응용통계연구 (2009)
22(4), 677–686

k-공간중위수 군집방법을 활용한 층화방법

손순철[1] · 전명식[2]

[1]고려대학교 통계학과, [2]고려대학교 통계학과

(2009년 2월 접수, 2009년 4월 채택)

요 약

표본조사에서 널리 쓰이는 모집단의 층화는 추정의 효율을 높이는 방법 중의 하나지만, 이상점을 포함하는 변수가 있는 경우에 여러 가지 문제점을 유발시킬 수 있다. 특히, 이상점이 존재하는 다변량 자료의 경우, 층화를 위한 k-평균 군집방법은 이상점에 매우 민감하여 추정의 효율을 떨어뜨릴 수 있다. 본 연구에서는 이상점이 존재하는 다변량 자료의 층화를 위해 k-평균 군집방법보다 강건하며 이상점을 따로 식별하는 과정이 배재된 k-공간중위수 군집방법을 제안한다. 기존 관련연구인 박진우와 윤석훈 (2008)과 동일한 자료에 대한 사례분석을 통해 층화과정들을 비교, 검토하였으며 이들의 효율성을 추정량의 분산을 통해 비교하였다.

주요용어: 네이만 배분, 다변량 층화, 이상점, k-공간중위수 군집방법, k-평균 군집방법.

1. 서론

표본설계에서 모집단에 대한 조사 및 추론의 효율을 높이기 위해 일반적으로 많이 쓰이는 방법 중의 하나로 층화(stratification)를 들 수 있다. 층화를 이용한 표본추출 방법은 층(stratum) 내에서는 동질적이고 층 간에는 이질적인 것이 바람직하다. 관심모수가 여러 개인 다변량 조사를 위한 층화방법으로, Golder와 Yeomans (1973), Jarque (1981)는 최적분리 군집방법인 k-평균 군집(k-means clustering)방법이나 주성분분석(principal component analysis) 등의 다변량기법을 활용한 층화방법을 제안하였다. 그러나, 표본조사에서 왜도가 심한 관심변수들을 많은 경우에 볼 수 있으며, 왜도가 심한 관심변수들은 이상점(outlier)들을 포함할 수가 있다. 왜도가 심한 단변량 모집단 층화와 관련하여 Lavallée와 Hidiroglou (1988)는 최적화에 필요한 효율적인 알고리즘을 제안하였다. 하지만, 다변량 층화에 대한 기존 연구들에서는 대부분 이상점이 존재하는 상황은 고려되지 않았다.

이상점이 존재하는 다변량 자료의 경우, 층화를 위한 k-평균 군집방법은 이상점에 매우 민감하게 반응하기 때문에 층화의 효율에 직접적인 영향을 주게 된다. 따라서, k-평균 군집방법을 사용하여 층화를 하고자 할 때에는 이상점에 대한 대처가 필요하다. 박진우와 윤석훈 (2008)은 이상점이 존재하는 다변량 조사를 위한 층화방법으로 이상점을 사전에 식별하여 식별된 이상점을 제외한 나머지 개체들로써 k-평균 군집방법을 사용하여 층화를 한 뒤, 이상점을 적절한 층에 재배치하는 방법을 사용하였다. 하지만 이러한 방법들은 이상점을 사전에 식별해야 하므로 이상점에 대한 명확한 정의가 없는 상황에서 연구자의 주관이 개입될 여지가 있다.

이 논문은 2007년 정부(교육인적자원부)의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임 (KRF-2007-314-C00039).

[2]교신저자: (136-701) 서울시 성북구 안암동 5-1, 고려대학교 통계학과 교수. E-mail: jhun@korea.ac.kr

본 연구에서는 이상점이 존재하는 다변량 조사의 층화방법으로 k-공간중위수 군집(k-spatial medians clustering)방법을 제안하고, 이 방법이 k-평균 군집방법보다 강건하며 이상점을 사전에 따로 식별할 필요가 없는 즉, 주관이 배제된 방법임을 보이고자 한다. 이를 위해 x_{hij}를 h층의 i번째 개체의 j번째 변수에 대한 값 ($h = 1, \ldots, L$, $i = 1, \ldots, N$, $j = 1, \ldots, p$), N을 모집단의 크기, n을 표본크기, L을 층의 수, N_h를 h층의 크기, n_h를 h층의 표본크기라 하자. 나아가 $\bar{x}_{hj}$를 h층의 j번째 변수의 표본평균, $\sigma_{hj}^2 = 1/(N_h - 1)\sum_{i=1}^{N_h}(x_{hij} - \bar{x}_{hj})^2$을 h층의 j번째 변수에 대한 분산 그리고 $W_h = N_h/N$이라 하면, j번째 변수의 모평균 θ_j에 대한 추정량은

$$\hat{\theta}_j = \frac{1}{N}(N_1\bar{x}_{1j} + \cdots + N_L\bar{x}_{Lj}) = \sum_{h=1}^{L} W_h\bar{x}_{hj}$$

이고, 그의 분산은

$$V(\hat{\theta}_j) = \frac{1}{N^2}\left[N_1^2V(\bar{x}_{1j}) + \cdots + N_L^2V(\bar{x}_{Lj})\right] = \sum_{h=1}^{L} W_h^2V(\bar{x}_{hj}), \quad \text{단 } V(\bar{x}_{hj}) = \frac{N_h - n_h}{N_h n_h}\sigma_{hj}^2$$

으로 표현된다.

