

총조사 자료를 이용한  
소지역 추정량 비교

## 목 차

제1절 서 론.....	1
1. 연구배경.....	1
2. 연구내용 및 범위.....	4
제2절 경제활동인구조사의 표본설계와 추정.....	5
1. 표본설계.....	5
2. 추정량.....	7
제3절 소지역 추정방법.....	8
1. 직접추정량.....	8
2. 합성추정량.....	9
3. 복합추정량.....	11
4. EBLUP 추정량.....	12
5. HB 추정량.....	14
제4절 모의실험.....	16
1. 표본추출.....	16
2. 추정량 계산.....	18
3. 권역별 추정.....	20
4. 평가 방법.....	24

제5절 모의실험 결과.....	25
1. 추정량 비교.....	25
2. 모의실험을 통해 발견한 점.....	46
제6절 결 론.....	51
1. 요약.....	51
2. 논의 및 향후 연구.....	53

## 표 목 차

<표 1> 시도별, 동읍면부별 경제활동인구조사 표본조사구 배정 현황	7
<표 2> 성별, 연령대별 연평균 실업률	10
<표 3> 모의실험 추정량 비교(경기)	30
<표 4> 모의실험 추정량 비교(강원)	32
<표 5> 모의실험 추정량 비교(충북)	33
<표 6> 모의실험 추정량 비교(충남)	34
<표 7> 모의실험 추정량 비교(전북)	35
<표 8> 모의실험 추정량 비교(전남)	36
<표 9> 모의실험 추정량 비교(경북)	37
<표 10> 모의실험 추정량 비교(경남)	38
<표 11> 모의실험 추정량 비교(제주)	39
<표 12> 추정량 비교 결과 요약	39
<표 13> 실업자 수가 관측되지 않은 횟수 및 지역 분포	50

## 그림 목차

[그림 1] 모의실험 추정 및 추정량 비교 과정 .....	19
[그림 2] 태안군, 영양군의 조사구별 실업률 분포 .....	23
[그림 3] 익산시, 오산시의 조사구별 실업률 분포 .....	23
[그림 4] 모의실험 추정량 분포(조사구 수 15개 이상 지역) .....	42
[그림 5] 모의실험 추정량 분포(조사구 수 10~14개 지역) .....	43
[그림 6] 모의실험 추정량 분포(조사구 수 5~9개 지역) .....	44
[그림 7] 모의실험 추정량 분포(조사구 수 4개 이하 지역) .....	45
[그림 8] 주민등록인구 25만 이상 지역의 ARB와 RMSE .....	47
[그림 9] 주민등록인구가 25만 이하이면서 실업자 수가 0으로 자주 관측되는 지역의 ARB와 RMSE .....	48

# 총조사 자료를 이용한 소지역 추정량 비교

## 제1 절 서 론

### 1. 연구배경

소지역 통계는 한 국가의 지역별 사회·경제상황을 파악하고, 그 지역 발전을 위한 정책을 결정하는 기초 자료로써 활용가치가 높다. 우리나라의 경우, 2000년을 전후로 신뢰할 만한 소지역 통계에 대한 수요가 급증하였다. 그러나 문제는 공급이 수요를 충족시키지 못하는 데 있다. 특히 고용 관련 분야에 있어서는 인구주택총조사(인구총조사)의 고용부문 조사결과를 제외하고는 표본조사를 포함하여 소지역별로 제공되는 통계는 전무한 실정이다. 이는 대부분의 표본조사가 전국 또는 대단위 지역(시도)별로 결과를 공표할 수 있도록 표본 규모를 설계하기 때문이다. 따라서 소지역 추정에서는 설계당시에 계획에 없던 지역을 추정해야 하는 어려움이 있으며 가장 큰 문제는 해당 소지역에 정도를 높일 수 있을 만큼 충분한 표본 수가 할당되지 않는다는 것이다. 즉, 표본설계규모보다 작은 소지역은 표본 규모가 너무 작기 때문에 직접 추정치를 사용할 경우 그 만큼 결과에 대한 정도(precision)와 효율성 측면에서 신뢰성이 떨어진다고 할 수 있다(김서영과 권순필, 2009).

이러한 문제를 해결하기 위해 다른 나라들은 1990년대 전후로는 추정방법을 바꾸기보다는 표본을 늘리는 방법을 선호하였다(Fabbris et al.,

## 2 총조사 자료를 이용한 소지역 추정량 비교

1988). 그러나 조사환경이 점점 열악해지고, 예산 문제가 대두되면서 표본 증가에 대한 부정적 측면을 해결할 방법이 필요하게 되었다. 이 때문에 소지역에 대하여 표본규모를 크게 확대하지 않고 표집오차의 증가를 줄일 수 있는 통계적 추정 방법을 찾기 위한 연구가 시작되었고, 지금도 활발히 연구가 진행되고 있다(Falorsi and Russo, 1987; Rao, 2003 등).

소지역 추정량은 통계작성 목적과 자료의 활용여부에 따라 크게는 표본조사결과로부터 직접 추정하는 직접추정량(direct estimator), 직접 추정결과에 추가정보를 이용하는 합성추정량(synthetic estimator)과 복합추정량(composite estimator) 등의 자료 기반 추정량(data based estimator)과 관심변수와 상관성이 높은 설명 변수가 존재할 때 사용가능한 모형 기반추정량(model based estimator)으로 나눌 수 있다(김정숙 등, 2008). 모형 기반 추정량은 자료 기반 추정량에 비해 우수한 것으로 알려져 있으며, 세부적으로는 회귀분석방법, 경험적 최량선형불편추정량(Empirical Best Linear Unbiased Predictor: EBLUP), 경험적 베이저안 추정법(Empirical Bayesian method: EB), 계층적 베이저안 추정법(Hierarchical Bayesian method: HB) 등이 있다. 이 추정 방법들은 경우에 따라서는 서로 상이한 결과를 보일 수 있다. 모형 기반 추정량은 충분한 설명변수가 있어야 한다는 것이 기본 가정이며 이 가정이 만족되지 않을 경우 모형 기반 추정량의 사용은 제한적일 수밖에 없다. 따라서 모형 기반 추정방법은 적절한 보조정보를 찾는 것이 성공적인 모형 선택의 핵심 요인이며 이 추가 정보를 찾기 위한 연구가 선행되는 것이 중요하다.

이처럼 소지역 추정을 위한 다양한 방법이 있는 만큼 소지역 통계 작성을 위한 추정 방법의 선택은 신중하게 결정되어야 한다. 추정 방법 선택 시에 고려할 사항은 추정에 사용될 자료의 특성과 추가 정보의 양이 얼마나 풍부하고 신뢰할 만한 가이다. 앞에서 설명한 바와 같이 특히 모형 기반 추정방법은 추가 정보가 얼마나 관심 변수를 설명할 수 있는냐에 따라 추정치 결과가 다를 수 있기 때문에 더욱 신중할 필요가 있다. 자료 기반 추정량은 이용할 만한 추가 정보가 없는 경우에 주로 사용되는 방법으로 대지역의 인구 특성(성별, 연령별 인구) 정보를 활용하는 경우가 많다. 그런데 이는 소지역과 대지역의 관심변수와 관련한 특성이 유사하다는 것을 가정하며 이를 만족하지 못할 경우에는 크게 편

향될 우려가 있다.

우리나라의 경우, 소지역 추정에 대한 관심과 연구가 증가하는 것에 비해 소지역 추정 방법에 의해 작성된 통계를 공식통계(official statistics)로 받아들일 만큼의 인식이 높지 않다. 하나의 원인은 추정방법 또는 보조정보의 선택에 따라 추정 결과가 다를 수 있다는 점에 대해 우려가 있다는 것을 들 수 있다. 이런 점에서 볼 때 소지역 추정 결과에 대한 평가는 간단한 일이 아니다. 왜냐하면 많은 추정 방법들이 개발되어 사용되고 있지만 모수(참값)를 모른다는 점에서 추정량의 효율성 및 정도 등에 대한 비교가 수월하지 않기 때문이다. 뿐만 아니라, 지금까지 직접 추정량에 익숙해져 있던 사용자들이 통계적 추론에 의한 소지역 추정량을 이해하기 어렵고 이들 추정 통계에 대한 부정적 견해도 무시할 수 없는 부분이다. 이에 연구자들은 소지역 추정량의 정도나 효율성 비교를 위해서 다양한 모의실험을 시도하고 있다. 가장 대표적인 방법이 인구총조사 자료를 이용한 실험이다. 이 실험의 목적은 인구총조사 자료의 관심값을 참값으로 간주하여 표본으로부터 계산한 소지역 추정값과 참값을 비교함으로써 다양한 추정방법들을 평가하고자 하는 것이다. 실험 방법은 인구총조사 자료로부터 가상의 표본을 여러 개 만들고 이들 표본으로부터 계산된 추정값을 인구총조사 자료값과 비교하는 것이다 (Falorsi 등, 1994; Ugarte 등, 2007; Meindl, 2007 등).

국내의 소지역 실업자 추정에 관한 연구는 90년대 중반부터 시작되어 지금까지 활발하게 진행되고 있다. 이상은·진영(1996), 박종태·이상은(2001)과 통계청(2002, 2006)은 자료 기반 추정결과를 제시하였다. 2000년을 전후로 모형에 관심이 높아지면서 이상은(2003), 김재두 등(2005)과 통계청(2006) 등은 모형 기반 추정량에 대해 추정치를 비교하였다. 뿐만 아니라 모형 기반 추정량과 관련하여 이상은(2006), 이강석·신기일(2008), 김정숙 등(2008)은 공간통계량과 이웃정보시스템을 활용하여 이들이 추정치에 미치는 영향을 비교·분석하였다.

이 연구들은 실제 조사 자료를 이용한 연구로서 정확한 모수를 모르는 상태에서 제안한 추정량의 효율성, 정확성과 정도 등을 모수 측면에서 비교·분석하는 데에는 한계가 있었다. 이에 본 연구는 인구총조사 자료를 참값으로 간주하고 이들로부터 추출한 가상의 소지역 표본들의

#### 4 총조사 자료를 이용한 소지역 추정량 비교

관측값으로 소지역 추정을 시도하였다. 이로부터 소지역 추정량에 따른 추정결과와 참값과의 차이를 비교함으로써 어떤 추정량이 참값을 가장 정확하고 효율적으로 추정하는지를 평가할 수 있다. 결론적으로 인구총조사 모의실험을 통한 소지역 추정량 비교는 다양한 추정량의 추정능력을 검증하는 것으로서 소지역 추정방법의 이용 가능성을 실증하는 데 그 의미가 있다.

## 2. 연구내용 및 범위

본 연구의 주목적은 우리나라 경제활동인구조사에서 공표를 목적으로 표본이 설계되지 않은 지역, 즉, 소지역에 대한 추정치를 생산하기 위한 다양한 소지역 추정량을 평가하고자 하는 것이다. 우리나라 경제활동인구조사(경활조사)를 이용한 고용통계 소지역 추정방법을 선택하는데 있어서 중요한 근거로 활용될 수 있을 것이다.

소지역은 우리나라 행정구역상 230여 시군구로 정의하되, 본 연구의 실제 추정은 우리나라 150여개 시군 지역만을 대상으로 한다. 이는 우리나라의 표본설계 단위는 행정구역상 16개 시도 단위를 기본으로 하고 있고, 이때 시군구는 자체적인 추정을 목표로 표본이 설계된 대지역에 비해 보다 작은 부차 지역에 해당된다. 여기서 굳이 150여개 지역으로 한정된 이유는 2008년 10월에 진행된 통계청의 시군구 고용조사의 범위가 160여개의 특·광역시와 시군지역을 대상으로 하였고, 이는 향후 이루어질 이 시군구 고용조사 결과와 추정결과의 비교 작업을 염두에 둔 것이다.

자료는 2005년 인구총조사 자료를 이용한다. 보조정보로는 2005년 11월 고용보험 실업급여 수급자 수를 우선적으로 고려한다. 실업급여 수급자 수가 여의치 않을 경우 대지역의 실업자 정보를 보조정보로 사용할 수 있다. 실제 분석은 2005년도 인구총조사 자료 중 10% 표본조사 자료로부터 여러 번 반복 추출하여 가상의 표본을 생성하여 수행하였다.

추정량은 직접추정량과 자료 기반 추정량인 합성추정량, 복합추정량

3개와 모형 기반 추정량인 EBLUP, HB의 2개를 포함하여 총 5개를 고려한다. 추정량 계산에 필요한 표집오차는 잭나이프(jackknife)방법으로 구한다. HB추정량은 몬테 칼로 마코프 체인(Monte Carlo Markov Chain: MCMC)에 의해 표본을 추출하고 이들의 추정치를 구한다.

각 추정량의 평가는 인구총조사 결과를 모수로 간주하고 인구총조사에서 추출한 표본에 대해서 소지역의 실업자를 추정한 후 각 추정량에 대한 평균상대편향(Average Relative Bias: ARB)과 상대평균제곱오차(Relative Mean Squared Error: RMSE)를 사용한다.

본 연구는 크게 5개의 절로 구성되었다. 2절에서 경제활동인구조사의 표본설계 방법과 직접추정량에 대해서 설명한다. 3절에서는 4개의 추정량과 추정량의 평가 방법에 대해서 설명한다. 4절에서는 실제 모의실험 결과를 제시하고 각 추정량의 성능을 비교·분석한다. 마지막으로 소지역 추정 결과를 토대로 결론 및 논의사항을 제시한다.

## 제 2 절 경제활동인구조사의 표본설계와 추정

본 연구는 실업자 수 추정을 위한 다양한 추정량의 비교를 목표로 하고 있다. 우리나라 실업자 수는 경찰조사에 의해 작성되고 있으며 향후 소지역의 실업자 수 추정도 이 경찰조사를 주요 정보로 이용하게 될 것이다. 따라서 본 연구는 인구총조사 자료를 이용한 모의실험이기는 하지만 모의실험의 표본을 생성할 때 가능한 경찰조사의 표본추출 방법과 동일한 방법을 사용하고자 하였다. 이에 모의실험의 표본추출 과정을 이해하기 위해서 경찰조사의 표본설계와 추정 방법을 살펴보기로 한다.

### 1. 표본설계

경찰조사는 층화2단 집락추출 방법으로 표본을 추출한다. 모집단은 16개 시도와 동부·읍면부를 고려하여 25개 층으로 나뉘고, 일차표집단

## 6 총조사 자료를 이용한 소지역 추정량 비교

위(Primary Sampling Unit: PSU)는 조사구, 이차표집단위(Secondary Sampling Unit: SSU)는 조사구 내 가구이다. 최종 표본가구 수는 PSU인 조사구 수와 SSU인 조사구 내 가구 수의 곱으로 나타낸다. 산출순서는 우선 조사구당 조사구역 및 가구 수를 결정하고, 이를 포함할 수 있는 표본조사구 수를 산출한다. 결국 경찰조사는 전국적으로 총 1,629개 표본조사구에서 약 32,000가구를 조사 대상으로 한다(통계청, 2007).

표집 방법을 좀 더 자세히 살펴보면, 경찰조사는 인구총조사의 10% 표본조사 자료를 모집단으로 이용한다. 이때 실제 조사가 매우 곤란한 208개 섬조사구, 7,528개 기숙시설 및 특수사회시설 조사구는 제외한 26,505조사구를 모집단 조사구로 사용하고 있으며 모집단의 반영도를 제고하기 위해 2005년 인구총조사 실시 이후부터 표본 개편 시까지(2005.11.1. ~ 2006.10.31) 신축된 아파트 조사구 자료(506개)를 추가하여 27,011개의 조사구를 최종 모집단으로 구성하였다.

경찰조사 결과의 공표 단위는 7개 대도시(특별시와 광역시)와 9개 도이다. 각 도 안에서는 동부와 읍면부가 고용형태 및 소득·소비구조에 차이를 보이는 점을 감안하여 다시 동부와 읍면부로 층화하여 최종적으로는 전국을 25개 지역으로 층화하여 구분한다.

표본조사구 추출은 실업자 및 가구 소득과 상관관계가 높은 주거형태와 경찰조사에 있어 이용상 중요도가 가장 높은 통계항목인 산업구조 및 경찰 상태를 조합하여 각 지역별로 분류기준에 따라 배열된 조사구 명부를 이용한다. 그런 다음 조사구 크기(가구 수)에 비례한 확률비례표집방법(Probability Proportional to Size sampling: PPS)으로 조사구를 추출한다.

표본조사구 내에서는 한 조사구를 5가구씩 조사구역으로 분할한 후 각 거처의 구역화가 완료되면 표본으로 뽑힌 1개 구역과 인접구역 3개, 즉, 4구역 20가구를 표본으로 조사하게 된다. 시도별, 동·읍면부별 경제활동인구조사 표본조사구 배정 현황은 <표 1>과 같다.

〈표 1〉 시도별, 동읍면부별 경제활동인구조사 표본조사구 배정 현황  
(단위 : 개)

시도	계	동부	읍면부	시도	계	동부	읍면부
전국	1,629	1,242	387	경기	198	135	63
				강원	90	45	45
서울	207	207	-	충북	81	45	36
부산	117	117	-	충남	90	45	45
대구	90	90	-	전북	81	45	36
인천	108	108	-	전남	90	45	45
광주	81	81	-	경북	108	54	54
대전	81	81	-	경남	99	54	45
울산	63	63	-	제주	45	27	18

## 2. 추정량

경활조사의 각 특성(특히, 실업자)은 원칙적으로 표본추출 당시의 추출률에 따라 추정하면 된다. 이는 경활조사의 표본이 확률표본이기 때문이다. 그렇지만 이 조사는 일회성 조사(one-shot survey)가 아니고 동일한 특성에 대하여 매월 조사하는 지속 조사(successive survey)이므로 모수추정은 모집단의 보조정보인 층별 추계인구를 활용한 비추정방법(ratio estimation method)을 사용한다.

비추정에 의한 직접 추정량은 식(3.1)과 같은 형태를 갖는다. 다만, 지역은 경활조사 결과의 공표 단위인 7개 대도시(특별시와 광역시)와 9개 도에 대한 최종 25개 층화 지역, 즉,  $i=1, \dots, 25$ 를 말한다(통계청, 2007).

