응용통계연구 제 **21**권 3호, *2008*년, *pp. 377 385*

이상점을 고려한 다변량 층화

박진우[1] 윤석훈[2]

요 약

여러 통계작성기관에서 실시하는 대부분의 표본조사들은 하나의 표본을 통해 서로 다른 여러 항목들을 조사하는 다목적조사이다. 다목적표본설계에서 층화변수들은 다변량이고 또한 서로 이질적인 속성을 지니는 관심변수들을 종합적으로 고려해야 하므로 층화는 매우 복잡한 양상을 띤다. 본 연구는 K-평균군집법을 적용한 다변량 층화에서 이상점의 효과를 지적하고, 층화 단계에서 사전에 이상점을 고려할 것을 제안하는 연구이다. 농촌생활지표조사를 위한 표본설계의 사례를 통해 이상점을 고려한 층화의 효과를 실증적으로 보인다.

주요용어: 다목적조사, 층화, 이상점, K-평균군집법.

1. 서론

표본설계에서 효율을 극대화시키기 위해 가장 널리 활용하는 방법 중의 하나로 층화(stratification)를 들 수 있다. 모집단 조사단위에 대해 유용한 보조정보들이 있을 때 이를 잘 활용하면 효과적으로 층화를 할 수 있다. 대부분의 전국 규모의 표본조사에서는 층화를 사용한다. 주요한 표본조사론 관련 문헌들 (Kish. 1965; Cochran, 1977; Sarndal 등, 1992; 박홍래, 2000)에서는 관심변수가 하나인 경우의 층화 문제를 주로 다루고 있다. 그런데 여러 기관에서 수행하는 전국적인 표본조사에서는 하나의 표본을 통해 여러 개의 서로 다른 항목을 조사하는 것이 일반적이다. 이와 같이 하나의 표본을 통해 서로 다른 복수의 관심변수에 대해 조사하는 것을 다목적조사(multipurpose survey)라고 하며 이때의 표본을 다목적표본이라고 한다. 농촌자원개발연구소에서 실시하는 농촌생활지표조사 (농촌진흥청, 2006)는 다목적표본조사의 한 예로서 한 표본조사를 통해 무려 100개가 넘는 항목들을 조사하게 된다.

다목적조사를 위한 표본설계에서는 관심변수들마다 특성이 서로 다르므로 어느 항목을 주로 고려하느냐에 따라 층화의 양상이 매우 달라질 수 있다. 다목적조사의 층화는 이론적으로 매우 복잡하고 어려운 편이지만 이에 대한 논의는 상대적으로 빈약한 편이다. Golder와 Yeomans (1973), Jarque (1981)는 다변량조사를 위한 층화에서 군집분석이나 주성분분석 등의 다변량기법을 사용하는 연구를 수행한 바 있다. 류제복 등 (2002)은 어가경제조사를 위한 표본설계에서 회귀나무를 사용한 층화방법을 사용하였고, 박진우 등 (2007)은 농기계조사를

1) (445-743) 교신저자. 경기도 화성시 봉담읍 와우리 산2-2, 수원대학교 통계정보학과, 부교수.
E-mail: jwpark@suwon.ac.kr
2) (445-743) 경기도 화성시 봉담읍 와우리 산2-2, 수원대학교 통계정보학과, 부교수.
E-mail: syun@suwon.ac.kr

위한 표본설계에서 요인분석(factor analysis)과 군집분석(cluster analysis)을 사용한 층화에 대해 연구한 바 있다. 다변량 기법을 사용한 기존의 층화 연구들에서는 층화변수들 중 이상점(outlier)이 존재하는 상황이 별도로 고려되지 않았다. 그러나 실제 층화를 위해 군집분석과 같은 다변량 기법을 이용할 경우, 이상점의 존재 여부에 따라서 층을 나누는 기준이 달라지는데 이것은 층화의 효율에 직접적인 영향을 끼치게 된다. 따라서 군집분석을 사용하여 층화를 하고자 할 때에는 층화를 하기 전에 이상점을 식별하여 적절히 처리하는 것이 필요하다.

