통계연구(2026), 제31권 제2호, 1-26/ DOI https://doi.org/10.22886/jkos.2026.31.2.1

지역별 고용조사 기반 다중 랜덤효과 혼합모형을 활용한 소지역 추정 연구[1)]

진선민[2)] . 이상인[3)]

요약

국가통계 이용 환경의 변화와 함께 지역·집단 단위로 세분화된 통계에 대한 수요가 지속적으로 확대되고 있다. 그러나 표본조사에 기반 한 공식통계에서 세분화 수준이 높아질수록 과소 표본 문제가 심화되어 직접추정의 불확실성이 급격히 증가하는 한계가 존재한다. 본 연구는 이러한 정책적·실무적 요구에 대응하여 소지역추정(SAE, small area estimation) 방법을 활용해 시군구×직업대분류 수준의 소득을 안정적으로 추정하는 방법을 검토하고자 한다. 2024년 지역별고용조사 자료를 이용해 직접추정, 단일 랜덤효과 모형, 교차분류 다중 랜덤효과 모형을 비교하고, 설계가중치와 모의실험 기반 평균제곱오차(MSE) 및 변동계수(CV)를 통해 성능을 평가하였다. 분석 결과, 다중 랜덤효과 모형은 표본수가 부족한 세분화 소지역에서 MSE와 CV를 유의하게 감소시켜 가장 높은 추정 안정성을 보였다. 또한, 시군구×직업대분류 소지역 추정결과를 바탕으로 다양한 유사도 분석을 수행한 결과, 평균소득이 유사한 지역 간에도 직업별 소득 구조에서 의미 있는 동형성과 이형성이 존재함을 확인하였다. 본 연구는 교차분류 다중 랜덤효과 모형이 국가통계의 고세분화 수요에 대응할 수 있는 실용적인 소지역추정 대안임을 시사한다.

주요용어 : 소지역추정, 혼합효과 모형, 다중 랜덤효과, 지역별 고용조사

1. 서론

정확하고 세밀한 지역 통계는 정책 결정과 사회·경제 분석의 핵심 기반이다. 그러나 표본조사는 통상 광역 단위나 대분류 집단을 중심으로 설계되기 때문에 시군구나 직업·산업과 같이 세분화된 집단 수준에서는 표본 수 부족 문제가 빈번하게 발생한다. 이로 인해 해당 수준에서의 직접추정(direct estimation)은 분산이 크게 증가하여 신뢰도가 급격히 저하되는 문제가 발생한다(Pfeffermann, 2013). 소지역 추정(small area estimation, SAE)은 이러한 한계를 보완하기 위해 표본조사 자료와 행정자료·센서스 등 외부 보조정보를 통계적 모형에 결합하여 표본이 적은 영역에서도 안정적인 추정치를 산출하는 방법론으로 특히 소득·고용과 같이 지역 간 이질성이 크고 정책적 활용도가 높은 지표에서 그 중요성이 크다.

본 논문에서 소지역(small area)은 표본조사에서 직접추정만으로는 충분한 정밀도

1) 본 논문은 진선민의 석사학위논문을 수정·보완하여 작성되었으며, 2022년도 과학기술정보통신부의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(RS-2022-NR068754)
2) 제1저자, 국가데이터처, E-mail: seonmin77@korea.kr
3) 교신저자, 충남대학교 정보통계학과, 부교수, E-mail: sanginlee44@gmail.com

(precision)를 확보하기 어려울 정도로 표본수가 작은 지리적 단위 또는 하위집단을 의미한다(Rao & Molina, 2015; Ghosh & Rao, 1994). 지리적 단위로는 시군구·읍면동·격자 등이 해당되며 하위집단으로는 연령×성별과 같은 인구학적 결합집단이나 특정 직업·산업 종사자, 연령별 소그룹 등이 포함된다(Pfeffermann, 2013). 소지역 추정에서는 지역과 집단을 세분화할수록 표본의 수가 감소하고, 통계적 불확실성이 커져 분석이 매우 어려워진다는 점이 기존 연구의 명백한 한계로 지적되어 왔다. 그러나 현재와 같이 사회·경제적 구조가 다양화되고 다층적으로 변화하는 환경에서는 단일 지역단위 혹은 대분류 집단에 머무르는 추정만으로는 정책 수립과 집단 특성 이해에 한계가 있다는 문제의식이 분명히 내포된다.

최근 국가통계에서는 공표 단위의 미시화가 빠르게 진행되고 있다. 지역별고용조사는 시군구 단위의 고용·소득 지표를 제공하고 있고, 인구주택총조사와 통계지리정보서비스 또한 세분화된 공간·인구·사회경제 통계를 지원하고 있다. 국제 통계기구에서도 SAE는 공식통계 생산을 위한 핵심 방법론으로 제시되고 있다(Eurostat, 2013; UN, 2020). 이러한 변화는 지역 맞춤형 정책 설계와 집단별 격차 분석의 필요성이 확대되고 있음을 반영한다. 그러나 표본조사의 구조적 제약으로 인해 세분화된 교차 영역(예: 시군구 × 직업)에서는 표본 희소성이 심화되고 추정 불확실성이 크게 증가한다. 이로 인해 통계 공표가 제한되거나 공표되더라도 변동계수(CV)가 과도하게 높아 활용 가능성이 낮아지는 문제가 지속적으로 제기되고 있다. 비록 세분화 조건에서 방법론적 난점이 상당하지만, 세분화된 집단 단위로 더 정확한 추정치를 산출할 수 있다면 소득·고용 등 주요 변수의 이질성과 특수성을 파악하고, 맞춤형 정책 설계의 토대를 제공한다는 큰 장점이 있다. 따라서 본 연구에서는 기존 통계의 한계를 인식하면서도 적극적인 분석 방법론의 도입을 통해 세분화된 범주 단위의 소지역 추정을 시도하였다.

2. 선행연구 검토

2.1 직접 추정을 활용한 소지역 추정 연구

소지역 추정 논의의 출발점은 직접추정(direct estimation)이다. 직접추정은 각 소지역에 속한 표본조사 자료만을 이용하여 해당 지역의 모수를 산출하는 설계기반(design-based) 접근으로 표본가중치를 적용한 평균이나 비율과 같은 확률표본추정량이 대표적이다. 이 방법은 표본설계에 근거한 불편성과 표준오차 산출의 명확성이라는 장점을 갖고 있으며 공식통계 생산 과정에서 오랫동안 기본적인 지역 통계 산출 방식으로 활용되어 왔다.

국내 국가통계 부문에서도 직접추정은 지역통계 산출의 핵심적 방법으로 자리 잡아 왔다. 인구주택총조사, 가계동향조사, 지역별고용조사 등 주요 표본조사에서 시도·시군구 수준의 지표는 전통적으로 직접추정치를 중심으로 공표되어 왔으며 이는 설계기반 해석 가능성과 통계의 투명성을 중시하는 공식통계의 원칙과 부합한다(김진·김재광, 2010; 김서영·권순필, 2010; 김순영, 2019). 해외에서도 American Community Survey(ACS),

Labour Force Survey(LFS) 등 주요 조사에서 직접추정은 소지역 지표 산출의 출발점으로 활용되어 왔다.

그러나 소지역 단위가 세분화될수록 직접추정의 한계는 구조적으로 뚜렷해진다. 표본수가 충분하지 않은 지역에서는 추정량의 분산이 급격히 증가하여 신뢰성이 저하되고, 특히 지역×집단과 같이 다차원적으로 세분화된 영역에서는 표본 희소성으로 인해 정책적 활용이 가능한 수준의 통계 품질을 확보하기 어렵다는 문제가 반복적으로 지적되어 왔다(Rao, 2003). 이에 따라 표본 정보와 외부 보조정보를 결합하여 추정의 정밀도를 제고하는 모형 기반 소지역 추정 접근이 대안으로 부상하였으며 그 대표적 출발점이 단일 랜덤효과를 도입한 Fay – Herriot 모형과 단위수준 혼합모형이다.

2.2 단일 랜덤효과 모형을 활용한 소지역 추정 연구

소지역 추정의 모형 기반 접근에서 가장 널리 사용되어 온 틀은 지역수준(area-level)의 Fay – Herriot(FH) 모형과 이에 기초한 경험적 최선선형불편예측(EBLUP)이다. Fay와 Herriot(1979)는 표본수가 작은 지역의 지표를 추정하기 위해 직접추정치와 보조정보를 결합하는 2단계 구조(관측식과 연결식)를 제안하고, 두 정보를 수축(shrinkage) 가중 방식으로 결합한 추정량을 제시하였다. 이후 FH 계열 모형은 소지역 추정 교과서와 실무 지침에서 사실상의 표준으로 자리 잡았으며 소득, 실업률, 건강지표 등 다양한 영역 평균 지표에 적용되어 직접추정 대비 낮은 평균제곱오차(MSE)와 안정적인 예측 성능을 보여 왔다(Jiang & Lahiri, 2006).

단위수준(unit-level) 접근에서도 단일 랜덤효과, 주로 지역 랜덤효과를 중심으로 한 모형이 주류를 이루어 왔다. Battese, Harter, Fuller(1988)는 개체수준 보조변수와 지역 랜덤효과를 결합한 중첩오차 모형(nested-error model; BHF)을 제안하여 농업조사 자료에서 카운티 단위 면적을 정밀 예측하였고, 이후 BHF 모형은 단위수준 SAE의 대표적 틀로 정착하였다. 실무 적용 과정에서는 비정규성 및 이분산성 문제에 대응하기 위해 로그 또는 Box – Cox 변환, 가중·이분산 BHF 모형, 비모수적 잔차 평활 등 다양한 변형이 축적되어 왔다. 빈도형·비율형 지표에 대해서는 일반화선형혼합모형(GLMM) 기반 단위수준 SAE와 모수적 부트스트랩을 통한 MSE 추정이 체계화되었다(Molina & Rao, 2010; Rao & Molina, 2015). 불확실성 추정 측면에서는 Prasad와 Rao(1990)가 EBLUP의 평균제곱오차에 대한 2차 근사식을 제시하여 표준오차 및 변동계수(CV) 보고의 이론적 기반을 마련하였다. 이후 Datta와 Lahiri(2000)는 최대우도법(ML), 제한최대우도법(REML) 등 다양한 분산성분 추정 방식 하에서의 MSE 평가를 포괄하는 통일적 틀을 정식화하였다.

