Journal of Korean
Data & Information Science Society
2006, Vol. 17, No. 1, pp. 1 ~ 11

Twostep Clustering of Environmental Indicator Survey Data

Hee-Chang Park[1)]

Abstract

Data mining technique is used to find hidden knowledge by massive data, unexpectedly pattern, relation to new rule. The methods of data mining are decision tree, association rules, clustering, neural network and so on.

Clustering is the process of grouping the data into clusters so that objects within a cluster have high similarity in comparison to one another. It has been widely used in many applications, such that pattern analysis or recognition, data analysis, image processing, market research on off-line or on-line and so on. We analyze Gyeongnam social indicator survey data by 2001 using twostep clustering technique for environment information. The twostep clustering is classified as a partitional clustering method. We can apply these twostep clustering outputs to environmental preservation and improvement.

Keywords : Clustering, Data mining, Environment information, Twostep custering

1. 서론

환경정보화의 목적은 최신 정보기술을 이용하여 환경관련 정보를 생산·수집·가공·처리·유통 또는 활용함으로써 환경행정업무의 효율화를 도모함에 있다. 환경데이터에 대해 그 동안 환경관련분야에서는 자료의 수집과 기초적인 분석방법, 다변량 분석방법 등에 중점을 두고 연구가 활발히 이루어지고 있다(이상훈(1995), 이용우(1998), 정상용 등(1998), 최성우와 송형도(2000), 문상기와 우남칠(2001), 환경부(2001, 2003, 2004), 김정태 등(2003)). 그러나 데이터의 양이 기하급수적으로 증가하고 있는 오늘날 방대한 양의 데이터베이스(database : DB)에 내재되어 있는 유용한 정보를 탐색하여

1) Professor, Department of Statistics, Changwon National University, Changwon, Kyungnam, 641-773, Korea
E-mail : hcpark@changwon.ac.kr

의미 있는 지식을 발견하기 위한 연구의 필요성이 대두되고 있으며, 이를 위한 도구가 데이터마이닝(data mining)이다.

데이터마이닝이란 방대한 양의 데이터 속에서 쉽게 드러나지 않는 유용한 정보를 찾아내는 과정으로, 대용량(massive)의 관측 가능한 데이터를 기반으로 숨겨진 지식, 기대하지 못했던 패턴, 새로운 법칙과 관계를 발견하고 이를 바탕으로 의사결정 등을 위한 정보로 활용하고자 하는 것이다. 데이터마이닝 기법으로는 군집분석(cluster analysis), 연결 분석(link analysis), 판별 분석(discrimination analysis) 등과 같은 기존의 통계 분석과 연관성규칙(association rule), 의사결정나무(decision tree) 기법, 신경망모형(neural network) 등의 분석 기법이 있다.

한편, 최근에는 연속형 자료와 범주형 자료를 동시에 고려한 군집분석기법이 개발되고 있는데, 이들 중 가장 대표적인 기법이 Chiu 등(2001)에 의해 제안된 twostep 군집분석방법이다. 이 알고리즘은 개체(records)간 거리에 근거하여, 서로 가까운 개체들은 같은 군집으로 묶이고 서로 상당히 떨어진 개체들은 다른 군집에 속하도록 하는 방법으로 사용자가 군집의 수를 범위로 정하면 그 안에서 알고리즘이 최적 해를 찾는 방법이다. 본 논문에서는 군집분석 기법 중 하나인 twostep 군집분석을 이용하여 분석을 실시하고자 한다.

일반적인 설문 조사 분석에서는 관심대상이 되는 문항에 대하여 일차원적인 분석을 실시하거나 데이터의 속성을 파악하기 위하여 인구통계학적 문항을 이용하여 세분화된 정보를 추출한다. 그러나 본 논문에서는 2001년 조사된 경상남도 사회지표조사 자료의 환경관련 설문에 대하여 내재되어 있는 정보를 추출하기 위해 인구통계학적 속성뿐만 아니라 환경문항과 관련성이 있을 것으로 추정되는 구분문항들을 추출하여 twostep 군집분석 기법을 적용하였다. twostep 군집분석 적용 및 분석을 통하여 환경정보화를 이루고 이 정보를 바탕으로 환경개선대책 수립과 환경 정책 결정에 필요한 의사결정 지원 등 효율적인 환경행정의 수행과 환경 정책 수립에 기여할 수 있게 한다. 본 논문의 2절에서는 twostep 군집분석에 대하여 기술하고 3절에서는 자료 구축 및 자료 탐색에 대하여 기술하며 4절에서는 twostep 군집분석의 분석 단계에 대하여 기술한다. 5절에서는 twostep 군집분석을 이용한 자료 분석 결과를 기술한 후, 6절에서 결론을 맺는다.

