2000 Proceedings of the Autumn Conference, Korean Statistical Society

환경 및 생태 모니터링 표본설계

김 선 웅[1], 류 제 복[2]

요 약

Hedayat et al.(1988)와 Stufken(1993) 등은 환경 또는 생태의 특성을 고려하여 인접 단위들을 추출확률을 최소화시킬 수 있는 표본설계방법을 제안하였다. 그리고 Hedayat et al.(1998)은 실제 조사에 있어 이들 방법보다 유용한 방법을 제시하였다. 본 논문에서는 이들 표본설계방법의 문제점들을 밝히고 이를 보완할 수 있는 방법과 추가적인 연구 방향 등을 제시한다.

1. 서 론

최근 공기, 물, 토양 등의 오염과 여러 동식물 종들의 멸종 등과 같은 환경적, 생태학적 측면의 연구에 대한 관심이 고조되어 있다. 이러한 영역에서의 연구를 수행하기 위해서는 환경 전문가나 생태학자들 뿐만 아니라 통계학자들의 참여가 있어야 하며 통계적 방법에 의한 자료 수집과 분석 등도 필수적이라 할 수 있다.

이러한 점에서 Cormack(1988)은 환경 과학 연구에 있어서 해결해야할 통계적 문제들은 수없이 많은 반면 이 분야에 참여하고 있는 통계학자들의 수는 극히 작다는 사실과 혁신적인 통계적 방법의 필요성을 강조하고 있다. 또한 Millard(1987)가 언급한 바와 같이 통계학자들과의 협력을 통하여 통계적 방법의 부적절한 사용으로 인한 오류는 크게 줄일 수 있다.

한편 환경적, 생태학적 연구와 관련하여 해결해야하는 주요 통계적 과제 중의 하나는 유한모집단에서의 표본추출방법에 관한 문제이다. 이는 환경 또는 생태 연구의 대상인 모집단의 특성상 고활동수준(high activity level)을 갖는 추출단위들이 아주 빈번히 서로 이웃하는 단위들로 구성되는 집락으로 나타나기 때문이다. 예를 들어 특정 지역의 희귀 동물들의 모집단 크기를 조사한다고 하자. 이를 위해서 그 지역을 아주 작은 단위들로 나눈 다음 이들 단위들 중 일부를 임의추출해서 각 단위들로부터 동물들의 수를 관찰하고 이들 자료를 바탕으로 모집단의 크기를 추정할 수 있다. 그런데 문제는 많은 종들이 사회적인 생활 습관을 가지고 있으므로 조사지역에 임의적으로 흩어져서 살지 않는다는 것이다. 다시 말해서 사회적 생활 구조 때문에 이 동물들을 관찰할 수 있는 단위들은 대개 서로 인접한 단위들로 구성되는 집락의 일부로서 고활동수준을 갖는다. 또 다른 예로서 화학공장의 생산과정에서 만들어지는 부산물들을 특정 지역들에 폐기한다고 하자. 만약 어떤 화학물질의 밀도가 적정 수준을 넘으면 해당 지역을 위험지역으로 분류한 후 적절한 대책을 강구하게 된다. 이를 위해서는 조사 지역을 작은 표본추출단위로 나눈 다음 이들 중 임의추출된 단위들의 흙을 주어진 절차에 따라 분석하는 방법을 생각할 수 있다. 이 경우 만약 추출된 단위가 특정 물질을 다량 포함하고 있다면 이웃하는 단위들도 거의 동일한 수준을 보일 것이다. 즉 앞의 예와 마찬가지로 고활동수준을 갖는 단위들이 서로 이웃하는 단위들로 구성되는 집락이 될 것이다.

1) (100-715) 서울 중구 필동 3가 26번지 동국대학교 통계학과
2) (360-764) 충청북도 청주시 상당구 내덕동 36번지 청주대학교 자연과학부 통계학 전공 교수

위의 두 가지 예에서 알 수 있는 것과 같이 연구 목적에 따라 표본추출방법을 2가지로 구분하여 생각해볼 수 있다. 하나는 모니터링을 위한 표본추출방법이고 다른 하나는 Thompson(1990)의 적응집락추출법(adaptive cluster sampling) 또는 Thompson(1991)의 층화적응집락추출법(stratified adaptive cluster sampling)을 이용하는 것이다. 예를 들어 일단 추출된 단위가 적정 수준이상의 화학물질을 포함하고 있어 이 단위의 주위에 있는 단위들 또는 그 추출단위가 속해있는 집락 내에 있는 단위들을 모두 조사하고자 한다면 후자의 방법을 이용할 수 있다. 그러나 단지 모니터링을 목적으로 한다면 적응집락추출법의 사용은 불필요하며 바람직하지도 않을 것이다.