### 제 3 절 소지역 추정방법

#### 1. 직접추정량

본 연구에서 사용할 직접추정량은 경찰조사의 추정 방법을 벤치마킹한 것이다. 즉, 모집단의 보조정보인 층별 추계인구를 활용한 비추정량 (ratio estimation)을 직접추정량으로 사용하는데, 직접 추정량  $\widehat{Y}_i^{dir}$ 의 형태는 다음과 같다.

$$\widehat{Y}_i^{dir} = \sum_j \sum_k \sum_l w_{ijkl} y_{ijkl} \quad (1)$$

여기서,  $w_{ijkl} = w_{ijkl}^0 \times \frac{X_{i,sa}}{\widehat{X}_{i,sa}}$  은 사후층화 가중값으로, 표본 설계당시의 설계가중값  $w_{ijkl}^0$ 에 대해 사후적으로 성별, 연령별 가중값을 곱한 형태이다. 설계가중값 및 사후층화 가중값 계산방법에 대해서는 통계청 (2006)을 참고할 수 있다. 그리고  $y_{ijkl}$ 는  $i$ 지역,  $j$ 번째 표본조사구 내  $k$ 번째 가구내  $l$ 번째 가구원의 관측값으로 우리의 관심인 실업특성을 나타낸다. 실제조사에서  $y$ 는 실업이면 1, 그렇지 않으면 0의 값을 가질 수 있다. 이상의 수식에서 사용된 첨자의 의미는 다음과 같다.

$i=1, \dots, 151$ , 지역 층

$j=1, \dots, n_i$  표본조사구,  $n_i$ 는  $i$ 번째 지역의 총 조사구수

$k=1, \dots, m_{ij}$  표본가구,  $m_{ij}$ 는  $i$ 지역의  $j$ 번째 조사구내 총 표본가구수

$l=1, 2, \dots, l_{ijk}$  표본가구원,  $l_{ijk}$ 는  $i$ 번째 지역의  $j$ 번째 조사구내  $k$ 가구의 가구원수

$X_{i,sa}$   $i$ 번째 지역의 성별, 연령별 15세 이상 모집단 추계인구수

$\widehat{X}_{i,sa}$   $i$ 번째 지역에서 조사된 성별, 연령별 15세 이상 인구수.

참고로 성별은 남녀 2개 그룹과 연령별로는 15세부터 5세 단위로 11개 그룹을 사용한다.

직접추정량은 불편추정량이기 때문에  $\widehat{Y}_i^{dir}$ 의 평균제곱오차는  $\widehat{Y}_i^{dir}$

의 분산과 같다.

## 2. 합성추정량

‘합성’이라는 용어는 대지역으로부터 계산된 추정량이 그 대지역을 구성하는 하부 지역 각각의 특성을 추정하는데 추가적인 정보로 사용된다는 것을 의미한다. 이때 소지역들은 추정된 특성에 대해 대지역의 특성과 동질(homogeneous)하다고 가정한다. 따라서 합성추정량(synthetic estimator)은 이미 다른 ‘유사지역’으로부터 정보를 빌려온 셈이다. 단순한 예로 보조정보가 없을 때 소지역 평균 각각을 추정하는데 전체평균(global mean)  $\bar{y}$ 를 이용할 수 있다.

합성추정량은 2단계 절차에 의해 구해진다. (i) 대지역 수준의 추정치는 대지역을 구성하는 하부그룹(즉, 소지역)에 의해 결정된다. (ii) 대지역 수준의 추정치는 관심 소지역 내에서 하부그룹 빈도에 비례해서 다시 소지역별로 조정된다. 이때 대지역내에서  $m_k$ 개의 소지역들을 포함하고 있는 부차관심영역인 각 시, 군 그룹 내의 소지역  $i$ 의 실업자에 대한 합성추정량,  $\widehat{Y}_i^{sm}$ 은 다음과 같다.

$$\widehat{Y}_i^{sm} = \sum_{g=1}^4 \frac{{}_g\widehat{X}_i}{{}_g\widehat{X}} {}_g\widehat{Y}^{dir}, \quad g=1,2,3,4 \quad (2)$$

여기서,  ${}_g\widehat{X} = \sum_{i=1}^{m_k} {}_g\widehat{X}_i$ ,  ${}_g\widehat{Y}^{dir} = \sum_{i=1}^{m_k} \sum_{h=1}^{n_i} {}_gM_{i g} y_{ij}$

$m_k$ 는 만약 소지역  $i$ 가 시이면,  $k=1$ , 즉,  $m_1$ 은 소지역  $i$ 가 포함된 대지역내의 시의 개수이고, 그렇지 않으면(소지역  $i$ 가 군인 경우),

$k=2$ , 즉,  $m_2$ 는 소지역  $i$ 가 포함된 대지역의 군의 개수,

$g$  = 성별(남/여) 및 연령대별(15~29세/30세 이상) 그룹의 4개 범주 수,

$y_{ij}$  = 성별, 연령별로 조사된  $i$ 소지역내의  $h$ 번째 조사구의 실업자수,

${}_gX_i$  = 성별, 연령별,  $i$  소지역의 15세 이상 추계인구,

$\widehat{X}_i$  = 성별, 연령별,  $i$  소지역의 경찰조사에서 15세 이상 조사인구,

${}_gM_i = {}_gX_i / \widehat{{}_gX}_i$  : 소지역  $i$ 의  $g$ 범주에 대한 승수

$\widehat{{}_gY}_i^{dh}$  = 성별, 연령별 범주의 소지역을 포함하는 대지역의 직접추정량.

식(0.2)에서  $\widehat{{}_gY}_i^{sm}$ 가 불편추정량이 되기 위해서는  $\frac{{}_gY}{{}_gX} = \frac{{}_gY_i}{{}_gX_i}$ 를 만족해야 하고, 이를 만족하지 못할 경우에는 편향추정량이 된다. 이 때,  $\widehat{{}_gY}_i^{sm}$ 의 평균제곱오차의 근사적 불편추정량은 다음과 같이 주어진다(통계청, 2002).

$$mse(\widehat{{}_gY}_i^{sm}) \approx (\widehat{{}_gY}_i^{sm} - \widehat{{}_gY}_i)^2 - \widehat{Var}(\widehat{{}_gY}_i) \quad (3)$$

합성추정량을 구할 때 성별, 연령별 범주를 4개 사용한 이유는 이들 범주 간에 실업특성이 크게 차이가 있기 때문이다. <표 2>는 2005년 경제활동인구조사에서의 연평균 실업률을 나타낸다. <표 2>에서 보는 바와 같이 남녀와 29세 이하와 30세 이상의 연령 간에 실업률의 변동이 크게 나타나고 있다. 이는 이들 범주 간에 실업특성의 분포가 크게 다르다는 것을 의미하며 소지역으로 갈수록 더 큰 변동을 보일 수 있음을 시사한다. 따라서 이들 범주 간을 나누어서 실업인구 특성을 고려하는 것이 타당할 것으로 보인다.

<표 2> 성별, 연령대별 연평균 실업률

성별	연령대별 연평균 실업률(%)						
	계	15~19세	20~29세	30~39세	40~49세	50~59세	60세 이상
남	4.0	13.2	9.0	3.5	2.7	3.0	1.8
여	3.4	11.8	6.5	3.1	2.3	1.8	0.7

합성추정량은 기본적으로 많은 수의 표본에 기반을 두기 때문에 적은 분산을 갖는다. 이처럼 합성추정량의 장점은 분산의 감소이지만, 만약 대영역 내에서 동질성에 대한 가정이 위배될 시에는 심각한 편향(bias)이 초래될 수 있다. 편향에 대한 감소(그러나 분산증가의 대가가 따름)는 복합추정량을 사용함으로써 보완될 수 있다. 합성추정량에 관한 논문은 Purcell and Linacre(1976), Gonzalez and Hoza (1978),

Ghangurde and Singh(1978) 등을 참고할 수 있다.

### 3. 복합추정량

복합추정량은 직접추정량(편향이 아주 적거나 없지만 큰 분산)과 합성추정량(작은 분산이지만 큰 편향)의 가중 합으로  $i$ 번째 소지역의 복합추정량  $\widehat{Y}_i^{com}$ 형태는 다음과 같다.

$$\widehat{Y}_i^{com} = \widehat{w}_i \widehat{Y}_i^{dir} + (1 - \widehat{w}_i) \widehat{Y}_i^{syn} \quad (4)$$

여기서  $\widehat{Y}_i^{dir}$ 은 직접추정량이고,  $\widehat{Y}_i^{syn}$ 은 합성추정량이다.  $w_i$ 는 상수로서  $0 \leq w_i \leq 1$  값을 갖는다. 복합추정량을 사용하기 위해 신중하게 고려해야 할 부분은 가중치  $w_i$ 의 선택이다. 가중치는 추정량의 평균제곱오차(MSE)를 최소화하는 값을 선택하는 것이 이상적이며, 다음의 근사식을 이용한다.

$$\widehat{w}_i = \frac{MSE(\widehat{Y}_i^{syn})}{MSE(\widehat{Y}_i^{syn}) + MSE(\widehat{Y}_i^{dir})} \quad (5)$$

그런데 이 경우, 식(3.3)의 편향 부분, 즉,  $(\widehat{Y}_i^{syn} - \widehat{Y}_i)^2$ 은 일반적으로 정확하게 측정할 수 없다. 그래서 한 가지 간단한 대안으로 가중치를  $w_i = f_i = n_i / N_i$ 로 하는 것이다. 그래서  $f_i$ 가 증가함에 따라 직접불편추정량에 더 많은 가중치를 부여하게 되는 것이다. 그러나 보통  $f_i$ 는 매우 작아서 이 가중치를 사용하는 것은 실제로 결과 변수의 변동과는 무관하게 합성추정량을 사용하는 것과 유사하게 된다. 이러한 방법을 표본 크기에 의존하는(Sample Sized Dependent: SSD) 추정량이라 한다. 본 연구에서 이 SSD는 고려하지 않는다. 이 외에도 그럴듯한 구체적 가중치에 관한 방법은 Ghosh and Rao(1994), Thomsen and Kleive Holmoy(1998)와 Marker(1999)를 참고할 수 있다.

또한 식(0.3)의  $\widehat{Y}_i^{syn}$ 의 평균제곱오차의 근사적 불편추정량은 소지역

에 배정된 표본조사구 수가 충분하지 못할 경우 음의 값이 나올 가능성도 있다. 그래서 새로운 대안으로 잭나이프 추정방법을 고려할 수 있다. 잭나이프 추정법은 추정량의 형태에 국한되지 않아 계산 및 응용이 쉽고 추정 결과도 신뢰할 만하다는 장점이 있다. 각 추정량들의 MSE 추정을 위한 추정식의 일반적인 형태는 다음과 같다(권순필, 2007).

$$mse_f(\widehat{Y}_{i.}) = \widehat{Var}_f(\widehat{Y}_{i.}) + [\widehat{Bias}_f(\widehat{Y}_{i.})]^2 \quad (6)$$

여기서,  $\widehat{Var}_f(\widehat{Y}_{i.}) = \frac{n_i - 1}{n_i} \sum_{h=1}^{n_i} [\widehat{Y}_{i.}(h) - \frac{1}{n_i} \sum_{h=1}^{n_i} \widehat{Y}_{i.}(h)]^2$ ,

$$\widehat{Bias}_f(\widehat{Y}_{i.}) = (n_i - 1) \left[ \frac{1}{n_i} \sum_{h=1}^{n_i} \widehat{Y}_{i.}(h) - \widehat{Y}_{i.} \right],$$

$\widehat{Y}_{i.}(h)$ 는  $h$ 번째 조사구를 제외하고 계산된 추정량이다. 합성, 복합추정량 계산 시 위 식에서  $\widehat{Y}_i$  대신  $\widehat{Y}_{i^{syn}}$ ,  $\widehat{Y}_{i^{com}}$ 를 대입하여 같은 방식으로 계산한다.

본 연구에서 고려한 합성추정량과 복합추정량들의 공통적인 특징은 그들의 유도과정에서 분포 가정이 없고 분산과 편향이 확률분포로부터 계산된 것이 아니라는 점에서 모형과 무관(model free)하다는 점이다. 또한 이들 자료 기반에 의한 합성과 복합추정법은 표본 수가 많고, 대지역과 소지역의 관심값에 대한 특성이 동일하다는 가정이 만족될 경우에는 사용할 만하다. 그런데 만약 이러한 가정이 만족되지 않는다면 소지역 추정치에 있어서 편향을 초래할 가능성이 높다는 점에 유의해야 한다.

#### 4. EBLUP 추정량

자료 기반 추정방법이 대지역의 직접추정값을 소지역의 특성값에 비례적으로 분배하는 개념을 사용하기 때문에 분산의 과소 추정과 특정지역에서의 편향문제가 크게 이슈화되었다. 그런데 소지역 추정값을 구하는 데 적절한 추가정보를 이용할 수 있다면, 이러한 문제는 통계적 모형

을 통해 해결할 수 있다. 국제조사통계연구회(IASS)의 위성 컨퍼런스인 ‘소지역추정연구회(SAE)’에서도 표본조사에서 모형 사용이 어떤 경우에는 반드시 필요하다는 것에 대해 꾸준히 논의되고 있다.

본 연구에서 다루는 3절과 4절의 EBLUP과 HB추정에 관한 부분은 2장에서 언급된 내용의 일부분을 재정리한 것이다. 가장 간단한 모형 중 하나가 Fay-Herriot 모형이고, 이는 소지역 추정에 종종 이용되는 모형이다. 본 연구도 추정량 비교를 위한 모의실험인 만큼 가능한 단순한 모형을 사용하여 모형 추정량과 자료 기반 추정량을 비교하고자 하였다. Fay-Herriot 모형에는  $x_i\beta + v_i$ 가 회귀직선 상에 있지 않고, 지역 효과  $v_i$ 만큼 직선에서 떨어져 있다고 가정한다. 구체적인 모형은 다음과 같다.

$$y_i = x_i\beta + v_i + e_i \quad e_i \sim \mathcal{N}(0, \psi_i), \quad v_i \sim \mathcal{N}(0, \sigma_v^2)$$

여기서,  $i$ 는 소지역,  $y_i$ 는 관측치(실업자 수 또는 실업률),  $x_i$ 는 설명변수(보조정보),  $v_i$ 는 지역 효과,  $e_i$ 는 표집오차를 나타낸다. 표집오차  $e_i$ 의 분산  $\psi_i$ 는 알려져 있다고 가정한다. 실제로는  $\psi_i$  대신에 관측 자료로부터 계산한 추정량  $\hat{\psi}_i$ 를 이용한다. 본 연구에서  $\hat{\psi}_i$ 는 일반적으로 사용되는 잭나이프방법에 의해 구한다. 지역효과  $v_i$ 의 분산  $\sigma_v^2$ 은 각 지역에서 동일하다고 가정한다. 알려진 분산  $\psi_i$ 와  $\sigma_v^2$ 에 대해서 모수의 최량선형 불편추정량(Best Linear Unbiased Estimator: BLUP)은

$$\hat{\theta}_i = \lambda_i y_i + (1 - \lambda_i) \mathbf{x}_i' \hat{\beta}$$

여기서  $\hat{\beta}$ 은 관측값으로부터 계산된  $\beta$ 의 최소제곱추정량이고,  $\lambda_i = \sigma_v^2 / (\psi_i + \sigma_v^2)$ 이다. 이때 가중치  $\lambda_i$ 는  $e_i$ 와  $v_i$ 의 분산비에 의해 결정된다. 만약 관측치의 표집오차  $\psi_i$ 가 커지면  $\lambda_i$ 는 0에 가깝게 되고 BLUP 추정량은 모형추정에 더 많은 영향을 받게 된다. 한편  $\sigma_v^2$ 이 커지면  $\lambda_i$ 는 1에 가까워지고 BLUP 추정량은 관측치의 영향을 더 받게 된다.

일반적으로  $\sigma_v^2$ 이 알려져 있다고 가정하지만, 실제로  $\sigma_v^2$ 은 모르는 경우가 많다. 이런 경우에는 그 값을 관측 자료로부터 추정하고, BLUP 추정량  $\hat{\theta}_i$ 에서  $\sigma_v^2$ 을 추정치  $\hat{\sigma}_v^2$ 으로 바꾸면 이것을 EBLUP 추정량이라 부른다. EBLUP의 MSE는  $\hat{\sigma}_v^2$ 을 최소제곱추정법으로 추정했다는 전제 하에 다음과 같이 구할 수 있다(Rao, 2003).

$$MSE(\hat{\theta}_i^{EBLUP}) = g_{1i}(\hat{\sigma}_v^2) + g_2(\hat{\sigma}_v^2)$$

$$\left( \begin{array}{l} g_{1i}(\hat{\sigma}_v^2) = \frac{\hat{\sigma}_v^2 \psi_i}{\psi_i + \hat{\sigma}_v^2} (= \bar{y}_i \psi_i) \\ g_2(\hat{\sigma}_v^2) = (\bar{y}_i)^2 \mathbf{x}_i' \left[ \sum_{i=1}^m \frac{\mathbf{x}_i \mathbf{x}_i'}{\psi_i + \hat{\sigma}_v^2} \right]^{-1} \mathbf{x}_i \\ \bar{V}(\hat{\sigma}_v^2) = [\hat{Q}]^{-1} = \left[ \sum_{i=1}^m \frac{1}{(\psi_i + \hat{\sigma}_v^2)^2} \right]^{-1} \end{array} \right)$$

### 5. HB 추정량

HB 추정량은 계층적 베이زي안 모형을 이용한다. 관측된 자료는 어떤 모수를 조건부로 한 조건부 모형이고, 이 모수는 다시 초모수(hyperparameter)라 불리는 추가적인 모수에 의존하는 확률분포로 모형화된다. 이때 계층적 베이زي안 분석은 깃스 표본자(gibbs sampler)를 사용한다. 깃스 표본자는 모형으로부터 모든 확률변수의 결합확률분포가 주어졌을 때, 저차원의 모든 조건부확률분포로부터 표본을 반복하여 충분히 생성하면 결국 생성된 표본이 결합확률분포와 주변확률분포에 대한 표본이 된다는 이론에 근거한다. 자세한 내용은 You 등(2003)과 김달호(2005)를 참고할 수 있다.

소지역 추정을 위한 계층적 베이지 모형은 다음과 같다.