본 연구의 목적은 다변량조사를 위한 층화에서 이상점을 사전에 고려하는 층화방법을 제안하는 것이다. 구체적으로 농촌생활지표조사를 위한 표본설계를 예로 들어 사전에 이상점을 고려하는 것의 필요성을 보이는 동시에 그 효과를 제시한다. 2장에서는 이상점을 고려한 층화방법을 소개하며 3장에서는 농촌생활지표조사의 예를 이용하여 2장에서 소개한 층화 방법의 효과를 실증적으로 소개한다. 마지막으로 4장에서는 본 연구의 결과를 간단하게 요약, 정리한다.

2. 이상점을 고려한 층화

표본설계에서 층화는 추출단위들에 대한 보조정보를 이용하여 추출단위를 몇 개의 군으로 나누는 작업이다. 다목적조사를 위한 표본설계에서는 자연스럽게 여러 개의 층화변수를 사용하게 된다. 그런데 다변량 층화변수가 존재할 때 이를 위한 표준적인 층화 방법이 딱히 마련되어 있지 않고 각 경우마다 표본설계자에 따라 나름의 합리적인 방법을 찾아야 하는 실정이다. 층화 작업은 기본단위의 속성에 따라 모집단을 몇 개의 부분집단으로 나누는 작업이라는 측면에서 본질적으로 다변량 분류기법인 군집분석(cluster analysis)과 맥을 같이한다(Golder와 Yeomans, 1973). 표본이론에서 층화의 효율을 평가하는 척도로써 추정량의 분산을 사용하는데 이것은 군집분석에서 군집화의 기준으로 평균제곱오차를 사용하는 K-평균군집법(K-means clustering method)과 Ward법에 해당된다. 따라서 층화를 다변량 자료를 활용하는 군집화(clustering) 작업이라고 본다면, 층화를 위해 자연스럽게 군집분석 기법을 도입하여 사용할 수 있다. 실제 Golder와 Yeomans (1973), Jarque (1981), 박진우 등 (2007)은 보조변수가 여러 개인 경우의 층화를 위해 요인분석, 주성분 분석, 군집분석 등의 다변량분석 기법을 활용한 바 있다. 다변량통계 기법을 층화에 적용하게 되면 연속형 층화변수인 경우 별도의 범주화하는 과정을 거치지 않고도 층화할 수 있다는 장점이 생긴다.

층화를 위해 군집분석을 사용하려고 할 때 생기는 문제 중 하나는 바로 이상점의 문제이다. 논의를 단순하게 하기 위해 먼저 단일변량의 연속형 층화변수가 주어지는 경우를 고려해보자. 이때 층화를 위해 가장 널리 활용되는 방법은 Dalenius와 Hodges (1959)의 누적도수제곱근법(이하 D-H법)이다. D-H법에서는 층화변수의 개별 자료가 도수로 변환되어 고려되므로 상대적으로 이상점의 영향이 완화된다. 즉, D-H법을 사용한 층화는 그 자체가 이상점을 고려한 층화라고 할 수 있다. 하지만 층화를 위해 군집분석을 사용한다면 이상점의 문제가 생긴다. 군집분석은 D-H법과는 달리 층화변수의 개별 자료를 직접적으로 사용하게 되므로 이상점이 존재할 때 그로 인해 큰 영향을 받는다. 대개 이상점은 별도의 군집으로 분류되기 쉬우므로 결과적으로 하나의 조사단위로만 구성되는 층이 생길 수 있다. 그런데 이런 현상은 일반적

인 표본설계의 층화에서는 바람직하지 않다. 따라서 층화를 위해 군집분석을 활용하려고 한다면 반드시 사전에 이상점이 존재하는 지를 고려하여 적절히 대처하는 것이 필요하다.

