평균 연결법과 K-means 혼합 클러스터링 기법을 이용한 공시지가 유사가격권역의 설정

이성규*, 홍성언*, 박수홍**

* 인하대학교 공과대학 지리정보공학과 대학원
** 인하대학교 공과대학 지리정보공학과 조교수

Yi Seong-kyu, Hong Sung-Eon, Park Soo-Hong

비교표준지를 이용하여 개별공시지가를 산정하는 우리나라 제도 하에서 가장 중요한 문제는 개별필지 주변의 표준지 중에서 어떤 표준지를 선택·이용하여 지가를 산정해야 하는가이다. 그러나 지침상에서는 비교표준지 선정시 매우 중요한 요인으로 작용하고 있는 유사가격권에 대하여 수치적인 기준이 아닌 모호한 개념상으로 규정하고 있어 비교표준지 선정에 있어 객관성과 정확성이 결여되고 있다.

본 연구에서는 현행 개념상으로만 규정하고 있는 유사가격권에 대하여 평균 연결법과 K-means 혼합 클러스터링 기법을 이용하여 유사가격권역을 정확하고 객관적으로 설정한다. 그리고 실제 사례지역을 선정하여 적용하여 봄으로써 방법론의 활용가능성과 타당성을 제시하고자 한다.

Ⅰ. 서론

우리나라는 1991년도부터 개별공시지가의 자동 산정이 일부지역에서 시범적으로 운용되다, 1996년도부터는 전국으로 확대·적용되어 실시되고 있다. 현재는 자동화 산정 프로그램인 ALPA(Automatic Land Price Appraisal System)에 의해 개별공시지가를 산정하고 있다. ALPA 시스템은 지가를 자동으로 산출한다는 장점을 가지고 있지만 최종 지가의 계산만이 가능하기 때문에 지가 산정을 위한 거의 대부분의 과정이 지가담당 공무원들에 의해 수작업으로 이루어지고 있다. (홍길순, 1998; 박정호, 1999).

수작업으로 인한 문제점은 여러 과정에서 나타나고 있으나 특히 우리나라와 같이 비교표준지를 이용하여 개별공시지가를 산정하는 제도 하에서 가장 중요한 문제는 개별필지 주변의 표준지 중에서 어떤 표준지를 선택·이용하여 지가를 산정해야 하는가가 가장 중요한 문제이므로 이 과정에서의 문제 발생을 최소화 하는 것이 가장 중요하다.

현행 지침상에서는 비교표준지 선정시 용도지역, 유사가격권, 토지이용상황, 도로접면 등을 고려하도록 되어있다. 유사가격권을 제외한 요소들은 전산 코드화가 되어 있어 용이하게 개별필지의 특성을 조사할 수 있다. 그러나 유사가격권의 경우는 비교표준지 선정시 지침상에서 상

당히 중요한 요소로 취급되고 있으면서도 어느 정도의 범위를 유사가격권으로 설정할 수 있는가에 대해서는 개념상으로만 모호하게 규정하고 있다(건설교통부, 2002). 따라서 정확한 비교표준지의 선정과 향후 지가산정 전 과정을 자동화하기 위해서는 객관적이고 수치적인 기준 적용을 통한 유사가격권의 설정이 필요하다.

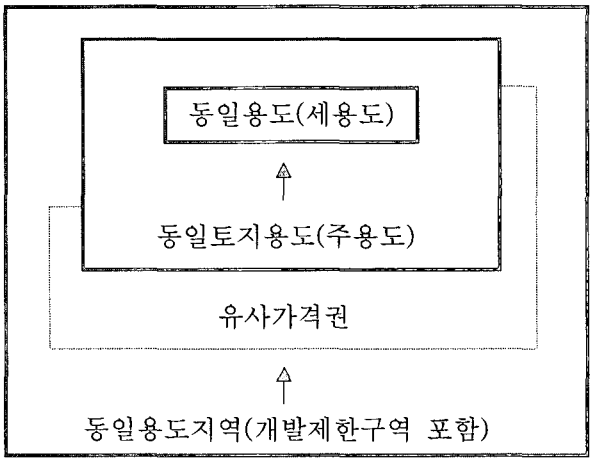
유사가격권의 설정에 관한 연구는 다양하지 못하며 다만, 비교표준지 선정 자동 선정 과정에서 유사가격 요소를 표준지와 개별지간의 지가차이로 유사가격권의 개념을 적용시키거나(박수홍·홍성언, 2003), 또는 유사가격을 도로접면 조건 등에 포함시켜 설정권을 해석한 선행연구가 있었다(박성규, 1999).