공간·시계열 구조를 갖는 자료의 경우에도, 다수의 연구는 여전히 단일 영역 랜덤효과를 기본으로 두되 그 공분산 구조를 확장하는 전략을 채택하였다. 즉, 랜덤효과의 차원을 늘리기보다는 하나의 영역 효과에 공간적(CAR/SAR) 또는 시간적(AR(1)) 상관을 부여하여 정보 차용을 강화하는 방식이 일반적이다(You & Zhou, 2011; Kawano, Parker & Li, 2025). 요컨대 기존 문헌의 주된 흐름은 FH(지역수준)와 BHF(단위수준)라는 단일 영역 랜덤효과 틀 안에서 보조정보 결합, 수축 예측, 불확실성 평가, 벤치마킹 절차를 점진적으로 정교화하는 과정으로 요약할 수 있다.

2.3 다중 랜덤효과 모형 적용 사례의 제한성

다중 랜덤효과 모형은 지역 외에 직업, 산업, 연령군 등 또 하나 이상의 분류축을 동시에 랜덤효과로 도입하여 교차(crossed) 또는 중첩(nested) 구조의 이질성을 함께 설명하려는 접근을 의미한다. 그러나 소지역 추정 문헌에서 지역×또 다른 분류축을 교차분류 다중 랜덤효과로 명시적으로 도입하고, 그로 인한 정밀도 개선을 체계적으로 검증한 실증 연구는 상대적으로 드물다. 기존 연구에서 '다중'이라는 표현은 반드시 교차분류 다중 랜덤효과의 도입을 의미하지는 않는다. <표 2.1>에 요약된 바와 같이, 지역수준 FH 계열에서는 시간·공간 의존성이나 상위 계층을 추가 구성요소로 결합하는 방식이 다수 보고되어 왔다(Rao & Yu, 1994; Marhuenda, Molina & Morales, 2013).

<표 2.1> 지역수준(area-level) 다중 구성요소 FH 계열

모형 구분	참고문헌	랜덤효과 구조	적용 사례	적합·성능 평가	교차분류 여부
시계열+횡단면 FH	Rao & Yu (1994)	지역 u_d + 시간 AR(1)·표본오차 AR(1)	고용·소득 등 시계열 영역평균	REML/ BLUP, MSE 근사	**아님** (지역 1축)
공간 FH	Cressie	지역 $u \sim$ CAR/SAR	인접영역 상관 존재 지표	REML /EBLUP	**아님** (지역 1축)
시공간 FH	Marhuenda, Molina & Morales (2013)	u에 공간(CAR/SAR) + 시간 AR(1)	행정·설문 지표	REML, 부트스트랩 MSE	**아님** (지역 1축)
다변량 FH	Benavent & Morales (2016)	벡터 u_d (지표 간 공분산)	여러 지표 동시 추정	REML/ EBLUP, MSE 근사	**아님** (영역 1축, 지표 다변량)
Three-fold FH	Marcis et al. (2023)	3계층 랜덤효과 (지역+상위집단 +잔차)	영역지표 일반	EBLUP, 부트스트랩 MSE, 진단	**아님** (영역 내부 3계층이지만 교차분류 아님)

또 다른 관점에서는 반응변수 분포의 이중성을 고려한 two-part 소지역 추정 접근이 제안되어 왔다. 이 접근은 반응변수가 0에 집중되어 있거나 강한 비대칭성을 보이는 경우, 발생 여부와 크기를 분리하여 각각을 별도의 하위 모형으로 설명하는 방식이다. Pfeffermann, Terryn, Moura(2008)는 문해력과 같은 지표를 대상으로 발생 여부와 크기를 분리한 two-part 모형에 영역 랜덤효과를 도입하고 MCMC 및 부트스트랩을 통해 불확실성을 평가함으로써, 기존 단일 분포 가정 하의 소지역 추정보다 예측 안정성을 향상시킬 수 있음을 보였다. 이러한 접근은 분포의 두 부분을 구조적으로 분리하는 데 초점을 두지만, 지역×직업과 같이 서로 다른 분류축을 동시에 모형화하는 교차분류 다중 랜덤효과를 직접 도입한 사례로 보기는 어렵다.

단위수준(unit-level) 소지역 추정에서도 유사한 경향이 관찰된다. 다수의 연구는 단일 지역 랜덤효과를 기본 구조로 유지한 채, 공간적 또는 시공간적 상관을 공분산 구조로 부여하여 인접 지역이나 인접 시점 간 정보 차용을 강화하는 변형을 제안해 왔다. 이러한 접근은 희박한 영역에서의 추정 안정화를 달성하는 데 기여하였으나, 임금이나 직업과 같이 세분화된 집단 수준에서 발생하는 이질성을 명시적으로 분해하는 데에는 한계가 있다. 실제로 지역×집단(예: 지역×직업) 교차 셀을 단위수준 SAE의 랜덤효과 구조에 명시적으로 포함시켜 적용한 응용 연구는 극히 제한적이다. 이러한 희소성은 단순한 연구 공백이라기보다는 현실적인 제약이 누적된 결과로 해석할 수 있다. 교차 셀 단위의 표본 수 부족은 분산 성분의 식별성을 약화시키고, 분산 추정의 불안정성을 초래한다. 또한 복합표본 설계 하에서 가중치·층화·집락 정보를 교차분류 다중 랜덤효과 구조와 일관되게 반영하는 데에는 상당한 방법론적 난이도가 따른다.

2.4 본 연구의 차별성

본 연구는 기존 소지역 추정 연구가 주로 지역 단위 평균 수준의 정확도 개선에 초점을 두어 온 한계를 넘어 시군구×직업분류라는 보다 세분화된 단위에서 소득을 추정하고, 그 수준과 구조를 함께 해석 가능한 분석 틀을 제안한다.

첫째, 분석 단위를 지역 단일 차원에서 확장하여 지역과 직업이라는 두 축이 교차하는 세분화 소지역을 실증적으로 다룬다. 이는 표본 규모가 급격히 감소하는 영역임에도 불구하고, 현실적인 표본조사 제약 하에서 지역 내부의 직업군별 소득 격차를 정책적으로 해석 가능한 수준까지 제시하고자 한 시도라는 점에서 차별성을 갖는다.

둘째, 이러한 세분화 단위의 이질성을 반영하기 위해 지역 효과와 직업분류 효과를 동시에 고려하는 다중 랜덤효과 모형을 활용한 소지역 추정 접근을 제안한다. 이는 지역 간 변이뿐 아니라 동일 지역 내 직업군 간 구조적 차이를 함께 반영할 수 있는 틀을 제공함으로써, 기존의 단일 차원 랜덤효과 중심 연구를 확장한다.

셋째, 소지역 추정 결과의 성과를 단순한 평균 수준의 근접성으로 평가하는 데서 벗어나, 정확도와 더불어 소득 구조의 형태를 함께 비교·판별하는 관점을 도입한다. 이를 통해 평균 소득이 유사한 지역이라 하더라도 직업군 구성에 따라 상이한 소득 구조를 가질 수 있음을 체계적으로 검토할 수 있는 기반을 마련한다.

넷째, 설계기반 추정과 모형기반 추정을 대립적으로 비교하는 데 그치지 않고, 직접추정, 모형기반 추정, 불확실성 평가, 구조 비교를 하나의 일관된 평가체계로 통합한다. 이는 실무에서 요구되는 재현성, 비교 가능성, 해석 용이성을 동시에 고려한 접근으로 기존 연구에서 분절적으로 다루어지던 요소들을 단일 프레임으로 연결한다.

다섯째, 소표본 영역에서의 추정 결과를 전국 상·하위 소득권, 인접 지역 등 정책적으로 의미 있는 비교 구도로 제시함으로써, 산업 구조가 특화된 지역이나 소멸 위험 지역에서 직업별 소득 격차가 갖는 정책적 함의를 구체적으로 제시한다. 이를 통해 소지역 추정 결과를 단순한 통계적 산출물에 그치지 않고, 지역 맞춤형 정책 논의로 확장할 수 있는 해석 가능성을 높인다.

3. 분석 방법

3.1 시군구×직업분류 소득 추정

본 연구는 시군구×직업분류 단위의 소득을 단위수준(unit-level)에서 추정하기 위해 2024년 지역별고용조사 원자료를 활용하여 설계기반 직접추정과 혼합모형 기반 소지역 추정을 비교 적용한다. 특히 지역과 직업분류의 이질성을 동시에 반영하는 다중 랜덤효과 혼합모형을 중심 방법론으로 설정하고, 이를 기존의 단일 랜덤효과 모형 및 직접추정과 동일한 기준에서 비교한다.

먼저 설계기반 직접추정은 각 시군구(d)×직업분류(j) 셀에서 관측된 응답값의 가중평균으로 소지역 (d,j)의 직접추정치는 식 (3.1)과 같이 가중평균으로 정의한다.

$$\hat{\theta}_{dj}^{D} = \frac{\sum w_i Y_i}{\sum w_i} \tag{3.1}$$

이는 표본설계를 충실히 반영한다는 장점이 있으나, 세분화된 셀에서는 표본 수 부족으로 인해 분산이 급격히 증가하는 한계를 가진다. 본 연구에서는 설계기반 직접추정을 기준(baseline)으로 설정하여 모형 기반 추정의 성능을 상대적으로 평가한다.

직접추정은 설계가중치(w)를 이용해 모수를 추정하므로 불편성(unbiasedness)을 보장하지만, 소지역 수준에서는 표본 수 부족으로 인해 분산이 크게 증가한다. 특히 다차원 교차 영역에서는 표본이 거의 존재하지 않는 경우도 빈번하게 발생하여 세분화된 집단의 추정에서는 직접추정치의 활용 가능성이 크게 제한된다. 반면, 모형 기반 소지역 추정은 이러한 한계를 극복하기 위해 조사자료와 보조정보를 통계모형으로 결합하고, 영역 간 정보 차용(borrowing of strength)을 통해 추정의 분산을 감소시키는 접근이다 (김순영, 2019; Rao & Molina, 2015). 이 가운데 혼합모형 기반 SAE는 고정효과로 보조변수의 체계적 영향을 설명하고, 랜덤효과로 소지역 간 이질성을 반영함으로써 실무적으로 가장 널리 활용되고 있다(Pfeffermann, 2013; Rao & Molina, 2015).