모니터링을 위한 표본설계에 관한 연구는 Hedayat et al.(1988), Stufken(1993), Hedayat 등(1998)에 의하여 이루어졌으며 특히 Hedayat et al.(1998)이 제안한 방법은 이전의 방법과 달리 실제조사에 있어 적절한 방법이라 할 수 있다.

본 논문에서 먼저 이들 방법들에 관하여 간략히 소개하고 Hedayat et al.(1998)이 제안한 표본설계의 문제점들을 언급한다. 또한 이 문제점들을 개선시킬 수 있는 방법을 제시하고 앞으로의 추가적인 연구과제에 대해서도 설명하기로 한다.

2. 기존 표본설계

본 논문에서 다루어지는 내용들은 환경 또는 생태 모니터링을 위하여 단순임의비복원추출(SRSWOR)을 대신할 수 있는 표본추출방법으로서 인접단위들의 추출확률을 조정하는 표본설계방법들에 관해 다룬다.

환경 또는 생태 모니터링을 하는데 있어 일단 모집단을 층화한 후 각 층을 일정 크기의 작은 단위로 나누고 적절한 표본추출방법을 사용하여 표본을 추출하는 것이 바람직하다. 그리고 모니터링이외의 추가적인 연구가 필요하면 이 표본추출방법들을 응용집락추출법 등과 연계하여 사용할 수 있을 것이다.

표본설계를 위한 기본 가정으로서 $R \times C$ 직사각형 구획(rectangular layout) 내에 N개의 단위들이 있는 것으로 간주한다. 즉 $N = RC$이다. 표본은 n개의 단위들로 구성되며 $1 \le n \le N$이다.

고정된 표본 크기(n)을 갖는 표본설계 d는 (S_d, P_d)로 표시한다. S_d는 각각 n개 단위들로 구성되는 모든 가능한 표본(s)들의 집합으로서 $s \in S_d$이며 P_d는 S_d에 관한 확률분포로서 모든 $s \in S_d$에 관하여 $P_d(s) > 0$을 만족한다.

d에 관한 일차와 이차 포함확률 π_i, π_{ij}, $i, j \in \{1, \cdots, N\}$은 각각 다음과 같이 정의된다.

$$\pi_i = \sum_{s \in S_d,\ s \ni i} P_d(s), \quad \pi_{ij} = \sum_{s \in S_d,\ s \ni i,j} P_d(s) \tag{2.1}$$

이 포함확률들은 Hedayat와 Sinha(1991)에서 다루어진 바와 같이 Horvitz-Thompson(1952) 추정량과 분산식을 표현하는데 있어 중요한 역할을 한다. 그리고 모니터링 표본설계에서 가장 관심을 갖게 되는 표본설계는 모든 i에 대하여 $\pi_i = n/N$이면서 π_{ij}는 단위 i와 j의 거리측도에 관한 비감소함수를 갖게 하는 방법이다. 이러한 이차포함확률을 갖는 표본추출계획을 세움으로서 고활동수준을 갖는 단위들이 소수의 집락 내에 서로 이웃하여 존재하거나 널리 산

재하여 있는 경우에 전체 조사 지역에 대하여 대표성 있는 표본을 추출할 수 있을 뿐만 아니라 Horvitz-Thompson 추정량의 분산을 줄일 수 있다. 다음은 이를 위해 제시된 기존의 방법들에 관한 설명이다.

N개의 단위들에 $1 \sim N$의 일련번호를 붙이거나 (l, m), $l = 1, \cdots, R$, $m = 1, \cdots, C$의 라벨을 부여하자. 이차원 라벨을 사용하는 경우 표본 s는 $((l_1, m_1), \cdots, (l_n, m_n))$이 되며 첫 번째 좌표 $l_1, \cdots, l_n$은 $\{1, \cdots, R\}$로 부터의 표본이고 $m_1, \cdots, m_n$은 $\{1, \cdots, C\}$로부터의 표본이다. 그러므로 $\{1, \cdots, R\}$과 $\{1, \cdots, C\}$로부터 각각 n개의 표본을 추출하는 표본설계 $d_R = (S_R, P_R)$과 $d_C = (S_C, P_C)$를 얻을 수 있다. 그리고 가능한 한 표본설계들 중 $\pi^R_{l_1} = n/R$ ($\pi^C_{l_1} = n/C$)과 l_1과 l_2간의 거리에 관한 비감소함수 $\pi^R_{l_1 l_2}$ ($\pi^C_{l_1 l_2}$)를 갖는 표본설계 d_R (d_C)을 고려한다. 단, $l_1, l_2 \in \{1, \cdots, R\}$, $l_1 \neq l_2$이다. 또한 앞으로 다루어지는 모든 표본설계는 단위 1과 R, 단위 1과 C는 각각 이웃하는 단위로 간주한다.

