The Korean Communications in Statistics **Vol. 14** No. 1, 2007, pp. 57–69

Modified Adaptive Cluster Sampling Designs*

Jeong-Soo Park;[1] Youn-Woo Kim[2] and Chang-Kyoon Son[3]

Abstract

Adaptive cluster sampling design is known as a sampling method for rare clustered population. Three modified adaptive cluster sampling designs are proposed. The adjusted Hansen-Hurwitz estimator and the Horvitz-Thompson estimator are considered. Efficiency issue of the proposed sampling designs is discussed in a Monte-Carlo simulation study.

Keywords: Hansen-Hurwitz estimator; Horvitz-Thompson estimator; interpolation; jumped adaptive cluster sampling design.

1. 서론

적응집락추출은 개체가 드물고 집락을 형성하고 있는 모집단에 대한 표본추출방법이다. 초기표본을 추출하고 관심변수의 값이 특정한 조건을 만족할 때 까지 이웃한 단위들을 조사하는 방법이다. 개체수가 적은 동식물 종의 조사 (생태학 분야) 나 오염도가 높은 지역의 조사 (환경과학 분야), 집중된 광물 또는 화석연료의 조사 (지구과학 분야) 및 희귀한 질병의 감염에 대한 조사 (역학 분야) 등 다양한 많은 분야에 응용될 수 있다 (이해용과 이필영, 2002; Thompson과 Seber, 1996; Thompson, 2004; Magnussen *et.al*, 2005; Noon *et.al*, 2006). 국내의 관련연구는 남궁평과 변종석 (2001), Lee (1998) 등이 있다. 최근의 국제적 연구로는 Dryver와 Thompson (2005), Felix-Medina와 Thompson (2004), Magnussen *et al.* (2005), Noon *et al.* (2006), Turk와 Borkowski (2005) 등을 들 수 있다.

기존의 적응집락설계는 초기 표본을 기반으로 한 단위씩 이동하면서 조사를 수행하기 때문에 개체들이 넓게 퍼져있는 경우에는 조사수가 급격히 증가함으로 인하여 시간과 비용이 많이 소요되는 문제점이 있다. 따라서 본 논문에서는 조사방법에 있어서 관심변수의 조건에 따라 이웃으로 한 단위 또는 두 단위씩 이동하는, 변경된 적응집락추출설계들 세가지를 제안하였다. 그리고 이들의 효율성과 모수추정에 대해 논의하였다.

* This research was supported by KOSEF (F01-2004-000-10351-0).

1) Professor, Department of Statistics, Chonnam National University, Gwangju 500-757, Korea. Correspondence : jspark@chonnam.ac.kr

2) Risk Analyst, Risk Management Department, Gwangju Bank, Gwangju 501-730, Korea.

3) Research Fellow, Korea Institute for Health and Social Affairs, Seoul 122-705, Korea.

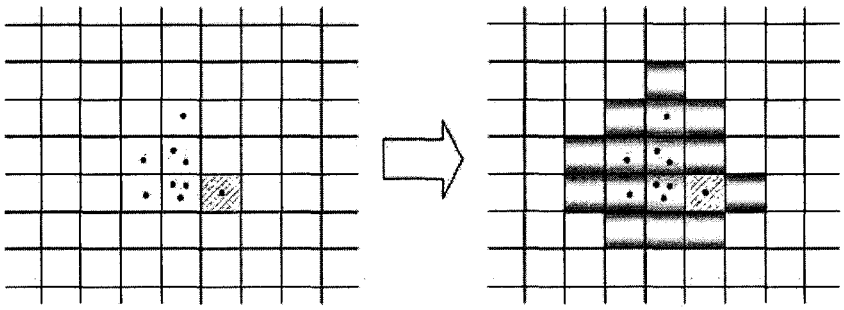
2절에서 기존의 적응집락 추출설계와 논문에서 제안한 추출설계들을 설명하였다. 3절에서는 모평균과 분산의 추정상의 중요사항을 기술하였고, 이를 이용하여 4절에서는 본 논문에서 제안한 표본추출방법들의 모수 추정과 MSE를 다루었다. 5절에서는 모의실험 결과와 추출설계들의 효율성에 대해 논의하였고 6절에서는 요약과 결론이 언급되었다.

2. 변경된 적응집락추출설계

우선 2차원 평면에서의 관심영역을 큰 사각형 모집단으로 설정하고, 이를 N개의 격자 또는 셀로 나눈 뒤, r번째 행과 c번째 열에 해당하는 셀의 관측치를 $y(r,c)$ 라 하자.

2.1. 단순 적응집락추출설계

기존의 단순 적응집락 추출설계 (Simple adaptive cluster design; SAD) 은 그림 2.1과 같이 초기표본으로 선택된 음영된 단위 (k,j)에서 특정한 값이 관측되었을 때 인접한 상하좌우의 단위인 $(k-1,j),(k+1,j),(k,j-1),(k,j+1)$ 번째 단위로 이동하여 값을 조사한다. 만일 이동한 단위에서 더 이상 관측 값이 없을 때는 이동을 멈춘다.