단위수준 혼합모형의 기본 형태로는 지역만을 랜덤효과로 포함하는 단일 랜덤효과 모형을 고려한다.

$$y_i = x_i^T \beta + u_{d(i)} + \varepsilon_i, \ \ i = 1,...n \tag{3.2}$$

$$u_{d(i)} \sim N(0, \sigma_u^2), \ \ \varepsilon_i \sim N(0, \sigma_\varepsilon^2)$$

여기서 β는 고정효과 계수, x_i는 보조변수 벡터, $u_{d(i)}$는 소지역 d의 랜덤효과, ε_i는 개체 오차이다. 이 모형은 지역 간 평균 차이를 반영할 수 있으나, 동일 지역 내 직업분류 간 구조적 이질성은 오차항에 흡수되는 한계를 가진다. 이러한 한계를 보완하기 위해 본 연구는 지역과 직업분류를 동시에 랜덤효과로 포함하는 교차분류(cross-classified) 다중 랜덤효과 혼합모형을 핵심 방법론으로 적용한다.

$$y_i = x_i^T\beta + u_{d(i)} + v_{j(i)} + \varepsilon_i,\ \ i = 1,....,n, \tag{3.3}$$

여기서 β는 고정효과 계수, $u_{d(i)}$는 지역 수준 랜덤효과, $v_{j(i)}$는 직업 대분류 수준 랜덤효과, ε_i는 개체 오차이며 지역 효과와 직업 효과는 서로 독립이라고 가정한다. 이 구조는 지역 간 정보 차용과 동시에 직업분류 간 정보 차용을 가능하게 하여 표본이 희박한 시군구×직업분류 셀에서도 추정 안정성을 제고한다. 다중 랜덤효과 모형의 핵심 장점은 분산 성분의 분해를 통해 변동의 원천을 구조적으로 구분할 수 있다는 점에 있다. 단일 랜덤효과 모형에서는 직업분류에 따른 체계적 차이가 오차항에 포함되지만, 다중 랜덤효과 모형에서는 이를 별도의 분산 성분으로 분리함으로써 추정의 효율성을 높인다. 이는 특히 지역 내부에서 직업 구조가 상이한 경우에 중요한 의미를 갖는다.

모형 기반 소지역 추정은 전체 표본자료에 적합된 모형을 기반으로 각 개인에 대한 예측값을 산출한 후, 개인별 예측값에 설계가중치를 적용하여 시군구×직업분류 단위의 소지역 평균 소득을 추정하는 방식으로 수행된다. 개인 i가 지역 d, 직업분류 j에 속할 때의 예측값은 식 (3.4)와 같이 계산된다.

$$\hat{y}_i = x_i^T\hat{\beta} + \hat{u}_d + \hat{v}_j \tag{3.4}$$

여기서, $\hat{u}_d$와 $\hat{v}_j$는 각각 지역 및 직업 대분류 수준 랜덤효과의 EBLUP 추정치이다. EBLUP은 분산성분을 자료로부터 추정한 후, 해당 추정값을 이용하여 랜덤효과의 조건부 기댓값을 계산하므로 개별 지역이나 직업분류의 표본수가 적을수록 전체 평균 방향으로 수축(shrinkage)하는 특성을 갖는다. 이를 통해 극단적인 직접 추정값을 완화하면서도 표본 정보가 충분한 집단의 특성은 상대적으로 잘 보존할 수 있다.

직접추정과 모형 기반 추정의 불확실성 평가는 2024년 지역별고용조사 표본자료를 의사모집단(pseudo-population)으로 간주하고, 실제 조사 설계와 유사한 절차를 반복 적용하여 소지역별 추정치의 변동성을 평가하는 설계기반 반복 표본추출 기반 모의실험을 통해 수행하였다(Prasad & Rao, 1990). 한편 설계가중치를 사용하는 소지역 추정량의 MSE 평가에 관한 방법론은 Torabi and Rao(2005, 2010)에 의해 체계적으로 논의된 바 있다. 지역별고용조사는 1단계에서 시·군·구를 가구 수 등 크기척도에 따른 확률비례계통추출(PPS)로 선정하고, 2단계에서 선정된 조사구 내에서 시작가구를 단순임의로 추출한 후 집락추출 방식으로 연속된 20가구를 표본으로 선정하는 다단계 집락표본설계를 따른다. 이러한 복합표본설계를 엄밀히 재현하는 것은 조사구 정보 없이는 불가능하므로 본 연구에서는 계산편의성과 재현성을 고려하여 근사 설계 표본추출 절차를 적용한다. 구체적으로, 의사모집단에서 시·군·구를 층으로 설정한 후 각 층 내에서 동일한 추출률(10%)로 단순임의추출을 수행하여 표본을 구성하고 각 표본에서 방법론별 소지역을 추정하는 절차를 100회 반복한다. 각 소지역 d에 대해 각 추정량 $\hat{\theta}_d$의 평균제곱오차(MSE)는 식 (3.5)와 같이 계산된다.

$$\widehat{MSE}(\hat{\theta}_d) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}_d^{(b)} - \theta_d^*)^2 \tag{3.5}$$

여기서 $\hat{\theta}_d^{(b)}$는 b번째 반복에서의 추정치, θ_d^*는 의사모집단에서의 평균소득이다. 추정치의 상대적 안정성을 평가하기 위해 변동계수(coefficient of variation, CV)를 함께 산출한다.

$$CV(\hat{\theta}_d) = \frac{\sqrt{\widehat{MSE}(\hat{\theta}_d)}}{|\hat{\theta}_d|} \times 100\,(\%) \tag{3.6}$$

CV는 영역 간 규모 차이를 제거한 상대적 불확실성 지표로 표본 규모가 작은 시군구 × 직업분류 셀에서 추정의 안정성을 비교하는 데 유용하다.

한편, 본 연구에서 적용한 층화 단순임의추출 근사 설계는 실제 지역별고용조사의 확률비례추출 및 집락 구조를 완전히 반영하지 못하는 한계가 있다. 이에 따라 분산 추정치는 실제 설계 기반 분산에 비해 과소추정될 가능성이 존재한다. 그러나 모든 방법론에 동일한 설계를 적용하여 모의실험 기반 성능지표(MSE, CV)를 산출하였으므로 직접추정법, 단일 및 다중 랜덤효과 모형 활용 추정법 간의 상대적 성능 비교는 일관된 기준에서 이루어졌다. 특히 다중 랜덤효과 모형은 이중 수축 효과를 통해 분산과 CV를 유의미하게 감소시키는지를 중심으로 평가한다. 본 연구는 지역과 직업분류라는 두 축의 이질성을 동시에 반영하는 다중 랜덤효과 혼합모형을 중심으로 구성되며, 이는 기존의 직접추정이나 단일 랜덤효과 모형이 갖는 한계를 구조적으로 보완하는 접근이다. 이러한 모형 구조는 세분화된 소지역 단위에서도 안정적인 추정을 가능하게 하는 기반을 제공한다.

다중 랜덤효과 혼합모형의 도입은 모형 복잡도를 증가시킬 수 있으므로, 본 연구에서는 과적합 가능성과 모형 적합성을 함께 점검한다. 이를 위해 단일 랜덤효과 모형과 다중 랜덤효과 모형에 대해 Akaike 정보기준(AIC)과 Bayesian 정보기준(BIC)을 산출하여, 적합도와 모형 복잡도의 균형을 비교 평가한다. AIC와 BIC는 식 (3.7)과 같다.

$$AIC = -2\log(L) + 2k, \tag{3.7}$$
$$BIC = -2\log(L) + k\log(n)$$

여기서 L은 최대우도, k는 유효 모수 수이다. 일반적으로 AIC와 BIC 값은 낮을수록 해당 모형의 데이터 적합도가 우수함을 의미한다.

3.2 시군구×직업분류별 소득 구조 파악

본 연구는 시군구×직업분류 수준의 소득을 추정하는 데 그치지 않고, 동일 또는 유사한 평균 소득 수준을 보이는 지역들 간에 소득 구조가 동형적인지, 혹은 이형적인지를 판별하기 위한 방법론적 틀을 함께 제안한다. 서론에서 지적했듯 동일 지역 내부라도 직업구성의 차이가 평균과 분포를 크게 바꿀 수 있어 시군구 평균만으로는

내부 이질성을 포착하기 어렵다. 특히 정책 현장에서는“평균이 비슷한 두 지역을 동일 처방으로 다뤄도 되는가?”라는 질문이 반복된다. 이에 본 장은 다중 랜덤효과 모형으로 추정한 결과를 사용해 다음의 연구 질문에 답하고자 한다.

(1) 인접하면서 평균소득이 유사한 시군구이라도, 직업분류별 소득 구조는 얼마나 어떻게 다른가
(2) 전국 상위/하위 소득권 시군구쌍의 직업별 고소득/취약 구조는 동형인가 이형인가

이를 위해 평균 수준 중심의 비교를 넘어, 직업분류별 소득 분포의 상대적 구성과 서열, 분포 거리, 일치 정도를 종합적으로 평가하는 다중 지표 접근을 적용한다. 직업분류는 8차 대분류 9개를 기본으로 하되, 각 조합의 신뢰도를 확보하기 위해 아래 기준을 적용하였다.

- (해석 포함 기준) 두 지역 모두에서 해당 직업의 상대표준오차(=모형기반 CV) $\leq$ 0.50
- (주의 표시) 어느 한쪽이라도 0.5 < CV $\leq$ 0.7인 경우 *로 표시하고 정성적 비교에 한정
- (제외 권고) CV > 0.70이거나 표본수 $n < 30$인 경우는 본문 비교에서 제외

이 기준을 통과한 공통 직업군에 대해, 수준(평균) 와 형태(서열) 정보를 분리하여 다음 절차로 비교·분류한다. 먼저 ① 정량 비교지표로 서열 유사도와 분포 유사도를 요약한다. 서열 유사도는 순위 상관(스피어만 ρ)으로 측정하며 ρ가 높을수록 두 지역의 직업별 서열 구조가 유사함을 뜻한다. 순위 상관계수는 두 지역별 서열의 순위 $R_i^{(A)}, R_i^{(B)}$라 하면 순위 상관 ρ는 식 (3.8)과 같다.

$$\rho = \frac{\sum (R_i^{(A)} - \overline{R}^{(A)})(R_i^{(B)} - \overline{R}^{(B)})}{\sqrt{\sum (R_i^{(A)} - \overline{R}^{(A)})^2}\sqrt{\sum (R_i^{(B)} - \overline{R}^{(B)})^2}} \quad \textbf{(3.8)}$$

순위 상관계수의 범위는 (-1, 1)로, 클수록 서열 구조가 유사하다고 볼 수 있다. 분포 유사도는 L1-share로 계산한다. 각 지역의 직업별 소득을 합으로 나눈 직업별 비중 S_A, S_B을 구한 뒤 $L1 - share \equiv \frac{1}{2}\sum |s_A - s_B|$로 정의하며, 값이 작을수록 분포 형태가 가깝다고 볼 수 있으며, 해석은 두 분포를 일치시키려면 최소한 L1-share 만큼의 비중(%p)을 재배치해야 한다는 뜻이다. 해석 기준은 다음과 같다: 동형은 대체로 ρ≥ 0.85이면서 L1≤0.03, 이형은 ρ≤0.80이거나 L1≥0.06, 그 사이 구간은 경계/중간형으로 보고 다른 지표와 함께 종합 판단한다.

