{MSE}(\hat{Y}_{syn,a})$ 와 $\widehat{MSE}(\hat{Y}_{dir,a})$ 에 비해 매우 작다고 가정할 수 있다면 다음과 같은 근사적인 가중치를 이용할 수도 있다.

$$\lambda_a^* = \frac{\widehat{MSE}(\hat{Y}_{syn,a})}{\widehat{MSE}(\hat{Y}_{syn,a}) + \widehat{MSE}(\hat{Y}_{dir,a})} \quad (7.16)$$

다. 모형 기반 추정량(Model-Based Estimator)

소지역 추정에 자주 이용되는 모형 기반 추정법(model-based estimation)으로는 EBLUP (empirical best linear unbiased prediction), EB(empirical Bayes), HB(hierarchical Bayes) 접근법 등이 있다. 최근에는 소지역 통계 작성을 위해 횡단면 조사자료(cross-sectional data)와 시계열자료(time series data)를 함께 추정과정에 이용하는 방법에 관한 연구가 활발히 진행되고 있다. 이 절에서는 횡단면 조사자료를 이용한 모형기반 추정량과 횡단면 조사자료와 시계열자료를 함께 이용하는 모형기반 추정량에 대한 최근의 연구들을 소개하기로 한다.

y_i를 i번째 소지역의 관심모수 θ_i에 대한 직접추정량, x_i를 모수 θ_i의 추정에 필요한 설명변수이고, 모형 $y_i=\theta_i+e_i$, $E(e_i)=0$ 를 가정할 때, 소지역 i에 대한 다음과 같은 선형회귀모형(linear regression model)을 고려할 수 있다.

$$\theta_i=\beta_0+\beta_1 x_i, \qquad i=1,2,\cdots,I\ , \tag{7.17}$$

여기에서 β_0와 β_1은 회귀모수를 나타낸다. 이때 θ_i에 대한 회귀합성추정량(regression synthetic estimator)은 다음과 같이 주어질 수 있다.

$$\hat{\theta}_{i(reg)}=\hat{\beta}_0+\hat{\beta}_1 x_i\ ,\ i=1,2,\cdots,I\ , \tag{7.18}$$

여기에서 $\hat{\beta}_0$와 $\hat{\beta}_1$은 결합모형 $y_i=\beta_0+\beta_1 x_i+e_i$ ($i=1,2,\cdots,I$) 로부터 계산되는 최소제곱추정량을 나타낸다. 조사 추정량 y_i들의 공분산을

추정할 수 있을 경우에는 일반화 가중최소제곱추정량을 이용할 수도 있다. 위의 회귀합성추정량은 조사 추정량 y_i들에 대한 가중치가 반영되지 않기 때문에 큰 편향이 발생할 수 있다. 반면, EB(Empirical Bayes) 추정량이나 EBLUP(Empirical Bayes Linear Unbiased Predictor)는 적당한 가중치가 부여되어 편향 발생이 다소 억제되는 결과를 얻을 수 있다.

이러한 편향에 대한 문제점을 해결하기 위해 Fay and Herriot(1979)는 모형 (7.17)을 다음과 같이 해당 소지역에 대한 랜덤효과 v_i를 갖는 모형으로 보완하였다.

$$\theta_i = \beta_0 + \beta_1 x_i + v_i \quad , \quad i=1,2,\cdots,I \, , \tag{7.19}$$

여기에서 v_i는 평균이 0이고 분산이 σ_v^2을 갖는 서로 독립인 정규분포를 따르는 확률변수, e_i는 평균이 0이고 분산이 σ_i^2인 서로 독립인 정규확률변수를 나타내며, σ_i^2은 기지인 값으로 가정된다. 이때 결합모형은 다음과 주어진다.

$$y_i = \beta_0 + \beta_1 x_i + v_i + e_i \quad , \quad i=1,2,\cdots,I \, . \tag{7.20}$$

위의 모형으로부터 θ_i의 EB 추정량은 직접조사추정량 y_i와 회귀합성추정량 $\hat{\theta}_{i(reg)} = \hat{\beta}_0 + \hat{\beta}_1 x_i$의 가중합으로 표현되며 다음 식과 같이 주어진다.

$$t_i(\hat{\sigma}_v^2, \mathbf{y}) = \omega_i y_i + (1-\omega_i)\hat{\theta}_{i(reg)} \, , \tag{7.21}$$

여기에서 $\omega_i = \hat{\sigma}_v^2 / (\hat{\sigma}_v^2 + \sigma_i^2)$, $\hat{\beta}_0$와 $\hat{\beta}_1$은 결합모형 하에서 추정된 가

중 최소제곱추정량을 나타내며, $\widehat{\sigma}_v^2$ 은 σ_v^2의 적률추정량 또는 최대우도 추정량 등이 이용될 수 있다. Fay and Herriot(1979)는 1970년 미국의 인구 주택 총 조사 자료로부터 인구 1000미만의 소지역에 대한 소득관련 추정에 식 (7.21)의 EB 추정량을 이용하였고, EB 추정량이 직접 조사 추정량이나 합성추정량에 비해 표본오차가 작다는 사실을 수치적으로 제시하였다.

횡단면자료를 이용한 소지역 추정방법은 조사 시기가 상이한 조사자료들의 정보를 모형에 반영시키는 것은 사실상 어렵다. Scott et al.(1977), Jones(1980), Tiller(1989) 등은 이러한 단점을 보완하기 위해 반복적인 월별 조사자료들의 정보와 센서스 및 행정자료를 모형에 포함시킨 횡단면 시계열 모형들을 소지역 추정 문제에 도입하였다.

θ_{it}, y_{it}와 x_{it}를 각각 조사시기 t에서 소지역 i에 대한 모평균, 직접조사추정값, i번째 소지역과 관계가 있는 연관변수라 할때, 우선 다음과 같은 모형을 고려한다.

$$y_{it} = \theta_{it} + e_{it}, \quad i=1,2,\cdots,I, \quad t=1,2,\cdots,T, \tag{7.22}$$

여기에서 표본오차 e_{it}의 평균은 0, 분산공분산행렬은 기지인 블럭대각행렬 Σ_i ($T\times T$ 행렬)로 가정한다. 모평균 θ_{it}에 관한 모형은 다양한 유형으로 설정될 수 있으며 다음과 같은 모형들이 고려될 수 있다.

(Ⅰ) $\theta_{it}=\beta_0+\beta_1 x_{it}+v_i+\varepsilon_{it}$, $i=1,2,\cdots,I$, $t=1,2,\cdots,T$,

여기에서 v_i는 소지역 고정효과, $\varepsilon_{it} \overset{ind}{\sim} N(0,\sigma^2)$.

(Ⅱ) $\theta_{it}=\beta_0+\beta_1 x_{it}+v_i+\varepsilon_{it}$, $i=1,2,\cdots,I$, $t=1,2,\cdots,T$,

여기에서 $v_i \overset{ind}{\sim} N(0,\sigma_v^2)$, $\varepsilon_{it} \overset{ind}{\sim} N(0,\sigma^2)$, $\{v_i\}$와 $\{\varepsilon_{it}\}$는 서로 독립이며, 모형(Ⅰ)과는 달리 v_i들이 랜덤효과로 가정되었다.

(Ⅲ) $\theta_{it}=\beta_0+\beta_1 x_{it}+v_i+u_t+\varepsilon_{it}$, $i=1,2,\cdots,I$, $t=1,2,\cdots,T$,

여기에서 $v_i \overset{ind}{\sim} N(0,\sigma_v^2)$, $u_t \overset{ind}{\sim} N(0,\sigma_u^2)$, $\varepsilon_{it} \overset{ind}{\sim} N(0,\sigma^2)$, $\{v_i\}$, $\{u_t\}$와 $\{\varepsilon_{it}\}$는 서로 독립이다. v_i는 소지역에 대한 랜덤효과, u_t는 조사시기에 대한 랜덤효과이다.

(Ⅳ) $\theta_{it}=\beta_0+\beta_1 x_{it}+v_i+u_{it}$, $u_{it}=\rho u_{i,t-1}+\varepsilon_{it}$, $|\rho|<1$,

여기에서 $v_i \overset{ind}{\sim} N(0,\sigma_v^2)$, $\varepsilon_{it} \overset{ind}{\sim} N(0,\sigma^2)$, $\{v_i\}$와 $\{\varepsilon_{it}\}$는 서로 독립이며, $\{u_{it}\}$는 $AR(1)$ 과정을 따른다. 모형 (Ⅳ)는 다음과 같이 시차모형(lag model)으로 재표현 가능하다.

$$\theta_{it}=\rho\theta_{i,t-1}+(1-\rho)\beta_0+\beta_1 x_{it}-\beta_1\rho x_{i,t-1}+(1-\rho)v_i+\varepsilon_{it}. \qquad (7.23)$$

모형 (Ⅳ)는 모평균 θ_{it}와 보조변수 x_{it}가 이전 조사시기에서의 값들

을 모형에 반영시키기 때문에 위의 네가지 모형들 중에서는 가장 현실적인 추정 모형이라 할 수 있다. 따라서 모형 (Ⅳ)를 이용한 결합모형을 고려한다면 다음과 같이 주어진다.

$$y_{it} = \beta_0 + \beta_1 x_{it} + v_i + (e_{it} + u_{it}) \ , \qquad (7.24)$$
$$u_{it} = \rho u_{i,t-1} + \varepsilon_{it} \ , \quad |\rho| < 1 \ ,$$

여기에서 $v_i \overset{ind}{\sim} N(0, \sigma_v^2)$, $\varepsilon_{it} \overset{ind}{\sim} N(0, \sigma^2)$, e_{it}는 평균이 0이고, 기지인 블록 대각 공분산행렬 $\Sigma = diag(\Sigma_1, \Sigma_2, \cdots, \Sigma_I)$를 갖는다. 캐나다 노동력조사에서는 표본 공분산행렬 Σ를 이용할 수 없기 때문에 Tiller(1989)는 복합오차 $w_{it} = e_{it} + u_{it}$를 $AR(1)$ 과정으로 취급한 후 다음과 같은 결합모형을 이용하여 소지역 추정을 시도하였다.

$$y_{it} = \beta_0 + \beta_1 x_{it} + v_i + w_{it} \ , \qquad (7.25)$$
$$w_{it} = \rho w_{i,t-1} + \varepsilon_{it} \ , \quad |\rho| < 1 \ ,$$

여기에서 $\theta_{it} = \beta_0 + \beta_1 x_{it} + v_i$, $v_i \overset{ind}{\sim} N(0, \sigma_v^2)$, $\varepsilon_{it} \overset{ind}{\sim} N(0, \sigma^2)$.

$\{y_{it}\}$를 $\boldsymbol{y} = (y_{11}, \cdots, y_{1T}; y_{21}, \cdots, y_{IT})^T = (\boldsymbol{y}_1^T, \cdots, \boldsymbol{y}_I^T)^T$로 표현하면 위의 모형 (25)는 다음과 같은 일반적인 혼합모형(mixed model)의 일종으로 볼 수 있다.

$$\boldsymbol{y} = X\beta + Zv + w \ , \ v \sim (0, \sigma_v^2 I), \ w \sim (0, \sigma^2(I \otimes \Gamma)), \qquad (7.26)$$

여기에서 $X^T=(X_1^T,\cdots,X_I^T)$, $Z=I\otimes 1_T$, $\beta=(\beta_0,\beta_1)^T$, X_i는 t번째 행이 $(1,x_{it})$로 주어지는 $T\times 2$ 행렬, I는 크기 I인 항등행렬, 1_T는 1을 원소로 갖는 t-벡터, Γ는 (i,j)번째 원소가 $\gamma_{ij}=(1-\rho^2)^{-1}\rho^{|i-j|}$ 인 $T\times T$ 행렬을 나타낸다.

β와 v의 일차결합 $\tau=k^T\beta+m^Tv$에 대한 $\tilde{\tau}$ $(=\tilde{\theta}_{it})$의 BLUP(best linear unbiased predictor)와 BLUP의 평균제곱오차(MSE)는 다음과 같이 주어진다(Henderson, 1975).

$$\tilde{\tau}=k^T\hat{\beta}+m^TZ^T\Sigma^{-1}(y-X\hat{\beta})(\sigma_v^2/\sigma^2)\ , \tag{7.27}$$

$$MSE(\tilde{\theta}_{it})=\sigma^2\{k^T(X^T\Sigma^{-1}X)^{-1}k+(\sigma_v^2/\sigma^2)m^Tm-(\sigma_v^2/\sigma^2)^2m^TZ^T\Sigma^{-1}AZm$$
$$-2(\sigma_v^2/\sigma^2)k^T(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Zm\}\ , \tag{7.28}$$

여기에서 $\Sigma=I\otimes[(\sigma_v^2/\sigma^2)J+\Gamma]$, $\hat{\beta}=(X^T\Sigma^{-1}X)^{-1}(X^T\Sigma^{-1}y)$, J는 원소들이 1로 구성된 $T\times T$ 행렬, $A=I-X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}$을 나타낸다.

식 (7.27)을 이용하여 결합모형 (7.25)의 $\theta_{it}(=\tau)$에 대한 BLUP 추정량을 계산하면 주요 항들은 다음과 같이 주어질 수 있다.

$$k^T=(1,x_{it}),\quad m^T=(0,\cdots,0,1,0,\cdots,0)_i\ ,$$

$$m^TZ^T\Sigma^{-1}(y-X\hat{\beta})=1_T'[(\sigma_v^2/\sigma^2)J+\Gamma]^{-1}(y_i-X_i\hat{\beta})\ .$$

식 (7.27)의 BLUP는 미지인 분산비(variance ratio) σ_v^2/σ^2와 자기상관계수(autocorrelation) ρ에 의존하므로 이들 값들은 추정되어야 한다. 모수들에 대한 BLUP는 다음과 같이 주어진다(Pantula and Pollack, 1985).

$$\hat{\rho}=\left[\sum_{i=1}^{I}\sum_{t=1}^{T-2}\hat{e}_{it}(\hat{e}_{i,t+1}-\hat{e}_{i,t+2})\right]\left[\sum_{i=1}^{I}\sum_{t=1}^{T-2}\hat{e}_{it}(\hat{e}_{it}-\hat{e}_{i,t+1})\right]^{-1} \quad (7.29)$$

$\hat{\sigma}_v^2$ 과 $\hat{\sigma}^2$ 은 다음 (7.30), (7.31), (7.32), (7.33)식의 정의로부터 유도될 수 있다.

$$z_{it}^{(1)}=z_{it}-z_{it}^{(2)}, \quad (7.30)$$

여기서 $z_{it}=\begin{cases} y_{it}-\hat{\rho}y_{i,t-1}, & t\geq 2 \\ f_1 y_{it}, & t=1 \end{cases}$, $z_{it}^{(2)}=c^{-1}d_i f_t$, $c=(1-\hat{\rho})[T-(T-2)\hat{\rho}]$,

$f_t=\begin{cases} 1-\hat{\rho}^2, & t=1 \\ 1-\hat{\rho}, & t\geq 2 \end{cases}$, $d_i=\sum_{t=1}^{T}f_t z_{it}$.

$$h_{0it}^{(1)}=h_{0it}-h_{0it}^{(2)}, \quad (7.31)$$

여기에서 $h_{0it}=\begin{cases} 1-\hat{\rho}, & t\geq 2 \\ f_1, & t=1 \end{cases}$, $h_{0it}^{(2)}=c^{-1}d_i f_t$, $f_t=\begin{cases} 1-\hat{\rho}^2, & t=1 \\ 1-\hat{\rho}, & t\geq 2 \end{cases}$,

$d_i=\sum_{t=1}^{T}f_t h_{0it}$.

$$h_{1it}^{(1)}=h_{1it}-h_{1it}^{(2)}, \quad (7.32)$$

여기에서 $h_{1it}=\begin{cases} x_{it}-\hat{\rho}x_{i,t-1}, & t\geq 2 \\ f_1 x_{it}, & t=1 \end{cases}$, $h_{1it}^{(2)}=c^{-1}d_i f_t$, $f_t=\begin{cases} 1-\hat{\rho}^2, & t=1 \\ 1-\hat{\rho}, & t\geq 2 \end{cases}$,

$d_i=\sum_{t=1}^{T}f_t h_{1it}$.

$$g_i=\sum_{t=1}^{T}f_t z_{it}, \quad f_{0i}=\sum_{t=1}^{T}f_t h_{0it}, \quad f_{1i}=\sum_{t=1}^{T}f_t h_{1it}. \quad (7.33)$$

종속변수를 $z_{it}^{(1)}$, 설명변수를 $h_{0it}^{(1)}$와 $h_{1it}^{(1)}$를 갖는 절편항이 없는 회귀

식을 적합시켰을 때 얻어지는 잔차제곱합을 $\hat{e}^T\hat{e}$라 하고, 종속변수 g_i, 설명변수 f_{0i}, f_{1i}를 갖는 절편항이 없는 회귀식에서 얻어지는 잔차제곱합을 $\hat{u}^T\hat{u}$라 할때, σ_v^2과 σ^2의 BLUP는 최종적으로 다음과 같이 주어진다.

$$\hat{\sigma}_v^2 = c^{-1}(I-2)^{-1}[\,\hat{u}^T\hat{u} - \hat{\sigma}^2(I-2)\,],$$

$$\hat{\sigma}^2 = [I(T-1)-2]^{-1}\hat{e}^T\hat{e}. \qquad (7.34)$$

따라서 θ_{it}의 EBLUP $\hat{\theta}_{it}$는 식 (27)에서 $\hat{\rho}$, σ_v^2, σ^2을 대체하여 얻을 수 있다.

만약 $p-1(\geq 2)$개의 x변수들이 모형에 포함되어 있다면, θ_{it}의 EBLUP $\hat{\theta}_{it}$는 다음과 같이 추정하면 된다. 우선 y_{it}와 $x_{1it}, \cdots, x_{p-1,it}$의 회귀식으로부터 $\{\hat{e}_{it}\}$을 추정한다.

다음으로 $\{h_{jit}, h_{jit}^{(1)}, h_{jit}^{(2)};, j=0,1,\cdots,p-1\}$을 앞서 언급된 바와 같이 1, x_{1it}, $\cdots$, $x_{p-1,it}$의 원소들로 정의한 후, $z_{it}^{(1)}$과 $h_{0it}^{(1)}, h_{1it}^{(1)}, \cdots, h_{p-1,it}^{(1)}$의 절편항이 없는 회귀식으로부터 $\hat{e}^T\hat{e}$을 추정한다. 같은 방법으로 $f_{ji}=\sum_{t=1}^{T} f_t h_{jit}$ $(j=0,1,\cdots,p-1)$를 계산한 후, g_i와 $f_{0i}, f_{1i}, \cdots, f_{p-1,i}$의 절편항이 없는 회귀식을 적합시켜 $\hat{u}^T\hat{u}$를 구한다. 마지막으로 (34)식에서 $I(T-1)-2$를 $I(T-1)-p$로, $I-2$를 $I-p$로 대체하여 BLUP $\hat{\sigma}_v^2$, $\hat{\sigma}^2$과 $\hat{\rho}$를 계산하면 EBLUP $\hat{\theta}_{it}$을 얻을 수 있고, (7.28)식으로부터

EBLUP $\hat{\theta}_{it}$의 MSE 값을 계산할 수 있다. 한편, 모형 (7.25) 하에서 조사 추정량 y_{it}의 MSE는 다음과 같이 주어진다.

$$MSE(y_{it}) = E(y_{it} - \theta_{it})^2 = V(w_{it}) = \frac{\sigma^2}{(1-\rho^2)}. \qquad (7.35)$$

$MSE(y_{it})$의 추정량은 위의 (7.35)식에서 σ^2, ρ를 각각 $\hat{\sigma}^2$, $\hat{\rho}$로 대체하여 얻을 수 있다.

라. 소지역 추정량들의 효율

$\hat{Y}_{M,a}(r)$을 소지역 추정법 M을 이용하여 추정된 r번째 반복에서 특성치 Y_a의 몬테카를로 추정값이라 하자. 이 때 n개의 소지역에 대한 평균제곱오차 추정값의 평균은 다음 식을 이용하여 계산할 수 있다.

$$Avg\ \widehat{MSE}_M = \frac{1}{n}\sum_a \sum_{r=1}^{R} \frac{(\hat{Y}_{M,a}(r) - Y_a)^2}{R} \qquad (7.36)$$

소지역 추정법 M을 이용하여 추정된 추정량들의 효율을 직접추정법 M_0를 이용한 추정량과 비교하여 나타낸다면 상대 효율은 다음 식을 이용하여 구할 수 있다.

$$Eff(M\ vs\ M_0) = \frac{Avg\ \widehat{MSE}_M}{Avg\ \widehat{MSE}_{M_0}} \qquad (7.37)$$

여기에서 $Avg\ \widehat{MSE}_{M_0}$는 직접추정법 M_0에 의해 추정된 추정값들에 대한 평균제곱오차의 평균을 나타낸다.

마. LFS의 소지역 통계 작성방법

현재 캐나다 LFS에서 소지역 통계 작성을 위해 사용되고 있는 추정량은 Drew et al.(1982)에 기초한 표본수 의존 추정량(sample size dependent estimator)이다. 우선 소지역 a에 대해서 일반화 회귀추정량으로 관심변수의 총계를 추정한다.

$$\hat{Y}_{GREG,a} = \hat{Y}_{e,a} + \hat{\beta}_a (X_a - \hat{X}_{e,a}) , \qquad (7.38)$$

여기에서 $\hat{Y}_{e,a} = \sum_{i \in s_a} \omega_i y_i$, $\hat{X}_{e,a} = \sum_{i \in s_a} \omega_i x_i$, $\hat{\beta}_a = \sum_{i \in s_a} \omega_i y_i x_i^T (\sum_{i \in s_a} \omega_i x_i x_i^T)^{-1}$

이다.

관심변수의 총계를 추정할 수 있으면 LFS에서 관심의 대상인 경제활동인구수, 실업자 수, 취업자 수, 실업률, 취업률 등을 모두 추정할 수 있게 된다. 추정과정에 이용된 가중치 ω_i는 설계 가중치와 무응답 가중치 조정이 반영된 것이며, LFS의 최종 가중치를 의미하는 것은 아니다. 일반적으로 ER 지역의 소지역 통계 작성에 사용되는 보조정보는 주(Province) 단위의 노동력 통계 작성에 이용될 수 있는 보조정보에 비해 제한적이라 할 수 있다.

Drew et al.(1982)는 LFS에 알맞은 소지역 추정량을 찾기 위해 여러 종류의 추정량들을 비교하였다. 여기에서 사용된 소지역은 CD(Census Division) 지역이고 각 소지역에 대해 다음과 같은 세 개의 범주에 대한 보조정보를 추정과정에 이용하였다.

(i) 연령 15~16세, 65세 이상 전체인구

(ii) 연령 17 ~64세 여성인구

(iii) 나이 17~64세 남성인구 .

소지역 a를 포함하는 주(Province) 단위에서 $\hat{\beta}$을 구하여 합성추정량 $\hat{Y}_{SYN,GREG,a}=\hat{\beta}X_a$을 계산하였다. 여기에서 $\hat{\beta}=\sum_{i\in s}\omega_i y_i x_i^T(\sum_{i\in s}\omega_i x_i x_i^T)^{-1}$이며, 주 단위에서 사용된 보조정보는 30개의 연령 및 성별그룹의 총 수, 각 ER 지역과 CMA 지역에 대한 인구 총계 등이다.

LFS에서 사용되고 있는 표본수 의존 추정량은 앞서 언급한 두 추정량의 가중평균 $\hat{Y}_{SSD,a}=\lambda_a\hat{Y}_{GREG,a}+(1-\lambda_a)\hat{Y}_{SYN,GREG,a}$ 을 취한다. 여기에서 가중치 λ_a는 $\lambda_a=\begin{cases}1, & \hat{N}_{e,a}\geq\delta N_a\\ \hat{N}_{e,a}/\delta N_a, & otherwise\end{cases}$ 이고, δ값은 2/3를 사용하고 있다. 만약 소지역에 대한 직접추정량이 목표 요구정도를 만족하고 있다면 합성추정량 부분에 대한 가중치는 0이 된다.