2. twostep 군집분석

twostep 군집분석은 개체(records)간 거리에 근거하여, 서로 가까운 개체들은 같은 군집으로 묶이고 서로 상당히 떨어진 개체들은 다른 군집에 속하도록 하는 방법으로 사용자가 군집의 수를 범위로 정하면 그 안에서 알고리즘이 최적 해를 찾는 것으로서 몇 가지 중요한 변수들에서 군집 간 차이가 뚜렷할수록 좋은 군집화라고 할 수 있다.

k-평균 군집분석은 연속형 자료만 모형화가 가능하지만 twostep 군집분석은 연속형 자료뿐만 아니라 명목형 자료에도 적용할 수 있어 자료형태에 제약을 받지 않으며 먼저 최초의 데이터를 대상으로 몇 개의 하위 군집 또는 그룹을 생성하는 과정을 거치게 되어 대용량 데이터에 유용하게 사용될 수 있다.

twostep 군집분석은 군집화를 위하여 두 단계의 과정을 거친다. 제 1단계는 순차적으로 개체들을 묶는 과정이다. 1개 개체가 들어오면 거리에 기준하여 알고리즘은 새

개체를 기존의 군집에 병합시킬 것인지 아니면 새 군집으로 받을 것인지를 결정한다. 그러다가 군집 수가 사용자가 지정한 수보다 커지게 되면 거리 기준을 상향 조정하여 군집간 거리가 새 기준에 미달하는 군집들을 병합시켜 총 군집 수를 줄인다. 1단계를 마치면 다수의 예비 군집들이 형성되어 진다.

제 2단계에서는 예비 군집들을 계층적 군집화(hierachical clustering) 시키는 과정으로 유사한 군집들을 계층적으로 응집시켜 몇 개의 군집 해를 만들고 BIC(Bayesian Information Criterion)와 AIC(Akaike Information Criterion)같은 통계적 기준을 적용시켜 가장 좋은 군집 해를 찾아 출력한다. BIC와 AIC는 식 (2.1), 식 (2.2)와 같다.

$$BIC(j) = -2\sum_{j=1}^{j}\xi_j + m_j \log(N) \tag{2.1}$$

$$AIC(j) = -2\sum_{j=1}^{j}\xi_j + 2m_j \tag{2.2}$$

여기서

$$\xi_i = -N_i\left(\sum_{k=1}^{K^A}\frac{1}{2}log(\hat{\sigma_k}^2 + \hat{\sigma_{ik}}^2) + \sum_{k=1}^{K^B}\hat{E}_{ik}\right),$$

$$m_j = J\{2K^A + \sum_{k=1}^{K^B}(L_k - 1)\},$$

$$\hat{E}_{ik} = -\sum_{l=1}^{L_k}\frac{N_{ikl}}{N_i}\log\frac{N_{ikl}}{N_i},$$

K^A : 연속형 변수의 총 수,

K^B : 범주형 변수의 총 수,

L_k : k번째 범주형 변수에 대한 범주의 수

이다.

twostep 군집분석에서 연속형 변수에 대하여는 정규분포를, 범주형 변수에 대하여는 다항분포를 적용하여 우도함수(likelihood function)를 활용하게 된다. 2단계 군집화에서 제 1단계는 순차적으로 진행되기 때문에 개체들의 진입 순서에 의존적이다. 따라서 개체들의 입력자료 파일내 순서가 임의적이어야 한다. twostep 알고리즘은 <그림 1>과 같다.

<그림 1> twostep 알고리즘

3. 자료 구축

1) 자료의 구조

경남사회지표조사의 자료구조는 크게 일반사항(인구통계학적 문항)과 도민의식조사부문으로 나누어져 있다. 일반사항은 조사응답자의 연령, 성별, 학력, 가구주와의 관계, 결혼 유무, 직업으로 구성되어 있으며, 도민의식조사부문은 소득・소비, 보건・체육, 주택, 환경・교통, 사회, 정보화, 문화・여가의 총 7개 분야로 구성되어 있다.

