응용통계연구 제 **20**권 2호, *2007*년, *pp. 387–394*

군집분석을 이용한 다목적 조사의 층화에 관한 연구

박진우[1] 윤석훈[2] 김진흠[3] 정형철[4]

요 약

본 연구는 여러 가지의 양적변수들을 조사하는 다목적, 다변량조사 표본설계에서 층화 문제를 다룬다. 다변량 층화변수를 사용하는 층화 방법으로 일변량 층화변수가 있을 때 사용하는 누적도수제곱근법을 독립적으로 여러 층화변수에 적용하는 방법, 군집분석을 이용하는 방법, 인자분석과 군집분석을 함께 이용하는 방법 등 세 가지 방법을 제시한다. 한편, 2001년 농업총조사 자료에 나타난 동·읍·면의 농기계별 보유대수 정보를 층화변수로 활용하여 세 가지 층화 방안의 효율을 실증적으로 비교하게 되는데 그 결과 인자분석과 군집분석을 함께 고려한 층화방법이 비교적 효율적인 것으로 나타났다.

주요용어: 다변량조사, 층화, 층의 경계점, 누적도수제곱근법, 요인분석, 군집분석.

1. 서 론

표본설계에서 층화 작업은 매우 중요한 부분이다. 어떤 층화변수를 사용하여 어떻게 층화하느냐에 따라 추정량의 효율이 달라지기 때문이다. 조사변수와 높은 상관관계를 갖는 층화변수가 존재할 경우 이를 이용하면 층화의 효과를 높일 수 있다. 층화변수가 연속형 변수일 때 층화는 층의 경계점을 정하는 문제로 귀착된다. 층화변수가 일변량인 경우의 층의 경계점을 정하는 방법으로 가장 널리 알려진 것은 Dalenius와 Hodges(1959)의 누적도수제곱근(the cumulative $\sqrt{f}$)법이다. 그 밖에 Eckman(1959)은 수치적인 계산법을 사용하는 또 다른 방법을 제안했으며, Hedlin(2000)도 Eckman법을 수정한 방법을 제안한 바 있다. 한편, Lavallee와 Hidiroglou(1988), Rivest(2002), Gunning과 Horgan(2004) 등은 치우친 분포의 모집단에 대한 층화경계점을 정하는 문제를 연구하였다.

일반적인 표본조사에서는 하나의 표본을 통해 여러 가지 변수들에 대한 조사를 하는 다변량조사(multivariate survey)를 실시하는 경우가 많다. 이 때 조사변수에 따라 층화의 기준이 서로 다를 수 있으므로 여러 개의 층화변수를 고려하는 것이 필요할 때가 있다. 박

1) (445-743) 경기도 화성시 봉담읍 와우리 산 2-2, 수원대학교 통계정보학과, 부교수
E-mail: jwpark@suwon.ac.kr
2) (445-743) 경기도 화성시 봉담읍 와우리 산 2-2, 수원대학교 통계정보학과, 부교수
E-mail: syun@suwon.ac.kr
3) (445-743) 경기도 화성시 봉담읍 와우리 산 2-2, 수원대학교 통계정보학과, 부교수
E-mail: jinhkim@suwon.ac.kr
4) (445-743) 경기도 화성시 봉담읍 와우리 산 2-2, 수원대학교 통계정보학과, 조교수
E-mail: jhc@suwon.ac.kr

진우(2000)는 농기계 판매액 조사를 위한 표본설계 연구에서 이런 문제를 지적한 바 있다. 이런 문제와 관련된 이론적인 연구들로서, Bryant 등(1960), Sitter와 Skinner(1994), Lu와 Sitter(2002) 등은 다차원 층화(multi-way stratification)에 관한 연구들을 하였으며, Golder와 Yeomans(1973), Jarque(1981)는 다변량 층화를 위해 군집분석(cluster analysis)을 사용하는 연구를 한 바 있다.

본 연구의 동기가 된 사례는 농기계 모니터링 시스템 마련을 위한 표본설계로서, 하나의 표본을 통해 트랙터, 콤바인, 이앙기, 관리기 등 여러 농기계들의 보유대수, 작업일수, 기계화율 등을 파악하고자 하는 표본설계이다(한국통계학회, 2006). 동·읍·면을 1차 추출단위로 하여 층화를 하는데 층화변수로 농업총조사 결과에 나온 동·읍·면의 농기계별 보유대수 정보를 활용한다. 논지역과 밭지역, 도시지역과 시골지역인지에 따라 농가가 보유하는 농기계의 종류가 서로 다른 편이므로 이런 점들을 모두 반영할 수 있으면서도 효율적인 층화를 해야 할 필요가 있다.