다음으로 ② 산점도(Bland - Altman)로 수준 차이의 영향을 제거하고 형태(서열) 차이에만 초점을 맞추기 위해 수준 차이를 표준화해 시각화한다. 직업별 추정소득 Y_A, Y_B를 지역별 평균과 표준편차로 z-점수화(Z_A, Z_B)하여 수준 차이의 영향을 제거한 뒤, x축을 공통 평균 $(Z_A + Z_B)/2$, y축을 표준화 차이 $(Z_A - Z_B)$로 둔다. 점들이 y=0선

부근에 밀집하면 두 지역의 수준 격차와 체계적 편의가 작아 동형 경향으로 본다. 반대로 절댓값이 큰 점들이 다수 존재하거나 한쪽으로 치우치면 이형 신호로 해석한다.

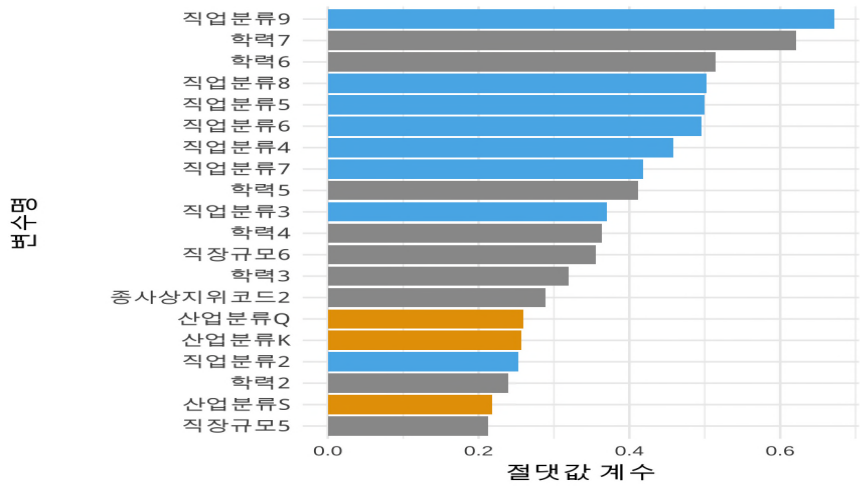
마지막으로 ③ 서열변화 그래프를 통해 형태 차이를 보완한다. 두 지역의 직업별 순위를 나란히 배치하고 동일 직업을 선으로 연결하여 교차의 유무와 크기를 확인한다. 교차가 없거나 극히 일부·경미하면 서열 구조가 유사한 동형에 가깝고, 교차가 많고 벌어짐이 크면 이형으로 본다. 특히 상위 1~3위권에서 반복적인 교차가 나타나면 우세축(제조 vs. 전문·사무)이 다르다는 근거가 된다. 이러한 지표들을 함께 고려하여 종합적으로 판정하고, 정리하면 산점도에서 0선 밀집, ρ이 높고 L1이 낮으며, 서열 교차가 없거나 경미하면 동형으로 판정한다. 산점도에서 체계적 이탈이 보이거나 ρ가 낮거나 L1이 큰 편이며 상위축 교차가 반복되면 이형으로 본다. 본 연구는 단일 지표에 의존하지 않고, 다중 지표 기반 구조 판정을 통해 평균 소득 중심의 비교가 간과할 수 있는 지역 내부 직업 분포의 차이를 체계적으로 드러내며, 다중 랜덤효과 모형을 통해 추정된 세분화 소지역 결과의 해석 가능성을 확장한다.

4. 분석결과 및 해석

4.1 시군구별 × 직업분류별 소득 추정

본 연구는 2024년 지역별고용조사 원자료를 이용하여 시군구 및 시군구×직업 대분류 수준의 평균 소득을 추정한다. 분석 대상은 최근 3개월 평균 임금이 관측된 임금근로자이고, 종속변수는 월평균 임금을 사용한다. 모형 적합 시에는 임금 분포의 비대칭성을 완화하기 위해 월평균 임금을 로그 변환한 값을 사용하고, 결과 해석 단계에서는 로그 변환에 따른 편의를 보정하기 위해 Duan의 스미어링 보정(Duan, 1983)을 적용하여 원척도 임금 수준의 추정 결과를 제시한다. 소지역은 시군구 $d = 1,...,D$ 와 직업대분류 $j = 1,...,J$ 의 교차조합 (d,j)으로 정의하며, 이는 표본 희소성이 가장 두드러지는 최소 정책 단위에 해당한다. 보조변수로는 성별, 연령, 교육수준, 혼인상태, 근로시간, 산업분류 등을 포함하였다. 소지역 소득 추정 결과의 비교를 위해 본 연구에서는 직접추정 결과를 기준으로 지역 단일 랜덤효과를 포함한 단위수준 혼합모형과 지역 및 직업분류를 동시에 고려한 다중 랜덤효과 혼합모형을 적용하였다. 이를 통해 지역 간 정보 차용뿐 아니라 동일 지역 내 직업분류 축에서의 이질성이 추정 결과에 미치는 영향을 함께 평가하였다. 먼저, 직업분류와 산업분류 중 어느 축이 임금 변동 설명에 더 중요한지를 확인하기 위해 전체 공변량을 포함한 선형모형에 LASSO를 적용한 결과, <그림 4.1>과 같이 직업분류 관련 계수의 상대적 중요도가 산업분류보다 일관되게 크게 나타났다. 이는 임금 수준의 변동이 산업 차이뿐 아니라 직업분류 축을 따라 구조적으로 분화되어 있음을 시사하며 직업분류를 추가 랜덤효과로 포함할 필요성을 뒷받침한다.

<그림 4.1> LASSO 계수 절댓값 비교



이러한 결과를 바탕으로 단일 랜덤효과 모형(지역)과 다중 랜덤효과 모형(지역+직업)의 적합도를 비교한 결과는 <표 4.1>에 제시하였다. 다중 랜덤효과 모형의 AIC(139,353.3)와 BIC(139,660.5)는 단일 랜덤효과 모형의 AIC(152,328.4), BIC(152,625.7)에 비해 현저히 낮게 나타났다. 이는 지역 효과에 더해 직업분류 수준의 이질성을 추가로 반영함으로써 관측 자료의 분산 구조를 보다 충실히 설명하고 있음을 의미한다.

<표 4.1> 모형의 AIC/BIC 비교 결과

모형	AIC	BIC
단일 랜덤효과(지역)	152,328.4	152,625.7
다중 랜덤효과(지역+ 직업)	139,353.3	139,660.5

이러한 결과는 소득 변동이 지역 간 차이에만 국한되지 않고 동일 지역 내에서도 직업분류 축을 따라 체계적으로 존재함을 시사하며, 해당 축을 모형화할 경우 소지역 단위에서 정보 차용의 범위가 확장되어 예측의 정확성과 안정성이 함께 제고될 수 있음을 보여준다.

본 연구의 관심 모수는 시군구 d와 직업분류 j의 조합으로 정의되는 소지역 (d,j) 평균소득 $\hat{\theta}_{dj}$이다. 분석 단위는 시군구와 직업 대분류(9개)의 교차 조합이며, 극단적 변동을 억제하기 위해 원표본 기준 각 조합의 최소 표본수를 일정 기준 이상으로 제한하였다. 이에 따라 총 1,374개 소지역 조합을 대상으로 추정 결과를 산출하였다. 직업 대분류는 관리자(1), 전문가(2), 사무(3), 서비스(4), 판매(5), 농림어업 숙련(6), 기능원 및 관련 기능(7), 장치·기계 조작·조립(8), 단순노무(9)이며, 군인(A)은 제외하였다. 농림어업 숙련(6)은 다수 시군구에서 표본수가 기준 미만으로 관측되어 대부분의 분석에서 제외되었다. 추정은 먼저 설계가중치를 반영한 직접추정치를 산출하여 소지역 추정의 기준선으로 설정한 후, 이를 바탕으로 모형기반 추정을 단계적으로 적용하는 방식으로 수행되었다. 모든 추정 방법에서 동일한 분석 대상과 변수 정의를 유지함으로써, 추정 결과 간 차이가 절차적 요인이 아니라 모형 구조에 기인하도록 설계하였다.

최종 모형은 성별, 연령대, 교육수준, 혼인상태, 근로시간, 사업체 규모, 산업분류, 종사상지위를 고정효과로 포함하고, 지역과 직업을 임의절편(random intercept)으로 포함한 다중 랜덤효과 혼합모형으로 설정하였다. 종속변수는 log(소득+1)이며, 고정효과 계수 추정 결과는 <표 4.2>에 제시하였다. 전반적으로 인구사회학적 특성과 산업·직업 특성이 소득 수준에 통계적으로 유의한 영향을 미치는 것으로 나타났으며, 이는 모형에 포함된 보조변수들이 소득 변동을 설명하는 데 적절하게 기여하고 있음을 시사한다. 또한 다중 랜덤효과 모형의 분산성분 추정 결과, 직업분류에 대한 임의 절편 분산(0.0452)이 지역에 대한 분산(0.0016)보다 현저히 크게 나타났다. 이는 임금 수준의 변동이 지역 간 차이보다 직업 구조에 의해 더 크게 설명됨을 시사한다. 이러한 과정을 통해 산출된 표본 상·하위 5개 시군구별 직업분류별 평균소득 추정 결과는 각각 <표 4.4>와 <표 4.5>와 같다. 추정한 소득의 단위는 만 원이다.