현재 LFS에서 소지역의 취업률과 실업률에 대한 추정값은 표본수 의존 추정량으로 구해진 추정치의 3개월 간의 평균값이 이용되고 있다. 캐나다 LFS에서는 6개월의 표본순환 방법을 이용하고 있기 때문에 3개월의 결과를 평균하게되면 결과적으로 1/3표본 수를 늘리는 효과를 갖게 된다. 만약 조사시점 간에 표본들이 정확히 일치한다면 추정의 효율 측면에서 이러한 이득은 기대할 수 없다.

연속조사에서 추정량의 정확도를 향상시키기 위해 서로 다른 시점에서 조사된 몇 개의 조사 자료를 풀링(pooling)하는 경우를 흔히 볼 수 있

다. 특히 시점이 다른 몇 차례의 조사결과를 결합하거나 이들의 평균을 구하는 것이 보통의 방법인데, 이러한 방법은 소지역 통계 작성 시 해당 소지역에 배당된 표본 크기가 매우 적어서 추정의 정확도가 크게 떨어지는 경우에 유용하다. 그러나 다른 시점의 조사결과들을 결합하여 산출되는 추정값은 개념상의 문제점을 항상 내재하고 있다.

대영역 내에서 추정된 소지역 추정값들의 합계는 대영역의 추정값과 일치해야 하지만, 대개의 경우 소지역 추정값들의 합계는 대영역의 추정값과 일치하지는 않는다. 따라서 최종적으로 총계를 일치시키는 다음과 같은 보정이 이루어져야 한다. 소지역 i에 대한 총계 Y_i의 추정량을 $\hat{Y}_i$, 해당 소지역을 포함하는 대영역 a에 대한 $\hat{Y}_i$의 합계를 $\hat{Y}(a)$라 하자. 이때 소지역 추정값들은 다음 (7.39)식의 $\hat{Y}_i^{ADJ}$를 통해 보정된다.

$$\hat{Y}_i^{ADJ} = \frac{\hat{Y}_i}{\hat{Y}(a)} \hat{Y}(a) , \qquad (7.39)$$

여기에서 $\hat{Y}(a) = \sum_{i \in a} \hat{Y}_i$ 이다.

바. 노동력 추정값의 분산추정

(1) 배 경

캐나다 노동력 조사(LFS)는 캐나다 통계청에서 실시하는 가장 큰 규모의 월 단위 가구 조사로써 주로 전국 단위, 주 단위 및 주 내의 소지역 단위에 대한 다양한 노동력 특성에 대한 추정값들을 생산하고 있다. 캐나다의 LFS는 6개의 순환 패널을 갖는 층화 다단계 연동교체 표본설계를 따른다. 매번 인구 센서스 후 LFS는 표본설계에서 부분적인 보완이 이루어져 왔으며, 특히, 1981년에는 표본추출, 데이터 수집 및 추정 방법론 등에서 포괄적인 보완이 이루어졌다. 이 시기에 주 내의 소지역에 대한 추정치의 신뢰도를 높이기 위한 사후층화 비추정 절차가 새로이 마련되었다.

이 논문은 분산추정의 방법론에 대한 연구결과를 요약하였다. 과거 LFS의 분산추정은 Keyfitz 절차를 일반화한 Woodruff의 계산법이 이용되었다(Woodruff 1971). 이 방법은 Platek and Singh(1976)의 논문에서는 Keyfitz 방법으로 불리운다.

LFS에서는 다음 소개되는 세가지 유형의 지역들이 표본설계에 반영된다. 주요도시들로 구성되어 있는 SR지역(self-representing area), 소규모 도시들과 시골을 포함하는 NSR지역(non-self-representing area)과 군사지역 등과 같은 특수지역들이 이러한 유형의 지역들이다. NSR지역과 특수지역들

에 대한 분산추정은 비추정방법에 Keyfitz 방법을 혼합하여 적용하였다. 이 단계 랜덤그룹 표본설계가 반영된 SR지역들에 대해서는 Rao, Hartley and Cochran(1962)과 Rao(1975)의 분산 추정량을 이용하였고, 이 방법을 Keyfitz 방법을 이용한 분산추정량과 비교하였다. 한편, 추정값들에 대한 분산 추정량들을 비 보정(ratio adjustment) 분산추정량들과 편향 및 안정성의 측면에서 비교하였다. 또한 반복 수의 증가에 따른 Keyfitz 분산추정량의 영향도 살펴보았다. 결론적으로 SR 지역에 대해서는 Keyfitz 방법이 훌륭한 대안으로 채택되었다.

(2) SR 설계에 대한 분산추정

2.1 SR 설계(SR Design)

LFS 표본설계에서 SR 지역들은 이단계 랜덤그룹 설계방식을 취하며, 일단계 추출단위(PSU)들은 확률비례추출로 추출되며 이단계 추출단위들은 계통추출이 이루어진다. 하나의 층에 N개의 일차추출단위(PSU)가 있고, j번째 PSU에 대한 크기 측도를 x_j, $j=1,2,\cdots,N$, 거주단위의 수를 M_j, 층에서의 추출률을 $1/W$, 층으로부터 추출된 PSU의 수를 n이라 하자. N개의 PSU는 n개의 그룹으로 랜덤하게 분할되고 i번째 랜덤 그룹은 N_i개의 PSU들을 포함한다. 여기에서 $\sum_{i=1}^{n} N_i = N$ 이다. 우선 다음의 수식을 정의하자.

$$p_j = \frac{x_j}{\sum_{t=1}^{N} x_t} \quad , \quad j=1,2,\cdots,N_t \ ,$$

$$\delta_{ij} = \begin{cases} 1 \ , & \text{if the } j\text{th PSU is in the } i\text{th group} \\ 0 \ , & \text{otherwise} \end{cases}$$

이때 $\pi_j = \sum_{j=1}^{N} \delta_{ij} p_j$ 는 i 번째 랜덤 그룹의 상대적인 크기를 나타낸다.

$a_{ij} = \delta_{ij} W p_j / \pi_{ij}$, $r_{ij} = a_{ij} - [a_{ij}]$ 이라 하고, $\{r_{ij}, j=1,2,\cdots,N\}$ 가 내림차순으로 정렬되었다고 할 때, 계통추출의 추출간격 W_{ij} 는 다음과 같이 정의할 수 있다.

$$W_{ij} = [a_{ij}] + 1 \ , \quad j=1,2,\cdots,R$$
$$= [a_{ij}] \ , \quad j=R+1,\cdots,N \ ,$$

여기에서 $R = \sum_{j=1}^{N} r_{ij}$, $\sum_{j=1}^{N} W_{ij} = W$, $i=1,2,\cdots,n$.

하나의 PSU는 n 개의 랜덤 그룹 각각으로부터 추출률 W_{ij} 에 비례하는 확률로 추출된다. i 번째 랜덤 그룹으로부터 추출된 j 번째 PSU는 $1/W_{ij}$ 의 비율로 부차추출된다. 이때 전체 추출률은 $1/W$ 이 된다. 각각의 랜덤 그룹은 1 에서 6 까지의 패널로 할당되고, 랜덤 그룹의 수 n 은 보통 6 의 배수이고 각각의 패널은 같은 수의 랜덤 그룹을 갖는다. 각 랜덤 그룹으로부터 하나의 PSU가 추출되므로 i 번째 랜덤 그룹으로부터 추출된 PSU에서 부차추출률은 $1/W_i$ 이 된다. 랜덤 그룹 i 에서 선택된 거주단위의 수는 m_i 로 나타내기로 한다.

2.2 분산 추정량(Variance Estimator)

하나의 층에 대한 특성치 y의 총계에 관심이 있다고 하자. j번째 PSU에서 k번째 거주단위에 대한 y값을 y_{ik} ($k=1,2,\cdots,M_j$)라 할 때, 총계 $Y=\sum_{j=1}^{N}\sum_{k=1}^{M_j}y_{ik}$는 $\hat{Y}=W\sum_{i=1}^{n}y_i$ 로 추정될 수 있다. 여기에서 y_i는 i번째 그룹에서 선택된 PSU로부터 추출된 m_i개의 거주단위에 대한 y값들의 합을 나타낸다. $\hat{Y}$의 분산추정량은 다음과 같은 방법으로 추정될 수 있다.

① Keyfitz의 분산 추정량(1957)

과거 표본설계에서 이용하였던 총계 추정량에 대한 분산 추정공식은 다음 (2.1)식과 같다.

$$\hat{V}_1(\hat{Y})=W^2(\sum_o y_i-\sum_e y_i)^2\ , \qquad (2.1)$$

여기에서 $\sum_o$는 홀수 패널들에 대한 합, $\sum_e$는 짝수 패널들에 대한 합을 나타낸다. 위의 (2.1)식을 일반화시킨 일반화 Keyfitz 분산 추정공식은 다음 (2.2)식과 같이 주어진다.

$$\hat{V}_2(\hat{Y})=W^2\frac{n}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2\ , \qquad (2.2)$$

여기에서 $\bar{y}=(1/n)\sum_{i=1}^{n}y_i$ 이며, $\hat{V}_2(\hat{Y})$가 효율성이나 안정성 측면에서

$\hat{V}_1(\hat{Y})$보다 선호될 수 있다.

② Rao, Hartley and Cochran의 분산 추정량(1962)

Rao, Hartley and Cochran의 분산 추정공식은 i번째 그룹으로부터 추출된 m_i개의 거주단위의 수가 고정되어있고, 거주단위들은 단순임의추출이 가정된 상태에서 유도된다. 분산 추정공식은 다음과 같이 주어진다.

$$\hat{V}_3(\hat{Y}) = A\sum_1^n \pi_i\left(\frac{M_i}{m_i}\frac{y_i}{p_i} - \hat{Y}\right)^2 + \sum_i^n \frac{\pi_i}{p_i} M_i^2\left(\frac{1}{m_i} - \frac{1}{M_i}\right)s_i^2 , \qquad (2.3)$$

여기에서 $A = \dfrac{\sum_1^n N_i^2 - N}{N^2 - \sum_1^n N_i^2}$, $s_i^2 = \dfrac{1}{m_i - 1}\sum_{k=1}^{m_i}(y_{ik} - \bar{y}_i)^2$. M_i는 i번째 그룹에서 선택된 PSU에 속해있는 거주단위들의 수이고 M_i개의 거주단위들 중 m_i개의 거주단위들이 계통추출로 추출되나 분산추정값은 단순임의추출 하에서 계산된다. i번째 그룹에서 선택된 PSU로부터 추출된 k번째 거주단위에 대한 y값이 y_{ik}이며 $\bar{y}_i$는 $\bar{y}_i = y_i/m_i$로 주어진다.

$\pi_i/p_i = W/W_i$ 이고, $M_i/m_i = W_i$ 이므로 (2.2)식의 분산 추정공식은 다음 (2.4)식과 같이 주어질 수 있다.

$$\hat{V}_3(\hat{Y}) = A\sum_1^n \pi_i\left(W\frac{y_i}{\pi_i} - \hat{Y}\right)^2 + W\sum_1^n\left(1 - \frac{m_i}{M_i}\right)M_i s_i^2 \qquad (2.4)$$

③ Rao의 분산 추정량(1975)

Rao의 분산 추정공식에서는 m_i개의 거주단위들이 단순임의추출로 추출되나, 표본크기 m_i를 확률변수로 취급하여 분산 추정공식을 유도하였다. Rao의 분산 추정공식은 다음 (2.5)식과 같이 주어진다.

$$\hat{V}_4(\hat{Y}) = A\sum_1^n \pi_i\left(W\frac{y_i}{\pi_i} - \hat{Y}\right)^2 + \sum_1^n\left\{\frac{\pi_i^2}{p_i^2} - A\left(\frac{\pi_i}{p_i^2} - \frac{\pi_i^2}{p_i^2}\right)\right\}\frac{M_i^2 s_i^2}{m_i} - \sum_1^n \frac{\pi_i}{p_i} M_i s_i^2 \quad (2.5)$$

$$= \hat{V}_3(\hat{Y}) + W^2\sum_1^n m_i s_i^2\left\{\left(1 - \frac{W_i}{W}\right) - A\left(\frac{1}{\pi_i} - 1\right)\right\} \quad (2.6)$$

Rao의 분산 추정공식은 이 단계 표본추출에서 확률표본크기가 가정되기 때문에 음의 값이 나올 가능성이 있음에 주의해야한다.

2.3 Monte Carlo Study

네 가지의 분산 추정량들의 편향과 상대적인 안정성을 검토하기 위한 몬테카를로 연구가 수행되었다. 이용된 자료는 1981년 센서스 자료 중 조사지역 내의 약 20%의 계통추출 표본이며 Halifax의 CMA(Census Metropolitan Area) 지역으로부터 19개의 층에 대해 검토가 이루어졌다. 추출률 $1/W$ 은 0.04로 주어진다. 19개 층에 대한 PSU의 수, 추출된 PSU의 수, 거주단위들의 수 및 기대 표본크기가 <표1>에 주어졌다.

<표1> 몬테카를로 연구를 위해 이용된 층

층	거주단위 수	PSU 수	추출된 PSU 수	기대 표본크기
1	737	49	6	29.5
2	490	33	4	19.6
3	745	45	6	29.8
4	720	34	6	28.8
5	621	37	6	24.8
6	630	38	6	25.2
7	503	31	4	20.1
8	340	23	4	13.6
9	472	33	4	18.9
10	468	33	4	18.7
11	367	28	4	14.7
12	390	23	4	15.6
13	626	36	6	25.0
14	650	39	6	26.0
15	350	22	4	14.0
16	736	46	6	29.4
17	573	35	6	22.9
18	773	48	6	30.9
19	866	64	8	34.6
합계	11,056	697	100	442.3

몬테카를로 기법을 이용하여 각 층에서 1,000 개의 표본이 독립적으로 생성되었다. t번째 몬테카를로 표본생성으로부터 층 h에 대한 총계 Y_h의 추정값을 $\hat{Y}_{ht}$ ($h=1,2,\cdots,19$, $t=1,2,\cdots,1000$), $\hat{V}_{jht}$ ($j=1,2,3,4$)를 $\hat{Y}_{ht}$ 의 네 개의 분산추정량이라 하고 다음을 정의하였다.

$$Y=\sum_{h=1}^{19} Y_h \ ,$$

$$\hat{Y}_t = \sum_{h=1}^{19} \hat{Y}_{ht}, \quad t=1,2,\cdots,1000,$$

$$\hat{V}_{jt} = \sum_{h=1}^{19} \hat{V}_{jht}, \quad j=1,2,3,4,$$

여기에서 $\hat{Y}_t$는 t번째 몬테카를로 표본생성에서 얻어지는 총계 Y의 추정값, $\hat{V}_{jt}\,(j=1,2,3,4)$는 분산 추정값을 나타낸다.

몬테카를로 기대값과 분산을 각각 E^*와 V^*로 표기할 때 T개의 몬테카를로 표본생성에 대한 기대값과 분산은 각각 다음과 같이 주어진다.

$$E^*(\hat{\theta}) = \frac{1}{T}\sum_{t=1}^{T} \hat{\theta}_t,$$

$$V^*(\hat{\theta}) = \frac{1}{T}\sum_{t=1}^{T} [\,\hat{\theta}_t - E^*(\hat{\theta})]^2.$$

위의 정의를 이용하여 $\hat{Y}$의 몬테카를로 분산 $V^*(\hat{Y})$과 분산추정량 $\hat{V}_j$의 몬테카를로 기대값 $E^*(\hat{V}_j)$와 몬테카를로 분산 $V^*(\hat{V}_j)$를 얻을 수 있다.

분산추정량 $\hat{V}_j$의 편향과 백분위 편향은 각각 다음과 같이 정의될 수 있다.

$$B_j = E^*(\hat{V}_j) - V^*(\hat{Y}),$$

$$PB_j = 100\,\frac{B_j}{V^*(\hat{Y})}, \quad j=1,2,3,4.$$

이때 $\hat{V}_j$의 평균제곱오차 MSE는 다음과 같이 주어진다.

$$MSE_j = V^*(\hat{V}_j) + B_j^2, \quad j=1,2,3,4.$$

Keyfitz 분산추정량 $\hat{V}_1$에 대한 $\hat{V}_j$의 상대효율은 다음과 같이 정의할 수 있다.

$$\text{Rel. Eff}(\hat{V}_j \text{ vs. } \hat{V}_1) = (MSE_1 / MSE_j)^{1/2}, \quad j=2,3,4.$$

분산추정량들에 대한 상대편향과 효율에 대한 결과는 <표2.1>과 <표2.2>에 주어졌다.

<표2.1> 총계 추정값에 대한 분산추정량들의 백분위 편향

구 분	백분위 편향(PB_j)			
	$\hat{V}_1$	$\hat{V}_2$	$\hat{V}_3$	$\hat{V}_4$
취업인구	23.4	24.5	-4.7	-6.3
실업인구	6.3	6.6	3.7	1.2
노동력인구	24.2	25.2	-5.1	-6.7

<표2.2> $\hat{V}_1$에 대한 $\hat{V}_2$, $\hat{V}_3$, $\hat{V}_4$의 상대효율

구 분	백분위 편향(PB_j)		
	$\hat{V}_2$	$\hat{V}_3$	$\hat{V}_4$
취업인구	1.51	3.22	3.11
실업인구	1.52	1.71	1.76
노동력인구	1.49	3.24	3.12

편향에 대해서는 분산추정량 $\hat{V}_1$과 $\hat{V}_2$가 유사하며, $\hat{V}_3$와 $\hat{V}_4$도 비슷한 결과를 보인다. 또한 $\hat{V}_1$과 $\hat{V}_2$의 편향은 취업인구와 노동력인구에서 비교적 큰 양의 편향값을 보이는 반면 $\hat{V}_3$와 $\hat{V}_4$의 편향은 상대적으

로 작은 값을 나타낸다. 효율성 측면에서 $\hat{V}_3$와 $\hat{V}_4$가 서로 유사하며 $\hat{V}_1$과 $\hat{V}_2$에 비해서는 월등한 효율을 보인다. $\hat{V}_1$과 $\hat{V}_2$중에서는 $\hat{V}_2$의 효율이 더 좋게 나타난다.

전체인구에 대한 비 추정값들의 분산추정량에 대한 결과가 다음 <표 3.1>과 <표3.2>에 주어졌다. 비 추정값들에 대한 분산추정량은 $\hat{V}_j^{(R)}$ ($j=1,2,3,4$)로 표기하였다.

<표3.1> 총계 추정값에 대한 분산추정량들의 백분위 편향

구 분	백분위 편향(PB_j)			
	$\hat{V}_1^{(R)}$	$\hat{V}_2^{(R)}$	$\hat{V}_3^{(R)}$	$\hat{V}_4^{(R)}$
취업인구	3.7	4.3	-1.1	-3.1
실업인구	5.3	5.5	4.0	1.4
노동력인구	4.5	5.0	-0.5	-2.5

<표3.2> $\hat{V}_1^{(R)}$에 대한 $\hat{V}_2^{(R)}$, $\hat{V}_3^{(R)}$, $\hat{V}_4^{(R)}$의 상대효율

구 분	백분위 편향(PB_j)		
	$\hat{V}_2^{(R)}$	$\hat{V}_3^{(R)}$	$\hat{V}_4^{(R)}$
취업인구	2.13	2.59	2.52
실업인구	1.57	1.71	1.76
노동력인구	2.08	2.56	2.51

$\hat{V}_1^{(R)}$과 $\hat{V}_2^{(R)}$의 편향이 취업인구와 노동력인구에서 $\hat{V}_1$과 $\hat{V}_2$에 비해 훨씬 작아진 사실을 확인할 수 있다. $\hat{V}_3^{(R)}$과 $\hat{V}_4^{(R)}$의 편향도 취

업인구와 노동력인구에서 $\hat{V}_3$ 와 $\hat{V}_4$ 보다 작은 값을 가지며 실업인구에서는 거의 변화가 발생하지 않았다.

몬테카를로 표본을 이용하여 네 가지 분산추정량들의 비 보정(ratio-adjustment) 추정값들에 대한 95% 신뢰구간을 살펴보았다. 추정값들의 95% 신뢰구간에 대한 포함비율이 다음 <표4>에 주어졌다.

<표4> 비보정을 갖는 총계 추정값들의 95% 신뢰구간 포함비율

구 분	포함 비율			
	$\hat{V}_1^{(R)}$	$\hat{V}_2^{(R)}$	$\hat{V}_3^{(R)}$	$\hat{V}_4^{(R)}$
취업인구	93.6	95.4	94.6	94.2
실업인구	94.3	95.1	95.3	95.0
노동력인구	93.2	95.3	94.6	94.2

취업인구, 실업인구 및 노동력인구의 모든 부분에서 네 가지 분산추정량들의 수행결과가 모두 적합한 것으로 나타났다.

편향의 관점에서는 비 보정 추정값들의 분산추정량들이 서로 상이한 결과를 나타내지는 않는다. 상대효율에서는 $\hat{V}_3^{(R)}$ 와 $\hat{V}_4^{(R)}$ 가 $\hat{V}_2^{(R)}$ 에 비해 효율이 높은 것으로 나타났다. 한편, $\hat{V}_1^{(R)}$ 의 자유도는 19 이고(각 층별로 1개의 자유도를 가짐) $\hat{V}_3^{(R)}$ 는 각 PSU가 하나의 반복으로 처리되어 81 개의 자유도를 갖게 된다. 따라서 비 보정 추정값들에 대한 Keyfitz 분산 추정량은 반복 수를 증가시키면 추정량의 안정성을 확보할 수 있다.

2.4 반복 수를 갖는 Keyfitz 분산 추정량

Keyfitz 방법의 효율성을 높이기 위해 6개의 순환 패널들이 반복표본들로 채택되었다. 6개의 순환 패널을 반복표본으로 이용한 분산 추정값들과 과거 표본설계를 이용한 2개의 반복표본으로부터 계산된 분산 추정값들이 비교되었다. 순환 패널이 반복표본으로 처리됨에 따라 기인된 중요한 관심 사항은 패널 편향으로부터 발생할 수 있는 분산 추정값들의 증가부분이다. 이러한 부분을 살펴보기 위해 '85년 3월부터 '87년 2월까지의 24개월의 LFS 자료가 이용되었다. 취업인구, 실업인구, 노동력인구의 24개월에 대한 분산 추정값들에 대한 평균과 표준편차를 계산하였다. 2개의 반복표본과 6개의 반복표본 하에서 얻어진 분산들에 대한 평균과 표준편차의 비는 24개의 CMA(Census Metropolitan Area) 지역들의 평균으로 산출하였고 다음 <표5>에 주어졌다.

<표5> 단순임의 분산추정값의 비교(LFS의 CMA지역 자료)

구 분	분산의평균에 대한 비 평균(average ratio) : 2 vs. 6 반복	분산의 표준편차에 대한 비 평균(average ratio) : 2 vs. 6 반복
취업인구	0.997	1.813
실업인구	0.995	1.515
노동력인구	1.003	1.833

6개의 반복표본을 이용한 분산이 2개의 반복표본의 분산보다 작은 값

을 나타낸다. 순환 패널을 반복으로 채택할 경우 분산 추정값들의 편향에 거의 영향을 미치지 않으며, 6개의 반복표본을 이용한 분산이 2개의 반복표본보다는 훨씬 안정적임을 확인할 수 있다. 즉 Keyfitz 방법에서 6개의 순환 패널을 반복으로 사용할 경우 심각한 편향은 발생하지 않으며 2개의 반복표본을 이용했을 경우보다는 효율이 증가됨을 확인할 수 있다.

(3) 비 추정값 탐색을 위한 분산추정방법

3.1 LFS에서 비 추정방법

과거 LFS에서는 사후층화 비 추정방법이 이용되었다. 무응답을 보정하기 위한 일종의 설계 가중치인 부차 가중치가 LFS 목표 모집단의 추정치들에 대해 비 보정되었다. 이러한 비 추정방법은 주 지역의 특성치에 대한 추정의 신뢰도를 높이는 결과를 보였으나 주 내의 소지역들에 대해서는 문제점을 안고 있었다. 주 내의 ER(Economical Region) 지역과 CMA(Census Metropolitan Area) 지역들에 대한 추정의 정확도를 높이기 위해 탐색적인 비 추정 절차가 채택되었다.

