

평균연결법과 K-means 혼합클러스터링 기법을 이용한 공시지가 유사가격권역의 설정

이성규* · 홍성언** · 박수홍***

A Similar Price Zone Determination of Public Land Price Using a Hybrid Clustering Technique

Seong-Kyu Yi*, Sung-Eon Hong**, and Soo-Hong Park***

요약 : 공시지가 유사가격권의 경우는 비교표준지 선정시 지침상에서 상당히 중요한 요소로 취급되고 있으면서도 실제적으로는 범위와 적용에 관해서 모호한 개념상의 규정을 두고 있다. 이러한 원인에 기인하여 비교표준지의 자동 선정이나 공시지가의 자동 산정에 있어 많은 문제점으로 작용하고 있다. 따라서 신속하고 정확한 비교표준지의 선정과 향후 지가산정 일련의 과정을 전산화하기 위해서는 자동화된 방식으로 유사가격권을 구획할 수 있는 객관적이고 합리적인 방법론이 필요하다. 본 연구에서는 개별공시지가 산정시 유사가격권 설정의 문제를 해결하고자 평균연결법과 K-means 혼합클러스터링 기법을 활용하여 유사가격권을 설정하여보고 이에 대한 타당성을 제시하고자 하였다. 이를 위해 실제 사례지역을 선정하고 실험한 결과 유사가격 권역 별로 군집화가 가능하였고, 현행 유사가격 권역과 많은 차이를 보이지 않아 방법론의 타당성을 제시할 수 있었다.

주요어 : 유사가격권, 비교표준지, 평균연결법, K-means 클러스터링, 혼합클러스터링

Abstract : Even though the similar land price zone is very important element in the public land appraisal procedure, the concept is implicitly described and applied into the actual land appraisal system. This situation makes it worse when applying for the automatic selection of a comparative standard land parcel. In addition, the division of similar land price zones requires the objective and reasonable process for improving ALPAS(Automatic Land Price Appraisal System), which becomes an issue today. To solve the similar land price zone determination problem that is caused by the lack of objective numerical standard, this study proposed a similar land price zone determination method using a hybrid clustering technique. Results showed that this hybrid clustering method that applied into the test area could easily detect similar land price zones with considerable accuracy levels, which are verified with some test statistics and real comparative standard land parcels done by manually.

Key Words : Similar Land Price Zone, Comparative Standard Land Parcel, Average-link method, K-means clustering, Hybrid Clustering

* 인하대학교 지리정보공학과 석사과정(Master Student, Dept. of GeoInformatic Engineering, Inha University), pla20@hanmail.net

** 청주대학교 복지·토지정보학부 전임강사(Full-time Lecturer, Dept. of Social welfare - Land Information, Cheongju University), hongsu2005@hanmail.net

*** 인하대학교 지리정보공학과 부교수(Associate Professor, Dept. of GeoInformatic Engineering, Inha University), shpark@inha.ac.kr

1. 서론

일반적으로 지리화적인 공간현상이나 복잡한 공간 문제로부터 일반화된 법칙이나 과학적 해답을 얻는 과정은 매우 복잡한 공간이론을 적용한 분석단계를 요구한다. 또한 공간을 대상으로 하는 분석은 데이터 용량도 방대하며, 분석과정도 매우 복잡하다. GIS의 활용은 방대한 양의 공간데이터의 처리 및 복잡한 분석 작업 등을 가능하게 하여 고차원적인 공간분석을 가능하게 하였다. GIS의 공간분석기법은 공간데이터베이스 내에 들어있는 공간데이터와 속성데이터를 이용하여 현실세계에서 발생하는 각종 문제를 해결하는 데 도움을 줄 수 있는 정보를 생성한다. 이와 같이 공간현상을 나타내는 각종 공간자료는 GIS의 공간분석기법을 통해 짧은 시간 내에 정확하고 객관적인 정보로 변환된다. GIS를 이용한 공간분석 결과는 공간법칙이나 공간이론을 찾는 탐구와 다양한 공간의사결정 문제를 해결하는데 사용되고 있다(국토연구원, 2004). 이러한 GIS 공간분석 기법의 발달로 인하여 공간 데이터의 양이 급격하게 증가하고 있는 환경에서 의미 있는 정보들을 추출 할 수 있는 공간패턴분석, 공간통계, 군집화 기법 등 다양한 공간분석 기법의 적용이 이루어지고 있다(이건학, 2004).