<표 4.2> 최종 모형 고정효과 계수 추정 결과 (요약)

변수	추정치(β)	SE	t value
성별(여성)	-0.1645	0.0021	-77.22
연령(30세이상~50세미만)	0.1474	0.0031	47.06
연령(50세이상~55세미만)	0.2113	0.0042	50.85
교육(전문대)	0.1762	0.0048	37.00
교육(4년제 대학교)	0.2520	0.0081	31.28
근로시간	0.0217	0.0001	228.18
사업체규모(300인 이상)	0.3669	0.0037	98.73
종사상지위(상용직)	0.3714	0.003	124.07
산업분류(금융 및 보험업)	0.3287	0.0069	47.38

<표 4.3> 다중 랜덤효과 모형의 분산성분 추정 결과 (REML)

랜덤효과	분산	표준편차	비율 (%)
지역	0.00157	0.0397	0.9
직업분류	0.04520	0.2126	25.0
잔차	0.13520	0.3677	74.1

※ 비율은 (분산/전체분산) × 100

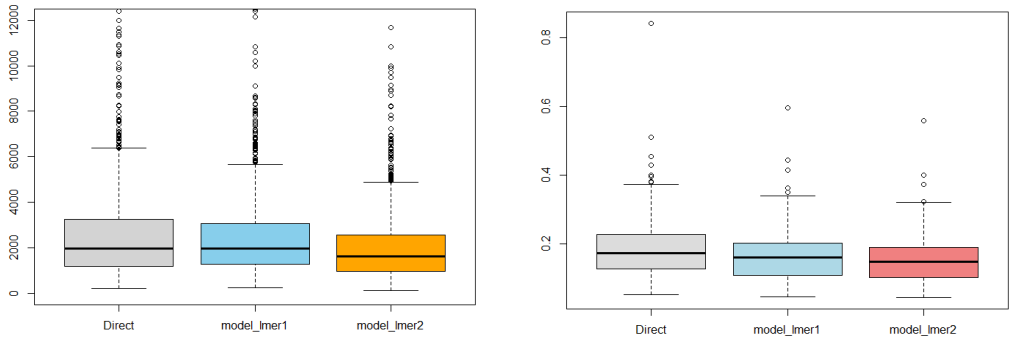
<표 4.4> 시군구별/직업분류별 표본수 상위 지역 추정량 결과 비교

시군구	직업분류	직접 추정법			단일 랜덤효과			다중랜덤효과		
		소득	MSE	CV	소득	MSE	CV	소득	MSE	CV
세종 세종시	2	437	724	6	391	2,791	5	417	943	5
	3	391	416	5	376	565	4	372	685	4
	4	228	914	13	232	941	13	215	1,090	13
	7	355	2,556	14	344	2,702	14	348	2,030	12
	8	356	1,648	11	379	1,597	9	355	896	8
	9	166	539	14	192	1,730	17	158	663	16
청주시 상당구	2	376	1,599	11	332	2,997	9	366	1,294	9
	3	366	786	8	345	869	6	351	673	6
	4	210	1,002	15	203	935	15	195	891	14
	5	229	2,077	20	218	1,770	18	219	1,575	17
	7	380	2,597	13	355	1,736	11	362	1,245	9
	8	354	634	7	361	558	6	341	509	6
	9	177	384	11	196	768	12	165	549	11
전주시 완산구	2	343	643	7	308	1,869	7	330	724	7
	3	347	646	7	337	494	6	337	481	6
	4	197	1,014	16	187	941	16	181	967	15
	5	250	1,942	18	245	1,461	15	238	1,328	13
	7	318	1,285	11	307	1,396	11	313	990	9
	8	371	2,284	13	375	1,124	9	351	1,550	8
	9	150	641	17	164	949	17	139	601	16
창원시 의창구	2	351	858	8	315	2,074	8	346	892	9
	3	376	946	8	359	751	6	368	573	6
	4	209	1,248	17	217	1,406	16	206	929	15
	5	266	3,056	21	273	2,837	19	263	1,950	17
	7	345	2,114	13	325	2,045	12	337	1,429	11
	8	343	773	8	357	627	6	341	437	6
	9	180	777	16	191	760	13	162	633	12
광주 광산구	2	354	1,713	12	315	2,025	8	346	854	8
	3	350	690	8	331	773	7	339	529	6
	4	210	1,031	15	205	923	15	196	890	14
	5	271	3,923	23	259	1,711	16	252	1,472	15
	7	334	1,018	10	337	925	9	345	949	8
	8	319	877	9	334	680	7	319	407	6
	9	165	585	15	176	593	12	148	590	12

<표 4.5> 시군구별/직업분류별 표본수 하위 지역 추정량 결과 비교

시군구	직업 분류	직접 추정법			단일 랜덤효과			다중랜덤효과		
		소득	MSE	CV	소득	MSE	CV	소득	MSE	CV
경북 의성군	2	269	9,911	37	252	3,169	20	280	2,606	18
	3	289	4,307	23	288	3,323	20	298	3,248	19
	4	187	2,098	24	171	1,514	19	167	1,602	19
	9	103	1,086	32	116	1,393	31	100	813	28
대구 군위군	3	296	5,319	25	287	4,719	24	295	4,190	22
	4	191	3,752	32	192	2,396	26	188	2,083	24
	9	96	731	28	106	876	28	92	622	25
경남 남해군	2	304	6,889	27	268	3,194	19	309	3,233	18
	3	338	9,163	28	312	4,111	18	318	3,683	18
	4	185	1,124	18	194	1,034	16	186	793	15
	9	139	1,102	24	159	1,693	24	132	999	22
경북 청도군	2	284	5,851	27	264	3,362	21	295	3,540	20
	3	316	3,931	20	297	3,429	19	307	3,122	18
	4	187	1,381	20	168	1,763	22	160	1,808	20
	8	254	2,359	19	284	2,672	16	278	2,009	15
	9	155	1,218	23	167	1,520	22	142	990	20
전남 구례군	2	302	9,476	32	271	4,140	22	302	4,110	21
	3	305	2,756	17	279	3,036	17	291	2,185	15
	4	219	2,643	24	197	2,430	22	192	2,357	20
	8	277	5,751	27	276	3,721	22	269	3,534	21
	9	159	1,356	23	163	1,611	25	135	1,419	21

<그림 4.2> 세 방법별 MSE, CV 비교



a. MSE 비교 b. CV 비교

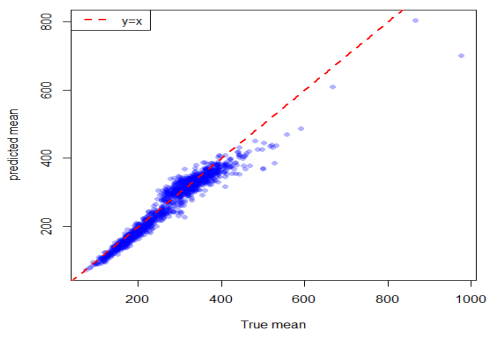
표본 규모에 따라 직접추정과 모형 기반 추정의 성능을 구분하여 살펴보면, 구간별로 차이가 일부 관찰된다. 상위 구간에서는 직접추정의 MSE가 다중 랜덤효과 모형보다 비슷하거나 일부 구간에서는 더 작게 나타났다. 이는 표본이 상대적으로 많은 경우 직접추정의 분산이 이미 작고, 수축이 유발하는 추가적 분산 감소 여지가 제한적이기 때문이다. 수축 추정은 분산을 줄이는 대신 소량의 편향을 도입하는 특성을 가지므로,

분산 절감 효과가 크지 않은 영역에서는 MSE 개선 폭이 제한되거나 일부 역전 현상이 나타날 수 있다. 반면 표본 하위 구간에서는 MSE가 전반적으로 직접 > 단일 > 다중 순으로 감소하는 전형적인 수축 효과가 확인되었다. 이는 표본이 극히 적은 셀에서 다중 랜덤효과 모형이 지역과 직업구성의 이질성을 동시에 반영함으로써 분산을 효과적으로 완화하였기 때문이다. 특히 시군구×직업분류와 같은 교차소지역 구조에서는 개별 셀의 표본 규모가 구조적으로 제한되는 경향이 있어, 다중 랜덤효과 모형의 안정화 효과가 더욱 두드러진다. 일부 셀에서 다중 랜덤효과 모형의 MSE가 단일 랜덤효과 모형보다 높거나 차이가 미미하게 나타난 사례가 관찰되었는데, 이는 수축 기반 모형에서 분산 감소와 편향 증가 간의 균형(trade-off)으로 설명될 수 있다. 특히 특정 시군구에서 직업구성이 상대적으로 균질하거나 직업분류별 표본 비중이 높은 경우에는 단일 랜덤효과 모형만으로도 충분한 설명력이 확보되며, 추가적 직업 랜덤효과의 이득이 제한적일 수 있다. 다만 이러한 사례는 전체 교차소지역 구조에서의 전반적 경향을 뒤집는 수준은 아니며, 전체 셀을 대상으로 할 경우 다중 랜덤효과 모형이 단일 랜덤효과 모형 대비 MSE 감소 경향성을 보였다.

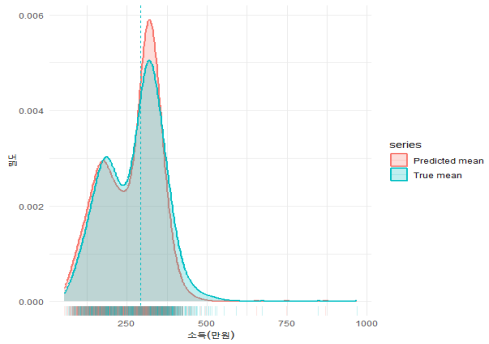
CV는 상·하위 모든 구간에서 일관되게 직접>단일>다중의 감소를 보인다. 이는 모형 기반 추정이 상대적 불확실성을 체계적으로 줄여 공표 기준(25% 이내 공표)을 보다 잘 충족시킴을 시사한다. 또한 <그림 4.2>의 상자그림으로 표본 규모에 상관없이 전체 소지역을 보면, MSE와 CV 모두가 직접(Direct)→단일(model_lmer1)→다중(model_lmer2)으로 갈수록 작아지며, 특히 다중 랜덤효과 모형이 직업구성 이질성을 랜덤효과로 흡수해 희소 셀 변동을 완화하는 효과가 뚜렷하다.

이러한 결과는 소지역을 더 세분화할수록 개별 셀의 표본 수가 급격히 줄어 분산이 커지는 현실과 맞물려, 소표본 셀에서는 다중 랜덤효과 모형이 MSE와 CV를 동시에 개선하는 실증적 이점이 일관되게 나타났다. 더 나아가, AIC/BIC 기준에서도 다중 랜덤효과 모형이 단일 모형 대비 우월하였고 <그림 4.3>의 산점도와 밀도 곡선에서도 전반적으로 다중 랜덤효과 모형이 모집단 평균과 높은 일치도가 확인된다. 밀도곡선(b) 역시 두 분포가 거의 중첩되어 중심과 분산이 유사함을 보이며, 다중 랜덤효과 모형의 예측은 체계적 편의 없이 모집단 평균을 잘 재현하는 것으로 해석된다.

