한국통계학회논문집
2010, 17권, 1호, 9-16

공간층화표본설계에 대한 보정

변종석[a], 손창균[1,b], 김종민[c]

[a]한신대학교 정보통계학과, [b]한국보건사회연구원, [c]미네소타대학교 통계학과

요 약

일반적으로 공간모집단에서의 표본설계에 대한 연구는 가정된 종속관계에 대해 설정된 모형 하에서 이루어지며, 이때 추정하고자 하는 모수들은 평균, 비율 그리고 면적 등이 될 수 있다. 본 연구에서는 연구대상이 지리적 조건이나, 모양에 의해 층화된 모집단에 대해 영역을 추정하고자 할 때, 공간적으로 관련이 있는 보조변수를 활용하여 가중치 조정방법을 제시하고, 이에 대한 효율성을 검증하고자 한다. 즉, 공간 추정량에 대한 보정추정과정을 적용하여 가중치 조정을 통한 추정량을 개선하고, 수치적 예제를 통해 제안된 추정량이 효율적임을 제시하였다.

주요용어: 공간표본추출, 보정, 보조정보, 가중치 조정.

1. 서론

일반적으로 공간모집단에서의 표본설계에 대한 연구는 가정된 종속관계에 대해 설정된 모형 하에서 이루어지며, 이때 추정하고자 하는 모수들은 평균, 비율 그리고 면적 등이 될 수 있다. 만일 어떤 지역에서 특정한 물질에 의해 오염된 면적이나 동식물의 분포정도 등과 같은 2차원 공간상에서 정의된 어떤 특성이 하나의 불규칙하고 큰 폐곡선의 형태로 존재하는 공간 모집단을 가정하고, 관찰점 또는 표본점들의 관찰 위치들 사이에 공간적인 종속관계가 존재하며, 집락의 형태를 띠며 분포한다고 하자. 그러면 관심대상이 되는 공간모집단이 조사지역의 지리적 여건이나 형태, 추정영역의 굴곡정도 등과 같은 기준에 의해 층화된 경우 전통적인 표본설계 방법보다는 공간적인 특성을 고려한 표본설계를 하는 것이 보다 바람직할 것이다. 관련된 연구로는 Quenouille (1949)가 처음으로 추출방법을 제안하였고, Koop (1990)은 Quenouille (1949)의 연구를 계통추출법을 적용하여 2차원공간상의 표본추출법을 제안하였다. 관심영역의 면적과 관련된 연구로는 Bellhouse (1981)과 Koop (1990)이 있다.

Cressie (1993)은 표본으로 관찰되지 않는 새로운 지역의 예측에 관심을 두고 있으며, 예측방법으로는 크리깅(kriging) 방법을 사용하였다. 공간 표본추출방법과 관련하여 전통적인 표본추출기법을 적용한 경우로는 Thompson (1990,1991)과 Thompson과 Seber (1996)의 적응집락 추출방법이 대표적인 예로 들 수 있을 것이다. 이와 같은 층화공간표본설계와 더불어 만일 조사지역에서 추정하고자 하는 특성과 관련된 보조정보를 이용할 수 있다면 보다 효율적으로 추정이 가능할 것으로 사료된다.

본 논문에서는 이러한 관점에서 공간층화추정량에 대해 살펴보고, 다음으로 조사지역의 보조정보를 이용한 보정된 공간층화추정량을 제안하고자 한다.

논문의 구성은 2절에서 전통적인 보정 추정과정에 대해 간단히 살펴보고, 3절에서는 공간층화추출설계에 대해 살펴보고자 한다. 4절에서는 공간층화추정량의 보정 추정과정을 제안하고, 5절에서는 수치적 예제로부터 제안된 추정량의 효율성을 살펴보고, 끝으로 6절에서는 결론을 다루고자 한다.