우리나라 개별공시지가는 산정의 경우, 기준이 되는 토지 즉, 비교표준지를 선택하고 비교표준지와 산정대상필지의 토지특성을 비교하여 서로 다른 특성을 찾아낸다. 그리고 서로 다른 토지특성에 대한 가격배율을 토지가격 비준표에서 추출한 후 비교표준지 가격(공시지가)에 가격배율을 곱하여 산정한다(건설교통부, 2004). 그런데 전국토의 방대한 지가를 조사함에 있어 조사인력의 부족과 실무자의 임의적 처리 등에 의해 공시지가의 적정성에 대한 논란이 부분적으로 제기되고 있다. 특히, 지가 자동산정 시스템인 ALPA(Automatic Land Price Appraisal System)는 지가만 계산하는 초보적인 수준으로 추가적인 자동화의 필요성이 대두되고 있다(문태현, 2000).

즉, 단순지가의 계산만이 가능한 시스템보다는 지가의 변화·예측과 공간분석이 가능한 GIS 활용 등 새로운 지가평가 방법의 필요성이 대두되고 있다. 또한, 새

로운 지가평가와 산정 일련의 과정에 대한 자동화로 UIS나 LMIS(Land Management Information System) 토지정보관리체계 사업 등에 통합될 필요성이 제기되고 있다.

우리나라는 개별필지 주변의 비교표준지를 이용하여 개별공시지가를 산정하고 있는데 현행 개별공시지가 산정 지침의 경우 비교표준지 선정 작업시 용도지역, 유사가격권, 토지이용상황, 도로접면 등을 고려하도록 되어있다. 이러한 요소들 중에서 유사가격권을 제외한 나머지 요소들은 비교적 용이하게 조사할 수 있어 자동화가 가능하나 유사가격권의 경우는 비교표준지 선정시 지침상에서 상당히 중요한 요소로 취급되고 있으면서도 실제적으로는 범위와 적용에 관해서 모호한 개념상의 규정만을 두고 있어 자동화의 한계성이 있다(박수홍 등, 2003). 때문에 신속하고 정확한 비교표준지의 선정과 향후 지가산정 일련의 과정을 전산화를 통해 UIS나 LMIS 등 GIS 관련 정보화 시스템에 통합되기 위해서는 무엇보다 자동화된 기법으로 유사가격권을 구획할 수 있는 객관적이고 합리적인 방법론이 필요하다.

공시지가 자동화와 유사가격권 설정에 관한 대표적인 선행 연구를 고찰해보면, 먼저 ALPA와 지가현황도면을 연계시켜 토지특성조사, 표준지 선정, 가격조정 등의 업무시 착오를 쉽게 발견·수정할 수 있는 시스템의 개발 연구, 또는 토지특성 자료를 자동으로 추출하고, 이를 ALPA와 연계하여 지가를 산정하는 시스템 개발에 대한 연구가 있었다(아우름솔루션, 2003). 그러나 이 연구에서는 도형자료를 활용한 시각화와 일반적인 토지특성 항목을 기준으로 개별필지의 토지특성만을 자동 추출하여 기존 ALPA 시스템을 지원하는 기능 외에 자동화 측면의 접근에서는 한계성이 있다.

