Journal of the Korean
Data & Information Science Society
2000, vol. 11, No. 1, pp. 111 ~ 126

소지역 통계 생산을 위한 추정방법

김영원 [1] · **성나영** [2]

요 약

지방자치제 실시에 따라 우리나라에서도 전국 또는 도 단위의 통계 뿐만 아니라 시·군·구 등의 소지역 통계에 대한 수요가 증대되고 있다. 하지만 정부통계 생산을 위해 실시되는 표본조사의 경우 시(특별시, 광역시) 및 도별 통계생산을 목적으로 하기 때문에 신뢰성 있는 소지역 통계를 산출하는 것이 불가능하고, 따라서 이런 소지역 통계생산을 위해 간접추정기법을 적극적으로 활용하는 것이 필요하다. 본 논문에서는 정부통계 생산을 위한 소지역 통계기법의 도입 및 활용 가능성을 검토해 보기 위해 인천광역시 숙박 및 음식점업의 총매출에 대한 구별 소지역 통계를 산출할 수 있는 여러 가지 간접추정방법을 제시하고, 아울러 도소매업 총조사 자료를 이용하여 제시된 간접추정량들의 효율성을 비교 분석해 보고자 한다.

주제어: 소지역 추정, 합성추정량, 복합추정량, EBLUP추정량, 도소매업 통계조사

1. 서 론

지방자치제 실시에 따라 우리나라에서도 전국 또는 도 단위의 통계 뿐만 아니라 시·군·구 등의 소지역 통계에 대한 요구가 증대되고 있다. 그러나 기존의 전국 단위나 도 단위 통계 산출을 위해 설계된 표본의 경우 소지역별 표본수가 극히 적고 심지어 특정 시·군·구 등에 대해서는 선정된 표본이 전혀 없을 수도 있다. 또한 현실적인 비용을 고려할 때 전국의 시·군·구 등의 소지역에 대한 신뢰성 있는 통계를 얻기 위해 새로운 표본조사를 실시하는 것은 불가능하다고 할 수 있다. 소지역 통계(small area statistic)란 이렇게 표본수가 극히 적어 우리가 일반적으로 사용하는 방법으로는 신뢰성 있는 추정이 불가능한 경우에 사용하는 통계분석기법이다. 경우에 따라 '소지역'이란 시나 군, 구와 같이 지리적으로 작은 지역을 나타낼 뿐 아니라 '소영역' 즉, 특정 연령-성별-인종 그룹과 같은 작은 부차모집단을 나타낼 수도 있다.

[1](140-742) 서울 용산구 청파동 2가 숙명여자대학교 통계학과 교수
[2](140-742) 서울 용산구 청파동 2가 숙명여자대학교 통계학과 대학원

우리나라 통계청에서 실시되는 도소매업 통계조사, 경제활동인구 통계조사 등의 주요 표본설계는 대부분 16개 시(특별시, 광역시) 및 도별 집계를 목적으로 표본설계되어 있고, 따라서 이런 표본조사에서의 표본크기는 이들 16개 시도지역을 기준으로 적정 수준의 정도(precision)를 유지하도록 결정되기 때문에 소지역에 해당하는 표본크기는 매우 작아지게 된다. 이 경우 각 지역에 해당하는 표본만을 사용하는 직접(direct)추정량은 소지역에 대해서는 신뢰성 있는 추정값을 제공하지 못한다. 이에 반해 간접(indirect)추정방법은 부족한 표본크기를 보완하는 방법으로 유사한 지역의 정보 또는 그 지역의 과거 정보를 이용하거나 센서스 또는 다른 통계조사의 관련 정보를 이용하는 것이다.

간접추정방법을 통한 소지역 통계분석기법은 11세기의 영국과 17세기의 캐나다에서 그 사용이 발견되고 있으며(Brackstone, 1987), 그 이후 이에 관한 다양한 연구결과들을 찾아볼 수 있으며, 현재 많은 나라에서 공식 정부통계를 생산하는 데 있어 이를 적극적으로 활용하고 있다. Ghosh와 Rao (1994)는 이런 소지역 추정기법에 대한 적용사례 및 기법들을 체계적으로 정리하고 있으며, Rao (1999)는 이 분야에서 개발되고 있는 모형기반(model-based) 추정기법들을 중심으로 최근 연구동향을 제시하고 있다.

이런 현실적인 요구에도 불구하고 아직 우리나라 통계청, 농림부, 한국은행 등 통계작성기관에서는 이러한 소지역 통계분석기법에 대한 연구가 미진한 상태이고 아직까지 이런 기법을 활용한 분석사례를 우리나라에서는 찾아볼 수 없다.

한편 성공적으로 소지역 통계분석기법을 사용하기 위해서는 표본조사에서 직접 얻어지는 관측결과 뿐만 아니라 기존의 센서스나 다른 통계조사를 통해 얻을 수 있는 관련 정보를 효과적으로 활용하는 것이 매우 중요하기 때문에 일반적으로 알려져 있는 간접추정기법을 그대로 사용하는 것이 아니라 적용사례에 따라 어떤 보조 정보를 활용하여 기존의 추정기법을 어떻게 수정보완하여 활용할 것인가를 심도 있게 검토하여 사례에 따른 최적의 적용방법을 찾는 작업이 필요하다. 따라서 소지역 추정기법의 활용 활성화를 위해서는 사례별로 심층적인 분석이 요구되며, 이런 관점에서 소지역기법의 도입 가능성을 검토해 보기 위해 비록 제한된 경우에 해당하지만 인천광역시의 숙박 및 음식점업에 대한 소지역 추정기법의 적용 사례에 대해 중점적으로 논의해 보고자 한다.