(a)초기의 표본선택 (b)단순 적응집락 추출

그림 2.1: 단순 적응집락추출설계

이러한 과정을 통해 조사된 단위들은 그림 2.1(b) 와 같이 관측값이 존재하는 단위들(m_i) 과 그들을 둘러싸고 있는 가장자리 단위들(a_i) 로 이루어지며, 이 단위들로 이루어진 영역을 네트워크라고 하고 A_i로 표기하자. 그러므로 네트워크 A_i내의 단위의 개수는 $n(A_i)=m_i+a_i$ 이다. 만약 어떤 초기표본에 관측값이 없으면 가장자리 단위 하나로 이루어진 네트워크가 형성된다. 또한 2개의 서로 다른 초기표본으로부터 적응추출과정을 진행하다가 서로 겹치게 되면 하나의 네트워크로 취급한다.

2.2. 건너뛴 적응집락추출설계

이 설계는 이웃을 한 단위씩 건너뛰며 조사하는 방법으로 초기 단위 (k,j) 로부터 $(k-2,j),(k+2,j),(k,j-2),(k,j+2)$ 번째 값을 순차적으로 조사한다. 그림 2.2는 본

논문에서 제안하는 설계 중 하나인 건너뛴 적응집락추출설계 (Jumped adaptive cluster design; JAD) 를 통해 표본을 추출해 나가는 과정을 보여주고 있다.

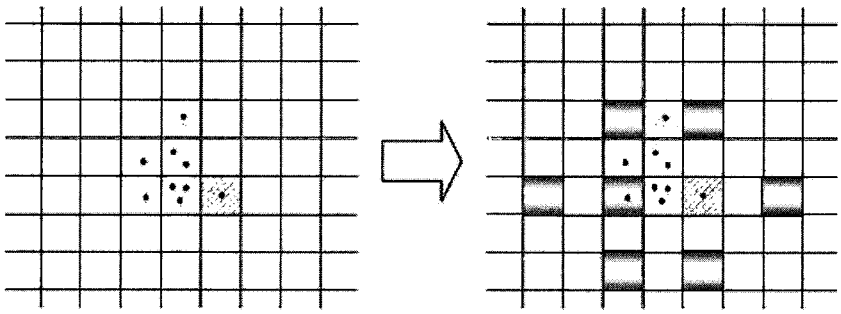


그림 2.2: 건너뛴 적응집락추출설계

네트워크 A_i나 m_i및 a_i의 결정은 SAD와 같은 방법을 따른다. 이 설계는 한 단위씩 건너뛰기 때문에 조사하는 총 네트워크의 수가 SAD에 비해 감소하는 특징이 있다. 그림 2.2에서 네트워크의 수는 SAD에 의해서는 $m+a=6+10=16$이고, JAD에 의해서는 $m+a=2+6=8$이다. 즉, 조사하는 표본단위의 수는 SAD에 비해 감소하지만, 건너뛴 사이 값을 조사하지 않으므로 정보를 손실할 수 있다.

이 설계와 관련하여 Hedayat과 Stufken (1998) 는 초기 표본단위의 근방에 있는 유사한 개체에 대한 포함확률을 조정하여, 초기표본과 유사한 개체가 선택되는 것을 통제함으로서 넓게 퍼져있는 개체들의 조사에 적용 가능한 표본추출설계를 제안하였다. JAD는 그들의 연구와 맥락을 같이하지만, 본 연구에서는 추출확률을 조정하는 것은 아니므로 그들의 연구와는 다른 접근이라 할 수 있다.

2.3. 내삽된 건너뛴 적응집락추출설계

JAD에서 건너뛴 사이 값의 정보를 손실할 수 있는 단점을 보완하기 위한 방법으로 건너뛴 사이 값에 대해 관측된 양쪽 단위의 평균값으로 대체해 주는 방법이다. 그림 2.3은 내삽된 건너뛴 적응집락추출설계 (JAD with Interpolation; $JADI$) 방법을 보여주고 있다. 그림 2.3에서 초기에 선택된 단위와 건너뛰어 관측한 단위 사이를 보정해주는 것을 보이고 있다. $n(A_i)=m_i+a_i+J_i$로서 m_i과 a_i의 결정은 JAD와 같고, J_i 는 내삽되는 네트워크의 개수를 나타낸다. 그림 2.3에서 $n(A_i)=3+7+2=12$이며, SAD 경우 $n(A_i)=7+11=18$이 된다.

2.4. 일반화 적응집락추출설계

SAD와 JAD를 혼합한, 일반화 적응집락추출설계 (Generalized adaptive cluster design; GAD) 방법으로서, 조사하는 단위의 특정한 조건 (관측값이 문턱치 c 보다 크냐 작냐) 에 따라 이웃을 건너뛰든지 또는 그냥 이웃을 조사하든지 하는 방법이다. 즉, 건너

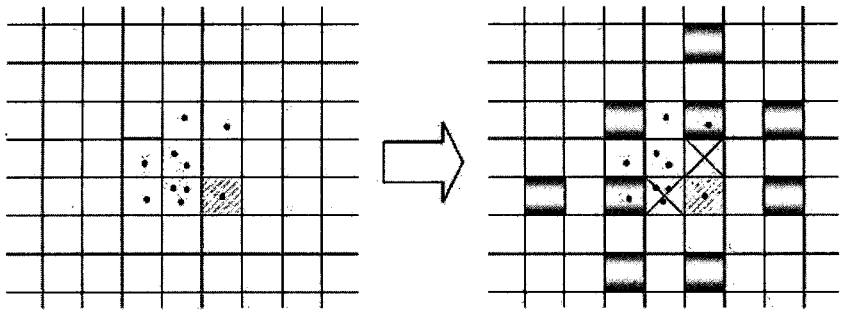


그림 2.3: 내삽된 건너뛴 적응집락추출설계

뛰으로 인한 정보손실을 막기 위해 조건을 만족하지 않는 단위에 대해서는 SAD로 조사해가는 방법이다.