본격적인 논의에 앞서 우선 2장에서 k-공간중위수 군집방법을 k-평균 군집방법과 비교하여 설명하고, 3장에서 기존 연구와의 비교를 위해 박진우와 윤석훈 (2008)에서 모집단으로 사용했던 농촌생활지표조사 자료 (2006)를 이용하여 관심변수들에 대한 평균을 추정하고, 모평균 추정량의 분산을 고려된 층화방법들에 대해 각각 구하였다. 이를 위한 층별 표본배분은 네이만 배분을 사용하였으며, 마지막으로 4장에서 본 연구의 결과를 요약하고 정리하였다.

2. 최적분리 군집방법

최적분리 군집방법은 주어진 기준에 따라 관찰값들을 몇 개의 최적군집으로 구분(partitioning)시키는 형태이다. 많은 경우 이 방법은 연구자에 의해 군집의 개수 k가 미리 결정되어 있으며, 주어진 최적 조건을 만족하는 k개의 점을 찾는 과정과 각 개체들을 할당하는 두 개의 단계로 이루어진다.

k-평균 군집방법

p차원 다변량 표본벡터를 각각 $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$으로 나타내면, 군집내 오차제곱합을 최소화하는 k-평균 군집방법은 다음과 같이 이루어진다.

단계 1: $1/N\sum_{i=1}^{N}\min_h||\boldsymbol{x}_i - \boldsymbol{b}_h||^2$을 최소화하는 k개의 점 $\boldsymbol{b} = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_k)$을 찾는다.
(단, $1 \le h \le k$, $||\cdot||$은 유클리디안 노옴)

단계 2: 각 $\boldsymbol{x}_i$를 가장 가까운 군집중심에 할당한다.

단계 1에서는 k개 군집의 초기값을 선택하여 결정된 초기 군집에 가장 가까운 개체들을 할당하여 형성된 군집들의 평균벡터 $\boldsymbol{b}$를 구하고, 재할당 과정을 $\boldsymbol{b}$가 거의 변하지 않을 때까지 반복하여 최종적으로 k개의 군집들의 군집중심벡터를 구한다. 이러한 k-평균 군집방법은 알고리즘이 간단하고 특히 큰 자료의 개체 군집화에도 상당히 효율적인 것으로 알려져 있다 (Milligan, 1980, 1981). 하지만, 알려진 바와 같이 k-평균 군집방법은 군집의 중심을 그 군집에 속한 개체들의 평균을 이용하여 구하기 때문에 이상점에 매우 민감하다. 따라서, 이상점이 존재하는 다변량 자료에 대한 군집방법으로는 절사된(trimmed) 평균을 이용하는 방법 (Cuests-Albertos 등, 1997)이나 본 연구에서 제안하는 k-공간중위수 군집방법을 고려해 볼 수 있다.

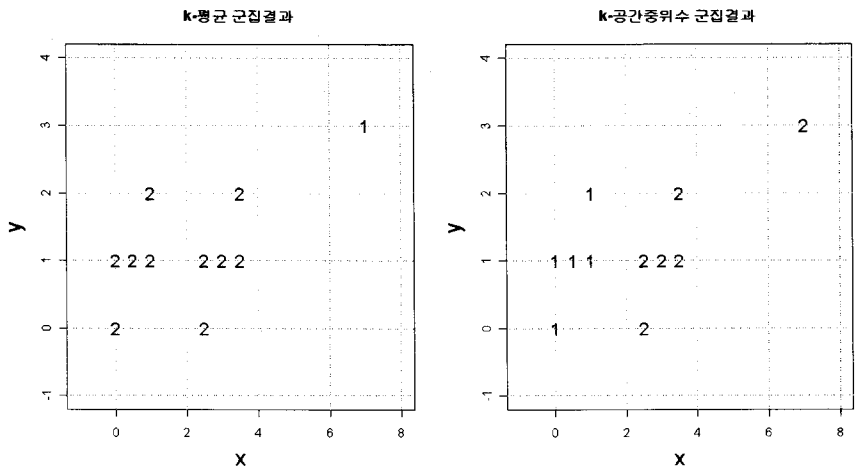

그림 **2.1.** 가상자료에 대한 군집분석

k-공간중위수 군집방법

앞서 k-평균 군집방법이 단계 1에서 군집내 오차제곱합을 최소화는 반면에 k-공간중위수 군집방법은 군집내 절대오차합 $1/N\sum_{i=1}^{N}\min_h\|\boldsymbol{x}_i-\boldsymbol{a}_h\|$ (단, $1\le h\le k$)을 최소화하는 방법을 사용하며, 이 때 구해진 군집의 중심은 군집의 공간중위수가 된다.