이를 위해 본 논문에서는 이 분야의 대표적인 논문 중에 하나인 Ghosh와 Rao (1994)의 결과를 참고로 우선 기존의 소지역 통계분석기법에서 적용되는 추정방법을 유형에 따라 간단히 소개하고, 1996년 12월 31일을 기준으로 통계청에서 실시한 도소매업 총조사 결과 중 인천광역시의 숙박 및 음식점업 자료를 기준으로 매년 통계청에서 실시하는 도소매업 및 서비스업 통계조사를 위한 표본설계에 따라 표본을 추출하는 사례연구를 통해 소지역에 해당하는 구별 숙박업 및 음식점업의 연간 총매출액에 대한 직접추정값과 여러 가지 형태의 간접추정값의 정확도를 비교분석하여 우리나라 정부 공식통계에 소지역 통계기법의 도입 및 활용 가능성을 검증해 보고자 한다.

2. 소지역 추정방법

2.1 합성추정량

Gonzalez(1973)에 따르면 합성추정값(synthetic estimate)은 다음과 같이 정의된다. "어느 한 대지역에 대해 표본조사를 통해 불편추정값이 얻어지고, 소지역의 (분포)특징이 대지역과 같다는 가정아래 이 추정값이 소지역에 대한 추정값을 유도하는데 사용될 때, 소지역에 대한 추정값을 합성추정값이라고 부른다."

이 추정방법의 특징을 살펴보면(Cassel 등, 1987), 첫째, 일정 표본크기의 한 지역에서 전통적인 방법으로 얻어진 추정량보다 더 정확한 추정량을 얻을 수 있다. 둘째, 표본에서 관찰값의 수가 너무 작아서 전통적인 방법으로는 추정값을 구할 수 없는 지역에 대해서도 추정값을 구할 수 있다.

한편 합성추정량을 사용하는 경우 요구되는 사항은 첫째, 관심변수와 관련이 있는 보조정보가 존재해야 한다. 관련이 높을수록 더 좋은 추정량이 만들어진다. 둘째, 모형가정이 만족되어야 한다. 즉, 대지역에서 관찰될 수 있는 관계가 소지역에 대해서도 만족해야 한다. 모형가정이 만족하지 않을 경우 편의가 발생하게 된다.

이 방법은 간단하면서도 일반적인 표본설계에 적용하기 쉽고, 유사지역으로부터 정보를 가져와 활용함으로써 정도의 개선 가능성이 있어서 소지역 추정에 많이 쓰인다.

이를 구체적으로 살펴보면, 모집단을 큰영역(large domain)으로 나누고 i지역, g영역에 대한 총계를 Y_{ig}로 나타내자. 예를 들어 관심변수가 사업체 매출액인 경우 모든 사업체는 조직형태 또는 사업장 면적 등의 지역 개념과 다른 별도의 특성에 따라 몇 개의 큰 영역으로 구분될 수 있다. 그러면 g영역에 대한 총계는 $Y_{\cdot g} = \sum_i Y_{ig}$가 되고, i지역에 대한 총계는 $Y_{i\cdot} = \sum_g Y_{ig}$가 된다. 표본조사로부터 영역 g에 대한 총계, $Y_{\cdot g}$의 직접 추정량, $\hat{Y}_{\cdot g}'$을 얻을 수 있으며, 소지역 i에 대한 총계 Y_i의 추정량으로 다음과 같은 합성추정량을 사용할 수 있다.

$$\hat{Y}_{(SYN),i} = \sum_g \left(\frac{X_{ig}}{X_{\cdot g}}\right) \hat{Y}_{\cdot g}' \tag{2.1}$$

여기서 X_{ig}는 i지역 g영역에서 보조변수의 총계, $X_{\cdot g}$는 g영역 전체에 대한 보조변수의 총계이고, $X_{\cdot g} = \sum_i X_{ig}$이다. 이런 보조변수에 관한 정보는 센서스 또는 다른 통계자료를 통해 얻는다고 가정한다. 지역별 추정량인 식(2.1)을 모두 합하면 모집단 총계 Y의 직접 추정량, $\hat{Y}'(= \sum_g \hat{Y}_{\cdot g}')$와 일치한다. 즉, $\sum_i \hat{Y}_{(SYN),i} = \hat{Y}' = \sum_g \hat{Y}_{\cdot g}'$.

식(2.1)에서 사용된 $\hat{Y}_{\cdot g}'$은 일반적으로 비추정량의 형태를 띤다. 즉, $\hat{Y}_{\cdot g}' = (\hat{Y}_{\cdot g}/\hat{X}_{\cdot g})X_{\cdot g}$이다. 여기서 $\hat{X}_{\cdot g}$와 $\hat{Y}_{\cdot g}$는 표본조사에서 직접 구한 추정량이다. 이 형태의 $\hat{Y}_{\cdot g}'$을 사용할 경우 (2.1)식의 합성추정량은 $\hat{Y}_{(SYN),i} = \sum_g X_{ig}(\hat{Y}_{\cdot g}/\hat{X}_{\cdot g})$이 된다. $\hat{Y}_{\cdot g}'$이 근사적으로 설계기반 불편성(design-based unbiased)을 갖는다면 $\hat{Y}_{(SYN),i}'$의 편의는 다음과 같다.

$$E(\hat{Y}_{(SYN),i}) - Y_i = \sum_g Y_{ig}\left(\frac{Y_{\cdot g}}{X_{\cdot g}} - \frac{Y_{ig}}{X_{ig}}\right)$$

위의 식에서 $Y_{.g}/X_{.g} = Y_{ig}/X_{ig}$일 때 편의가 0이 됨을 알 수 있다. 특수한 경우로 X_{ig}를 i지역, g영역의 모집단 크기, N_{ig}로 놓으면 $\bar{Y}_{ig} = \bar{Y}_{.g}$일 때 편의가 0이 되는 것이다. 이러한 가정은 매우 제약적이어서 합성추정량의 편의의 발생은 불가결하다고 할 수 있다. 편의가 발생하지만 식(2.1)을 참고로 하면 이 추정량의 분산은 $\hat{Y}'_{.g}$의 분산에 따라서만 달라짐을 알 수 있다.