본 연구에서는 현행 개념상으로만 규정하고 있는 유사가격권에 대하여 평균 연결법(average linkage method)과 K-means 혼합 클러스터링 기법을 이용하여 유사가격권역을 정확하고 객관적으로 설정한다. 그리고 실제 사례지역을 선정하여 적용하여 봄으로써 방법론의 활용가능성과 타당성을 제시하고자 한다. 구체적으로, 연구에서는 군집화 인자로 당해 년도 개별공시지가와 필지간 공간상의 거리를 이용하였다. 그리고 계층적인 클러스터링 기법(average linkage method)을 이용하여 최적의 군집수를 객관적으로 설정하였고, 이렇게 설정된 군집수를 비계층적인 클러스터링 기법(K-means)에 적용하여 유사가격권을 설정하였다. 그리고 산출된 군집별 클러스터링 결과에 대하여 다양한 통계 기법을 적용하여 유사가격권 설정을 위한 최적의 군집수를 제시하고자 한다.

II. 유사가격권과 혼합 클러스터링 기법

1. 유사가격권의 개념

비교표준지의 선정 기준은 그림 1과 같이 조사대상 토지가 일반토지인 경우, 조사대상 토지와 동일한 용도지역(개발제한구역 포함)안에 있는 유사가격권의 표준지 중에서 토지이용상황(주용도)이 같은 표준지를 선정하도록 되어 있다. 만일, 동일한 용도지역 내 토지이용상황이 같은 유사가격권의 표준지가 없는 경우에는 토지이용상황이 다르더라도 조사대상 필지 인근의 토지이용상황을 감안하여 유사가격권의 표준지를 선정하여야 한다고 규정하고 있다(건설교통부, 2003).



<그림 1> 비교표준지 선정 개념도

이렇게 현행 지침상에서는 비교표준지 선정시 중요 요인으로 유사가격권이라는 기준을 규정하고 있다. 그러나 유사가격권의 가격적 편차, 범위 등에 대해서는 수치적이고 구체적인 기준으로 제시하지 못하고 있다. 단지 땅값의 형성요인(도로조건, 건축규제, 주변여건 등)이 비슷하여 유사한 가격대를 형성하는 지역적 범위를 말한다고 개념적으로 정의하고 있다. 이렇게 모호한 유사가격권의 규정으로 비교표준지 선정에 있어 객관성과 합리성이 결여되고 있으며, 비교표준지 선정 자동화에 문제점으로 지적되고 있다.

선행 비교표준지 자동 선정에 관한 연구에서도 유사가격권의 모호성으로 인하여 비교표준지를 자동화된 방법론으로 선정할 경우 적정 범위를 객관적이고 수치적인 기준으로 정립하기 곤란하여 자동화 구현에 한계성을 지적하였다. 그리고 현행 수작업에 의한 비교표준지 선정과 자동 선정 결과가 정확하게 일치하지 않는 주요원인 중의 하나가 바로 유사가격권의 정의·설정의 문제로 지적하고 있다(박수홍·홍성언, 2003).

이렇듯 유사가격권의 중요성에 비해 그에 대한 명확한 수치적인 기준을 제시하지 못하고 있어 지가 산정 관련 자동화에 있어 객관적인 적용의 어려움이 있다. 그러므로 현행 관련 법률이나 지침에서 기술하고 있는 유사가격권에 관한 범위를 보다 명확히 할 필요성이 있다.

2. 혼합 클러스터링 기법

클러스터링 분석의 방법은 각 측정치 사이의 유사성의 척도로써 무엇을 사용하느냐에 따라 여러 가지로 분류될 수 있다. 유사성의 척도로 사용되는 것에는 각 측정치 사이의 상관계수와 거리가 있다. 이중에 많은 분석법에서 사용하는 것이 거리에 의한 것이고, 거리를 이용하는 방법에는 Mahalanobis Distance, Minkowski, Euclidean Distance 등이 있다(Kachigan, 1986; 송문섭·조신섭, 1997). 그리고 이 유사성 척도에 기초해 집단내의 변량에 대한 집단간의 변량을 최대화시키는 방법(algorithm)에 따라 계층적(hierarchical)방법과 비계층적(Non-hierarchical or disjoint or K-means)방법이 있다(Hair·Anderson·Tatham, 1987).