2) 자료의 선정

자료의 선정에서는 환경관련 문항과 인구통계학 속성 관련문항, 집단구분문항을 추출하였다. 환경관련 문항으로는 지역의 상수도 오염도, 지역의 하수도 오염도, 지역의 소음진동 오염도, 지역의 악취오염도, 지역의 대기 오염도, 지역의 토양 오염도, 쓰레기 종량제 봉투의 사용만족도, 경남발전과 환경보존 선호도, 환경오염의 해결주체 문항을 추출하였고 인구통계학 속성 관련문항으로는 연령, 성별, 학력, 가구주와의 관계, 결혼유무, 직업, 조사지역, 설문지 작성자, 시군명, 월평균 소득, 월평균 저축, 거주 주택 소유형태, 거주 주거형태, 주택 사용방수, 주택면적을 추출하였으며 집단구분 문항으로 향후 2~3년 이후 가구소득, 의료시설과 서비스 만족도, 주관적 건강상태, 거주지 주택규모의 만족도, 거주지 일조통풍의 만족도, 거주지 상하수도 시설의 만족도, 주관적 사회계층, 자치단체의 업무 인지 방법, 자치단체의 주민여론 반영도, 시책추진의 우선적 처리사항을 추출하였다

3) 자료의 정제

자료의 정제는 데이터의 충실도와 분석 결과의 정확성을 기하기 위해 실시되는 것으로 무응답 또는 조사를 하지 않은 부분 즉, 결측치를 분석대상에서 제외시킨 후 새로운 DB를 구축하는 과정이다. 경남 사회 지표 조사에서는 조사문항에 대하여 응답자가 무응답을 하거나 조사되지 않은 응답에 대해서는 결측치로 처리한 다음 정제된

DB를 구축하였다.

4) 자료의 변환

사회지표조사의 자료에 대하여 환경 관련 문항에 대하여 분석 결과의 해석을 용이하게 하기 위해 21번 문항(쓰레기 종량제 봉투의 사용 만족도)을 1. 아주불만 2. 불만 3. 보통 4. 만족 5. 아주만족으로 변환하였다. 구분문항에 대하여 2번 문항(향후 2~3년 이후 가구소득)을 1. 많이 나빠짐 2. 조금 나빠짐 3. 전과동일 4. 조금 좋아짐 5. 많이 좋아짐으로 변환하였으며, 3번 문항(의료시설과 서비스 만족도), 5번 문항(주관적 건강상태), 22번 문항(주관적 사회계층), 28번 문항(자치단체의 주민여론 반영도)을 2번 문항과 마찬가지로 보기문항을 1→5, 2→4, 4→2, 5→1로 변환하였다.

4. 모형 구축 및 분석 단계

본 논문에서는 twostep 군집분석 모형의 구축을 위해 SPSS의 Clementine 8.0을 사용하였으며, twostep 군집분석 모형 구축 및 분석 단계는 다음과 같다.

[단계 1] 변수 선정 : 군집화에 사용할 변수를 선정하고 속성분석에 사용할 변수를 선정한다. 9개의 환경 관련 문항에 대하여 twostep 군집분석을 실행하고 인구통계학 문항과 구분 문항으로 각 군집에 대한 속성을 파악하기 위한 변수로 사용한다.

[단계 2] 군집의 수 결정 : twostep 군집분석은 사용자가 군집의 수를 범위로 정하면 그 안에서 알고리즘이 최적 군집 수를 찾는 것으로 최소 군집수와 최대 군집수를 결정한다. 본 논문에서는 최대 군집수를 14개, 최소 군집수를 2개로 지정한다.

[단계 3] 모형 생성 : twostep 군집분석 모형을 생성한다. 모형 생성 결과 5개의 군집으로 모형이 생성되었고 각 군집에 대한 해석 결과 군집의 특성이 명확하게 구분되었다.