본 연구의 목적은 다변량조사를 위한 편리하면서도 효율적인 층화방법을 제시하는 것이다. 주어진 층화변수를 가지고 층을 구분하는 층화 작업은 본질적으로 데이터를 특정한 변수를 가지고 여러 개의 군집으로 나누는 군집분석과 일치한다. Jarque(1981)는 다변량 층화를 위해 주성분분석을 먼저 실시하고 거기서 파악된 제 1 주성분 값을 가지고 층화를 하는 방법을 소개한 바 있는데, 본 연구에서는 요인분석(factor analysis)를 통해 주요 요인들을 파악한 후 그 요인에 대해 군집분석을 실시하여 층화를 하는 방법을 제안하고자 한다. 구체적으로 농기계 관련 조사를 위한 표본설계의 예를 통해 층화 방법의 효율을 살펴보고자 한다. 2절에서는 몇 가지 층화전략을 소개하며, 3절에서는 농업총조사 데이터를 활용하여 제안한 전략들의 효율을 비교하였다. 마지막으로 4절에서는 전체적인 연구결과를 정리하였다.

2. 다변량 층화변수를 이용한 층화전략

여러 양적 조사변수들을 한꺼번에 조사하기 위한 다목적 표본설계(multipurpose sampling design)에서는 층화를 위해 여러 가지 전략을 고려할 수 있다. 본 연구에서는 네 가지 전략을 소개하고 비교하고자 하는데 한 가지를 제외하고는 모두 군집분석법을 사용하게 된다. 효율적인 층화란 가급적이면 같은 층 안에서는 동질적, 서로 다른 층들 간에는 이질적이 되도록 모집단 단위들을 묶는 것을 의미한다. 한편, 군집분석은 주어진 자료의 특성을 고려하여 가급적 유사한 단위들을 같은 군집으로 묶는 것이 목적이므로, 군집화와 층화는 본질상 동일한 성격을 지닌다고 할 수 있다. 따라서 본 연구의 주된 관심은 다변량 층화변수들이 주어질 때 층화의 문제를 군집화의 문제로 여기고자 하는 것이다. 효율적인 층화란 추정량의 분산을 최소화시킬 수 있도록 층을 나누는 것을 의미한다. 층화를 군집화의 문제로 볼 때, 분산을 최소화시키는 층화는 곧 평균제곱오차를 최소화시키는 군집화와 일치한다고 할 수 있다. 따라서 본 연구에서는 추출단위의 군집화를 위해 Ward(1963)가 제안한 군집방법을 고려하고자 하는데, 이 방법은 군집화로 인해 생기는 정보의 손실(군집 전후의 전체 제곱오차의 차이)을 최소화하는 방법이기 때문에 효율적인 층화와 개념적으로

일치한다(Jarque, 1981).

층화전략 1: 요인점수를 사용한 군집방법

여러 개의 층화변수들이 서로 독립이 아니라 어떤 상관구조를 지니므로 요인분석을 통해 변수들에 내재된 요인을 파악한 후 개별단위들의 요인점수(factor score)를 구해 군집화하는 전략이다. 이 전략의 특징은 층화 과정에서 사람에 의한 주관이 개입될 여지가 없이 통계적인 절차에 의해 기계적으로 층이 나눠진다는 점이다.

층화전략 2: 요인그룹 변수들의 합을 사용한 군집방법

요인분석을 통해 축약된 몇 개의 요인을 찾아낸 후 각 요인에 대해 높은 부하값(loading)을 지니는 변수들의 합을 새로운 변수로 삼아 군집화를 실시하는 방법이다.

층화전략 3: 직접적인 층화변수를 사용한 군집방법

이 전략은 층화변수들을 직접적으로 사용하여 군집화하는 방법이다. 요인분석을 이용한 별도의 변수 축약 과정을 생략하고 오직 군집 만을 사용하여 층화한다.

층화전략 4: 누적도수제곱근법을 사용

가장 쉽게 생각할 수 있는 것은 가장 중요하다고 생각되는 몇 가지 층화변수를 정한 후 각 층화변수에 대해 독립적으로 누적도수제곱근법을 적용하여 층의 경계점을 정하고 그것들을 결합시키는 방법이다. 이 방법은 층화변수들 사이의 상관관계를 무시한다는 점과, 고려하는 층화변수의 수가 많아질수록 층의 수가 지나치게 커진다는 문제를 지니고 있다. 그럼에도 불구하고 다른 세 가지 층화전략과의 비교를 위해 이 방법을 사용한다.