그림 2.4는 관측값에 특정조건 $c = 2$를 적용하여 조사단위의 관측값이 c보다 크면 JAD를 적용하고, 그 이하이면 SAD를 적용하여 네트워크를 형성하는 과정을 보여주고 있다. 이 경우 $n(A_i) = 8 + 13 = 21$ 가 된다. 이 결과 네트워크의 수는 그림 2.4를 SAD방법으로 조사하는 결과와 동일하게 나타난다.

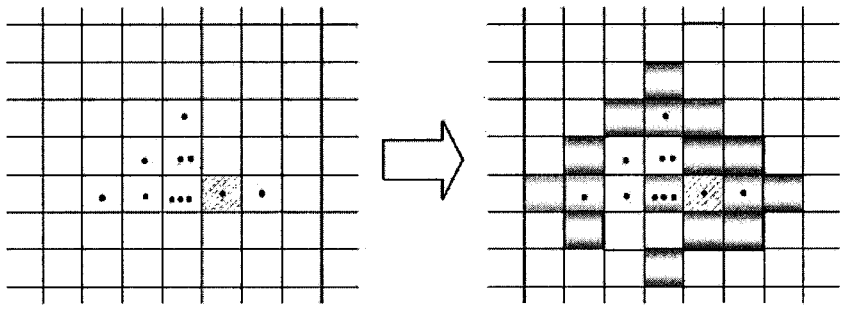


그림 2.4: 일반화 적응집락추출설계

3. 모평균과 분산의 추정 문제

모수추정에 이용 가능한 추정량으로는 Hansen-Hurwitz (HH) 추정량과 Horvitz-Thompson (HT) 추정량이 있다. HH추정량은 n개의 단위를 복원 추출할 경우 단위 i의 추출확률 p_i를 모든 단위에 대해 안다면 이용가능하다. 마찬가지로 HT추정량도 표본내의 단위 i에 대한 포함확률 (inclusion probability) π_i를 알 수 있다면 이용 가능하다. 그러나 이들 (추출확률과 포함확률) 을 모든 표본단위에 대해 알 수 없는 경우가 일반적이다. 따라서 적응집락추출에서는 대체적인 방법으로 HH추정량을 위해서는 p_i 대신에 한 단위의 네트워크가 초기표본에 교차되는 확률을 이용하는 수정된 방법을 사용하고,

HT추정량을 위해서는 π_i대신에 네트워크와 교차되는 초기표본의 확률을 이용한다.

본 절에서는 -특히 4절의 전개를 위해- 우선 Thompson과 Seber (1996) 에 의해 제안된 추정량과 분산 및 분산 추정량의 성질에 대해 살펴보고자 한다.

3.1. 초기 교차확률을 이용한 모평균 추정

단위 i가 최종 표본에 포함되기 위해서는 네트워크 A_i의 임의의 단위가 초기 표본의 일부로 선택 되거나 단위 i의 네트워크의 한 단위가 가장자리 단위로 선택될 때이다. m_i를 네트워크 A_i의 단위들의 수라 하고, a_i를 가장자리 단위의 총 수라 하자. 그러면 단위 i가 표본에 포함될 확률은 다음과 같이 정의 된다.

$$\pi_i = 1 - \left[\binom{N - m_i - a_i}{n_1} \Big/ \binom{N}{n_1} \right]. \tag{3.1}$$

모든 표본단위에 대해 이 확률을 알고 있다면, 다음과 같은 HT 추정량을 이용할 수 있다.

$$\hat{\mu}_{HT} = \frac{1}{N} \sum_{i=1}^{n} \frac{y_i}{\pi_i}. \tag{3.2}$$

만일 표본의 임의의 집락에서 단위 i가 가장자리 단위이면, 모든 집락이 표본에서 빠지게 됨으로 a_i를 알 수 없다. 이러한 문제를 해결하기 위해 π_i에서 a_i를 없애는 부분적인 포함확률 (partial inclusion probability) 을 다음과 같이 정의한다.

$$\pi_i' = 1 - \left[\binom{N - m_i}{n_1} \Big/ \binom{N}{n_1} \right]. \tag{3.3}$$

부분 포함확률 π_i'는 단위 i에 대한 네트워크 A_i에 초기표본이 교차한 확률로 해석된다. 따라서 초기교차확률에 근거한 μ의 비편향추정량은 다음과 같다.

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} \frac{y_i I_i'}{\pi_i'}. \tag{3.4}$$

여기서 I_i'은 초기표본이 네트워크 A_i와 교차하면 1, 그렇지 않으면 0을 갖는다.