2.2 복합추정량

직접추정량의 불안정성과 합성추정량의 편의를 서로 보완할 수 있는 방법으로 두 추정량의 가중평균을 사용하는 방법을 생각할 수 있는데, 이렇게 두 추정량의 가중평균을 사용하는 것이 복합추정량(composite estimator)이다. 복합추정량의 일반적인 형태는 아래와 같다.

$$\hat{Y}_{(COMP),i} = w_i\hat{Y}_{1i} + (1-w_i)\hat{Y}_{2i}$$

여기서 $\hat{Y}_{1i}$은 직접추정량이고, $\hat{Y}_{2i}$은 간접추정량으로, 예를 들어 $\hat{Y}_{1i}$은 불편추정량 $(\hat{Y}_i)$을, $\hat{Y}_{2i}$은 합성추정량 $(\hat{Y}_{(SYN),i})$을 사용할 수 있다. w_i는 적당하게 선택된 0과 1사이의 가중값이다.

복합추정량을 활용하기 위해서는 적절한 가중값 w_i를 결정하는 과정이 매우 중요하다. 이를 위해 Purcell과 Kish (1979) 등이 제시한 것과 같이 MSE를 최소로 하는 최적가중값을 찾는 방법이 있다.

한편 최적가중값 외에 표본크기에 의해 결정하는 단순가중값이 있다. 단순가중값은 모집단 크기와 표본크기 또는 보조변수의 총계에 따라서만 가중값이 달라진다. 먼저 Drew 등 (1982)은 다음과 같은 가중값을 사용하는, 표본크기의존추정량(sample-size dependent estimator)을 제안했다.

i지역 모집단 크기, N_i의 직접 불편추정량을 $\hat{N}_i$이라고 할 때,

$$w_i(D) = \begin{cases} 1, & \hat{N}_i \geq \delta N_i \text{ 일 때} \\ \hat{N}_i/(\delta N_i), & \text{다른 경우} \end{cases} \tag{2.2}$$

여기서 N_i는 알고 있는 값이고, δ는 합성추정량의 기여도를 결정해 주는, 임의로 선정되는 상수값이다. 예를 들어 Canadian Labour Force Survey의 경우 $\delta = 2/3$를 사용하고 있다. 또한 Sarndal과 Hidiroglou (1989)도 식(2.2)와 유사한 가중방법을 제안하고 있다.

2.3 BLUP 추정량

BLUP(best linear unbiased prediction) 추정량은 Henderson (1950)이 제안한 추정량으로 선형불편추정량 중 MSE를 최소로 하는 추정량이다. 즉, 최소분산불편추정량(BLUE, best

linear unbiased estimator)과 유사한 개념의 추정량이다. 그리고 EBLUP(empirical BLUP) 추정량은 BLUP 추정량에서 분산항에 추정값을 대입하여 얻는 추정량이다.

Ghosh와 Rao (1994)는 다음과 같은 내포오차회귀모형(nested error regression model)을 사용한 EBLUP 추정량을 제시하고 있다.

$$y_{ij} = x_{ij}^T\beta + v_i + e_{ij}, \quad i = 1, \ldots, m, \ j = 1, \ldots, N_i \tag{2.3}$$

여기서 y_{ij}는 i지역의 j번째 관찰값을 나타내고, v_i는 $E(v_i) = 0$, $V(v_i) = \sigma_v^2$인 iid 확률변수로 랜덤효과(random effects)를 나타낸다. e_{ij}는 $e_{ij} = \tilde{e}_{ij}k_{ij}$이고, $\tilde{e}_{ij}$는 $E(\tilde{e}_{ij}) = 0$, $V(\tilde{e}_{ij}) = \sigma^2$인 iid 확률변수이며 k_{ij}는 알려진 상수이다. 그리고 N_i는 i지역의 모집단 크기를 나타낸다.

한편 i지역 평균, $\bar{Y}_i$는 표본에 뽑힌 부분과 표본에 뽑히지 않은 부분으로 나누어 아래와 같이 표현할 수 있다.

$$\bar{Y}_i = f_i\bar{y}_i + (1 - f_i)\bar{y}_i^* \tag{2.4}$$

여기서 $f_i = n_i/N_i$이고, $\bar{y}_i$는 표본에 뽑힌 원소들의 평균, $\bar{y}_i^*$는 표본에 뽑히지 않은 원소들의 평균을 나타낸다. 결국 $\bar{Y}_i$의 추정이란 표본에 뽑히지 않은 원소들의 평균을 예측하는 것과 같아진다.

$\bar{Y}_i$의 BLUP 추정량은 다음과 같이 얻어진다. 먼저 표본에 뽑힌 원소들에 대해 (2.3)의 모형을 사용하여 $\bar{x}_{ij}^T\tilde{\beta} + v_i$의 BLUP 추정량을 구한다. 다음으로 위에서 구한 $\bar{x}_{ij}^T\tilde{\beta} + v_i$의 BLUP 추정량을 식(2.4)의 $\bar{y}_i^*$에 대입하여 $\bar{Y}_i$의 BLUP 추정량을 얻는다.