탐색적 추정절차는 보정값들의 수열을 통해 수행된다. 먼저 부차가중치가 주 내의 소지역의 인구를 참조하여 보정되며, 이 후 성별-연령대별 범주를 반영한 주 수준의 보정값이 최종 가중치에 적용된다. 이러한 절차는 한번 더 수행되어 한쌍의 가중치가 추가적으로 생성된다. W_0를 부차가중치라 하고 (W_1, W_2)와 (W_3, W_4)를 2회 반복으로부터 생성된 가중치

들의 쌍이라 하자. 노동력 특성값들은 W_4를 이용하여 추정된다. 주 지역의 성별-연령대별 그룹들에서 주변 총계 W_4는 상응하는 그룹들의 외부 인구 추정치와 정확히 일치하나 주 내의 소지역에 대해서는 반드시 그렇지는 않다. 그러나 그 차이는 매우 작게 나타난다.

3.2 1회 반복 비 추정값에 대한 분산공식

1회 반복 비 추정값들에 대한 분산공식을 유도하면 다음과 같다. 여기에서 적용되는 기본적인 방법론은 선형적인 형태의 부차 가중치를 얻을 때까지 테일러 전개 근사식을 연속적으로 적용하는 방법이다. 세부적인 유도과정을 소개하면 다음과 같다.

$Y^{(0)}, Y^{(1)}, Y^{(2)}$를 주 지역에서 W_0, W_1, W_2에 근거하여 추정된 노동력 특성값 y의 추정값들이라 하자. 이때 $Y^{(2)}$는 다음 (3.1)식과 같이 주어질 수 있다.

$$Y^{(2)} = \sum_a \frac{Y_a^{(1)}}{P_a^{(1)}} P_a \ , \qquad (3.1)$$

여기에서 $Y_a^{(1)}$은 주 지역에서 성별-연령대별 그룹 a에 대한 특성치 y의 W_1 가중추정값, $P_a^{(1)}$은 주 지역에서 성별-연령대별 그룹 a에 대한 인구의 W_1 가중 추정값, P_a는 주 지역에서 성별-연령대별 그룹 a에 대한 인구의 외부 추정치를 나타낸다.

F_a를 $F_a = Y_a^{(1)} / P_a^{(1)}$이라 할 때, $(E(Y_a^{(1)}), E(P_a^{(1)}))$에서 F_a에 대한 일계 테일러 근사식을 구하면 다음과 같이 주어진다.

$$F_a \doteq \frac{E(Y_a^{(1)})}{E(P_a^{(1)})} + \frac{1}{E(P_a^{(1)})}\{Y_a^{(1)} - E(Y_a^{(1)})\} - \frac{E(Y_a^{(1)})}{\{E(P_a^{(1)})\}^2}\{P_a^{(1)} - E(P_a^{(1)})\}.$$

이때 $Y^{(2)}$의 분산에 대한 테일러 근사식은 다음 (3.2)식과 같이 주어질 수 있다.

$$V(Y^{(2)}) = V(\sum_a F_a P_a)$$

$$\doteq V\left\{ \sum_a \frac{P_a}{E(P_a^{(1)})} (Y_a^{(1)} - R_{Y_a}^{(1)} P_a^{(1)}) \right\}, \qquad (3.2)$$

여기에서 $R_{Y_a}^{(1)} = E(Y_a^{(1)}) / E(P_a^{(1)})$을 나타낸다.

다음으로 W_1 가중 추정값 $Y_a^{(1)}$과 $P_a^{(1)}$은 W_0 가중 추정값의 항으로 다음 (3.3)식과 같이 나타낼 수 있다.

$$Y_a^{(1)} = \sum_s \frac{Y_{sa}^{(0)}}{P_s^{(0)}} P_s, \qquad (3.3)$$

$$P_a^{(1)} = \sum_s \frac{P_{sa}^{(0)}}{P_s^{(0)}} P_s,$$

여기에서 s는 CMA 또는 ER 지역을 나타내며, P_s는 지역 s의 인구를 나타낸다. (3.3)식의 $Y_a^{(1)}$과 $P_a^{(1)}$을 (3.2)식에 대입하여 W_0 가중 추정값의 비에 대한 일계 테일러 근사식을 구하면 다음 (3.4)식과 같이 주어질 수 있다.

$$V(Y^{(2)}) \doteq V\left[\sum_a \frac{P_a}{E(P_a^{(1)})} \sum_s \frac{P_s}{E(P_s^{(0)})} \{(Y_{sa}^{(0)} - R_{Y_{sa}}^{(0)} P_s^{(0)}) - R_{Y_a}^{(1)}(P_{sa}^{(0)} - R_{P_{sa}}^{(0)} P_s^{(0)})\} \right] \qquad (3.4)$$

여기에서 $R^{(0)}_{Y_{sa}} = E(Y^{(0)}_{sa}) / E(P^{(0)}_{s})$, $R^{(0)}_{P_{sa}} = E(P^{(0)}_{sa}) / E(P^{(0)}_{s})$이다.

위의 (3.4)식은 다음 (3.5)식과 같이 축약된 형태로 다시 표현할 수 있다.

$$V(Y^{(2)}) \doteq V\left\{ \sum_{s} \sum_{h \in s} \sum_{i=1}^{n_h} \sum_{a} (Z^{(0)}_{Y_{shia}} - R^{(1)}_{Y_a} Z^{(0)}_{P_{shia}}) \right\}$$

$$= V\left(\sum_{s} \sum_{h \in s} \sum_{i=1}^{n_h} D^{(0)}_{shi} \right), \qquad (3.5)$$

여기에서 $D^{(0)}_{shi} = \sum_{a} \left(Z^{(0)}_{Y_{shia}} - R^{(1)}_{Y_a} Z^{(0)}_{P_{shia}} \right)$,

$$Z^{(0)}_{Y_{shia}} = \frac{P_a}{E(P^{(1)}_a)} \frac{P_s}{E(P^{(0)}_s)} (Y^{(0)}_{shia} - R^{(0)}_{Y_{sa}} P^{(0)}_{shi}),$$

$$Z^{(0)}_{P_{shia}} = \frac{P_a}{E(P^{(1)}_a)} \frac{P_s}{E(P^{(0)}_s)} (P^{(0)}_{shia} - R^{(0)}_{P_{sa}} P^{(0)}_{shi})$$

이고, h는 s에 속하는 층을 나타내며, i는 층 h에서 반복을 나타낸다.

식 (3.5)에서 $\left\{ \sum_{i=1}^{n_h} D^{(0)}_{shi} \right\}$는 부차가중치들에 의해 결정되므로 독립성을 가정할 수 있다. 따라서 식 (3.5)는 다음 (3.6)식과 같이 표현될 수 있다.

$$V(Y^{(2)}) \doteq V\left(\sum_{h \in s} \sum_{h} \sum_{i=1}^{n_h} D^{(0)}_{shi} \right)$$

$$= V\left(\sum_{h} \sum_{i=1}^{n_h} \sum_{s \ni h} D^{(0)}_{shi} \right) \qquad (3.6)$$

여기에서 $\sum_{s \ni h}$는 층 h를 포함하고 있는 주 내의 모든 소지역들에 대한 합을 나타낸다. $D^{(0)}_{hi}$를 $D^{(0)}_{hi} = \sum_{s \ni h} D^{(0)}_{shi}$와 같이 정의하면, 위의 (3.6)식은 다음과 같이 주어진다.

$$V(Y^{(2)}) \doteq V\left(\sum_h \sum_{i=1}^{n_h} D_{hi}^{(0)}\right) \tag{3.7}$$

$\left\{\sum_i D_{hi}^{(0)}\right\}$는 부차가중치에 의해 결정되므로 이 변수들은 독립성을 가정할 수 있으며, 분산은 다음 식으로부터 추정될 수 있다.

$$\hat{V}(Y^{(2)}) \doteq \sum_h \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} (D_{hi}^{(0)} - \overline{D}_h{}^{(0)})^2 \tag{3.8}$$

여기에서 $\overline{D}_h{}^{(0)} = \frac{1}{n_h} \sum_{i=1}^{n_h} D_{hi}^{(0)}$이다. 그러나 이 표현식에서는 기대값이 포함되어있고 이러한 값들은 미지의 값이므로 추정값으로 대체하여 분산의 추정값을 근사적으로 계산할 수 있으며 이를 이용한 분산 추정값은 최종적으로 다음 (3.9)식과 같이 주어진다.

$$\hat{V} \doteq \sum_h \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} (D_{hi}^{(2)} - \overline{D}_h{}^{(2)})^2 \tag{3.9}$$

여기에서 $D_{hi}^{(2)} = \sum_{s \ni h} D_{shi}^{(2)}$,

$$\overline{D}_h{}^{(2)} = \frac{1}{n_h} \sum_{i=1}^{n_h} D_{hi}^{(2)},$$

$$D_{shi}^{(2)} = \sum_a (Z_{Y_{shia}}^{(2)} - R_{Y_a}^{(2)} Z_{P_{shia}}^{(2)}),$$

$$Z_{Y_{shia}}^{(2)} = \frac{P_a}{P_a^{(1)}} \frac{P_s}{P_s^{(0)}} \left(Y_{shia}^{(0)} - \frac{Y_{sa}^{(0)}}{P_s^{(0)}} P_{shi}^{(0)}\right) = Y_{shia}^{(2)} - \frac{P_{shi}^{(0)}}{P_s^{(0)}} Y_{sa}^{(2)},$$

$$Z_{P_{shia}}^{(2)} = \frac{P_a}{P_a^{(1)}} \frac{P_s}{P_s^{(0)}} \left(P_{shia}^{(0)} - \frac{P_{sa}^{(0)}}{P_s^{(0)}} P_{shi}^{(0)}\right) = P_{shia}^{(2)} - \frac{P_{shi}^{(0)}}{P_s^{(0)}} P_{sa}^{(2)},$$

$$R_{Y_a}^{(2)} = \frac{Y_a^{(1)}}{P_a^{(1)}} = \frac{P_a}{P_a^{(1)}} \frac{Y_a^{(1)}}{P_a} = \frac{Y_a^{(2)}}{P_a}.$$

위의 분산 추정공식 (3.9)식은 노동력 특성값들의 W_2 가중 추정값들에

대한 추정공식이며 W_0와 W_2의 두 가중치의 값을 요구한다.

3.3 2회 반복 비 추정값의 분산 추정

2회 반복 비 추정값에 대한 분산공식은 3.2절을 응용하여 테일러 급수전개의 연속적인 적용으로 얻을 수 있다. 그러나 2회 반복에 기인한 분산 추정공식은 매우 복잡한 형태를 취하기 때문에 1회 반복에 기인한 분산 추정공식이 오히려 합리적일 수 있다. 1회 반복 분산 공식은 한 쌍의 가중치 (W_0, W_2)를 이용한다. 여기에서는 (W_0, W_2) 대신에 (W_0, W_4)를 이용하였다. W_2보다는 W_4가 노동력 추정값들의 CV값들에 강한 영향력을 주지 않기 때문이다. 주 지역인 Nova Scotia 지역의 1981년 센서스 자료를 이용하여 몬테카를로 시뮬레이션이 수행되었다. 각각의 몬테카를로 표본에서 LFS 표본설계가 매 단계의 표본 추출을 통해 검증되었다. 1,000개의 몬테카를로 표본들이 독립적으로 추출되었다. 각각의 몬테카를로 표본들에 대해 2회 반복표본 비 추정값($Y^{(4)}$), 1회 반복 분산 추정량과 이의 CV 추정값을 이용한 분산 추정값($\hat{V}(Y^{(4)})$)과 95% 신뢰구간 ($Y^{(4)} \pm 1.96\sqrt{\hat{V}(Y^{(4)})}$)을 주 지역과 주 지역내의 소지역들에 대해 계산하였다. 또한 1,000개의 CV 값들의 평균을 계산하여 실제값과 매우 유사한 몬테카를로 CV 값들과 비교하였다. 결과는 <표6.1>에 주어졌다. <표6.1>의 모든 셀에 대해서 CV 값의 차이는 8% 미만으로 나타났고, 21개의 셀 중 13개의 셀에서는 4%미만의 값을 갖는다. 한편 신뢰구간의 포함범위는 <표6.2>에 주어졌다. 취업인구와 노동력인구에 대한 95% 신뢰

구간의 포함범위는 만족한 값을 보이나 실업인구에 대해서는 다소 낮은 포함범위를 나타내나 여전히 받아들일만한 결과를 보인다.

<표6.1> 1회 반복 분산추정량의 평균 *CV*값과 몬테카를로 *CV*값

구 분	ER 210	ER 220	ER 230	ER 240	ER 250	CMA Halifax	Province (Nova Scotia)
				Average CV's			
취업인구	3.52	3.46	3.14	3.05	1.96	2.01	1.08
실업인구	10.36	12.28	13.13	13.43	10.35	10.55	5.27
노동력인구	2.98	3.17	2.85	2.73	1.77	1.83	0.91
				Monte Carlo CV's			
취업인구	3.48	3.35	2.95	2.86	1.97	1.99	1.11
실업인구	10.90	12.71	13.28	13.37	11.12	11.31	5.59
노동력인구	2.76	3.08	2.76	2.53	1.72	1.74	0.92

<표6.2> 1회 반복 분산추정량에 의한 95% 신뢰구간의 포함범위

구 분	ER 210	ER 220	ER 230	ER 240	ER 250	CMA Halifax	Province (Nova Scotia)
취업인구	94.5	92.8	94.0	94.7	94.7	94.9	92.5
실업인구	92.1	90.7	91.4	91.8	92.7	92.7	93.1
노동력인구	96.2	93.0	93.6	95.2	95.2	96.0	94.0

(4) 결 론

비 보정을 하지 않은 추정값들의 Keyfitz 분산 추정법은 매우 큰 양의 편향을 가지며 효율성도 크게 떨어진다. 반면에 탐색적 비 보정 추정방법

은 상대적으로 작은 편향을 가지며 효율성도 크게 향상되는 것으로 확인되었다. 이 논문에서 소개된 비 보정 추정값들에 대한 분산 추정방법들은 무시할 수 있을 정도의 작은 편향을 갖는다. 한편 Keyfitz 방법은 반복 수를 증가시킬 경우 다른 분산 추정방법에 비해 효율을 크게 향상시킬 수 있었다. LFS 자료에서 6개의 순환 패널을 반복으로 취급하여 Keyfitz 방법을 적용시켜본 결과 순환 패널 편향에 기인한 Keyfitz 추정분산의 편향은 발생하지 않았다. Keyfitz 방법에 의해 유도된 1회 반복 분산공식은 2회 반복의 탐색적 비 추정값들에 대해서 매우 합리적인 분산 추정값들을 제공하며 신뢰구간에 대한 포함범위도 좋은 특성을 나타냈다.

8. 소지역 추정 적용 사례

8.1 스위스 노동력조사의 소지역 실업률 추정에서 모형 기반 추정량(Using Model-Based Estimation to improve the Estimate of Unemployment on a Regional Level in the Swedish Labor Force Survey)

(1) 배 경

지방 및 지역사회계획을 위한 안정된 정보를 얻기 위해서는 매우 큰 규모의 조사나 인구총조사를 시행하는 것이 바람직할 수 있다. 그러나 이 일은 대개 매우 막대한 비용이 들기 마련이므로 다른 대안을 찾아야만 한다. 국제적으로 특별히 미국과 캐나다에서는 작은 지역의 추정에 관한 새로운 방법을 개발하는데 매우 큰 관심을 가지고 있다.

1.1 합성 추정량

오랫동안 존재해왔고 최근에 더 개발되고 향상되어온 하나의 방법은 이른바 합성 추정이다. "합성"이라는 단어는 몇 개의 다른 의미로 사용되지만 Gonzalez (1973)에 의해 제시된 설명이 가장 일반적으로 받아들여져 왔다. : "하나의 비편향 추정값이 넓은 지역에 대한 표본조사로부터 얻어졌을 경우, 소지역이 더 넓은 지역과 같은 특성을 가지고 있다는 가정하에서, 이 추정값이 부분지역에 대한 추정값을 유도하기 위해 사용되었을 때 이러한 추정값들을 합성추정값이라고 한다." Sarndal(1984)에 따

르면 이 정의는 합성추정이라고 알려져 있는 방법의 기반을 이루는 두 가지 생각을 표현하고 있다. : 그 추정값은 부분 추정값들의 혼합이며 이러한 부분 추정값들은 순수하다기 보다는 모형에 근거하고 있다.

아래와 같은 특성들이 주어지면 이 추정 방법이 연구 중인 지역영역에서 추정을 위한 적합한 방법이 될 수 있다.

(a) 한 지역에서 어떤 표본크기에 대해 실제 지역에서 나타나는 표본으로부터 얻은 정보만이 사용되었을 때 전통적인 방법으로 구한 추정량보다 더 좋은 정도(precision)를 갖는 추정량을 얻어낼 수 있다.
(b) 표본에 전통적인 방법으로는 추정값을 전혀 얻어낼 수 없을 만큼 관찰치의 수가 매우 적은 지역에 대해서 조차도 추정값들을 얻어낼 수 있다.

합성추정에 있어서는 아래의 요구조건들이 반드시 충족되어야 한다.

(c) 모집단의 각 단위에 대해 연구 대상이 되는 변수와 상관이 있는 보조정보가 있어야만 한다. 상관의 정도가 높아질수록 더 좋은 추정량을 얻을 수 있다.
(d) 모형에 대한 가정이 만족되어야 한다. : 큰 모집단 그룹들에서 관측되는 관계들이 작은 지역에 대해서도 역시 유효해야한다.

연구 대상이 되는 모집단은 보조변수에 따라 많은 그룹으로 나누어진

다. 보조변수와 연구대상이 되는 변수와의 관계에 대한 정보는 표본으로부터 얻어진다.

각 지역에 대한 합성 추정값은 그 지역의 부분 그룹들로부터 얻어진 추정값들의 가중합으로 계산된다. 가중치는 그 지역의 모집단 부분그룹에 있는 단위의 수로 구성된다. 그 절차는 앞서 언급했던 모형가정에 근거하고 있다. : 큰 그룹들에 있어서의 관계들은 작은 지역들에 대해서도 마찬가지로 유효하다. 만일 이 모형 가정이 맞다면 합성 추정은 그 자체로는 추정을 위한 기반이 충분하지 않은 소지역에서 추정에 필요한 정보들을 큰 그룹들로부터 "빌려온다"는 것에 강점이 있다. 한편, 만약 모형가정이 맞지 않다면 합성 추정법은 그 자신의 "순수한" 형태에서 편향 추정값을 생산한다. 그러나, Sarndal(1984)이 제시했던 추정량들은 모형가정이 틀린 경우에서조차 비편향이었다. 비편향성은 표본의 크기가 작은 소지역에서 정도(precision)가 낮을 수 있는 항을 더함으로서 얻어진다.

1.2 적용

이 논문에서 언급하고 있는 연구는 연구중인 지방영역에서 AKU에서 사용된 정의에 따른 실업율의 추정값을 향상시키려는 하나의 시도라고 할 수 있다. AKU는 매달 2만 2천명에게 노동시장에서의 그들의 상태에 관해 인터뷰를 한 표본조사이다. 소지역에서 실업률을 추정하는 것은 어려운 일이다. 왜냐하면 그 지역에서 표본 안에 있는 실업자의 수가 적을 뿐만 아니라 종종 하나도 없을 때도 있기 때문이다. 전체 표본크기는 크다고 할 수 있지만 그렇다고 할지라도 상대적으로 넓은 지역에 대

한 추정값들의 정도(precision)는 낮다. 일반적으로 지방 수준에서 현재 사용되고 있는-보통의 표본평균이 관련된-기법들에 의해 추정값들을 구하려 하는 것은 무의미한 일이다.

이 연구에서 사용되고 있는 보조정보는 각 지방자치단체(AMS)의 직업소개소에 의해 등록된 구직자의 수에 관한 통계량들로 구성되어 있다. AMS와 AKU의 통계량들간의 관계와 차이에 대한 설명은 3.3절에서 주어질 것이다. 현재의 연구에 있어서의 계산은 AMS 표본들이 AMS 등록자들에 맞추어 온 세 개의 다른 달들을 언급한다. 경험적 자료에 대한 더 자세한 설명은 3.2절에 나온다.

연구는 같은 경험자료들에 대해서 몬테칼로 시뮬레이션에 의해 수행되었으며[Cassel, Raback and Sarndal(1983)] 추정값들은 지방자치단체들의 군집들(소위 주(州))에 대해서 계산되었다. 이 연구의 결과는 현재 사용되고 있는 추정량보다 약 25%더 효율적인 비편향 합성추정량을 구하는 것이 가능하다는 것을 보여준다.

(2) 추정량에 대한 설명

아래 단락은 개인의 AKU와 AMS 상태를 설명하는데 사용된 표기법을 보여준다. 실업을 AKU에서는 AKU1으로 표기한다; 취업은 AKU0로 표기한다. AMS에의 등록 상태는 AMS1으로 표기한다; 미등록은 AMS0로 표기한다.

지역영역에서 AMS1의 수는 AMS로부터 주어지는 매달의 통계량들로부터 알 수 있다. AMS0는 모집단에서 AMS1을 뺌으로써 계산할 수 있

다. 우리의 목적은 AKU의 정의에 따라 주와 지방자치단체에서 16-64세인 사람들의 실업률을 구하는 것이다. P_0(AKU1)을 그 비율을 표기하는데 사용하기로 한다.

우리는 세 가지 유형의 추정량들에 대해 연구하였는데 이들은 아래의 특성들을 가지고 있다:

$\hat{P}_{1q}$는 각 지역 q에서 보조정보 없이 구해진 표본평균이다. 즉,

$$\hat{P}_{1q} = \frac{1}{n_q} \sum_{k \in s_q} Y_k$$

여기서

n_q = 표본에서 지역 q 에 있는 단위들의 수

s_q = 표본 s 에서 지역 q에 속하는 단위들의 집합

Y_k = 1, 만일 개인 k가 AKU1일 경우

= 0, 그 외의 경우

이 추정량은 비편향 추정량이며 s_q에 속하는 단위들만을 사용한다. n_q가 작아질수록 정도(precision)은 감소한다.

$\hat{P}_{2q}$는 비수정 합성추정량이다. 이것은 본래의 합성추정량이며 AMS에 등록된 사람들을 이용한다. 어떤 지역에 대한 추정값은 더 큰 그룹에서 관찰된 관계에 근거하는데 그 지역에서도 그 관계가 그대로 유효함을 가정하며 이 가정이 틀린 경우, 그 추정량은 편의를 가지게 된다. 이 추정량은 표본 전체를 이용한다.

모집단은 보조정보를 이용하여 H집단으로 나누어진다. 각 부집단 h

에서 실업율은 AKU 정의에 따라 부집단 h에 속한 단위들 사이에서 추정된다. 이 모수는 θ_h, $h=1, \ldots, H$ 로 표기하기로 한다.

추정량 $\hat{P}_{2q}$는 모수의 추정량 $\hat{\theta}_h$들의 가중합으로 만들어지며 이 가중은 부집단에 속한 지역의 각각의 비율이다. ; 즉,

$$\hat{P}_{2q} = \sum_{h=1}^{H} (N_{hq}/N_{.q})\hat{\theta}_h$$

그리고 $\hat{\theta}_h = \sum_{k \in s_h} Y_k/n_h$. 여기에서

N_{hq} = 모집단에서 부집단 h와 지역 q에 속하는 개인들의 수

N_q = 모집단에서 지역 q에 속하는 사람들의수

n_h = 표본에서 부집단 h에 속하는 사람들의 수

s_h = 표본 s에서 부집단 h에 속하는 부분

부집단의 수 H 와 그들의 정의는 모형가정에 의존한다. 3절에서 지금까지 시도했던 다른 모형들이 설명될 것이다.

