효과적인 적응집락추출계획

김연우[1], 손창균[2], 박정수[3]

요 약

보통 생태학 분야 등에 적용될 수 있는 적응집락추출계획(adaptive cluster sampling plan)을 수정하여, 표본의 크기 면에서 더 효율적인 Jumped 및 일반화 적응집락추출계획을 제안하였다. 이러한 계획 하에서 Hansen-Hurwitz(HH)와 Horvitz-Thompson (HT) 추정량으로 모수를 추정하였다. 제안한 새로운 계획들을 시뮬레이션을 통하여 기존의 계획과 비교하였다.

1. 서 론

적응추출(Adaptive Cluster Sampling, Thompson and Seber, 1996)은 초기에 추출한 추출단위를 대상으로 조사하는 도중 그 표본단위에 인접하고 있는 표본조사 단위의 조사를 추가적으로 조사할 것인지에 대해 조사결과를 보아가면서 결정하는 표본추출법이다. 본 논문에서는 이와같이 군집의 형태로 모여있는 자료에 대하여 기존에 연구되어진 *SAD* 방법과 *JAD* 방법에 대해 시뮬레이션을 통해 여러 가지 조건 하에서 HH 추정량 와 HT 추정량을 비교했다. 또한 *JAD*를 보완하는 방법인 *JADi* 와 *SAD*와 *JAD* 방법을 혼용한 *GAD* 표본추출 방법을 통해 표본 선택의 효율성에 대해 논의했다.

2. 적응추출에서의 추정

2.1 Horvitz-Thompson(HT) 추정량

$$\widehat{\mu}_{HT} = \frac{1}{N}\sum_{k=1}^{N}\left(\frac{y_k^* I_k}{\alpha_k}\right), \qquad I_k = \begin{cases} 1 & y_i \geq 1 \\ 0 & otherwise \end{cases}$$

$$E[\widehat{\mu}_{HT}] = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i}{\alpha_i}\right)E[I_i] = \frac{1}{N}\sum_{i=1}^{N} y_i = \mu$$

$$\alpha_{ij} = 1 - \frac{\binom{N-m_k}{n_1} + \binom{N-m_h}{n_1} - \binom{N-m_k-m_h}{n_1}}{\binom{N}{n_1}}$$

여기서 α_{ij}는 두 네트워크 간에 결합확률이다(Thompson, 1996).

---

1) 전남대학교 통계학과 박사과정.
2) 경기도 화성시 봉담읍 상리 14번지, 협성대학교 교양학부 교수.
3) 전남대학교 통계학과 교수, E-mail: jspark@chonnam.ac.kr. 본 연구는 한국과학재단 국제공동연구사업 (과제번호: F01-2004-000-10351-0)에 의해 지원 받았음.

2.2 Hansen-Hurwitz(HH) 추정량

모집단에서 네트워크의 수를 K라고 정의하자. 또한 A_k는 k번째 네트워크이고, m_k는 네트워크 A_k안에 들어있는 셀들의 수가 된다. 네트워크 안에서 발견된 셀에 대한 관측값들의 평균 w_k 과 전체평균 $\hat{\mu}_{HH}$는

$$w_k = \frac{1}{m_k}\sum_{j\in A_k} y_i \quad , \qquad \widehat{\mu_{HH}} = \frac{1}{n_1}\sum_{i=1}^{n_1}\frac{1}{m_i}\sum_{j\in A_i} y_j = \frac{1}{n_1}\sum_{i=1}^{n_1} w_i = \overline{w}$$

이 된다. w_i는 네트워크 A_i안에서 관측된 m_i들의 평균값을 의미하게 된다.

2.3 표본의 크기

또한 부가추출에서 각 방법에 따라 표본 크기의 기대값을 비교하기 위해, 한 셀당의 표본의 크기의 기대값을 $E[V]$라고 하면,

$$E[V] = N - \frac{1}{N!}\sum_{i=1}^{N}(N-m_i-a_i)!(N-n_1-m_i-a_i+1)\times\cdots\times(N-n_1)$$

로써 각 방법들에 대한 한 개의 셀에 대한 기대 샘플크기를 구할 수 있다. 표본 기대값에는 edge 영역을 포함하여 실제로 관측을 위하여 방문했던 지역들을 모두 포함시키기 때문에 샘플의 영역이 더 확장되게 된다. 각 표본추출의 방법에 대한 기대 표본의 크기는 $m_i + a_i$ 의 차이에 따라 달라지게 된다. 그러므로 한 셀당 표본의 크기 $E[V]$는 다음과 같이 표현 될 수 있으며,

$$SAD\ a_i \geq GAD\ a_i \geq JAD\ a_i = JADi\ a_i\ .$$

시뮬레이션의 결과를 통해 표본의 크기를 비교하겠다.