Hedayat et al.(1988)과 Stufken(1993)은 다음과 같은 포함확률을 갖는 표본설계 d_R을 제안하였다.

$$\pi^{\alpha}_{l_1} = n/R, \tag{2.2}$$

$$\pi^{\alpha}_{l_1 l_2} = \begin{cases} 0 & , l_1 - l_2 \equiv \pm 1, \cdots, \pm\alpha \,(\mathrm{mod}\, R) \\ \dfrac{n(n-1)}{R(R-2\alpha-1)} & , \text{ 그밖에} \end{cases} \tag{2.3}$$

여기서 α는 양의 정수이며 표본설계 d_C에 관해서도 동일하다.

표본설계는 d^{α}, $\alpha = 0, 1, \cdots$로 표시하며 $\alpha = 0$인 경우 위의 포함확률들은 단순임의비복원추출에서의 포함확률과 일치한다. 이 표본설계는 단지 2개의 이차포함확률만을 가질 수 있으므로 서로 다른 다양한 포함확률을 갖도록 하기 위하여 다음과 같은 표본설계도 고려할 수 있다.

M은 양의 정수이고 양수 $\beta_0, \cdots, \beta_M$은 $\sum_{i=0}^{M} \beta_i = 1$을 만족하는 양수들이라 하자. 그러면 β_{α}, $\alpha \in \{0, 1, \cdots, M\}$의 확률로 크기 n인 표본을 추출하는 표본설계 d_{α}을 세울 수 있으며 이때 포함확률은 다음과 같이 표현된다.

$$\pi_{l_1} = n/R, \tag{2.4}$$

$$\pi_{l_1 l_2} = \sum_{\alpha=0}^{M} \beta_{\alpha}\, n(n-1)\, \delta_{\alpha}(l_1, l_2)/(R(R-2\alpha-1)) \tag{2.5}$$

여기서

$$\delta_{\alpha}(l_1\ l_2) = \begin{cases} 0 & , l_1 - l_2 \equiv \pm 1, \cdots, \pm\alpha \,(\mathrm{mod}\, R) \\ 1 & , \text{ 그밖에} \end{cases}$$

이다.

한편 Hedayat et al.(1998)은 Fuller(1970)가 제안한 임의층경계(random stratum boundaries)를 갖는 표본추출방법을 모니터링 표본설계에 적용하였다. Fuller의 방법은 $R = nu$을 가정하며 여기서 u는 양의 정수이다. R개의 단위들을 크기가 u인 n개의 층으로 나누기 위하여 우선 $r(\in \{1, \cdots, u\})$을 임의로 선택한다. 이때 i번째 층은 다음과 같다,

$$S_i = \{ r + (i-1)u + j : j = 1, \cdots, u \} \tag{2.6}$$

이 n개의 층으로부터 각각 1개의 단위를 임의추출 한다. 그러면 다음과 같은 포함확률을 얻는다.

$$\pi_{l_1} = n/R, \tag{2.7}$$

$$\pi_{l_1 l_2} = \begin{cases} \alpha/u^3 & , l_1 - l_2 \equiv \pm 1, \cdots, \pm\alpha \,(\mathrm{mod}\, R),\ 0 < \alpha < u \\ 1/u^2 & , \text{그밖에} \end{cases} \tag{2.8}$$

여기서 $l_1, l_2 \in \{1, \cdots, R\}$, $l_1 \neq l_2$이다.