$\hat{P}_{3q}$은 수정된 합성 추정량이며 역시 보조변수를 사용한다. 이것은 $\hat{P}_{2q}$와 같이 모형가정에 근거하고 있으며 각 지역에서 모형으로부터의 편차를 추정한다. 이 추정량은 잔차를 고려하기 때문에 모형에 결점이 있는 경우에도 (근사적으로) 비편향 추정량이다. 이 추정량은 두 개의 부분으로 구성되는데 $\hat{P}_{2q}$와 동일한 순수 모형 부분과 수정부분이다. 모형부분에서는 표본 전체로부터의 정보가 이용되고 수정부분에서는 지역으로부터의 정보와 전체 표본으로부터의 정보가 이용된다. 우리는 $\hat{P}_{3q}$를 $\hat{P}_{2q}$와 $\hat{P}_{1q}$와 비교할 것이다.

$$\hat{P}_{3q} = \hat{P}_{2q} + \sum_{s_q} e_k / \Pi_k$$

여기서 Π_k는 k단위의 포함 확률이고 $e_k = Y_k - \hat{Y}_k$ 이다. 만약 단위 k가 집단 h, $h=1,\ldots,H$ 에 속한다면 $\hat{Y}_k = \hat{\theta}_h$ 가 된다.

(3) 모 형

이 장에서는 AMS와 AKU간의 관계를 자세히 다루려 하는데 이는 많은 모형을 구축하는데 있어서 이 관계를 이용하기 위함이다; 그 후에는 이 모형들을 자세히 살펴볼 것이다.

3.1 AMS와 AKU간의 공변량

<그림1>은 AMS 변수가 AKU변수 사이에 시간에 걸쳐서 매우 강한 상관관계가 있음을 보여준다.

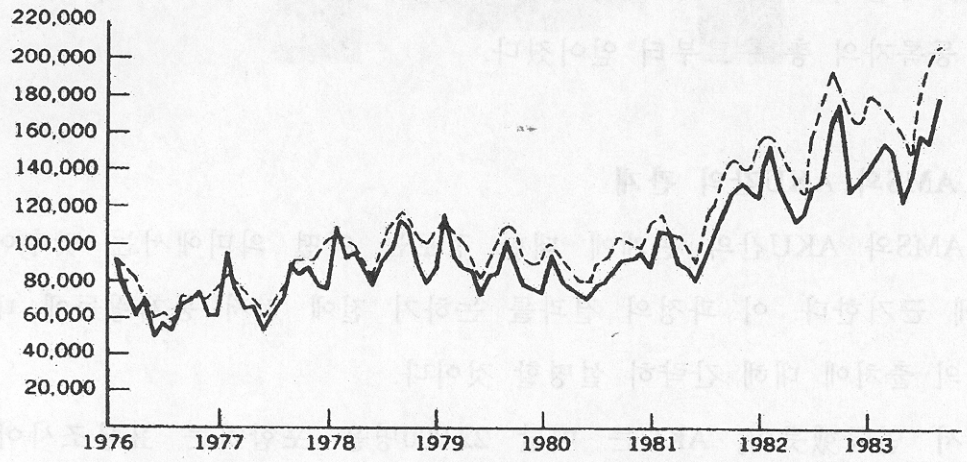


Figure 1. The number of unemployed measured by AMS (dotted line) and AKU (solid line) in 1976–1983.

3.2 연구를 위한 경험적 지지

AKU에 따른 실업자의 수와 AMS에 등록된 구직자의 수간의 관계를 자세히 알아보기 위해서는 AKU와 AMS가 맞추어진 자료(matching-data)가 사용되었다. 1982년 11월, 1983년 1월 그리고 1983년 5월의 AKU자료에 AMS의 구직자 등록자료가 더해졌다. AKU자료는 AKU 조사에서 측정을 위해 선택된 주의 다음에 오는 화요일에 나타나므로 AMS등록에 대해 맞추어졌다. 그러나 AMS 통계량에 달마다 계산된 등록 버전은 그 달의 마지막으로 일한 날을 언급한다. AKU 표본에 있는 62,661명 중에 2,128명이 AMS에 구직자로 등록되어 있었다. AKU에서의 무응답 때문에, AKU 조사에서 인터뷰를 했고 동시에 AMS에 등록된 사람의 수는 1,926명으로 감소하였다.

등록된 사람들 중 모집단에서 성별, 나이 그리고 지방자치단체에 의해

누락된 사람의 수는 1982/83 RTP(the Register of the Total Population)과 AMS 등록자의 총 수로부터 얻어졌다.

3.3 AMS와 AKU간의 관계

AMS와 AKU간의 관계에 대한 정보는 어떤 의미에서는 맞추어진 자료에 근거한다. 이 과정의 결과를 논하기 전에 먼저 통계량들에 대한 두 개의 출처에 대해 간략히 설명할 것이다.

앞서 언급했듯이, AKU는 매달 22,000명을 포함하는 표본조사이다. 전화인터뷰에 의해서 어떤 기간동안 개인의 취업 상태에 대한 자료가 모아진다. 한 개인이 실업상태로 고려되는 것은 다음의 경우들인데 조사기간동안 실업 상태에 있으면서 동시에 (1) 구직중이거나 지난 60일 동안 만든 직업신청서의 결과를 기다리는 중인 경우 또는 (2) 월급을 받지 않는 상태에서 임시로 쉬고 있는 직업으로의 재고용을 기다리는 경우 또는 (3) 아프지 않은 상황에서라면 직업을 구하는 경우이다. AKU 조사에서 무응답 비율은 약 6%정도이다.

AMS 통계량은 구직자로 등록되어 있으면서 10일 이내에 일할 준비가 되는 직업이 없는 사람들의 총 수에 근거한다. 등록의 기본 요소는 "statistics card"인데 이것은 직업소개소에 의해 각 개인에게 뽑혀진다.

신청자가 직업소개소를 재방문 할 때마다 등록정보는 갱신되며 약 4주안에 다시 방문하라는 약속을 받게 된다. 따라서 만일 지원자의 취업상태의 변화가 일어나게 되면 직업소개소가 이러한 사실을 알게 되기까지 약 4주의 시간이 지나게 된다. 이 시간 경과 때문에 어떤 단위에 대

해서는 등록에 대한 정보가 자료에 포함되지 않을 수도 있다.

이러한 면에서 볼 때 실업에 대한 통계량의 두 가지 출처들은 서로 다르다고 할 수 있다. <표 1>은 AMS와 AKU간의 관계를 요약해 놓은 것이다. 이 표에 나타난 숫자들은 맞추는 절차(matching-procedure)로부터 얻은 자료들에 근거하고 있다. 약 66,000단위들이 이 맞우는 작업에 관련되어 있기 때문에 이들은 믿을 만한 추정으로 보여진다. AKU에 따르면 AMS에 등록된 사람들 중에서 62%가 실업상태에 있다.

합성추정량을 만들기 위해 사용된 첫 번째 모형(이후 모형 A)은 $H=2$인 그룹에 근거한다. 더 정확히 말하자면, 국가적 수준에서 맞추어진 자료로부터 각각 62%와 0.9%로 추정된 모수 θ_1(θ_1은 AMS1에 포함되면서 AKU1에도 포함되는 비율이다)과 θ_2(θ_2는 AMS0에 포함되면서 AKU1에도 포함되는 비율이다)는 각 지방에 대해서도 거의 비슷해야 한다. 이러한 비율들이 국가 전체의 값들과 일치하지 않는 지방들에서 합성추정량은 편향된 추정값을 산출할 것이다.

Table 1. The Relation between AMS and AKU[a]

AKU	AMS		
	AMS0 (Not Registered Applicants for Work)	AMS1 (Registered Applicants for Work)	Total
AKU0 (Employed)	96.2	1.1	97.3
AKU1 (Unemployed)	0.9	1.9	2.7
Total	97.0	3.0	100.0

[a] This table shows percentage figures for the cross-classification of AMS status with AKU status.

그 모형이 얼마나 적용가능한가를 판단하기 위해서 AMS와 AKU가 일치하지 않는 원인들을 살펴보아야 할 필요가 있다. 만일 AMS와 AKU의 실업에 대한 정의에 서로 차이가 없다면 <표 1>의 비대각 셀에 있는 값들은 0이 되어야 할 것이다. 그러나 표1의 비대각 셀에 있는 값들은 0이 아니다. 다음 절에서 우리는 이 현상에 대한 설명을 찾아볼 것이다. 이러한 고려들은 모형구축의 기반을 형성한다.

3.4 AMS와 AKU 통계량들 간의 차이에 대한 몇 가지 이유들 및 이러한 차이들이 모수 θ_1과 θ_2에 미치는 영향

AMS와 AKU 통계량들 간에 관찰되는 차이점들에 대해서 두 가지

가능한 원인을 살펴보려고 한다.

(a) AMS와 AKU간 정의의 차이들과 AMS 등록 갱신 과정

직업신청자들의 등록 갱신에 있어서 발생하는 시간 간격 때문에 등록부는 더 이상 직업을 구하지 않는 사람들을 많이 포함한다. 이러한 사람들의 수는 노동 시장의 조건에 의존한다. AMS등록 자료에 따르면 1982년에 스웨덴 전체 인구 중에서 4주의 기간동안 일자리 구하는 것을 그만둔 사람의 수가 20,000에서 40,000명 정도가 된다. 이 변이는 계절적 패턴은 따른다. 결국 더 이상 일자리를 구하지 않는 사람들의 수가 감소할 때 θ_1이 증가해야 한다. 그러나 이 계절적 패턴의 결과로서 θ_1은 $\pm 1\%$도 변하지 않는 다고 생각된다. 만일 시간 차이가 전혀 없다면, 다시 말해서, 만일 모든 등록된 사람들이 실제로 여전히 일자리를 찾고 있다면 θ_1은 아마도 최대한 65%까지 오를 것이다.

맞추어진 자료의 등록자들 중에서 62%는 AKU에 따르면 실업 상태인 것으로, 26%는 취업 상태인 것으로 그리고 12%는 노동자가 아닌 것으로 분류되었다. 앞서 등록갱신의 과정은 이러한 차이를 작은 부분에서만 설명할 수 있다는 것을 살펴보았었다. 실업의 정의들은 어떤 사람이 AMS의 의미에서는 일자리를 구하고 있더라도 동시에 AKU의 의미에서는 여전히 취업상태일 수도 있는 그런 것들이다. AMS에서 "직업이 없는 구직자들"이라는 단어는 현재 직업이 없으며 10일 안에 일을 가질 준비가 되어 있는 사람들을 의미한다. 이 그룹에는 10일 안에 직업을 그만 두는 사람들과 부업을 하는 사람들도 포함된다. 한편, AKU는 조사하는 주 동

안 적어도 1시간동안 일을 했을 경우 이러한 사람들은 취업자로 간주한다. AMS에서는 신청자가 10일 이하의 단기간 직업을 가지고 있는 경우 그들이 전임으로 일을 해왔을 수도 있다는 사실에도 불구하고 직업이 없는 신청자들로 간주된다. 따라서 이 그룹에 있는 모든 사람들은 단기간 직업을 가지고 있는 신청자가 될 가능성도 있다. 임시직을 가지고 있는 신청자들의 수를 모르기 때문에 그들의 존재가 어느 정도까지 26%의 취업상태를 설명할 것인가를 추정할 수는 없다. 노동시장에 새로 들어왔거나 다시 진입한 사람들은 아마도 임시직을 가진 신청자들의 큰 부분을 차지하고 있을 것이다

(b) Cause due to Age-Specific Reasons

우리의 경험적 자료들은 취업상태인 등록된 지원자들의 비율은 55세 이하인 경우 더 나이가 많은 사람들인 경우에 비해 그리고 여성인 경우 남성인 경우보다 더 많은 부분을 차지한다는 것을 보여준다(표2). AKU에 따른 등록된 실업상태의 지원자들에 관해서 말하자면 약 40% 정도는 단지 그들의 존재를 나타내기 위해 직업소개소를 방문하는 사람들로 구성되어 있다고 생각된다. 회사들이 그들의 직원을 감축할 때, 종종 58세에 이른 사람들에게 조기 은퇴 수당을 우선적으로 지급한다. 이 제안을 수락한 사람들은 직업소개소에 등록이 되며 국가로부터 연금을 받을 수 있는 60번 째 생일날까지 그 곳에 나타난다. 원칙적으로 그런 사람들은 오직 그들의 삶을 보장하기 위해서 등록이 되어 있다는 사실에도 불구하고 노동시장의 처분에 따르게된다. 그러나 이 사람들을 인터뷰해보면 이들

은 이미 노동시장을 떠났다고 말할 것이다. 이러한 사실은 <표 2>에서 볼 수 있다.

<표 1>에서 보았듯이, 많은 실업인구들이 AMS에 등록되어 있지 않다. 물론 AMS에 등록되지 않은 상태에서 직업을 구하고 있을 수도 있다. : AKU에 따르면(1983년의 평균값) 실업자의 약 10%(15,000명)가 AMS에 등록되지 않은 채 직업을 구하고 있었다. AKU에서 실시한 인터뷰에 따르면 AMS에 등록된 사람이라고 주장하는 사람들 중에서조차 실제로는 등록되지 않은 채 직업을 구하고 있는 사람들도 얼마정도 있었을 것으로 보인다.

Table 2. The Proportions of People Unemployed, Employed, and Not in the Labor Force, among Individuals Registered by AMS, Classified by Sex and Age(%)

Age	Sex					
	Male			Female		
	Unemployed θ_1	Employed	Not in the Labor force	Unemployed θ_1	Employed	Not in the Labor force
16-24	66.8	25.4	7.8	57.5	33.6	8.9
25-54	67.3	26.3	6.4	57.3	31.0	11.7
55-64	68.2	11.7	20.1	55.4	14.0	30.6

무엇보다도, 재정적 권유가 없기 때문에 AMS에 등록되지 않은 채 일자리를 구하고 있는 사람들은 바로 노동시장에 처음 들어온 사람들일 것이라고 추측하는 것도 일리가 있다. <표 3>은 신청자로 등록되지 않

은 사람들 중 실업자의 비율(θ_2)을 보여준다. <표 3>에서 본 것처럼 θ_2는 연령에 따라 변화한다. θ_1과 θ_2는 성별과 연령 그룹들에 따라 변화하므로 만약 모집단이 성별과 연령 그룹별로 나누어진다면(아래 모델 E를 보라) 모형가정은 아마도 더 실제적이 될 것이다. 그러나 우리의 결과는 그룹의 수 H가 늘어남에 따라 추정된 분산의 크기가 반드시 줄어드는 것은 아니라는 것을 보여준다.

Table 3. Proportion of Unemployed Among Individuals Not Registered as Applicants by AMS(%)

Age	Male	Female
16-24	1.5	1.7
25-54	0.8	0.7
55-64	0.5	0.4

3.5 Alternative Classification of the Population

개인의 노동시장 상황에 더욱 강조점을 둔 모집단 분류방법에 대한 하나의 대안은 일종의 재정적 보상으로부터 만들어질 수 있다. 예를 들어 지원자들은 세 개의 그룹으로 나누어 질 수 있다. 첫 그룹에 있는 지원자들은 노동 시장 보상을 현금으로 받고(KAS), 두 번째 그룹은 기금 보상이라고 불리는 또 다른 형태의 보상을 받으며 세 번째 그룹은 보상을 받지 않는 것이다. 이 그룹들에서 AMS에 의해 등록된 개인들 가운데 AKU에 따른 실업율은 각각 60%, 66% 그리고 54%가 된다. 전체 비율은

62%이다.

다른 종류의 보상은 개인의 연관성에 대한 분류를 노동시장에 동시에 제공한다. 기금 보상을 받은 신청자들은 시장에 가장 강한 연관성을 가진다. 보상을 받지 않은 신청자들은 넓은 의미에서 보면 노동시장에 새로 진입한 사람들이며 현금으로 보상을 받는 신청자들은 이 두 집단 사이 어딘가로 분류될 수 있다.

이 세 그룹의 분포는 아마도 노동 시장의 지역에 따라 다를 것이다. 예를 들어 주력 산업이 문을 닫은 지역과 신청자의 대부분이 학생인 지역을 비교해 보라. 그룹들의 분포는 아마도 계절에 따라 달라질 것이다. : 여름이 끝난 직후 많은 사람들이 시장에 많이 진입하게 될 것이라고 생각된다. 그 그룹들은 또한 시장의 상태에 따라서도 많이 변할 것이다. : 잘 설립된 시장이라고 할지라도 불황에 영향을 받을 것이며 이는 신청자 중에 많은 사람들이 보상을 받을 자격을 갖게 됨을 의미한다.

이 모집단의 분류는 더욱 실제적인 모형을 만드는 것뿐만 아니라 시간에 걸쳐 나타나는 연관성들의 변이를 줄여서 높은 정도(precision)로 모수들을 추정하는 일의 더욱 가능하게 하는데 그 목적이 있다.(모형 E를 보라)

3.6 Description of the Model Used

모형들은 모집단이 나누어지는 그룹의 수 H이 수와 보조 변수들이 이용되는 방법에 의해 구분된다.

모형 A는 모집단을 두 그룹으로 나눈다. 즉, $H=2$이다. 한 그룹은 AMS0로 나머지 하나는 AMS1로 구성된다.

모형 B는 $H=48$개의 그룹을 이용한다. 그 그룹들은 AMS0와 AMS1을 지방에 따라 세분하여 만든다.(스웨덴은 24개의 지방으로 나누어진다.) 이 모형은 AMS와 AKU의 상태간의 관계가 각 지방에 따라 다르다고 가정한다.

모형 C는 AMS의 상태, 성별 그리고 연령에 관한 정보를 이용하며 $H=12$이다. 여섯 개의 연령/성별 그룹들이 두 개의 AMS상태 범주와 결합되어 있다.

우리는 모형 C^{1}과 C^{11}도 시도해 보았다. 모형 C^{1}은 모형 C를 더욱 정교하게 만든 것이다; 지방의 정보가 더해졌다. 모형 C^{11}은 성별과 연령에 관한 정보만을 이용하므로 모형 C보다 덜 정교하다. 우리는 모형 C가 모형 C^{11}에 비해 평균적으로 더 작은 신뢰구간을 생성한다는 점에서 더 나은 모형이며 모형 C^{1}에 의해서는 실제적으로 향상되는 바가 없음을 알게 되었다. 따라서 여기에서는 모형 C하에서의 추정량들에 대해서만 서술할 것이다.

모형 D는 AMS의 상태와 소위 H-지역(스웨덴은 8 H-지역으로 나누어지는데 이들 각 지역은 유사한 구조를 갖는 지방으로 구성된다)에 관한 정보를 이용한다. $H=16$이다.

모형 E에서 $H=4$이다. 그룹1은 AMS0로 구성되어 있다. 나머지 세 개의 그룹들은 AMS1 그룹을 기금 보상을 받는 개인들, 현금 보상을 받는 개인들 그리고 보상을 받지 못하는 개인들로 세분해서 만들어진다.

(4) 결 과

4.1 모형 모수의 추정값

합성추정량 $\hat{P}_{2q}$와 $\hat{P}_{3q}$는 특정한 그룹들 안에서 AKU에 따른 실업률의 추정값들의 선형결합으로 만들어진다. $\hat{\theta}$을 그룹 h에서 AKU1의 비율이라고 하고 $\hat{\theta}_h$를 그 추정값이라고 하자. 우리는 연구 대상이 되는 대부분의 모형들에 대한 이 추정값들을 주는 것으로부터 시작할 것이다. 만일 보조정보가 효과가 있다면 그룹에 따라 $\hat{\theta}_h$간에는 많은 차이가 있을 것이다. 이 적용에서는 각기 다른 시점에 대해 $\hat{\theta}_h$들이 안정적이어야만 한다. <표 4>는 $\hat{\theta}_h$가 어떻게 달라지는지를 보여준다.

모형 A에서는 추정값 $\hat{\theta}_1$과 $\hat{\theta}_2$가 각각 AMS1에서 AKU1의 비율과 AMS0에서 AKU1의 비율과 관련이 있다.

모형 B에서는 $\hat{\theta}_{1j}$ 와 $\hat{\theta}_{2j}$ 가 지방 j에 대해서 모형 A에서와 같은 비율과 관련이 있다.

모형 C에서 추정값 $\hat{\theta}_{21}$은 연령이 16-24인 남성 그룹의 AMS0 중 AKU1의 비율과 연관되어 있고 $\hat{\theta}_{22}$는 연령이 25-54인 남성 그룹의 AMS0 중 AKU1의 비율과 연관되어 있으며 $\hat{\theta}_{23}$은 연령이 55-64인 남성 그룹의 AMS0 중 AKU1의 비율과 연관되어 있다. $\hat{\theta}_{24}$- $\hat{\theta}_{26}$은 $\hat{\theta}_{21}$- $\hat{\theta}_{23}$과 같은 연령대 여성 그룹의 각 대응되는 비율과 연관되어 있다. $\hat{\theta}_{11}$- $\hat{\theta}_{16}$은 전과 같은 연령/성별 그룹의 AMS1 중 AKU1의 비율과 연관되어 있다.

모형 E에서 $\hat{\theta}_2$는 AMS0 중 AKU1의 비율과, $\hat{\theta}_{11}$은 현금 보상을 받는 AMS1 중 AKU1의 비율과, $\hat{\theta}_{12}$는 기금 보상을 받는 AMS1 중 AKU1의 비율과 그리고 $\hat{\theta}_{13}$은 보상을 받지 못하는 AMS1 중 AKU1의 비율과 연관되어 있다.

여기서 한 가지 짚고 넘어가야 할 것은 이전의 값들은 표본 변이를 전제로 한, 실제 모수값이 아닌 추정값들이다. 그룹의 수가 많아지면 정도(precision)가 줄어드는데 이는 추정값이 더 적은 관측치들에 의해 구해지기 때문이다. 이 사실은 어려운 문제를 말해준다 : 모형은 실재와 일치해야 하지만(그래서 꽤 자세해야 하지만) 동시에 정도(precision)를 고려할 때 모형은 적은 모수를 가지고 있는 것이 바람직하다는 것이다. <표 5>를 통해 대부분의 효과적인 보조정보는 이미 모형 A에서 소개되었다는 것을 알 수 있다. 다른 모형들에 있는 추가보조정보들이 얼마나 정확도에 영향을 줄 수 있을 것인가 하는 것은 매우 중요한 질문이다. 모형 편의를 줄이기 위해 추가 정보를 소개하는 것은 어쩌면 정도(precision)에 있어서 받아들일 수 없는 손실을 가져올지도 모른다.