객관적인 기준에 의해 유사가격권을 설정하기 위해 이 연구에서 사용되어진 클러스터링 분석 방법은 비계층적 군집 분석 방법의 하나인 K-means 클러스터링 분석(K-means cluster analysis)이다. 이 방법은 계층적 군집 분석 방법에 비하여 계산 속도가 빠르고 대량의 자료에서 군집을 발견하는데 상당히 효과적이다(이근수·김삼근, 1991; Brain, 1993).

이 기법은 MacQueen에 의하여 제안된 알고리듬으로서, 우선 패턴을 k개의 군집으로 나눈 후 군집에 포함되어 있는 패턴들의 평균을 클러스터의 중심값으로 계산한다. 그리고 이 중심값과 각 패턴과의 거리를 계산한 후 가장 거리가 가까운 클러스터에 패턴을 포함시키는 방법으로 조건은 다음의 식과 같다(조형기·민준영, 1996; 김윤식·모경주, 2000).

$$x_i \in c_j \quad ||x_{i-z_j}||^2 < ||x_{i-z_k}||^2$$

여기서, $1 \le i \le N$, $1 \le k \le c$, $j \ne k$ 이며, N은 패턴 수, c는 군집 수, z는 군집의 중심값을 나타낸다.

일반적으로 이 기법은 사례(자료)의 수가 수백 개를 넘을 경우에는 K-means 클러스터링 분석(K-means cluster analysis)을 사용하는 것이 바람직 하다(정충영·최이규, 1998).

이 기법은 군집의 수 k와 초기 군집 중심에 따라 그 결과가 달라지기 때문에 초기 군집수 및 군집 중심을 결정하는 것이 중요한데 이는 주어진 패턴에서 처음 k 개의 군집수를 추출하여 군집의 중심값을 주는 방법과 임의로 k개를 추출하여 중심값을 주는 방법이 있다. 후자의 경우 사전에 주어진 군집수 k가 원 데이터 구조에 적합하지 않거나, 군집의 개체 분류시 처음 선정한 군집 중심점(seed)들의 영향을 많이 받으므로 부적절한 위치에 군집 중심이 위치한 경우 좋은 군집화 결과를 얻기 힘들다. 대부분의 K-means 클러스터링 분석은 최적 군집수 결정을 위해서 다양한 군집수에 대한 반복적인 실험이 요구되므로 연구에서는 이러한 군집수 결정에 있어 계층적인 클러스터링 기법을 혼합 이용하여 효용성을 높였다. 즉, 비계층 군집분석의 경우 군집수를 사용자가 사전에 정의하여 주어야 하기 때문에 군집수의 결정에 있어 주관성의 개입 소지가 있고, 군집수를 잘못 설정함으로써 클러스터링의 정확도가 낮아 질 수 있다. 따라서 연구에서는 객관적이고 정확한 유사가격권의 설정을 위해 계층적 군집분석 기법 중 평균 연결법을 적용하여 먼저 적정 군집수를 산출하고, 이를 다시 비계층 분석기법 즉, K-means 클러스터링 기법을 이용하여 유사가격권을 설정하는 혼합 클러스터링 기법을 이용하였다.