다목적조사에서 층화변수가 다변량일 때도 이상점이 존재할 수 있다. 따라서 군집분석법을 이용하여 층화를 하고자 한다면 사전에 먼저 이상점을 식별하는 것이 필요하다. 다변량 자료에서 이상점을 식별하는 방법으로는, 기본적으로 자료의 분포를 가정하고 그에 따라 적절한 이상점 식별 규칙을 제시한 표준적인 방법들 (Davies와 Gather, 1993; Rocke와 Woodruff, 1996; Becker와 Gather, 1999)과 간편한 그래프를 이용하여 시각적으로 이상점을 식별하는 방법 (Granquist, 1995), 군집분석에서 군집의 수를 실제 원하는 층의 수보다 많게 한 후 아주 작은 수의 조사단위들만 포함되는 군집을 이상점으로 간주하는 방법 등이 있다. 표본 이론의 주류라고 할 수 있는 설계기반(design-based) 접근법에서는 가급적 분포의 가정을 꺼리고 있는 점을 감안한다면 층화를 위한 이상점 식별 방법으로는 그래프를 이용하는 방법이나 군집분석을 이용하는 방법이 유용하다고 할 수 있다. 본 연구에서는 다변량 자료에서의 이상점 식별을 위해 산점도행렬(scatter plot matrix)을 활용한다. 그래프를 통해 이상점을 식별하는 경우 이상점의 기준이 애매할 수 있으므로 아주 명확하게 드러나는 점들만 이상점으로 제거하는 방안, 이상점의 기준을 넓게 하여 보다 많은 점들을 제거하는 방안 등을 동시에 검토하는 것이 바람직할 것이다.

사전에 나누고자 하는 층의 수가 정해져 있으면 바로 K-평균군집법으로 층을 나누면 되는데 일반적으로 층의 수를 몇 개로 해야 할지 정해져 있지 않다. 이럴 때에는 현실적으로 가능한 여러 가지 층의 수를 모두 고려하여 군집분석을 실시한 후 여러 통계량들을 비교, 검토하여 최적의 층수를 결정하는 것이 바람직하다.

산점도행렬을 그려 시각적으로 확연하게 구분되는 의심점에 해당되는 자료를 이상점으로 간주하여 일차적으로 이 점들을 제외시킨 나머지 자료들을 가지고 군집분석을 수행한다. 이상점을 제외시킨 자료들에 의한 층화가 이루어지고 나면 사전에 제외시켜 두었던 이상점 자료들을 이미 형성된 층들의 특성과 비교하여 가장 비슷하다고 판단되는 층에 소속시킨다. 이렇게 할 경우 이상점 자체는 층 내의 다른 단위들과 다소 이질적일 수는 있겠지만 그래도 이상점으로 인한 전체적인 층화 과정의 왜곡은 막을 수 있어 결과적으로 층화의 효율을 높일 수 있다.

3. 농촌생활지표조사를 위한 층화의 예

3.1. 농촌생활지표조사

농촌생활지표 조사는 "농촌주민들의 생활수준을 종합적이고 체계적인 방법으로 질적, 양적으로 파악하여 농촌주민의 삶의 질 향상을 위한 연구 및 지도사업과 관련 정책수립에 필요한 기초자료로서의 농촌생활지표를 작성"하는 데에 그 목적을 두고 있다 (농촌진흥청, 2006). 농촌생활지표란 "농촌생활의 전반적인 사항에 관한 한정되고 대표적인 항목의 계량화된 척도"로서 인간의 기본적 생활영역과 건강, 경제, 문화, 여가, 교육 등의 다양한 생활영역을 포함하고 있어 농촌지역 주민의 생활수준을 양적, 질적으로 파악할 수 있는 자료이다. 농촌생활지표 조사는 2000년 이후 매년 실시되는 조사로서 농촌생활지표 10개 부문을 2년 주기로 나누어 조사를 해오고 있다. 표본크기는 전국 400개 마을 2,000 가구이다. 이 조사는 전국의