본 논문은 한신대학교 학술연구비 지원에 의하여 연구되었음.
[1] 교신저자: (122-705) 서울시 은평구 불광동 진흥로 268번지, 한국보건사회연구원 부연구위원.
E-mail: chkson@kihasa.re.kr

2. 보정추정량

유한 모집단인 $U = \{1, 2, \ldots, N\}$을 고려하고, y_k를 모집단의 k단위에 대해 y의 값이라 하자. 그러면, 모평균은 $\mu_y = \Sigma_{k \in U} y_k / N$이 된다. 이러한 모집단에서 크기 n인 확률 표본 s를 추출한다고 하자. 그러면 모집단의 각 단위들이 표본에 포함될 확률은 각각 $v_k = P(k \in s)$, $v_{kl} = P(k \& l \in s)$이며, 이들은 모두 양의 값을 가진다. $k \in s$인 단위에 대해, 데이터로서 (x_k, y_k)를 관찰 가능하고, 이때, 보조변수벡터 $x_k = (x_{k1}, x_{k2}, \ldots x_{kJ})'$는 y_k와 연관되어있다고 가정한다. 보조변수 x_k의 모평균을 $\mu_x = \Sigma_{k \in U} x_k / N$이라 하고, 이 값은 기지라고 가정한다. 이러한 정보는 센서스나 행정 자료로부터 얻을 수 있다. 모평균에 대한 Horvitz-Thompson(HT) 추정량은 $\hat{\mu}_{yHT} = \Sigma_{k \in s} d_k y_k / N$이며, 여기서 $d_k = 1/v_k$이다.

보정의 목적은 기지의 보조정보를 이용하여 기존의 추출가중치(sampling weight) d_k를 조정하는 것으로서 가능하면 새로운 가중치를 w_k라 할 때 새로운 가중치가 기존의 추출가중치 d_k에 근접하도록 조정하는 것이다. 즉 기지의 보조변수 x_k에 대해 제한조건 $\Sigma_{k \in s} w_k x_k / N = \mu_x$하에서 거리함수 $\Sigma_s G(w_k, d_k) = \Sigma_s (w_k - d_k)^2 / d_k$를 최소로 하는 새로운 가중치 w_k를 구하는 것이다.

Deville과 Sarndal (1992)은 다음과 같이 정의되는 라그란쥐 함수를 최적화하는 방법으로 보정추정 과정을 제안하였다.

$$\text{Min} \sum_s d_k G\left(\frac{w_k}{d_k}\right) - \lambda\left(N^{-1} \sum_s w_k x_k - \mu_x\right), \tag{2.1}$$

여기서 라그란쥐 승수 벡터로서 $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_J)'$이다.

식 (2.1)에 정의된 라그란쥐 함수를 w_k에 대해 편미분하여 0으로 놓고 정리하면, 새로운 가중치는 다음과 같이 구해진다.

$$w_k = d_k(1 + x_k'\lambda) = d_k F(x_k'\lambda), \tag{2.2}$$

그러면 다음과 같이 정의되는 보정방정식으로부터 라그란쥐 승수벡터 λ를 구할 수 있다.

$$\frac{1}{N} \sum_s d_k (1 + x_k'\lambda) x_k = \mu_x, \tag{2.3}$$

보정방정식 (2.3)을 λ에 대해 정리하면 다음과 같다.

$$\lambda = (\mu_x - \hat{\mu}_{xHT}) \left(N^{-1} \sum_s d_k x_k x_k'\right)^{-1}. \tag{2.4}$$

결과적으로 관심변수에 대한 모평균 μ_y은

$$\hat{\mu}_{yGREG} = \frac{1}{N} \sum_s w_k y_k = \hat{\mu}_{yHT} + (\mu_x - \hat{\mu}_{xHT}) \hat{B}_s, \tag{2.5}$$

이때, $\hat{B}_s = (\Sigma_s d_k x_k' y_k)(\Sigma_s d_k x_k x_k')^{-1}$이고, $\hat{\mu}_{xHT} = \Sigma_{k \in s} d_k x_k / N$이다.