<그림 4.3> 모집단 평균과 다중 랜덤효과 모형 예측값 평균 분포 비교



a. 모집단 평균과 예측값 평균 산점도



b. 모집단 평균과 예측값 평균 분포(밀도 곡선)

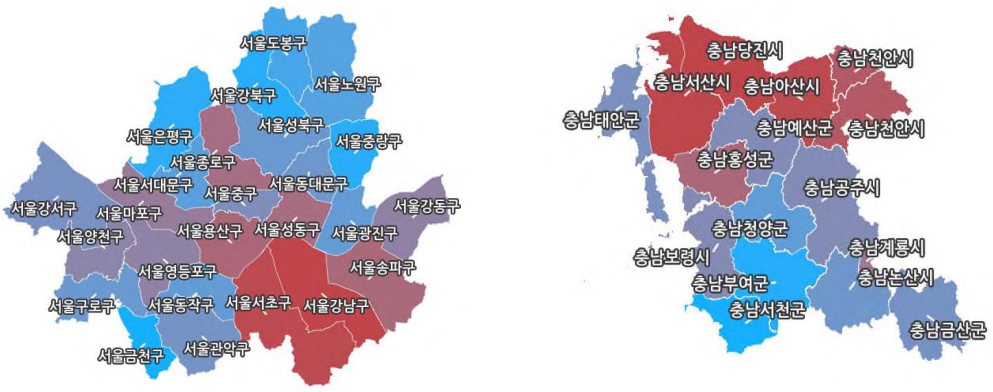
추가적으로 전체 교차소지역 수준에서의 정량적 개선 효과를 확인한 결과, 1,374개 교차소지역 중 약 75%에서 다중 랜덤효과 모형의 MSE가 직접추정보다 낮게 나타나 전반적인 오차 감소 경향이 확인되었다. 또한 공표 기준(CV≤25%) 충족 비율은 직접추정 72.7%에서 다중 랜덤효과 모형 87.6%로 약 14.9%p 증가하였다.

종합하면, 본 연구에서 적용한 모의실험 기반 MSE·CV 비교는 단순한 모형 적합도 지표(AIC, BIC)를 넘어, 표본설계를 반영한 불확실성 평가를 통해 공식통계 맥락에서 소지역 추정치의 신뢰성과 공표 가능성을 실증적으로 판단할 수 있는 기준을 제공한다. 특히 표본 희소성이 심화되는 교차 소지역에서는 다중 랜덤효과 모형이 정보 차용을 통해 분산을 효과적으로 감소시키고 추정 안정성을 유의미하게 개선함을 확인하였다. 이는 지역단위뿐 아니라 직업분류 차원의 이질성을 동시에 고려하는 것이 소지역 소득 추정의 품질을 제고하는 데 핵심적임을 시사한다.

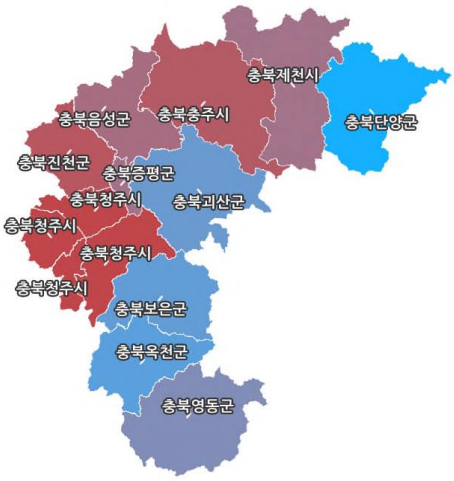
4.2 인접한 시군구별/직업분류별 소득 구조

본 절은 4.1절의 다중 랜덤효과 모형 추정치를 활용하여 평균소득이 유사한 인접 시군구 후보군의 직업분류별 소득 구조가 동형인지 이형인지를 판별해 “평균이 비슷하면 같은 정책으로 충분한가”라는 실무적 질문에 정량 근거를 제시한다. 후보군은 시군구별 평균소득으로 1차 식별하고, <그림 4.4> 지도 시각화를 통해 지리적 인접성을 교차 확인하여 선정하였다. 시군구 평균소득은 표본 규모가 충분한 수준이므로 직접추정치를 사용하였으며 시군구×직업분류 단위의 표본이 극소한 수준에서는 추정 안정성을 고려해 다중 랜덤효과 모형 추정치를 적용하였다. <그림 4.4>의 지도 시각화에서 붉은색 계열은 고소득, 파란색 계열은 저소득을 의미하며 동일 도(특·광역시) 내 인접 시군구들의 상대적 위상을 직관적으로 확인할 수 있다. 본 절에서는 <그림 4.4>의 대표적인 사례 중심으로 결과를 해석한다. 전국 단위로 확장 시 많은 후보군들을 찾을 수 있다.

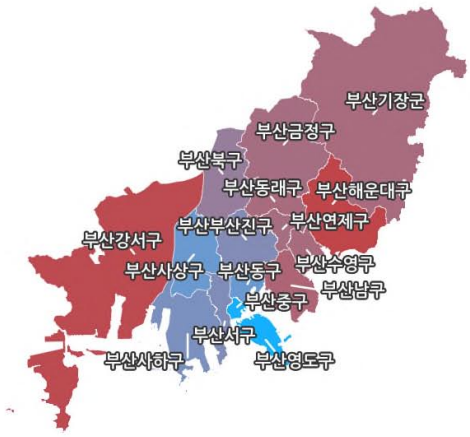
<그림 4.4> 시도별 평균소득 시각화



a. 서울 구별 평균소득 시각화 b. 충청남도 시군구별 평균소득 시각화



c. 충청북도 시군구별 평균소득 시각화



d. 부산 구별 평균소득 시각화

(i) 동형: 부산 사하구-부산 사상구

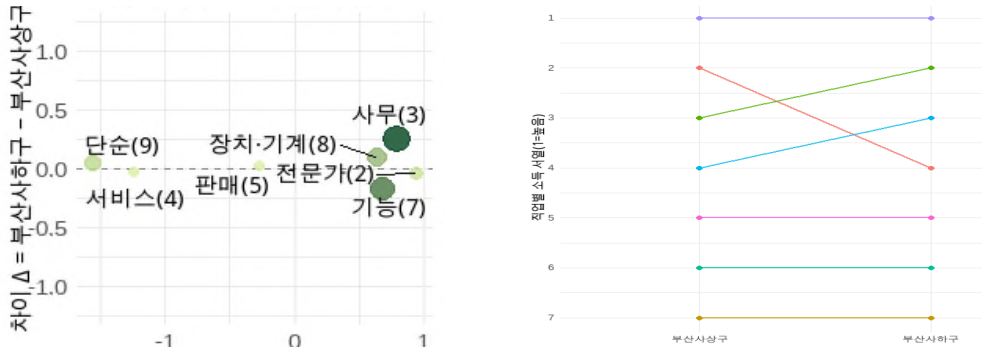
<표 4.6> 부산 사하구 - 부산 사상구 직업분류별 평균 (단위: 만원)

부산 사하구 (257)	전문가(326) > 사무(326) > 장치·기계(309) > 기능원(302) > 판매(240) > 서비스(168) > 단순(148)
부산 사상구 (254)	전문가(329) > 기능원(315) > 사무(308) > 장치·기계(302) > 판매(238) > 서비스(170) > 단순(144)

부산 사하구-부산 사상구는 인접 지역이며, <표 4.6> 기준 직업별 평균 수준이 사하구는 257만원, 사상구는 254만원으로 전반적으로 유사하다. 공통 7개 직업군의 서열 상관(ρ)=0.865로 높은 편이며, 분포차 지표인 L1-share=0.013($\approx$ 1.31%p 재배치)로 분포 형태가 매우 유사하다. <그림 4.5>의 a. Bland - Altman 관점에서도 점들이 $y=0$ 선 주변에 고르게 분포해 체계적 괴리가 거의 없고, 일부 직업에서만 미세 차이가 관찰된다.

구체적으로 사무(3)는 사하구가 사상구보다 약 +18만원, 기능원(7)은 −13만원, 장치·기계(8)는 +7만원 정도의 차이를 보여 소폭의 교차가 있으나, 전문가·판매·서비스·단순은 사실상 근접하다. 즉, 허용 CV 기준을 충족하는 공통 직업군 전반에서 상대적 위치와 분포 형태가 거의 동일하므로, 사하구 - 사상구는 동형으로 분류하는 것이 타당하다.

<그림 4.5> 부산 사하구-부산 사상구 산점도 및 서열 변화



a. 부산 사하구 - 부산 사상구 산점도 b. 부산 사하구 - 부산 사상구 서열 변화

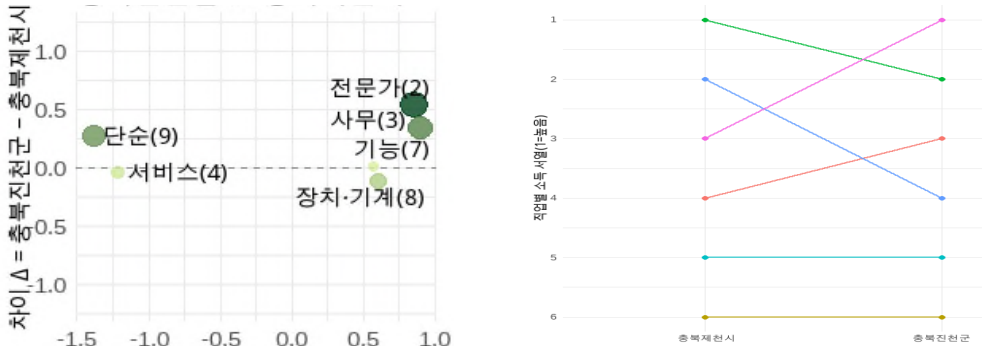
(ii) 이형: 충북 진천군-충북 제천시

<표 4.7> 충북 진천군-충북 제천시 직업분류별 평균 (단위: 만원)

충북 진천군 (295)	전문가(367) > 사무(362) > 기능원(321) > 장치·기계(319) > 서비스(168) > 단순(167)
충북 제천시 (266)	사무(334) > 장치·기계(328) > 전문가(322) > 기능원(320) > 서비스(171) > 단순(144)

<표 4.7>의 소지역별 추정치를 바탕으로 계산한 공통 직업군 6개 기준 순위 상관계수가 ρ = 0.714으로 서열 유사성이 낮으며, 분포차는 L1-share = 0.032로 약 3.2%p의 비중 재배치가 필요하다. <그림 4.6>의 a. 산점도(Δ=진천 - 제천) 와 b. 서열변화 그래프를 보면 직업별 차이는 전문가·사무가 양(+) 편차, 장치·기계가 음(−) 편차로 갈라지며, 상위 1~3위권에서 반복 교차가 확인된다. 두 후보군 모두 분포 비중(L1)은 비교적 가깝지만, 서열 구조의 유사성이 낮아 이형으로 판정하며 정책 우선순위가 달라질 수 있다.