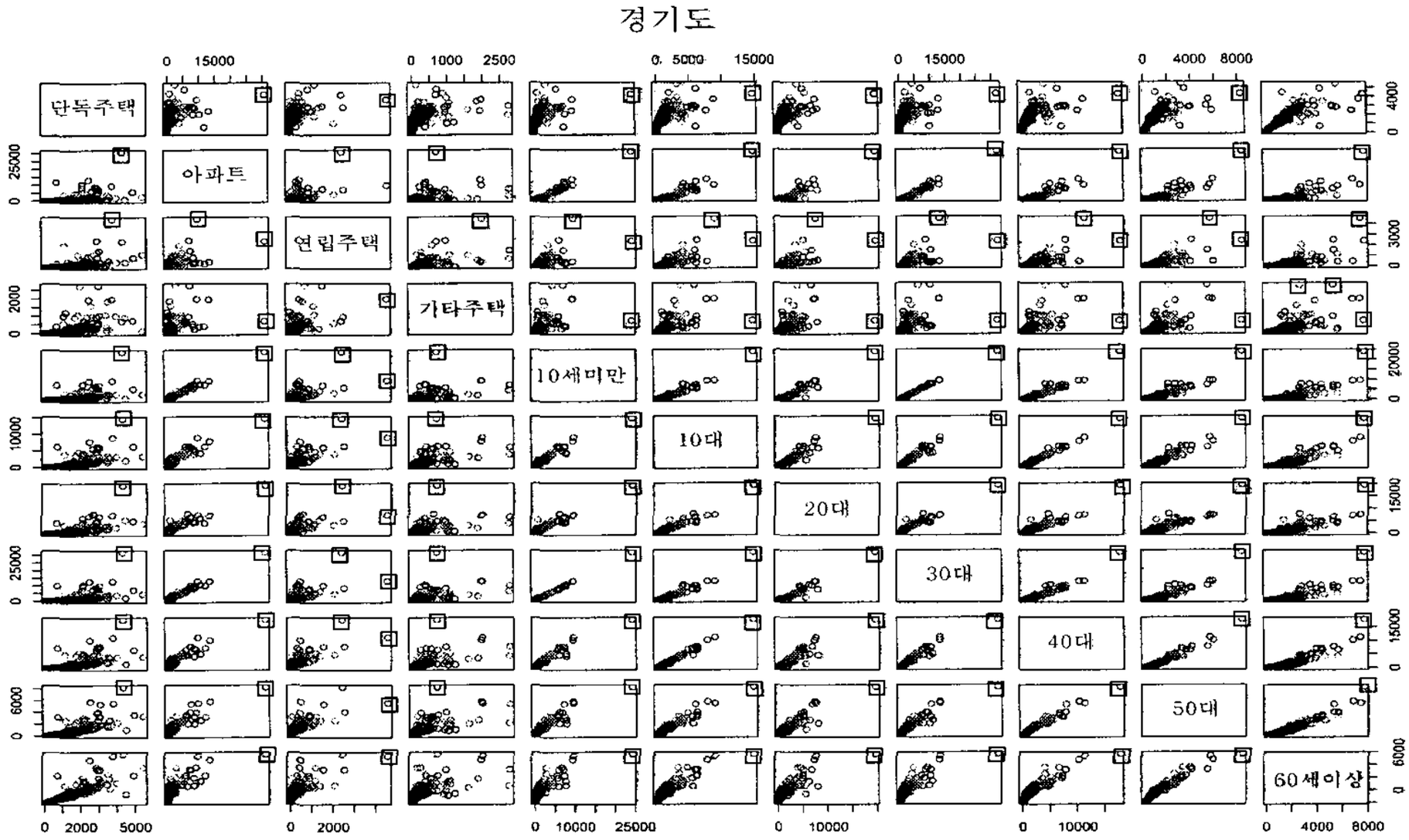

그림 3.1: 경기도 층화변수들의 산점도행렬

읍/면 거주 가구를 모집단으로 하는데 표본설계를 위한 추출틀로는 2005년 인구주택총조사의 읍/면 자료가 사용되었고, 1차추출단위는 읍/면이다.

층화를 위해 활용 가능한 층화변수로는 읍/면의 주택유형별(단독주택, 아파트, 연립주택, 기타주택) 가구수와 연령대별(10세미만, 10대, 20대, 30대, 40대, 50대, 60세이상) 인구수 등이 있다. 이 변수들은 서로·독립적이지 않고 복잡한 연관 관계를 지닌다. 또한 도별 지역통계 작성을 해야 하므로 도별로 따로 층화를 하게 된다.