공간중위수는 단변량에서의 중위수 개념을 다변량에 확장한 것으로서 관측치의 절대적 크기와는 무관하게 중위수에 대한 관측치의 상대적 방향 즉, 중위수와 관측치 사이의 각도에 의해서만 결정된다 (Brown 1983). 따라서, 이와 같은 공간중위수를 활용한 k-공간중위수 군집방법은 k-평균 군집방법에 비해 이상점에 대해 강건한 성질을 갖게 된다 (Jin, 1999).

예시

이상점이 존재하는 임의의 2-차원 자료를 통해 k-평균 군집방법에 비해 k-공간중위수 군집방법이 이상점에 강건함을 보이고자 한다. 다음 그림 2.1은 11개의 자료 점과 이 점들에 대해 각각 k-평균 군집방법과 k-공간중위수 군집방법을 적용시킨 결과이다.

그림 2.1에서 개체들의 소속 군집을 '1'과 '2'는 표시하였다. 위 그림에서 알 수 있듯이 $x=2$를 기준으로 두 개의 분리된 집단이 있을 때, 이상점이라고 할 수 있는 하나의 점 (7, 3)이 추가되면 k-평균 군집방법과 k-공간중위수 군집방법은 전혀 다른 결과를 보여준다. 즉, k-공간중위수 군집방법은 이상점에 강건하기 때문에 각 개체들의 소속집단을 유지하면서 이상점을 자연스럽게 군집 '2'에 할당하는 반면, k-평균 군집방법은 이상점에 영향을 많이 받기 때문에 이상점을 제외한 모든 개체를 하나의 집단(군집 '1')에 할당하였음을 알 수 있다.

3. 농촌생활지표 조사 자료의 층화

본 연구에서는 연구목적상 박진우와 윤석훈 (2008)과 동일한 모집단을 사용하여 군집방법을 활용한 층화방법의 효율을 비교, 검토하였다.

표 **3.1.** 모집단의 자료구조

관측치	행정구역별				주택유형별(가구수)				연령대별(인구수)			
	행정코드	광역시/도	시/군/구	읍/면/동	단독	아파트	연립	기타	~10	10~19	···	60~
1	2331011	인천광역시	강화군	강화읍	3,558	778	610	1,846	2,213	2,695	···	3,874
2	2331031	인천광역시	강화군	선원면	1,022	980	70	47	870	789	···	1,343
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮
1410	3932032	제주도	남제주군	표선면	2,729	113	88	204	1,172	1,144	···	1,995

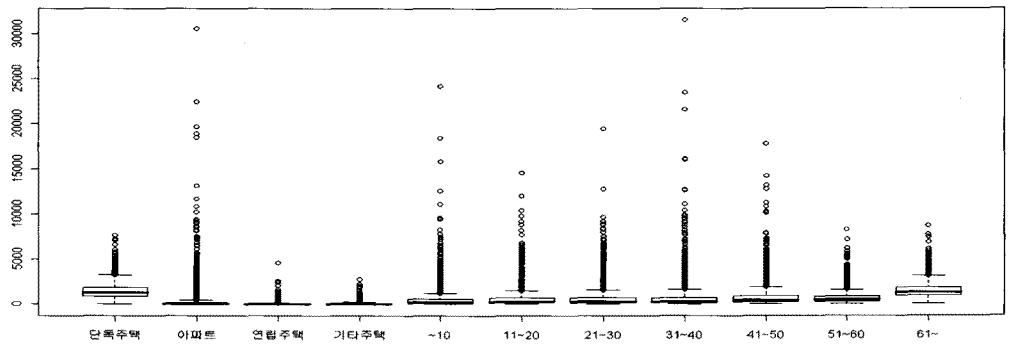

그림 **3.1.** 모집단의 관심변수 박스플롯

표 **3.2.** 모집단의 관심변수 평균 및 분산

	주택유형별(가구수)				연령대별(인구수)						
	단독	아파트	연립	기타	~10	10~19	20~29	30~39	40~49	50~59	60~
평균	1,514	537	71	105	679	714	718	885	916	74	1,515
분산	870,249	3,505,354	51,331	51,320	2,212,474	1,467,648	1,691,758	3,798,931	1,997,022	564,854	870,384

3.1. 모집단

2005년 인구주택총조사에서 농촌지역으로 분류된 1,410개의 읍/면/동에 대해서 주택유형별(단독주택, 아파트, 연립주택, 기타주택) 가구수와 연령대별(10세미만, 10대, 20대, 30대, 40대, 50대, 60세이상) 인구수 등을 조사한 자료의 구조는 다음 표 3.1과 같으며, 본 연구에서는 주택유형별 가구수와 연령대별 인구수를 관심변수로 하여 이들의 평균을 도별로 추정하고 추정량의 분산을 구하고자 한다.