3. 농기계 관련 표본설계를 통한 층화전략의 비교

3.1. 농기계 모니터링 시스템 마련을 위한 표본설계

2절에서 소개한 네 가지 층화전략들의 효과를 비교하기 위해 농기계 모니터링 시스템 마련을 위한 표본설계의 예를 사용한다. 이 표본설계는 농기계의 보유대수, 작업일수, 기계화율 등 농기계 관련 조사를 위한 다목적 표본을 추출하기 위한 것이다(한국통계학회, 2006). 이 표본설계를 위한 1차 추출단위는 전국의 2,995개 동·읍·면인데 추출틀로 2000년 농업총조사(통계청, 2001)의 동·읍·면 통계를 사용한다. 층화변수는 각 동·읍·면의 경운기, 트랙터, 콤바인, 바인더, 관리기, 건조기, 이앙기, 정미기의 보유대수이다. 먼저 8개 층화변수들 간의 상관계수를 계산한 것이 다음의 표 3.1에 나와있다. 이 표를 보면 기종에 따라 상관계수의 크기가 서로 다른 편인데, 트랙터와 콤바인, 콤바인과 이앙기 등 일부 기종들은 상당히 큰 상관관계가 있다. 반면 바인더 같은 경우 어느 기종과도 상관계수의 크기가 그리 크지 않은 편이다. 그러나 상관행렬을 보고 한 눈에 변수들의 특성을 설명하기는 쉽지 않다.

표 3.1: 층화변수들 간의 상관계수 행렬

	경운기	트랙터	콤바인	바인더	관리기	건조기	이앙기	정미기
경운기	1.00	0.81	0.79	0.49	0.81	0.55	0.85	0.78
트랙터		1.00	0.92	0.30	0.57	0.44	0.87	0.68
콤바인			1.00	0.31	0.49	0.44	0.92	0.74
바인더				1.00	0.40	0.11	0.46	0.46
관리기					1.00	0.49	0.58	0.54
건조기						1.00	0.49	0.40
이앙기							1.00	0.77
정미기								1.00

표 3.2: 8개 변수의 요인 적재행렬

변수	요인 1	요인 2
경운기	0.552	0.838
트랙터	0.813	0.420
콤바인	0.952	0.308
바인더	0.215	0.431
관리기	0.252	0.807
건조기	0.316	0.434
이앙기	0.811	0.509
정미기	0.595	0.537
요인 공헌도	3.093	2.547

8개 층화변수들 간의 관계를 파악하기 위해 요인분석을 실시하여 구한 최종적인 요인 패턴행렬의 결과가 표 3.2에 나와 있다. 요인분석을 실시하는 이유는 농기계 보유대수들 간의 내부적 상호의존 관계를 그 저변에 내재하는 소수 몇 개의 공통요인으로 축약하여 새로운 개념을 지니는 변수로 해석하고자 하는 것이다. 즉, 8개 농기계의 보유대수를 좀 더 낮은 차원에서 새로운 개념의 변수로 축약하여 해석하고자 하는데 의미를 두고자 한다. 요인분석 결과 8개 농기계 보유대수에 대한 변수들의 특징을 파악할 수 있는데, 제 1 요인은 트랙터-콤바인-이앙기-정미기 등으로 대표되는 규모가 큰 기계라는 요인으로, 제 2 요인은 경운기-관리기-바인더-건조기 등의 중소형 기계라는 요인으로 파악되었다. 위의 요인분석 결과를 근거로 하여 앞 절에서 소개한 네 가지 층화전략을 생각하게 된 것이다. 이때, 층화전략 2에서는 대형기종수(=트랙터+콤바인+이앙기+정미기)와 중소형기종수(=경운기+바인더+관리기+건조기)라는 두 개의 새로운 변수에 대해 군집분석을 적용하게 된다.