3.2. 이상점 식별

층화에 앞서 도별로 이상점이 존재하는 지의 여부를 파악하는 것이 필요하다. 본 연구에서는 간편하게 이상점을 식별하기 위해 산점도행렬(scatter plot matrix)을 이용한다. 도별로 각각 11개 층화변수를 이용한 산점도행렬을 그렸는데 다음의 그림 3.1은 경기도의 산점도행렬 그림이다. 그림 3.1에서 작은 사각형으로 표시한 점들은 이상점으로 의심되는 자료의 점인데 확인한 결과 대부분의 점은 화성시 태안읍에 해당되는 것이었다.

산점도행렬에 의해 각 도별 이상점을 찾아 정리한 결과가 표 3.1에 나와 있다. 전국 1,410개의 1차추출단위 중 33개가 이상점으로 파악되었는데 해당 지역은 모두 다른 지역에 비해 도시화가 상당히 이루어진 인구밀집 지역이었다. 이상점으로 파악된 지역의 자료는 층화를 위한 군집분석의 대상에서 일단 제외시켰다가 나중에 형성된 층들 중 속성 상 가장 유사한 층에 포함시키게 된다.

표 3.1: 도별 확인된 이상점

도	읍/면 수	이상점 수	비고
경기	160	1 (화성시 태안읍)	인천시 포함
강원	114	2 (원주시 문막읍, 홍천군 홍천읍)	
충북	103	4 (청원군 내수읍, 옥천군 옥천읍, 진천군 진천읍, 증평군 증평읍)	
충남	170	4 (천안시 직산읍, 천안시 목천읍, 연기군 조치원읍, 홍성군 홍성읍)	
전북	159	4 (완주군 삼례읍, 완주군 봉동읍, 고창군 고창읍, 부안군 부안읍)	
전남	229	6 (순천시 해룡면, 광양시 광양읍, 화순군 화순읍, 해남군 해남읍, 영암군 삼호읍, 영광군 영광읍)	
경북	247	6 (대구시 달성군 화원읍, 포항시 남구 오천읍, 포항구 북구 흥해읍, 경산시 하양읍, 경산시 진량읍, 칠곡군 석적면)	대구시 포함
경남	216	6 (부산시 기장군 기장읍, 울산시 울주군 범서읍, 마산시 내서읍, 김해시 장유면, 거제시 신현읍, 양산시 웅상읍)	부산시, 울산시 포함
제주	12	0	
계	1410	33	

표 3.2: 도별 층 수

도	경기	강원	충북	충남	전북	전남	경북	경남	제주	계
층 수	4	3	3	4	4	5	6	4	1	34

3.3. 층화

본 연구에서는 층화를 위해 K-평균군집법(K-means clustering method)을 사용한다. K-평균군집법은 각 1차추출단위의 11개 층화변수 값을 11차원 공간상의 점으로 간주하고 이 점들을 K개의 군집으로 나누되 각 군집의 중심과 그 군집에 속한 점들과의 거리의 제곱의 전체 합이 최소가 되도록 군집을 나누는 방법이다.

이상점을 제외시킨 1,377개의 1차추출단위 자료를 활용하여 각 도별로 층의 수를 1~7개로 하여 층화한 후 각각에 대해 11개 층화변수의 분산을 계산하였다. 이 결과를 이용하여 도별로 적절한 층수를 결정하였는데 표 3.2는 최종적으로 결정된 도별 층의 수를 나타낸다. 제주도만 별도로 층을 나누지 않았고 나머지 도들은 3~6개의 층으로 결정됐다. 다음으로는 앞서 제외시켜 놓았던 33개 이상점들을 이미 구분한 지역별 층 중 적절한 층에 포함시킨다. 경기도를 예로 들면, 화성시 태안읍을 제외시킨 후 4개의 층으로 나누었는데, 층 4가 전반적으로 인구수나 아파트 가구수가 많은 조사단위로 구성되어 있었다. 따라서 가장 인구수가 많은 이상점 태안읍을 층 4에 포함시켰다. 이렇게 함으로써 층화가 완성된다.