그리고 박성규(1999), 박수홍 등(2003)의 연구가 있는데 이 연구들은 비교표준지 자동 선정 과정에서 유사가격 요소를 표준지와 개별지간의 지가차이로 유사가격권의 개념을 적용시키거나 또는 유사가격을 도로접면 조건 등에 포함시켜 모호한 개념으로 규정되어 있는 유사가격권을 보다 명확함으로써 자동화 가능성을 제시하였다. 그러나 유사가격 권역의 구획이나 주변의 지가 분포 등이 고려되지 않는다는 한계성을 찾

을 수 있다. 최근에는 K-평균 군집화 기법을 이용하여 유사 가격권을 설정한 연구가 있었다(이성규 · 홍성언, 2004). 이 연구의 경우는 군집화 기법으로 유사가격권을 구획하였으나 군집 수를 사용자가 직접 정의해주어야 하므로 최적의 군집 수 설정과 이에 따르는 객관성에 한계성을 지니고 있다.

본 연구에서는 객관적인 기준과 수치적인 기준의 부재로 많은 문제점을 발생시키고 있는 유사가격권 설정의 문제를 해결하고자 평균 연결법(average linkage method)과 K-means 혼합 클러스터링 기법을 이용하여 유사가격권을 설정하고 이에 대한 적용 가능성을 제시하고자 한다. 구체적인 목적은 다음과 같다.

첫째, 군집분석의 이론적 탐색을 거쳐 평균 연결법과 K-means 혼합 클러스터링 모형을 정립한다. 둘째, 실험지역 적용을 통하여 객관화된 최적의 군집수를 결정하고, 결정된 최적의 군집수를 적용하여 유사가격권을 설정한다. 셋째, 군집화의 재현성 평가를 통하여 방법론의 타당성을 제시하고, 설정된 유사가격권 내 현행 개별필지와 비교표준지와의 일치율을 분석하여 봄으로써 방법론의 실제 적용 가능성을 제시하고자 한다.

2. 개별 공시지가 산정 방법과 유사가격권 설정의 문제점

1) 공시지가의 개념과 산정 체계

공시지가란 합리적이고 일관성 있는 지가정보체계를 세우기 위해 부동산가격 공시 및 감정평가에 관한 법률에 따라 산정하여 공시되는 땅값을 말한다. 공시지가의 종류에는 표준지공시지가와 개별지공시지가가 있다. 그 중에서 표준지공시지가란 표준지의 적정가격을 감정평가사가 평가하고 건교부장관이 공시한 표준지의 단위면적당(m^2)가격이다. 개별지공시지가는 표준지공시지가와 토지가격 비준표를 기초로 하여 공무원이 산정한 지가로서 시·군·구청장이 결정·공시한 지가를 말한다(건설교통부, 2004).

구체적으로 개별공시지가는 건설교통부장관이 매년

공시하는 표준지공시지가를 기준으로 시장·군수·구청장이 조사한 개별토지의 특성과 비교표준지의 특성을 비교하여 건설교통부장관이 개발·공급한 「표준지와 지가 산정 대상 토지의 지가형성요인에 관한 표준적인 비교표(토지가격 비준표)」상의 토지특성 차이에 따른 가격배율을 산출하고, 이를 표준지 공시지가에 곱하여 산정한 후 감정평가사의 검증을 받아 토지소유자 등의 의견수렴과 시·군·구 토지평가위원회의 심의 등의 절차를 거쳐 시장·군수·구청장이 결정·공시하는 개별토지의 단위 면적당 가격을 말한다.

일반적으로 개별공시지가는 ① 개별필지의 토지특성을 조사한 후에 ② 산정의 기준이 되는 토지 즉, 비교표준지를 선택하고, ③ 비교표준지와 산정대상필지의 특성을 비교하여 서로 다른 특성을 찾아낸 다음, ④ 서로 다른 토지특성에 대한 가격배율을 토지가격 비준표에서 추출한 후 ⑤ 비교표준지 가격(공시지가)에 가격 배율을 곱하여 산정 하게 된다.