[단계 4] 군집별 속성 분석 : 각 군집에 대한 인구통계학 문항과 구분 문항에 대하여 차이가 있는지 분석하기 위하여 명목형 자료에는 교차표에 의한 카이제곱 검정을 실시하였으며, 연속형 자료에 대해서는 분산분석을 실시한 후 사후 검정으로 Scheffe 검정법을 사용하여 군집별 속성을 파악한다.

5. 군집 분석 결과 및 속성 분석

박희창과 조광현(2005)은 2001년 경남사회지표자료에서 환경관련 문항 중 연속형 문항에 대하여 k-평균 군집분석을 실시하였다. 환경 관련 문항 중 연속형 문항인 상수도 오염도, 하수도 오염도, 소음진동 오염도, 악취 오염도, 대기 오염도, 토양 오염도, 쓰레기 종량제 봉투 만족도의 문항에 대하여 k-평균 군집분석을 실시한 결과 3개의 군집으로 k-평균 군집분석을 실시하였을 때가 군집의 특성이 명확하게 구분되었으

며 군집 1의 집단은 환경 문항에 대한 응답이 긍정적인 집단으로 분류되고 군집 3은 환경 문항에 대한 응답이 보통으로 군집 2의 집단은 환경 문항에 대한 응답이 부정적인 집단으로 분류되었다. 본 논문에서는 연속형 문항뿐만 아니라 경남발전과 환경보존 선호도, 환경오염의 해결주체의 명목형 문항을 추가하여 twostep 군집분석을 실시하였다. twostep 군집분석의 결과는 <표 1>과 같다.

k-평균 군집분석에서는 3개의 군집화가 되었으나 <표 1>에서 보는 바와 같이 twostep 군집분석에서는 5개의 군집으로 군집화가 되었다. k-평균 군집분석에서는 환경관련 문항에 대한 보통의 응답을 한 하나의 군집이 twostep 군집분석에서는 명목형 변수에 의해 3개의 군집으로 세분화 된 것을 알 수 있다. 이는 k-평균 군집분석보다 명목형 변수를 사용하는 twostep 기법이 군집화를 더 세분하게 해준다는 것을 의미한다. 이를 자세하게 살펴보면 군집 1 집단은 오염도에 대하여 보통의 응답을 하였고 경남발전과 환경보존 선호도에 대해서는 환경보존 하에 경제개발의 빈도수가 가장 높으며 환경오염의 해결주체에 대해서는 주민의 빈도수가 가장 높다. 군집 2 집단은 오염도에 대해서는 긍정적으로 응답을 하였고 경남발전과 환경보존 선호도에 대해서는 환경보존의 빈도수가 가장 높으며 환경오염의 해결주체에 대해서는 공동의 빈도수가 가장 높다. 군집 3 집단은 오염도에 대해서는 부정적으로 응답을 하였고 경남발전과 환경보존 선호도에 대해서는 환경보존의 빈도수가 가장 높으며 환경오염의 해결주체에 대해서는 공동의 빈도수가 가장 높다. 군집 4 집단은 오염도에 대해서는 보통의 응답을 하였고 경남발전과 환경보존 선호도에 대해서는 환경보존 하에 경제개발의 빈도수가 가장 높으며 환경오염의 해결주체에 대해서는 공동의 빈도수가 가장 높다. 마지막으로 군집 5 집단은 오염도에 대해서는 보통의 응답을 하였고 경남발전과 환경보존 선호도에 대해서는 단기적 경제개발 장기적 환경보존의 빈도수가 가장 높으며 환경오염의 해결주체에 대해서는 공동의 빈도수가 가장 높다.