Deville과 Sarndal (1992)에 의해 제안된 일반화회귀 추정량(GREG)은 다음과 같은 점근분산을 가진다.

$$\text{AV}(\hat{\mu}_{yGREG}) = \frac{1}{N^2} \sum_{k \in U} \sum_{l \in U} (v_{kl} - v_k v_l)(d_k E_k)(d_l E_l), \tag{2.6}$$

여기서 $E_k = y_k - x_k'B$이고, 이때 B는 이론적인 "모집단 적합(census fit)"를 나타내는 정규방정식으로 표현되는 다음 식을 만족한다.

$$\left(\sum_U x_k x_k'\right) B = \sum_U x_k y_k. \tag{2.7}$$

Deville과 Sarndal (1992)은 다음과 같은 분산추정치를 제안하였다.

$$\hat{V}(\hat{\mu}_{yGREG}) = \frac{1}{N^2} \sum_{k \in s} \sum_{l \in s} \frac{(v_{kl} - v_k v_l)}{v_{kl}} (w_k e_k)(w_l e_l), \tag{2.8}$$

여기서 $e_k = y_k - x_k'\hat{B}_s$는 표본으로부터 구한 잔차들이다.

3. 층화 공간추출

공간적인 특성을 갖는 2차원상의 조사지역에서 사전에 정의된 어떤 특정한 변수에 대한 관심영역의 분포 면적이나 비율을 추정하고자 한다. 동식물의 분포나 특정물질의 오염지역 등과 같은 변수들의 특징은 서로 군락(집락)을 형성하여 분포하는 특징을 지니기 때문에 일정한 범위 안에서 인접된 위치들 간의 관계를 가진다.

공간모집단에 대한 추론을 위해 기본적으로 필요한 가정은 첫째, 추론하고자 하는 관심지역이나, 영역의 형태가 하나의 불규칙한 폐곡선으로 표현되며, 둘째 공간모집단에 존재하는 여러 특성이나 추론하고자 하는 형태와 같은 적절한 층화기준에 의해 공간모집단을 층화한다. 셋째, 층화한 공간을 칸(cell)으로 구분할 때 사각형의 격자로 분할한다. 이때 분할된 사각형의 형태는 정사각형으로 가정한다. 이때 각 층간의 칸의 크기는 할당된 층의 표본크기에 의해 달라진다.

또한 공간모집단으로부터 표본을 추출하기 위해 필요한 기본가정으로는 층화공간표본추출설계를 이용하여 표본점을 추출하는 방법으로는 전통적인 표본추출방법을 고려한 복합표본추출설계는 고려하지 않으며, 단지 층화추출설계만을 고려한다. 또한 각 층내에서 간 칸들에 대한 표본점추출방법은 동일한 공간추출법을 적용하고, 분할된 칸에서 하나의 표본점을 추출한다. 층화표본추출을 위한 기호를 다음과 같이 정의하자.

P_h: h층의 모비율,

P: 관심영역에 대한 모비율,

A_{ah}: h층에서의 관심영역의 면적,

A_h: h층의 면적,

A: 모집단 면적,

n_h: h층에서의 표본점의 수.

이와 더불어 각 층에 대해 칸으로 분할된 경우에 P_{hij}를 h층의 (i, j)칸의 모비율, A_{ahij}를 h층의 (i, j)칸의 관심영역의 면적, A_{hij}를 h층의 (i, j)칸의 면적이라 하자. 또한 N_h는 h층의 모집단 칸 수이다. 각 h층에 대한 층별 가중값은 $W_h = A_h/A = A_h/\Sigma_h A_h$이며, h층에서의 관심영역의 비율은 $P_h = A_{ah}/A_h$이고, 이를 이용하여 모집단 비율은 다음과 같이 표현된다.

$$P = \frac{\sum_h A_h P_h}{\sum_h A_h} = \sum_h W_h P_h. \tag{3.1}$$

이와 관련된 모집단 전체면적에 대한 모비율 P의 추정량은 다음과 같다.

$$\hat{p} = \sum_h \frac{A_h}{\sum_h A_h} p_h = \sum_h W_h p_h, \tag{3.2}$$

여기서 $p_h = n_h^{-1}\Sigma_i\Sigma_j z_{hij}$이며, 만일 (i, j)번째 사각형에서 표본점이 관심영역에 속하면 $z_{hij} = 1$, 그 외에는 0인 베르누이 확률변수이다.