표 3.3: 네 가지 층화전략 별 평균 농기계 보유대수 추정량의 표준오차

기종 / 전략	경운기	트랙터	콤바인	바인더	관리기	건조기	이앙기	정미기
층화전략 1	83.67	9.16	1.63	9.04	33.11	31.04	22.86	55.13
층화전략 2	**38.32**	7.85	1.77	**6.55**	26.41	27.76	**16.05**	27.46
층화전략 3	52.64	8.49	1.77	7.33	**23.87**	**24.28**	17.09	**26.84**
층화전략 4	137.28	**3.43**	**1.33**	10.49	60.82	43.17	18.29	59.02

3.2. 층화전략의 비교

앞에서 소개한 네 가지 층화전략에 따라 전국의 동·읍·면을 층화한 후, 2001년 농업총조사의 동·읍·면별 8개 농기계 보유대수를 조사변수로 하여 각 층화전략 하에서의 평균추정량의 분산을 계산한다. 이때 층의 수는 4개로 했는데 이는 실제 농기계 모니터링 시스템 구축을 위한 표본설계의 결정을 그대로 사용한 결과이다. 다음의 표 3.3은 네 가지 층화전략별로 평균추정량의 표준오차를 계산한 결과이다. 기종별로 표준오차가 가장 작은 값은 굵은 글씨체로 표시하였다. 층화전략 1에서 3까지는 요인분석과 군집분석을 활용한 전략이고, 층화전략 4는 트랙터에 대해 누적도수제곱근법을 사용한 전략이다. 각 층화전략별로 자세히 살펴보자. 먼저 상대적인 비교를 위해 의도적으로 포함시킨 층화전략 4의 결과를 살펴보자. 트랙터만을 층화변수로 고려한 전략이므로 예상대로 트랙터, 그리고 트랙터와 상관이 매우 높은 콤바인에 대해서는 가장 효율적인 전략으로 나타났다. 하지만 그 밖의 다른 기종들에 대해서는 효율적이지 못한 것으로 드러났다. 특히 경운기, 관리기, 건조기, 정미기 등 중소형 기종의 경우 전략 4는 바람직하지 못한데, 관련성이 많지 않은 트랙터만을 사용하여 층을 나눈 것이므로 이런 결과가 나온 것은 당연하다. 나머지 세 가지 층화전략은 모두 군집분석을 이용한다는 면에서는 공통점이 있는데 그 중 요인분석에서 나온 요인점수를 가지고 군집분석을 실시하는 층화전략 1은 조사 데이터를 직접적으로 사용하는 층화전략 2와 3에 비해 전반적으로 효율적이지 않은 것으로 나타났다. 층화전략 1의 경우, 조사된 데이터를 직접 사용하지 않고 그것을 요인점수라는 새로운 변수로 환산하여 사용하는데 이 과정에서 데이터 고유의 특성이 일정 부분 손상되어 원래 데이터의 속성을 왜곡시키는 측면이 발생하기 때문이다. Jarque(1981) 논문의 사례에서도 이와 유사한 결과가 나온 바 있는데, 조사 데이터를 직접 사용하여 군집화한 경우가 주성분분석의 제 1 주성분 값으로 군집한 경우에 비해 더 효율적이라는 결과가 나왔다. 한편 층화전략 2와 3은 전반적으로 효율이 비슷한 편인데 상대적으로 중요도가 높은 기종인 경운기, 트랙터, 콤바인, 바인더, 이앙기 등에서는 층화전략 2가 조금 나은 것으로 나타났다.

위의 분석 결과에 따라 농기계 모니터링 시스템 구축을 위한 표본설계에서는 최종적으로 층화전략 2를 채택하였다. 다음의 표 4.1는 층화전략 2에 따라 나누어진 4개 층별 현황을 행정구역별로 정리한 표이다. 각각의 층이 대농형, 중농형, 소농형, 도시형으로 확연하게 구분되므로 합리적인 층화라고 여겨진다.

표 4.1: 행정구역별 4개의 층에 분리된 동·읍·면의 개수

	층 1 (대농지역)	층 2 (중농지역)	층 3 (소농지역)	층 4 (도시지역)	총계
서울	0	0	0	190	190
부산	0	3	9	122	134
대구	1	3	8	115	127
인천	0	7	10	97	114
광주	0	1	7	74	82
대전	0	0	5	70	75
울산	0	8	7	41	56
경기	6	48	88	298	440
강원	1	16	76	94	187
충북	2	37	63	51	153
충남	13	77	80	36	206
전북	0	25	130	93	248
전남	21	77	123	74	295
경북	36	103	110	85	334
경남	26	59	111	116	312
제주	7	3	10	22	42
전국	113	467	837	1,578	2,995

4. 맺음말

표본설계에서 층화는 매우 널리 사용되는 방법인 동시에 효율을 높이기 위해 아주 중요한 작업이다. 본 연구에서는 여러 개의 양적 층화변수들이 주어지는 다변량조사를 위한 층화전략을 검토하였다. 일변량 층화 문제에서 가장 널리 사용되는 누적도수제곱근법을 여러 변수에 대해 독립적으로 사용하는 층화방법, 군집분석을 이용한 층화방법, 요인분석과 군집분석을 동시에 사용한 층화방법 등을 소개한 후, 2001년 농업총조사의 동·읍·면 농기계 보유대수 자료를 사용하여 각 층화방법의 효율을 비교하였다. 그 결과 요인분석과 군집분석을 함께 활용하는 층화방법이 비교적 만족스런 층화전략임을 알 수 있었다.