모집단의 관심변수들은 그림 3.1의 박스플롯에서 볼 수 있듯이 왜도가 심한 편이기 때문에 이상점이 존재하는 다변량 자료의 형태를 보이며, 이러한 경향은 9개의 도 모두에서 유사하다.

표 3.2는 모집단의 관심변수들에 대한 평균 및 분산을 나타낸 것이다.

3.2. 군집방법별 층화효과 비교

모집단의 관심변수들의 평균에 대한 추정을 위해 층화임의추출방법을 이용하고자 한다. 농촌생활지표조사 자료는 도별 지역통계를 작성해야 하므로 도별로 따로 층화를 하였으며, 도별 층의 개수는 같은 자료에 대한 기존 연구인 박진우와 윤석훈 (2008)의 도별 층의 개수와 동일하게 택하였다. 모집단에서 도

표 3.3. 도별 층별 개체수(괄호 안은 해당 도에서의 군집방법별 군집(층)번호)

도	읍/면/동 수		k-공간중위수 군집방법	k-평균 군집방법	박진우·윤석훈
경기	4	160	68(1), 56(2), 24(3), 12(4)	120(1), 28(2), 11(3), 1(4)	95(1), 38(2), 15(3), 12(4)
강원	3	114	48(1), 44(2), 22(3)	90(1), 19(2), 5(3)	74(1), 28(2), 12(3)
충북	3	103	70(1), 25(2), 8(3)	77(1), 20(2), 6(3)	67(1), 21(2), 15(3)
충남	4	170	92(1), 38(2), 26(3), 14(4)	120(1), 33(2), 10(3), 7(4)	94(1), 43(2), 17(3), 16(4)
전북	4	159	72(1), 48(2), 35(3), 4(4)	87(1), 54(2), 14(3), 4(4)	58(1), 50(2), 36(3), 15(4)
전남	5	229	88(1), 73(2), 43(3), 19(4), 6(5)	126(1), 78(2), 19(3), 4(4), 2(5)	68(1), 63(2), 54(3), 23(4), 21(5)
경북	6	247	69(1), 67(2), 44(3), 33(4), 21(5), 13(6)	158(1), 56(2), 20(3), 7(4), 5(5), 1(6)	83(1), 73(2), 36(3), 24(4), 18(5), 13(6)
경남	4	216	125(1), 68(2), 19(3), 4(4)	165(1). 38(2), 9(3), 4(4)	126(1), 57(2), 18(3), 14(4)
제주	1	12	12(1)	12(1)	12(1)
계	34	1410	1410	1410	1410

※ 동일한 '도'내에서 밑줄 친 군집은 같은 읍/면/동으로 구성됨.

별 층의 개수를 k로 하여 k-평균 군집방법과 k-공간중위수 군집방법을 각각 수행한 결과 다음 표 3.3과 같이 도별로 층별 개체수를 얻을 수 있었다. 하나의 군집은 하나의 층과 같은 개념이기 때문에, 각 군집의 크기는 층의 크기 N_h와 같으며, 군집들의 갯수의 총합은 층의 수 L과 같게 된다. 즉, 표 3.3에서 $N_1 = 68$, $N_2 = 56, \ldots$, $N_{34} = 12$, $L = 34$가 된다.

표 3.3을 보면 k-평균 군집방법의 결과 소수의 개체가 하나의 군집을 형성한 경우, 해당 군집에 속하는 개체(지역)들은 이상점으로 예상할 수 있으며, 이 지역들을 구체적으로 살펴보면 표 3.4와 같다. 이상점으로 예상할 수 있는 표 3.4에 있는 지역들의 변수들은 앞의 표 3.2와 비교하면 매우 큰 값들을 갖고 있으며, 군집 내 개체수가 작을수록 더욱 큰 값들을 갖고 있기 때문에 이상점으로 판단할 수 있다. 즉, k-평균 군집방법은 이상점에 민감하기 때문에 멀리 떨어진 소수의 개체가 하나의 군집을 형성하였음을 알 수 있다.