<그림 4.6> 충북 진천군-충북 제천시 산점도 및 서열 변화



a. 충북 진천군 - 충북 제천시 산점도 b. 충북 진천군 - 충북 제천시 서열 변화

4.3 전국 상·하위 소득권 직업분류별 소득 구조 파악

앞에서는 지리적 인접성에 따른 소득구조의 형태를 파악하였다. 그러나 정책·연구 현장에서는 지리적 인접성과 무관하게 전국 단위에서 평균 수준이 비슷한 지역들을 비교해야 하는 경우가 빈번하다. 본 절의 분석은 추정치 선택에 있어 앞 절과 동일한 기준을 따르되, 시군구 평균소득은 직접추정치로, 직업분류까지 포함한 소지역 수준에서는 다중 랜덤효과 모형 추정치를 활용한다. 신뢰 가능한 비교를 위해 양 지역 모두 CV≤0.50인 직업군만 포함하는 동일 기준을 적용하여, 전국 상위권(고소득) 및 하위권(저소득) 지역군의 형태(서열)·분포 유사도를 체계적으로 점검한다. 정량 평가 지표는 앞 장과 동일하다.

4.3.1 상위권

상위권 비교는 수도권 핵심구를 포함한 전국 고소득 상위대 시군구를 대상으로 하였다. 상위권 지역들은 공통적으로 전문가(2)·사무(3) 비중이 높고, 장치·기계(8)가 일정 수준 이상을 차지하지만, 상대 서열과 격차(△)는 후보군에 따라 달라진다.

(i) 동형: 서울 서초구–서울 강남구

<표 4.8> 서울 서초구 - 서울 강남구 직업분류별 평균 (단위: 만원)

서울 서초구 (471)	전문가(469) > 사무(398) > 판매(282) > 서비스(202) > 단순(189)
서울 강남구 (480)	전문가(486) > 사무(375) > 판매(307) > 서비스(243) > 단순(158)

<표 4.8>에서 서초구와 강남구는 순위 상관계수 ρ =1로 서열이 완전 일치하며 분포차 또한 0.041으로, 약 4.1%p의 비중만 재배치하면 두 지역의 직업별 소득 분포 형태가 일치하여 동형으로 판정한다.

(ii) 이형: 서울 양천구–울산 동구

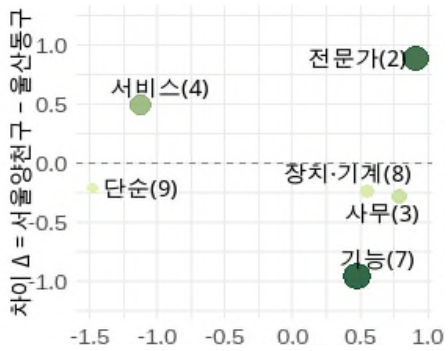
<표 4.9> 서울 양천구–울산 동구 직업분류별 평균 (단위: 만원)

서울 양천구 (359)	전문가(419) > 사무(355) > 장치·기계(336) > 기능원(297) > 서비스(218) > 단순(155)
울산 동구 (326)	기능원(383) > 사무(381) > 장치·기계(357) > 전문가(339) > 서비스(174) > 단순(174)

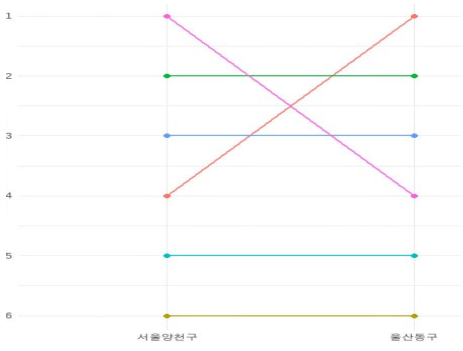
<표 4.9>를 보면 서울 양천구–울산 동구는 평균 수준은 유사하나, ρ = 0.464, L1 = 0.074로 두 지역의 직업별 소득 분포 형태를 일치시키려면 총 7.4%p 정도의 비중 재배치가 필요함을 시사한다. <그림 4.7>의 a에서 점들이 0선에서 체계적으로 이탈하고 산포가 넓다. 울산 동구는 기능·제조(7·8) 축이, 서울 양천구는 전문·사무(2·3) 축이

상대적으로 우위를 차지하고 <그림 4.7> b에서 상위 1~3위권에서 반복 교차가 나타나 이형으로 판정된다.

<그림 4.7> 서울 양천구 - 울산 동구 산점도 및 서열 변화 (상위권 이형)



a. 서울 양천구-울산 동구 서열 변화



b. 서울 양천구-울산 동구 산점도

4.3.2 하위권 (이형)

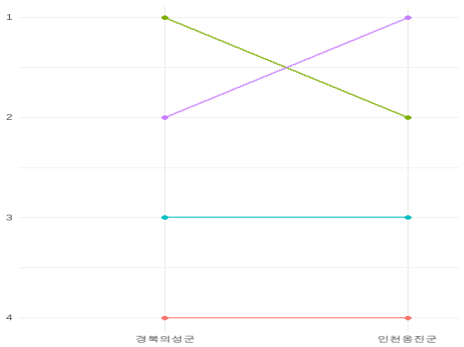
<표 4.10> 경북 의성군-인천 옹진군 직업분류별 평균 (단위: 만원)

경북 의성군 (185)	사무(298) > 전문가(280) > 서비스(167) > 단순(100)
인천 옹진군 (187)	전문가(370) > 사무(303) > 서비스(212) > 단순(101)

<표 4.10>를 보면 두 지역이 하위권 평균임에도 순위 상관계수는 0.8로 직업분류별 서열과 수준이 살짝 어긋나고 분포차는 0.061로 크다. 또한 <그림 4.8>의 a의 산점도가 0선에서 멀고 <그림 4.8>의 b에서 상위 직업군에서 서열이 교차한다. 옹진은 전문가·사무가 우세하고, 의성은 사무가 우세하고 전문가가 약해 두 후보군의 우세축이 다른 이형으로 분류 가능하다.

<그림 4.8> 경북 의성군 - 인천 옹진군 산점도 및 서열변화 (하위권 이형)

a. 경북 의성군-인천 옹진군 산점도



b. 경북 의성군 - 인천 옹진군 서열 변화

시군구×직업분류 단위에서 소득 구조를 비교·분석한 결과, 평균 소득 수준이 유사한 지역이라 하더라도 직업분류별 소득의 상대적 배열과 우세 축은 충분히 다를 수 있음이 확인된다. 이는 지리적으로 인접한 지역 간 비교와 전국 상·하위 소득권 사례 모두에서 공통적으로 관찰되는 특징이다. 즉, 공간적 근접성이나 평균 소득의 유사성이 직업별 소득 구조의 동일성을 보장하지 않으며, '평균 유사 = 구조 유사'라는 가정이 성립하지 않을 수 있음을 시사한다.

먼저 인접한 지역의 사례를 보면, 평균 소득 수준이 비슷하고 생활권을 공유하는 지역이라 하더라도 직업분류별 소득 구조에서는 뚜렷한 차이가 반복적으로 나타난다. 특히 전문·사무 축(2·3)과 제조·기능 축(7·8) 간의 상대적 우위 차이, 그리고 제조 내부에서의 장치·기계(8)와 기능원(7) 간 서열 분화가 일관되게 관찰된다. 이는 시군구 평균 소득만으로는 지역 내부의 노동시장 구조와 직종별 임금 결정 메커니즘을 충분히 포착하기 어렵다는 점을 보여준다.

이러한 현상은 전국 상·하위 소득권에 속한 지역 간 비교에서도 동일하게 확인된다. 예컨대 서울 양천구와 울산 동구는 평균 소득 수준이 근접함에도 불구하고 직업분류별 분포 형태의 상관이 낮고(ρ = 0.643), 상위 직종의 서열 교차가 관찰되어 구조적 이형성이 뚜렷하게 나타난다. 반면 서울 서초구와 강남구는 직업별 상대 위치와 격차가 거의 일치하여 매우 높은 상관(ρ = 1)을 보이며, 서열 교차가 관찰되지 않는 동형 구조를 보인다. 이는 동일한 평균 소득권으로 분류되는 지역이라 하더라도, 내부의 직업별 소득 구조가 동질적인 경우와 이질적인 경우로 구분될 수 있음을 시사한다. 나아가 이러한 동형 - 이형의 구분은 평균 수준에 기반한 일률적 정책 설계보다는, 지역별 소득 구조의 성격에 따라 상이한 정책 접근이 요구될 수 있음을 암시한다.

상위 소득권과 하위 소득권 모두에서 구조적 이형성을 가르는 주요 분기점은 전문·사무 축과 제조·기능 축의 상대적 우위에서 발생한다. 한 지역은 전문·사무직 중심의 소득 구조를 보이는 반면, 다른 지역은 제조·기능직의 상대적 비중이 높은 구조를 보일 수 있으며, 이 경우 평균 소득이 유사하더라도 임금 구조는 본질적으로 다르게 나타난다. 이러한 차이는 단순한 수준 비교를 넘어, 지역 노동시장의 성격과 정책 수요가 상이함을 시사한다.

세분화된 시군구×직업분류 단위의 소득 구조를 함께 고려할 경우, 유사한 평균을 가진 지역이라 하더라도 정책 우선순위와 개입 방식은 달라질 수 있으며 세분화 소지역 추정에서 안정적으로 수행 가능한 다중 랜덤효과 모형을 통해 지역 맞춤형 정책 설계와 직종별 지원 전략 수립을 기대할 수 있다.

5. 결론

본 연구는 국가통계의 소지역 추정에 다중 랜덤효과 혼합모형을 적용하여, 지역단위 이상의 세분화된 단위에서도 안정적인 소득 추정이 가능함을 실증적으로 보이는 데 목적이 있다. 특히 시군구×직업분류라는 고차원 교차 단위를 소지역으로 정의하고, 지역과 직업 두 축에서 동시에 정보공유(borrowing of strength)를 수행함으로써, 기존

직접추정이나 단일 랜덤효과 모형이 갖는 불안정성을 유의미하게 개선한다.