- $\theta_i$ 에 대한 조건부  $y_i | \theta \sim \mathcal{N}(\theta, \sigma^2)$

◦  $\beta, \sigma_v^2$ 에 대한 조건부  $\Theta | \beta, \sigma_v^2 \sim \mathcal{N}(X^t \beta, \sigma_v^2)$

◦ 이때  $\beta$ 와  $\sigma_v^2$ 은 독립이며,  $\sigma_v^2 \sim IG(a, b), \beta \propto 1$

김스 표본자를 사용하기 위한 조건부확률분포는 다음과 같다. 이때,  $y=(y_1, \dots, y_m)^t, X^t=(x_1, x_2, \dots, x_m), \Theta^t=(\theta_1, \theta_2, \dots, \theta_m)^t$ 이며  $m$ 개의 소지역  $i=1, 2, \dots, m$ 에 대해서,

$$\begin{aligned} \beta | y, \sigma_v^2, \Theta &\sim \mathcal{N}(X^t X)^{-1} X^t \Theta, \sigma_v^2 (X^t X)^{-1}, \\ \Theta | y, \beta, \sigma_v^2 &\sim \mathcal{N}((1-r_i)y_i + r_i x_i^t \beta, \sigma_v^2 (1-r_i)), \quad r_i = \sigma_v^2 / (\sigma_v^2 + \sigma_i^2), \\ \sigma_v^2 | y, \beta, \Psi, u, \Theta &\sim IG(a + m/2, b + \sum_{i=1}^m (\theta_i - x_i^t \beta)^2 / 2). \end{aligned}$$

주어진 조건부확률분포에 근거하여 김스 표본자를 사용한 베이지안 계산을 할 때, Gelman과 Rubin(1992)의 길이가  $2d(=d+D)$ 인  $L$ 개의 병렬체인을 고려한다. 병렬체인을 사용하면 김스 표본자의 수렴여부를 쉽게 검증할 수 있기 때문이다.

Gelfand와 Smith(1991)에서 제안한 바와 같이 이들 사후분포로부터 각 확률변수에 대한 조건부사후분포의 평균과 분산을 구한다. 이때에는 MCMC방법에 의한 라오-블랙웰 추정량(Rao-Blackwellized estimator)을 이용한다.  $i$  소지역의 사후분포에 대한 평균( $E[\Theta_i | y_i]$ )과 분산( $V[\Theta_i | y_i]$ )의 라오-블랙웰 추정량은 다음과 같은 형태를 갖는다(You 등, 2003).

$$\begin{aligned} E[\Theta_i | y_i] &= \frac{1}{LD} \sum_{l=1}^L \sum_{k=d+1}^{2d} [(\sigma_v^{-2(lk)} + \Sigma_i^{-1}) \times (\Sigma_i^{-1} \mathbf{y}_i + \sigma_v^{-2(lk)} (X_i \beta^{(lk)} + \mathbf{u}_{i^{(lk)}}))], \\ V[\Theta_i | y_i] &= \frac{1}{LD} \sum_{l=1}^L \sum_{k=d+1}^{2d} (\sigma_v^{-2(lk)} + \Sigma_i^{-1}) \end{aligned}$$

$$\begin{aligned}
 & + \frac{1}{(LD)^2} \sum_{i=1}^L \sum_{k=d+1}^{2d} [(\sigma_v^{-2(lk)} + \Sigma_i^{-1}) \times (\Sigma_i^{-1} \mathbf{y}_i + \sigma_v^{-2(lk)} (X_i \boldsymbol{\beta}^{(lk)} + \mathbf{u}_{i^{(lk)}}))] \\
 & \quad \times [(\sigma_v^{-2(lk)} + \Sigma_i^{-1}) \times (\Sigma_i^{-1} \mathbf{y}_i + \sigma_v^{-2(lk)} (X_i \boldsymbol{\beta}^{(lk)} + \mathbf{u}_{i^{(lk)}}))]^t \\
 & - \frac{1}{(LD)^2} \left[ \sum_{i=1}^L \sum_{k=d+1}^{2d} [(\sigma_v^{-2(lk)} + \Sigma_i^{-1}) \times (\Sigma_i^{-1} \mathbf{y}_i + \sigma_v^{-2(lk)} (X_i \boldsymbol{\beta}^{(lk)} + \mathbf{u}_{i^{(lk)}}))] \right] \\
 & \quad \times \left[ \sum_{i=1}^L \sum_{k=d+1}^{2d} [(\sigma_v^{-2(lk)} + \Sigma_i^{-1}) \times (\Sigma_i^{-1} \mathbf{y}_i + \sigma_v^{-2(lk)} (X_i \boldsymbol{\beta}^{(lk)} + \mathbf{u}_{i^{(lk)}}))] \right]^t.
 \end{aligned}$$

여기서,

$L$ : 동시에 생성할 마코프 체인의 수

$d$ : burn in까지 각 마코프 체인 안에서 생성된 난수의 총 수

$D$ : burn out 후에 생성된 난수의 총 수이다.

## 제 4 절 모의실험

### 1. 표본추출

우리나라 행정구역상 소지역은 전체 시·군·구로 정의될 수 있다. 본 연구에서는 전체 시·군 지역만을 소지역으로 정의한다. 본 연구에서 행정구역상 ‘구’ 지역을 소지역 정의에서 제외한 이유는 첫째, 특·광역시외의 경우 지리적으로 도 단위 지방보다는 규모가 작고 고용 형태가 유사하다는 점에서 굳이 ‘구’ 단위 추정이 필요하지 않을 것으로 판단되었고, 둘째, 2008년 10월 ‘시군구 고용조사’에서 특·광역시 내의 ‘구’는 조사 단위에 포함되지 않았기 때문이다. 본 연구는 향후 이 ‘시군구 고용조사’ 결과와 소지역 추정 결과의 비교를 염두에 둔 부분도 있다.

모의실험에 사용된 최종 소지역은 도에 속하는 전체 158개의 시군 중에서 표본조사구가 배정되어 있지 않은 계룡시, 신안군, 울릉군과 표본조사구 수가 1개인 증평군, 완도군, 진도군, 군위군을 제외한 151개 시군이 해당된다. 표본추출방법은 가능한 경찰조사의 표집 방법과 추정 방법에 근사하도록 설계하였다. 왜냐하면 가상(virtual)의 모의실험이긴 하지만, 모의실험의 환경을 현재 관심의 대상인 소지역과 그 소지역의 표본 수, 관심값의 특성 등에 최대한 맞추어 줌으로써 연구의 현실성(reality)을 높이기 위함이다.

모집단은 2005년 인구총조사의 10% 표본조사 자료를 이용한다. 모집단에서 조사가 곤란한 조사구는 제외하였다. 대부분 경찰조사에서 사용하는 모집단과 유사하나, 다른 점은 인구총조사 실시 이후부터 표본 개편 시까지(2005.11.1. ~ 2006.10.31) 신축된 아파트 조사구 자료(506개)를 사용하지 않았다는 것이다. 그래서 이번 모의실험에서의 모집단 표본조사구 수는 경찰조사의 모집단 표본조사구 수보다는 적게 구성되어 있다.

표본조사구의 추출은 경찰조사는 25개 지역 층에서 조사구의 규모(가구 수)에 비례하는 확률비례표집 방법을 이용한다. 이때 25개 층 내 시군의 표본조사구 개수는 보다 상위 대지역의 표본을 추출하는 과정에서 임의로 배정된 것이다. 이에 비해 모의실험은 경찰조사에서 배정된 시군의 표본조사구 개수만큼 151개 시군 각각에서 조사구 크기에 비례하는 확률비례표집 방법으로 표본조사구를 추출한다. 즉, 모의실험은 경찰조사의 표본조사구 개수와 동일하게 뽑기 위해서 처음부터 151개 시군에 경찰조사의 표본조사구 개수만큼 표본을 할당하는 방법을 사용한다. 예를 들면, 경찰조사에서 경기도는 확률비례표집에 의해 총 198개의 표본조사구가 할당되었고, 시군별로는 수원시에 16개, 성남시 15개 등 31개 시군에 각각의 표본조사구 수가 할당되었다. 이와 동일한 개수만큼 시군의 표본조사구 수를 추출하기 위해 모의실험에서는 수원시 16개, 성남시 15개 등 처음부터 표본조사구 수를 할당한 후 추출하였다.

표본가구는 경찰조사에서는 조사구를 5가구씩 조사구역으로 분할한 후 각 거처의 구역화가 완료되면 표본으로 뽑힌 1개 구역과 인접구역 3

개, 즉, 4구역 20가구를 추출한다. 그러나 이러한 구역화가 모의실험 연구에서는 쉽지 않기 때문에 본 연구에서는 각 표본조사구 내에서 20가구를 단순임의표집방법(SRS)으로 추출한다. 사실 모의실험과 같이 랜덤하게 가구를 추출하는 것이 보다 합리적인 방법이긴 하지만, 현장조사 시 조사상의 어려움에 직면할 수 있기 때문에 편의상 경찰조사와 같은 방법이 표집방법으로 선호되기도 한다.

본 연구에서는 이와 같은 표본 추출과정을 1,000번 반복하여 1,000개의 표본을 생성하여, 각각의 표본에 대한 추정량을 구하고 추정결과를 평가한다.

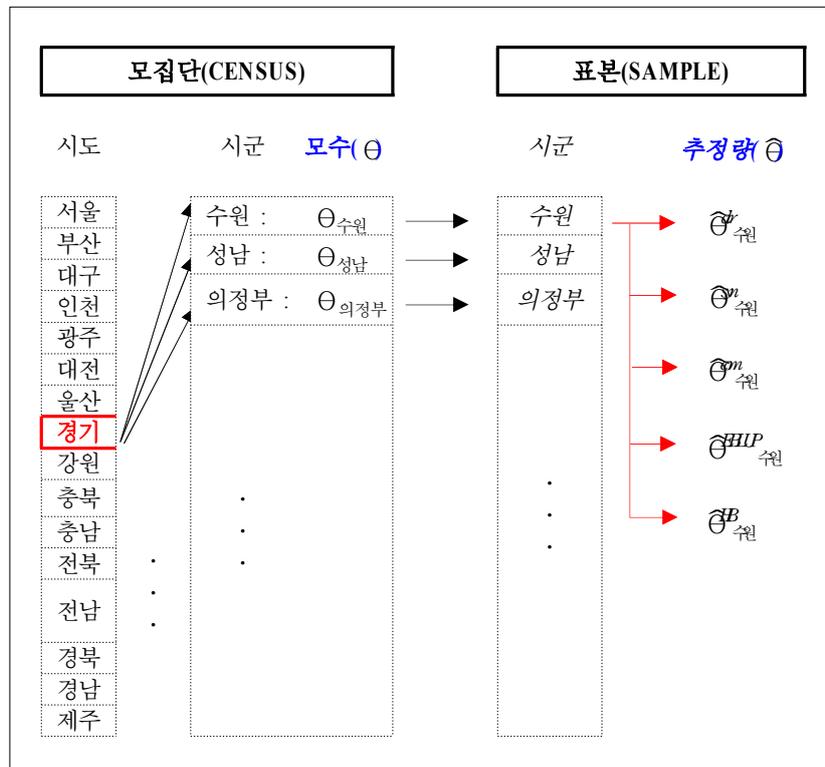
## 2. 추정량 계산

각 시군별 표본조사구의 실업자 수를 대상으로 직접추정량, 합성추정량, 복합추정량, EBLUP, HB 등 5개의 추정량을 계산한다. EBLUP과 HB 추정량 계산을 위한 모형은 소지역 추정 모형 중 가장 단순한 형태를 가지는 Fay-Herriot모형을 이용한다. 이 모형에 지역 정보와 시간 정보를 사용하여 모형추정치보다 효과적으로 추정할 수 있는 방법도 있다. 그러나 본 연구는 인구총조사 자료를 모집단으로 이용하기 때문에 시계열형태 모형 등은 고려할 수 없기도 하다. 여기서 실업자 수는 이산형(discrete) 자료로서 실제로 한쪽으로 치우치는(skewed) 성향이 강하게 발생한다. 따라서 모의실험에서는 이런 경향을 완화하고자 원자료를  $\log(\text{실업자수})$ 로 변환하여 추정하였다.

각 추정량의 비교를 위한 참값으로 인구총조사의 10% 표본조사 자료의 실업자 수를 사용한다. 3절에서 설명한 각 추정방법에 의해 5개의 추정량,  $\hat{\theta}_{\text{소지역}}^d$ ,  $\hat{\theta}_{\text{소지역}}^m$ ,  $\hat{\theta}_{\text{소지역}}^m$ ,  $\hat{\theta}_{\text{소지역}}^{\text{EBLUP}}$ ,  $\hat{\theta}_{\text{소지역}}^{\text{HB}}$ 을 계산한다. 그리고 [그림 2]와 같이 모집단의 각 소지역 모수( $\theta_{\text{소지역}}$ )값과 각각 추정량들을 평가기준에 의해 비교함으로써 각 소지역 추정량들의 성능을 비교한다.

단, 본 연구는 다양한 소지역 추정방법에 따른 추정량들의 성능을 비교하고자 하는 주된 목적임을 앞에서 언급하였다. 따라서 참값이 있는 인구총조사 자료를 활용하여 이로부터 가상의 표본을 만들고, 이들 표본자

료로부터 추정한 추정결과들이 참값과 얼마나 근사하게 추정되었는지를 파악하는 것이 연구의 핵심 목표이다. 이에 모의실험에 의한 추정결과와 경찰조사와의 직접적인 비교는 불가능하다는 점에서 본 연구의 결과를 해석하는 데 착오가 없기를 당부한다. 실제로 인구총조사와 경찰조사의 경찰 항목 간의 직접적인 비교는 어렵다. 왜냐하면, 두 조사의 조사시기, 조사대상, 표본규모 및 설계, 조사원 성향이 다르기 때문에 두 통계를 직접 비교할 수 없음을 유념해야 한다.



[그림 1] 모의실험 추정 및 추정량 비교 과정

지금까지 설명한 각 지역에서의 모의실험 과정을 간단하게 요약해보면 다음과 같다.

STEP1. 표본추출 : R개

√ 시군에 할당된 개수만큼의 표본조사구 추출(PPS)

(모집단  $k=151$ 개 시군  $\rightarrow n_h$ 개만큼 각 시군의 할당된 표본 추출)

√ 표본가구 추출 : 각 조사구마다 20가구를 랜덤 추출

STEP2. 실업자수 추정 : R=1,000개의 표본에 대해 5개의 추정량 계산

$$, \hat{Y}_{i,r}^u, \hat{Y}_{i,r}^m, \hat{Y}_{i,r}^{HLP}, \hat{Y}_{i,r}^B (r=1, \dots, 1,000) \text{ 계산}$$

STEP3. 추정량 평가 : 모수값에 대한 상대편향, 평균제곱오차 계산

### 3. 권역별 추정

본 연구는 효율적인 소지역 추정을 위해 다양한 시도를 하였다. 특히 우리나라의 경우, 지리적으로는 좁은 편이지만, 좁은 면적에 비해 특정 권역별로 경제·사회적 변화가 매우 심하다고 할 수 있다. 지역별 실업자는 지역별 경제활동 상황과 관련이 있기 때문에 경제활동이 활발한 지역, 인구밀집지역, 대도시 인근 지역들에서 각 소지역들은 유사한 성향을 보일 수 있다. 따라서 이런 특성을 고려하여 151개 시군을 유사한 특성을 가질 것으로 판단되는 몇 개의 권역으로 나누어 추정해 보았다. 모의실험에서 고려한 권역은 다음과 같다.

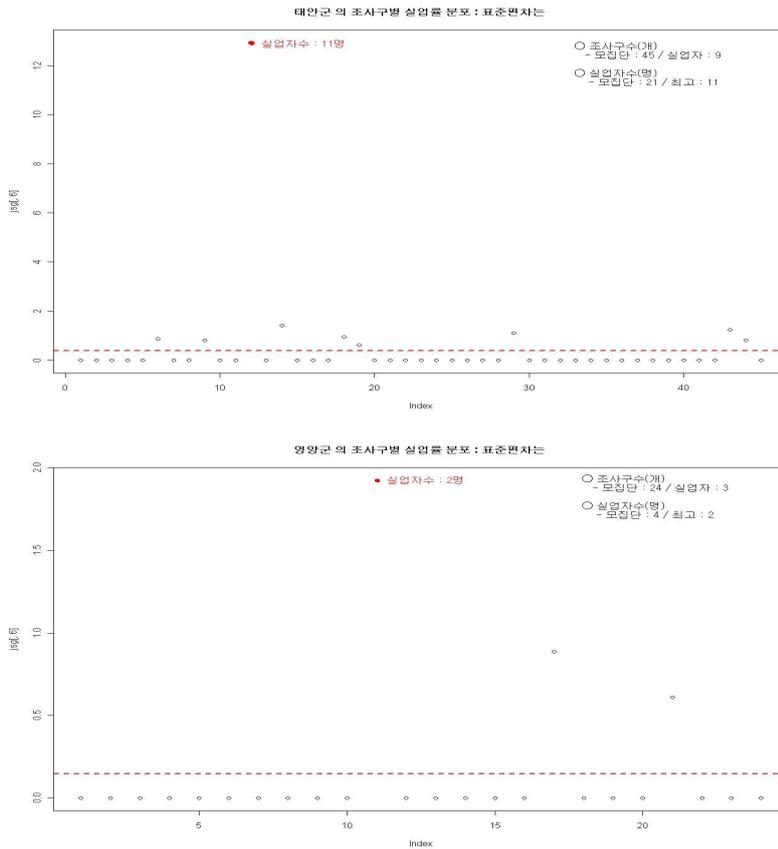
- ① 151개 시군을 한 개의 권역으로 간주하여 소지역 추정량 계산
- ② 151개 시군을 시와 군의 2개 권역으로 간주하여 각 권역별로 추정량 계산
- ③ 151개 시군을 인구규모 25만 명을 기준으로 25만 명 이상과 미만의 2개 권역으로 간주하여 각 권역별로 추정량 계산

위의 3가지 각 경우에 대해 모형 추정량 EBLUP과 HB 추정량이 각각 계산되고 비교되었다. ③의 권역은 ②와 유사하지만, 우리나라 행정 구역상 ‘시’라 하더라도 인구규모가 작은 지역이 많고, 그 경제활동 역시 ‘군’지역과 비슷한 지역들이 많기 때문에 ②와는 약간 다른 경향을 보일 수 있다. 특히 인구규모가 작은 지역은 실업자 수도 상대적으로 적은 경우가 많고, 표본조사구 또한 적게 추출될 가능성이 높다. 또한 실업자 수가 적은 지역의 경우, 조사구에 따라 실업자 분포가 매우 다를 수 있고, 이는 모의실험에서 분산이 커지게 되는 원인이 될 수 있다. 실제로 어떤 지역의 경우, 각 조사구에 포함된 실업자 수가 상당히 차이를 보이는 경우가 발생하고 이런 지역은 모의실험 추정치의 편향이나 분산이 매우 크게 나타날 가능성이 있다.