한편 $\bar{x}_{ij}^T\tilde{\beta} + v_i$의 BLUP 추정량은 다음과 같다.

$$\bar{x}_i^{*T}\tilde{\beta} + \gamma_i(\bar{y}_{(w)i} - \bar{x}_{(w)i}^T\tilde{\beta})$$

여기서 $\tilde{\beta}$는 β의 BLUE이고, $\gamma = \sigma_v^2/(\sigma_v^2 + \frac{\sigma^2}{w_{i.}})$이다. 그리고 $w_{i.} = \sum_{j=1}^{n_i} w_{ij}$, $w_{ij} = k_{ij}^{-2}$이며, $\bar{y}_{(w)i}$, $\bar{x}_{(w)i}$는 가중값 w_{ij}를 갖는 가중평균, $\bar{x}_i^*$는 X_i^*의 평균을 나타내는 벡터이다(Prasad and Rao, 1990). BLUE $\tilde{\beta}$는 $\{(y_{ij} - \gamma_i\bar{y}_{(w)i})/k_{ij}, (x_{ij} - \gamma_i\bar{x}_{(w)i})/k_{ij}\}$와 같이 변환된 자료를 가지고 구한 최소자승추정량이다(Fuller and Battese, 1973).

결국 $\bar{Y}_i$의 BLUP 추정량은 다음과 같다.

$$\hat{\bar{Y}}_{(BLUP),i} = f_i\bar{y}_i + (1 - f_i)\{\bar{x}_i^{*T}\tilde{\beta} + \gamma_i(\bar{y}_{(w)i} - \bar{x}_{(w)i}^T\tilde{\beta})\} \tag{2.5}$$

식 (2.5)에 (σ_v^2, σ^2)의 일치추정량인 $(\hat{\sigma}_v^2, \hat{\sigma}^2)$을 대입하면 EBLUP(empirical BLUP) 추정량, $\hat{\bar{Y}}_{(EBLUP),i}$이 얻어진다. Fuller와 Battese(1973)는 'method of fitting constants'를 이용해서 (σ_v^2, σ^2)의 추정량, $(\hat{\sigma}_v^2, \hat{\sigma}^2)$을 다음의 절차에 따라 구한다.

먼저, $(y_{ij} - \bar{y}_{(w)i})/k_{ij}$을 $(x_{ij} - \bar{x}_{(w)i})/k_{ij}$에 회귀시켜 자유도 v_1인 잔차제곱합, $SSE(1)$을 구한다. 다음으로 y_{ij}/k_{ij}을 x_{ij}/k_{ij}에 회귀시켜 잔차제곱합, $SSE(2)$를 구한다. 그러면 $\hat{\sigma}^2 = v_1^{-1}SSE(1)$이고, $\hat{\sigma}_v^2 = max(\tilde{\sigma}_v^2, 0)$이다. 그리고 $\tilde{\sigma}_v^2 = \eta_*^{-1}[SSE(2) - (n - p)\hat{\sigma}^2]$이다. 여기서 $\eta_* = \sum_i w_{i.}(1 - w_{i.}\bar{x}_{(w)i}^T A_1^{-1}\bar{x}_{(w)i})$, $A_1 = \sum_i\sum_j w_{ij}x_{ij}x_{ij}^T$이다.

3. 사례연구

우리나라 통계청에서는 도·소매업 실태 파악을 위해 5년마다 전수조사에 해당하는 '도·소매업 총조사'(통계청, 1998)를 실시하고 있고, 매년 표본조사에 해당하는 '도·소매업 및 서비스업 통계조사'(통계청, 1997a)를 실시한다. 도·소매업 및 서비스업 통계조사는 전국 단위나 시·도 단위의 통계를 위해 표본설계되기 때문에 도·소매업 총조사가 실시되는 해에 대해서는 소지역 통계의 산출이 가능하지만 그렇지 않은 해에 대해서는 이러한 소지역 통계를 제시하지 못하고 있는 실정이다. 그러나 지방자치제의 실시 등 다양한 이유에 따라 소지역 통계에 대한 수요가 날로 증가하고 있다.

한편, 매년 '사업체 기초통계조사'(통계청, 1997b)가 모든 사업체에 대해 실시되고 있고, 이를 통해 모든 사업체에 대한 종사자수를 파악할 수 있다. 따라서 사업체 기초통계조사 결과 얻어진 모든 사업체에 대한 종사자수를 효과적으로 활용하면 도·소매업 및 서비스업 통계조사 결과를 가지고 효율적인 소지역 통계를 얻을 수 있다.

본 사례연구에서는 방대한 총조사자료를 활용해야 하기 때문에 적절한 규모라고 판단되는 인천광역시로 연구대상 지역을 한정하였고, 아울러 다른 도·소매업종에 비해 각 소지역으로 사업체가 널리 산재되어 있는 숙박업 및 음식점업을 연구대상 업종으로 선정하였다.

3.1 표본설계 현황

현행 도·소매업 및 서비스업 통계조사의 대략적인 표본설계는 다음과 같다. 산업세분류 및 16개 시·도 지역별로 부차모집단을 설정하고, 종사자수에 의해 층화를 한다. 예를 들어, 호텔업과 종사자수 40인 이상의 사업체에 대해서 전수층으로 분류하고, 나머지는 표본층으로 분류한다. 전수층에서는 모든 사업체를 표본으로 추출하며 표본층에서는 판매액 순서에 의해 자료를 나열한 후 산출된 표본규모만큼 계통추출방법에 의해 추출한다.