2) 유사가격권의 설정 방법과 문제점

비교표준지의 선정 기준을 고찰해 보면, 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 토지 이용상황이 같은 유사가격권의 표준지가 없는 경우에

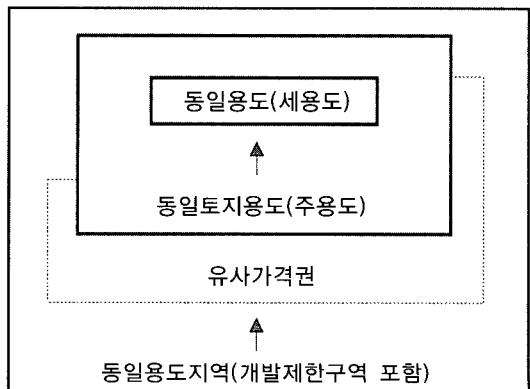


그림 1. 비교표준지 선정 개념도

는 토지이용 상황이 다르더라도 조사대상 필지 인근의 토지이용상황을 감안하여 유사가격권의 표준지를 선정하여야 한다고 규정하고 있다(건설교통부, 2004)(그림 1).

이렇게 현행 지침상에서는 비교표준지 선정의 중요요인으로 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 자동화된 방법론으로 비교표준지를 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 등, 2003; 이성규·홍성언, 2004).

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.

3. 혼합 클러스터링 모형의 설계와 재현성 평가 방법

1) 혼합 클러스터링 모형 설계

(1) 클러스터링 적용 개요

군집분석의 방법은 각 측정치 사이의 유사성의 척도로서 무엇을 사용하느냐에 따라 여러 가지로 분류될

수 있다. 유사성의 척도로 사용되는 것에는 각 측정치 사이의 상관계수와 거리가 있다. 이중 많은 분석법에서 사용하는 것이 거리에 의한 것이고, 거리를 이용하는 방법에는 Mahalanobis Distance, Minkowski, Euclidean Distance 등이 있다. 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 *et al.*, 1987).

객관적인 기준에 의해 구체적으로 유사가격권을 설정하기 위해 이 연구에서 사용되어진 군집 분석 방법은 대량의 자료에서 군집을 발견하는데 상당히 효과적인 것으로 알려져 있는 비계층적 군집 분석 방법의 K-means 클러스터링 분석과 객관적인 군집 수 선정에 있어 다양한 기준 통계치 정보를 제공해 주는 계층적 군집 분석 방법의 평균연결법(average-linkage)을 혼합·적용하였다. 이에 대한 구체적인 내용은 다음과 같다.

(2) K-means 클러스터링

K-means 클러스터링 기법은 MacQueen에 의하여 제안된 알고리즘으로서, 우선 패턴을 k개의 군집으로 나누는 후 군집에 포함되어 있는 패턴들의 평균을 클러스터의 중심 값으로 계산한다. 클러스터의 중심값과 각 패턴과의 거리를 계산한 후 가장 거리가 가까운 클러스터에 패턴을 포함시키는 방법으로 조건은 다음의 식과 같다(MacQueen, J. B, 1967; 조형기 등, 1996; 김윤식 등, 2000).

$$x_i \in C_j \quad \|x_i - z_j\|^2 < \|x_i - z_k\|^2$$

여기서, $1 \leq i \leq N$, $1 \leq k \leq c$, $j \neq k$ 이며, N은 패턴 수, c는 군집 수, z는 군집의 중심값을 나타낸다. 일반적으로 사례(자료)의 수가 수백 개를 넘을 경우에는 K-means 클러스터링 분석을 사용하는 것이 바람직하다. 이 방법은 크게 다음과 같은 두 가지 절차를 따르고 있다. 첫째는 초기 군집의 수 k를 규정하여 결정된 초기 군집에 각 개체들을 할당하는 것이며, 둘째는 군집의 일부 개체들 또는 전체를 기준에 따라 최적분리에 이를 때까지 해당하는 규칙에 재할당하는 것이다.

이 기법을 사용하기 위해 우선 군집 중심의 초기값(initial seed)을 찾아야 하는데, 일반적으로 이것은 p 개의 변수로 기술되는 n 개의 자료에 대해 행렬을 구성한 후, p -차원(variable) 공간에서 주어진 군집의 개수와 같은 k 개의 점을 선택함으로써 시작된다. 이렇게 군집 중심의 초기값이 결정되면 다음 단계로 각 개체는 적절한 근접도(proximity measure)에 근거하여 k 개의 초기 중심에 가장 가까운 군집에 할당하고, 그 군집의 중심은 추가적으로 할당된 구성원에 의해 다시 계산된다. 이러한 할당과정은 반복되며 군집내의 흐트러짐이 큰 군집은 다시 분리되고, 작은 군집들은 서로 병합되는 형식으로 진행된다.