식 (3.4) 로부터 교차확률 π_i'은 k번째 네트워크에서 단위 i에 대해 α_k로서 동일하다. 따라서 식 (3.4) 는 다음과 같다.

$$\hat{\mu} = \frac{1}{N} \sum_{i=k}^{K} \frac{{y_k}^* I_k}{\alpha_k} = \frac{1}{N} \sum_{i=k}^{\kappa} \frac{{y_k}^*}{\alpha_k}. \tag{3.5}$$

여기서 K는 모집단에서 서로 다른 총 네트워크의 수이며, κ는 표본에서 서로 다른 네트워크의 수이다. 또한 I_k는 초기표본이 k번째 네트워크 A_k와 교차하면 1, 그렇지 않으면 0을 갖는다.

만일 k번째 네트워크에 x_k가 존재한다면, 교차확률은 다음과 같다.

$$\alpha_k = 1 - \left[\binom{N - x_k}{n_1} / \binom{N}{n_1} \right]. \tag{3.6}$$

또한 j와 k번째 네트워크가 교차할 확률은 다음과 같다.

$$\alpha_{jk} = 1 - \left[\binom{N - x_j}{n_1} + \binom{N - x_k}{n_1} - \binom{N - x_j - x_k}{n_1} \right] / \binom{N}{n_1}. \tag{3.7}$$

따라서 분산과 분산 추정량은 각각 다음과 같다.

$$V(\hat{\mu}) = \frac{1}{N^2} \sum_{j=1}^{K} \sum_{i=1}^{K} \left(\frac{\alpha_{jk} - \alpha_j \alpha_k}{\alpha_j \alpha_k} \right) {y_j}^* {y_k}^*, \tag{3.8}$$

$$\hat{V}(\hat{\mu}) = \frac{1}{N^2} \sum_{j=1}^{\kappa} \sum_{i=1}^{\kappa} \frac{{y_j}^* {y_k}^*}{\alpha_{jk}} \left(\frac{\alpha_{jk}}{\alpha_j \alpha_k} - 1 \right). \tag{3.9}$$

3.2. 초기 교차수를 이용한 모평균 추정

일반적으로 n개의 단위를 복원 추출하고, 추출결과로 나온 단위 i의 확률 p_i가 모든 단위에 대해 알려져 있다면, 모평균 μ의 HH추정량은 다음과 같이 정의된다.

$$\hat{\mu}_{HH} = \frac{1}{nN} \sum_{i=1}^{n} \frac{y_i}{p_i}. \tag{3.10}$$

그러나 적응 추출설계에서 i단위의 확률 p_i를 모든 단위에 대해 알 수 없기 때문에 HH추정량을 수정하여 어떤 단위의 네트워크가 초기표본에 의해 교차되는 확률을 p_i대신 이용하게 된다 (Thompson과 Seber, 1999). 새로운 변수 w_i를 다음과 같이 정의하자.

$$w_i = \frac{1}{m_i} \sum_{j \in A_i} y_j. \tag{3.11}$$

그러면, 모평균 μ에 대해 수정된 추정량 ${\widetilde{\mu}_{HH}}^*$은 다음과 같이 정의된다.

$$\begin{aligned} {\widetilde{\mu}_{HH}}^* &= \frac{1}{n_1} \sum_{i=1}^{n_1} \frac{1}{m_i} \sum_{j \in A_i} y_j \\ &= \frac{1}{n_1} \sum_{i=1}^{n_1} w_i = \overline{w}. \end{aligned} \tag{3.12}$$

여기서 w_i는 네트워크 A_i안에 있는 m_i개의 관측치들의 평균이다.

수정된 추정량 $\widetilde{\mu}_{HH}{}^*$은 y_i값 보다는 w_i들의 모집단에서 n_1개의 단순임의 표본을 추출하여 구한 표본평균이고, 비편향 추정량이 된다. 이와 더불어 단순임의 추출설계로부터 $\widetilde{\mu}_{HH}{}^*$의 분산과 분산 추정량은 다음과 같다 (Thompson and Seber, 1996).

$$V(\widetilde{\mu}_{HH}{}^*) = \frac{N-n_1}{Nn_1(N-1)}\sum_{i=1}^{N}(w_i-\mu)^2, \tag{3.13}$$

$$\hat{V}(\widetilde{\mu}_{HH}{}^*) = \frac{N-n_1}{Nn_1(n_1-1)}\sum_{i=1}^{n_1}(w_i-\widetilde{\mu}_{HH}{}^*)^2. \tag{3.14}$$

4. 제안된 추출방법에 따른 모수추정

모평균 μ에 대한 추정량 식 (3.12) 와 그에 대한 분산 및 분산추정량 식 (3.13) 과 (3.14) 는 SAD를 적용한 경우에 초기 교차수를 이용한 추정량이다. 따라서 만일 본 논문에서 제안한 표본추출방법을 적용할 경우 각각의 방법에 따라 추정량의 성질이 변화할 것이다. SAD의 경우 식 (3.12) 의 모평균 μ에 대한 추정량 $\widetilde{\mu}_{HH}{}^*$은 비편향 추정량이며, 식 (3.14) 의 분산 추정량 역시 식 (3.13) 에 대한 비편향 추정량이다 (Thompson과 Seber, 1996). 따라서 본 절에서는 제안한 표본추출방법에 따른 추정량 및 분산과 분산 추정량의 성질을 살펴보고자 한다.

4.1. 포함확률

기본적으로 SAD로부터 단위 i의 추출확률과 단위 i와 교차하는 초기확률을 안다면, 제안된 방법에 따라 이 확률을 적용하여 그에 따른 추정량과 분산 및 분산 추정량을 구할 수 있을 것이다. 이러한 관점에서 다음과 같은 포함확률을 정의할 수 있다.