그리고 Hedayat et al.(1998)은 이차포함확률을 지정하는데 있어서 단위 (l_1, m_1)와 (l_2, m_2)간의 거리에 관한 구체적인 함수, 즉 $D((l_1, m_1), (l_2, m_2))$을 이용하였고 $\alpha_1 + \alpha_2$, $\max(\alpha_1, \alpha_2)$ 등을 사용 가능한 거리함수의 예로 제시하였다. 여기서 $l_1 - l_2 \equiv \pm\alpha_1 \,(\mathrm{mod}\, R)$, $m_1 - m_2 \equiv \pm\alpha_2 \,(\mathrm{mod}\, C)$이다. 그리고 다음을 만족하는 포함확률을 갖는 표본설계를 구하였다.

$$\pi_{(l_1, m_1)} = n/N, \tag{2.9}$$

$$\pi_{(l_1, m_1)(l_2, m_2)} \propto \alpha_1 + \alpha_2 \tag{2.10}$$

그런데 식(2.9)와 식(2.10)을 만족하는 표본설계를 얻기 위해서 고정점 (l_0, m_0)로부터 거리가 $x = \alpha_1 + \alpha_2$만큼 떨어져있는 점들의 개수 n_x에 관한 식들을 구할 수 있다. Hedayat et al.은 R과 C의 홀수 또는 짝수 여부 그리고 R과 C의 대소 관계에 의해 5가지 경우로 구분하여 이 n_x의 값들을 구했다.

그리고 $\pi_{(l_1, m_1)(l_2, m_2)} \propto \alpha_1 + \alpha_2$을 만족시키는 이차포함확률을 얻기 위하여 다음 식을 세운다.

$$\begin{aligned} K &= \sum_{(l_2, m_2) \neq (l_1, m_1)} \pi_{(l_1, m_1)(l_2, m_2)} \Big/ \sum_x x\, n_x \\ &= n(n-1)/RC \sum_x x\, n_x \end{aligned} \tag{2.11}$$

예를 들어 $R = C$이고 홀수인 경우 이차포함확률은 다음과 같다.

$$\pi_{(l_1, m_1)(l_2, m_2)} = K(\alpha_1 + \alpha_2) = 2n(n-1)(\alpha_1 + \alpha_2)/R^3(R-1)(R+1) \tag{2.12}$$

지금까지 기존 표본설계방법들에 관하여 설명하였는데 이들 방법들에 관한 문제점을 살펴보면 다음과 같다.

Fuller(1970), Hedayat et al.(1988), Stufken(1993) 등의 방법은 Hedayat et al.(1998)의 방법과 비교할 때 그 실행 방법은 간단한 반면 단위간의 거리함수를 사용하지 않으므로 실제 조사를 하는데 있어 단위간의 거리가 증가함에 따라 이차포함확률이 달라지는 것을 구체화하기 어려운 단점이 있다. 그리고 Hedayat et al.(1998)이 제시한 방법은 실행이 다른 방법에 비해 다소 복잡하기는 하지만 모니터링 표본 설계를 위한 적절한 방법으로서 유용하게 사용될 수 있다. 그러나 이 방법은 일차포함확률 n/N을 유지하면서 이차포함확률을 만족시킬 수 있는 표본추출계획이 실제로 존재하지 않을 수 있다는 문제가 있다.

예를 들어 $n = R = C = 3$인 간단한 경우에도 이러한 문제를 쉽게 확인할 수 있다. 이는 $R = C$이고 홀수인 경우로서 이차포함확률은 식 (2.12)로부터 얻을 수 있다. 그리고 $\pi_{(l_1,m_1)(l_2,m_2)} = \sum_{s\in S_d,\ s\ni (l_1,m_1)(l_2,m_2)} P_d(s)$ 이므로 84개의 모든 가능한 표본 s의 추출확률 $P_d(s)$을 구할 수 있다. 이때 각 표본의 추출확률을 살펴보면 전체 84개의 추출확률 중 6개가 $P_d(s) < 0$이라는 것을 알 수 있다. 이는 표본추출계획을 세울 수 없다는 것을 의미한다.

3. 제안 방법

앞의 예를 통하여 알 수 있는 바와 같이 Hedayat et al.(1998)의 방법은 비교적 간단한 경우에도 표본설계가 존재하지 않을 수 있다. 그런데 이러한 문제 외에 한 가지 고려할 수 있는 것은 현재의 $\pi_{(l_1,m_1)(l_2,m_2)}$는 $\pi_{(l_1,m_1)(l_2,m_2)} \propto \alpha_1 + \alpha_2$을 만족시키는 이차포함확률이기는 하나 각 표본 $s = ((l_1, m_1), \cdots, (l_n, m_n))$에 관한 거리함수를 표현하고 있는 것은 아니라는 점이다. 그러므로 각 표본 s 내에 있는 단위 상호간의 전체적인 거리함수로서 다음 식을 생각해볼 수 있다.

$$D_s(l,m) = \sum_{i<j}^{n} (\alpha_{l_{ij}} + \alpha_{m_{ij}}) \tag{3.1}$$

여기서 $l_i - l_j = \pm \alpha_{l_{ij}} (\bmod R)$, $m_i - m_j = \pm \alpha_{m_{ij}} (\bmod C)$이다.