본 사례연구에서는 1997년에 실시한(1996년도 기준) 도·소매업 총조사 결과에서 나온 인천광역시의 숙박 및 음식점업 자료를 1998년에 실시된 도·소매업 통계조사 표본설계에 따라 표본을 추출하여 소지역 통계분석기법을 이용해 구별 숙박업 및 음식점업에 대한 총매출액을 추정해 보고자 한다. 1998년에 실시된 도·소매업 통계조사를 위한 표본설계는 전국 및 16개 시·도 지역별 산업소분류별 추정을 목적으로 하는 표본설계이기 때문에 시·도내 소지역에 해당하는 구별 숙박업과 음식점업(숙박 및 음식점업의 산업소분류)에 대한 신뢰성있는 추정값을 얻을 수 없다는 한계를 갖고 있다.

숙박 및 음식점업은 산업세분류에 의해 각각 숙박업 및 식당업, 주점업, 다과점업으로 분류된다. 산업세분류 업종별로 산출된 인천광역시의 숙박 및 음식점업의 표본크기는 390개이다. 이 중 전수층의 사업체 개수는 18개이다. 본 사례연구에서는 효율성 비교를 위해 전수층에 속하는 사업체를 제외하고 실제적으로 의미가 있는 표본층에 속하는 숙박업의 1,582개 업체, 음식점업의 24,893개 업체를 연구대상 모집단으로 설정하고 소지역 통계

기법을 적용하였다. 전수층을 제외한 모집단 및 표본크기는 〈표 1〉과 같고, 각 구별 모집단 크기 및 전수층을 제외한 표본추출 결과 구별로 뽑힌 표본크기는 〈표 2〉와 같다.

〈표 1〉 전수층을 제외한 산업세분류 업종별 모집단 및 표본크기

중분류업종	숙박 및 음식점업						계
소분류업종	숙박업		음식점업				
세분류 업종	숙박업	소계	식당업	주점업	다과점업	소계	
모집단 크기	1582	1582	15386	6667	2840	24893	26475
표본 크기	123	123	103	79	67	249	372

〈표 2〉 모집단 크기 및 구별 표본추출 결과

		숙박업	식당업	주점업	다과점업
중구	모집단 크기	249	1402	474	316
	표본크기	22	14	8	9
동구	모집단 크기	100	675	292	106
	표본크기	6	3	6	4
남구	모집단 크기	419	3051	2012	625
	표본크기	34	18	27	16
연수구	모집단 크기	49	1184	334	165
	표본크기	5	6	2	3
남동구	모집단 크기	87	2289	972	425
	표본크기	6	15	9	8
부평구	모집단 크기	255	2969	1296	573
	표본크기	17	19	15	13
계양구	모집단 크기	73	1329	500	242
	표본크기	5	10	6	5
서구	모집단 크기	42	1667	625	268
	표본크기	4	12	6	5
강화군	모집단 크기	119	714	137	101
	표본크기	10	4	0	3
옹진군	모집단 크기	189	106	25	19
	표본크기	14	2	0	1

3.2 추정량

본 사례연구에서는 숙박업과 음식점업 각각에 대해 구별 연간 총매출액의 단순 직접추정값과 앞 절에서 소개한 비합성(ratio synthetic)추정값, 표본크기의존추정값, EBLUP 추정값을 구하고자 한다.

본 자료에서 종사자수와 연간 총매출액의 상관계수는 숙박업의 경우 약 0.94이고, 음식점업의 경우 약 0.67로 나타났다. 따라서 종사자수와 연간 총매출액은 상당히 높은 상관관계를 갖는다고 할 수 있다. 그리고 앞에서 언급한 바와 같이 매년 실시되는 사업체 기초통계조사를 통해 모든 사업체에 대한 종사자수를 얻을 수 있다. 따라서 본 사례연구에서는 연간 총매출액의 추정을 위해 보조변수로 종사자수를 사용하고자 한다.

앞에서 소개된 추정기법을 기초로 본 사례연구를 위해 사용된 추정량을 정리하면 다음과 같다. 여기서 X는 종사자수를, Y는 연간 총매출액을 나타낸다.

3.2.1 단순 직접추정량

층의 개수가 H인 층화추출에서 일반적으로 사용되는 i지역 총계에 대한 단순 직접추정량은 다음과 같다.

$$\hat{Y}_{(DR),i} = \sum_{h=1}^{H} \bar{y}_{ih} N_{ih}$$

여기서 $\bar{y}_{ih}$는 i지역, h층의 단순평균이고, N_{ih}는 i지역, h층의 모집단 크기이다. 본 연구에서 소지역을 나타내는 i는 구를 의미하며 h는 산업세분류 업종을 의미한다.

3.2.2 비합성추정량

비합성추정량은 합성추정량의 한 형태로, 관심지역(i구)을 포함하는 대지역(인천광역시)에서의 관심변수와 보조변수의 비, Y/X와 관심지역에서의 비, Y_i/X_i가 같다는 가정아래 관심변수를 추정하는 것이다. 추정량은 아래와 같다.

$$\hat{Y}_{(RS),i} = \frac{y}{x} X_i$$

여기서 y와 x는 각각 표본 전체 사업체에서의 매출액 및 종사자수 총계를 나타내고, X_i는 i구의 종사자수 총계를 의미하며, 이는 매년 통계청에서 실시되는 사업체 기초통계조사를 통해 얻을 수 있다.