정리 4.1 JAD의 경우 포함확률 $\pi_i^{(JAD)}$ 는 다음과 같다.

$$\pi_i^{(JAD)} = 1 - \left[\binom{N-s_i}{n_1}\Big/\binom{N}{n_1}\right]. \tag{4.1}$$

여기서 s_i는 표본에서 단위 i와 교차하는 네트워크의 총 수로서 초기표본에 의존하는 값이다.

정리 4.2 $JADI$의 경우 포함확률 $\pi_i^{(JADI)}$ 는 다음과 같다.

$$\pi_i^{(JAD)} = 1 - \left[\binom{N-s_i-J_i}{n_1}\Big/\binom{N}{n_1}\right]. \tag{4.2}$$

여기서 s_i는 정의4.1와 같고, J_i는 내삽되는 네트워크의 수이다.

정의 4.1로부터 다음과 같은 정리를 도출할 수 있다.

정리 4.3 GAD의 네트워크의 수 $A_i^{(GAD)}$ 는 SAD의 총 네트워크의수 $A_i^{(GAD)}$ 와 같다.

이를 위해 SAD를 제외한 3가지 추출방법에 대한 표본평균을 각각 $\widetilde{\mu}^{(JAD)}$, $\widetilde{\mu}^{(JADI)}$, $\widetilde{\mu}^{(GAD)}$ 라고 표시하자. 또한 그에 따른 MSE 또한 3가지 추정량에 대해 같은 형식으로 표현하기로 한다.

4.2. 모수 추정과 MSE

4.2.A. JAD 방법

만일 단위 i에 대한 포함확률을 알 수 있다면, 정의 4.1의 식 (4.1) 으로부터 다음과 같이 모평균 μ에 대한 추정량 $\widetilde{\mu}_{HT}^{(JAD)}$ 을 도출할 수 있다.

$$\hat{\mu}_{HT}^{(JAD)} = \frac{1}{N}\sum_{i=1}^{n}\frac{y_i}{\pi_i^{(JAD)}}. \tag{4.3}$$

단위 k에 대한 네트워크의 교차확률 $\alpha_k^{(JAD)}$ 라 하고, 이 값을 알고 있다면 다음과 같이 $\widetilde{\mu}^{(JAD)}$ 을 다시 표현 할 수 있다.

$$\widetilde{\mu}^{(JAD)} = \frac{1}{N}\sum_{i=k}^{\kappa}\frac{{y_k}^*}{\alpha_k^{(JAD)}}. \tag{4.4}$$

여기서 단위 k의 교차확률은 다음과 같다.

$$\alpha_k^{(JAD)} = 1 - \left[\binom{N-{s_k}^*}{n_1}\Big/\binom{N}{n_1}\right]. \tag{4.5}$$

또한 j와 k번째 네트워크가 교차할 확률은 다음과 같다.

$$\alpha_{jk}^{(JAD)} = 1 - \left[\binom{N-{s_j}^*}{n_1} + \binom{N-{s_k}^*}{n_1} - \binom{N-{s_j}^*-{s_k}^*}{n_1}\right]\Big/\binom{N}{n_1}. \tag{4.6}$$

이때, ${s_i}^*$와 ${s_j}^*$는 각각 가장자리 단위를 제외하고, 단위 i와 단위 j에 교차하는 네트워크수를 나타낸다.

추정량 식 (4.4) 이 모평균 μ의 비편향 추정량이면, 식 (3.8), (3.9) 에서의 분산 및 분산 추정량의 형식을 적용할 수 있지만, 실제로 식 (4.3) 과 (4.4) 가 μ의 비편향 추정량임을 완성된 형태의 수식으로증명하기란 어렵다. 왜냐하면 표본의 군락 형태에 따라 건너뜀의 방향이 다양하기 때문에, 그에 따른 교차확률을 완성된 수식으로 표현하기가 어렵기 때문이다. 따라서 식 (4.3) 과 (4.4) 에 대한 추정량의 성질은 몬테카를로 실험을 통해 수치적인 방법으로 유도하고자 한다. 이와 같은 관점에서 $\widetilde{\mu}^{(JAD)}$ 의 MSE는 다음과 같이 정의될 수 있다.

$$MSE(\widetilde{\mu}^{(JAD)}) = E(\widetilde{\mu}^{(JAD)} - \mu)^2. \tag{4.7}$$

모수추정을 위해 포함확률 또는 교차확률을 이용하는 방법 이외에 식 (3.11) 과 같이 네트워크의 초기교차수를 이용하여 모평균 μ를 추정할 수 있다. 이를 정의하면 다음과 같다.

$$\widetilde{\mu}_{HH}^{(JAD)} = \frac{1}{n_1}\sum_{i=1}^{n_1}\frac{1}{m_i}\sum_{j\in s_i} y_j = \frac{1}{n_1}\sum_{i=1}^{n_1} w_i = \overline{w}^{(JAD)}. \tag{4.8}$$

여기서 w_i는 네트워크 s_i안에 있는 m_i개 관측치들의 평균이다.