- step 1. 주어진 표본에서 처음 k 개의 개체들을 초기 k 개 군집의 중심으로 선택
- step 2. 임의로 k 개의 점들을 선정하여 초기 값으로 사용
- step 3. k 개의 초기 중심점 선택
- step 4. 각 개체를 가장 가까운 중심점을 갖는 군집으로 할당한 후 새로운 군집의 중심점 계산
- step 5. 각 개체의 할당에 변화가 없을 때까지 위의 단계를 반복하여 최종적으로 k 개의 군집 형성

(3) 객관적인 군집수의 선정

앞서 언급한 K-means 클러스터링 기법이 대용량의 데이터를 처리하는데 최적화된 기법으로 많이 사용되고 있지만 초기에 군집수 k 를 설정해야 하는 어려움이 있다. 연구에서는 이러한 한계성을 극복하고자 계층적 군집분석 결과 얻어지는 통계량의 변화를 관찰하여 군집 수를 결정하는 방법을 이용하였다.

계층적 클러스터링 방법은 합쳐지는 두 군집간의 거리를 무엇으로 정하느냐에 따라 두 군집간의 최단거리를 기초로 하는 단일연결법(single linkage method),

최장거리를 기초로 하는 완전연결법(complete linkage method), 두 군집에 있는 모든 개체들 사이의 거리의 평균을 기초로 하는 평균연결법(average linkage method), 두 군집간의 평균거리를 기초로 하는 중심연결법(centroid linkage method), 각 병합의 단계에서 생기는 잔차 제곱합(error sum of squares)을 기초로 하는 Ward 법 등이 있다(마상진, 2003)(그림 2).

계층적 분석 기법의 가장 큰 장점은 최적의 군집 수에 대한 정보의 획득이 가능하다는 것이다. 즉, 단계별 군집 결과를 시각화한 트리(tree)형태의 덴드로그램(dendrogram)을 그린 후 군집이 병합되는 과정에서의 거리가 상대적으로 큰 변화를 보일 경우에 대한 군집 개수를 취하는 방안이 있다. 그리고 어떤 판정기준을 최적화 시키는 군집방법에서 군집의 개수에 대응되는 판정기준의 값(통계량)을 플랫폼에 급격한 변화를 보이는 곳에 대응되는 군집의 개수를 결정하는 방법이 일반적으로 많이 쓰인다(Kaufman and Rousseeuw, 1990).

이러한 종류의 계층적 클러스터링 기법들은 적용 대상의 목적에 적절하게 이용이 되는데 연구에서는 단일 연결법과 완전연결법의 경우 이 연구에서 유사가격권 설정을 위한 객체와 군집 혹은 군집과 군집 간의 결합에 있어 단적인 경우(최단-최장 거리)를 토대로 군집을 형성하므로 군집화에 문제가 있어 이는 배제하였다. 나머지 기법 즉, 평균연결법과 중심연결법, Ward 법을 중심으로 사전 실험을 거쳐 기법의 선정이 이루어졌다. 적용결과 Ward 법은 최적 군집수 선정을 위한 통계량 분석에서 군집수의 해석이 어려웠다. 그리고 평균연결법과 중심연결법은 통계량 분석의 결과는 유사하나 평균연결법이 통계량에서 좀더 명확한 군집수 정보를 나타내어 연구에서는 K-means 클러스터링과 평균연결법을 이용했다. 평균연결법에 대한 구체적인 알고리즘은 다음과 같다.