층간 독립이라는 가정 하에서 추정량 $\hat{p}$의 분산은 다음과 같다.

$$\begin{aligned} V(\hat{p}) &= \sum_h W_h^2 V(p_h) + \sum_{h \neq h'} W_h W_{h'} \mathrm{Cov}(p_h, p_l) \\ &= \sum_h W_h^2 V(p_h). \end{aligned} \tag{3.3}$$

추정된 비율추정량을 이용하여 면적에 대한 추정량과 분산을 구해보면 h층에서의 면적을 추정한 후 모집단에 대한 전체면적을 추정하면 된다. h층에서 추정된 면적 추정량과 층의 가중치를 이용하여 모집단 수준의 면적 추정량과 분산의 계산이 가능하게 된다.

h층에서의 면적 추정량은 다음과 같이 추정된다.

$$\hat{A}_{ah} = A_h p_h. \tag{3.4}$$

따라서 전체면적 추정량은 다음과 같다.

$$\hat{A}_a = A\hat{p} = A\sum_h W_h p_h = \sum_h A_h p_h. \tag{3.5}$$

각 층마다 독립적으로 표본을 추출한다면 면적 추정량의 분산은 다음과 같다.

$$V\left(\hat{A}_a\right) = V\left(\sum_h A_h p_h\right) = \sum_h A_h^2 V(p_h). \tag{3.6}$$

4. 층화공간 추정량의 보정추정

층별 가중치 W_h를 보정하기 위해 보조변수를 x라 하고, 이는 층별 표본비율 p_h와 관련되어 있다고 하자. $\bar{x}_h$와 $\bar{X}_h$를 h층에 대해 표본과 모집단의 평균이라 하자. 또한 모집단 평균 $\bar{X} = \Sigma_h W_h \bar{X}_h$를 정확히 안다고 가정하자. 그리고 p_h과 P_h는 h층에 대해 표본과 모집단 비율이라 하자. 목적은 보조변수 x를 이용하여 $P = \Sigma_h W_h P_h$추정하고자 한다. 그러면 새로운 층별 가중치를 W_h^*이라 하고, 이는 다음의 (4.2) 보정 방정식의 조건하에서 식 (4.1)을 최소로 한다.

$$G(W_h^*, W_h) = \sum_h \frac{(W_h^* - W_h)^2}{q_h W_h}, \tag{4.1}$$

이에 대한 제한조건은 다음과 같다.

$$\bar{X} = \sum_h W_h^* \bar{x}_h. \tag{4.2}$$

라그란쥐 방법을 이용하여 새로운 보정가중치 W_h^*는 다음과 같다.

$$W_h^* = W_h + \frac{W_h q_h \bar{x}_h}{\sum_h W_h q_h \bar{x}_h^2}\left[\bar{X} - \sum_h W_h \bar{x}_h\right] = W_h\left(1 + \frac{q_h \bar{x}_h}{\sum_h W_h q_h \bar{x}_h^2}\left[\bar{X} - \sum_h W_h \bar{x}_h\right]\right) = W_h g_h, \tag{4.3}$$

여기서 $g_h = 1 + q_h\bar{x}_h(\Sigma_h W_h q_h \bar{x}_h)^{-1}[\bar{X} - \Sigma_h W_h \bar{x}_h]$으로서 g-가중치이며, q_h는 고정된 값이다.