경기도의 층수는 4개인데, 층화 후 각 층별 11개 층화변수의 평균과 표준편차를 계산한 결과가 표 3.3이다. 이를 더욱 일목요연하게 파악하기 위해 각 변수값들의 분포를 층별 상자그림으로 나타낸 것이 그림 3.2에 나와 있다. 이 표와 그림을 관찰해보면, 대부분의 변수에 대해 층별 분포가 확연하게 다른 것을 알 수 있다. 좋은 층화란 층간에 이질적인 것을 의미하므로, 위의 결과는 층화가 효과적으로 이루어졌다는 것을 시사한다.

표 3.3: 경기도 읍/면의 층별, 변수별 기초통계량

변수		층 1		층 2		층 3		층 4	
		평균	SD	평균	SD	평균	SD	평균	SD
가구수	단독주택	1235.12	475.35	2420.40	662.44	2754.47	1060.03	2908.92	1084.40
	아파트	68.95	144.13	911.87	633.33	3518.80	1340.38	10437.1	6722.52
	연립주택	47.66	70.54	296.76	230.79	738.13	561.50	1271.42	1297.49
	기타주택	119.35	172.24	424.92	341.01	805.33	653.86	950.33	809.76
인구수	10세미만	378.76	228.94	1411.08	506.15	3474.87	691.86	8604.67	5080.34
	10대	487.18	292.05	1520.47	450.69	3330.47	647.20	6545.08	2891.02
	20대	491.20	332.92	1785.87	793.30	3491.07	674.08	6923.75	4164.38
	30대	515.73	313.39	1966.37	617.19	4686.13	857.90	11258.8	6656.94
	40대	692.20	350.93	1987.95	517.21	4171.07	777.66	8042.42	3704.19
	50대	585.60	255.73	1350.21	348.70	2403.20	475.28	4072.00	1834.56
	60세이상	1171.42	462.36	2213.37	564.35	3280.87	922.76	4970.58	1708.75