3.2.3 표본크기의존추정량

표본크기의존추정량은 두 가지 추정량을 적당한 가중값을 주어 결합하는 복합추정량의 한 가지 형태로써, 다음과 같이 표본크기에 따라 가중값을 결정하는 추정량이다.

$$Y_{(SD),i} = \begin{cases} \hat{Y}_{(REG),i} = \hat{Y}_{(DR),i} + (y/x)(X_i - \hat{X}_{(DR),i}), & w_i \geq W_i \text{ 일 때} \\ \frac{w_i}{W_i}\hat{Y}_{(REG),i} + (1 - \frac{w_i}{W_i})\hat{Y}_{(RS),i}, & w_i < W_i \text{ 일 때} \end{cases}$$

여기서 $\hat{Y}_{(REG),i}$는 종사자수 X를 보조변수로 사용하여 i지역 표본조사자료만을 이용하여 구한 회귀추정량(regression estimator)으로 또다른 형태의 직접추정량에 해당하고, $\hat{Y}_{(DR),i}$과 $\hat{X}_{(DR),i}$는 표본에서 구한 각각의 단순 직접추정량이다. 그리고 w_i는 n_i/n이며 W_i는 N_i/N이다. 이 형태는 식(2.2)에서 $\delta = 1$인 경우에 해당된다.

3.2.4 EBLUP 추정량

종사자수가 많아질수록 총매출액의 분산은 커질 것이다. 이러한 점을 고려하여 모형(2.3)에서 $k_{ij} = x_{ij}^{1/2}$으로 놓고, $w_{ij} = k_{ij}^{-2} = 1/x_{ij}$으로 가정하면, BLUP 추정량은 다음과 같다.

$$\hat{Y}_{(BLUP),i} = y_i + (N_i - n_i)\{\bar{x}_i^{*T}\tilde{\beta} + \gamma_i(\bar{y}_{(w)i} - \bar{x}_{(w)i}^T\tilde{\beta})\}$$

여기서 y_i는 표본에서 얻어지는 i지역의 해당업종의 연간 총매출액(표본총계)이고, $\tilde{\beta}$는 β의 BLUE이며, $\gamma_i = \sigma_v^2/(\sigma_v^2 + \frac{\sigma^2}{w_i.})$이다. 위의 식에 2절에 소개된 Fuller와 Battese(1973)의 'method of fitting constants'를 이용해서 구한 $(\sigma_v^2,\ \sigma^2)$의 추정량, $(\hat{\sigma_v}^2,\ \hat{\sigma}^2)$을 대입하면 EBLUP 추정량이 된다.

3.3 추정결과

숙박업의 경우 세분류업종이 하나로 분류되므로 관계없으나 음식점업의 경우 세분류업종이 식당업, 주점업, 다과점업으로 분류되고, 이 세분류업종을 부차모집단으로 하여 세분류업종별로 표본크기를 결정하였다. 간접추정량은 본래 표본설계에 의존하지 않고 사용될 수 있지만 본 사례연구에서는 음식점업의 경우 세분류에 따른 층화개념을 반영하여 세분류업종별 추정값을 각각 구한 후 이를 모두 합하여 음식점업의 추정값을 산출하는 방법도 가능하다. 따라서 세분류에 따른 층화를 반영하지 않고 추정값을 구해보고, 또한 층화를 반영하는 추정값도 함께 구했다.

먼저 숙박업의 구별 연간 총매출액에 대한 추정 결과는 〈표 3〉과 같다. 표에서 괄호안의 값은 참값과의 오차(추정값-참값)를 나타내고 있다.

〈표 3〉 숙박업 연간 총매출액 추정값 및 오차 (단위: 만원)

구	참값	단순 직접 추정값	비합성추정값	표본크기의존 추정값	EBLUP 추정값
중구	651776	593299 (-58477)	965407 (313632)	646096 (-5680)	666169 (14393)
동구	252554	169583 (-82971)	318170 (656162)	196778 (-55776)	196149 (-56405)
남구	2897933	2864974 (-32959)	2499600 (-398333)	2974385 (76452)	2866792 (-31141)
연수구	346820	259896 (-86924)	400982 (54162)	383241 (36421)	366708 (19888)
남동구	624460	1080975 (456515)	531737 (-92723)	1042118 (417658)	838310 (213850)
부평구	1511758	1362000 (-149758)	1072191 (-439567)	1343112 (-168646)	1403250 (-108508)
계양구	313570	182208 (-131362)	326887 (13317)	263170 (-50400)	296123 (-17447)
서구	253910	239400 ((-14510)	200491 (-53419)	211070 (-42840)	226257 (-27653)
강화군	331328	341982 (10654)	540454 (209126)	337841 (6513)	382954 (51626)
옹진군	106227	93487 (-12739)	706077 (599850)	206534 (100307)	260231 (154004)

음식점업의 구별 연간 총매출액에 대한 추정 결과는 〈표 4〉와 〈표 5〉에 나타나는데, 〈표 4〉는 표본설계의 층화개념을 반영하지 않고 구한 추정결과이고, 〈표 5〉는 층화개념을 반영하여 구한 추정결과이다.

〈표 4〉와 〈표 5〉에서 강화군과 옹진군에 대해 단순 직접추정값이 나타나지 않은 것은 산업세분류로 나눈 부차모집단 중 주점업에 대해 강화군과 옹진군에서 표본이 전혀 뽑히지 않았기 때문이다. 표본크기의존추정값도 직접추정값을 일부 사용하기 때문에 역시 구할 수 없다. 반면 표본이 전혀 없는 소지역에서도 비합성추정값과 EBLUP 추정값을 얻을 수 있음을 볼 수 있다.