반면에 k-공간중위수 군집방법은 2장에서 살펴본 바와 같이 k-평균 군집방법보다 층 내의 동질성 확보 측면에서 이상점에 강건한 성질이 있으므로 이상점들이라 여겨지는 지역들을 가장 가까운 군집에 포함시킬 것이라 예상할 수 있다. 실제 표 3.3을 살펴보면, k-평균 군집방법은 경기도에서 안중읍을 포함한 11개 지역과 태안읍(1개 지역)을 각각 군집3과 군집4로 나누었지만, k-공간중위수 군집방법에서는 이들 12개 지역이 하나의 군집4를 형성하였다. 전라남도에서도 k-평균 군집방법은 해룡면을 포함한 4개 지역과 광양읍을 포함한 2개 지역을 각각 군집4와 군집5 두 개의 군집으로 나누었지만, k-공간중위수 군집방법은 이들을 합친 6개 지역을 하나의 군집5로 만들었다. 마찬가지로, 경상북도에서도 k-평균 군집방법에 의해서는 군집4, 군집5, 군집6으로 나뉘었던 지역들이 k-공간중위수 군집방법에서는 하나의 군집6을 이루었다.

박진우와 윤석훈 (2008)은 이상점을 미리 탐색적 방법 등을 통하여 식별한 뒤 이상점이라고 판단된 점들을 제외하고 나머지 개체들로 k-평균 군집방법을 수행하고, 제외시켰던 이상점들을 다시 가장 속성이 유사한 군집에 할당하는 방법을 사용하였기 때문에 표 3.3에서 볼 수 있듯이 군집결과가 k-평균 군집방법의 결과와 매우 다르다. 하지만, k-공간중위수 군집방법의 결과와 비교했을 때, 이상점을 의심할 수

표 3.4. 이상점으로 예상할 수 있는 지역

도	군집(층) 번호			읍	주택유형별(가구수)				연령대별(인구수)						
	ksm	박·윤	km		단독	아파트	연립	기타	~10	10~19	20~29	30~39	40~49	50~59	60~
경기	4	4	4	태안	4,309	30,580	2,421	684	24,182	14,591	19,478	31,536	17,770	8,228	7,706
전북	4	4	4	삼례	3,737	2,436	37	38	2,155	2,321	3,850	2,764	2,119	1,756	3,184
				봉동	2,934	1,780	98	73	2,329	1,714	2,324	2,754	1,781	1,469	2,586
				고창	3,595	1,980	488	235	2,706	3,281	1,666	2,847	3,022	2,087	3,126
				부안	3,823	2,503	101	312	2,679	2,736	1,783	2,899	3,016	2,249	3,711
전남	5	5	4	해룡면	2,889	4,424	0	113	4,441	3,596	2,415	5,250	3,461	1,481	2,913
				해남	4,578	2,981	324	569	3,845	3,588	2,392	4,340	4,115	2,642	3,782
				삼호	2,101	4,637	15	115	4,136	2,410	3,419	5,204	2,630	1,338	1,744
				영광	4,051	2,057	419	700	3,055	3,228	2,247	3,467	3,372	2,469	4,009
			5	광양	5,288	8,527	428	352	6,329	6,794	6,418	7,902	7,006	4,069	5,193
				화순	4,649	9,026	105	309	7,757	5,494	5,677	9,165	5,642	3,323	5,525
경북	6	6	5	안강	5,325	5,566	77	262	4,400	4,629	3,472	5,818	5,261	3,454	5,894
				오천	4,230	7,285	519	1,287	6,473	5,225	5,749	8,489	6,927	3,886	3,733
				흥해	4,696	6,446	418	465	4,637	5,501	6,158	6,318	6,012	3,616	5,432
				하양	5,713	4,605	56	101	3,411	5,679	7,038	4,565	4,324	2,968	4,151
				진량	4,769	9,093	87	286	6,500	5,805	8,321	9,073	5,085	2,565	3,874
			6	화원	4,124	10,800	275	709	8,254	8,179	6,678	11,123	10,131	4,971	4,568
경남	4	4	4	내서	2,336	19,672	157	172	12,555	11,994	9,125	16,098	13,203	6,129	4,903
				장유면	2,460	22,460	17	219	18,407	9,130	8,797	23,508	10,814	5,139	4,901
				신현	5,570	18,504	1,034	1,577	15,812	10,425	12,756	21,638	14,210	5,509	3,277
				웅상	4,541	18,892	252	870	11,115	9,775	9,632	16,068	12,753	7,163	8,649

※ ksm: k-공간중위수 군집방법, 박·윤: 박진우·윤석훈, km: k-평균 군집방법

있는 지역들은 표 3.4의 군집번호를 통해 알 수 있듯이 유사한 군집 결과를 보였고, 표 3.3의 밑줄 친 경기도의 12개 지역과 경상북도의 13개 지역은 k-공간중위수 군집방법과 동일하게 하나의 군집을 형성하였다. 이와 같이 k-공간중위수 군집방법은 이상점을 사전에 식별하는 등의 별도의 과정을 거칠 필요가 없으므로, 이상점에 대한 명확한 정의가 없는 상황에서 연구자의 주관이 개입될 여지를 배제할 수 있다.