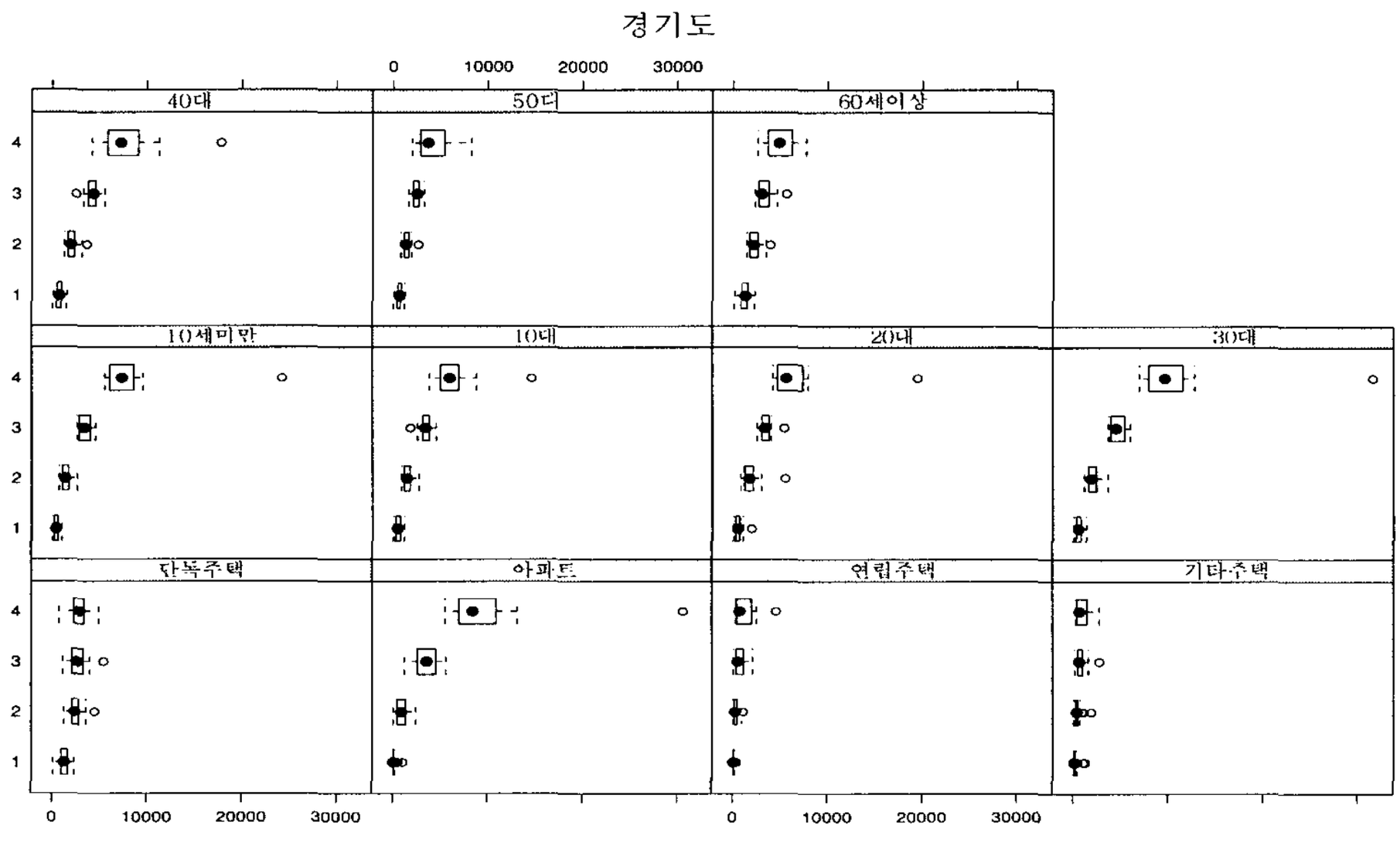

그림 3.2: 경기도 층화변수값들의 층별 분포

3.4. 층화효과 비교

본 연구에서 소개한 층화전략의 효과를 살펴보기 위해 농촌생활지표조사를 위한 추출틀 자료에 있는 11개 변수와 1차추출단위별 가구수, 경제활동인구수(20대에서 50대 인구수) 등의 변수에 대해 모평균 추정량에 대한 추정분산값을 구하여 비교하기로 한다. 실제 농촌생활지표조사에서와 같이 200개의 1차추출단위 표본을 추출하는 경우에 대해 각각 단순임의추출, 이상점을 고려한 층화임의추출, 이상점을 고려하지 않은 층화임의추출에서의 추정량의 분산

표 3.4: 여러 변수별 층화효과 비교

변수		모평균 추정량에 대한 추정분산		
		단순임의추출	층화임의추출(이상점 고려)	층화임의추출(이상점 미고려)
가구수	단독주택	3734.05	585.50	753.55
	아파트	15040.70	65.47	230.95
	연립주택	220.25	9.10	11.52
	기타주택	220.20	25.03	25.42
	총가구	32004.18	843.75	1323.87
인구수	10세미만	9493.24	74.83	223.26
	10대	6297.36	212.91	254.80
	20대	7258.96	317.96	352.24
	30대	16300.38	110.43	375.92
	40대	8568.78	244.40	336.73
	50대	2423.66	154.60	201.24
	60세이상	3734.63	565.91	725.88
	경제활동인구	119123.30	2352.90	3816.81

을 계산하여 비교한 결과가 다음의 표 3.4에 나와 있다. 층화임의추출의 경우 표본배분은 네이만배분법에 따랐다.

표 3.4에 따르면, 고려한 모든 변수에 대해 단순임의추출보다 층화임의추출에 의한 추정분산의 값이 월등하게 작은 것을 알 수 있다. 이것은 군집분석을 이용한 다변량 층화의 효과가 뛰어났음을 시사한다. 다음으로 이상점을 고려한 층화와 고려하지 않은 층화의 추정분산을 비교하면, 모든 변수에 대해 이상점을 고려했을 때가 더 효율적인 것으로 나타났다.

이상의 결과를 통해 다목적표본설계를 위한 층화 과정에서 이상점을 고려하는 것이 필요하다는 사실을 알 수 있다. 다변량 층화변수들의 활용이 가능한 경우 군집분석과 같이 잘 알려진 다변량 기법을 사용하여 층화를 하면 매우 편리하다. 그렇다고 해서 방대한 층화변수 자료들의 특성에 대한 검토 없이 무조건 군집분석을 하는 것은 바람직하지 못하다. 이상점으로 인해 층화가 왜곡되는 것을 방지하기 위해서는 사전에 이상점 유무를 확인하여 적절한 조치를 취하는 것이 효과적이다.