또한 이러한 각 표본의 거리함수를 표본설계에 반영하도록 하기 위하여 선형계획법(linear programming)을 이용할 수 있는 다음과 같은 목적함수 ϕ와 제약조건들을 세울 수 있다.

$$\phi = \sum_{s\in S_d} D_s(l,m) P_d(s) \tag{3.2}$$

i) $\pi_{(l_i,m_i)(l_j,m_j)} = 2n(n-1)(\alpha_1 + \alpha_2)/R^3(R-1)(R+1)$

ii) $P_d(s) \geq 0$

결과적으로 식 (3.2)을 최대화시키는 해를 얻을 수 있는데 이들 해는 제약조건 i), ii)를 만족시키면서 각 표본 s에 관한 거리함수를 최대한 반영시킬 수 있는 표본추출계획이다.

4. 결론 및 차후 연구 방향

본 논문에서는 기존 방법들의 문제점을 설명하였고 Hedayat et al.(1998)이 제안한 이차포함확률을 활용하면서 그 문제점들을 개선시킬 수 있는 표본설계방법을 제시하였다. 그러나 Hedayat et al.(1998)이 제시한 방법과 본 논문에서 제안한 개선 방법을 사용하는데 있어서 아직 해결해야 할 문제점이 남아 있다. 이는 Hedayat et al.(1998)가 제시한 표본설계방법은 $\pi_{(l_1,m_1)(l_2,m_2)} \geq 0$은 만족시키지만 Horvitz-Thompson 추정량의 분산 추정량의 값이 비음이 되도록 하기 위한 조건인 $\pi_{(l_1,m_1)}\pi_{(l_2,m_2)} \geq \pi_{(l_1,m_1)(l_2,m_2)}$을 항상 만족시키지 못하기 때문이다. 따라서 모니터링을 위한 표본설계라 할지라도 특정 모집단에만 사용되어야하는 제약이 따르므로 이를 보완할 수 있는 새로운 방법의 연구의 필요하다.

아울러 Hedayat et al.(1988), Stufken(1993), Hedayat et al.(1998)의 방법에 대해 관리적 표본추출법(controlled sampling)의 활용 방법을 모색해보는 것도 적절하리라 생각된다. 환경 또는 생태 연구를 하는데 있어 기존에 제안된 다양한 관리적 표본추출법을 이용함으로서 표본설계자가 조사의 실용적인 측면, 즉 조사 비용 또는 표본의 대표성 등을 감안하여 표본설계를 할 수 있기 때문이다.

참고문헌

[1] Cormack, R. M. (1988). Statistical challenges in the environmental sciences: A personal view. *Journal of the Royal Statistical Society*, A, Vol. 151, 201-210.

[2] Fuller, W. A. (1970). Sampling with random stratum boundaries. *Journal of the Royal Statistical Society*, B, Vol. 32, 209-226.

[3] Hedayat, A., Rao, C. R., Stufken, J. (1988). Sampling designs excluding contiguous units. *Journal of Statistical Planning and Inference*, Vol. 19, 159-170..

[4] Hedayat, A., and Sinha, B. K.. (1991). *Design and Inference in Finite Population Sampling*. Wiley, New York.

[5] Hedayat, A., and Stufken, J. (1998). Sampling designs to control selection probabilities of contiguous units. *Journal of Statistical Planning and Inference*, Vol. 72, 333-345

[6] Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe, *Journal of the American Statistical Association*, Vol. 47, 663-685.

[7] Millard, S. P. (1987). Environmental monitoring, statistics, and the law : Room for improvement. *American Statistician*, Vol. 41, 249-253.

[8] Stufken, J. (1993). Combinatorial and statistical aspects of sampling designs to avoid the selection of adjacent units, *Journal of Combinatorial Information and System Sciences*, Vol. 18, 81-92.

[9] Thompson, S. K. (1990). Adaptive cluster sampling, *Journal of the American Statistical Association*, Vol. 85, 1050-1059.

[10] Thompson, S. K. (1991). Stratified adaptive cluster sampling, *Biometrika* 78, 389-397.

[11] SAS/OR User's Guide, *SAS Institute Inc.* (1989).