3. 제안된 적응추출계획

나열한 모든 표본추출 방법은 단순임의 추출을 통해 뽑은 셀이 관측값을 가지고 있을 때 그 값을 중심으로 인근으로 범위를 확장해 가면서 값을 조사하는 방법으로 각각의 특성을 기술하였다. 모집단을 여러개의 격자 형태로 구분할 수 있다. 그리고 각 격자들은 각각이 고유의 열과 행의 값을 가지게 된다. 그 위치는 (row, col)으로 표시하기로 한다.

3.1 단순 적응 설계(Simple Adaptive Design)의 형태

초기의 샘플 (i,j)로부터 $(i-1,j),(i+1,j),(i,j-1),(i,j+1)$번째 값을 추출하는 방식으로 무작위 추출을 통해 표본이 되는 셀을 구하고 되고 그 셀에 관측값이 존재할 경우 주변으로 확대하는 방법이다. 모든 셀은 이웃하는 셀들과 근접하여 있으며 이를 '주변 셀'이라고 한다. 표본의 크기 n_1(음영이 있는 셀)은 단순 임의 추출법을 사용하여 발생된다. 선택된 셀에서 발견된 값을 y라고 했을 때 이 y값이 조건 c를 만족한다면, 즉 $y_i > c$라면 선택된 셀의 주변 셀들을 추가적으로 추출 하게 된다. 이 프로세스는 주변셀에서 y값이 조건 c를 만족하지 않는 경계선에 이를 때까지 계속된다. 이 경계선을 이루는 셀을 edge라 한다. 초기에 선택되어진 셀이 조건 c를 만족하지 않는 경우 더 이상 주변 셀들은 추가되지 않고 크기 1인 군집을 이루게 된다. 만약 초기에 선택되어진 셀이 edge 셀일 경우도 크기가 1인 군집으로 인식하며 표본추출은 종료되게 된다. 적응추출을 통해 수집된 셀들은 조건 c를 만족하는 셀들(m_i)과 그 edge(a_i) 셀들로 이루어져 있으며, 이를 통합하여 네트워크 A_i $(m_i + a_i)$라 한다.

3.2 건너뜀 적응 설계(Jumped Adaptive design)의 형태

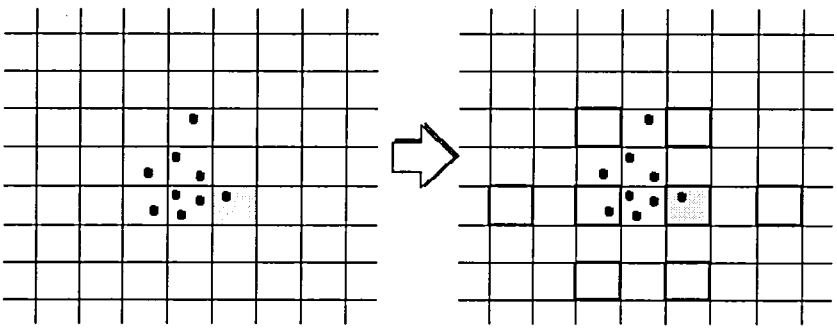


그림2) 건너뛴 적응 설계

표본이 군집의 형태를 이루고 있다는 조건하에서 한 단계씩 건너 뛰어 샘플링을 하는 조사방법이다. 초기의 샘플 (i,j)로부터 $(i-2,j)$, $(i+2,j)$, $(i,j-2)$, $(i,j+2)$번째 값을 추출하는 방식으로 조건 c와 상관없이 관측값의 유무에 따라서 이동하며 표본을 추출하는 방식이다.