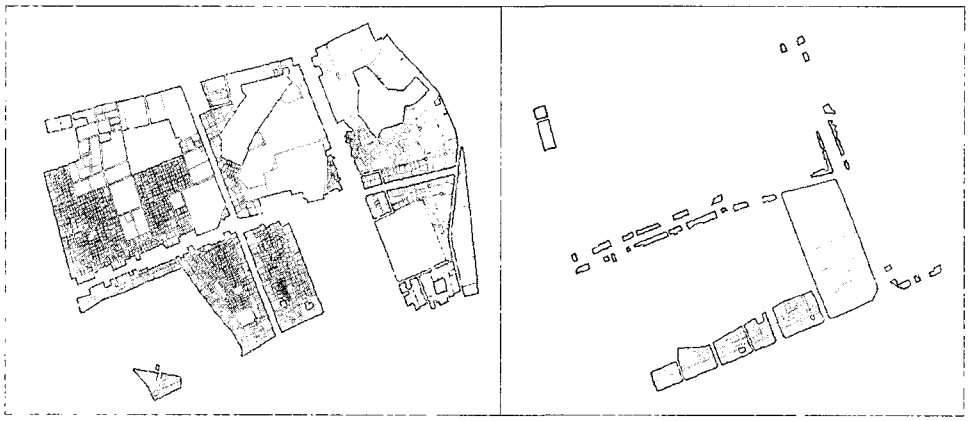
Ⅲ. 실험 및 분석

1. 실험

평균 연결법과 K-means 혼합 클러스터링 기법을 이용하여 실제 유사가격권을 설정하고자 서울시 강남구 삼성동 일부 지역(일반주거지역, 일반 상업지역)을 연구지역으로 선정하였다. 데이터는 강남구에서 구축하여 놓은 토지특성 도면을 이용하였다. 데이터 구축시 특수토지나 공공용지의 경우는 일반토지와 비교표준지 선정하는 기준이 상이하기 때문에 이를 포함할 경우 군집화에 오류가 발생할 수 있기 때문에 이는 배제하였다(그림 2). 클러스터링을 위한 도구로

SAS Institute 社의 통계프로그램인 SAS(Statistical analysis system)를 이용하였다.

연구에서는 군집분석 수행 전에 두 가지 전처리(Preprocessing)과정을 수행하였다. 첫째, 유사 가격권 설정 인자로 지가와 거리(X, Y) 두 가지 인자를 적용하였는데 두 인자들의 분포 형태가 동일하지 않기 때문에 군집분석 수행 전에 변수의 표준화 과정을 수행하였다(mean=0, std=1 의 정규분포 형태로 변환). 둘째, 군집형성 과정에 이상치(outlier or noise)가 다수 존재할 경우 군집이 부적절하게 형성되기 때문에 군집 분석 수행 전에 이상치를 제거해 주었다. 이상치 제거를 위해 다수의 군집수를 적용하여 군집이 형성될 때 빈도가 낮은 데이터 들이 군집을 형성하는 경우 이를 이상치로 보고 제거 하였다. 연구에서 결측치(missing value)에 의한 영향은 군집분석 전 무손실 데이터임을 확인 하였으므로 배제 하였다.



<그림 2> 연구대상지역
(좌 : 강남구 삼성동 일반주거지, 우 : 강남구 삼성동 일반주거지)

군집형성 과정에서 개체가 군집으로 묶이면 개체와 새로 만들어진 군집과의 유사성(similarity)을 계산하여 최적의 군집수를 계산하게 되는데 연구에서는 중심 연결법과 평균 연결법 및 워드 기법을 적용한 결과 중심 연결법과 평균 연결법 적용 후 형성된 군집모형이 비슷하며 군집 개수 설정을 위한 통계치 정보 해석이 워드 기법의 경우 적절치 못한 것으로 산출되어 평균 연결법과 중심 연결법 중에 평균 연결법을 적용하였다. 그리고 입체군집기준(CCC; Cubic Clustering Criterion), PST2(pseudo Hotel ling's T^2) 등의 통계치 분석을 통하여 최적의 군집수를 결정하였다. 이렇게 최적의 군집수를 결정한 후 K-means 클러스터링 기법을 이용하여 유사가격권을 설정하였다.

2. 결과 분석

표 1은 연구지역의 일반 상업지역과 일반 주거지역에 대하여 평균 연결법(average linkage)에 의한 CCC 값이 국부적 최고점을 보이며 PST2 값을 만족(주변보다 상대적으로 값이 큰 군집수보다 하나 더 많은 군집수)하는 5개의 군집수를 선정하였다. 표 2는 이렇게 산출된 최적의 군집수에 대해 K-means 클러스터링 기법을 이용하여 유사가격권을 설정한 결과이다. 적정 군집의 개수는 입체군집기준(CCC)의 값이 국부적 최대값(local peak)을 보이는 곳을 찾는 것이다. 일반 상업지역의 경우 군집수가 17개, 일반 주거지역의 경우 군집수가 42개 일 때 그 이후