Table 4 . Estimated θ_h in Different Groups According to Models A, B, C, and E

Model A	θ_1	θ_2
November	0.607	0.007
January	0.595	0.008
March	0.657	0.007

Model B	November		January		March	
j	$\hat{\theta}_{1j}$	$\hat{\theta}_{2j}$	$\hat{\theta}_{1j}$	$\hat{\theta}_{2j}$	$\hat{\theta}_{1j}$	$\hat{\theta}_{2j}$
1	0.525	0.004	0.490	0.010	0.584	0.006
2	0.655	0.005	0.489	0.009	0.775	0.004
3	0.728	0.009	0.719	0.014	0.487	0.014
4	0.594	0.005	0.376	0.006	0.657	0.006
5	0.709	0.004	0.596	0.006	0.582	0.004
6	0.575	0.006	0.654	0.019	0.754	0.005
7	0.752	0.004	0.631	0.006	0.710	0.007
8	0.623	0.008	0.624	0.007	0.753	0.005
9	0.716	0.001	0.532	0.003	0.722	0.001
10	0.740	0.012	0.636	0.003	0.716	0.009
11	0.605	0.010	0.695	0.008	0.580	0.008
12	0.490	0.004	0.716	0.008	0.495	0.009
13	0.585	0.006	0.607	0.007	0.570	0.010
14	0.483	0.015	0.562	0.006	0.743	0.011
15	0.730	0.011	0.416	0.004	0.686	0.006
16	0.604	0.003	0.512	0.011	0.668	0.006
17	0.623	0.004	0.612	0.007	0.697	0.007
18	0.700	0.010	0.693	0.009	0.588	0.012
19	0.403	0.005	0.523	0.013	0.709	0.005
20	0.706	0.005	0.461	0.007	0.716	0.001
21	0.662	0.007	0.810	0.007	0.630	0.009
22	0.487	0.003	0.660	0.013	0.709	0.001
23	0.425	0.008	0.713	0.011	0.662	0.011
24	0.603	0.010	0.540	0.006	0.813	0.013

Model C	$\hat{\theta}_{11}$	$\hat{\theta}_{12}$	$\hat{\theta}_{13}$	$\hat{\theta}_{14}$	$\hat{\theta}_{15}$	$\hat{\theta}_{16}$
November	0.665	0.657	0.709	0.576	0.579	0.432
January	0.604	0.682	0.620	0.589	0.519	0.547
March	0.735	0.684	0.716	0.560	0.620	0.648
	$\hat{\theta}_{21}$	$\hat{\theta}_{22}$	$\hat{\theta}_{23}$	$\hat{\theta}_{24}$	$\hat{\theta}_{25}$	$\hat{\theta}_{26}$
November	0.016	0.005	0.003	0.016	0.007	0.002
January	0.014	0.010	0.002	0.017	0.009	0.002
March	0.014	0.010	0.003	0.018	0.006	0.002

Model E	$\hat{\theta}_0$	$\hat{\theta}_{11}$	$\hat{\theta}_{12}$	$\hat{\theta}_{13}$
November	0.007	0.571	0.637	0.572
January	0.008	0.618	0.635	0.498
March	0.007	0.616	0.717	0.552

4.2 지역단위에서의 실업율 추정

이 연구의 주된 목적은 연구 대상이 되는 지역으로서 Municipalities를 이용하여 P_0(AKU1)을 추정하는 것이다. 앞서 언급했던 것처럼 AKU 조사는 매달 스웨덴에 살고 있는 16-24세의 사람들 중 22,000명의 표본을 뽑아서 실시하는 것이다. 이 표본은 지방, 성별, 결혼여부 그리고 시민여부에 따라 층화되며 하나의 층 안에서 사람들은 임의 추출된다. 표준오차는 잘 알려진 층화 표본의 분산 공식에 의해 바로 계산된다. 본 논문에서는 이 공식을 위해 지면을 할애하지 않으려고 한다. <표 5>는 선택된 몇 개 지방에 대해서 1982년 11월 모수의 추정값들과 표준편차들을 보여준다.

각 지방에 대해서 <표 5>는 거주자의 수, 표본에 포함된 사람의 수, 추정값 $\hat{P}_{1q}$과 더불어 그 옆 괄호 안에 $\hat{P}_{1q}$의 표준편차를 보여준다. 그리고 이어서 $\hat{P}_{2q}$와 $\hat{P}_{3q}$에 대해서도 대응되는 추정값들이 나타나 있다. 합성 추정량 $\hat{P}_{2q}$와 $\hat{P}_{3q}$는 네 개의 다른 모형 $A-D$에 대해 각각 계산되었다. 각각의 모형에 대해 $\hat{P}_{2q}$의 표준오차가 $\hat{P}_{1q}$와 $\hat{P}_{3q}$의 그것에 비해 현저하게 작다는 것을 알 수 있다. $\hat{P}_{2q}$의 표준오차는 그룹의 수가 많아질수록 증가한다.AMS 정보를 사용한 모형에서 $\hat{P}_{3q}$의 표준오차는 $\hat{P}_{1q}$의 표준오차에 비해 작다.

각 지방에 대해 P_0(AKU1)의 구간을 계산할 수 있다. 예를 들면 Botkyrka에 대해 95%신뢰 한계를 이용하여 $\hat{P}_{1q}$에 의한 구간은

(0.05-4.34)%, 모형 A에서 $\hat{P}_{2q}$에 의한 구간은 (1.65- 1.91)% 그리고 모형 A에서 $\hat{P}_{3q}$에 의한 구간은(0.56-4.00)이 된다. $\hat{P}_{3q}$를 이용한 구간은 $\hat{P}_{1q}$을 이용한 구간에 비해 다소 짧다. 그러나 $\hat{P}_{2q}$를 이용한 구간은 충분히 짧다 : 나머지 두 개는 매우 넓다. 다른 모형들에 대한 비교 역시 비슷한 결과를 보여준다. $\hat{P}_{2q}$는 모형에 매우 크게 의존하며 만일 모형이 유효하지 않을 때에는 편의가 생길 수도 있다는 사실을 유념해야 한다. 그러한 경우에 구간의 신뢰수준은 미지의 편의 때문에 알 수가 없게 된다.

Table 5. Estimates of P_0(AKU1) with Estimated Standard Errors(Within Brackets) for Selected Municipalities using Models A-D for AKU in November 1982

Municipality	Number of Inhabitants 16-64 Years	n_q	$\hat{P}_{1q}$	Models			
				A		B	
				$\hat{P}_{2q}$	$\hat{P}_{3q}$	$\hat{P}_{2q}$	$\hat{P}_{3q}$
Botkyrka	44,119	175	2.24(1.12)	1.78(0.07)	2.28(0.88)	1.37(0.17)	2.37(0.97)
Stockholm	422,694	1,358	0.79(0.24)	1.63(0.06)	0.86(0.23)	1.24(0.16)	0.90(0.23)
Tierp	12,028	57	0.00	2.05(0.07)	-0.07(0.98)	2.02(0.44)	-0.02(1.05)
Katrineholm	19,715	108	3.42(1.75)	2.71(0.09)	5.20(1.33)	3.33(0.51)	5.20(1.28)
Kinna	5,897	21	3.99(4.27)	2.33(0.08)	0.09(3.19)	2.17(0.41)	0.24(4.36)
Gnosjo	5,514	13	7.73(7.39)	1.42(0.06)	2.88(3.67)	1.25(0.27)	2.41(2.72)
Gotland	34,492	636	1.95(0.54)	1.94(0.07)	2.19(0.47)	2.16(0.46)	2.16(0.47)
Malmo	149,195	450	3.65(0.88)	3.81(0.12)	4.32(0.72)	4.11(0.43)	4.29(0.77)
Goteborg	278,468	910	1.88(0.45)	2.81(0.09)	2.56(0.39)	2.67(0.33)	2.55(0.39)
Laxa	5,218	15	6.51(6.36)	3.38(0.10)	4.78(2.96)	3.22(0.55)	4.71(2.84)
Arjeplog	2,494	8	19.96(14.11)	3.36(0.12)	11.88(5.54)	4.18(0.56)	11.87(5.51)

Municipality	Number of Inhabitants 16-64 Years	n_q	$\hat{P}_{1q}$	Models			
				C		D	
				$\hat{P}_{2q}$	$\hat{P}_{3q}$	$\hat{P}_{2q}$	$\hat{P}_{3q}$
Botkyrka	44,119	175	2.24(1.12)	1.92(0.08)	2.23(0.87)	1.39(0.17)	2.36(0.97)
Stockholm	422,694	1,358	0.79(0.24)	1.68(0.07)	0.93(0.24)	1.25(0.16)	0.90(0.23)
Tierp	12,028	57	0.00	2.11(0.07)	-0.21(1.14)	2.16(0.15)	-0.06(1.06)
Katrineholm	19,715	108	3.42(1.75)	2.77(0.09)	5.21(1.32)	2.88(0.18)	5.24(1.31)
Kinna	5,897	21	3.99(4.27)	2.36(0.08)	0.16(3.93)	2.27(0.2541)	0.03(3.20)
Gnosjo	5,514	13	7.73(7.39)	1.50(0.07)	3.11(3.94)	1.48(0.13)	2.72(3.22)
Gotland	34,492	636	1.95(0.54)	2.03(0.08)	2.27(0.47)	1.69(0.21)	2.18(0.47)
Malmo	149,195	450	3.65(0.88)	3.85(0.12)	4.34(0.77)	3.81(0.36)	4.27(0.77)
Goteborg	278,468	910	1.88(0.45)	2.88(0.09)	2.70(0.38)	2.90(0.28)	2.47(0.39)
Laxa	5,218	15	6.51(6.36)	3.42(0.11)	4.93(3.18)	3.59(0.22)	4.77(2.60)
Arjeplog	2,494	8	19.96(14.11)	3.85(0.13)	11.58(5.37)	3.83(0.35)	11.73(5.41)

$\hat{P}_{1q}$와 $\hat{P}_{3q}$에 의한 구간이 매우 넓기 때문에 구간의 하단은 종종 0보다 작게 나오기도 한다. 추정값 $\hat{P}_{3q}$는 음수로까지 나타나기도 한다(Tierp 지방). 이러한 받아들일 수 없는 추정값들은 그 지역으로부터의 표본이 AMS1과 AKU0의 성질을 가진 단위를 많이 포함하고 있을 때 발생한다. 이런 경우 $\hat{P}_{3q}$에서 수정항의 영향은 비정상적으로 커지게 된다. 대응하는 $\hat{P}_{1q}$는 종종 0의 값을 가지게 되며 이 때 $\hat{P}_{1q}$을 이용하여 신뢰구간을 구하는 것은 불가능하다.

모형 A는 모형에서는 C에 비해 약간 짧은 신뢰구간이 만들어진다. 모형 A와 C는 모형 B와 D에 비해서 짧은 신뢰구간을 생성한다. 만일 짧은 신뢰구간을 만들어내는 능력에 의해 모형의 적합도를 측정한다면

모형 A가 가장 좋다. 그러나 AMS의 상태, 성별 그리고 나이에 관한 정보를 이용하는 모형 C는 모형 A에 비해 더욱 자세하며 따라서 모형 C가 모형 A에 비해 더 실제적인 모형이 될 것으로 기대된다. 대신 얼마만큼의 정도(precision)를 잃게 되는데 이는 추정해야 할 모수의 수가 더 많아지기 때문이다.

<표 6>은 많은 지방들에서 세 개의 측정기간 동안 추정된 $\hat{P}_{1q}$, $\hat{P}_{2q}$ 그리고 $\hat{P}_{3q}$의 값들을 보여준다. 첫 번째 열에서 우리는 P(AMS1)이 월별로 어떻게 변하는지를 알 수 있다. $\hat{P}_{2q}$와 $\hat{P}_{3q}$의 완비성은 각 모형에 포함되어 있다. 지방의 이름 옆에 있는 괄호 안에 있는 수는 그 지역에서의 평균 관측치 수를 나타낸다. 1월과 3월에 대해서 우리는 11월과 같은 결과를 이끌어낼 수 있다. $\hat{P}_{1q}$와 $\hat{P}_{3q}$에 대한 큰 표본 변이가 크기 때문에 월별 추정값의 차이가 많이 난다. P(AMS1)의 큰 월간 변이는 다소 패턴이 없어 보이기도 하지만 일반적으로 비율은 11월에서 1월로 갈수록 증가하고 1월에서 3월로 갈수록 감소한다.

Table 7. Comparisons of P(AMS1) and of the Estimates of P(AKU1) for Selected Municipalities[a]

			Models							
			A		B		C		D	
P(AMS1)	$\hat{P}_{1q}$		$\hat{P}_{2q}$	$\hat{P}_{3q}$	$\hat{P}_{2q}$	$\hat{P}_{3q}$	$\hat{P}_{2q}$	$\hat{P}_{3q}$	$\hat{P}_{2q}$	$\hat{P}_{3q}$
Sollentuna(115)										
N	0.90	2.27	1.20	2.95	0.86	2.99	1.79	3.01	0.87	2.98
J	1.16	0.83	1.51	1.21	1.52	1.10	1.65	1.32	1.57	1.11
M	1.01	3.74	1.41	3.66	1.25	3.83	1.52	3.70	1.19	3.85
Stockholm(1325)										
N	1.61	0.79	1.63	0.86	1.24	0.90	1.68	0.93	1.25	0.90
J	1.65	1.90	1.80	2.34	1.76	2.27	1.88	2.46	1.81	2.27
M	1.51	1.08	1.73	1.58	1.53	1.54	1.79	1.69	1.47	1.55
Enkoping(100)										
N	2.94	1.29	2.43	2.58	2.43	2.72	2.48	2.62	2.57	2.67
J	3.34	1.85	2.79	1.40	2.51	1.50	2.89	1.45	2.87	1.36
M	3.00	3.25	2.70	1.97	2.71	1.69	2.79	2.08	2.85	1.83
Nykoping(179)										
N	2.47	1.89	2.14	2.58	2.69	2.67	2.20	2.67	2.26	2.62
J	2.88	2.94	2.52	2.67	3.45	3.60	2.61	3.86	2.59	3.70
M	2.44	4.42	2.34	3.84	2.51	4.12	2.42	3.84	2.47	3.75
Hultsfred(51)										
N	4.50	8.22	3.36	5.32	3.76	4.16	3.43	4.96	3.58	4.91
J	4.49	2.29	3.46	3.59	3.37	3.68	3.54	3.69	3.60	3.69
M	4.18	6.37	3.47	4.90	3.65	4.73	3.55	4.90	3.67	4.76
Gotland(636)										
N	2.13	1.95	1.94	2.19	2.16	2.16	2.03	2.27	1.69	2.18
J	2.25	1.92	2.15	2.08	2.10	2.11	2.27	2.20	1.90	2.14
M	4.18	6.37	3.42	4.90	3.65	4.73	3.55	4.90	3.67	4.76
Malmo(450)										
N	5.24	3.64	3.81	4.22	4.11	4.29	3.85	4.39	3.81	4.26
J	5.49	3.84	4.05	4.63	4.63	4.77	4.15	4.77	4.09	4.69
M	5.40	3.87	4.26	3.92	3.92	4.40	4.32	4.65	4.11	4.38

[a] N≡November; J≡January; M≡March.

<표 7>은 추정값들에 대한 시간에 따른 변이를 다른 각도에서 보여준다 : 이 표는 크고 작은 차이의 빈도를 나타낸다. $\hat{P}_{1q}$와 $\hat{P}_{3q}$에 대해 월간에 큰 차이들의 빈도는 $\hat{P}_{2q}$에 비해 현저하게 높다.

Table 8. The Frequency of Large and Small Differences in P(AMS1) and in Estimates of P(AMS1) between Months[a]

			Models							
			A		B		C		D	
Difference	P(AMS1)	$\hat{P}_{1q}$	$\hat{P}_{2q}$	$\hat{P}_{3q}$	$\hat{P}_{2q}$	$\hat{P}_{3q}$	$\hat{P}_{2q}$	$\hat{P}_{3q}$	$\hat{P}_{2q}$	$\hat{P}_{3q}$
From November to January										
below -3.0%	2	16	1	11	1	11	1	10	0	10
-3.0～-1.5%	7	13	3	14	4	14	4	15	4	15
-1.5～-0.3%	11	7	6	13	6	18	6	15	9	15
-0.3～ 0.3%	25	15	36	13	35	11	35	11	38	12
0.3～ 1.5%	25	9	37	13	38	11	38	12	33	13
1.5～ 3.0%	21	12	13	15	15	12	15	15	14	14
over 3.0%	6	23	2	21	2	23	2	21	3	21
From January to March										
below -3.0%	1	18	1	18	1	18	1	19	1	18
-3.0～-1.5%	8	7	3	12	4	4	3	12	4	9
-1.5～-0.3%	19	9	10	13	17	9	9	13	10	12
-0.3～ 0.3%	50	21	64	19	40	18	66	16	63	18
0.3～ 1.5%	12	9	16	12	32	15	16	13	15	12
1.5～ 3.0%	5	13	3	13	4	13	2	11	4	12
over 3.0%	1	19	1	11	2	10	1	12	1	11

(5) 결 론

다른 몇 개의 추정량들에 대한 비교를 통해서 우리는 비록 $\hat{P}_{3q}$의

정도(precision)가 $\hat{P}_{1q}$에 비해 좋기는 하지만 이는 지방의 수준에서 실제적으로 사용하기에 충분하지 않다. 이것은 연구된 모든 모형에 대해서 마찬가지이다. 이러한 관점에서 볼 때 비수정 합성 추정량 $\hat{P}_{2q}$만이 합리적인 선택이라고 할 수 있다. 그러나 $\hat{P}_{2q}$는 모형이 잘 맞지 않는 경우 편향 추정량이 되고 다른 모형들에 대해서 이 편향의 크기에 대한 정보를 알아내기가 쉽지 않다. 지방 수준의 사회계획에 있어서 $\hat{P}_{2q}$의 추정값은 현존하는 대안에 관련하여 평가되어야 한다. 다시 말하면, 모형 C에 근거한, 미지의 모형 편향을 가질 가능성이 있는, 합성추정량에 의해서 구한 AKU에 따른 실업율의 추정값을 사용하는 것이 더 나은가? 또는 그 지방에 대한 AMS 통계량을 사용해야 하는가? 특별히 고안된 평가연구에 의해 P(AMS1)를 추정할 때, 다른 지방들에서 $\hat{P}_{2q}$에 대한 모형 편향에 관한 정보를 얻어내는 것이 가능해졌다.

8.2 주와 소지역 단위의 취업과 취업률 추정량에서 회귀 기법 적용(Use of Regression Techniques for Developing State and Area Employment and Unemployment Estimates)

(1) 방법론적 목적

이 연구에서 통계적 모형을 세우는 주된 목적은 (1)CPS와 비교하여 최소의 연간 오차를 갖는 주(state) 수준의 통계량을 만드는 것이고, 둘째로 주 내부의 횡단면 기대확률분포에 대응하는 지역수준의 통계량을 만

드는 것이며, 셋째로 지역의 경향(trend), 순환(cycle), 계절효과(seasonal movement)의 기댓값에 대응하는 추정량을 생산하기 위한 경제 메커니즘에 숨어있는 구조를 찾아내는 것이다.

(2) 통계적 체계의 제한사항(Constraint on Statistical system)

주와 지역에 대한 BLS 노동력 추정량을 추정하기 위해 사용되고 있는 통계시스템은 다음과 같은 제약사항을 가지고 있다. (1) CPS는 비교를 위한 통계적 기준이다. 왜냐하면, 바람직한 통계적 성질을 가지고 있고, 취업률과 실업률에 대한 공식적인 개념과 정의를 측정하기 때문이다. (2) 어떠한 주나 지역에 대한 공식적인 추정량은 위에서 설명한 BLS신뢰도 기준을 만족시키는 (월별 또는 연간) 추정량을 만들 수 있을 만큼 표본의 크기가 충분할 때, CPS로부터 직접 유도할 수 있다. (3) CPS 추정량은 11개의 큰 주와 2개의 대도시 지역의 월 통계량과 나머지 39개 주와 콜롬비아 지역의 연평균 통계량을 기초로 하여 사용된다. (4) 나머지 39개 주와 콜롬비아 지역을 앞으로는 "nondirect-use states"라고 하고 세 번째에 언급한 지역의 월 통계량은 같은 주에 대해서는 CPS 연평균추정량으로 사용할 것이다. (5) 모든 nondirect-use substate 지역의 추정량들을 합산하여 공식적인 주의 추정값이 되도록 기계적으로 계산해야 한다. 국가전체에 대한 추정량을 만들기 위해 주의 추정량을 합산할 필요는 없다. 주와 지역에 대한 월 노동력을 산출하기 위해 ESA(Employment Security Agencies)에 의해 현재 사용되고 있는 방법인 BLS의 일반적인 절차는

BLS의 월별취업과 소득보고서의 기술적 노트에서 제공한다.

(3) 연구 전략(Research Strategy)

우리가 지금까지 수행해 온 방법론적인 접근은 대부분 연구범위 안에서만 유용한 데이터의 질만 요구해 왔었다. 물론 어떤 지역에 대해 프로그램하기 적합한 데이터는 충분히 큰 표본으로부터 구할 수 있을 것이다. 왜냐하면 이러한 일련의 과정은 불편성, 최소분산, 엄격한 통계검정에 대한 민감도와 같은 바람직한 통계적 성질을 가지고 있기 때문이다. 그리고 널리 인정되는 정의와 개념을 가지고 취업률과 실업률을 공식적으로 추정한다. 결론적으로 만약에 우리가 비용에 제한이 없다면 모든 주와 지역에 대한 CPS 데이터를 사용할 수 있을 것이다. 하지만, 한정된 예산에서는 11개의 주와 2개의 큰 지역에 대해서만 CPS 데이터를 직접 사용할 수 있다. 따라서, 11개의 주와 2개의 큰 지역 이외의 지역에 대해서는 이러한 통계적 성질을 가지며 CPS 데이터를 최대한 사용할 수 있도록 CPS의 표본데이터와 UI시스템(Unemployment Insurance System : 실직 보험 시스템)으로부터 계정데이터(accounting data)를 결합하였다.
이러한 방법을 사용하면 우리의 관심은 월별 CPS 데이터를 이용할 수 있는 지역들에 집중된다. 현재 우리는 모든 주와 200개의 SMSA(Standard Metropolitan Statistical Areas : 표준 도시 통계 지역)에 대한 데이터를 가지고 있다. 이러한 지역에 대해서는 단일 회귀방정식(single-equation regression techniques)을 사용하여 만족할 만한 결과를 얻어왔다.. 이러한

회귀방법은 첫째로, 바람직한 통계적 성질을 갖도록 하는 절차를 사용하고, 둘째로 CPS데이터를 최대한 사용해야 한다는 우리들의 두 가지 목적에 부합되는 것이다.

(4) 회귀 모형(Regression MOdel Specification)

우리가 개발하고 테스트해 온 방정식은 다변량 선형 통계 모형에 기초하고 있다.

$$CPS = B_0 + B_1X_1 + \cdots + B_kX_k + e$$

CPS : 종속변수로서 CPS 데이터로부터의 취업률과 실직률에 대한 월별 통계량

$X_1 \sim X_k$: 독립변수, UI 시스템으로부터 나온 값, CPS에서의 표본 추정값, 독립적인 인구추정값

$B_0 \sim B_k$: 독립변수의 계수

e : 모형의 오차항, 여기에서는 독립변수, 종속변수에서의 측정오차와 모든 결측 독립변수들의 결합효과라고 할 수 있다.

Note : 본 논문에서 대문자는 모수 또는 확률변수를 뜻하고 소문자는 모수의 추정값을 나타낸다.

전통적인 OLS(Ordinary Least Square) 방법을 사용하여 계수를 추정하기 위해서는 (1) 등분산성 즉, $Var(e_i) = \sigma^2$, (2) 공분산=0 , $Cov(e_i, e_j) = 0$ 이라는 가정이 필요하게 된다. 하지만 모형의 종속변수는 사회경제적 시계열 표본 통계량들이기 문에 가정 (1)과 (2)에 위배된다는 것을 알 수 있

다. 하지만, 이론적으로는 변수변환 과정을 거쳐 오차항들이 가정을 충족시키도록 한다면 OLS를 사용할 수 도 있다. 이러한 두 단계 과정을 WLS(Weighted Least Square)가로 부르며, 이러한 가중값은 자기상관곗와 CPS, SMSA 표본분산으로부터 추정된다. 이러한 가중치 자체가 추정되게 되면 실질적인 추정과정은 EGLS(Estimated Generalized Least Square)라고 부르기도 한다.

(5) 주 실업통계 모형(State Unemployment Models)

이 연구의 주된 목적은 주에 대한 데이터를 기초로 하여 각 주에 해당되는 취업률과 실업률 예측을 위한 모형을 찾아내는 것이다. 이러한 세분화된 모형은 Handbook 방법보다 지역 노동시장의 상황을 더 잘 반영할 수 있다는 이점을 가지고 있다. 또한 "nondirect-use states"에 대해 횡단면 시계열 데이터에 기초한 모형보다 더 쉽게 이해할 수도 있고, 모든 사람들이 받아들일 수 있는 모형을 만들 수 있다.