- step 1. 맨 처음 단일 개체를 포함하는 N 개 군집과 대칭인 $N \times N$ 거리행렬 $D=d_{ij}$ 를 구성한다.
- step 2. 최단거리가 되는 두 개 군집 쌍을 U, V 라 할 때 이 둘 사이의 거리 d_{UV} 를 계산한다.
- step 3. 두 군집 U, V 를 병합하고 이들로 구성된 군집 (UV)의 거리행렬에서 두 군집에 대응하는 행과

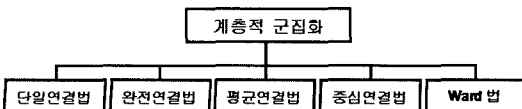


그림 2. 계층적 군집화 방법

열을 모두 삭제하고 (UV)에 대응하는 행과 열에 대해 나머지 군집들과의 거리를 새로 계산한다.

step 4. 앞의 단계2와 단계3을 총 $N-1$ 번 반복한다. 매회 반복 후 서로 병합된 군집 및 각 병합 단계에서의 수준(거리)을 표시해 둔다.

(4) 혼합 클러스터링 모형 설계

연구에서는 그림 3과같이 유사가격권역 설정을 위해 혼합클러스터링 모형을 설계하였다. 이를 구체적으로 기술하면, 먼저 비계층적 군집 분석(non-hierarchical clustering method) 방법인 K-means 클러스터링 분석(K-means cluster analysis)은 계층적 군집 분석 방법에 비하여 계산 속도가 빠르고 대량의 자료에서 군집을 발견하는데 상당히 효과적인 장점을 가지고 있다. 그러나 군집의 개수와 중심점을 주는 방법으로 주어진 패턴에서 처음 k 개의 군집수를 추출하여 군집의 중심을 주는 방법 외에 임의로 k 개의 군집수를 추출하여 중심을 주는 방법의 경우 사전에 주어진 군집의 수 k 가 원 데이터 구조에 적합하지 않거나, 군집의 개체 분류시 처음 선정한 군집 중심들의 영향을 많이 받으므로 부적절한 위치에 군집 중심이 위치한 경우 좋은 군집화 결과를 얻기 힘들다. 따라서 K-means 클러스터링 기법이 가지고 있는 문제점을 극복하기 위해 군집수 결정에 있어 계층적인 군집 분석 방법인 평균연결법을 이용하였다. 이렇게 클러스터링 기법을 혼합 설계함으로써 반복적인 군집 수에 관한 실험 없이도 통계적이고 객관적으로 군집수 정의할 수 있어 단일 기

법을 적용할 경우 보다 효율성을 높였다. 그리고 클러스터링 시행에 따른 다양한 통계량 즉, PSF, PST2, CCC, RMSSTD, SPR, RS에 대해서는 실험부분에 자세히 기술하였다.

2) 재현성(reproducibility) 평가

군집화가 이루어지고 나면 이에 대한 객관적인 정확도의 평가가 요구된다. 군집화는 흔히 차원 수가 큰 자료를 대상으로 하기 때문에 면밀한 자료탐색(data exploration)을 하더라도 최적 군집 수에 대한 감을 갖기가 힘들기 때문에 형성된 군집화 결과가 재현성(일치도)을 갖는가를 확인할 필요가 있다. 동일한 메커니즘에서 생성된 독립적인 새 데이터 셋을 동일한 방식으로 자체 군집화한 결과가 기존 군집화를 적용한 결과와 유사하다면 재현성이 있는 것이므로 군집화는 바람직하며, 반면에 두 결과가 상당부분 불일치한다면 그런 군집화는 바람직하지 않다고 본다(허명희 · 이용구, 2003). 자료 분할(data partitioning) 방식은 동일한 군집화 방법의 반복(replication)을 가능하게 해주므로 이를 활용하여 재현성 평가를 할 수 있다(Milligan, 1996). 재현성 평가 절차는 다음과 같다.

- Step1. 주어진 자료를 training data와 test data로 분할한다.
- Step2. training data를 주어진 k 개의 군집을 갖도록 클러스터링 하여 모델(training data based model)을 형성하고 이를 test data에 적용하여 분류한다.
- Step3. test data를 단계2와 동일한 방식으로 군집화 하여 모델(test data based model)을 형성한다.
- Step4. test data의 두 군집화 결과를 토대로 교차분류표(cross-table)를 작성한다. 적용된 군집화가 의미 있는 것이라면 교차분류표를 이루는 행과 열은 강한 대응성을 보인다.