표 3.5는 경기도 모집단의 읍/면/동에 대하여 군집방법별로 각각 형성된 군집들의 평균 프로화일을 나타내었다. k-평균 군집방법과 k-공간중위수 군집방법 모두 군집간의 뚜렷한 차이를 보이고 있어 각 군집을 하나의 층으로 택할 수 있으나, k-평균 군집방법의 경우 군집4가 하나의 개체로 구성되어 있는 점이 결점으로 생각된다.

각각의 군집방법별로 얻어진 군집들을 층으로 하여 층화효과를 추정량의 분산을 통해 알아보고자 한다. 층별 표본배분은 네이만 배분(Neyman allocation) 방법을 따랐고, 표본크기는 $n = 200$으로 하였다.

변수별 네이만 배분을 이용한 표본 배분

일변량 자료에 대한 층화추출의 경우 네이만의 최적배분방법을 이용하면, h층의 j번째 변수에 대한 배분값은

$$n_{hj} = \frac{N_h \sigma_{hj}^2}{\sum_{h=1}^{L} N_h \sigma_{hj}^2} \times n, \quad h = 1, \ldots, L, \; j = 1, \ldots, p$$

표 **3.5.** 경기도의 군집방법별, 변수별 평균 프로화일

변수		k-공간중위수 군집방법				k-평균 군집방법			
		군집1 (68)	군집2 (56)	군집3 (24)	군집4 (12)	군집1 (120)	군집2 (28)	군집3 (11)	군집4 (1)
주택유형별 (가구수)	단독주택	1,058	1,982	2,820	2,909	1,446	2,753	2,782	4,309
	아파트	26	473	2,738	10,437	184	2,566	8,606	30,580
	연립주택	20	173	660	1,271	85	595	1,167	2,421
	기타주택	69	299	757	950	167	697	975	684
연령대별 (가구수)	~10	263	952	2,939	8,605	536	2,766	7,189	24,182
	10~20	341	1,124	2,829	6,545	650	2,714	5,814	14,591
	20~30	338	1,210	3,174	6,924	691	2,999	5,782	19,478
	30~40	357	1,324	3,985	11,259	743	3,744	9,415	31,536
	40~50	519	1,483	3,565	8,042	911	3,377	7,158	17,770
	50~60	469	1,069	2,134	4,072	719	2,024	3,694	8,228
	60~	999	1,855	3,034	4,971	1,361	2,903	4,722	7,706

※ 괄호() 안은 각 군집에 속하는 읍/면/동의 수

표 **3.6.** 네이만 배분을 이용한 층별 표본배분

도	k-공간중위수 군집방법	k-평균 군집방법
경기	8(68), 9(56), 9(24), 15(12) → 9(68), 10(56), 10(24), 12(12)	25(120), 10(28), 9(11), 제외(1)
강원	2(48), 10(44), 4(22)	10(90), 4(19), 2(5)
충북	6(70), 4(25), 4(8)	7(77), 4(20), 2(6)
충남	7(92), 5(38), 7(26), 9(14)	14(120), 8(33), 3(10), 2(7)
전북	3(72), 2(48), 4(35), 1(4)	4(87), 3(54), 1(14), 1(4)
전남	5(88), 3(73), 5(43), 5(19), 4(6)	8(126), 8(78), 5(19), 2(4), 1(2)
경북	3(69), 2(67), 5(44), 1(33), 5(21), 10(13)	11(158), 6(56), 4(20), 4(7), 2(5), 제외(1)
경남	10(125), 12(68), 13(19), 2(4)	20(165), 8(38), 6(9), 2(4)
제주	5(12)	5(12)

※ 괄호() 안은 모집단 층의 크기, →은 밑줄 친 군집에서 발생한 잉여표본 재배분 결과

로 구할 수 있다. 그러나 단독주택 가구수와 60세 이상 인구수 변수를 제외한 모든 변수들에서 앞 절에서 언급한 태안읍의 분산이 매우 커서 이 지역이 속한 층에서는 네이만 배분 결과 모집단의 크기보다 표본크기가 더 커지는 문제점이 발생하였다. 특히, k-평균 군집방법에서는 태안읍과 화원읍이 각각 하나의 층을 이루기 때문에 층별 분산을 알 수 없어 네이만 배분법을 적용할 수가 없었다.