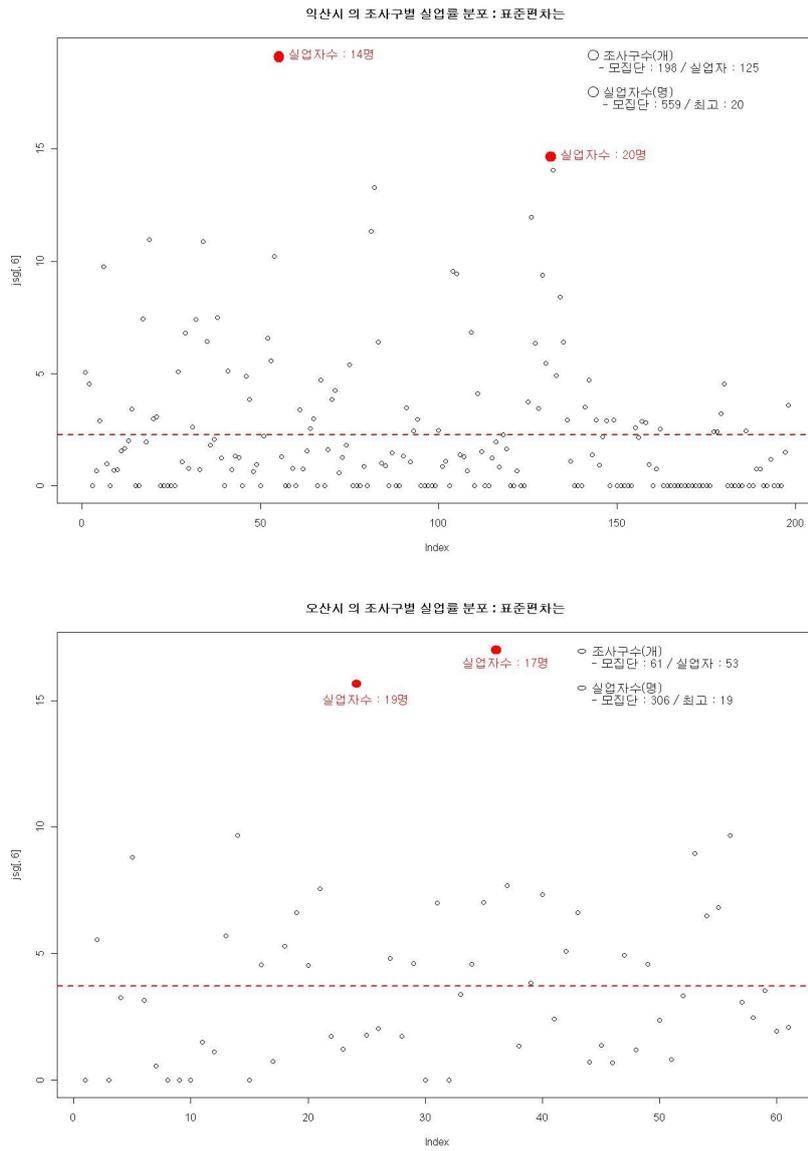
[그림 2]와 [그림 3]은 각각 조사구별로 실업자가 편중되어 있는 지역과 고루 퍼져있는 지역의 조사구별 실업률 분포를 산점도로 나타낸 것이다. 하나의 예로, 태안군과 영양군의 경우를 보면, 모집단에 비해서 실업자가 있는 조사구의 비중이 각각 20%, 12.5%로 상당히 작다. 또한 특정 조사구 내에 전체 실업자의 과반을 차지하는 실업자가 존재하고 있어 해당 조사구의 표집 여부에 따라 추정값에 중대한 영향을 미친다는 것을 알 수 있다. 즉, [그림 2]의 태안군, 영양군의 조사구별 실업률 분포 그림을 보면 태안군의 총 모집단 조사구는 45개인데, 이 중에서 실업자를 최소한 1명 이상 포함하는 조사구는 9개에 불과하다. 게다가 45개 모집단 조사구에 포함된 총 실업자 수는 21명인데 반해 굵은 점으로 표현된 조사구 1개에만 11명의 실업자가 포함되어 있다. 즉, 태안군 전체 실업자의 50% 이상이 한 조사구에 집중되어 있는 것이다. 영양군의 경우는 기본적으로 실업자가 매우 적은 지역으로 모집단 조사구 21개 내에 포함된 실업자 수는 단 4명뿐이다. 이 4명의 실업자가 겨우 4개의 조사구에 2명, 1명, 1명으로 분포되어 있어 표본조사로 실업자를 포착하기가 쉽지 않다. 반면 [그림 3]의 익산시와 오산시의 조사구별 실업률 분포를 보면 실업자가 있는 조사구의 비중이 각각 63.1%, 86.9%로 상당히 높고 앞의 태안군과 영양군에 비해 특정 조사구에 많은 비중의 실업자 수가 포함되어 있지 않다.

22 총조사 자료를 이용한 소지역 추정량 비교

이는 태안군과 영양군같이 지역의 규모 및 표본 수가 작으면서 특정 조사구에 많은 비중의 실업자 수가 포함되는 경우 매번 구성되는 모의 실험 표본으로부터 편향된 결과를 도출할 수 있음을 시사한다. 뿐만 아니라 태안군과 영양군의 이 같은 특성은 우리나라 대부분의 ‘군’ 단위 지역에서 비슷하게 나타나는 현상이다. 따라서 이들 지역들에 대해서 소지역 추정을 적용할 때에는 모집단 및 표본의 규모를 늘리고 추정이 용이한 실업자 수를 확보할 수 있도록 비슷한 지역들을 집락화하는 등의 노력을 기울여 분석함으로써 소지역 추정의 효과를 높일 수 있을 것으로 기대한다.



[그림 2] 태안군, 영양군의 조사구별 실업률 분포



[그림 3] 익산시, 오산시의 조사구별 실업률 분포

#### 4. 평가 방법

소지역 추정량의 평가는 다음과 같은 평균상대편향(ARB)과 상대평균제곱오차(RMSE)를 이용하겠다. 본 연구에서는  $R=1,000$ 번의 모의실험을 수행한다.

$Y_i$ 는 소지역  $i$ 의 모집단의 특성값, 즉, 모수이며,  $\hat{Y}_{i,r}$ 은  $r$ 번째 표본으로부터 계산된 소지역  $i$ 의 추정량이다. 소지역  $i$ 에 대한 추정량의 평균상대편향은 다음과 같다.

$$ARB_i = \frac{1}{R} \left( \sum_{r=1}^R \frac{\hat{Y}_{i,r}}{Y_i} \right) \times 100, \quad R=1,000.$$

전체  $m$ 개의 소지역 추정량에 대한 평균상대편향의 평균은 다음과 같이 구할 수 있다. 즉,

$$\overline{ARB} = \frac{1}{m} \sum_{i=1}^m ARB_i, \quad m \text{은 소지역 개수.}$$

상대편향, ARB는 100에 가까울수록 좋고 이는 상대적으로 편향이 작음을 의미한다. 상대편향은 추정량이 참값으로부터 얼마나 편향되었는지를 판단할 수 있다.

소지역  $i$ 의 추정량에 대한 상대평균제곱오차는 다음과 같다.

$$RMSE_i = \frac{\sqrt{MSE_i}}{Y_i} \times 100, \quad R=1,000.$$

이때, 평균제곱오차는  $MSE_i = \frac{1}{R} \sum_{r=1}^R (\hat{Y}_{i,r} - Y_i)^2$ 이다. 소지역의 모든 표본에 대한 추정량의 상대평균제곱오차의 평균은

$$\overline{RMSE} = \frac{1}{m} \sum_{i=1}^m RMSE_i, \quad m \text{은 소지역수.}$$

이다. 상대평균제곱오차는 0에 가까울수록 좋으며, 표본에 따른 추정량들의 퍼짐 정도를 판단하기 위한 값이다.

그런데 평균상대편향의 평균,  $\overline{ARB}$ 과 상대평균제곱오차의 평균  $\overline{RMSE}$ 는 일반적인 평균의 특성과 마찬가지로 어떤 지역에서 편향 또

는 산포(dispersion)가 매우 클 때 이 값이 전체 평균에 영향을 줄 수 있다. 이런 경우, 앞에서 언급된 것과 같은 영양군, 태안군 등의 큰 ARB와 RMSE 때문에 전체적으로 소지역 추정 결과에 대한 평가가 왜곡될 수 있다. 그러므로 평가 기준으로  $\overline{ARB}$ ,  $\overline{RMSE}$ 를 추정량의 평가 결과로 해석할 때에는 이 점에 유의해야 한다.

이런 경우, 이상치에 영향을 덜 받고 추정량들을 평가할 수 있도록 본 연구에서는 151개 시군의 추정량들의 결과에 대해 다음과 같은 4개의 평가 방법을 이용해 보기로 한다.

- ① 직접추정량에 비해 다른 추정량의 ARB가 좋은 지역 수
- ② 직접추정량에 비해 다른 추정량의 RMSE가 작은 지역 수
- ③ 각 추정량의 ARB가 가장 좋은 지역 수
- ④ 각 추정량의 RMSE가 가장 작은 지역 수

## 제5 절 모의실험 결과

### 1. 추정량 비교

1,000번의 모의실험을 통해 5개의 추정량  $\hat{Y}_{i,r}^d$ ,  $\hat{Y}_{i,r}^{sm}$ ,  $\hat{Y}_{i,r}^{om}$ ,  $\hat{Y}_{i,r}^{HLP}$ ,  $\hat{Y}_{i,r}^B$ 에 대한 추정 및 분석결과를 비교해보자.

추정량과 그 평가에 대한 보다 구체적인 수치나 현상은 <표 3> ~ <표 11>, <표 12>, [그림 4] ~ [그림 7]을 통하여 직접 확인할 수 있다. <표 3> ~ <표 11>은 각 추정방법들에 의해 계산된 추정치에 대한 ARB와 RMSE를 제시하였다. 결과는 편의상 경기, 강원, 충북, 충남, 전북, 전남, 경북, 경남, 제주의 9개 '도' 단위로 나누고, 각 도 내의 시군별로 정리하였다. <표 12>에서는 <표 3> ~ <표 11>의 결과를 토대로 평가 기준에 따라 상대적으로 우수하게 추정된 지역의 개수를 제시하였다. [그림

4] ~ [그림 7]은 각각 표본조사구 수가 15개 이상인 지역 10~14개와 5~9개, 4개 이하인 지역 중 일부 지역에 대해 추정량들의 분포 현황을 상자 그림(box plot)으로 나타낸 것이다. 이를 통해 모의실험에 대한 다양한 결과를 시각적으로 살펴봄으로써 전체적인 추정량들의 경향을 쉽게 이해할 수 있다.

분석에 앞서 몇 가지 용어에 대한 설명을 하자면, EB1, EB2, EB25, HB1, HB2, HB25는 각각 4절에서 설명된 바 있는 권역에 따라 달리 추정된 추정량들이다. EB1, HB1은 권역 ①을 이용하여 추정된 EBLUP과 HB 추정량, EB2, HB2는 권역 ②(시 vs. 군 모형), EB25, HB25는 권역 ③(2005년 주민등록인구 25만 이상 vs. 미만 모형)을 이용하여 추정된 EBLUP과 HB 추정량이다. 여기서 EB25와 HB25는 다른 모형추정량들과 추정 경향이 크게 다르지 않아 지면상 자세한 추정 결과는 <표 4>에는 제시하지 않고 그림에만 나타내기로 한다.

#### 가. 총체적 비교

결과적으로, 소지역 추정을 위해 합성·복합추정량, 모형추정량의 사용 가능성이 충분히 확인되었다고 볼 수 있다. 즉, 직접추정량보다는 대 지역의 정보를 빌려오거나(borrow strength) 추가의 보조 정보(auxiliary data)를 이용하는 방법이 보다 효율성이 높은 것으로 나타났다. 추정량들간의 비교를 좀 더 쉽게 하기 위해 <표 12>에 각 추정량이 우세한 지역의 수를 정리하였다. 이를 보면, 직접추정량은 ARB가 가장 좋은 반면, 추정값들의 분산(RMSE)이 제일 크다. 이는 직접추정량이 비편향추정량이지만, 표본 수가 작을 경우에는 큰 분산을 가진다는 특성이 그대로 반영된 당연한 결과라 할 수 있다. 그러나 RMSE 측면에서 보면, 합성추정량은 151개 시군 중에서 122개, HB1과 HB2에도 각각 141개, 145개 지역에서 직접추정량보다 작은 RMSE를 갖는다. 합성, 복합, EBLUP, HB 중 RMSE면에서 상대적으로 그 성능이 좋지 않은 EBLUP의 경우도 100개 이상의 지역에서 직접추정량보다 RMSE가 작게 나타났다. 또한 5개의 추정량 중 RMSE가 가장 작은 추정량으로 나타난 지역이 직접추정량은

3개인데 비해 HB1이 46개, EB2가 30개, 복합추정량이 28개 순으로 나타났다. 전체적으로 볼 때 RMSE 측면에서는 모형추정량인 HB1이 그 중 우세하다. 자료 기반인 합성·복합추정량과 모형 기반인 HB와 EB는 ‘시’의 인구 규모가 큰 지역에서는 추정값에 큰 차이는 없지만, 실업자 수가 0으로 자주 관측되는 과소 규모 지역에서는 HB 추정량이 합성·복합추정량보다 덜 편향되거나 RMSE가 약간 더 작은 경향이 있다. 특히 HB2는 다른 추정량들에 비해 ARB와 RMSE 측면에서 많이 개선되었음을 알 수 있다(그림 9참고). 이런 지역에 대해서는 ‘시’단위 지역의 정보까지 이용해서 모형 추정을 하는 것보다는 비슷한 성향의 지역 정보를 이용해 모형추정을 하는 것이 효율적일 것으로 보인다.

자료 기반 추정량과 모형 기반 추정량은 추정에 있어서 서로 다른 특성을 보이는 것으로 나타났다. 즉, 직접추정량의 편향이 심할 경우, 합성·복합추정량은 다른 경우보다 크게 과소 또는 과대 추정되는 경향이 있다. 예를 들어, ‘연기군’은 직접추정치 ARB가 111.87로 참값에 대해 약 12% 정도 과대 추정하고 있다. 이에 반해 합성과 복합추정치의 ARB는 각각 42.80, 43.68로 약 50% 이상 과소 추정하는 것을 알 수 있다. 이에 반해, 이 지역의 EB1과 EB2는 114.36, 91.12, HB1과 HB2의 ARB는 98.6, 76.22로 합성·복합추정량에 비해 ARB가 상당히 개선됨을 알 수 있다. 청송군, 과천시, 홍천군 등에서도 이런 현상이 나타난다. 또한 직접추정량은 상당히 안정적인데 반해, 평택시, 화성시, 양구군, 해남군, 강진군, 나주시, 광양시, 화순군, 상주시, 거제시, 안성시, 서귀포시 등 많은 지역에서 합성·복합추정치가 상당히 편향되어 추정되고 있다. 그리고 이들 지역들에서 HB1 또는 HB2는 합성과 복합추정치들과 그 편향 정도가 비슷하거나 훨씬 낮게 나타난다. 더구나 RMSE는 HB1과 HB2에서 훨씬 작게 나타남을 알 수 있다.

모형추정량인 HB와 EBLUP은 시 단위 지역에서는 상당히 우수한 결과를 보였다. 그러나 군 단위 지역에서의 추정 결과가 일관적이지 않고 다소 혼란스러운 면이 있다. 특히 실업자 수 분포가 희박하고, 조사구 수가 적은 지역에서는 추정치의 분산이 크다는 것이 문제이다. 물론, 이런 지역들에서는 합성·복합추정량도 편향 정도와 분산 정도가 크게 나

타나고 있기는 하다. 그런데 이들 지역을 시군 권역으로 나누거나 인구 규모로 나누어 추정할 경우 모형 추정 결과가 상당히 개선되었다. 특히 HB2의 경우 ARB와 RMSE 측면에 많은 개선이 있었다. 즉, HB1, EB1보다는 HB2, EB2가 더 좋은 결과를 보인다는 것은 지역적 특성을 잘 고려하여 권역을 나눈다면 이런 지역에서는 합성·복합추정량보다는 모형 추정량의 개선 가능성이 높다는 것을 의미한다.

#### 나. 권역별 비교

권역별 추정결과를 보면, 몇 개 지역을 제외하면 시, 군(또는 인구 25만 기준)을 나누어 모형을 추정하는 것보다는 단일 모형으로 추정하는 것이 전체적으로는 약간 더 나은 경향이 있다. 이는 모형 내의 성분들을 추정하기에 2개의 모형보다는 단일 모형이 보다 많은 관측치를 제공하기 때문으로 보인다. 그런데 전체 경향에 비해 시와 군 지역별로는 약간의 차이가 드러난다. ‘시’ 지역은 권역으로 나누면 그렇지 않을 때보다 편향과 RMSE 측면에서 약간 커지긴 하지만, 그 차이는 무시할 만하다. 그러나 ‘군’ 지역은 권역별 추정에서의 추정결과가 더 우세한 것으로 나타났다. <표 3> ~ <표 11>을 보면, 모집단의 실업자 수가 적고 조사구 수도 적은 서천, 홍성, 태안, 당진, 완주, 영덕 등 대부분의 ‘군’ 지역에서 HB1보다는 HB2의 추정치의 편향과 RMSE가 감소함을 알 수 있다.

시와 군지역별 추정 결과에 대해 HB1과 HB2를 비교하여 좀 더 구체적으로 설명해보자. 권역별로 추정하면 전체적 편향은 권역화를 하지 않았을 때에 비해 훨씬 줄어드는 경향이 있다. RMSE 측면에서는 권역화를 했을 때와 하지 않았을 때의 결과가 전체적으로 비슷한 경향을 보였다. 특히 권역화했을 때의 RMSE는 전체 151개 시군에 대해서 그 변동성이 다른 추정량들에 비해 상당히 안정적인 경향을 보이는 특성이 발견되었다. 그런데, 표본 규모별로 보면, 표본 수가 작은 지역들에서는 편향 면에서는 권역화를 했을 때가 그렇지 않았을 때보다 편향이 작지만 참값에 대해 다소 과소 추정하는 경향이 있다. 게다가 표본 수가 매우 작은 경우 즉, 표본조사구 수가 2개 또는 3개인 지역들에서는 권역화

를 하지 않았을 때에 비해 RMSE가 상대적으로 큰 값을 갖는 지역들이 많은 것으로 나타났다. 권역화를 하지 않은 HB1의 RMSE는 몇 개 지역(서천, 영양, 진안, 보성, 의성, 해남군 등)을 제외하면 대체로 안정적인 것을 알 수 있다. 참고로 이들 지역들은 표본조사구가 2개 또는 3개 지역이면서 참값이 매우 작은 값을 갖는 경우에 해당된다.