3.3 보완된 건너뜀 적응 설계(Jumped Adaptive interpolation design)

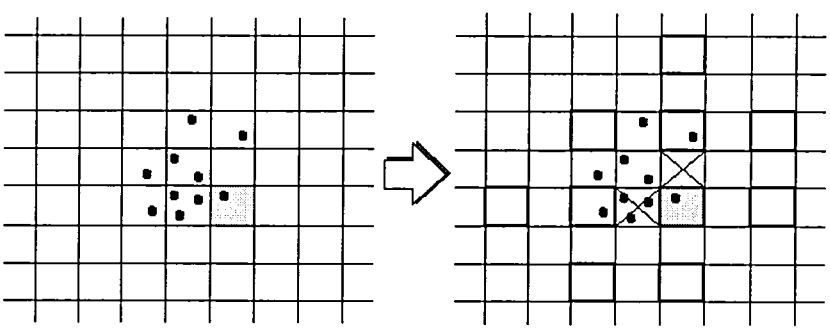
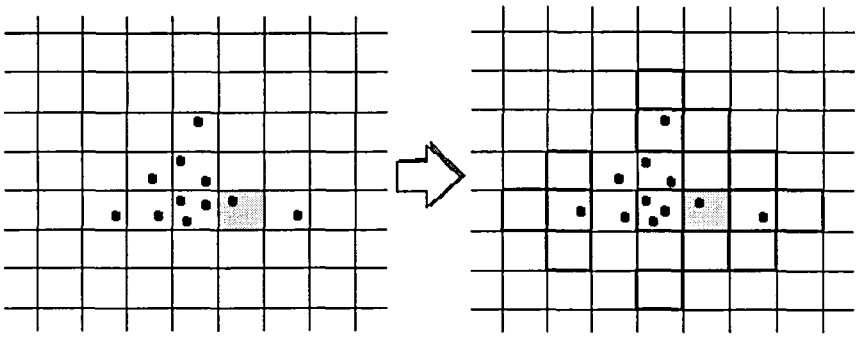


그림3) 보완된 건너뜀 적응 설계

*JAD*의 단점을 보완하기 위하여 관측된 셀의 사이 값에 양측 셀의 평균값으로서 보정을 해 주는 방법이다. 초기의 샘플 (i,j)로부터 $(i-2,j),(i+2,j),(i,j-2),(i,j+2)$ 번째 값을 추출하고, 다음과 같이 그 사이값을 보정해 주게 된다. *SAD* 의 경우 모든 관측값을 다 조사하게 됨으로 조사방식의 효율성이 낮아질 수 있고, 반면에 *JAD* 의 경우는 데이터의 특성을 지나치게 축소시킬 수 있다는 점을 보완하기 위해 방문하지 않은 사이값에 대해 산술평균 값으로 보정을 해 주는 방법이다.

3.4 일반화된 적응 설계(Generalized Adaptive design)의 형태 (c=2인 경우)

GAD는 한계값에 따라 *SAD* 와 *JAD* 을 병행하여 사용하는 방법이다. 한계값 $c=2$로 해 주었을 경우 표본추출을 한 셀에서 관측값이 c보다 크게 관측될 경우 *JAD* 를 적용하고 그 이하의 관측값이 나올 경우에는 *SAD* 를 적용하여 표본추출을 시도하는 방법이다. 초기의 샘플 (i,j)로



부터

그림4) 일반화된(혼합형) 적응 설계

$$\begin{cases} (i-1,j),(i+1,j),(i,j-1),(i,j+1) & \text{if } y_i<c \\ (i-2,j),(i+2,j),(i,j-2),(i,j+2) & \text{if } y_i\geq c, \end{cases} \qquad \begin{cases} c\to 1 & \Rightarrow SAD \\ c\to \infty & \Rightarrow JAD \end{cases}$$

가 된다. 조사되는 관측값을 보고 다음 진행상황을 결정하는 표본추출방법이다.