5.1 실업률(Unemployment Rate)

전체 실업은 경험실업과 신규실업으로 구분할 수 있다. 전자의 사람들은 현재의 실업 이전에는 일자리가 있었던 사람들이고, 후자의 사람들은 그렇지 않은 사람들이다. 따라서, 전자의 사람들은 실업에 대한 보상을 받을 자격이 있지만 반면에 후자의 사람들은 그렇지 못하다. 이것은 매우 중요한 사실이다. 왜냐하면, 예측변수에 대한 데이터의 주요원천은 그 주의 UI시스템의 결과로부터 나온 행정용 데이터이기 때문이다. 따라

서 앞의 4개 범주를 반영하는 행동모형을 지정하는 것은 현실적이지 못하다. 왜냐하면 우리의 주 데이터 출처는 기껏해야 두 개의 변수에 기초한 아주 희박한 정보만을 포함하기 때문이다. 따라서 실업모형에 대한 항등식은 다음과 같이 정의할 수 있다.

$$CPSU = EXP + ENT$$

CPSU : 전체실업, EXP : 경험실업, ENT : 신규실업

이 분류법은 근로자들은 논리적 경제적으로 구분하는 것이며 더 중요한 것은 주 수준에서 알고 있는 데이터의 제한에 대한 문제에 주의를 기울이고 있다는 것이다. 모형에 있는 변수들은 두 개의 그룹을 대표할 수 있도록 선택된다. (하지만 주의 표본자료가 이러한 목적에 대해서는 신뢰할 수 없기 때문에 각각의 모형은 경험실업과 신규실업으로 발전되지 않는다) 주의 실업모형에 있어 수준 대신에 비율을 종속변수로 선택하였다. 그 이유로는 (1) 경제 변수에서 비율이 더 중요하다; (2) 비율은 매우 상호작용이 활발한 변수이다. (3) 비율을 사용하는 경우 수준을 사용하는 것보다 회귀식을 합하는 것이 더 용이하다. 종속변수의 특수성 때문에 독립변수 또한 비율로 지정할 수 있다. 첫 번째 설명변수는 주의 보험급여를 받는 실직자들의 비율이다 즉 직업을 잃었거나 결과적으로 보험급여를 받는 경험 노동자들에 대한 변수이다. 확실히 이러한 변수들은 전체 총실업과 연관되어 있고 경제순환동안의 총실업 변동의 많은 부분을 설명할 수 있다. 하지만 이러한 것이 대부분 주에 해당하는 것은 아니다. 우리는 추가적인 기대 변동부분을 설명할 수 있는 자료로부터 UI통계량을 확인하는데 실패하였다. 그러나 많은 대부분의 주에 있어 주기적으로

민감한 산업 고용 자료를 찾아낼 수 있었다. 그 결과 취업률 변수가 모형에 추가되었다. 동시에 이러한 변수들은 전체실업률에서 유효한 부분의 변동을 설명하였다. 다음 단계로 신규실업에 의해 발생되는 비율에 있어 대부분의 변동을 설명해 줄 수 있는 통계량을 지정하는 것이다. 이렇게 함으로써 예측오차를 감소시킬 수 있다.

UI시스템은 다음과 같은 조건을 필요로 한다. 노동자가 벌어들이는 금액은 최소한의 금액을 충족시켜야 하고 기간은 얼마간의 기간을 만족해야 한다.(UI자료의 장점과 단점 등과 같은 더 자세한 내용은 Blaustein(1979)을 참고하기 바란다) 정의에 의하면 신규실업은 전임의 취업경력이 없거나 노동을 하고자 하는 의욕이 없는 사람을 말하며 따라서 실직보험을 받을 자격이 없다. 그 결과 신규실업 변수가 전체에서 차지하는 변동을 설명할 수 있는 UI시스템의 지역자료를 찾을 수 없다. 이러한 신규실업들에 대해 설명할 수 있는 계절 가변수의 사용은 대부분 주에서는 성공적이지 못하였는데 그 이유는 아마도 종속변수의 표본오차 때문이었다고 생각된다.

CPS 자료로부터 나라 전체와 지역에 대한 신규실업률을 조사하는 것은 신규실업의 계절패턴이 횡단면적으로 매우 유사하다는 것을 알 수 있다. 따라서 우리는 모든 주에 있어서 신규실업요인에 의해 발생되는 변동을 설명할 수 있는 예측변수로서 나라 전체의 CPS 신규실업률을 이용하였다.

앞의 표기 방식을 따르면 시간 t에 대한 주의 실업률 회귀모형은 일반적으로 다음과 같이 주어진다.

$$CPSR_t = b_0 + b_1 CLMS_t + b_2 EMP_t + b_3 T_t + \sum_{j=1}^{12} c_j SEAS_t + e_t \ , \quad t = 1, \cdots, n$$

n : 표본의 월 관측치의 개수

$CPSR$: CPS 월별 주 실업률

$CLMS$: 주 실업보험의 혜택을 받는 비율(CES)

$$= \frac{\text{실업보험신청자수}}{\text{월급근로자수(비농업분야)}}$$

EMP : 주 CES 산업 취업에 기초한 비율(ratio) 또는 지수(index)

T : 선형추세변수 : $t=1, n;\ T_1=1,\ T_n=n$

SEAS : 계절 가변수, 0또는 1의 값을 갖는 이진형 변수 D와 USENT(국가전체의 월 신규실업율) 의 곱. 예를 들어 만약에 1월이면 $D_1 = 1,\ D_2 = D_3 = \cdots = D_{12} = 0$, 따라서 $D_1 USENT_t = USENT_t$가 된다.

$b_0 \sim b_3$, c_j : 모형의 계수

e : 모형의 오차항

신규실업 변수는 예측계열에서의 계절성을 강조하기 위해 단일 시계열 변수 대신에 12개의 계절 가변수로 지정되었다.(지역 CPS 신규실업 비율을 이용한 추가적인 테스트는 아직 완성되지 않았다) 독립변수 CLMS는 비록 많은 주의 모형에서 큰 설명력을 보이고 있지는 않지만, 모든 주의 모형에 포함되어 있다. 만약에 추세변수(T)가 없다면 모든 주에 대해 같은 모형 변수를 가지게 된다. 그러나, 주에 의해 취업변수를 지정하게 되면 모형은 주에 따라 다양해지게 된다. 선형추세변수에 대해 두 가지의 설명이 가능해진다. (1) 다른 예측변수에 의해 포착되지 못한 단기간의 짧은 기간동안 실업률에 있어 꾸준한 증가가 있었다. (2) 같은

기간동안 발생한 보험실업률과 CPS 취업률 사이의 뚜렷하게 진행되고 있는 수렴은 추세변수에 의해 포착되었다.

5.2 예측변수의 측정오차(Measurement Error in Predictor Variables)

독립변수로서 표본데이터를 사용하는 것은 변수문제에 있어 오차를 발생시킨다. 이러한 문제점과 해결방안을 Judge et al.(1980)과 Rubinfeld(1976)가 제안하였다. 우리의 모형에 있어 계수의 추정은 편향되어 있고 불일치성을 가지고 있다. 하지만 만약에 이러한 변수들을 제거하게 되면 좀 더 심각한 문제에 직면하게 된다. 계수의 값들은 우리의 목적에 맞는 본질적인 경제적 의미가 없다. 그리고, 더 중요한 것은 본질적으로 측정오차를 가지고 있는 변수를 포함하는 검정법은 모형의 예측정확도를 향상시킨다. Wonnacott(1970)는 만약에 모수의 추정에 관심이 있다면 이때 가장 중요한 것은 변수의 오차라고 주장하였다. 하지만 단지 예측에 관심이 있다면 그 때는 OLS추정량이 가장 최적의 예측치가 될 것이라고 주장하였다. 위에서 언급한 것처럼 이 모형의 가장 큰 목적은 최소의 MSE를 갖는 추정량을 예측하는 것이다. 따라서 이 연구에서는 예측변수와 오차항은 독립이라는 가정 하에서 변수의 오차는 고려하지 않았다.

(6) 지역 모형(Area Models)

앞에서 언급한 것처럼 지역에 대한 노동력 추정에 사용할 수 있는

모형은 각 지역에 대한 신뢰할만한 데이터의 유용함수이다. 프로그램목적으로 미국은 대부분 단일 카운티로 구성된 약 2500개의 상호배반적인 LMS(Labor Market Areas ; 노동시장지역)로 나누었다. LMA를 인구가 더 조밀하게 나눈 지역이 SMSA이다. 모형이 시간이 지남에 따라 진부해지는 것을 막기 위해 지역 방정식의 검정은 SMSA의 부분, 일반적으로 CPS 자료가 유용한 큰 200개의 SMSA(이하 "CPS LMA")으로 제한하였다. 더 인구가 조밀한 LMS는 이 작업에서 주요 LMA로 간주될 것이다.

이론적으로 "nondirect-use state" 각각에 대해 회귀식이 존재하는 것처럼 추정을 위한 이러한 LMA 각각에 대한 분리된 회귀식을 가지고자 한다. 하지만 대부분의 소지역에 대한 CPS 표본 추정량은 존재하지 않고 소지역에 대해 유용한 CPS 자료는 신뢰성에 의심이 간다. 따라서 이러한 LMA에 대한 각각의 회귀식 접근 방법은 불가능하게 된다. 다행히도 이러한 것을 대신할 수 있는 각각의 소지역에 대한 적당한 추정량을 추정할 수 있는 많은 통계적 방법들이 존재한다.

6.1 실업자(Unemployment)

Substate에 대한 실업률 추정에 대한 방정식은 BLS의 계약하에 1980년에 수행된 MIPR(Mathematical Policy Research)의 연구에 의존하게 된다. Ericksen(1974)의 연구에 의하면, MPR은 추정 목적을 위한 공동의 추정치를 형성하기 위하여 횡단면적으로 방정식 변수에 월별 시계열 데이터를 결합하는 회귀 표본 방식을 이용하여 모형을 개발하였다. MIPR은 독립변수들에 대한 데이터와 CPS 지역데이터의 결합에 있어 각각의 지역에 대한 실업율을 구하기 위해 사용되는 단일 방정식의 계수를 추정하는데

이용된다. 하지만 각각의 지역에 대한 CPS 데이터가 표본으로 사용하기에는 아직 신뢰할 만한 수준이 되지는 못한다.

BLS 지역모형은 MPR 모형을 개량한 모형이다. 종속변수와 독립변수들은 서로 모수의 크기가 다른 지역의 데이터를 합하였을 때 발생하는 문제를 완화시키기 위해 비율로서 지정하였다. 모형에서 선택된 독립변수들은 추정에 사용될 지역 설명변수의 설명력을 약간 반영할 것이다. 어떤 지역에 대한 모형은 다음과 같이 정의될 것이다.

$$CPSR_i = B_{0i} + \sum_{1}^{k} B_k X_{ik} + \sum_{1}^{12} C_j Z_j + e_i$$

여기서

$i = 1, m$ (m = 지역의 개수)

$CPSR_i$: 공동데이터에서 i번 째 SMSA에 대한 CPS 전체 실업율

B_{0i} : 절편항, i번째 지역에 대해서는 1, 다른 지역에 대해서는 0의 값을 갖는 지시변수

B_k, C : 기울기

Z_j : 계절 가변수, $Z_j = 1 (j = 1, 6, 7)$, 0 그 이외의 경우

e_i : 오차항

모집단의 나이비율과 지역 CES 취업률을 포함한 많은 독립변수들이 테스트되어 졌다. 위에서 나온 모형들은 전반적으로 좋은 결과를 보여주었다. 그럼에도 불구하고 절편항은 모형의 독립변수에 의해 설명되어지지 않는 지역과 지역사이의 구조차이를 찾아내기 위하여 공동의 200개 SMSA의 모든 지역을 걸쳐 이동하는 것을 허용하였다. 위에서 설명하였

듯이 이 모형은 공분산 모형이다. 더 자세히 말하자면 지역과 지역사이의 차이를 설명하기 위하여 $(m-1)$개의 가변수를 사용하는 취소제곱모형이다. 많은 대안의 방법들을 시도해 보았지만 향상된 결과를 보여주지는 못했다. CPS, SMSA자료를 사용하기 위해서는 이분산성의 가정을 갖는 WLS방법을 사용해야 한다. 잔차들 사이의 자기상관을 발견하기 위해 Goldberger(1962)의 BLUP(Best Linear Unbiased Predictor)의 변형이 추정회귀모형에 결합되었다. 일반적으로 t시간에서 i번째 지역에 대한 모형은 다음과 같이 정의된다.

$$\widehat{CPSR}_{it} = b_{0i} + b_1 CLMS_{it} + b_2 EMPOP_{it} + b_3 STUR_t + \sum_{1}^{12} C_j Z_{jt} + \hat{p}_i \widehat{e}_{it}$$

여기서

$$\widehat{e}_t = \sum_{k=0}^{3} \hat{e}_{t-k}/4$$

이고, $\hat{p}_i$ 는 i번째 지역에대한 추정 자기상관계수

추정시계열요소 pe는 예측계열에서 랜덤오차를 줄이기 위해 그리고 잔차가 서로 자기상관되어 있는 경우에는 예측오차를 줄이기 위해 4개월의 평균에 기초한다. 우리는 변수선택문제에 있어 오차를 가지게 된다. 왜냐하면 측정오차를 가지고 있는 주의 변수를 사용하기 때문이다. 앞서 언급한 주 모형의 이러한 상쇄반응은 지역 모형에서 이러한 변수들을 선택하게 한다.

(7) 모형의 검정(Test of Model Objectives)

7.1 연간 검정(Annual Test)

각각의 모형에 대한 성능을 테스트하는 주요 측도로 추정오차를 사용한다. Drapper & Smith(1981)는 변수선택을 위한 표준 경영경제 분석 절차를 따를 때, 처음에 결정해야 할 것은 표집기간 동안의 추정오차를 최소화시키는 모형을 선택하는 것이다. 왜냐하면 우리가 지정할 수 있는 최고모형보다 CPS의 체계적 부분에 가까운 추정량을 만들어 내야하기 때문이다. 하지만 특히 변수들간에 서로 상관되어 있고 구조가 연속된 추정기간 안에서 통계적으로 유의하게 이동하는 경우에는 그러한 모형을 가지는 최소의 오차를 갖는 추정량을 표집기간 밖에서는 만들어 내지는 못한다. 우리가 테스트하고자 하는 표집기간은 1976-1983년 사이의 주 자료와 1978-1983년 동안의 LMAs 자료이다. 이미 알고 있듯이 이 기간 동안에는 계속되는 경기의 후퇴가 있었다. 게다가 표본설계가 많은 주에 있어 바뀌었고 두 개의 주요 예측변수에 있어 큰 변동이 있었다. 이렇게 역동적이며 계획되지 않은 자료에 대해 모형을 적합시키는 경우에는 잘못된 모형이 나오게 되거나 새로운 데이터가 표본에 추가될 때 연속된 기간 동안에 계수들이 통계적으로 유의할 정도로 바뀔 수 있다. 우리는 이러한 최적주의원리에 빠지지 않도록 주의해야 할 것이다. 최적주의 원리란 Picard와 Cook(1944)가 제안한 것으로 MSE와 같이 표본의 검정통계량에 기초하여 모형의 예측력을 평가하는 방법이다. 따라서 모형선택을 위한 표본 내의 검정도 고려하지 않을 뿐 아니라 평가의 주요수준도 고려하지 않는다.

위에서 언급한 것처럼 우리의 목적은 CPS와 비교하여 최소의 오차를

갖는 표본추정기간의 추정값을 만들어내는 것이다. CPS 연평균 추정량의 39개 "nondirect -use state" 와 콜롬비아 지역의 공식통계량이기 때문에 각각의 모형에 대해 결정적인 테스트는 CPS와 비교해 연 추정오차에 집중하게 된다. 따라서 우리가 따르게 되는 결정규clr은 최소의 연추정오차나 RMS(Relative Mean Square)를 가지는 추정량을 갖는 모형을 선택하는 것이다. 이러한 테스트는 각각의 모형에 대한 예측력을 평가하는 검정력이 강한 검정방법이다.

앞서 언급했듯이, 미국 국내의 경기는 1976년 이후에 2개의 경제순환을 갖는다. 단일 표본외 검정에 기초하여 더 나은 모형을 선택하기 위해서는 이러한 연속되는 기간동안에 다른 대안의 방법은 더 나은 결과를 낳지 못하였다. 그 결과 우리는 1980-1983년 사이의 데이터를 이용해 각각의 모형을 테스트해보았다. 1980년에 대한 예측은 1976-1979년의 자료에 기초하였고 1981년에 대한 예측은 1976-1980년의 데이터를 사용하였다. 이 연구에서 언급된 모형은 전체 1980-1983년 동안의 최저평균 RMS 오차를 갖는 추정량을 갖는 모형이다.

7.2 월별 검정(Monthly Test)

우리는 지역의 계절 패턴과 경제순환을 반영하는 추정량을 만들 수 있는지 결정하기 위해 앞의 순서화 작업으로부터 선택된 모형을 테스트해 보았다. 이러한 검정방법은 매우 어려운 점이 많다. 왜냐하면, 예측계열에 있어 이러한 효과를 직접적으로 테스트할 수 있는 지역데이터가 존재하지 않기 때문이다. 주와 SMSA CPS 월간 자료는 이러한 목적을 사용되지만 월별 추정량의 큰 표본오차 때문에 이러한 자료의 사용은 검

정력이 약해지게 된다. 다른 서로 상관된 지역 자료는 간접적으로 이러한 패턴을 가진 일반적인 대응관계를 발견하기 위해 사용되지만 이러한 것이 통계적 검정을 수행하기 위한 것은 아니다. 따라서, 예측계열 자체에 대해서만 테스트할 수 있고 통계적으로 받아들일 수 있는 성질을 가지고 있는지에 대해서 생각해 볼 수 있다. 이러한 검정을 위해서 전통적인 시계열분석방법을 생각할 수 있다. 이론적으로 시계열은 다음과 같은 3개의 구성요소를 가지고 있다. (1)추세-순환 (2) 계절효과 (3) 불규칙변동. 다른 조건들이 같다면 올바르게 지정된 모형은 지역의 계절패턴과 경제순환 이동에 대응하는 추정량을 만들 수 있다. 모형이 지역데이터에 기초하고 있다는 점에서 독립변수, 종속변수는 이러한 패턴과 높은 상관관계를 가질 가능성이 많다. 그러나 모형의 몇몇 변수들은 측정오차 특히, 종속변수의 표본오차를 포함하고 있다. 그 결과 모형의 예측계열은 우리가 측정하고자 하는 규칙적 패턴을 발견하지 못하게 하는 많은 랜덤오차를 포함하게 된다. 이 시점에서 우리는 우리의 모형을 세우는데 있어서의 방법론적 목적을 다시 언급할 필요가 있다. 비경제적 움직임을 최소로 하는 시계열을 만들어야 할 뿐만 아니라 경제메커니즘에 숨어있는 움직임까지 잡아내야 한다. 물론 우리는 시스템에서 임의 충격을 기대할 수도 있지만 이러한 것은 불규칙적을 일어나야만 한다. 하나의 검정방법은 예측계열에서 백색잡음효과나 랜덤오차를 측정하는 것이다. 이 검정을 위해 사용할 수 있는 통계량은 월별 변화의 퍼센트의 표준오차와 Census X-11에서 계절효과를 조정하기 위한 프로그램으로부터 계열의 불규칙요소의 상대부분과 하나의 시간간격에 대한 계열의 자기상관함수 등

이다. 이러한 통계량 등을 이용하여 모형의 성능을 평가하는 것은 적절한 것이지만 백문이 불여일견이란 속담처럼 직접모형에 적용하여 보는 것이 더 좋은 판단방법인 것 같다.

7.3 지역 모형(Area Models)

지역 모형에 대한 검정은 별로 정교하지 못하다. 왜냐하면 특별히 지정한 SMSA에 대한 CPS 데이터의 신뢰성이 일반적으로 그 area에 대한 모형의 성능을 평가하기에는 부족하기 때문이다. 따라서 공동의 모든 지역에 대해 모형의 성능을 평가해 줄 수 있는 지시자로서 결합가중평균, MSE, 누적빈도분포 등을 사용하였다. 1980년 Census 데이터를 이용해 주 내부의 기대분포를 검정하는 것은 그리 간단한 일은 아니다. 많은 사람들이 검정결과의 해석을 어렵게 하는 표본오차와 비표본오차의 차이에 대해서 잘 알고 있지 못한. 따라서 이 글에서 이러한 검정력과에 대해 평가하고자 한다.

7.4 모형의 안정성 검정(Test for Model Stability)

우리는 우리의 자료가 기간이 짧고, 그 기간동안에 불규칙한 변동을 보이기 때문에 잘못된 모형이나 모형의 계수가 불안정한 결과를 보이게 된다. 이러한 모형에 대한 안정성평가는 모형평가라는 의미를 갖게 된다. Snee(1977)는 모형의 예측력 평가와 계수의 안정도 평가의 한 방법으로 교차타당도방법에 대해 논의한바 있다. 이 방법은 전체 데이터를 알고리즘에 의해 추정 자료와 예측자료라는 2개의 자료로 분할하는 방법이다. 추정 자료는 모형을 추정하기 위해 쓰이는 자료이다. 이러한 표본외 검

정은 Geisser(1975)에 의해 제안된 자료분할방법이 축약된 형태이다. N개의 관측치를 갖는 데이터셋을 모형추정을 위한 (N-n)개의 자료와 n개의 모형검정을 위한 자료로 분할하는 방법이다. 우리의 모형에서 자료를 분할하기에 적당한 변수는 시간변수이다. 왜냐하면 우리의 모형은 시간에 의존하는 자료를 이용하기 때문이다. 앞에서 언급한 것처럼 1년에 해당하는 n=12의 연속 관측치를 검정기간으로 설정하였다.

시간의 흐름에 따라 모형이 진부해지는 것을 막기 위해 검정은 가능한 시간관련부문에 제한하였다. 결과적으로 우리는 N월별 데이터 포인트의 모든 (k-1)개의 부집단을 이용하여 최적의 모형을 추정하였다. 우리는 k=N/12, k번째 연도에 대한 월별 추정량을 예측할 것이다. 이러한 모형의 성능은 전체 RMS오차나 계수의 안정도 평가에 기초하여 판단하게 될 것이다. 이렇게 제한된 검정에 의해 4년 동안의 데이터를 이용해 조사해 본 결과 최적의 모형이 항상 각각의 테스트년도에 있어 최적의 모형이 되는 것이 아니라는 것을 알았다. 이것은 모형이 종속변수의 정보나 변동을 모두 잡아내지 못했다는 것을 의미한다. 우리는 새로운 자료가 유용하다면 모형 평가를 계속할 것을 기대한다.

(8) 예측모형에서 랜덤오차에 대한 조정 (Adjustment to Reduce Random Error in Predicted Series)

우리는 이 연구의 초기에 다른 것들이 모두 같다는 조건 아래 많은 주에 대한 CPS 종속변수의 표본오차는 경제분석에 적합한 계열을 만들어

내는 모형을 찾아내기 어렵다는 것을 가정하였다. 게다가 초기 시계열 검정은 연추정오차를 줄이기 위해 BLUP 예측모형을 사용하는 것은 월 추정에 있어 랜덤오차를 증가시킨다는 것을 보여주었다. 왜냐하면 자기 회귀오차는 CPS에서 표본오차를 반영하기 때문이다. 그 결과 우리는 예측 모형에서 랜덤오차를 줄이기 위한 다양한 방법을 생각해 보았다. 첫 번째 방법은 이동평균방법을 이용하여 BLUP 예측모형에서 시계열요소를 평활하는 방법이었고 두 번째 방법은 모형의 구조요인을 이용하거나 비중심화 MARC(Moving Average Ratio Correction)를 이용하여 유도된 월별 예측 추정량을 조정하는 것이다. 두 번째 방법은 Handbook 방법에서 체계적 오차에 대한 수정에 있어 현재 사용되고 있는 방법으로 표본외의 연 예측요차를 줄이는 효과를 보여준다. 시간이 지남에 따라 모형이 진부해지는 것을 방지하기 위해 실업률에 대해서는 6개항 MARC를 취업률에 대해서는 4개항 MARC를 이용하는 방법이 최적의 결과를 보여주었다. MARC와 관련되어 테스트한 예측모형은 자기 회귀항을 포함하지 않는다.