이와 함께 $\widetilde{\mu}_{HH}^{(JAD)}$ 가 만일 μ의 비편향 추정량이면 식 (3.13) 과 (3.14) 의 분산 및 분산 추정량에서 추정량을 대체하여 계산할 수 있으나, 앞에서 언급한 바와 같이 JAD의 경우 단위 i에 교차하는 네트워크의 방향이 군락의 형태에 의존함으로 추정량의 비편향성을 보장할 수 없다. 따라서 다음과 같이 추정량의 MSE를 계산하기로 한다.

$$MSE(\widetilde{\mu}_{HH}^{(JAD)}) = E(\widetilde{\mu}_{HH}^{(JAD)} - \mu)^2. \tag{4.9}$$

4.2.B. JADI 방법

정의 4.2로부터 포함확률이 식 (4.2) 를 적용한 모평균 μ의 추정량은 다음과 같다.

$$\hat{\mu}_{HT}^{(JADI)} = \frac{1}{N}\sum_{i=1}^{n}\frac{y_i}{\pi_i^{(JADI)}}. \tag{4.10}$$

이와 더불어 교차확률을 알 수 있다면, 추정량은

$$\widetilde{\mu}^{(JADI)} = \frac{1}{N}\sum_{i=k}^{\kappa}\frac{{y_k}^*}{\alpha_k^{(JAD)}} \tag{4.11}$$

이고, 이때 식 (4.5) 와 (4.6) 과 유사하게 각각의 교차확률을 정의할 수 있다.

$$\alpha_k^{(JADI)} = 1 - \left[\binom{N-{s_k}'}{n_1}\Big/\binom{N}{n_1}\right]. \tag{4.12}$$

또한 j와 k번째 네트워크가 교차할 확률은 다음과 같다.

$$\alpha_{jk}^{(JADI)} = 1 - \left[\binom{N-{s_j}'}{n_1} + \binom{N-{s_k}'}{n_1} - \binom{N-{s_j}'-{s_k}'}{n_1}\right]\Big/\binom{N}{n_1}. \tag{4.13}$$

여기서 $s_i' = s_i + J_i$로서 단위 i의 교차수와 그에 따른 내삽된 단위 수의 합으로 표현된다. JAD와 같은 맥락에서 $\widetilde{\mu}^{(JADI)}$ 의 MSE는 다음과 같이 정의된다.

$$MSE(\widetilde{\mu}^{(JADI)}) = E(\widetilde{\mu}^{(JADI)} - \mu)^2. \tag{4.14}$$

한편 단위 i에 대한 네트워크의 초기 교차수를 이용하여 추정량을 계산하면 식 (4.8) 과 같은 형식으로 μ의 추정량을 다음과 같이 정의할 수 있다.

$$\widetilde{\mu}_{HH}^{(JADI)} = \frac{1}{n_1}\sum_{i=1}^{n_1}\frac{1}{m_i}\sum_{j\in s_i'} y_j = \frac{1}{n_1}\sum_{i=1}^{n_1} w_i = \overline{w}^{(JADI)}. \tag{4.15}$$

여기서 w_i는 네트워크 $s_i' = s_i + J_i$안에 m_i개 관측치들의 평균이다. 또한 $\widetilde{\mu}_{HH}^{(JADI)}$ 에 대한 MSE는 식 (4.9) 와 같은 형식을 갖는다.

4.2.C. GAD 방법

GAD는 총 네트워크의 수가 SAD와 동일함으로 결과적으로 초기표본이 SAD를 적용한 경우와 같다면, GAD추정량은 SAD추정량의 성질을 그대로 유지함으로 $\widetilde{\mu}^{(GAD)}$은 μ의 비편향 추정량이며, 분산과 분산 추정량은 식 (3.8) 과 (3.9) 와 같다. 이와 함께, 초기 교차수를 이용한 추정량의 경우 또한 SAD와 동일한 성질을 유지한다.

5. 모의실험에 의한 효율성 비교

5.1. 난수발생

본 논문에서 고려한 표본추출계획들에 대해 수정된 HH추정량 $\widetilde{\mu}_{HH}{}^*$의 비편향성을 보장할 수 없고, 또한 MSE에 대한 이론적 비교도 어려우므로 우리는 추출방법들에 대해서 몬테카를로 모의실험으로 MSE 값을 계산하여 효율성을 비교하였다.

먼저 관심영역을 20×20개의 격자로 나눈다. 적응집락추출을 구현하기 위해 자료가 군락의 형태로 모여 있는 모집단을 랜덤으로 생성해야 한다. 이를 위해 Diggle (1983) 의 포아송 군집과정 (Poisson cluster process) 을 사용하여 다음 절차에 따라 난수를 발생하였다 (Consiglio와 Scanu, 2001).

단계 1. $U(0,1) \times U(0,1)$로부터 n_1개의 난수를 발생시킨 후 이것을 각각 $(x_1, y_1), \cdots, (x_{n1}, y_{n1})$이라 표시한다.

단계 2. 포아송 분포 $P(\lambda)$로부터 n_1개의 난수를 발생시킨 후 이것을 $m_1, m_2, \ldots, m_{n1}$이라고 표시하자. 즉 m_i는 i번째 군집 내에 존재하는 개체의 수를 의미한다 ($\lambda = 10, 30$).

단계 3. $i = 1, \ldots, n_1$에 대하여, 이변량 정규분포 $BN(x_1, y_i, \sigma, \sigma, 0)$ 으로부터 m_i개의 난수를 발생시킨다 ($\sigma = 0.8, 1.2, 1.6$).