2024년 지역별고용조사 자료를 활용한 실증 분석 결과, 직접추정은 표본 규모가 충분한 일부 영역에서는 경쟁력을 보이지만, 시군구×직업분류 수준으로 세분화될수록 분산이 급격히 증가하고 추정치의 변동성이 커지는 한계를 보인다. 단일 랜덤효과 혼합모형은 이러한 불안정성을 일정 부분 완화하지만, 직업 간 이질성이 큰 상황에서는 수축 효과가 충분히 작동하지 못하는 경우가 발생한다. 이에 비해 다중 랜덤효과 혼합모형은 지역과 직업을 동시에 고려한 부분풀링을 통해 희소 영역에서도 추정치를 안정화하며, 평균제곱오차(MSE)와 변동계수(CV)를 기준으로 함께 비교할 때 전반적으로 가장 우수한 성능을 보인다. 이는 소지역 추정의 핵심 과제인 정확도와 공표 가능성을 동시에 충족하는 방법론적 이점을 명확히 보여준다.

본 연구는 추정의 정확도 비교에 그치지 않고, 시군구×직업분류별 소득 구조를 함께 분석함으로써 평균 중심의 소지역 추정이 갖는 한계를 보완한다. 인접한 지역 간 비교와 전국 상·하위 소득권 사례를 종합한 결과, 평균 소득 수준이 유사한 지역이라 하더라도 직업분류별 소득의 상대적 배열과 비중은 상당히 다를 수 있음이 확인되었다. 지리적으로 인접한 지역임에도 전문·사무직 중심 구조와 제조·기능직 중심 구조가 대비되는 사례가 관찰되었고, 전국 상·하위 소득권 내부에서도 직업군 간 소득 격차의 형태와 서열이 상이한 이형 사례가 다수 확인되었다. 이는 공간적 근접성이나 평균 소득 수준만으로는 지역 노동시장과 소득 분포의 실질적 특성을 충분히 설명하기 어렵다는 점을 시사한다.

이러한 구조 분석 결과는 다중 랜덤효과 모형을 통해서 안정적으로 확보 가능한 세분화 소지역 정보의 정책적 가치를 부각한다. 평균 수준이 유사한 지역이라 하더라도 직업군별 소득 구조에 따라 정책 수요와 개입 방향은 달라져야 하며, 구조적으로 동형인 지역군은 정책 경험의 공유와 확산이 가능하고, 이형인 지역군은 각자의 우세 직업 축에 기반 한 차별화된 전략이 요구된다. 예컨대 전문·사무직 중심 구조를 가진 지역에는 고숙련 인력 유치와 지식기반 산업 육성이, 제조·기능직 중심 구조를 가진 지역에는 공정 고도화, 직업훈련 및 전환교육 강화가 보다 적합할 수 있다. 이는 소지역 추정 결과를 단순 공표 통계가 아닌 정책 설계의 근거 자료로 활용하기 위해서는 구조 정보가 함께 제공될 필요가 있음을 의미한다.

본 연구에는 몇 가지 한계와 향후 연구 과제가 남아 있다. 첫째, 본 연구는 시군구×직업분류 교차 소지역 수준에서 다중 랜덤효과 모형의 적용 가능성을 평가하는 데 목적이 있어 지역 효과와 직업 효과를 독립적으로 가정한 기본 모형을 사용하였다. 다만 특정 지역에 특정 직업이 집중되는 구조를 고려할 경우, 지역과 직업 간 상호작용을 반영한 교차 랜덤효과 또는 상관구조를 포함하는 확장 모형이 보다 정교한 설명력을 제공할 수 있다. 그러나 교차 소지역 단위에서 표본이 매우 적은 경우가 다수 존재하는 자료 구조를 고려할 때, 이러한 확장 모형은 분산성분의 안정적 추정에 제약이 발생할 가능성이 존재한다. 이에 대한 체계적인 비교 분석은 향후 연구 과제로 남겨둔다. 둘째, 본 연구는 모의실험을 통해 각 방법론의 평균제곱오차(MSE)를 비교함으로써 소지역 추정량의 전반적인 변동성을 평가하였다. 그러나 모형 기반 소지역 추정을 단일 표본에 적용할 경우, 랜덤효과 분산성분 추정의 불확실성이 최종 추정량의 변동성에 추가적인

영향을 미칠 수 있다. 이에 분산성분 추정의 변동성을 보다 정교하게 반영한 MSE 추정 방법에 대한 검토는 향후 연구 과제로 남겨둔다. 셋째, 랜덤효과의 정규성 가정, 오차의 이분산성, 비선형 관계 가능성 등으로 인한 모형 오적합 위험이 존재한다. 향후 혼합분포 랜덤효과, 이분산 모형화, 스플라인 기반 비선형성 도입 등을 통해 모형의 강건성을 강화할 필요가 있다. 넷째, 본 연구는 횡단면 자료에 기반하므로 시간에 따른 소득 구조의 변화와 동태적 효과를 포착하지 못했다. 공간 상관과 시계열 구조를 결합한 시공간 소지역 추정으로의 확장은 중요한 연구 과제가 될 것이다.

종합하면, 본 연구는 다중 랜덤효과에 기반 한 소지역 추정과 소득 구조 비교를 하나의 일관된 절차로 통합함으로써 평균 중심의 단순 비교를 넘어 직업별 이질성의 형태와 서열을 정량적으로 해석하는 분석 틀을 제시한다. 이는 시군구×직업분류 수준까지 확장된 세분화 소지역 추정이 국가통계의 정확도 제고와 정책 활용 가능성을 동시에 충족할 수 있음을 보여 주는 실증적 근거로서 의미를 갖는다.

(2026년 2월 2일 접수, 2026년 3월 4일 수정, 2026년 3월 17일 채택)

참고문헌

김서영, 권순필 (2010). 소지역 실업자수 추정을 위한 로지스틱 선형혼합모형 기반 EBLUP 타입 추정량 평가.

김순영 (2019), 소지역 추정 기법의 최신 동향 연구. 통계개발원.

김진, 김재광 (2010). 시군 실업통계 작성을 위한 소지역 추정모형. Communications for Statistical Applications and Methods (응용통계논문집), 17(3), 337 – 347.

Battese, G. E., Harter, R. M., & Fuller, W. A. (1988). *An error-components model for prediction of county crop areas using survey and satellite data.* Journal of the American Statistical Association, 83(401), 28 – 36.

Benavent, R., & Morales, D. (2016). Multivariate Fay – Herriot models for small area estimation. Computational Statistics & Data Analysis, 94, 372 – 390.

Datta, G. S., & Lahiri, P. (2000). *A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems.* Statistica Sinica, 10(2), 613 – 627.

Duan, N. (1983). Smearing estimate: A nonparametric retransformation method. Journal of the American Statistical Association, 78(383), 605 – 610.

Eurostat (2013). *Guidelines on Small Area Estimation for City Statistics.*

Fay, R. E., & 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(366a), 269 – 277.

Ghosh, M., & Rao, J. N. K. (1994). *Small area estimation: An appraisal. Statistical Science,* 9(1), 55 – 76.

Marcis, L., Morales, D., Pagliarella, M. C., & Salvatore, R. (2023). Three-fold Fay – Herriot model for small area estimation and its diagnostics. Statistical Methods & Applications, 32(5), 1563 – 1609.

Marhuenda, Y., Molina, I., & Morales, D. (2013). Small area estimation with spatio-temporal Fay-Herriot models. Computational Statistics & Data Analysis, 58, 308 – 325.

Prasad, N. G. N., & Rao, J. N. K. (1990). The estimation of mean squared error of small-area estimators. Journal of the American Statistical Association, 85(409), 163 – 171.

Pfeffermann, D. Benedicte Terryn and Fernando A.S. Moura (2008) *Small area estimation under a two-part random effects model with application to estimation of literacy in developing countries,* Survey Methodology, Statistics Canada

Pfeffermann, D. (2013). *New important developments in small area estimation. Statistical Science,* 28(1), 40 – 68.

Rao, J. N. K., & Yu, M. (1994). Small-area estimation by combining time-series

and cross-sectional data. The Canadian Journal of Statistics, 22, 511 - 528.

Rao, J. N. K., & Molina, I. (2015). *Small Area Estimation (2nd ed.)*. Wiley.

Sho Kawano, Paul A Parker, Zehang Richard Li (2025) *Spatially selected and dependent random effects for small area estimation with application to rent burden,* Journal of the Royal Statistical Society Series A: Statistics in Society

Torabi, M., & Rao, J. N. K. (2005). *Mean squared error estimators of small area means using survey weights.* Proceedings of the Survey Methods Section, SSC Annual Meeting, June 2005.

Torabi, M., & Rao, J. N. K. (2010). *Mean squared error estimators of small area means using survey weights.* Canadian Journal of Statistics, 38(4), 598 - 608.

United Nations (2020). *Handbook on Small Area Estimation for Poverty, Inequality and Vulnerability.*

You, Y., & Zhou, Q. M. (2011). *Hierarchical Bayes small area estimation under a spatial model with application to health survey data.* Survey Methodology, 37(1), 25 - 37.

Small Area Estimation Based on the Regional Employment Survey Using Mixed Models with Multiple Random Effects[4)]

Seonmin Jin[5)] · Sangin Lee[6)]

Abstract

As demand for finer-grained and micro-level official statistics has increased, survey-based statistics face growing uncertainty at highly disaggregated regional and group levels due to sample sparsity. This study addresses this challenge by applying small area estimation (SAE) techniques to produce stable income estimates at the municipality×major occupation level. Using data from the 2024 Regional Employment Survey, we compare direct estimation, a single random-effects model, and a crossed multiple random-effects model, evaluating performance using Monte Carlo simuation-based mean squared error (MSE) and coefficient of variation (CV). The results show that the multiple random-effects model substantially improves estimation stability in areas with small sample sizes. Structural comparisons further reveal heterogeneous occupational income patterns across regions with similar average income levels. Overall, this study demonstrates that crossed multiple random-effects models offer a practical approach for meeting the increasing demand for micro-level official statistics.

Key words : Small-area estimation, Mixed-effects models, Multiple random effects, Regional Employment Survey

4) This paper is a revised and extended version of the master's thesis of Seonmin Jin, and it was supported by the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (RS-2022-NR068754)

5) First Author, Ministry of Data and Statistics, E-mail: seonmin77@korea.kr

6) Corresponding author, Professor, Dept. of Statistics, Chungnam National University, E-mail: sanginlee44@gmail.com