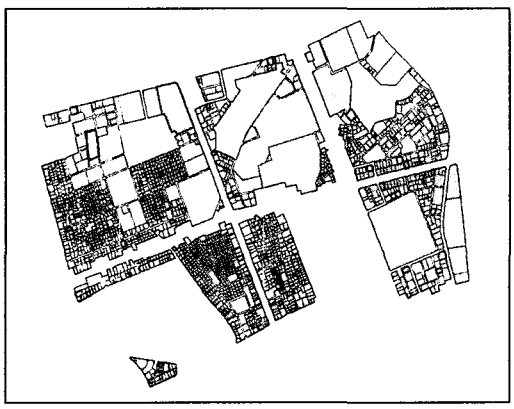
로 꾸준히 CCC 값이 증가하고는 있지만 일정한 수준으로 증가-감소하는 패턴을 보이므로 그 이상의 군집수는 의미가 없는 것으로 분석된다. 즉 이러한 상태를 국부적 최대값(local peak)으로 판단하였다. 이러한 패턴은 초기 군집중심점을 기준으로 반복을 수십 번 거쳐 나온 K-means 분석 결과에 나타난 CCC 값에서도 명확히 드러난다. 따라서 최적 군집수를 일반 상업지역의 경우 17개, 일반 주거지역의 경우 42개로 결정하였다. 물론 이러한 결과는 클러스터링에 관한 정확도 측면에 기반하여 유사가격권을 설정한 것으로 실질적인 군집수 결정을 위해서는 군집화 권역별 현행 비교표준지와의 일치율 등의 분석이 요구된다. 그림 3은 일반주거지역에 대하여 군집수를 42개로 설정하였을 경우의 유사가격권 구획결과이고, 그림 4는 일반상업지역에 대하여 군집수를 17개로 설정하였을 경우 유사가격권 구획결과를 나타낸 것이다.

<표 1> 평균연결법에 의한 최적의 군집 수 산출 결과

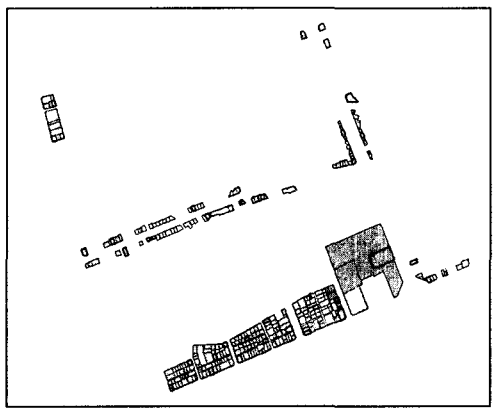
구 분	군집 수	CCC	PST2
일반 상업지역	7	23.7	90.0
	11	29.3	29.5
	17	42.6	17.1
	28	47.2	18.7
	34	52.2	20.5
일반 주거지역	13	5.17	40.8
	22	5.84	18.3
	27	17.4	96.6
	36	27.7	20.7
	42	32.5	3.1

<표 2> K-means 기법을 적용한 유사가격권 설정 결과

구 분	군집수	CCC	PSF
일반 상업지역	7	27.258	422.49
	11	32.110	482.64
	17	44.300	741.72
	28	48.696	856.59
	34	53.701	1042.75
일반 주거지역	13	20.557	1066.72
	22	28.587	1072.17
	27	32.228	1083.55
	36	35.078	1057.69
	42	44.742	1202.82



<그림 3> 일반주거지역 군집화 결과(k=42)



<그림 4> 일반상업지역 군집화 결과(k=17)

Ⅳ. 결 론

본 연구에서는 현행 모호한 규정으로 많은 문제점을 발생시키고 있는 유사가격권 설정의 문제를 해결하고자 평균 연결법과 K-means 혼합 클러스터링 기법을 이용하여 유사가격권을 설정해본 결과 다음과 같은 결론을 얻을 수 있었다. 먼저, 가장 객관적이고 정확하게 유사가격권을 설정하기 위해 평균 연결 기법을 이용하여 최적의 군집 수를 결정하고, 이러한 군집 수들을 기준으로 K-means 클러스터링 기법을 이용하여 실제 유사가격권을 구획하고 이에 대한 통계치 검증을 통하여 최적의 군집 수를 결정하여 보았다. 즉, 입체군집기준 값을 분석하여 일반 상업지역의 경우 군집수가 17개, 일반 주거지역의 경우 군집수가 42개 일때가 가장 높은 정확도로 유사가격권이 설정됨을 제시할 수 있었다.

이러한 군집화 기법은 유사가격 권역별로 평균지가, 표준편차 등의 정보를 수치적으로 파악할 수 있으므로 비교표준지 선정에 있어 보다 객관적이고 정확성을 기할 수 있을 것이고, 향후 비교표준지 선정 자동화시 유사가격권을 정확하게 반영할 수 있을 것으로 기대된다. 그리고 주변지가에 비해 가격이 매우 고지가인 필지들은 군집화가 이루어지지 않거나 아주 작은 수로 군집화가 이루어지는 것을 분석할 수 있어 지역적으로 문제가 되고 있는 지가불균형 지역을 관리하거나 해결하는데 효율적으로 이용될 수 있을 것으로 사료된다.