따라서 P의 보정추정량 $\hat{p}_{cal}$은

$$\hat{p}_{cal} = \sum_h W_h g_h p_h \tag{4.4}$$

이며, $\hat{p}_{cal}$의 분산은 다음과 같다.

$$V(\hat{p}_{cal}) = V\left(\sum_h W_h g_h p_h\right) = \sum_h W_h^2 g_h^2 V(p_h). \tag{4.5}$$

이와 더불어 전체 면적에 대한 보정추정량은 비율 추정량을 직접 이용해 다음과 같이 나타낼 수 있다.

$$\hat{A}_{cal} = A\hat{p}_{cal} = A\sum_h W_h g_h p_h = \sum_h A_h^* p_h. \tag{4.6}$$

또한 보정추정량 $\hat{A}_{cal}$의 분산은 다음과 같이 구할 수 있다.

$$V\left(\hat{A}_{cal}\right) = V\left(\sum_h A_h^* p_h\right) = \sum_h A_h^{*2} V(p_h). \tag{4.7}$$

5. 수치적 예제

모의실험을 위해 $N = 10{,}000$개의 난수 y_i^*와 x_i^*를 $U(0,1)$에서 발생시키고, $S_y^2 = 50$, $S_x^2 = 50$으로 고정하여 다음과 같은 모형 하에서 난수를 생성하였다 (Bartley 등, 1983).

관심변수: $y_i = 3.0 + \sqrt{S_y^2 + (1-\rho^2)}y_i^* + \rho S_y x_i^*$

보조정보: $x_i = 4.0 + S_x x_i^*$

모의실험을 간단히 하기위해 y_i의 크기에 따라 모집단을 2개의 층으로 구분하였다. 각 y_i단위는 보조변수 x_i와 상관관계를 가진다. 따라서 각 층에 속한 단위의 비율을 추정하는 과정에 보조정보를 이용한 보정 추정과정을 적용함으로써 추정량의 효율성을 알아보고자 하였다.

효율성을 비교하기 위해 비율 추정량에 대한 분산과 제안된 면적 추정량의 분산의 상대적인 비로 표현되는 상대효율을 다음과 같이 정의하자.

$$\mathrm{RE}(\hat{p}_{cal}) = \frac{1}{T}\sum_{t=1}^{T}\mathrm{RE}_t(\hat{p}_{cal}), \tag{5.1}$$

여기서 $\mathrm{RE}_t(\hat{p}_{cal}) = \hat{V}_t(\hat{p})/\hat{V}_t(\hat{p}_{cal})$이다.

표 1: 모의실험 결과

상관계수	반복수(T)	추정값($\hat{p}_{cal}$)	표준편차$\left(\sqrt{\hat{V}(\hat{p}_{cal})}\right)$	편향($b(\hat{p}_{cal})$)	상대효율(RE)
	100	0.4000931	0.000787973	−0.000093096	0.7138377
$\rho = 0.3$	1,000	0.4000139	0.000805527	−0.000013864	0.7157397
	2,000	0.3999900	0.000795530	9.9896234E−6	0.7156654
	100	0.3998951	0.000872600	0.000104900	1.0054812
$\rho = 0.7$	1,000	0.4000027	0.000793100	−2.73E−06	1.0058572
	2,000	0.3999987	0.000825200	1.30E−06	1.0056613
	100	0.4000502	0.000877661	−0.000050180	1.5372587
$\rho = 0.9$	1,000	0.3999721	0.000826085	0.000027888	1.5250761
	2,000	0.3999818	0.000804389	0.000018235	1.5252971

이와 함께 편향은 다음과 같이 정의하였다.

$$b(\hat{p}_{cal}) = \frac{1}{T}\sum_{t}^{T} b_t(\hat{p}_{cal}), \tag{5.2}$$

여기서 $b_t(\hat{p}_{cal}) = E(\hat{p}_{cal}) - P$이다.

단, 모집단의 비율은 $P = 0.4$로 고정하고, 칸의 수는 실험의 편의상 4개로 한정하였다. 또한 층의 수는 매우 간단하게 2개의 층으로 구분하였다. 모의실험에서 반복수는 각각 $T = 100, 1{,}000, 2{,}000$번을 수행하였고, 그에 따른 추정값의 표준편차 및 편향을 각각 구하였다. 또한 제안된 방법의 효율성을 알아보기 위해 다음 표 1과 같이 상대효율을 계산하였다.