<표 1> twostep 군집분석 결과

항목 \ 군집	군집 1	군집 2	군집 3	군집 4	군집 5
상수도 오염도	3.122	3.987	2.768	3.093	3.105
하수도 오염도	2.972	3.872	2.635	2.943	2.966
소음진동 오염도	2.915	3.955	2.537	2.904	2.959
악취 오염도	2.973	3.973	2.605	3.027	3.025
대기 오염도	3.219	4.296	2.733	3.231	3.218
토양 오염도	3.294	4.234	2.825	3.284	3.256
쓰레기 종량제 봉투 만족도	2.584	2.605	2.396	2.494	2.52
경남발전과 환경보존 선호도	1 : 1.6% 2 : 39.2% 3 : 56.3% 4 : 3.0%	1 : 32.9% 2 : 8.0% 3 : 3.8% 4 : 55.3%	1 : 34.4% 2 : 2.5% 3 : 0.6% 4 : 62.5%	1 : 0.3% 2 : 0.0% 3 : 99.7% 4 : 0.0%	1 : 0.0% 2 : 100% 3 : 0.0% 4 : 0.0%
환경오염의 해결주체	1 : 23.2% 2 : 62.0% 3 : 14.8% 4 : 0.1%	1 : 13.1% 2 : 26.6% 3 : 3.5% 4 : 56.9%	1 : 13.6% 2 : 27.8% 3 : 0.4% 4 : 58.2%	1 : 0.1% 2 : 0.0% 3 : 0.0% 4 : 99.9%	1 : 0.1% 2 : 0.1% 3 : 0.0% 4 : 99.8%
군집 레코드 수	2121	1079	1751	3137	1799

다음은 각 군집에 대한 인구통계학 문항과 구분 문항에 대하여 차이가 있는지 분석하기 위하여 명목형 자료에는 교차표에 의한 카이제곱 검정을 실시하였고 연속형 자료에 대해서는 분산분석을 실시하고 사후 검정으로 Scheffe 검정법을 사용하였다.

교차분석 시 특정 범주를 파악하기 위하여 잔차의 분석을 실시하였다. 잔차란 실측도수와 기대도수의 차를 말하는 것으로 잔차가 큰 곳은 특정적인 곳으로 간주한다. 그러나 잔차 그대로는 대소의 절대적인 평가가 불가능하여 식 (5.1)과 같이 정의되는 수정된 잔차를 이용하여 분석한다.

$$d_{ij} = e_{ij}/\sqrt{V_{ij}} \tag{5.1}$$

여기서 $e_{ij} = (f_{ij} - t_{ij})/\sqrt{t_{ij}}$(표준화잔차)이고, $V_{ij} = (1 - n_{i.}/N)(1 - n_{.j}/N)$(표준화잔차의 분산)이다. 수정된 잔차 d_{ij}는 평균이 0, 표준편차가 1인 정규분포를 근사적으로 따르며 $|d_{ij}|$ 가 2이상인 곳은 특정적인 곳으로 간주한다(노형진(2001)).

1) 교차표에 의한 분석

(1) 각 군집과 주거별 주택 소유형태 교차표

<표 2>에서 보는 바와 같이 군집 2의 집단은 주거별 주택소유형태에서 자가의 응답 비율이 높은 것으로 나타났고 군집 3의 집단은 전세와 월세의 응답비율이 높은 것으로 나타났다. 군집 4의 집단은 보증부월세와 기타의 응답비율이 높은 것으로 나타났으며, 군집 5의 집단은 전세의 응답비율이 높은 것으로 나타났다.

<표 2> 주거별 주택 소유형태에 대한 군집간 비교

			주거별 주택소유 형태				
			자가	전세	보증부월세	월세(사글세)	기타(무상 등)
twostep 5군집	군집 1	빈도	1425	433	103	87	73
		수정된 잔차	.2	.6	-.1	.3	-1.8
	군집 2	빈도	842	126	29	36	46
		수정된 잔차	8.2	-7.2	-3.6	-1.2	.3
	군집 3	빈도	1109	381	95	113	53
		수정된 잔차	-3.6	2.0	1.1	5.8	-2.5
	군집 4	빈도	2069	642	174	101	151
		수정된 잔차	-1.5	.8	2.0	-2.7	2.4
	군집 5	빈도	1179	394	83	59	84
		수정된 잔차	-1.5	2.2	-.6	-1.7	1.3
전체		빈도	6624	1976	484	396	407

(2) 각 군집과 자치단체의 업무인지 방법의 교차표

군집 1의 집단은 자치단체의 업무인지 방법에서 홍보자료의 응답 비율이 높은 것으로 나타났고 군집 2의 집단은 반상회 및 통반장, 주의사람의 응답비율이 높은 것으로 나타났다. 군집 3의 집단은 기타의 응답비율이 높은 것으로 나타났고 군집 4의 집단

과 군집 5의 집단은 대중매체의 응답비율이 높은 것으로 나타났다.