따라서, 본 연구에서는 k-평균 군집방법 결과 생성된 층에서 태안읍과 화원읍으로 각각 구성된 층을 제외하고, 나머지 32개의 층에 표본배분을 하였고, 각각 방법별로 잉여 표본($N_h < n_h$인 경우)이 발생하면 동일 도내의 다른 층에 표본크기에 비례하게 재배분하였다. 예를 들어, 50~60세 인구수 변수에 대한 네이만 배분 결과 및 잉여표본에 대한 수정 결과는 표 3.6과 같다.

이제 방법별 층화효과를 비교하고자 9개 도에 대하여 각 변수별로 모평균 추정량 $\hat{\theta}_j$의 분산 $V(\hat{\theta}_j)$을 구하였다. 표 3.7을 보면, 우선 단순임의추출에 비해 층화임의추출에 의한 분산이 매우 작은 것을 통해 층화임의추출의 효율이 단순임의추출보다 좋다는 것을 알 수 있다. 또한 층화방법별 분산을 비교하여 보면, k-공간중위수 군집방법은 k-평균 군집방법에 비해 모든 변수에서 더 작은 분산을 갖고 있기 때문에 모집단에 대한 추정의 효율이 좋을 것으로 판단된다. 특히, k-평균 군집방법을 이용한 층화에서는 1개 지역이 하나의 군집을 이루었던 경기도 태안읍과 경상북도 화원읍이 모집단에서 빠진 상태이기 때문에,

표 3.7. 관심변수별 모평균 추정량에 대한 분산

변수		모평균 추정량에 대한 분산		
		단순임의추출	층화임의추출	
			k-공간중위수 군집방법	k-평균 군집방법
주택유형별 (가구수)	단독주택	3734.05	584.89	848.24
	아파트	15040.70	222.56	226.01
	연립주택	220.25	13.68	15.98
	기타주택	220.20	22.19	26.03
연령대별 (인구수)	~10	9493.24	188.14	223.19
	10~20	6297.36	237.12	253.42
	20~30	7258.96	329.73	346.98
	30~40	16300.38	310.61	378.81
	40~50	8568.78	282.56	347.92
	50~60	2423.66	163.15	224.02
	60~	3734.63	560.03	813.04

표 3.8. 네이만 배분의 평균을 이용한 층별 표본배분

도	k-공간중위수 군집방법	k-평균 군집방법
경기	6(68), 12(56), 13(24), 19(12) → 7(68), 15(56), 16(24), 12(12)	26(120), 15(28), 10(11), 제외(1)
강원	2(48), 9(44), 4(22)	9(90), 4(19), 2(5)
충북	5(70), 5(25), 3(8)	6(77), 5(20), 2(6)
충남	6(92), 5(38), 8(26), 8(14)	11(120), 8(33), 5(10), 3(7)
전북	2(72), 2(48), 4(35), 1(4)	3(87), 4(54), 2(14), 1(4)
전남	5(88), 2(73), 3(43), 5(19), 4(6)	6(126), 7(78), 4(19), 1(4), 1(2)
경북	3(69), 2(67), 6(44), 1(33), 5(21), 9(13)	8(158), 7(56), 4(20), 3(7), 3(5), 제외(1)
경남	7(125), 12(68), 13(19), 4(4)	15(165), 10(38), 7(9), 4(4)
제주	5(12)	5(12)

※ 괄호() 안은 모집단 층의 크기, →은 밑줄 친 군집에서 발생한 잉여표본 재배분 결과

변수별로 태안읍과 화원읍에서 다른 지역의 값들과 크게 다르지 않을수록 k-공간중위수 군집방법보다 효율이 떨어진다고 할 수 있다. 즉, k-공간중위수 군집방법은 k-평균 군집방법에는 포함되지 않은 태안읍과 화원읍(이상점들)을 포함시키고도 항상 k-평균 군집방법보다 좋은 효율을 나타냈다고 할 수 있다. 또한 k-공간중위수 군집방법은 이상점을 사전에 식별하지 않고도 이상점을 사전에 식별해야 하는 박진우와 윤석훈 (2008)의 방법과도 비슷한 층화 효율을 가짐을 확인하였다.

네이만 배분의 평균을 이용한 표본배분

앞에서, 층별로 네이만 배분을 이용하여 표본배분을 할 때, 11개의 변수를 각각 층화변수로 이용하여 층별 표본을 얻었다. 따라서, 이러한 방법은 관심변수별로 층별 표본크기가 다르기 때문에, 여기서는 층별 표본크기를 11개의 변수별 네이만 배분 결과 나타난 층별 표본크기의 평균 $n_h = 1/p \sum_{j=1}^{p} n_{hj} (h = 1, \ldots, L)$을 이용하였고 (Schuenemeyer, 1975), 그 결과는 표 3.8과 같다.