4. 맺음말

표본설계에서 층화는 널리 사용되고 있으며 설계의 효율을 높이기 위해 아주 중요하게 다루어진다. 실제 현장에서 다루는 대부분의 층화는 다목적조사를 염두에 둔 것으로서 다변량 층화변수를 사용하여 이루어진다. 그런데 대부분의 표본조사 관련 문헌들은 하나의 관심변수 조사를 염두에 둔 층화를 주로 다루고 있을 뿐이다. 이런 까닭에 다목적조사를 위한 층화는 표준적인 방법 없이 설계자에 따라 임기응변식으로 이루어지는 실정이다.

본 연구에서 다목적조사를 위한 층화를 위해 K-평균군집법을 사용하는 것을 제안하였다. 다목적조사에서의 층화를 위해 다변량기법을 사용한 일부 선행연구들이 있었다. 그런데 기존의 연구에서는 층화를 위한 보조정보에 이상점이 존재하는 경우를 따로 고려하지 않고 있다.

본 연구에서는 K-평균군집법을 층화에 적용할 때 이상점으로 인한 문제를 지적하고, 이상점을 고려하는 층화방법을 제안하였다. 농촌생활지표조사를 위한 표본설계의 예를 통해, 이상점을 식별하여 제거한 후 K-평균군집법을 적용하여 층화한 경우 그렇지 않은 경우보다 층화의 효율을 더욱 높일 수 있다는 사실을 보였다.

참고문헌

농촌진흥청 (2006). <2006 농촌생활지표>, 농촌진흥청.

류제복, 김영원, 박진우 (2002). 어가경제조사 표본설계, <통계분석연구>, **7**, 159–183.

박진우, 윤석훈, 김진흠, 정형철 (2007). 군집분석을 이용한 다목적 조사의 층화에 관한 연구, <응용통계연구>, **20**, 387–394.

박홍래 (2000). <통계조사론>, 영지문화사.

Becker, C. and Gather, U. (1999). The masking breakdown point of multivariate outlier identification rules, *Journal of the American Statistical Association*, **94**, 947–955.

Cochran, W. G. (1977). *Sampling Techniques*, John Wiley & Sons, New York.

Dalenius, T. and Hodges, J. L. (1959). Minimum variance stratification, *Journal of the American Statistical Association*, **54**, 88–101.

Davies, L. and Gather, U. (1993). The identification of multiple outliers, *Journal of the American Statistical Association*, **88**, 782–792.

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, *Applied Statistics*, **30**, 163–169.

Kish, L. (1965). *Survey Sampling*, John Wiley & Sons, New York.

Rocke, D. M. and Woodruff, D. L. (1996). Identification of outliers in multivariate data, *Journal of the American Statistical Association*, **91**, 1047–1061.

Sarndal, C. E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*, Springer, New York.

[2008년 2월 접수, 2008년 3월 채택]

Multivariate Stratification under Consideration of Outliers

Jin Woo Park[1] Seokhoon Yun[2]

Abstract

Most of the sample surveys conducted by several statistics preparation agencies are multipurpose surveys inquiring into several distinguishing items through a single sample. In a multipurpose sample design, the stratification tends to be very complex since the stratification variables which are both multivariate and heterogeneous must be considered collectively. In this paper we point out an outlier effect in a multivariate stratification to which the K-means clustering method is applied and propose to consider outliers prior to the stratification step. We also show an empirical stratification effect under consideration of outliers through a case study of sample design for The Rural Living Indicators.

Keywords: Multipurpose survey, stratification, outlier, K-means clustering method.

1) Corresponding author. Associate Professor, Dept. of Applied Statistics, University of Suwon, Suwon 445-743, Korea.
E-mail: jwpark@suwon.ac.kr
2) Associate Professor, Dept. of Applied Statistics, University of Suwon, Suwon 445-743, Korea.
E-mail: syun@suwon.ac.kr