포아송 군집과정을 이용하여 생성된 난수를, 예를 들어 그려보면 그림 5.1과 그림 5.2와 같은 산포를 나타난다.

5.2. 적응추출을 구현하는 알고리즘

수정된 적응집락 추출의 과정을 컴퓨터로 구현하기 위해 미로 탐색 알고리즘을 수정하여 적용하였다 (이석호, 1993). 이는 이차원 배열로 표현된 미로에서 미로 속의 탐색자의 위치를 행과 열로 언제나 표현할 수 있다. 또한 한번 방문한 지역은 다시 방문하지 않는 형태를 잘 구현한다. 먼저 모든 셀을 0으로 마크한 후, 샘플링을 하게 되는 (또는 이미 조사한) 셀은 -1값을 줘서, 차후에 -1을 갖는 셀은 다시 방문하지 않는다. 이 알고리즘은 초기에 추출된 셀에서 관측값이 발견되었을 경우 그 위치값 (r, c) 와 관측값을 창고 (stack) 에 저장한 후 시계방향으로 주변 셀을 탐색하여 값을 조사한다. 일단 시

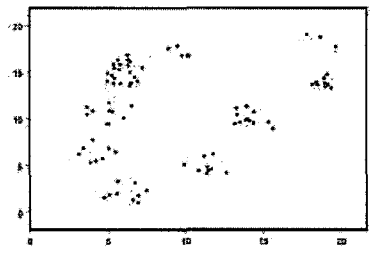


그림 5.1: $n_1 = 10$, $\sigma = 0.8$, $\lambda = 10$

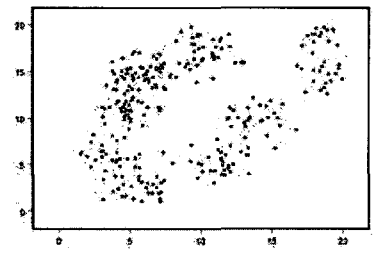


그림 5.2: $n_1 = 10$, $\sigma = 1.2$, $\lambda = 30$

계방향에 의한 탐색이 끝나더라도 아직 조사가 안 된 셀이 존재하게 되므로, 그 끝난 지점에서 저장된 값을 참조하고 이제 시계반대 방향으로 이동하면서 -1로 마크되지 않은 셀을 조사하게 된다.

원래의 미로 알고리즘은 현재의 셀에서 기본적으로 주변 8개의 셀을 고려하지만, 본 연구에서는 주변 4개의 셀만 고려하는 알고리즘으로 바꿔서 이용했다. 이 알고리즘은 컴퓨터 모의실험에서 뿐만아니라 현장에서 실제로 적응추출을 사용하는 데에도 적용할 수 있다. 수정된 적응집락 추출의 과정을 구현한 컴퓨터 프로그램은 김연우 (2005) 에서 구할 수 있다.

5.3. 모의실험 결과

표 5.1은 모의실험에서 구해진 추정값과 MSE이다. 각 방법에 대해서 비편향성에 대해 확정적으로 말할 수는 없지만 대체로 SAD가 다른 방법들보다 작은 편차를 보였다. MSE에서도 큰 차이를 보이지는 않았지만, 약간의 차이를 해석하자면 σ가 작을 때 SAD가 편향이 가장 작은 것으로 나타났다. 또한 $\lambda = 30$이고, σ가 클 때에는 JAD가 SAD보다 MSE가 작게 나타났다. 이는 모집단의 분포 형태에 따라 추출방법들의 효율성이 달라지는데, 군락의 크기가 크고 넓게 퍼진 경우에는 JAD가 SAD보다는 정도가 조금 높게 나타난 반면, 군락의 분포가 밀집된 경우에는 기존의 방법이 더 효율적임을 의미한다. 이러한 결과는 어떤 방법이 다른 방법들에 비해 항상 효율적이지는 않으며, 모집단의 형태에 따라 효율적인 추출방법을 적용하는 것이 바람직한 방법이라는 것을 나타내는 것이다. 한편 $JADI$는 JAD에 비해 MSE가 작아지지 않았으며, GAD는 SAD보다는 MSE 가 크게 나타났지만, JAD보다는 작은 값을 갖는 것으로 나타났다.

원래 본 연구는 JAD를 이용하여 SAD보다 조사하고자 하는 표본의 크기를 줄이려는 목적을 가졌으므로, 실제 모의실험에서의 표본의 크기의 평균을 계산한 결과가 표 5.2와 같다. 표 5.2로 부터 $JADI$는 JAD와 같은 표본의 크기를 갖는 반면에 JAD가 SAD와 GAD보다 평균적으로 약간 작은 표본 규모를 갖는 것을 볼 수 있다. 그러나 예상과는 달리 줄어든 표본의 규모는 그다지 크지 않은 것으로 나타났다.