8.3 실업보험 자료와 LFS 자료를 이용한 소지역 추정: 변형된 Fay-Herriot 방법의 적용

(Combining Unemployment Benefits Data and LFS Data to Estimate ILO Unemployment for Small Areas: an Application of a Modified Fay-Herriot Method)

(1) 배 경

정부 공식통계는 경제 정책, 재원 분배 및 정책 결정 등에 참고 자료로 이용된다. 대 영역에 대해서는 공식 통계의 정보가 이용자들에게 제공되나 소지역에 대해서는 그렇지 못한 실정이다. 최근 영국 내에서는 소지역 통계 작성에 대한 요구가 꾸준히 제기되고 있고 특히 노동시장 동향에 대한 측도 개발이 시급히 요구되고 있다. 노동력 조사(LFS)는 노동시장 정보 파악에 중요한 역할을 담당하고 있으나 직접조사 추정값들은 소지역 추정에는 한계를 안고 있다. 이 논문은 LFS로부터 소지역 추정의 신뢰도를 향상시킬 수 있는 모형 기반 추정량들을 소개한다.

영국의 LFS는 세 달을 주기로 연속 조사가 실시된다. LFS는 약 60,000 조사가구 단위를 갖는 대규모 조사로써 16세 이상의 약 150,000명의 인구에 대해 조사가 이루어진다. 영국의 LFS는 국제노동기구(ILO)의 요구조건을 만족하도록 표본설계되어 있으며 이를 통해 실업통계가 작성된다. LFS 표본설계에서 표본은 단순임의추출로 추출되며 주로 국가 수준의 추정값을 생산하도록 설계되어 있다. 일년에 한번 소지역 단위인 UA지역(Unitary Authority)과 LAD지역(Local Authority)에대한 추정값들이 작성된다. 이 연구에서 소개되는 내용은 현재 영국 통계국(ONS)에서 진행하고 있는 소지역 추정법과 밀접한 관계가 있다.

소지역 추정법은 소지역에 대한 직접 조사 추정값들이 신뢰성에 문제가 있거나계산될 수 없을 때 이용할 수 있는 통계적인 기법으로써 인근 지역의 보조정보를 빌려 소지역의 특성값을 추정하는 간접 추계 방법이다. 이 연구에서 이용한 주요 보조정보는 실업보험을 청구한 사람들의 수

이다. 실업보험 자료는 행정 시스템에 의해 획득되기 때문에 표본오차가 없고 지역적 범주로 또는 성별-연령대별 범주들로 다양하게 분류될 수 있다. 실업보험 지급 청구자 수와 ILO 실업자 수와는 시기에 따라 약간의 차이는 있지만 강한 상관성을 나타낸다. 시기에 따라 발생하는 차이는 주로 행정 시스템의 변경 또는 경제 사이클의 변화 등에 기인한다. ONS는 Southampton 대학과 연계하여 LFS자료와 실업보험 청구자 수의 자료를 결합하여 소지역 추정값의 신뢰성을 확보할 수 있는 연구를 진행하고 있으며, 특히 UA 또는 LAD 지역에 대한 실업자 수를 추정하는 SPREE 방법(Purcell and Kish, 1980), 로지스틱 모형에 근거한 일종의 변형된 Fay-Herriot 방법(1979)과 Multi-level 모형화 방법(Goldstein, 1995)과 같은 세 가지 추정방법에 대해 연구를 진행하고 있다. SPREE 방법보다는 로지스틱 모형에 근거한 변형된 Fay-Herriot 방법과 Multi-level 모형화 방법이 더 좋은 효율을 나타내며, 여기에서는 변형된 Fay-Herriot 방법에 초점을 맞추어 소개한다.

(2) 소지역 추정 방법

앞으로 소개되는 수식에서 첨자 i와 j는 각각 UA 지역과 LAD 지역에 대한 성별-연령대별 그룹을 나타내며, 첨자 g와 h는 각각 UA 지역과 LAD 지역들을 나타낸다. 표본 자료는 LFS 추정값들과 각 지역 내에서 성별-연령대별 그룹으로 분류된 각 셀들에 대한 실업보험 청구자 수의 자료들로 이루어져 있다.

N_{ig}를 셀 (i, g)에서의 인구 총계, U_{ig}를 같은 셀에서의 실업자 총계라 할 때, 이 셀에서의 실업률은 $Z_{ig} = U_{ig}/N_{ig}$이다. 일반적으로 실업률 Z_{ig}는 g번째 지역의 특성값들에 의해 결정된다. Z_{ig}의 기대값과 분산을 각각 $E(z_{ig}) =$ $_{ig}$, $Var(z_{ig}) =$ $_{ig}(1$ $_{ig})/N_{ig}$라 하자. g번째 지역의 특성값들이 $E(Z_{ig})$의 값에 미치는 영향을 열거하기 위해 로지스틱 모형이 이용되었다. 이용된 로지스틱 모형은 $\text{logit}(\pi_{ig}) =$ $\boldsymbol{x}_{ig}^{T}\boldsymbol{\beta}$ 이다. 여기에서 벡터 $\boldsymbol{x}_{ig}$는 소지역 g에서 i번째 성별-연령대별 그룹에 대한 속성들을 나타내며 알고있는 값이다.

N^{*}_{ig}를 셀 (i, g)에서의 인구 총계에 대한 LFS 추정값이라 하고, U^{*}_{ig}를 실업자 수에 대한 추정값이라 할 때, LFS 실업률 추정값은 $Z^{*}_{ig} = U^{*}_{ig}/N^{*}_{ig}$로 나타낼 수 있다. 성별-연령대별 그룹들이 합리적으로 정의되어 그룹 내에서 추출된 조사단위들에 대한 표본 가중치들에서 변동이 거의 발생하지 않는다고 가정할 수 있다면 셀 (i, g)에서의 실업률에 대한 LFS 추정값들은 표본 실업률로 근사될 수 있다. LFS 표본은 단순임의추출 표본이므로 다음 식들이 성립한다.

$$E(Z^{*}_{ig} \mid Z_{ig}) = Z_{ig}\ , \tag{1.A}$$

$$Var(Z^{*}_{ig} \mid Z_{ig}) = [\,(N_{ig} \quad n_{ig})/(N_{ig}-1)\,][\,Z_{ig}(1 \quad Z_{ig})/n_{ig}\,] \tag{1.B}$$

$$= Z_{ig}(1 \quad Z_{ig})/n^{*}_{ig}$$

여기에서 $n^{*}_{ig} = n_{ig}(N^{*}_{ig}-1)/(N^{*}_{ig}-n^{*}_{ig})$이고, n_{ig}는 셀 (i, g)의 LFS 표본크기를 나타낸다. 주어진 Z_{ig}에 대해 Z^{*}_{ig}와 $\boldsymbol{x}_{ig}$의 독립성을 가정한다면 위의 식들은 다음과 같이 주어질 수 있다.

$$E(Z^*_{ig} \mid Z_{ig}) = E[\, E(Z^*_{ig} \mid Z_{ig},\ \boldsymbol{x}_{ig}) \mid \boldsymbol{x}_{ig}\,] = E(Z_{ig} \mid \boldsymbol{x}_{ig}) = \quad _{ig}\ , \qquad (2.A)$$

$$Var(Z^*_{ig} \mid \boldsymbol{x}_{ig}) = E[\, Z_{ig}(1 \quad Z_{ig}) / n^*_{ig} \mid \boldsymbol{x}_{ig}\,] + Var(Z_{ig} \mid \boldsymbol{x}_{ig}) = \quad _{ig}(1 \quad _{ig}) / n^{**}_{ig} \qquad (2.B)$$

여기에서 $n^{**}_{ig} = n^{oo}_{ig}[1+(n^{oo}_{ig}-1)/N^*_{ig}]^{-1}$이다. 일반적으로 n_{ig}는 N_{ig}에 비해 상대적으로 작은 값을 갖기 때문에 식 (2.A)와 (2.B)는 Z^*_{ig}에 대한 일종의 근사 이항 로지스틱 모형을 정의하기 위하여 $_{ig}$에 대한 로지스틱 항과 결합될 수 있다. 이 모형은 표본크기 $n^o_{ig} = round(n^{**}_{ig})$와 표본 실업자 수 $m_{ig} = round(n^o_{ig} \times Z^*_{ig})$를 입력값으로 갖는 로지스틱 회귀 소프트웨어를 이용하여 실제 표본 자료에 적합될 수 있다. 여기에서 $\boldsymbol{\beta}$의 추정값과 $Var(\boldsymbol{\beta})$의 추정값 $v(\boldsymbol{\beta})$를 이끌어 낸다. 이때 Z_{ig}의 추정량은 $\pi_{ig} = \text{antilogit}(\ \boldsymbol{x}_{ig}{}^T \boldsymbol{\beta})$이 된다. 그러나 이 추정량은 불편성을 만족하지는 않는다. 따라서 편의를 보정한 형태의 추정량은 다음 (3)식과 같이 주어질 수 있다.

$$\pi_{ig} = \pi_{ig}\left[1 - \frac{1}{2}(1-\pi_{ig})(1-2\pi_{ig})(\ \boldsymbol{x}_{ig}{}^T v(\boldsymbol{\beta})\ \boldsymbol{x}_{ig})\right] \qquad (3)$$

이때 소지역 g에서 실업자 총계에 대한 추정량은 다음 (4)식과 같이 주어진다.

$$\theta_g = \sum_{i \in g} \alpha_{ig} N^*_{ig} \pi_{ig} \qquad (4)$$

소지역 g와 h에 대한 모형기반 추정량들 간의 추정 공분산은 다음 (5)식을 통해 계산될 수 있다.

$$c(\theta_g, \theta_h) = \sum_{i \in g} \sum_{j \in h} \alpha_{ig} N^*_{ig} \pi_{ig}(1-\pi_{ig})(\ \boldsymbol{x}_{ig}{}^T v(\boldsymbol{\beta})\ \boldsymbol{x}_{jh})\, \pi_{jh}(1-\pi_{jh}) N^*_{jh} \alpha_{jh} \qquad (5)$$

(3) LFS 자료를 이용한 적용결과

1995년~'96년과 1998년~'99년의 LFS 자료들을 이용하여 UA지역과 LAD지역에 대한 실업률 추정값들을 계산하였다. 주요 보조변수로써 실업보험 청구자 수를 이용하였다. 이러한 보조변수 및 성별-연령대별 범주의 6개 그룹, 지리적인 권역으로 분류된 12개 그룹과 사회-경제적 분류 집락인 7개 그룹에 대한 척도들이 모형에 포함되었다.

UA지역과 LAD지역에 대한 실업률 추정값은 기존의 직접추정방법보다는 2절에서 언급한 모형기반 추정방법이 상대적으로 변동이 작고 훨씬 안정적으로 나타났다. 모형기반 추정량의 추정오차에 대한 LFS 조사 추정량의 추정오차 비의 평균값은 1995년~'96년 자료에서는 5.5049, 1998년~'99년 자료에서는 5.3851 의 값을 나타내며, 모형기반 추정방법이 상대적으로 작은 변동을 나타낸다는 사실을 확인하였다.

(4) 추가 연구

2절에서 소개된 것과 같은 합성추정 형태의 모형기반 추정량은 소지역들 간의 변동을 설명할 수 없는 문제점을 안고 있으며, 일반적으로 이러한 방법으로 추정된 추정오차는 과소 추정되는 경향이 있다. 이러한 문제점은 소지역에 대한 랜덤효과를 모형에 반영한 Multi-level 모형을 통해 어느 정도 해소할 수 있으며 이러한 연구가 현재 영국 통계국에서 진행되고 있다. 대상 모형은 $\mathrm{logit}(\pi_{ig}) = \boldsymbol{x}_{ig}^{T}\boldsymbol{\beta} + u_g$와 같은 모형이다. 여기

에서 지역 명시 변수 $\{u_g\}$는 평균이 0이고 분산이 σ_u^2인 확률변수로 가정된다. EBLUP 형태의 성분 추정값 $\pi_{ig}=\text{antilogit}(\boldsymbol{x}_{ig}^T\boldsymbol{\beta}+u_g)$에 기반을 둔 (4)식의 소지역 추정값들은 표본 크기가 큰 UA 및 LAD 지역에서는 LFS 추정값들과 유사하며, 표본 크기가 작은 UA 및 LAD 지역에서는 고정효과를 같는 추정값들과 유사한 경향을 나타낸다. 이러한 추정량이 갖는 실제적인 문제는 추정량의 평균제곱오차(MSE) 계산이 쉽지만은 않다는 데에 있다. 현재 영국 통계국에서는 하나의 절충안으로써 다음과 같은 분산 추정공식을 고려하고 있다.

$$v(\theta_g)=\sum_{i\in g}\sum_{h\in g}\alpha_{ig}N_{ig}^{*}\pi_{ig}(1-\pi_{ig})\left[\sigma_u^2+\boldsymbol{x}_{ig}^T v(\boldsymbol{\beta})\,\boldsymbol{x}_{hg}\right]\pi_{hg}(1-\pi_{hg})N_{hg}^{*}\alpha_{hg}\ , \quad (6)$$

여기에서 σ_u^2은 랜덤효과 모형 적합에서 추정되는 소지역 간의 추정분산을 나타낸다.

(5) 결 론

LFS 직접 추정값들은 실업보험 청구자료와 같은 이용 가능한 보조정보를 통해 개선될 수 있다는 사실을 이 논문에서는 보여주고 있다. 이 논문에서 고려된 방법론은 추정값의 정확도를 개선시키기는 하나, 모형에 랜덤효과를 포함시키는 문제와 EBLUP 형태의 추정량들의 평균제곱오차를 추정하는 방법 및 비 추정 방법 등은 여전히 해결되어야 할 문제점으로 남게 된다. 이러한 문제를 해결하기 위한 연구가 현재 영국 통계국 및 Southampton 대학 연구진들에 의해 진행되고 있다.

9. 소지역 추정 적용방안

9.1 기본 개념

대규모 통계조사에서 국가단위와 도단위의 추정치를 구하기 위한 조사설계에서도 국가단위와 도단위의 영역만을 고려하는 통계조사는 거의 없다. 이와같은 통계조사에서도 다양한 교차분류영역과 세분된 영역에 대한 추정값을 계산할 필요가 있을 수 있다. 이와같은 대부분의 경우에는 보다 더 세분화된 영역(소지역)의 추정값의 요구되는 정확도의 수준에 대해서 조사설계 또는 추정과정에서 특별한 관심을 갖지 못하였다. 교차분류 영역이 희소 부차모집단이거나 지리적 소지역에 대한 추정값에 대해서 질적인 수준의 의문을 갖게되거나 아예 추정값을 계산할 수 없는 소지역으로 생각될 경우에는 문제가 심각해질 것이다. 만일에 소지역의 추정값을 해당 영역의 통계조사 자료를 이용하여 계산해야한다면 조사설계, 추정과정 및 조사과정을 포함한 전체적인 종합계획을 개발해야 할 것이다.

조사설계와 추정과정을 중점적으로 다루기 위해서 연구영역을 도개로 나누어 설명하겠다.

(1) 계획된 영역(Planned Domain)

표본조사설계에서 개별적인 층, 층의 그룹들을 조사설계시에 분석단위로 고려하여 목표수준을 참작한 표본들을 연구한 경우이다. 즉 카나다

의 경우에는 주의 하부영역인 경제영역(ER : Economic Region), 실직보험 영역(UIR : Unemployment Insurance Region)과 건강관리영역(HPR : Health planning Region)등은 조사설계연구시에 분석단위로 고려되고 있으며 이런 영역외에도 규모가 큰 카운티, 인구조사구역과 주내의 세분된 영역등이 여기에 해당된다.

(2) 비계획 영역(Unplanned Domain)

표본조사설계시에는 분석단위로 생각하지 못했으나 조사자료분석과정등에서 통계작성의 필요성이 대두된 소지역을 말하며 이런 영역들은 설계 층을 교차분류하거나 좀더 세분화된 주내부의 영역인 경우이다. 이런 경우에는 조사된 표본의 수가 적거나 아예 없을 수 있으므로 추정과정에서 신뢰성의 문제가 제기될 것이다.

다음에는 종합계획에서 연구되어야할 사항에 대해서 살펴보자.

① 설계계획

노동력조사와 같이 지속적이고 주기적인 조사로부터 자료의 요청이 많은 경우에는 매 5년마다 표본을 재설계하고 있다. 이런 주기적인 대규모 통계조사 설계에서는 과거 요구된 소지역 통계들을 참고로 앞으로 필요하게된 내용까지 고려하여 종합적인 표본설계를 연구하게 된다. 또한 특수목적의 통계조사에서는 조사목적에 합당한 사항들을 고려하고 이를 만족할 수 있는 분석단위를 연구해야할 것이다. 이와같은 두가지 경우에는 표본설계자는 국가단위와 주단위의 추정값 뿐만아나라 관심영역의 추

정값에 대한 요구된 정확도를 만족할 수있게 설계를 연구해야할 것이다. 종합계획의 첫단계에서는 소지역 자료의 공급측면에서 연구영역이 조사설계단계에서 사전에 분석단위로 인식되어지는 정도에 따라서 계획된 영역으로 취급되도록 하는 것이다. 만일에 비용의 제한으로 인하여 어떤 소지역에 대해서는 신뢰할 만한 추정값을 계산할 수없다면 다음 세가지 중에서 하나를 선택해야한다. 첫째는 영역을 통합하여 신뢰할 수 있는 추정값을 계산하는 것이고, 두 번째는 다른 통계조사자를 결합하여 추정값을 계산하는 방안을 연구하는 것이며, 세 번째는 의뢰기관과 협의하여 신뢰성에 심각한 문제가 있는 소지역 추정값은 공표하지 않도록 하는 것이다. 어떤 영역에 대해서는 사전에 인식되지 않을 수도 있다. 이러한 비계획영역에 대해서 별도의 특별한 추정법을 적용하여 추정값을 계산해야 할 것이다.

② 표본설계

실제로 표본설계에서 단일관심주제를 다루는 경우에도 국가단위 또는 주 단위의 수준에서도 최적이 되기 쉽지 않다. 통상적으로 이론적으로나 조사실행적인 면에서 제한사항을 만족시키 위해서 표본추출과정과 자료수집과정에서 다양한 타협적인 방법들이 적용된다. 자료의 요구수준에 따라서 연구영역의 추정값을 계산하는 과정에서 타협적인 방법을 심도있게 다루어야 할 것이다. 표본설계단게에서 소지역 자료의 요구를 반영할 수 있는 두가지 방안인 표본배분과 표본집락정도를 설명하겠다.

③ 표본배분전략(Allocation Strategy)

일반적으로 국가단위의 추정값에 대한 최적표본배분법은 주 단위에 대해서는 주별 모집단의 크기에 비례하도록 표본을 배분하는 것이다. 이 경우에 크기가 작은 주에서는 추정값의 신뢰도에 문제가 있을 수있으므로 절충적인 표본배분법이 적용되고 있다. 교차분류등 세분된 국가단위의 추정값에 대해서는 중요시하는 정도에 따라서 상이한 절충배분법을 적용할 수 있을 것이다 상대적으로 큰 주에 대해서는 표본의 수가 줄어드는 것에 대한 추정값의신뢰도의 영향이 적을 것이지만 규모가 적은 주에서는 표본수가 약간만 증가해도 추정값과 해당 주의 자료에 대한 신뢰도와 영향이 상당히 클 것이다.

주 단위내의 영역에 대해서도 위에서 언급한 표본수의 증감에 따른 영향의 형상은 같을 것이다.대부분의 경우에 최적배분이 단조롭기 때문에 설계하는 사람들은 규모가 큰 지역에서 규모가 작은 지역으로 표본을 재배분하는 특성을 이용할 수 있을 것이다.

④ 집락화(Clustering)

대규모의 가구조사에서는 일반적으로 국가단위와 주 단위의 xhdrpfidddd에 대한 비용적인 효율성을 제고하기 위해서 1차 추출단위들 상대적으로 크게 한다. 이런 설계에서는 집락을 크게 함에 따라서 어떤 영역에는 많은 표본이 배정되고 반면에 어떤 영역에는 표본이 하나도 배분되지않을 가능성이 있으므로 비계획된 영역에 대한 통계 생산에 결함이 생길 수 있다.중요하게 생각되는 영역의 추정값의 신뢰도를 고려한다

면 표본의 집락화를 최소화해야할 것이다.집락화에서 중요하게 다루어져야할 요인들은 추출틀의 선택,추출단위의 선정과 크기,층의 크기와 개수와 추출단계등이며 이런 목표를 달성하기 위해서 조사실행의 제한조건을 가능한 최소로 해야한다.

⑤ 추정법(Estimation)

특정한 통계조사에서 계획과 표본설계단계에서 영역의 추정값에 아무리 많은 주의를 기울렸더라도 소규모 영역에서 적절한 신뢰도를 갖춘 추정값을 계산하기 위해서는 특수한 추정법이 필요하게 될 것이다. 최근에는 유사영역의 자료 또는 정보를 관심영역의 것과 결합하는 합성추정법(Synthetic Estimator)이 많은 관심을 끌고 있다. 합성추정법은 관심영역과 주변 유사영역의 특성이 같다는 전제조건에서 유용하지만 만일에 전제조건에서 조금만 이탈되더라도 설계편향이 심각하므로 이들의 이용에 많은 문제점이 있다. 확률적인 표본학자들은 합성추정법의 특성을 살리면서 설계편향을 조절할 수 있는 직접추정량과 합성추정량을 결합한 복합추정량을 제안하고 잇다. 경험적 베이즈 추정법과 다른 기법들에서도 복합추정량에서 각 성분별 가중값을 조정하는 방안이 연구되어 왔다.

9.2 표본설계시 고려할 사항

일반적으로 소지역 추정문제는 추정법에서 다루어져야할 내용으로 생각하고 있지만 앞에서도 언급했듯이 표본설계 시에 고려할 수 있다면 많은 문제들을 완화시킬 수 있을 것이다. 카나다 노동력조사의 경우를 사례

로 하여 설명하겠다. 2장에서 층화와 추출단위 구성에 대한 개념적인 설명은 하였지만 여기서는 소지역 추정법 적용 측면에서 다시 언급하겠다.

카나다 노동력 조사 설계는 각 단계별로 다양한 추출방법을 적용하여 최종추출단위인 가구를 59,000가구 추출하여 매월 조사하고 있으며 표본가구로 한번 선정되면 6개월간 조사되고 다른가구로 교체되는 체계이다. 또한 1차 추출단위와 집락도 일정한 주기로 연동교체하고 있다. 카나다의 10개주에 대해서 각 주별로 경제구역으로 구분하였으며 경제구역은 대표영역과 비대표영역으로 나누고 있다 대표영역은 중규모또는 대규모의 도시로 구성되었고 비대표영역은 나머지 경제구역으로 구성되었다.

층화와 표본선정은 각 영역내에서 이루이지고 두 영역별로표본추출단계수와 추출단위가 다르게 적용되는 복잡한 표본조사이다. 예를 들면 도시지역에서는 2단계 층화 추출법을 적용하지만 도시외의 지역에서는 3단계 층화추출법을 적용하고 있다.

(1) 추출단계와 추출단위

지역추출틀은 일반적으로 집락화된 추출법과 연계되어있다. 에를 들면 1차 추출단위는 여러개의 2차 추출단위들을 포함하고 있다. 만일에 2차 추출단위의 리스트를 이용할 수있다면 주어진 리스트에서 표본을 직접 추출하게 된다면 표본의 집락의 정도를 약화시킬 수 있을 것이다. 이런 경우에는 효과를 높임으로써 추정값의 신뢰도를 제고할 뿐만아니라 비계획된 영역의 소지역 추정값의 정도를 크게할 수 있을 것이다. 후자의 경우에는 표본들이 골고루 분포되게 하여 비계획된 영역에서도 표본이 선

정될 가능성이 크게 될 것이다. 반대로 집락화된 표본설계에서는 어떤 여역에서는 충분한 표본이 선정되어 추정값의 신뢰성을 크게하지만 유사한 다른 영역에서는 아주 적은 수의 표본이 선정되거나 아예 표본이 없는 경우에는 좋은 추정값을 계산할 수 없다.