이처럼 권역화에 따른 HB2는 표본규모가 매우 작은 경우에 대해 편향은 줄일 수 있지만 RMSE가 전반적으로 다소 크게 나타나는 경향을 보인다. 특이한 사실은 표본 수가 적은 군 단위 지역들의 경우 대부분이 농촌 지역들로서 실업자 수가 매우 희박한 지역들이라는 공통성을 갖는다는 것이다. 뿐만 아니라 이들 지역들에서는 보조정보로 사용된 고용보험 수급자 수 또한 지역별 변동이 상당히 있을 수 있다는 특성도 동시에 내포하고 있다. 따라서 이들 표본규모가 매우 작은 지역들에 대해서는 지역적 특성을 고려한 세부적 연구를 통해 추정치의 신뢰도를 확보할 수 있을 것으로 보인다.

수치적 결과를 제시하지는 않았지만, 인구 25만 명 기준으로 권역을 나누었을 때, 25만 명 ‘시’ 단위 지역에서는 모형 추정치의 편향과 RMSE가 보다 개선되는 경향이 있었다. 그런데, 25만 명 미만 지역에서는 ‘군’ 지역과 마찬가지로 크게 개선되지는 않았다. 이는 우리나라 인구규모가 작은 지역의 지역적 특성이 행정구역상으로 분명하게 분리될 수 없음을 의미한다고 볼 수 있다. 오히려, 이들 지역들의 경제활동의 성향이 행정구역 경계를 넘어서 인근 지역과 보다 큰 동일 권역을 형성하는 것으로 보이며 따라서, 서로 유사한 특성을 갖는 이러한 지역들을 권역으로 형성하여 추정에 이용한다면 이들 지역들의 추정치도 개선될 수 있을 것으로 본다.

따라서 비슷한 성향의 지역들에 적합한 동일 모형을 사용하여 보다 모형 적합도를 높일 수 있다는 측면에서 향후 지역별 고용·경제 특성 연구를 통하여 권역별 모형화 작업을 시도해 보는 것도 의미있는 작업이 될 수 있을 것이다.

〈표 3〉 모의실험 추정량 비교(경기)

지역	조사 구수	상대편향( <i>ARB</i> )								상대평균제곱오차( <i>RMSE</i> )							
		직접	합성	복합	EB1	EB2	HB1	HB2	직접	합성	복합	EB1	EB2	HB1	HB2		
수원시	16	97.61	89.13	89.22	87.00	89.85	95.43	92.36	30.19	14.00	13.95	19.42	15.78	21.81	21.29		
성남시	15	100.41	83.73	83.89	89.37	92.32	98.01	94.74	31.93	18.26	18.15	18.39	14.87	22.71	21.51		
의정부시	6	98.33	69.92	69.96	77.84	79.53	84.82	83.48	44.38	30.88	30.84	24.25	21.36	27.93	26.74		
안양시	10	100.91	78.42	78.50	90.40	92.95	95.17	92.47	40.85	22.94	22.87	16.29	11.96	24.14	22.78		
부천시	13	101.25	80.73	80.89	95.66	98.82	101.27	98.11	30.61	20.86	20.73	16.63	13.65	21.69	20.49		
광명시	6	98.20	75.79	75.81	104.41	106.70	98.97	99.03	37.60	25.35	25.33	13.97	10.60	23.66	21.60		
평택시	8	101.56	171.19	171.09	174.63	178.14	138.17	142.80	66.31	73.19	73.09	77.76	79.33	50.82	52.92		
동두천시	4	111.22	52.67	52.67	77.35	77.89	80.00	83.98	56.99	47.62	47.62	26.73	25.96	35.80	32.49		
안산시	9	101.78	116.85	116.79	130.76	134.37	116.97	115.77	47.95	20.43	20.39	36.25	37.29	33.06	30.28		
고양시	13	99.39	112.90	112.80	105.95	109.07	102.47	100.84	32.46	17.10	17.01	17.02	15.16	22.64	21.14		
과천시	3	113.04	153.24	153.23	165.15	165.62	116.85	137.87	118.55	55.38	55.38	75.12	75.39	37.41	53.92		
구리시	4	99.33	97.13	97.13	101.56	102.91	89.31	93.60	62.32	10.06	10.05	13.91	11.51	27.30	24.89		
남양주시	11	101.14	147.13	147.06	157.68	161.00	123.85	127.33	46.44	49.35	49.27	61.02	62.28	35.80	36.80		
오산시	4	100.17	60.39	60.39	89.66	90.80	89.30	91.95	42.69	40.06	40.06	16.43	14.15	25.92	23.14		
시흥시	5	103.25	82.66	82.69	104.05	106.41	101.72	101.33	57.85	19.17	19.15	13.93	10.52	25.64	23.30		
군포시	5	102.20	91.53	91.54	114.87	117.14	104.52	106.03	50.30	12.41	12.40	20.63	19.33	24.18	22.33		

〈표 3〉 모의실험 추정량 비교(경기) (계속)

지역	조사 구수	상대편향(ARB)							상대평균제곱오차(RMSE)							
		직접	합성	복합	EB1	EB2	HB1	HB2	직접	합성	복합	EB1	EB2	HB1	HB2	
의왕시	4	101.32	119.91	119.90	162.80	164.83	124.77	133.50	59.57	23.19	23.19	66.76	67.52	38.69	44.47	
하남시	4	99.30	76.70	76.70	88.27	89.27	88.02	91.50	45.23	24.51	24.51	17.58	15.64	30.47	27.40	
용인시	10	100.53	142.15	141.98	145.57	149.45	119.63	120.01	42.92	44.45	44.28	50.02	51.59	33.82	32.37	
파주시	4	96.21	189.76	189.71	224.74	228.49	181.15	191.31	82.36	91.72	91.67	127.92	129.98	87.62	96.72	
이천시	5	102.28	135.54	135.53	120.41	121.82	105.86	112.39	60.87	38.00	37.98	26.97	26.59	27.43	27.56	
안성시	5	96.24	287.37	287.31	243.59	246.09	179.69	200.75	72.53	189.52	189.47	148.55	150.05	85.36	106.07	
김포시	4	100.33	145.84	145.82	181.56	184.25	144.37	153.83	74.84	48.08	48.06	84.96	86.15	54.65	61.96	
화성시	6	102.04	206.71	206.64	231.11	235.10	163.63	174.11	58.69	108.66	108.59	134.25	136.50	75.64	84.21	
광주시	4	104.45	157.79	157.77	182.75	185.48	146.13	154.95	67.49	59.87	59.85	86.13	87.34	56.66	63.49	
양주시	4	105.23	82.60	82.61	103.89	105.18	97.24	101.29	59.57	19.23	19.23	15.12	13.24	27.30	24.85	
포천시	5	109.89	123.99	123.98	115.57	116.60	98.15	106.25	74.61	26.98	26.98	24.52	24.18	29.53	28.53	
여주군	4	99.21	78.56	81.03	131.12	103.86	102.20	79.61	70.54	41.72	42.62	41.48	44.22	27.14	48.27	
연천군	2	105.61	137.13	137.42	178.59	177.60	126.61	86.99	123.20	72.37	71.88	98.75	101.95	56.00	68.57	
가평군	2	122.45	155.26	156.47	246.65	247.07	169.12	112.59	139.86	89.16	90.52	168.87	174.14	84.88	72.95	
양평군	3	100.01	98.86	99.43	132.34	114.91	100.80	76.96	83.16	44.61	45.00	47.40	41.58	27.77	51.18	

〈표 4〉 모의실험 추정량 비교(강원)

지역	조사 구수	상대편향(ARB)							상대평균제곱오차(RMSE)							
		직접	합성	복합	EB1	EB2	HB1	HB2	직접	합성	복합	EB1	EB2	HB1	HB2	
춘천시	13	101.95	108.88	108.44	57.59	58.22	71.32	71.77	40.84	21.02	21.03	43.34	42.55	35.49	34.14	
원주시	15	100.63	115.84	114.96	92.37	93.73	89.12	91.11	39.96	25.75	25.29	14.52	11.55	23.78	21.44	
강릉시	13	102.00	103.40	103.34	98.38	99.80	96.20	97.83	30.93	18.44	18.14	13.33	10.44	20.73	18.96	
동해시	5	106.90	84.66	84.80	93.65	94.31	90.97	96.71	50.42	21.37	21.27	18.62	17.78	27.68	24.60	
태백시	3	100.95	51.93	52.05	61.79	62.00	61.39	67.75	66.78	48.93	48.82	41.04	40.83	51.72	46.13	
속초시	4	101.49	93.58	93.63	95.79	96.51	83.72	92.40	65.99	17.68	17.71	19.45	19.45	33.68	29.23	
삼척시	4	107.81	126.38	126.39	139.77	140.45	111.81	127.82	90.40	34.45	34.48	49.09	49.32	28.06	38.04	
홍천군	5	81.75	189.46	187.45	175.42	173.13	113.66	74.38	94.89	105.76	103.66	95.27	96.78	26.83	44.01	
횡성군	2	89.52	125.71	125.96	163.16	162.94	111.64	74.98	120.07	45.41	45.59	83.83	87.71	35.91	60.35	
영월군	3	105.28	117.61	117.60	181.04	168.84	121.16	79.03	142.79	39.48	39.75	97.72	89.09	31.42	40.64	
평창군	3	107.72	67.56	68.03	80.43	86.04	68.13	66.94	78.81	38.19	37.82	35.77	40.72	45.97	60.31	
정선군	3	108.43	73.26	73.83	222.72	181.07	158.85	107.13	84.50	34.48	34.20	132.56	107.17	66.13	47.91	
철원군	3	102.94	94.48	94.66	132.37	135.70	95.60	79.08	74.81	28.49	28.38	56.63	64.75	28.68	49.60	
화천군	3	102.92	165.42	165.18	144.64	223.47	88.30	78.92	76.90	81.97	81.68	92.56	282.45	29.46	50.63	
양구군	2	99.83	318.82	318.68	256.22	402.75	139.35	99.72	197.76	238.30	238.12	213.75	563.71	53.11	56.54	
인제군	3	114.32	125.60	125.54	121.33	165.88	81.57	74.75	96.19	45.02	44.89	64.09	155.93	32.92	52.55	
고성군	3	110.17	51.90	52.07	77.57	94.38	60.31	59.88	87.18	50.49	50.33	40.78	59.86	49.96	63.42	
양양군	3	115.43	90.47	90.68	152.73	157.51	103.12	74.63	119.01	28.51	28.49	75.60	85.85	31.52	57.10	

〈표 5〉 모의실험 추정량 비교(충북)

지역	조사 구수	상대편향(ARB)							상대평균제곱오차(RMSE)						
		직접	합성	복합	EB1	EB2	HB1	HB2	직접	합성	복합	EB1	EB2	HB1	HB2
청주시	30	97.50	84.37	87.95	67.80	69.52	90.57	88.53	20.11	21.08	19.36	33.47	31.09	18.94	19.55
충주시	12	100.11	137.47	135.23	78.04	78.91	81.81	83.96	39.66	44.19	42.10	24.93	23.50	27.00	24.36
제천시	8	98.54	119.20	118.78	90.40	91.19	83.02	88.32	49.69	28.04	27.70	18.03	16.76	27.08	23.05
청원군	6	101.87	92.13	92.37	247.74	180.51	174.73	113.01	67.06	31.58	31.22	152.79	139.39	82.08	51.47
보은군	2	95.99	218.63	218.58	220.25	228.81	141.06	87.69	185.82	141.33	141.19	143.83	159.64	52.91	51.06
옥천군	4	93.94	64.31	64.94	108.77	94.35	83.19	62.51	78.66	41.77	41.38	29.81	32.25	26.48	51.74
영동군	3	112.72	171.28	171.04	177.13	169.25	124.86	84.86	132.43	91.77	91.35	95.13	90.52	44.12	59.04
진천군	3	117.06	65.99	66.72	133.56	113.12	101.18	71.55	123.44	40.45	40.20	47.19	41.63	33.10	60.57
괴산군	4	94.23	182.93	182.98	321.58	290.21	213.98	121.28	147.39	104.88	104.89	239.66	212.88	119.63	52.75
음성군	4	106.79	108.12	108.12	274.79	214.70	185.45	118.09	76.57	36.96	36.80	183.39	149.29	95.11	54.45
단양군	4	99.35	85.51	85.57	156.29	150.79	107.40	82.02	76.79	33.01	32.91	75.24	73.32	28.59	46.98

〈표 6〉 모의실험 추정량 비교(충남)

지역	조사 구수	상대편향( <i>ARB</i> )							상대평균제곱오차( <i>RMSE</i> )						
		직접	합성	복합	EB1	EB2	HB1	HB2	직접	합성	복합	EB1	EB2	HB1	HB2
천안시	26	100.89	82.16	84.69	124.17	127.09	105.47	106.14	27.67	24.61	23.10	29.06	28.97	22.02	20.99
공주시	7	98.18	110.15	110.05	86.57	86.98	73.36	81.55	74.74	24.94	25.02	22.62	22.10	32.44	25.81
보령시	7	97.23	82.43	82.54	88.50	88.90	72.51	81.67	73.89	24.56	24.56	22.18	21.69	32.20	24.85
아산시	9	105.70	150.59	149.33	219.60	222.49	156.62	170.14	68.20	59.31	58.17	123.27	124.86	66.87	78.78
서산시	8	99.27	147.87	147.26	167.01	168.21	123.66	138.12	51.95	56.76	56.10	73.58	74.16	33.05	45.16
논산시	6	100.38	99.65	99.65	101.62	102.29	84.23	93.25	65.39	20.69	20.73	19.89	19.41	26.82	23.02
금산군	3	122.86	80.81	81.17	148.70	132.43	111.00	80.68	129.96	37.26	37.23	63.72	54.43	37.67	60.97
연기군	2	111.87	42.80	43.68	114.36	91.12	98.60	76.22	98.87	59.68	58.95	28.19	38.93	36.65	61.49
부여군	4	102.32	159.42	158.87	216.05	189.60	152.85	96.21	117.36	87.07	86.15	129.73	109.57	60.79	48.45
서천군	2	86.82	150.83	150.81	363.84	308.56	247.83	136.19	128.43	79.11	78.79	279.03	234.67	152.47	66.83
청양군	2	87.98	87.51	87.75	104.18	115.94	68.44	48.10	140.36	37.40	37.46	41.18	59.78	40.53	65.79
홍성군	4	111.87	164.39	163.65	260.49	226.80	181.11	115.81	117.58	91.39	90.42	174.60	147.83	89.06	60.09
예산군	3	98.60	94.25	94.44	171.90	145.36	126.46	85.85	95.45	37.70	37.54	83.47	68.42	39.90	49.95
태안군	2	55.17	162.62	162.31	286.57	266.69	187.57	104.64	111.04	89.76	89.36	205.49	189.05	94.25	46.20
당진군	5	100.29	117.86	117.89	245.07	194.81	176.74	105.39	113.23	49.75	49.64	154.10	124.82	82.11	49.58

〈표 7〉 모의실험 추정량 비교(전북)

지역	조사 구수	상대편향(ARB)							상대평균제곱오차(RMSE)							
		직접	합성	복합	EB1	EB2	HB1	HB2	직접	합성	복합	EB1	EB2	HB1	HB2	
전주시	22	97.50	68.97	72.27	60.96	62.47	85.75	83.15	26.36	33.40	30.75	39.85	37.91	24.41	25.20	
군산시	10	99.73	127.69	126.87	103.29	104.85	94.75	97.70	46.07	35.99	35.28	13.91	11.26	24.37	21.94	
익산시	11	100.16	93.92	94.19	90.03	91.73	91.51	91.66	39.74	17.94	18.00	14.99	10.95	23.83	21.78	
정읍시	6	95.40	135.68	135.33	87.37	87.91	75.03	82.91	62.62	43.47	43.10	21.16	20.42	32.01	25.84	
남원시	4	104.29	272.01	271.01	150.00	150.67	115.45	133.40	91.91	179.14	178.11	59.37	59.65	30.15	43.57	
김제시	5	92.43	199.34	198.66	139.43	140.21	105.73	120.60	68.51	105.77	105.08	48.34	48.59	23.24	32.42	
완주군	4	93.80	89.63	90.09	290.29	229.18	209.73	123.95	84.46	46.50	46.06	199.63	162.12	113.86	56.95	
진안군	2	78.86	219.19	219.47	408.81	486.25	240.48	147.12	182.26	172.62	172.63	355.01	483.55	149.88	83.19	
무주군	3	109.06	74.84	75.18	147.71	195.35	92.43	74.22	114.64	48.70	48.37	85.71	181.28	35.20	60.70	
장수군	2	87.74	182.63	182.52	231.93	332.50	130.75	91.40	159.58	128.71	128.14	179.18	394.76	52.51	61.36	
임실군	3	75.39	171.88	171.22	235.05	272.68	142.82	92.09	111.82	119.49	117.75	166.67	230.17	53.09	46.09	
순창군	2	90.65	88.60	89.79	143.28	161.63	88.50	55.47	278.09	50.14	53.00	71.95	103.73	21.32	50.79	
고창군	3	89.33	85.15	86.38	147.10	144.46	99.74	65.32	133.87	48.97	49.28	67.53	68.60	31.26	60.77	
부안군	4	95.15	65.24	66.06	137.16	130.29	97.80	72.71	82.07	49.19	48.60	56.54	53.84	22.22	46.18	

〈표 8〉 모의실험 추정량 비교(전남)