향후 연구내용으로는 첫째 연구에서 설정한 유사가격권역에 대하여 권역별 현행 개별필지에 대한 비교표준지 일치율을 분석하여 방법론의 타당성을 제시할 것이다. 둘째 연구에서는 유사가격권을 구획하는데 있어 지가와 거리라는 두 가지 인자를 사용하였으나 이외에 영향을 줄 수 있는 다른 토지특성 인자에 대한 연구도 함께 병행할 것이다.

참고문헌

1) 건설교통부, 2003, 2003년도 적용 개별공시지가 조사·산정 지침.
2) 국토개발연구원, 1997, "공시지가의 균형성 제고 방안", 국토연 97-23
3) 구자훈·김성희, 1999, "GIS를 활용한 개별공시지가 산정 및 도로개설에 따른 토지보상비 산정 방법론", 한국 GIS 학회지, 제7권 제1호.
4) 김기형·전명식, 1991: SAS 군집 분석. 자유 아카데미, p.68.
5) 김성호·조성빈·백승익, 2002, "자료융합방법의 성과에 대체수준이 미치는 영향에 관한 연구 : 몬테카를로 시뮬레이션 접근방법", 경영과학, 제16권 제1호, pp.129-141.
6) 김성희·정병호·김재경, 2002, 의사결정 분석 및 응용, 서울:영지문화사.
7) 김윤식·모경주·윤인섭, 2000, "클러스터링 기법을 이용한 공정 데이터의 압축 저장 기법에 관한 연구", 한국가스학회지, 제4권 제4호, pp.58-64.
8) 나성호, 2002, "고객세분화를 위한 군집분석 기법 중 K-평균 군집분석과 코호넨 네트워크의 분류 성능에 관한 연구", 서울대학교 대학원 석사학위 논문, pp.4-14.
9) 박수홍·홍성언·김현석·김정엽, 2003, "공간 다기준 의사결정 방법을 이용한 개별공시지가 비교표준지 선정", 한국GIS학회지, 제11권 제1호, pp.1-11.
10) 박성규, 1999, "토지 평가의 자동화를 위한 GSIS의 적용에 관한 연구", 조선대학교 대학원 박사학위 논문.

11) 박정호, 1999, "공시지가제도에 관한 연구", 동의대학교 대학원 석사학위 논문.
12) 안종욱, 2000, "개별공시지가의 효율적인 산정을 위한 GIS 적용에 관한 연구", 건국대학교 행정대학원 석사학위, pp.40-43.
13) 이근수·김삼근, 1991, "군집화 알고리즘에 관한 고찰", 안성대학교 산업대학원 논문집, 제23권, pp.175-185.
14) 이효상, 2001, "개별공시지가산정을 위한 토지특성테이블 구축에 관한 연구", 명지대학교 산업대학원 석사학위 논문.
15) 원덕진·김상윤·김경익·민경덕, 2000, "황해상 해무 발생시의 기상 및 해양 요소의 특성 분석", 한국기상학회지, 제36권 제6호 pp.631-642.
16) 마상진, 1997, "고등학교 학생의 학습방식에 관한 군집분석", 서울대학교 농업교육 대학원 석사학위 논문.
17) 조형기·민준영·최종욱, 1996, "클러스터링 방법을 이용한 차종인식 모형", 한국정보처리학회 논문지, 제3권 제2호, pp.369-380.
18) 홍길순, 1998, "개별공시지가 제도의 발전방향에 관한 연구", 중앙대학교 대학원 석사학위 논문.
19) 허명회·이용구, 2003, "클레멘타인을 활용한 K-평균 군집화 결과의 재현성 평가", SPSS White Paper.
20) Brian S. Everitt, 1993, Cluster analysis, 3rd edit., Halsted Press, p.170.
21) MacQueen. J.B, 1967, "Some Methods for Classification and Analysis of Multivariate Observations," Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1, pp.281 -297.
22) Milligan. G.W. and Cooper. M.C, 1985, "An Examination of Procedures for Determining the Number of Clusters in a Data Set," Psychometrika, 50, pp.159 -179.
23) Milligan. G.W. and Cooper. M.C, 1987, "A Study of Variable Standardization," College of Administrative Science Working Paper Series, Columbus OH, The Ohio State University, pp.87 -63.
24) Pollard, D, 1981, "Strong Consistency of k -Means Clustering," Annals of Statistics, 9, pp.135 -140.
25) Sarle. W.S, 1983, Cubic Clustering Criterion, SAS Technical Report A-108, Cary, NC: SAS Institute Inc.