<표 3> 자치단체의 업무인지 방법에 대한 군집간 비교

			자치단체의 업무 인지방법					
			대중매체	홍보자료	인터넷	반상회 및 통반장	주의사람	기타
twostep 5군집	군집 1	빈도	1332	224	52	179	287	47
		수정된 잔차	-2.4	2.9	-.2	.2	1.0	-.1
	군집 2	빈도	554	93	27	166	207	31
		수정된 잔차	-10.0	-.4	.0	8.9	6.5	1.4
	군집 3	빈도	1146	134	52	140	223	56
		수정된 잔차	.4	-2.1	1.4	-.6	-.2	2.9
	군집 4	빈도	2184	274	68	189	359	63
		수정된 잔차	6.5	-.5	-1.5	-5.7	-2.9	-1.1
	군집 5	빈도	1212	161	49	151	199	26
		수정된 잔차	2.3	.0	.6	.1	-2.6	-2.6
전체		빈도	6428	886	248	825	1275	223

(3) 각 군집과 결혼유무와의 교차표

군집 2의 집단은 결혼유무에서 사별의 응답비율이 높은 것으로 나타났고 군집 3의 집단은 이혼의 응답비율이 높은 것으로 나타났다. 군집 4의 집단은 유배우의 응답비율이 높은 것으로 나타났으며, 군집 5의 집단은 미혼의 응답비율이 높은 것으로 나타났다.

<표 4> 결혼유무에 대한 군집간 비교

			결혼유무			
			미혼	유배우	사별	이혼
twostep 5군집	군집 1	빈도	479	1433	175	34
		수정된 잔차	1.8	.1	-2.1	-1.4
	군집 2	빈도	180	670	200	27
		수정된 잔차	-3.8	-3.9	10.9	1.3
	군집 3	빈도	358	1179	167	47
		수정된 잔차	-.8	-.1	.1	2.4
	군집 4	빈도	661	2192	230	52
		수정된 잔차	-.1	3.6	-4.9	-1.5
	군집 5	빈도	412	1189	161	35
		수정된 잔차	2.0	-1.3	-.8	-.1
전체		빈도	2090	6663	933	195

(4) 각 군집과 시군과의 교차표

군집 1의 집단은 시군명에서 통영시, 사천시, 합천군의 응답비율이 높은 것으로 나타났고 군집 2의 집단은 진해시, 사천시, 밀양시, 거제시, 의령군, 창녕군, 하동군, 산청군, 함양군, 거창군의 응답비율이 높은 것으로 나타났다. 군집 3의 집단은 창원시, 마산시, 함안군의 응답비율이 높은 것으로 나타났고 군집 4의 집단은 진주시, 김해시, 양산시의 응답비율이 높은 것으로 나타났으며 군집 5의 집단은 김해시, 고성군, 거창군, 합천군의 응답비율이 높은 것으로 나타났다.

<표 5> 시군간에 대한 군집간 비교

			시군명																		
			창원시	마산시	진주시	진해시	통영시	사천시	김해시	밀양시	거제시	양산시	의령군	함안군	창녕군	고성군	남해군	하동군	산청군	함양군	거창군	합천군
twostep 5군집	군집1	빈도	355	277	195	104	141	102	235	66	127	126	29	47	51	50	38	36	25	21	35	61
		수정된 잔차	.6	-1.2	-3.1	1.6	5.3	2.0	.4	-2.4	1.0	-.7	.6	-.2	-.4	1.2	1.0	-1.2	-1.4	-2.2	-2.2	2.1
	군집2	빈도	49	68	69	60	54	62	44	113	111	45	34	30	67	22	21	74	59	37	40	20
		수정된 잔차	-11.1	-7.6	-5.2	2.2	.8	3.0	-7.6	11.4	7.2	-3.0	6.0	1.2	8.2	.0	1.1	12.0	11.2	5.5	3.3	-1.0
	군집3	빈도	376	312	192	87	53	31	201	49	68	99	14	89	28	24	23	28	12	18	22	25
		수정된 잔차	6.4	5.3	-.1	1.5	-3.4	-5.3	1.0	-2.9	-3.4	-1.2	-1.8	8.7	-2.7	-2.2	-.9	-1.4	-3.1	-1.8	-3.2	-2.6
	군집4	빈도	536	465	427	120	134	144	369	85	155	221	30	35	74	51	55	35	23	51	74	53
		수정된 잔차	1.4	1.9	5.6	-1.5	-.9	1.9	2.0	-4.6	-1.8	2.2	-1.7	-5.2	-.7	-2.0	1.1	-4.4	-4.3	.7	.4	-2.7
	군집5	빈도	300	247	209	53	67	60	222	86	88	128	15	23	29	55	17	27	31	22	54	66
		수정된 잔차	.4	-.2	.9	-3.1	-1.8	-1.7	2.3	1.8	-1.4	1.7	-1.7	-3.1	-2.7	3.4	-2.3	-1.7	.8	-1.1	2.3	4.4
전체		빈도	1616	1369	1092	424	449	399	1071	399	549	619	122	224	249	202	154	200	150	149	225	225