지역	조사 구수	상대편향( <i>ARB</i> )							상대평균제곱오차( <i>RMSE</i> )							
		직접	합성	복합	EB1	EB2	HB1	HB2	직접	합성	복합	EB1	EB2	HB1	HB2	
목포시	14	101.50	116.84	115.96	89.03	90.00	82.54	86.56	51.90	30.17	29.85	17.37	15.54	27.76	24.25	
여수시	16	95.52	60.20	62.09	70.37	71.43	80.36	80.41	33.26	41.84	40.20	31.03	29.40	27.56	26.22	
순천시	13	100.18	85.16	86.15	72.90	73.86	81.93	82.72	35.92	23.52	23.08	28.97	27.49	27.53	25.70	
나주시	5	99.44	228.05	226.87	131.92	131.99	97.11	116.65	88.86	137.27	136.07	47.75	47.46	22.01	30.36	
광양시	6	94.47	219.62	217.55	199.53	201.07	147.87	168.37	83.22	128.57	126.43	105.66	106.54	53.88	73.82	
담양군	2	82.44	76.92	77.10	142.24	132.85	98.97	65.61	97.46	40.53	40.34	60.21	55.26	23.71	51.46	
곡성군	2	106.52	123.56	123.69	110.78	146.55	71.01	54.83	169.36	59.22	59.20	54.48	124.63	52.12	73.60	
구례군	2	90.02	80.25	80.42	87.17	114.37	57.85	46.00	119.82	40.35	40.38	43.52	89.20	56.42	74.72	
고흥군	4	94.66	190.14	190.03	252.80	255.44	163.25	103.63	117.86	123.32	123.03	175.85	184.07	72.55	51.52	
보성군	2	62.25	344.01	344.13	459.23	498.23	277.01	158.62	190.74	287.79	287.82	399.60	460.43	184.42	84.79	
화순군	3	94.69	42.88	43.38	106.44	88.89	84.34	63.98	88.81	60.08	59.66	26.29	34.05	28.70	54.66	
장흥군	2	102.44	95.88	96.24	119.24	141.93	78.07	58.26	141.23	42.75	42.88	54.58	94.62	41.34	67.63	
강진군	2	100.49	152.31	152.43	176.90	206.51	114.10	80.85	136.80	85.00	84.98	106.76	158.33	44.75	66.64	
해남군	4	102.82	357.58	356.89	479.03	462.94	308.35	178.36	155.97	301.84	300.88	408.99	397.81	215.10	105.78	
영암군	2	110.97	88.90	89.27	130.01	131.43	90.04	59.47	203.53	40.06	40.22	53.89	59.63	59.34	82.68	
무안군	2	95.80	76.67	76.86	110.31	117.68	78.17	59.58	126.94	40.47	40.34	42.34	54.98	48.67	76.01	
함평군	2	118.54	116.66	116.69	119.28	156.23	74.79	58.04	203.83	53.30	53.21	60.09	132.74	44.26	74.17	
영광군	3	96.79	144.30	144.22	160.87	171.68	109.17	79.09	98.00	76.53	76.26	85.39	104.40	32.38	51.57	
장성군	2	123.84	64.21	64.62	121.04	118.47	84.34	51.68	291.78	45.27	45.32	43.61	71.28	71.12	90.43	

〈표 9〉 모의실험 추정량 비교(경북)

지역	조사 구수	상대편향(ARB)							상대평균제곱오차(RMSE)							
		직접	합성	복합	EB1	EB2	HB1	HB2	직접	합성	복합	EB1	EB2	HB1	HB2	
포항시	17	101.39	81.07	82.00	88.28	90.23	94.51	93.38	31.39	22.61	22.00	16.25	12.06	21.43	20.11	
경주시	11	99.33	112.25	112.07	122.25	124.18	103.85	107.46	39.24	21.08	20.90	27.22	26.71	22.11	21.52	
김천시	6	103.94	140.47	140.32	134.16	135.45	111.97	121.71	72.26	45.90	45.77	40.44	40.61	26.96	31.60	
안동시	8	101.73	84.97	85.12	55.91	56.35	63.25	65.89	54.48	19.92	19.84	45.20	44.67	43.84	40.54	
구미시	13	102.74	93.96	94.20	171.57	175.96	127.05	129.32	35.18	15.51	15.38	75.39	77.71	37.80	38.63	
영주시	5	103.24	149.16	149.03	110.05	110.64	91.44	102.45	72.70	54.31	54.18	24.66	24.45	24.88	23.52	
영천시	4	95.03	97.65	97.72	112.36	113.29	97.54	107.10	70.59	15.35	15.36	23.06	22.65	26.46	25.43	
상주시	6	102.26	231.60	230.98	155.11	155.99	117.19	133.44	71.95	136.51	135.87	63.69	64.15	30.26	42.74	
문경시	4	97.08	124.61	124.56	162.19	163.22	124.06	141.27	77.45	31.48	31.43	69.62	70.11	37.40	50.94	
경산시	8	103.59	84.65	84.88	102.27	103.94	99.19	101.37	46.39	20.04	19.92	13.28	10.18	23.95	21.61	
의성군	3	106.11	344.70	343.72	464.48	415.86	311.73	174.97	172.31	284.53	283.16	386.93	344.32	217.17	98.73	
청송군	2	88.07	359.66	359.78	265.06	379.89	150.80	103.37	224.35	300.19	300.24	215.52	459.68	75.39	71.81	
영양군	2	144.57	441.22	441.18	424.41	634.73	234.80	165.20	312.86	387.70	387.63	398.26	858.83	146.62	111.99	
영덕군	2	98.59	139.11	139.26	254.55	248.56	169.11	103.16	144.35	69.96	70.00	175.44	173.20	83.15	62.10	
청도군	2	104.80	74.51	74.84	175.25	156.79	124.72	79.11	163.11	40.06	40.10	89.71	76.76	49.45	64.71	
고령군	2	89.32	111.09	111.50	241.63	227.22	160.87	96.46	137.96	47.43	47.89	159.78	148.59	68.45	45.99	
성주군	2	79.21	97.86	98.08	121.73	126.90	83.73	60.01	109.07	40.64	40.56	48.93	59.22	32.88	61.60	
칠곡군	3	108.25	59.09	60.87	189.88	136.85	146.46	96.59	98.12	48.58	47.95	94.24	100.46	58.17	57.47	
예천군	2	113.26	92.97	93.52	80.29	88.56	58.12	45.99	162.23	39.32	39.44	36.88	44.09	58.69	77.89	
봉화군	3	100.65	164.48	164.46	198.88	217.80	127.59	87.76	120.18	94.09	93.89	125.16	156.80	41.34	51.06	
울진군	2	114.29	69.85	70.79	152.70	132.73	117.65	81.31	110.09	41.83	41.31	66.16	55.61	42.94	61.08	

〈표 10〉 모의실험 추정량 비교(경남)

지역	조사 구수	상대편향(ARB)							상대평균제곱오차(RMSE)							
		직접	합성	복합	EB1	EB2	HB1	HB2	직접	합성	복합	EB1	EB2	HB1	HB2	
창원시	16	101.08	101.61	101.60	115.32	118.21	105.91	105.61	31.88	15.51	15.44	21.87	20.99	23.00	21.36	
마산시	13	97.60	78.38	78.91	86.90	88.95	93.05	91.56	29.80	24.70	24.28	17.39	13.31	20.93	20.01	
진주시	10	101.28	103.02	103.02	63.03	64.06	79.55	78.46	37.47	16.04	16.03	37.83	36.39	29.82	29.19	
진해시	4	101.98	113.98	113.94	124.46	125.97	103.03	109.11	58.16	22.35	22.31	30.31	30.09	26.07	26.07	
통영시	4	103.64	125.88	125.83	94.83	95.42	86.81	94.09	57.67	32.33	32.28	19.09	18.39	29.32	25.79	
사천시	4	107.15	98.60	98.64	77.18	77.72	66.75	73.63	106.46	15.42	15.46	27.01	26.28	39.00	32.60	
김해시	10	102.42	100.18	100.23	129.47	132.55	112.82	113.25	44.59	15.32	15.44	34.06	34.39	30.01	28.19	
밀양시	4	108.30	108.95	108.95	75.80	76.29	66.08	73.52	95.61	19.23	19.26	28.20	27.54	43.23	37.31	
거제시	6	107.13	208.33	207.82	151.75	153.30	117.53	128.83	78.55	112.89	112.38	56.86	57.36	31.65	39.30	
양산시	6	101.29	85.50	85.58	119.77	121.91	107.37	110.07	43.22	19.54	19.47	24.81	24.09	25.30	24.45	
의령군	2	108.61	95.87	96.34	212.42	190.07	150.73	94.81	159.23	44.06	44.22	127.06	109.68	63.38	60.66	
함안군	2	112.08	61.18	61.96	191.08	154.71	143.88	93.47	119.61	47.56	47.11	100.55	81.65	58.50	60.71	
창녕군	2	87.35	112.25	112.40	170.79	158.05	117.31	73.57	137.41	52.31	52.22	87.13	78.31	31.14	48.73	
고성군	3	94.58	99.99	100.25	166.00	147.78	118.86	80.81	104.19	45.16	44.94	80.36	68.24	34.63	51.76	
남해군	2	91.96	154.76	154.54	290.27	281.27	188.16	112.20	151.58	89.29	88.89	212.00	206.72	96.67	57.99	
하동군	2	99.78	59.04	59.99	85.05	81.94	68.14	55.74	107.56	48.96	48.43	30.91	33.74	49.95	71.06	
산청군	2	97.02	107.84	107.81	145.23	160.57	91.00	58.77	212.19	50.30	50.22	72.56	99.00	32.76	64.51	
함양군	2	87.05	136.79	137.71	149.79	168.97	92.71	57.56	227.76	72.67	73.83	78.04	111.19	19.20	48.32	
거창군	3	98.78	182.92	182.61	302.92	271.21	206.88	121.18	145.87	117.02	116.47	219.89	193.15	113.39	60.53	
합천군	2	113.30	137.29	137.47	233.67	209.04	162.22	97.95	207.81	73.97	73.81	148.69	129.01	76.13	69.81	



#### 다. 표본규모별 비교

이제 표본규모별로 소지역 추정량들을 비교해보자. 151개 전체 지역을 그림으로 제시할 수 없기 때문에 표본 규모별로 2개씩 지역을 임의로 선택하여 [그림 4] ~ [그림 7]에 제시하였다. [그림 4]는 조사구 수가 15개 이상인 원주시와 포항시, [그림 5]는 조사구 수가 10~14개인 안양시, 마산시, [그림 6]은 조사구 수가 5~9개인 의정부시, 정읍시, [그림 7]은 조사구 수가 4개 이하인 옥천군, 무안군의 각 추정방법에 의한 추정값 분포를 상자그림으로 나타낸 것이다. 그림에서 점선은 해당 지역의 모집단의 실업자 수(참값)를 나타낸 것이고, 상자그림의 상자 부분에는 1000개 추정값의 50%가 포함되어 있다.

전체적으로 표본규모와 상관없이 직접추정량은 비편향이고 다른 추정량들에 비해 큰 분산을 갖는다. 그리고 대체로 직접추정량은 참값을 포함하고 표본규모가 작아질수록 편향되는 경향이 있다. 이를 통해 직접추정량이 큰 변동을 갖는다는 것을 시각적으로 확인할 수 있다. 상자안에 참값을 포함하지 않는 경우는 순창군, 보성군, 장성군, 청송군, 영양군, 함양군처럼 규모가 작으면서 실업자가 포함된 조사구 수도 적은 지역이다. 이들 지역에서 직접추정값의 분포가 참값을 포함하지 않는 이유는 표본 내에 실업자가 아예 관측되지 않아 1,000번의 모의실험에서 추정값이 0값을 가지는 경우가 500번을 넘기 때문이다(<표 12> 참고).

합성과 복합추정량은 그 추정값이 직접추정치에 비해 다소 편향되고, 분산은 직접추정량에 비해 상당히 줄어든 것을 알 수 있다. 이는 이들 추정량들이 대지역의 표본 정보를 가져다 씌으로써 표본 수를 늘리는 것과 같은 효과를 가지기 때문이다. 그러나 복합추정량은 경우 직접추정량과 합성추정량의 단점들을 보완하기 위한 추정량으로써 이들의 가중평균 형태지만, 표본이 작은 소지역 추정에서는 상대적으로 합성추정량에 더 많은 가중을 주게 되어 있다. 따라서 복합추정량의 분포 형태가 합성추정량의 그것과 크게 다르지 않음을 알 수 있다. 또한 모든 지

역에서 그런 것은 아니지만 최소한 그림에 나타난 지역들에서는 합성과 복합추정량은 참값을 대부분 포함하지 못하는 경향이 있고, 실제로 <표 3> ~ <표 11>의 결과에서도 표본규모가 작은 지역에서 이들 자료 기반 추정량들은 다소 불안정한 결과를 보인다는 것을 알 수 있다.

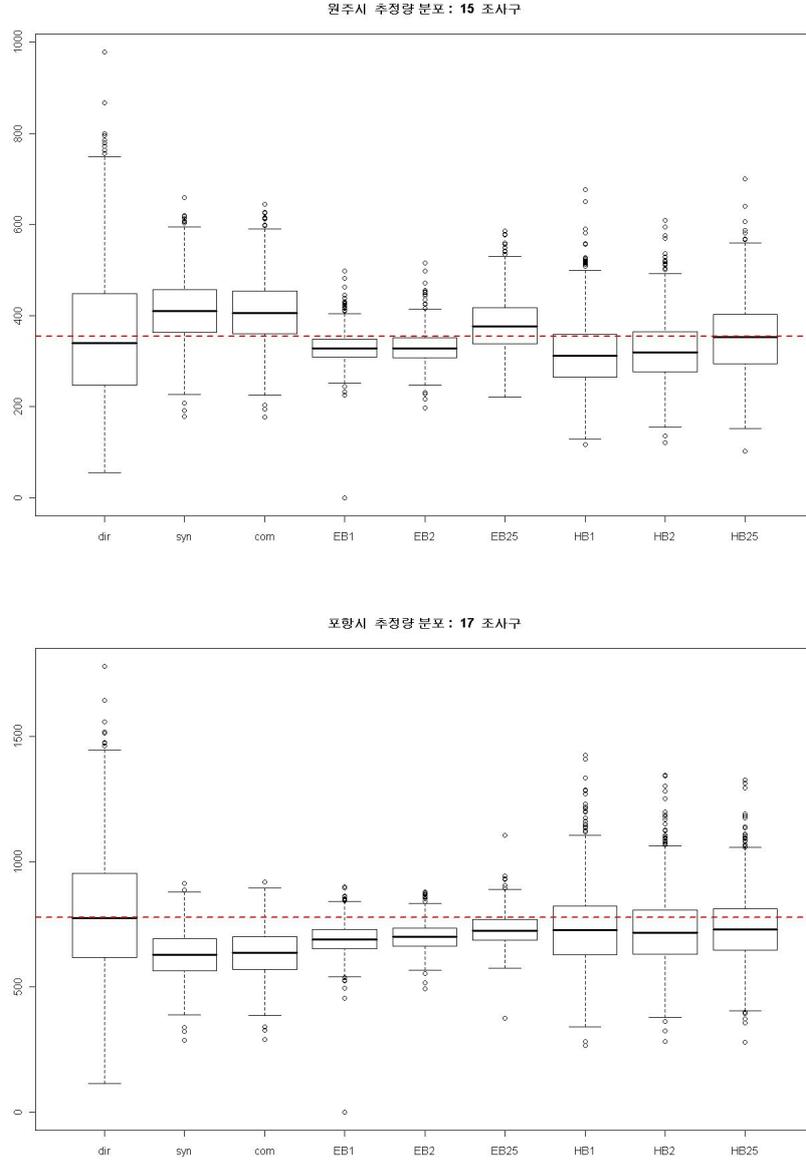
모형추정량은 직접추정량보다는 편향되었지만, 합성과 복합추정량에 비해서는 덜 편향되었다. 추정치의 분산은 직접추정치보다는 당연히 줄어들고, 합성과 복합추정량에 비해서는 표본규모가 큰 지역에서는 유사하거나 다소 큰 반면, 표본 규모가 작은 지역에서는 작은 경향이 있다. 이는 이미 뽑힌 표본을 상대로 소지역 추정을 시도할 경우, 직접추정량보다는 상대적으로 변이가 작기 때문에 해당 추정량을 신뢰할 수 있는 근거가 될 수 있다. 모형추정량의 경우 권역별 추정이 가능한 것처럼 다양한 모형의 적용이 가능하다는 장점<sup>1)</sup>이 있기 때문에 소지역에 적합한 추정량을 찾는 시도를 보다 다각적으로 해볼 수 있다. 현재 사용한 실업급여 수급자 수와 같은 보조정보 외에도 대지역의 경제활동인구조사 결과, 공간통계 등을 보조 정보에 대입해 봄으로써 최선의 추정량을 탐색해 볼 수 있다.

---

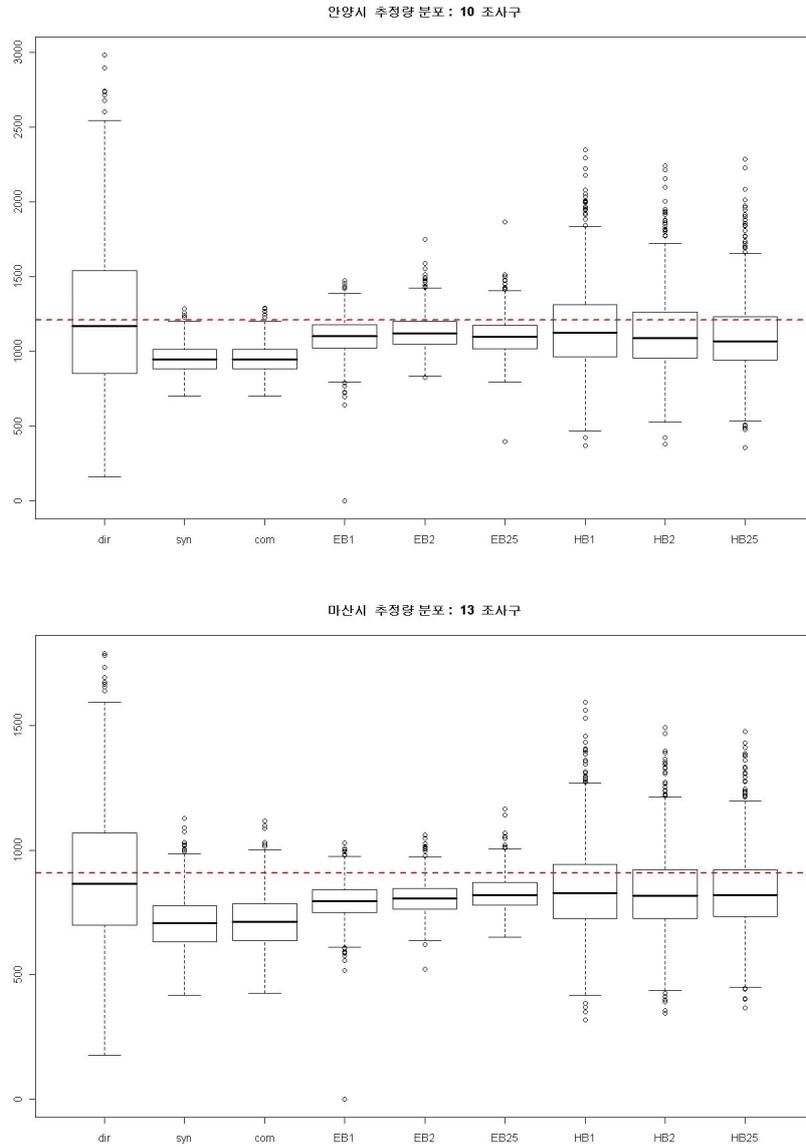
1) Rao(2000)는 그의 연구에서 소지역 추정을 위한 모형 기반 추정방법의 장점으로 다음 네 가지를 기술하였다.