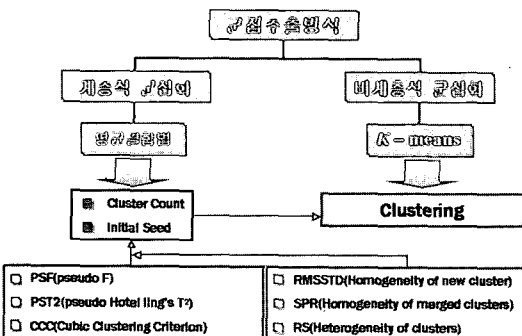


그림 3. 혼합 클러스터링 모형

훈련자료와 검증자료에 대하여 좀 더 구체적으로 기술하여 보면 다음과 같다. 분석하고자하는 data set을 분할하는 자료분할(data partitioning) 방식을 거치게 되는데 훈련자료와 검증자료를 어떻게 분할하며 얼마

나 산정하는지의 기준이 명확하게 정해진 것은 아니지만 주로 무작위추출(random sampling)방법을 사용하며 전체 data set의 1/2 이나 1/3 데이터를 검증자료로 사용한다. 무작위추출 방법은 모집단의 모든 표본이 동등하게 추출될 가능성이 높고 모든 구성요소들의 성격이 모두 서로 비슷하며, 하나의 표본이 추출되는 것이 다른 표본이 추출될 확률에 아무런 영향을 주지 않는 경우나, 모집단의 단일특성에 대해 분석하려는 경우에 사용되는 추출방법이다. 이 연구에서는 해당 용도지역을 기준으로 지번 순서로 나열했을 때 홀수-짝수 번째에 속하는 데이터를 각각 훈련자료와 검증자료로 설정하였다. 지번부여 방식의 성격상 인접지번의 거리 및 지가 차이는 서로 비슷하다는 가정 하에 전체 data set을 훈련자료와 검증자료 양쪽 모두에 균등하게 배분하였다.

자료 분할 방식을 이용한 Rand Index(Rand, 1971)는 군집화의 재현성을 평가하는 객관적인 지표로 많이 사용된다(김성호 등, 2002). Rand Index는 얼마나 군집의 유사한 개체들이 같은 군집 혹은 상이한 개체들이 서로 다른 군집에 소속되어 있는가를 나타내는 지표라 할 수 있다. Rand Index의 분자는 동일한 두 개체 쌍이 동일한 군집에 소속되어 있는가 혹은 상이한 군집에 소속되어 있는가의 빈도 수이며 분모는 군집 모형을 이루는 전체 개체 쌍의 수이다. Rand Index는 2×2 분할 표를 사용하여 구할 수 있다(표 1). 표에서 짝지어진 데이터 쌍의 모든 경우의 수는 n(n-1)/2개이며, 이들은 다음의 두 범주로 분류된다.

- 1 : 짝지어진 쌍이 클러스터링 결과에서 하나의 클러스터에 속한 경우의 수
- 0 : 짝지어진 쌍이 클러스터링 결과에서 서로 다른

표 1. Rand Index(2×2 분할 표)

		클러스터링 결과 B	
		1	0
클러스터링 1 결과 A	0	a	b
	1	c	d

클러스터에 속한 경우의 수

Rand Index는 총 개체 쌍(a+b+c+d) 중 일치하는 쌍(a+d)의 비율이며 식은 다음과 같다.

$$\text{Rand Index} = \frac{a+d}{a+b+c+d}$$

Rand Index에서 두 군집화 결과 간의 유사도(similarity) 분석은 두 개의 군집이 정확하게 일치한다면 Rand Index는 1.0의 값을 군집의 구성원들 간에 일치 가능성이 전혀 이루어지지 않는다면 Rand Index는 0의 값을 가지게 되는데 값이 클수록 군집화결과 간의 유사도가 높다는 것을 의미한다. 즉, 군집화가 올바르게 정확하게 수행되었다는 것을 의미하게 된다.