이외에도 학력에 대한 각 군집간의 비교결과, 군집 1의 집단과 군집 3의 집단은 고졸의 응답 비율이 높은 것으로 나타났고 군집 2의 집단은 무학과 초졸의 응답비율이 높은 것으로 나타났다. 군집 4의 집단은 전문대졸, 대학재학, 대졸, 대학원이상의 응답비율이 높은 것으로 나타났으며 군집 5의 집단은 대학재학과 대졸의 응답비율이 높은 것으로 나타났다. 또한 직업에 대해서는 군집 1의 집단은 고위임직원 및 관리자, 서비스종사자의 응답비율이 높은 것으로 나타났고 군집 2의 집단은 농업, 임업 및 어업 숙련 종사자, 무직의 응답비율이 높은 것으로 나타났다. 군집 3의 집단은 장치·기계 조작 및 조립 종사자, 무직의 응답비율이 높은 것으로 나타났고 군집 4의 집단은 전문가, 가정주부, 기타의 응답비율이 높은 것으로 나타났으며 군집 5의 집단은 사무종사자, 단순노무종사자, 학생의 응답비율이 높은 것으로 나타났다. 조사지역에 대해서는 군집 2의 집단은 농촌지역, 어촌지역의 응답비율이 높은 것으로 나타났고 군집 3의 집단은 상가지역, 주거지역, 공업지역의 응답비율이 높은 것으로 나타났으며 군집 4의 집단은 주거지역의 응답비율이 높은 것으로 나타났다. 마지막으로 주거하는 주택형태에 대해서는 군집 2의 집단은 거주하는 주택형태에서 단독주택의 응답 비율이 높은 것으로 나타났고, 군집 4의 집단은 아파트의 응답비율이 높은 것으로 나타났다.

2) 분산분석에 의한 분석

각 군집에 대하여 연속형 문항에 대한 분산 분석을 실시하였다. 분산분석 실시 결과 현재 주택의 사용 면적의 문항을 제외하고는(유의수준 0.05) 모든 문항이 각 군집에 대한 평균차가 유의하다는 결과가 도출되었고 각 군집간의 차이를 알아보기 위하여 사후 검정을 실시하였다(Scheffe 검정).

각 문항에 대한 분산 분석의 결과를 구체적으로 살펴보면 월평균 소득에 대한 분산분석 결과 군집 2의 집단이 월평균 소득이 낮게 나타나고 있으며 군집 4의 집단은 월평균 소득이 높게 나타나고 있다. 월평균 저축에 대해서는 군집 2의 집단이 월평균