본 논문은 몇 가지 한계를 지닌다. 먼저, 여러 층화변수들이 동일한 중요성을 갖는 것으로 간주하였다. 실제 표본조사에서는 여러 변수들의 중요도가 다른 것이 일반적이다. 그러므로 각 변수별 비중을 서로 다르게 고려하는 군집분석을 하는 것이 바람직할 것인데 본 논문에서는 그런 경우까지 고려하지는 못했다. 또한, 농기계 보유대수 자료라는 특정한 사례에 대해서만 각 층화전략을 비교하였다. 향후 보다 다양한 사례를 통해 본 논문에서 제시한 층화전략의 효용성에 대한 검증이 이루어지는 것이 필요하다. 마지막으로, 기존의 연구들에서 제시된 보다 다양한 여러 방법들을 종합적으로 적용하여 비교하지 못하고 제한된

방법에 대해서만 비교하였다.

참고문헌

박진우 (2000). A sampling design of the agricultural machine estimated sales survey, *The Korean Communications in Statistics*, **8**, 375–382.

통계청 (2001). <2000 농어업총조사>, CD.

한국통계학회 (2006). <농기계 모니터링 시스템 표본설계>, 학술연구용역보고서.

Bryant, E. C., Hartley, H. O. and Jessen, R. J. (1960). Design and estimation in two-way stratification, *Journal of the American Statistical Association*, **55**, 105–124.

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

Eckman, G. (1959). An Approximation useful in univariate stratification, *The Annals of Mathematical Statistics*, **30**, 219–229.

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

Gunning, P. and Horgan, J. M. (2004). A new algorithm for the construction of stratum boundaries in skewed populations, *Survey Methodology*, **30**, 159–166.

Hedlin, D. (2000). A procedure for stratification by an extended Eckman rule, *Journal of Official Statistics*, **16**, 15–29.

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

Lavallee, P. and Hidiroglou, M. (1988). On the stratification of skewed populations, *Survey Methodology*, **14**, 33–43.

Lu, W. and Sitter, R. R. (2002). Multi-way stratification by linear programming made practical, *Survey Methodology*, **28**, 199–207.

Rivest, L. P. (2002). A generalization of the Lavallee-Hidiroglou algorithm for stratification in business surveys, *Survey Methodology*, **28**, 191–198.

Sitter, R. R. and Skinner, C. J. (1994). Multi-way stratification by linear programming, *Survey Methodology*, **20**, 65–73.

Ward, J. H. (1963). Hierachical grouping to optimize an objective function, *Journal of the American Statistical Association*, **58**, 236–244.

[2006년 12월 접수, 2007년 2월 채택]

A Study on the Use of Cluster Analysis for Multivariate and Multipurpose Stratification

Jin Woo Park[1] Seokhoon Yun[2] Jinheum Kim[3] Hyeong Chul Jeong[4]

ABSTRACT

This paper considers several stratification strategies for multivariate and multipurpose survey with several quantitative stratification variables. We propose three methods of stratification based on, respectively, the method of cumulative frequency square root which is the most popular one in univariate stratification, cluster analysis, and factor analysis followed by cluster analysis. We then compare the efficiency of those methods using the Dong-Eup-Myun data of the holding numbers of farming machines, extracted from the 2001 Agricultural Census. It turned out that the method based on cluster analysis with factor analysis would be a relatively satisfactory strategy.

Keywords: Multivariate survey, stratification, method of cumulative frequency square root, factor analysis, cluster analysis.

1) Associate Professor, Department of Applied Statistics, The University of Suwon, Suwon 445-743, Korea E-mail: jwpark@suwon.ac.kr
2) Associate Professor, Department of Applied Statistics, The University of Suwon, Suwon 445-743, Korea E-mail: syun@suwon.ac.kr
3) Associate Professor, Department of Applied Statistics, The University of Suwon, Suwon 445-743, Korea E-mail: jinhkim@suwon.ac.kr
4) Assistant Professor, Department of Applied Statistics, The University of Suwon, Suwon 445-743, Korea E-mail: jhc@suwon.ac.kr