- ① 지역 특성을 모형에 반영할 수 있다.
- ② 모형의 타당성을 표본 자료로부터 확인할 수 있다.
- ③ 횡단면 혹은 시계열 자료같은 복잡한 경우를 다룰 수 있다.
- ④ 추정량의 변이에 대하여 안정적인 지역별 추정량을 구할 수 있다.

42 총조사 자료를 이용한 소지역 추정량 비교

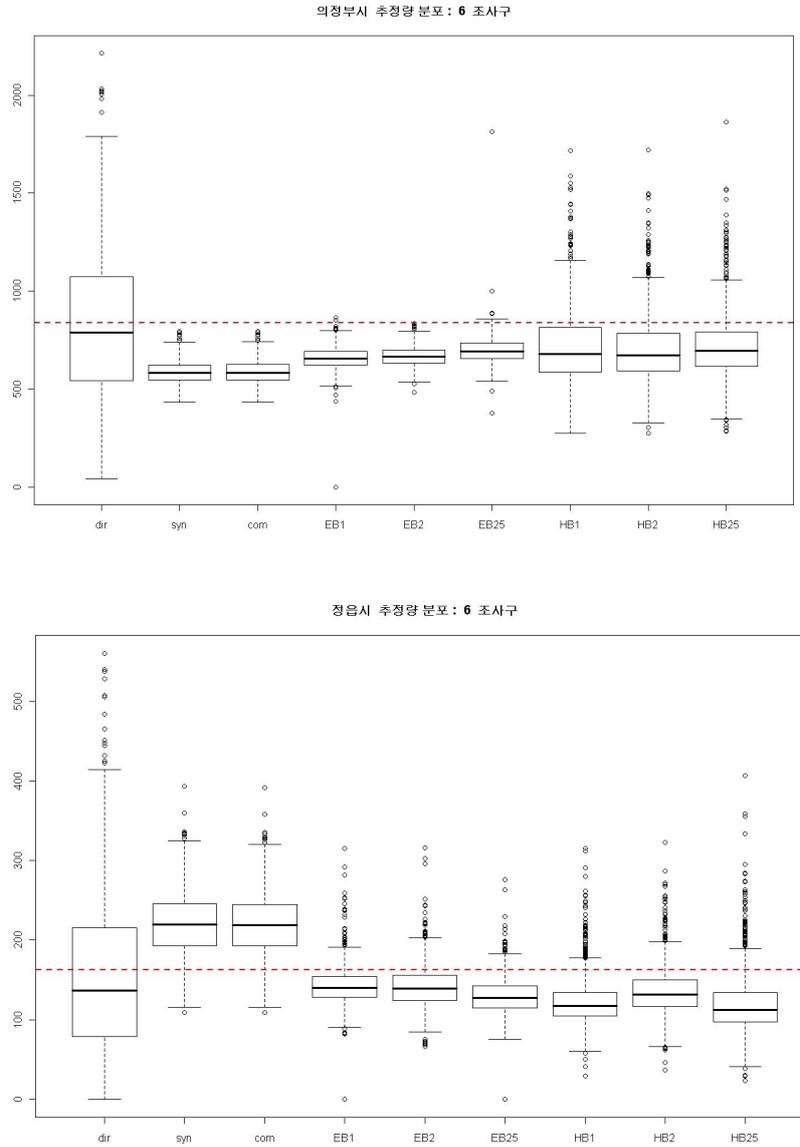


[그림 4] 모의실험 추정량 분포(조사구 수 15개 이상 지역)

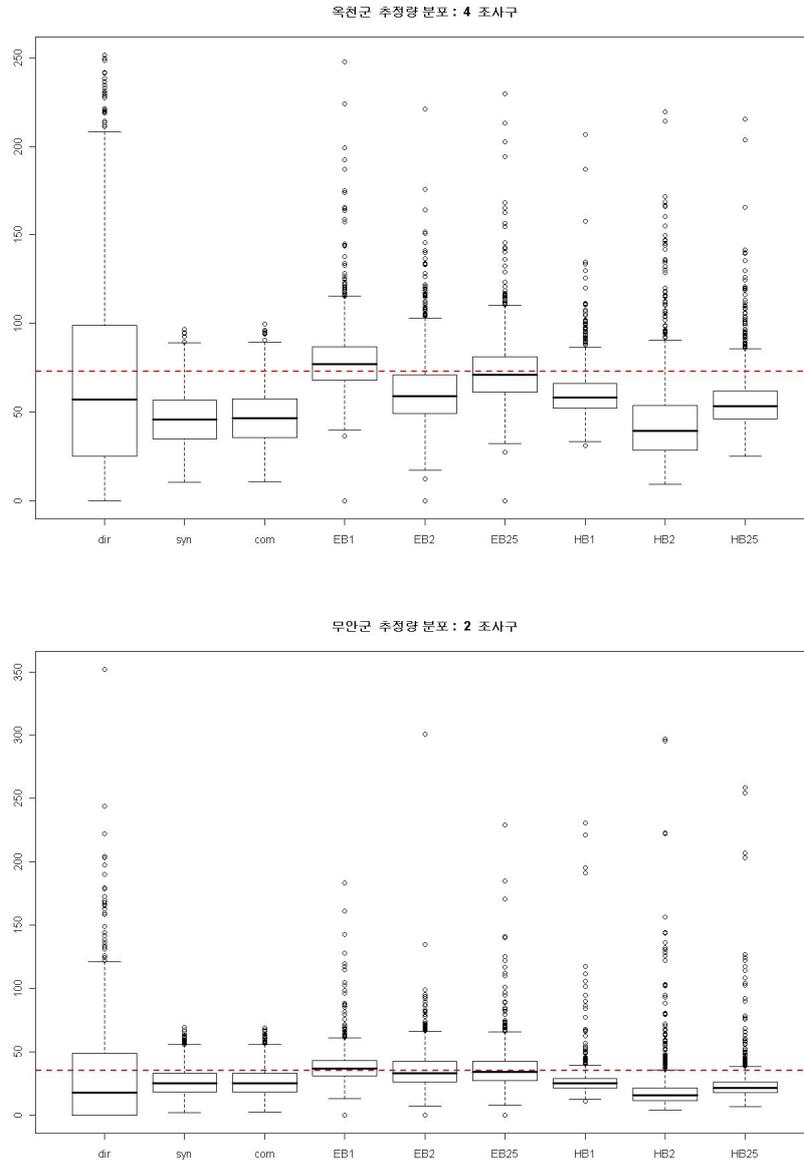


[그림 5] 모의실험 추정량 분포(조사구 수 10~14개 지역)

44 총조사 자료를 이용한 소지역 추정량 비교



[그림 6] 모의실험 추정량 분포(조사구 수 5~9개 지역)



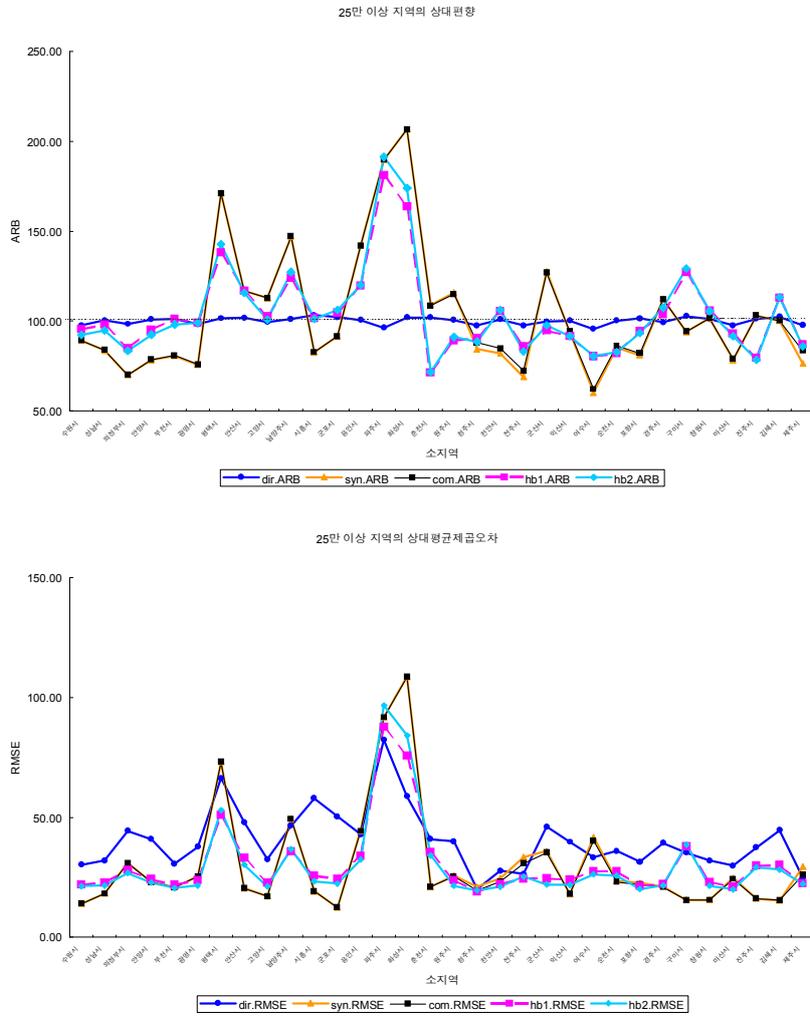
[그림 7] 모의실험 추정량 분포(조사구 수 4개 이하 지역)

## 2. 모의실험을 통해 발견한 점

지금까지의 모의실험 결과를 통해 몇 가지 중요한 사실을 발견할 수 있다. 이는 우리나라와 같은 환경에서 소지역별 추정을 하는 데 반드시 고려되어야 할 사항이라 생각된다. 중요한 점은 실업자를 추정하는 데 있어서 우리나라 조사구에 포함된 실업자 분포는 상당히 변동이 심하다는 것이다. 이는 경찰조사의 표본 선정에도 영향을 미치는 요소이며, 만약 실업자가 아주 적거나 전혀 없는 조사구가 표본조사구로 선정되었을 경우 그 지역의 실업자는 매우 과소 또는 실업자가 전혀 없는 것으로 추정될 수 있다. 물론 이러한 특성까지 경찰조사 표본 조사구 선정에 반영할 수는 없지만, 최소한 추정하는 데 있어서 또는 결과를 이해하는 데 있어서 이런 특성은 충분히 검토되고 반영되어야 한다.

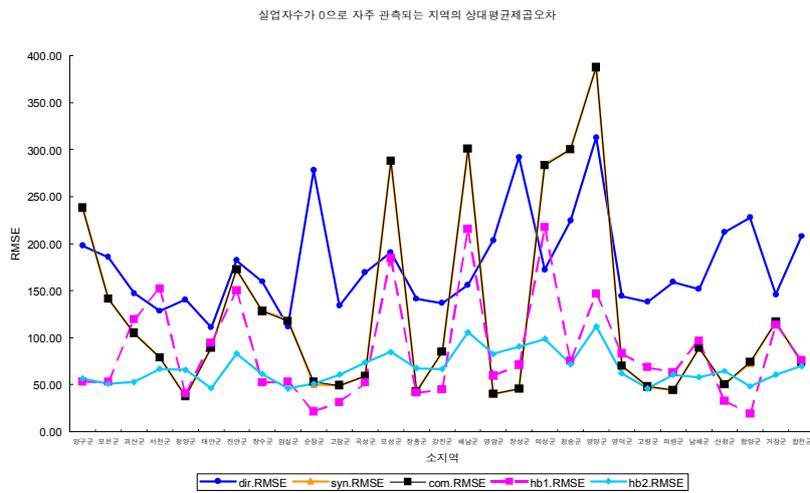
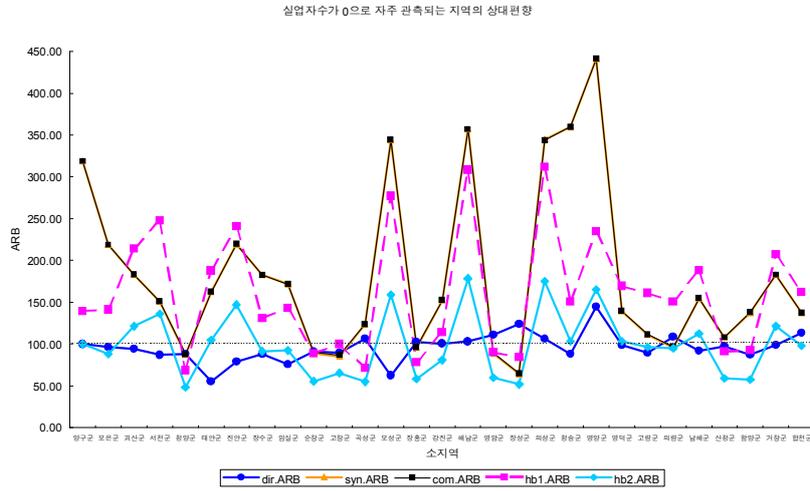
우선 소지역 추정은 표본 수가 많고 특성값이 고루 분포되어 있는 지역에서의 추정은 그렇지 않은 지역에 비해 보다 효율적인 것으로 나타났다. [그림 8]과 [그림 9]는 각각 주민등록인구 25만 이상인 지역의 추정량들과 1,000번의 실험 중 실업자 수가 0으로 관측된 횟수가 적어도 500번 이상인 지역들의 ARB와 RMSE를 나타낸 것이다(<표 13> 참고).

주민등록인구 25만 이상 지역(32개)은 모두 ‘시’ 지역으로 적어도 해당 지역 내에서는 거점 도시의 역할을 하고 있는 곳들이며, 표본조사구 수가 제일 많은 도시는 청주로 30개의 조사구를 가지고 있으며, 가장 작은 도시는 4개가 배정되어 있는 경기도 파주로, 평균적으로 12.6개의 표본조사구를 가지고 있다. 해당 지역들의 경우 각 추정량들의 ARB가 비슷한 정도를 보이고 있으며 RMSE도 크게 흔들림이 없음을 알 수 있다. 그러나 평택, 파주, 화성처럼 표본조사구 수가 작은 지역은 다른 지역들에 비해 상대적으로 큰 ARB와 RMSE를 가지는 것을 알 수 있다. 그럼에도 이들 지역이 [그림 9]의 실업자 수가 0으로 자주 관측되는 지역에 비해 나은 ARB와 RMSE를 가지는 것은 상대적으로 특성값들이 지역 내 조사구들에 골고루 분포되어 있기 때문으로 보인다.



[그림 8] 주민등록인구 25만 이상 지역의 ARB와 RMSE

48 총조사 자료를 이용한 소지역 추정량 비교



[그림 9] 주민등록인구가 25만 이하이면서 실업자 수가 0으로 자주 관측되는 지역의 ARB와 RMSE

실업자 수가 0으로 자주 관측되는 지역들(29개)은 모두 군 지역으로 표본조사구 수는 제일 많은 지역이 4개, 제일 적은 지역이 2개로 평균적으로 2.3개의 표본조사구를 가지고 있다. 이들 지역의 특성은 직접추정량이 다른 지역들에 비해 편향된 지역들이다. 특히 태안, 진안, 임실, 보성 등의 지역에서 그 경향이 심한데, 이는 기본적으로 표본조사구 수가 상당히 작아 특성값이 관측되지 않은 경우가 많이 발생되었기 때문이다. 즉 1,000개의 직접추정량 중에서 다수가 실업자 수를 0으로 관측하였기 때문에 모집단의 실업자 수에 비해 과소하게 추정된다.

이렇듯 표본 수가 너무 작은 지역의 경우, 집락화 등을 통해 지역특성이 유사한 지역을 묶어서 모집단 및 표본의 규모를 늘려줌으로써 추정 능력을 향상시킬 수 있을 것으로 본다. 한편, 이들 지역의 경우에는 RMSE도 상당히 크다. 영양군의 경우 각 추정량들의 ARB와 RMSE가 모두 다른 지역에 비해 상당히 큰 편이다. [그림 2]의 산점도를 통해 살펴보면 영양군은 전체 24개 조사구 가운데 단 3개의 조사구에서만 실업자 수가 관측되고 있으며, 그 중에서 2개의 조사구만을 뽑아 추정에 이용한다. 이들 같이 모수값을 구성하는 특성값들이 일부 조사구에만 편중되어 있는 경우, 즉, 모집단 규모가 작고 표본이 적게 할당되었으면서 특정 경제활동 상태로 구분이 가능한 지역들에 대해서는 단순 모형의 적용보다는 지역 특성에 맞는 보조 정보를 찾아내서 모형을 구축하는 작업이 필요할 것으로 보인다.