4. 시뮬레이션

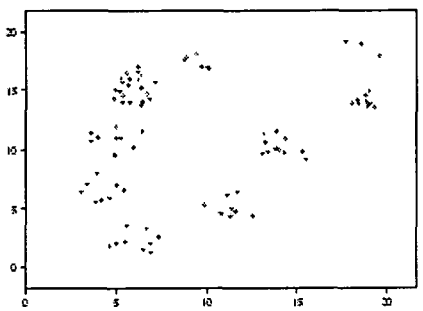


그림 5) $n=10,\ \sigma=0.8,\ \lambda=10$

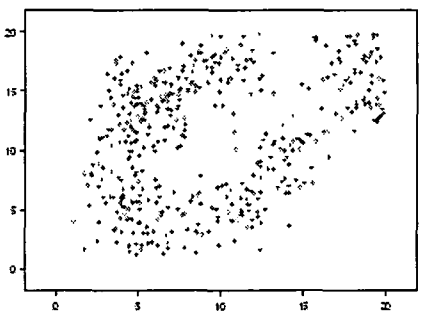


그림 6) $n=10,\ \sigma=1.6,\ \lambda=50$

생성된 난수 집락의 분포 형태

4.1 난수의 생성

Diggle(1983)의 포아송 군집과정을 사용하여 다음 절차를 이용하여 난수를 생성하였다.

step 1) 균일 분포로부터 (x, y)의 쌍을 이루게끔 x, y에 대해서 각각 n개의 난수를 생성한다.

step 2) 단위 면적에서 발생하는 난수를 생성하기 위하여 포아송 분포로부터 $\lambda=10,30,50$인 경우의 난수를 각각 n개 생성한다.

step 3) 균일분포를 통해 생성된 난수 (x,y)의 쌍 n개를 이변량 정규분포 μ_x, μ_y에 대입한다. 난수의 쌍은 독립이므로 $BN(\mu_x,\mu_y,\sigma,\sigma,0)$을 따르게 되며, 각각에 대해서 $\sigma=0.8, 1.2, 1.6, 2.0$를 대입하여 step2에서 포아송 분포에 의해 발생된 난수의 수만큼 이변량 정규분포를 따르는 난수를 발생한다.

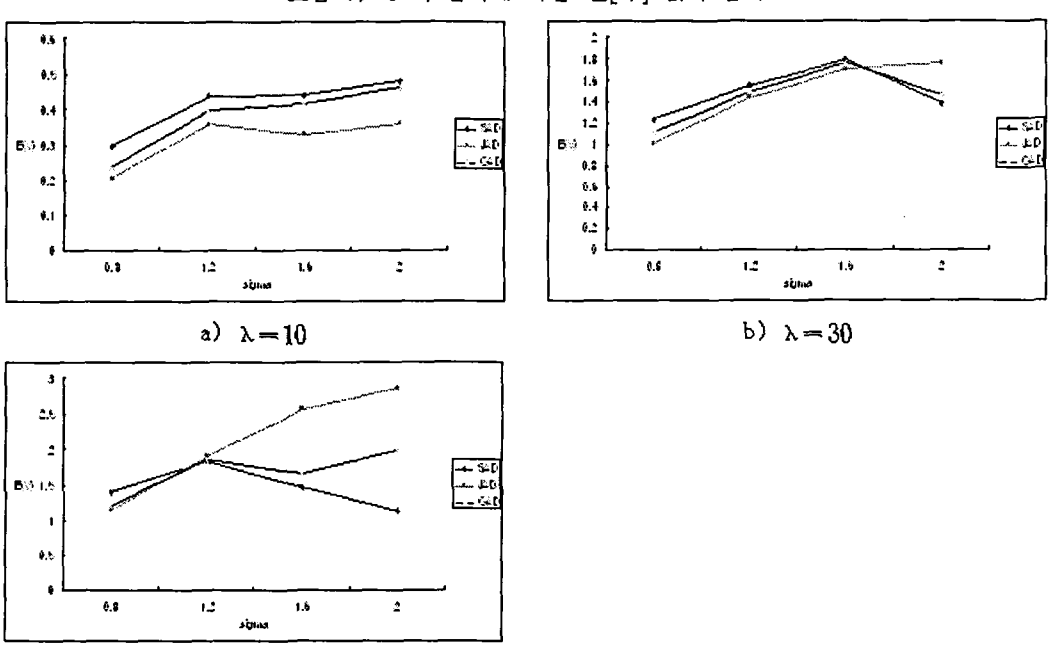
step 4) 조건 n개를 만족할 때까지 반복하여 step 2)로 이동

포아송 군집추출과정을 이용하여 생성된 난수로 plot를 그려보면 아래와 같은 형태로 값들이 배치되게 된다. 여기서 n은 균일 분포를 통해 발생한 난수를 의미하며, σ는 표준편차를, λ는 포아송분포에 근사시켜 n을 중심으로 난수의 발생된 난수를 수를 뜻한다.