또한, 각 방법별로 관심변수의 모평균 추정량 $\hat{\theta}_j$에 대한 분산 $V(\hat{\theta}_j)$을 구한 결과는 표 3.9와 같다. 앞 절의 결과와 마찬가지로 k-공간중위수 군집방법을 이용한 층화의 효율이 항상 k-평균 군집방법보다 비슷하거나 좋다는 것을 알 수 있다.

표 **3.9.** 관심변수별 모평균 추정량에 대한 분산

변수		모평균 추정량에 대한 분산		
		단순임의추출	층화임의추출	
			k-공간중위수 군집방법	k-평균 군집방법
주택유형별 (가구수)	단독주택	3734.05	687.92	1005.13
	아파트	15040.70	288.17	291.20
	연립주택	220.25	15.99	20.79
	기타주택	220.20	26.22	32.11
연령대별 (인구수)	~10	9493.24	201.10	232.78
	10~20	6297.36	248.21	260.03
	20~30	7258.96	364.98	387.34
	30~40	16300.38	325.29	386.13
	40~50	8568.78	294.91	362.87
	50~60	2423.66	175.53	238.91
	60~	3734.63	681.18	973.46

4. 결론

표본설계에서 모집단의 층화는 추론의 효율을 높이기 위해 일반적으로 많이 쓰이는 방법 중의 하나지만, 모집단의 변수들은 실제 많은 경우 이상점을 가지고 있다. 이상점이 존재하는 다변량 자료의 경우, 층화를 위한 k-평균 군집방법은 이상점에 매우 민감하게 반응하는 단점이 있으며 이를 사용하기 위해서는 이상점을 사전에 식별해야하는 과정이 필요하다. 본 연구에서 제안한 k-공간중위수 군집방법을 활용한 층화방법은 k-평균 군집방법보다 강건하며 이상점을 따로 식별할 필요가 없는 주관이 배제된 방법이다. 기존 연구와의 비교를 위해 박진우와 윤석훈 (2008)과 동일한 자료에 대한 사례분석을 통해 제안된 층화방법의 효율성을 네이만 배분을 사용하여 구한 추정량의 분산을 통해 입증하였다.

참고문헌

농촌진흥청 (2006). <2006 농촌생활지표>, 농촌진흥청.

박진우, 윤석훈 (2008). 이상점을 고려한 다변량 층화, <응용통계연구>, **21**, 377–385.

통계청 (2006). <2005 인구주택총조사>, 통계청.

Brown, B. M. (1983). Statistical uses of the spatial median, *Journal of the Royal Statistical Society. Series B,* **45**, 25–30.

Cuests-Albertos, J. A., Gordaliza, A. and Matran, C. (1997). Grand tour and projection pursuit, *Journal of Computational and Graphical Statistics*, **4**, 155–172.

Golder, P. A. and Yeomans, K. A. (1973). The use of cluster analysis for stratification, *Applied Statistics*, **22**, 213–219.

Jarque, C. M. (1981). A solution to the problem of optimum stratification in multivariate sampling, *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **30**, 163–169.

Jin, S. (1999). A study on the partitioning method for cluster analysis, 박사학위논문, 고려대학교.

Lavallée, P. and Hidiroglou, M. A. (1998). On the stratification of skewed populations, *Survey Methodology*, **14**, 33–43.

Milligan, G. W. (1980). An examination of the effect of six types of error perturbation of fifteen clustering algorithms, *Psychometrika*, **45**, 325–342.

Milligan, G. W. (1981). A review of Monte Carlo tests of cluster analysis, *Multivariate Behavioral Research*, **16**, 379–407.

Schuenemeyer, J. H. (1975). Maximum eccentricity as a union-intersection test in multivariate analysis, Geogia University, Athens.

Stratification Method Using k-Spatial Medians Clustering

Soonchul Son[1] · Myoungshic Jhun[2]

[1]Department of Statistics, Korea University; [2]Department of Statistics, Korea University

(Received February 2009; accepted April 2009)

Abstract

Stratification of population is widely used to improve the efficiency of the estimation in a sample survey. However, it causes several problems when there are some variables containing outliers. To overcome these problems, Park and Yun (2008) proposed a rather subjective method, which finds outliers before k-means clustering for stratification. In this study, we propose the k-spatial medians clustering method which is more robust than k-means clustering method and also does not need the process of finding outliers in advance. We investigate the characteristics of the proposed method through a case study used in Park and Yun (2008) and confirm the efficiency of the proposed method.

Keywords: k-means clustering, k-spatial medians clustering, multivariate stratification, Neyman allocation, outliers.

This work was supported by a Grant from the Korea Research Foundataion(KRF-2007-314-C00039).

[2]Corresponding author: Professor, Department of Statistics, Korea University, Seoul 136-701, Korea.
E-mail: jhun@korea.ac.kr