모의실험 결과 관심변수와 보조변수간의 상관계수의 변화에 따라 상대효율(RE)에 차이가 있으며, 전체적으로 모비율 0.4에 대한 편향이 거의 0에 가깝게 추정되었다. 결과적으로 공간표본추출에서 관심변수와 상관관계가 큰 보조정보를 이용함으로서 추정의 효율을 높일 수 있음을 알 수 있다.

6. 결론

본 논문에서는 조사개체가 어떤 지역에서 특정한 물질에 의해 오염된 면적이나 동식물의 분포정도 등과 같은 2차원 공간상에서 정의되는 경우에 전통적인 표본조사설계보다는 공간적인 특성을 고려한 표본설계가 바람직 할 것이다. 또한 해당지역의 특정 개체에 대한 보조정보를 이용할 수 있다면, 개체들에 대한 표본정보만을 이용하는 것보다 효율적일 것이다. 이러한 관점에서 본 논문은 전통적인 공간층화 표본 추출설계를 고려한 공간 추정량에 대해 기지의 보조정보를 이용한 보정을 실시하여 추정량의 효율성을 높이고자 하였다. 제한된 모의실험을 통해 제안된 추정량이 보다 효율적임을 알 수 있었다.

향후 공간 표본 추출설계에 대해 2차원 보조정보가 존재하는 경우로 확장하여 이용 가능한 보정 추정량을 제안하고자 한다.

참고 문헌

Bartley, P., Fox, B. L. and Schreage, L. E. (1983). *A Guide to Simulation*, Springer-Verlag, New York.

Bellhouse, D. R. (1981). Area estimation by point-counting techniques, *Biometrics*, **37**, 303–312.

Cressie, N. A. C. (1993). *Statistics for Spatial Data*, John Wiley & Sons, New York.

Deville, J. C. and Sarndal, C. E. (1992). Calibration estimators in survey sampling, *Journal of the American Statistics Association*, **87**, 376–382.

Koop, J. C. (1990). Systematic Sampling of two-dimensional surfaces and related problems, *Communication in Statistics-Theory and Methods*, **9**, 1701–1750.

Quenouille, M. N. (1949). Problems in plane sampling, *The Annals of Mathematical Statistics*, **20**, 355–375.

Thompson, S. K. (1990). Adaptive cluster sampling, *Journal of the American Statistics Association*, **85**, 1050–1059.

Thompson, S. K. (1991). Adaptive cluster sampling: Design with primary and secondary units, *Biometrics*, **47**, 1103–1115.

Thompson, S. K. and Seber, G. A. F. (1996). *Adaptive Sampling*, Wiley, New York.

2009년 10월 접수; 2009년 12월 채택

Calibration for Spatial Stratified Sampling Design

Jong-Seok Byun[a], Chang-Kyoon Son[1,b], Jong-Min Kim [c]

[a]Department of Statistics and Information, Hanshin University
[b]Korea Institute for Health and Social Affairs
[c]Division of Science and Mathematics, University of Minnesota, Morris

Abstract

The sampling design for the spatial population studies needs a model assumption of a dependent relationship, where the interesting parameters can be the population mean, proportion and area. We know that the study of an interested spatial population, which is stratified by a geographical condition or shape, and the degree of distort of an estimation area is much useful. In light of this, if auxiliary information of the target variable such as wasted area contaminated by some material and the degree of distribution of animal or plants is available, then the spatial estimator might be improved through the calibration procedure. In this research, we propose the calibration procedure for the spatial stratified sampling in which we consider the one and two-dimensional auxiliary information.

Keywords: Spatial sampling, calibration, auxiliary information, weighting adjustment.

This work was supported by Hanshin University Research Grant.

[1] Corresponding author: Institute for Health and Social Affairs, San 42-12, Bulgwang-dong, Eunpyeong-gu, Seoul 122-705, Korea. E-mail: chkson@kihasa.re.kr