〈표 4〉 음식점업 연간 총매출액 추정값(단위: 만원)-층화 반영않음

구	참값	단순 직접 추정값	비합성추정값	표본크기의존 추정값	EBLUP 추정값
중구	11429778	14233955 (2804177)	11978908 (549130)	11322873 (-106905)	10204032 (-1225746)
동구	3636442	2645297 (-991145)	4195280 (558838)	2196664 (-1439778)	3686995 (-50553)
남구	32317017	28320879 (-3996138)	28248927 (-4068090)	30379451 (-1937566)	32122706 (-194311)
연수구	9527831	13218273 (3690442)	10153855 (626024)	13396928 (3869097)	11042136 (1514305)
남동구	17918531	15741559 (-2176972)	17845911 (-72620)	16326485 (-1592046)	18011259 (92728)
부평구	27353132	28539979 (1186847)	23923742 (-3429390)	25550313 (-1802819)	25036989 (-2316143)
계양구	8259985	5891445 (-2368540)	9287114 (1027129)	6571421 (-1688564)	8092850 (-167135)
서구	10542330	12364918 (1822588)	11512529 (970199)	11919712 (1377382)	11181525 (639195)
강화군	3726971	-	4834155 (1107184)	-	4422835 (695864)
옹진군	540954	-	696374 (155420)	-	680697 (139743)

3.4 효율비교

본 사례연구에서는 총조사 자료를 이용함으로써 각 소지역별 총매출액에 대한 참값을 알 수 있다. 이 참값을 이용하여 각 추정방법의 오차를 측정할 수 있게 된다. 오차에 대한 측정은 각 소지역에서 발생하는 오차를 종합적으로 반영한 다음과 같은 오차제곱평균(average squared error)을 이용하였다.

$$ASE = \frac{1}{m}\sum_{i=1}^{m}(est. - Y_i)^2$$

여기서 Y_i는 i지역의 해당 산업 연간 총매출액의 참값, $est.$는 그에 대한 추정값을 의미하며 m은 추정대상 소지역 수를 나타낸다. 오차제곱평균을 구한 결과는 〈표 6〉과 같다.

〈표 5〉 음식점업 연간 총매출액 추정값(단위: 만원)-층화 반영

구	참값	단순 직접 추정값	비합성추정값	표본크기의존 추정값	EBLUP 추정값
중구	11429778	14233955 (2804177)	11878890 (449112)	11056857 (-372921)	10626431 (-803347)
동구	3636442	2645297 (-991145)	4154400 (517958)	2765504 (-870938)	3746753 (110311)
남구	32317017	28320879 (-3996138)	28432655 (-3884362)	29885042 (-2431975)	32413827 (96810)
연수구	9527813	13218273 (3690442)	10025937 (498106)	13822454 (4294623)	11149887 (1622056)
남동구	17918531	15741559 (-2176972)	17732772 (-185759)	16127872 (-1790659)	17617985 (-300546)
부평구	27353132	28539979 (1186847)	23771317 (-3581815)	25641699 (-1711433)	24934798 (-2418334)
계양구	8259985	5891445 (-2368540)	9193403 (933418)	6512314 (-1747671)	8501925 (241940)
서구	10542330	12364918 (1822588)	11409684 (867354)	12073280 (1530950)	11459928 (917598)
강화군	3726971	-	4731976 (1005005)	-	4609111 (882140)
옹진군	540954	-	687850 (146896)	-	694057 (153103)

〈표 6〉 오차제곱평균

	단순 직접추정값	비합성추정값	표본크기의존 추정값	EBLUP 추정값
숙박업	2.69	8.73	2.28	0.90
음식점업(1)	6.69	3.26	3.92	1.01
음식점업(2)	6.69	3.13	4.60	1.09

여기서 음식점업(1)은 층화개념을 반영하지 않은 경우이고, 음식점업(2)는 층화개념을 반영한 경우를 나타낸다.

오차제곱평균을 통해 EBLUP 추정량이 가장 효율적임을 알 수 있다. 음식점업의 경우 간접 추정방법을 사용할 경우 오차제곱평균이 작아진다. 그러나 숙박업의 경우 비합성추정값의 오차제곱평균이 가장 크다. 모형가정이 잘 맞지 않을 경우 합성추정량은 편의가 발생한다. 본 연구에서 합성추정량을 구하기 위해 사용한 모형가정은 소지역인 구에서의 총매출액과 종사자수의 비가 대지역인 인천광역시 전체에서의 비와 같다는 것인데, 숙박업의 경우 일부 구에서는 그러한 모형가정이 잘 맞지 않아 합성추정량인 비합성추정량을 사용한 추정에서 편의가 크게 발생한 것으로 생각된다.

음식점업에서 층화개념을 반영한 경우와 반영하지 않은 경우에 있어 오차제곱 평균에 약간의 차이가 나타나긴 하나 어느 쪽이 더 우수하다고 단정할 만큼 뚜렷한 차이는 나타나지 않는다.

4. 결 론

현재 우리나라 통계청에서 시행되고 있는 대부분의 표본조사는 16개 시·도지역을 기준으로 표본설계되므로 소지역에 대한 추정을 하기에 표본수가 극히 적다. 따라서 소지역 추정을 할 경우 안정적인 추정값을 제시하지 못한다. 표본수가 적은 소지역에 대해 안정적인 추정값을 제시할 수 있는 방법이 소지역 통계분석기법이다.

최근에는 다양한 이유로 소지역 통계에 대한 수요가 증가하고 있다. 그러나 우리나라의 경우 소지역 통계분석에 관한 연구가 미진하고, 이용사례도 찾아볼 수 없다.

본 논문에서는 소지역 통계분석에 사용되는 간접추정량들을 유형별로 소개하고, 사례분석을 위해 1996년 인천광역시의 숙박업 및 음식점업의 자료를 이용하여 구별 연간 총매출액을 추정하고 효율을 비교해 보았다.