노동력조사에서 집락을 축소하기 위한 두가지 선택사항을 살펴보자. 첫째는 대규모의 도시에서 지역추출틀을 주소지를 리스트로한 리스트 틀로 교체할 수 있는 가능성의 검토이고, 둘째는 농촌지역과 소규모 도시지역에서 표본추출단계을 줄인 것이다. 1991년 카나다 센서스에서 포함범위를 개선하기 위해서 고안한 주소록에는 센서스조사구내에 거주하는 사람들의 주소,전화번호와 지리적인 정보가 포함되어 있다. 한가지 선택사항은 주소록에서 주거지를 층화추출법으로 선정하는것이며, 이표본은 사후 센서스의 주소록에 포함되지않은 주거지에 대한 보완적인 절충을 위해서 주어진것이며 추가주거지를 보완하는데 장애가되는 요소를 보정하기위한 연구들이 계속되고 있으며 이와같은 정보들은 차후의 센서스와 통계조사를 위해서 분석되어야 할 것이다.

도시이외의 지역에서 추출단위를 변경하고 표본추출단계를 줄이는 선택사항은 보다 더 좋은 소지역 추정값을 얻기 위해서 집락을 축소하는 방안을 유지하는 것이다. 면대면 면접조사에서 전화와 컴퓨터보조면접조사로 자료수집기법의 변경으로 과거조사와 비교한 비용-분산의 분석은 더 이상 적절하지 않다. 지금은 노동력조사의80%정도는 전화면접으로 이루어지고 있으며, 전화면접조사의 증가로 여행비용이 경감될 뿐만아니라 1차 추출단위를 생략하고 직접인구조사구를 추출할 수 있게될 것이다.

(2) 층 화

1차 추출단위에 대한 언급에서와 마찬가지로 대규모 층을 소규모 층으로 교체하는 방안이 층화에 대한 연구 내용이다. 우리가 바라는 것은 재정의한 영역이나 비계획된 영역이 완벽한 층에 포함되는 것이고 이렇게 함으로써 영역 내에서 표본의 크기가 보다 더 안정될 것이다.

요구된 추정값을 구하는데는 겹쳐진 지역의 영역이 있을 수 있다. 예를 들면 카나다의 각 주는 경제영역과 실직보험영역으로 분할하였다. 층으로 분할된 것들의 교차에 의해서 생성되는 영역들을 취급할 때 지리적으로 영역이 겹친 영역이 생길 것이다. 예로써 71개 경제영역과 61개 실직보험영역에서 133개 교차영역이 생기지만 이는 다룰 수 있는 개수이나 어떤 경우에는 교차영역의 개수가 너무 많아서 효과적으로 다루기가 힘든 경우도 있다. 여기서는 몇 개의 교차영역들은 소규모영역이므로 쓸모없는 층이 될 것이다.

소규모의 층들을 갖는 집락을 줄이는 결합방안이 소지역 추정의 요구조건을 충족시킬 수 있는 보다 더 좋은 조사설계가 완성되기를 바란다.

(3) 표본 배정

표본설계이전에 소지역의 정의들이 주어졌다면 조사설계과정에서 고려하여 이들을 계획된 영역으로 다룰 수 있게 할 수 있다. 표본설계자는 소지역에 충분한 표본을 배정하여 각 소지역에서 신뢰할 만한 추정값을 계산할 수 있게 할 수 있을 것이다. 노동력조사와 같은 대규모조사에서

위와 같은 접근은 최소한 이론적으로는 수많은 소지역 추정값들의 계산이 가능하게 할 것이다. 매월 59,000가구를 조사한다고 가정하면 하나의 지역에서 분기별로 신뢰할만한 추정값을 생산하기 위해서는 매월 100가구를 조사하면 충분할 것이다. 신뢰할 만한 통계를 생산하기에 충분한 표본의 배분을 생각했을 때 카나다 전체를 600개의 겹쳐지지 않은 지역으로 분할할 수 있으므로 이들 지역의 합집합들은 매월 신뢰할 만한 통계를 생산하기에 충분한 표본을 가질 것이다.

여러 가지 다양한 표본배분전략을 생각할 수 있을 것이며, 위에서 아래로 하향식 접근에서는 먼저 주별로 표본크기를 결정하고 세부영역별로 표본을 배정하는 방식이지만 주별로 주어진 표본크기로는 세부영역별 추정값을 신뢰할 만한 수준에서 계산이 가능하지 않을 수가 있다. 아래에서 위로 상향식 전략에서는 각 소영역의 통계를 목표정도를 충족시킬 수 있도로 먼저 소영역에서 표본크기를 결정하는 방식이며, 이때는 각 소영역에서는 어느 정도 충분한 표본이 배정될 수 있으나 주 단위에서 보면 위에서 아래로 하향식에서 결정된 표본의 크기 보다 표본수가 더많아 질수 있을 것이다. 두전략중 어느것을 이용할지라도 최초의 표본배분은 조정되어야할 것이다. 결과적으로 표본배분은 동일배분과 비례배분의 절충방식처럼 될 것이다. 실제로 표본설계자는 주 단위의 요구 신뢰정도와 몇몇 지역의 집합의 세부영역의 요구 신뢰 정도를 반복적으로 고려하면서 전체적인 비용과 조사 실행성을 기준으로 표본배분을 결정할 것이다.

노동력조사의 재 설계에서 적용한 기법은 다른 통계조사에서도 유용하게 사용할 수 있을 것이다. 표본배분은 두 단계로 실행되었다. 먼저

42,000가구의 핵심표본을 국가와 주단위 수준에서 신뢰할 만한 추정값을 생산하도록 배정하고, 나머지 17,000가구를 대부분의 세부 소영역에서 신뢰할 만한 추정값을 계산할 수 있도록 배정하고 나서 모든 계획된 영역에서 신뢰할만한 추정값을 생산할 수 있도록 배분을 조정하는 것이다. 이와 같은 절충방법은 주 단위에서 손실은 최소로하고 세부 소영역단위의 이득은 극대화하도록 적용된다. 예를 들면 Ontario와 Quebec에서 실업자 추정값에 대한 변동계수는 2.8과 2.6퍼센트에서 3.2와 3.0퍼센트로 변하고 카나다 실업자 추정값에 대한 변동계수도 1.36퍼센트에서 1.51퍼센트로 약간 커졌다. 그러나 주단위의 최적화 배분법을 적용한다면 실직보험영역에 대한 추정값의 변동계수가 17.7퍼센트까지 큰 경우가 있지만 절충식 배분법을 적용한 후에는 가장 나쁜 경우의 변동계수가 9.4퍼센트로 개선되었다.

표본의 재 배분은 한 영역에서 다른 영역으로 표본을 이동시키는 범위내에서 이루어질 것이다. 예를 들면 대규모 주에서 1,000가구의 표본을 줄여서 소규모 주로 재배분한다면 대규모 주의 추정값의 신뢰도는 허용할 만한 수준에서 낮아지지만 소규모 주에서 추정값의 신뢰도는 괄목할 만한 크기로 개선될 것이다. 이와 같은 표본의 재 배분은 동일한 주 내부에서 소영역간에 이루어질 것이다.

(4) 기타 고려사항

표본설계자는 설계된 표본이 사용중인 기간에 계획된 영역의 정의가 바꾸어져서 비 계획된 영역으로 새롭게 정의되는 상황이 주어진다는 것

을 고려해야할 것이다. 예를 들면 실직보험영역은 1995년 표본설계 후 2년 내지 3년이 지난 후에 정의가 변경되었다. 이런 상황을 표본설계과정에서 고려하기 위해서는 인구 조사구와 같이 안정적인 표준화된 영역들의 집합으로 추출단위를 구성한 후에 소지역의 정의가 변경되면 새로운 영역을 표준화된 소영역의 합집합으로 재정의 함으로써 일관성 있는 소지역 추정법을 적용할 수 있다. 노동력 조사에서도 이와 같은 개념을 적용하였다.

새로운 내용으로 갱신전략도 한가지 대안적인 방법이 될 수 있다. 최초에 추출했던 표본과 새로 추출해야할 표본의 중복이 최대가 되도록 표본을 선정하는 것도 좋은 대안이 될 수 있다. 이와같은 전략을 적용한다면 새로운 표본으로 작성되어야할 목록이 최소화되기 때문에 새로운 조사원의 고용최소화 함으로써 조사현장에서 혼란을 줄일 수 있는 장점이 있다.

10. 결 언

정보화 사회에서 신속 정확한 통계의 필요성이 강조되고 있을 뿐만 아니라 우리나라에서는 1997년 IMF라는 환란의 위기를 맞이하여 이를 극복하는 과정에서 모든 중장기 정부정책의 입안 시에는 객관적이고 신뢰성있는 통계를 기초로 해야한다는 자명한 진리를 큰 대가를 치르고 알게 되었다. 특히 IMF사태를 극복하면서 구조 조정이라는 개혁조치가 강행되면서 젊은이들의 최대 관심은 실업이라는 통계가 되었다. 국가단위의 실업자수와 실업률도 중요하지만 거주지역의 실업자수와 실업률에 대한 통계의 생산도 요청되어 왔으나 현재의 경제활동 조사의 체계 하에서는 시도단위의 고용관련 통계조차도 신뢰성에 대해서 논란이 많은데 시군구의 고용관련통계의 작성은 엄두조차 내기 어려운 상황이다. 그러나 2000년 인구주택총조사를 분석한 후에는 새로운 경제활동인구조사의 표본을 설계해야하므로 이 시기에 맞추어 시군구 단위의 고용관련통계를 작성하는데 기초가 될 소지역 추정법을 연구하고 이의 실용화를 위해서 최근 몇 년간 기초적인 이론연구와 방법론적인 타당성에 대한 사례연구들이 진행되었다. 하지만 미국이나 카나다에서는 고용통계 뿐만 아니라 다양한 분야에서 카운티 또는 더 세부적인 영역에 대한 통계작성을 위해서 소지역 추정법을 10년 이상을 연구하여 오고 있다. 우리는 시간과 노력을 좀 더 효율적으로 이용하기 위해서 카나다의 실업률에 관련된 소지역 통계 작성 현황을 살펴보고 우리나라의 여건에 적용할 수 있는 내용을 정리하고 이론적으로나 방법론적으로 개선 발전시켜야할 내용에 대해서는 심층

적인 연구를 목적으로 카나다 노동력 통계조사의 지침서를 중심으로 우리나라에서 경제활동인구조사의 자료를 이용하여 시군구 실업통계 작성 연구에 도움이 될만한 내용들을 번역하여 편집하였다. 먼저 캐나다 노동력 조사 배경과 목적을 요약하여 실업통계 관련 상황을 설명하였으며 다음에는 표본설계내용과 조사구 관리 방법을 살펴보았다. 특히 소지역을 작성 단위별로 추정값의 신뢰도 관리를 위한 표본배분과정을 심도있게 살펴보았다. 표본 조사구의 순환에 따른 추정량의 계산 방법과 가중치를 계산하는 절차를 살펴보았고, 조사된 자료의 품질관리는 비표본오차를 줄이는데 있어서 핵심적인 사항이므로 결측값의 보완대책과 무응답률의 관리 및 보완방안 등을 요약하였다. 소지역의 노동력 통계작성에 이용되는 추정법을 설계기반 추정량, 간접 추정량과 모형기반 추정량으로 대별하여 요약 정리하였으며 캐나다에서는 설계기반 추정량의 신뢰도에 문제가 없을 정도로 조사된 표본이 충분한 경우에는 우선적으로 설계기반 추정량을 적용하는 것을 원칙으로 하고 있다. 설계기반 추정량으로는 사후층화 추정량, 비 추정량과 회귀 추정량을 언급하였고, 간접추정량으로는 합성 추정량, 복합추정량과 표본 수 의존 복합추정량을 설명하였으나 캐나다에서는 표본 수 의존 복합 추정량이 폭넓게 적용되고 있지만 이론적으로 논란이 되고 있다. 모형기반 추정량은 EBLUP(Empirical Best Linear Unbiased Prediction), 경험적 베이즈 추정량과 계층적 베이즈 추정량이 언급되었다. 여러 가지 추정량들이 언급되었으나 이들의 효율성 비교에 대한 상대효율도 요약하였다. 현재 카나다 노동력의 소지역 통계작성에 적용되고 있는 표본 수 의존 복합추정량의 절차를 자세하게 설명하였고,

노동력 통계작성에서 적용될 수 있는 추정량의 분산추정에 대한 별도의 연구논문을 번역하여 포함하였다. 다음에는 캐나다 외의 나라에서 소지역의 실업통계작성을 위해서 연구된 논문을 요약하였다. 마지막으로 우리나라에서 경제활동인구조사에서 시군구의 실업통계를 작성할 수 있도록 표본설계를 개편할 때 참고가 될만한 내용을 소지역 추정법 적용방안이라는 제목으로 자세하게 진행과정별로 설명하였다.

본 번역은 통계청의 표본설계와 소지역 추정법에 관심을 가지고 있는 사람들을 위해서 쉽게 풀이 할려고 노력했으나 부족한 점이 있을 수 있다고 생각된다. 이런 보완해야할 사항은 앞으로 연구를 진행하면서 수정 보완하여 지방화시대에 요구되는 소지역 단위의 통계를 경제적이고 효율적으로 작성하는 기틀을 마련하도록 하겠다.

부 록

부 록

표A1. 1998 LFS 표본의 주별 세분화

Provinces	Strata			Households			Source		
	Urban	Rural	Apart.	Urban	Rural	Apart.	Total	Core	EI
Newfoundland	35	13	0	1151	836	0	1987	1987	0
Prince Edward Island	17	12	0	603	818	0	1421	1421	0
Nova Scotia	46	27	2	1753	1714	90	3557	2611	946
New Brunswick	45	20	0	1800	1261	0	3061	2604	457
Quebec	209	38	12	7693	2383	283	10358	5458	4900
Ontario	298	40	34	11717	2158	1634	15508	7179	8329
Manitoba	59	14	6	2481	1152	273	3906	3255	651
Saskatchewan	56	21	1	2471	1537	63	4072	3408	664
Alberta	75	13	6	2903	914	174	3991	3991	0
British Columbia	114	12	7	3882	854	234	4113	4113	856
Canada	954	210	68	36454	13627	2751	36027	36027	16803

표A2. 1998 LFS 표본의 EIER별 세분화

EIER	Strata			Households			Source		
	Urban	Rural	Apart.	Urban	Rural	Apart.	Total	Core	EI
020	15	0	0	465	0	0	465	465	0
021	20	13	0	686	836	0	1522	1522	0
123	17	12	0	603	818	0	1421	1421	0
224	8	2	0	294	145	0	439	397	42
225	7	6	0	287	342	0	629	555	74
226	22	0	2	716	0	90	806	806	0
227	4	7	0	235	466	0	701	490	211
228	5	12	0	221	761	0	982	363	619
329	32	4	0	1270	225	0	1495	1090	405
330	13	16	0	530	1036	0	1566	1514	52
433	8	9	0	323	541	0	864	455	409
434	17	0	2	730	0	45	775	510	265
435	15	0	0	566	0	0	566	0	566
436	3	3	0	238	368	0	606	224	382
437	13	0	0	722	0	0	722	153	569
438	13	5	0	590	325	0	915	401	514
439	80	0	9	1971	0	202	2173	2131	42
440	20	15	0	667	670	0	1336	732	604

EIER	Strata			Households			Source		
	Urban	Rural	Apart.	Urban	Rural	Apart.	Total	Core	EI
441	6	3	0	305	300	0	605	295	310
442	9	3	0	507	179	0	686	316	370
443	12	0	1	534	0	36	570	169	401
444	13	0	0	540	0	0	540	72	468
546	18	0	5	684	0	191	875	571	304
547	16	9	0	553	527	0	1080	387	693
548	29	14	0	901	561	0	1462	657	805
549	18	0	2	780	0	94	874	132	742
550	66	0	16	1971	0	827	2798	2798	0
551	14	0	4	604	0	133	737	448	289
552	17	0	1	760	0	56	816	258	558
553	15	0	2	641	0	157	798	282	516
554	7	4	0	290	185	0	475	136	339
555	11	0	2	558	0	81	638	167	471
556	17	0	2	720	0	95	815	238	577
557	5	2	0	353	227	0	580	262	318
558	20	8	0	843	529	0	1372	439	933
559	17	0	0	760	0	0	760	73	687
560	15	0	0	712	0	0	712	52	660
561	13	3	0	587	129	0	716	279	437
664	45	0	6	1622	0	273	1895	1895	0
665	10	11	0	522	742	0	1264	1139	125
666	4	3	0	337	410	0	747	221	526
767	18	0	0	767	0	0	767	562	205
768	16	0	1	709	0	63	773	623	150
769	14	17	0	581	1023	0	1604	1604	0
770	8	4	0	414	514	0	928	619	309
871	25	0	3	861	0	85	946	946	0
872	29	0	3	970	0	89	1059	1059	0
873	21	13	0	1072	914	0	1986	1986	0
975	12	4	0	530	218	0	748	748	0
976	60	0	6	1573	0	196	1769	1769	0
977	13	0	1	612	0	38	649	351	298
978	21	4	0	628	356	0	984	627	357
979	8	4	0	539	280	0	819	618	201
Canada	954	210	68	36454	13627	2751	52830	36027	16803

표A3. LFS 개편에서 고 소득층에 대한 층화결과

CMA	Dwellings	Strata	Clusters	Median Income	Average Income
Montreal	15,237	3	83	$121,881	$132,818
Ottawa	6,558	2	39	$111,729	$116,973
Toronto	35,433	4	185	$144,387	$156,477
Hamilton	6,584	1	34	$101,875	$107,130
London	4,036	1	21	$108,009	$108,604
Winnipeg	7,543	2	42	$96,763	$100,264
Calgary	7,501	1	41	$123,066	$131,543
Edmonton	5,835	1	28	$111,334	$118,600
Vancouver	16,483	3	89	$119,777	$122,739
Total	105,210	18	562	$122,765	$132,217

표A4. 주별 표본 배분과 CV 비교

Province	Old Sample				Redesigned Sample				Current Sample			
	Core		Total		Core		Total		Core		Total	
	Size	CV	Size	CV	Size	CV	Size	CV	Size	CV	Size	CV
Newfoundland	2,240	5.4	2,582	5.0	2,240	5.1	2,240	5.1	1,884	5.8	1,884	5.8
Prince Edward Island	1,421	6.6	1,421	6.6	1,421	6.1	1,421	6.1	1,421	6.1	1,421	6.1
Nova Scotia	3,101	5.2	4,002	4.8	3,102	4.9	4,050	4.5	2,609	5.5	3,557	5.1
New Brunswick	3,095	5.3	3,441	5.0	3,096	5.2	3,480	5.0	2,604	5.7	2,988	5.3
Quebec	6,474	4.0	11,356	3.4	6,436	3.7	11,590	3.1	5,413	4.2	10,567	3.5
Ontario	8,517	3.8	17,388	3.2	8,473	3.4	17,206	2.8	7,125	3.7	15,858	2.9
Manitoba	3,276	6.3	3,897	6.1	3,869	5.0	4,428	4.8	3,254	5.6	3,813	5.3
Saskatchewan	4,527	5.0	4,563	5.0	4,053	5.0	4,107	5.0	3,409	5.6	3,463	5.6
Alberta	5,205	4.6	5,225	4.6	4,745	4.4	4,745	4.4	3,991	4.9	3,991	4.9
British Columbia	4,454	5.7	4,975	5.1	4,875	4.6	5,583	4.4	4,100	5.2	4,808	4.8
Canada	42,310	1.96	58,850	1.67	42,310	1.73	58,850	1.50	35,810	1.96	52,350	1.62

참 고 문 헌

참 고 문 헌

[1] Brisebois, F. and Mantel, H.(1996). Month-in-sample effects for the Canadian Labour Force Survey. SSC Annual Meeting, June 1996, *Proceedings of the Survey Methods Section.*

[2] Brodeur, M., Montigny, G. and Berard, H.(1995). Challenges in developing the National Longitudinal Survey of Children. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

[3] Cochran, W.G.(1977). *Sampling Techniques*, 3rd Edition, John Wiley and Sons, New York.

[4] Chen, E.J., Gambino, J., Laniel, N. and Lindeyer, J.(1994). Design and estimation issues for income in the redesign of the Canadian Labour Force Survey. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

[5] Choudhry, G.H. and Rao, J.N.K.(1989). Small Area Estimation Using Models that Combine Time Series and Cross-Sectional Data. *Proceedings of the Statistics Canada Symposium on Analysis of Data in Time, October* 1989.

[6] Dufour, J., Simard, M., Allard, B. and Ray, G.(1996). Redesign of the Labour Force Survey Sample: impact on data quality. Statistics Canada, internal document.

[7] Drew, J.D., Belanger, Y. and Foy, P.(1985). Stratification of the Canadian Labour Force Survey. *Survey Methodology*, 11, 95-110.

[8] Drew, J.D., Singh, M.P., Choudhry, G.H.(1982). Evaluation of Small Area Estimation Techniques for the Canadian Labour Force Survey. Survey Methodology, 8, 17-47.

[9] Friedman, H.P. and Rubin, J.(1967). On some invariant criteria for grouping data. *Journal of the American Statistical Association*, 62, 1159-1178.

[10] Hartley, H.O. and Rao, J.N.K.(1962). Sampling with unequal probabilities and without replacement. *Annals of Mathematical Statistics*, 33, 350-374.

[11] Kennedy B., Drew J.D., and Lorenz P.(1994). The Impact of Nonresponse Adjustment on Rotation Group Bias in the Canadian Labour Force Survey. Presented at the 5th International Workshop on Household Survey Nonresponse. Ottawa, Canada.

[12] Lemaitre, G.E. and Dufour J.(1987). An Integrated Method for Weighting Persons and Families. *Survey Methodology*, 13, 199-297.

[13] Lorenz, P.(1995). Labour Force Survey-Head Office Hot deck Imputation System Specifications-Version 3. Statistics Canada, internal document.

[14] Mantel, H., Laniel, N., Duval, M.C. and Marion, J.(1994). Cost modelling of alternative sample designs for rural areas in the Canadian Labour Force Survey. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

[15] Mian, I.U.H. and Laniel, J.(1994). Sample allocation for the Canadian Labour Force Survey. *Proceedings of the Survey Research Methods Section*,

American Statistical Association.

[16] Rao, J.N.K., Hartley, H.O. and Cochran, W.G.(1962). A simple procedure for unequal probability sampling without replacement. *Journal of the Royal Statistical Society, B*, 24, 482-491.

[17] Sarndal, C.E., Swensson, B. and Wretman J.(1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.

[18] Sheridan, M., Drew, D. and Allard, B.(1996). Response rate and the Canadian Labour Force Survey: Luck or Good Planning? Proceedings of Statistics Canada Symposium 96 of Nonsampling Errors, 67-75.

[19] Simard, M. and Dufour, J.(1995). Impact of the introduction of Computer-Assisted Interviewing as the new Labour Force Survey data collection method. Statistics Canada, internal document.

[20] Singh, M.P., Drew, J.D., Gambino, J.G. and Mayda, F.(1990). *Methodology of the Canadian Labour Force Survey*, 1984--1990. Statistics Canada. Catalogue Number 71-526.

[21] Singh, M.P., Gambino, J. and Mantel, H.(1994). Issues and strategies for small area data (with discussion). *Survey Methodology*, 20, 3-22.

[22] Singh, A.C., Kennedy, B., Wu, S. and Brisebois, F.(1997). Composite estimation for the Canadian Labour Force Survey. *Proceedings of the Survey Research Methods Section,* American Statistical Association.

[23] Statistics Canada(1998). Guide to the Labour Force Survey. Available on the internet at www.statcan.ca/english/concepts/labour/index.htm

[24] Sunter, D., Kinack, M., Akyeampong, E. and Charette, D.(1995). Redesigning the Canadian Labour Force Survey Questionnaire: Development and Testing. Statistics Canada internal document.

[25] Tambay, J.L. and Catlin, G.(1995). Sample Design of the National Population Health Survey. *Health Reports*, 7, 29-38.

[26] Wolter K.M.(1985). *Introduction to Variance Estimation.* Springer-Verlag, New York.