실업자 수가 0으로 자주 관측된 지역의 특성을 살펴보면, 표본조사구 수가 작은 지역일수록 실업자 수가 한 명도 관측되지 않는 경우가 많다. <표 13>을 보면 1,000개 중 적어도 200개 이상의 표본에서 실업자 수가 한 명도 관측되지 않은 지역은 56개(37%)이다. 물론 이들 지역에서 실제 실업자 수가 0인 지역은 없다. 따라서 소지역의 작은 표본으로 인한 큰 변이성 때문에 직접추정량을 이용하여 실업자 수를 추정하는 것은 심각한 편향 문제를 발생시킨다. 이 경우, 다른 추정량을 계산하는데도 문제가 발생하는데, 직접추정량의 실업자 수가 0으로 추정되는 경우 해당 지역에서의 추정량에 대한 MSE를 구할 수 없다. 따라서

〈표 13〉 실업자 수가 관측되지 않은 횟수 및 지역 분포

실업자 수가 관측되지 않은 횟수	지역 수	비중	지역(조사구 수)
0	35	23.18	수원(16), ..., 용인시(10), 춘천시(13), ..., 강릉시(13), 청주시(30), 충주시(12), 천안시(26), 전주시(22), ..., 익산시(11), 목포시(14), ..., 순천시(13), 포항시(17), ..., 경산시(8), 창원시(16), ..., 김해시(10), 제주시(20), 서귀포시(25)
1~100	44	29.14	평택시(8), ..., 여주군(4), 동해시(5), 태백시(3) ..., 고성군(3), 제천시(8), ..., 단양군(4), 공주시(7), ..., 논산시(6), 정읍시(6), 김제시(5), 김천시(6), ..., 칠곡군(3), 진해시(4), ..., 양산시(6)
101~200	16	10.60	양평군(3), 삼척시(4), 정선군(3), 화천군(3), 인제군(3), 옥천군(4), 진천군(3), 금산군(3), 연기군(2), 예산군(3), 남원시(4), 부안군(4), 나주시(5), 광양시(6), 화순군(3), 영천시(4)
201~300	11	7.28	홍천군(5), 영월군(3), 양양군(3), 홍성군(2), 당진군(5), 완주군(4), 무주군(3), 울진군(2), 함안군(2), 고성군(3), 하동군(2)
301~400	5	3.31	영동군(3), 부여군(4), 담양군(2), 무안군(2), 영광군(3)
401~500	11	7.28	연천군(2), 가평군(2), 횡성군(2), 구례군(2), 고흥군(4), 함평군(2), 청도군(2), 성주군(2), 예천군(2), 봉화군(3), 창녕군(2)
501~600	16	10.60	괴산군(4), 서천군(2), 청양군(2), 임실군(3), 고창군(3), 곡성군(2), 장흥군(2), 강진군(2), 해남군(4), 의성군(3), 영덕군(2), 고령군(2), 의령군(2), 남해군(2), 거창군(3), 합천군(2)
601~700	6	3.97	양구군(2), 보은군(2), 태안군(2), 장수군(2), 영암군(2), 산청군(2)
701~800	4	2.65	진안군(2), 순창군(2), 장성군(2), 영양군(2)
801~900	3	1.99	보성군(2), 청송군(2), 함양군(2)
계	151	100	

이러한 관측값을 이용하여 추정되는 합성·복합추정량 및 모형추정량들의 추정값과 MSE를 계산하는 작업이 어렵기 때문에 소지역 추정에 많은 문제를 야기한다. 이와 같은 문제는 경제활동인구조사와 같이 한번 뽑힌 표본으로 지속 조사를 하는 경우에도 필연적으로 발생하는 문제로 보이며 그에 대한 대책이 필요하다.

이와 관련 통계청(2002)은 소지역 추정에 있어서 표본 수가 작은 지역에 대해 집락화를 시도한 바 있다. 본 연구의 결과를 통해 보더라도 행정구역상의 소지역뿐만 아니라 유사 특성을 가지는 경제권역, 노동시장권역, 산업권역 등을 고려한 추정도 고려해볼 만하다. 이와 같은 지역권역화 연구는 향후 우리나라 소지역 추정 연구에 시사하는 바가 클 것이다.

## 제6절 결론

지금까지 모의실험을 통해 소지역 추정량들을 비교해 보았다. 본 절에서는 모의실험 결과를 종합적으로 요약하고, 연구를 통해 발견된 여러 가지 현실적인 사안들에 대한 해결방안을 모색하고자 한다. 그리고 소지역별 실업자 추정을 위한 최적의 방법을 선택하기까지 진행되어야 할 연구과제들을 중심으로 논의하고자 한다.

### 1. 요약

다양한 소지역 추정 방법에 대한 모의실험 결과를 정리하면 다음과 같다. 추정방법은 직접추정량, 합성과 복합추정량의 자료 기반 추정량, EBLUP과 HB의 모형 기반 추정량 등 5가지 방법을 고려하고, 이들 추정량들에 대한 성능은 ARB와 RMSE를 통해 평가하였다.

그 결과 소지역 통계를 위해 직접추정량보다는 합성·복합추정량, 모형추정량 등을 사용하는 것이 효율성 측면에서 보다 나은 것으로 나

타났다. 일반적으로 직접추정량은 비편향추정량이지만, 인구 및 표본 규모가 많이 작은 지역에서는 직접추정량이 오히려 ARB 측면에서도 향상된 면을 보이지 않는 경우도 있었다. RMSE 측면에서는 직접추정량을 제외한 나머지 추정량들이 월등히 나은 결과를 보였다.

모의실험 결과에 따르면 자료 기반 추정량과 모형 기반 추정량 중, 어떤 추정량이 월등히 낫다고 단언하기는 어렵지만 모형 기반 추정량이 다소 우세한 것을 알 수 있다. 확실한 것은 인구 규모가 어느 정도 이상인 ‘시’ 이상 지역에서는 모형추정량이 편향과 효율성 측면에서 자료 기반 추정량보다 낫다는 것이다. 그리고 인구규모 또는 실업자 수가 적은 지역, 즉, ‘군’ 단위 지역에서는 모든 추정량들의 효율성이 떨어지는데, 합성과 복합추정량이 상대적으로 더 떨어진다. 이는 이들 추정량들이 그 소지역을 포함한 대지역의 실업자 수를 소지역의 인구에 비례해서 배분하는 방식을 사용하는 데서 기인한다 할 수 있다. 이로부터 우리나라와 같이 소지역별로 변동성이 큰 지역들의 특성이 대지역의 특성과 유사하지 않다는 것을 짐작할 수 있다. 즉, 소지역의 인구규모 및 실업 특성이 대지역과 맞지 않은 경우에는 자료 기반 추정량은 상당한 편향을 초래한다. 실제로 모의실험 결과에서도 이와 같은 현상을 확인할 수 있었으며 특히 인구규모가 큰 지역에서도 직접추정치에 비해 편향되는 경향이 큰 지역들이 있다는 것은 우려스러운 부분이다. 그러나 이런 지역들을 제외한 대도시 지역에서의 합성과 복합추정량은 소지역 추정량으로서 사용할 만하다.

HB 추정량은 모형추정량 중에서는 가장 좋은 추정량으로, ‘시’ 이상 지역에서 그 추정치의 효율성이 높다. 뿐만 아니라, 추정치의 효율성이 떨어지는 ‘군’ 지역에서는 소지역들을 유사한 특성을 갖는 권역들로 묶어서 추정함으로써 추정치를 개선할 수 있다는 점은 상당히 고무적인 측면이라 하겠다. 또한, 모형 기반 추정량의 경우 인구학적 보조 정보만을 빌려다 쓰는 자료 기반 추정량에 비해 보조 정보의 탐색·적용 가능성이 높고, 권역별 모형 적용 등 다양한 시도를 모형 내에서 적용해 볼 수 있다는 점에서 상당히 긍정적이라고 판단된다.

본 모의실험 연구는 소지역 추정 연구를 진행하는 데 상당한 의미가 있었다. 몇 가지 발견한 사실은 실업자 수처럼 관측이 쉽지 않은 특성값의 경우 0으로 관측되는 경우가 많기 때문에 실제값에 비해 과소 또는 과대 추정될 수 있는 위험이 크다는 것이다. 이런 경우 실업자 수에 대한 MSE 계산도 불가능하며 이는 표본조사구가 1개인 경우에도 마찬가지로 발생하는 현상이었다. 이러한 현상은 ‘군’ 단위의 모형추정량에 적지 않은 영향을 끼쳤을 것으로 보인다. 왜냐하면, Fay-Herriot 모형에서 지역효과 변동항( $v_j$ )에 대해 평균이 0이고 분산은  $\sigma_v^2$ 으로 동일하다고 가정하였기 때문에 변동이 큰 지역들의 특성을 충분히 반영할 수 없었을 것으로 보인다. 뿐만 아니라 이 연구에서는 지역변동항을  $b_j=1$ 로 고정하였기 때문에 변동이 큰 지역의 특성이 전체 지역의 특성을 축소시키는 경향이 있었던 것으로 판단된다. 이는 향후 적절한 지역별  $b_j$ 값을 찾음으로써 추정치를 개선할 수 있을 것으로 보인다. 또한 해당 소지역의 시계열 자료를 활용하는 것도 이러한 단점을 보완할 수 있는 방법이 될 것이다.

또한 다른 지역이나 보조 정보를 이용하는 것이 무조건 효율적이지는 않았다. 대지역과 소지역 간 특성이 현저히 다르거나, 보조 정보의 소지역에 대한 설명력이 떨어지는 경우에는 오히려 추정이 더 좋지 않은 결과를 보여주기도 하였다. 이런 경우에는 Time series & Cross-sectional 모형이나 Time series 모형과 같이 동일 지역 내에서 조사를 지속하면서 발생하는 시계열 정보를 이용하는 것이 보다 효율적일 수 있다.

## 2. 논의 및 향후 연구

모의실험 결과는 그동안 소지역 추정 연구를 진행하면서 가질 수밖에 없었던, 전통적인 표본을 이용한 추정이 아닌 보조 정보와 모형을 이용한 소지역 추정 결과가 과연 믿을 만한지에 대한 의구심을 해소시켜

줄 수 있었다. 즉, 추정된 소지역 실업자 수가 얼마나 모수값에 근접한 지 혹은 모수값을 중심으로 어떻게 분포되어 있는지 등을 자료로부터 직접적으로 확인할 수 있었고, 이를 통해 이들 추정량의 활용 가능성을 증명시켜주었다. 그렇지만 최종 추정방법을 선택하기까지 추가적으로 검토되어야 할 사항은 여전히 많이 남아있다. 이를 위해 향후 지속되어야 할 소지역 추정 과제들을 언급하고자 한다.

첫째, 보다 세부적이면서도 현실적인 문제점들에 대한 해결 방안이 연구되어야 한다. 특히 소지역 추정에서 아주 작은 값 또는 결측값(missing value) 처리에 대한 연구가 필요하다. 실업자 수와 같이 특성치가 낮은 빈도로 조사되는 경우 소지역 추정은 특정 표본 값에 크게 영향을 받게 된다. 또한 어떤 경우에는 실업자가 한 명도 관측되지 않을 수 있다. 그렇게 되면 소지역의 직접추정값은 '0'이 된다. 이 경우 인근 지역 정보를 이용해 보정을 시도해 볼 수도 있으나, 결국 오차의 정도를 제시할 수 없는 결정적 단점이 있기 때문에 추정량으로써의 쓸모는 없게 된다.

둘째, 소지역들의 지역적 특성 연구가 필요하다. 특성이 다른 소지역들을 하나의 모형으로 추정하는 것보다는 특성이 유사한 지역들을 집락화하여 규모를 키우거나 권역화를 이용한 모형 추정을 시도해보면 추정량의 정도를 높일 수 있을 것으로 기대된다.

셋째, 최선의 보조 정보 탐색이 병행되어야 한다. 모형추정량의 경우, 모형 내의 구성 성분의 다양한 변화를 통해 여러 가지 추정량을 얻어낼 수 있지만, 그러기 위해선 보다 설명력 있는 보조 정보를 찾아내는 노력이 필수적이다.

넷째, 실제 자료를 이용한 지속적인 추정 연구가 필요하다. 본 연구는 모의실험의 특성상 모수를 알고 있는 상태에서 반복 표집을 통해 얻어진 추정량들의 효율성을 모수와 직접 비교해 볼 수 있었으나 실제 조사 자료를 이용한 연구에서는 이러한 인공성이 제거될 수밖에 없다. 다행히도 고용통계 분야에서는 통계청에서 2008년 10월에 대규모 시군구 고용통계조사를 실시하였다. 이 조사 결과를 gold standard로 이용해서

추정량들의 효용성을 판단해 본다면 좋은 결과를 얻을 수 있는 기회가 될 것으로 보인다.

마지막으로, 가장 기본적이면서 연구를 위한 시간과 노력이 많이 필요한 과제는 보다 효과적이며 세련된 모형을 찾는 것이다. 모형추정량의 경우 보조정보, 지역 변이, 시계열 자료 이용 등의 다양한 방법을 시도해 볼 수 있다. 일련의 이런 시도들을 통해 소지역에 보다 적합한 모형을 개발한다면 소지역 추정 연구에 가장 튼튼한 기본이 만들어지는 것이기 때문이다.

본 연구는 모의실험이라는 점에서 실제 조사 자료와 완전히 동일한 표본이나 관측값을 사용한 것은 아니지만, 발생가능한 다양한 표본을 최대한 고려한 객관적인 실험이라는 점에서 높이 평가할 만하다. 더구나 반복 실험을 통해 소지역 추정량들이 보여주는 성능(performance)을 실증적으로 확인할 수 있는 좋은 기회가 되었다. 어떤 특정한 추정량이 가장 좋다고 말할 수는 없지만, 이러한 연구를 통해 지역의 집락화 필요성, 권역별 분석, 실업자 수가 관측되지 않을 경우의 실제적 확인 그리고 그로 인한 과소 추정 가능성 등을 확인할 수 있었으며, 이러한 작업에서 소지역 추정을 위해 향후 고려해야 할 많은 시사점 및 과제를 얻을 수 있었다.

본 연구에서 보여준 다양한 소지역 추정량들에 대한 평가가 이후 연구의 기초자료로써 활용되어 소지역 추정 분야에 있어 일보 전진이 있기를 기대한다.

## 참고문헌

- 김달호(2005), "R과 Winbugs를 이용한 베이지안 통계분석", 자유아카데미.
- 김서영, 권순필(2009), "모형 기반 소지역 실업자 수 추정", 통계개발원.
- 김정숙, 황희진, 신기일(2008), "이웃정보시스템을 이용한 공간 소지역 추정량 비교", 응용통계연구, 21, 855-866.
- 박종태, 이상은. (2001), "소지역 추정법에 관한 비교연구", 한국데이터정보과학회지, Vol. 12, No. 2, 47-55.
- 이강석, 신기일(2008), "격자자료분석을 위한 이웃정보시스템의 비교, 응용통계연구, 21, 387-397.
- 이상은, 진영. (1996), "변동계수를 이용한 소지역 통계의 안정성 검토", 통계분석연구, 창간호, 23-40.
- 이상은(2006), "공간통계량을 활용한 베이지안 자기포아송 모형을 이용한 소지역 통계, 응용통계연구, 19, 421-430.
- 통계청(2002), "소지역 추정법에 의한 시·군·구 실업통계 개발", 용역 연구 보고서.
- \_\_\_\_\_ (2005), "2005 인구주택총조사 조사지침서 표본조사용(종합편)"(내부자료).
- \_\_\_\_\_ (2006). "소지역 실업통계 표본을 위한 표본규모 검토", 통계연구 발표 자료집, 108-140.
- \_\_\_\_\_ (2007), "가구부문 표본개편 결과"(내부자료).
- 허문열 외 4인(2007), "R과 통계계산", 박영사.
- Boonstra, H.J.(2007), "A comparison of several design-based and model-based estimators for municipal unemployment rates", Proceedings of the Section on Small Area Estimation, IASS Satellite Conference.
- Fabbris, L., Russo, A., and Sanetti, I.(1988), "Storia e proposte in tema di campionamento a livello regionale, provinciale e sub-provinciale per l'indagine sulle forze di lavoro.", Rapporto di ricerca FOLA, 4. Dipartimento

di Scienze Statistiche, Universita' di Padova.

- Falorsi, P.D. and Russo, A.(1987), "Un metodo di stima sintetica per piccoli domini territoriali nelle indagini ISTAT sulle famiglie", Atti del Convegno della Societa' Italiana di Statistica, Perugia, Italia, 11-20.
- Falorsi, P.D., Falorsi, S. and Russo, A.(1994), "Empirical Comparison of Small Area Estimation Methods for the Italian Labour Force Survey", Survey Methodology, vol. 20, No. 2, pp. 171-176.
- Fay, R.E., and Herriot, R.A.(1979), "Estimates of income for small places: An application of James-Stein procedures to census data", Journal of the American Statistical Association, 74, 269-277.
- Gelman, A. and Rubin, D.B.(1992), "Inference from iterative simulation using multiple sequences (with discussion)". Statistical Science, 7, 457-511.
- Gelfand, A. and Smith, A.F.M.(1991), "Gibbs sampling for marginal posterior expectations", Communications in Statistics - Theory and Methods, 20, 1747-1766.
- Ghangurde, P.D. and Singh, M.P.(1978). "Evaluation of efficiency of synthetic estimates", Proceedings of the Social Statistics Section, American Statistical Association, 52-61.
- Gonzalez, M.E. and Hoza, C.(1978), "Small area estimation with application to unemployment and housing estimates", Journal of the American Statistical Association, 73, 7-15.
- Ghosh, M. and Rao, J.N.K.(1994), "Small Area Estimation: an Appraisal(with Discussion)", Statistical Science, 9, 55-93.
- Marker, D.A.(1999), "Organization of Small Area Estimators Using a Generalized Linear Regression Framework", Journal of Official Statistics, 15, 1-24.
- Meindl, B.(2007), "Estimating unemployment-rates for small areas - A simulation - based approach", Paper presented at the Journal of Official Statistics, 15, 1-24.
- Purcell, N.J. and Linacre, S.(1976), "Techniques for the Estimation of Small Area

Characteristics" Paper presented at the IASS Satellite Conference on Small Area Estimation.

Rao, J.N.K. (2003), "Small area estimation", Wiley.

Thomsen, I. and Holmoy, A.M.K.(1998). "Combining data from surveys and administrative record systems", The Norwegian experience, *Inter.Statist. Rev.*, 66, 201-221.

Ugarte, L., Militino, A.F., Prado, C., Vicente, J.S., Ayestaran, M.(2007), "Small Area Estimation in Official Statistics: The Labour Force Survey of The Basque Country", *Proceedings of the Section on Small Area Estimation, IASS Satellite Conference.*

You, Y. and Rao, J.N.K.(2003), "Pseudo Hierarchical Bayes Small Area Estimation Combining Unit Level Models and Survey Weights", *Journal of Statistical Planning and Inference*, 111, 197-208.