인천광역시의 숙박 및 음식점업의 자료를 이용한 사례연구 결과 간접추정량이 직접추정량보다 효율적인 것으로 나타났고, 그 중에서 EBLUP 추정량의 효율이 가장 높게 나타났다.

본 논문에서는 소지역 추정기법의 도입 가능성을 검토해 보기 위해 극히 제한된 사례를 대상으로 간접추정량의 효율성을 검증해 보았다. 따라서 정부 공식 통계 생산을 위한 소지역 추정기법의 활성화를 위해서는 통계조사별로 전국 자료를 대상으로한 좀더 심층적인 검토가 필요할 것이다. 아울러 본 연구에서 고려하지 않았지만 최근 소지역 추정기법으로 상당한 연구가 진척되고 있는 모형을 기반으로 한 계층적 베이즈(hierachical Bayes)방법(Ghosh 등, 1999, 1996, 1994; Datta 등, 1999; Rao, 1999)의 활용방안에 대한 검토가 요망된다.

참 고 문 헌

1. 통계청 (1997a). *1995*년 기준 도소매업 및 서비스업 통계조사보고서, 통계청.

2. 통계청 (1997b). *1996*년 기준 사업체기초통계조사보고서, 통계청.

3. 통계청 (1998). *1996* 도 · 소매업총조사보고서, 통계청.

4. Brackstone, G. J. (1987). Small Area Data : Policy Issues and Technical Challenges. In *Small Area Statistics* (Platek, R., Rao, J. N. K., Sarndal, C. E. and singh, M. P.) 3-20. Wiley, New York.

5. Cassel, C. M. Kristiansson, G., Raback, G. and Wahlstrom, S. (1987). Using Model-Based Estimation to Improve the Estimate of Unemployment on a Regional Level in the Swedish Labor Force Survey. In *Small Area Statistics* (Platek, R., Rao, J. N. K., Sarndal, C. E. and Singh, M. P.), 141-159. Wiley, New York.

6. Datta, G. S., Lahiri, P., Maiti, T. and Lu, K. L. (1999). Hierachical Bayes Estimation of Unemployment rates for the U. S. states. *Journal of the American Statistical Association*, 94, 1074-1082.

7. Derw, D., Singh, M. P. and Choudhry, G. H. (1982). Evaluation of Small Area Estimation Techniques for the Canadian Labour Force Survey. *Survey Methodology*, 8, 17-47.

8. Fuller, W. A. and Battese, G. E. (1973). Transformations for Estimation of Linear Models with Nested-Error Structure. *Journal of the American Statistical Association*, 68, 626-632.

9. Ghosh, M., Mangia, N. and Kim, D. H. (1996). Estimation of Median Income of Four-person Families : A Bayesian Approach. *Journal of the American Statistical Association*, 91, 1423-1431.

10. Ghosh, M., Natarajan, K., Waller L. A. and Kim, D. (1999). Hierarchical Bayes GLMs for the Analysis of Spatial Data : An Application to disease mapping. *Journal of Statistical Planning and Inference*, 75, 305-318.

11. Ghosh, M. and Rao, J. N. K. (1994). Small Area Estimation : An Appraisal. *Statistical Science*, 9, 55-93.

12. Gonzalez, M. E. (1973). Use and Evaluation of Synthetic Estimators. In *Proceedings of the Social Statistics Section*, 33-36. Amer. Statist. Assoc., Washington, DC.

13. Herderson, C. R. (1950). Estimation of Genetic Parameters. *The Annals of Mathematical Statistics*, 21, 309-310.

14. Prasad, N. G. N. and Rao, J. N. K. (1990). The Estimation of the Mean Squared Error of Small-Area Estimators. *Journal of the American Statistical Association*, 85, 163-171.

15. Purcell, N. J. and Kish, L. (1979). Estimation for Small Domain. *Biometrics*, 35, 365-384.

16. Rao, J. N. K. (1999). Some Recent Advances in Model-Based Small Area Estimation. *Survey Methodology*, 25, 175-186

17. Sarndal, C. E. and Hidiroglou, M. A. (1989). Small Domain Estimation: A Conditional Analysis. *Journal of the Americal Statistical Association*, 84, 266-275.

Application of In-direct Estimation for Small Area Statistics

Young-Won Kim [3] · **Na-Young Sung** [4]

Abstract

Small area estimation is becoming important in survey sampling due to a growing demand for reliable small area statistics.

In estimating means, totals, and other parameters for small areas of a finite population, samplie sizes for small areas are typically small because the overall sample size is usually determined to provide specific accuracy at a much higher level of aggregation than that of small area. The usual direct estimators that use the only information which is gotten from the sample in a given small area provide unreliable estimates. However, indirect estimators utilize the information from the areas related with a given small area, that is, borrow strength from other related areas, and so give more accurate estimates than direct estimators.

In this paper we investigate small area estimation methods such as synthetic, composite and empirical best linear unbiased prediction estimator, and apply them to real domestic data which is from the Survey of Hotels and Restaurants in In-Chon as of 1996 and then evaluate the performance of these methods by measuring average squared errors. This evaluation shows that indirect estimators, which are small area estimation methods, are more efficient than direct estimator.

Key Words and Phrases: Small Area Estimation, Synthetic Estimator, Composite Estimator, EBLUP Estimator, Survey of Wholesale & Retail Trade and Service Industry

[3]Professor, Department of Statistics, Sookmyung Women's University, Seoul 140-742
[4]Graduate, Department of Statistics, Sookmyung Women's University, Seoul 140-742