4.2 시뮬레이션 결과

그림7)은 σ와 모집단의 형태에 따라서 셀 당의 평균 기대값이 어떻게 차이가 나는지에 관해 그래프를 그려본 결과이다. 여기서 *JAD*와 *JADi*는 실제로 같은 셀로 이루어 있기 때문에 같은 것으로 간주하고 *SAD*, *GAD* 세 가지 방법에 대해서 비교해 보았다. 그림 a)은 개체수가 적은 경우 즉, $\lambda=10$인 경우로 σ의 크기와 상관없이 전 범위에서 *JAD*를 통한 샘플링 추출 방법이 가장 효율적인 결과를 나타내고 있다. 그림 b)은 $\lambda=30$인 경우로 개체수를 증가시킨 후에 각 셀당의 기대값을 그래프로 그린 경우로써 $\sigma=2.0$인 경우를 제외하고는 앞서와 마찬가지로 *JAD*를 통한 표본추출 방법이 가장 좋은 것으로 나타났다. $\sigma=2.0$인 경우는 자료가 너무 넓게 퍼지는 경우로써 군집적인 형태를 띄지 않게 된다. 그 결과 오히려 *JAD*방법을 통한 샘플링 추출은 다른 네트워크까지 그 범위를 확장시켜서 기대샘플을 추출하게 됨으로 적응추출을 적용하지 않다. 그림 c)는 $\lambda=50$인 경우로써 개체가 많은 경우를 가지고 적응추출을 적용해본 결과이다. 대부분의 경우에서 *SAD* 방법을 통한 추출이 효율적인 경우로 나타나며 *JAD*는 큰 값을 갖게 된다.

그림 7) σ 의 변화에 따른 $E[V]$ 값의 변화



a) λ=10

b) λ=30

c) λ=50

4. 결론 및 토의

기존의 사용되어 지고 있는 *SAD*방법을 좀 더 효율적으로 개선하고자 새로운 추정방법들을 제안하였다. 시뮬레이션 결과 자료가 모든 지역에서 HH추정량 보다는 HT추정량이 참값을 추정하는데 있어서 좋은 결과를 가지고 있다. 실험설계에 있어서 일반적으로 논의가 되어지고 있는 *SAD* 방법은 추정에 있어서는 좋은 결과를 가지고 있지만 조사하는 범위가 너무 광범위하여 효율적인 측면에서는 그리 좋은 방법이 아닐 수 있다. 비교를 통해 4가지의 표본조사 방법의 시간 비용적 효율성을 고려해 보았을 때 다음과 같은 크기를 갖음을 알 수 있다.

$$eff(ACD) < eff(GAD) < eff(JCD) = eff(JCDi)$$

결국 최소의 표본추출로써 좀더 좋은 추정을 할 수 있는 방법을 추구한다는 면에서, 시뮬레이션의 결과에서 볼 수 있듯이 *JAD*, *JAD(i)*, *GAD* 방법을 통한 추정이 효율적임을 볼 수 있다.

참고문헌

이해용, 이필용 (2002), 표본조사입문, 교우사.

성내경 (2003), 표본조사방법론, 자유아카데미.

Anderson-Freed, H.S. (1992), *Data Structures in C*, 이석호 역, 사이텍 미디어.

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

Upton, G. and Fingleton, B. (1985), *Spatial Data Analysis by Example,* Volume 1, John Wiley & Sons, Inc.

Lawson, A.B. and Denison, D.G.T. (2002), *Spatial Cluster Modelling*, Chapman and Hall/CRC.

Salehim, M. (2003), Comparison between Hansen-Hurwitz and Horvitz-Thompson estimators for adaptive cluster sampling, *Environmental and Ecological Statistics* 10, pp.115-127.

Hedayat, A.S. and Sinha, J.(1998). Sampling designs to control selection probabilities of contiguous units. *Jour. Statist. Plann. Infer.* 72, pp.333-345.