표 5.1: 집락내의 평균 개체 수(λ)와 집락의 넓이(σ)에 따른 추정치

σ	$\lambda=10, n_1=10$						$\lambda=30, n_1=10$					
	μ	추정량	SAD	JAD	$JADI$	GAD $(c=2)$	μ	추정량	SAD	JAD	$JADI$	GAD $(c=2)$
0.8	0.225	$\widetilde{\mu}$	0.226	0.222	0.221	0.225	0.750	$\widetilde{\mu}$	0.748	0.758	0.751	0.753
		$MSE(\widetilde{\mu})$	0.044	0.048	0.050	0.046		$MSE(\widetilde{\mu})$	0.533	0.536	0.541	0.535
1.2	0.223	$\widetilde{\mu}$	0.223	0.218	0.224	0.220	0.730	$\widetilde{\mu}$	0.732	0.720	0.723	0.727
		$MSE(\widetilde{\mu})$	0.046	0.047	0.049	0.046		$MSE(\widetilde{\mu})$	0.537	0.537	0.540	0.537
1.6	0.220	$\widetilde{\mu}$	0.220	0.226	0.227	0.222	0.725	$\widetilde{\mu}$	0.726	0.722	0.721	0.725
		$MSE(\widetilde{\mu})$	0.048	0.047	0.047	0.047		$MSE(\widetilde{\mu})$	0.539	0.537	0.537	0.539

표 5.2: 추출방법에 따른 평균 표본의 크기

	$\lambda=10, n_1=10$			$\lambda=30, n_1=10$		
σ	SAD	JAD	GAD	SAD	JAD	GAD
$\sigma=0.8$	19.98	15.37	16.52	36.37	29.60	31.93
$\sigma=0.8$	21.39	16.68	17.87	38.46	31.56	34.84
$\sigma=0.8$	24.70	18.41	20.03	41.86	33.39	37.18

6. 결론

군락의 형태가 크고 넓게 퍼져있을 경우, 기존의 적응집락추출계획보다 조사 단위수를 줄일 수 있도록 이웃을 건너뛰는 방식의 변경된 적응집락추출설계들을 제안하였다. 이들의 경우 모수의 추정과 몬테카를로 방법에 의해 계산된 MSE 값을 이용하여 효율성을 비교하였다.

모의실험의 결과, 모집단의 분포 형태에 따라 추출방법들의 효율성이 달라지는데, 군락의 크기가 크고 넓게 퍼진 경우에는 건너뛴 적응집락추출방법 (JAD) 가 기존의 추출방법 (SAD) 보다는 정도가 조금 높게 나타난 반면, 군락의 분포가 밀집된 경우에는 기존의 방법이 더 효율적이다는 것이다. 이러한 결과는 어떤 방법이 항상 다른 방법에 비해 효율적이지는 않으며, 단지 모집단의 형태에 따라 효율적인 추출방법을 적용하는 것이 현실적임을 의미한다. 표본의 크기 면에서는 JAD가 SAD보다 그 크기를 줄일 수 있다는 점에서 약간 효율적이라고 말할 수 있다. 모집단의 형태가 대략 어떻게 구성되어 있을 것인가에 대한 사전연구를 통하여 적절한 적응집락추출계획을 수립하는 것이 바람직할 것으로 판단된다.

감사

건설적인 제안을 해주신 심사위원들에게 감사드린다.

참고문헌

김연우 (2005). 효율적인 적응추출 계획. 전남대학교 대학원 석사학위논문.

남궁평, 변종석 (2001). Optimal design of the adaptive searching estimation in spatial sampling. <한국통계학회논문집>, **8**, 73–85.

이석호 (1993). <C로 쓴 자료구조론>. 사이텍미디어, 서울.

이해용, 이필영 (2002). <표본조사입문>. 교우사, 서울.

Consiglio, L. D. and Scanu, M. (2001). Some results on asymptotics in adaptive cluster sampling. *Statistics and Probability Letters*, **52**, 189–197.

Diggle, P. J. (1983). Some models for bivariate point patterns. *Journal of Royal Statistical Society*, Ser. B, **45**, 11–21.

Dryver A. L. and Thompson S. K. (2005). Improved unbiased estimators in adaptive cluster sampling. *Journal of Royal Statistical Society*, Ser. B, **67**, 157–166.

Felix-Medina M. H. and Thompson S. K. (2004). Adaptive cluster double sampling. *Biometrika*, **91**, 877–891.

Hedayat, A. S. and Stufken, J. (1998). Sampling designs to control selection probabilities of contiguous unit. *Journal of Statistical Planning and Inference*, **72**, 333–345.

Lee, K. J. (1998). Two-phase adaptive cluster sampling with unequal probabilities selection. *Journal of Korean Statistical Society*, **27**, 265–278.

Magnussen S., Kurz W., Leckie D. G. and Paradine D. (2005). Adaptive cluster sampling for estimation of deforestation rates. *European Journal of Forest Research*, **124**, 207–220.

Mohammad, S. M. (2003), Comparison between Hansen-Hurwitz and Horvitz-Thompson estimators for adaptive cluster sampling. *Environment and Ecological Statistics*, **10**, 115–127.

Noon B. R., Ishwar N. M. and Vasudevan K. (2006). Efficiency of adaptive cluster and random sampling in detecting terrestrial herpetofauna in a tropical rainforest. *Wildlife Society Bulletine*, **34**, 59–68.

Thompson, W. L. (2004). *Sampling Rare or Elusive Species.* Island Press, Washington DC.

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

Turk P. and Borkowski J. J. (2005). A review of adaptive cluster sampling: 1990-2003. *Environmental and Ecological Statistics*, **12**, 55–94.

[Received July 2006, Accepted November 2006]