저축이 낮게 나타나고 있으며 군집 3의 집단은 월평균 저축이 높게 나타나고 있다. 향후 2~3년 이후 가구 소득의 변화에 대한 분산분석 결과 군집 2의 집단이 부정적으로 나타났으며 군집 1의 집단과 군집 5의 집단은 긍정적으로 나타났다. 의료시설과 서비스 만족도에 대한 분산분석 결과 군집 2의 집단은 긍정적으로 나타나고 있다. 주관적 건강상태에 대한 분산분석 결과 군집 2의 집단은 나쁘다라고 나타났다. 거주지 주택규모의 만족도에 대한 분산분석 결과 군집 2의 집단은 긍정적으로 나타났으며 군집 3의 집단은 부정적으로 나타나고 있다. 거주지 일조 통풍의 만족도에 대한 분산분석 결과 군집 2의 집단은 긍정적으로 나타났으며 군집 3의 집단은 부정적으로 나타나고 있다. 거주지 상하수도 시설의 만족도에 대한 분산분석 결과 군집 2의 집단은 긍정적으로 나타났으며 군집 3의 집단은 부정적으로 나타나고 있다. 현재 주택의 사용방수에 대한 분산분석 결과 군집 3의 집단이 주택의 사용 방수가 적은 것으로 나타났다. 주관적 사회계층에 대한 분산분석 결과 군집 3의 집단이 주관적 사회계층이 낮다고 생각하는 것으로 나타났다. 자치단체의 주민여론 반영도에 대한 분산분석 결과 군집 2의 집단이 반영도가 높다고 나타나고 있고 군집 3의 집단이 반영도가 낮다고 나타나고 있다. 연령에 대한 분산분석 결과 군집 2의 집단이 연령이 높은 것으로 나타났다.

6. 결론

본 논문에서는 2001년 조사된 경상남도 사회지표조사 자료의 환경관련 설문에 대하여 자료의 구축 및 탐색을 통하여 일차원적인 정보를 얻을 수 있었고 각 문항에 대하여 전반적인 속성을 파악할 수 있었다. 그러나 이 정보만으로는 데이터 속에 내제되어 있는 정보를 추출하기에는 부족하였다. 환경관련 문항에 대하여 인구통계학적 문항뿐만 아니라 환경문항과 관련성이 있을 것으로 추정되는 구분문항들을 추출하여 twostep 군집분석에 적용, 환경관련 문항에 대하여 비슷한 속성을 가지는 집단으로 구분하였고 각 집단간의 다양한 속성의 차이를 알아 볼 수 있었으며, 쉽게 드러나지 않는 유용한 정보를 추출할 수 있었다. 또한 twostep 군집분석은 연속형 자료뿐만 아니라 명목형 자료를 사용하여 군집화를 실시할 수 있어 k-평균 군집분석에 의한 군집결과보다 더욱 세분화된 군집 결과를 도출할 수 있었다. 향후 이 정보를 바탕으로 다양한 속성들 간의 분석을 실시, 환경 정보화를 통하여 환경개선대책 수립과 환경 정책 결정에 필요한 의사결정 지원 등 효율적인 환경행정의 수행과 환경 정책 수립에 기여할 수 있을 것이다.

참고문헌

1. 김정태, 정진도, 김광석(2003). 여름철 충청남도 서북부 지역에서의 대기오염물질 농도 분포특성에 관한 연구, 대한환경공학회 2003 춘계학술발표회 논문집, 1326-1328.
2. 노형진(2001). 한글 SPSS 10.0에 의한 조사방법 및 통계분석, 형설출판사.
3. 문상기, 우남칠(2001). 통계분석을 이용한 지하수위 변동 특성 분류, 한국지하수토양환경학회 01 추계학술발표회논문집, 2001권, 155-159.

4. 박희창, 조광현 (2005). 사회지표조사 자료의 K-평균 군집분석, *Journal of the Korean Data Analysis Society*, 제 7권 제 2호, 465-476.
5. 이상훈(1995). 수질자료의 추세분석을 위한 비모수적 통계검정에 관한 연구, 환경영향평가, 제4권 제2호, 93-103.
6. 이용우(1998). 폐기물 배출량의 지역간 차이에 관한 분석, 대한지리학회 33권 2호, 209-224.
7. 정상용, 강동환, 심병완(1998). 부산지역 지하수의 수질오염 특성, 한국지하수토양환경학회 98 공동심포지엄 및 추계학술발표회 논문집, 1998권, 86-92.
8. 최성우, 송형도(2000). 다변량 통계분석법을 이용한 대구지역 부유분진의 오염원 기여도 추정, 한국환경위생학회지, 제26권 제4호, 1-8.
9. 환경부(2001). 전국폐기물통계조사.
10. 환경부(2003). 환경통계연감.
11. 환경부(2004). 환경백서.
12. Chiu, T., Fang, D., Chen, J., Wang, Y., and Jeris, C. (2001). A Robust and Scalable Clustering Algorithm for Mixed Type Attributes in Large Database Environment. *In Preceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 263-268.

[2005년 12월 접수, 2006년 2월 채택]