4. 실험 및 결과분석

1) 실험

평균 연결법과 K-means 클러스터링 기법을 이용하여 유사가격권을 설정하고자 서울시 강남구 삼성동 일부 지역(일반주거지역, 일반 상업지역)을 연구지역으로 선정하였다(그림 4). 데이터는 강남구에서 구축하여 놓은 토지특성 도면을 이용하였다.

실험에 이용된 연구대상지역과 이에 따른 데이터는 선행 연구(이성규·홍성언, 2004)에서 이용하였던 것과 동일하도록 하였다. 선행 연구에서는 K-means 군집화 기법만을 이용하여 유사가격권역을 구획하고자 하였으나 연구에서는 보다 개선된 혼합 클러스터링 기법을 이용하여 객관적인 군집 수의 정의와 이를 통한 정확도 향상(현행 비교표준지와의 부합률 기준)을 목적으로 하였다. 그래서 선행 연구성과와 명확한 비교가 가능하도록 동일한 지역과 동일한 데이터를 이용하여 실험함으로써 보다 명확하게 개선된 방법론을 제시하고자 하였다.

데이터 구축 시 특수 토지나 공공용지의 경우는 일반토지와 비교표준지를 선정하는 기준이 상이하기 때문에 이를 포함할 경우 군집화에 오류가 발생할 수 있

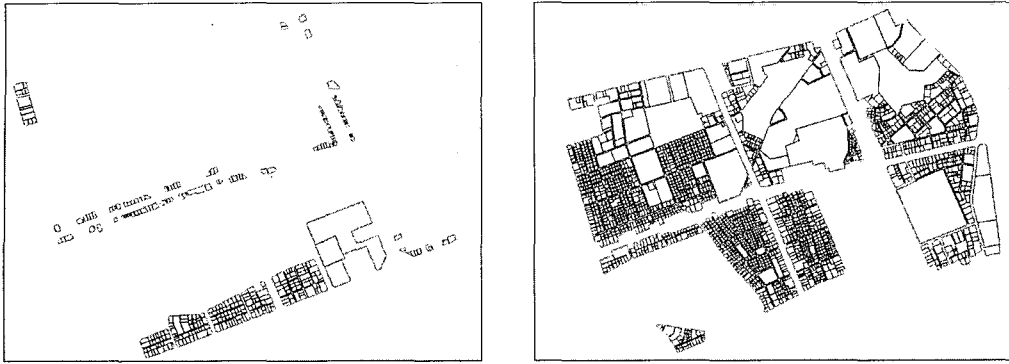


그림 4. 연구대상지역(좌 : 일반상업지, 우 : 일반주거지)

기 때문에 대상 지역에서 배제하였다.

혼합클러스터링 기법을 이용한 유사가격권 설정 절차로는 먼저, 군집분석 수행 전에 전처리(preprocessing) 과정을 거쳤다. 그리고 유사 가격권 설정 인자로 지가와 거리(X, Y) 두 가지 인자를 적용하였는데 두 인자들 간의 유사성 거리를 계산할 때 주의할 점은 변수값을 표준화해야 한다는 것이다. 만약 변수값을 표준화하지 않는다면 변수값의 단위에 따라 상이한 결과를 초래할 수 있기 때문이다(Milligan and Cooper, 1987). 데이터의 특성상 거리의 편차 정도에 비해 가격의 편차 정도가 크게 나타나므로 가격에서의 차이가 대상간의 거리를 완전히 좌우하게 되는 경향을 보인다.

유사가격권 설정을 위해 거리에 의한 차이도 무시할 수 없는 것이므로 자료의 표준화를 통해 선정된 변수들의 가중치를 표준화해 주어야 한다. 데이터의 결측치(missing value)에 의한 영향은 군집분석 전 무손실 데이터임을 확인 하였으므로 배제하였다.

평균연결법을 적용하여 하위 군집을 상위 군집으로 순차적으로 묶는 과정에서 결합한 결과를 다양한 통계치를 기준으로 통계치의 변동이 유의적인 부분을 찾아 선정 기준에 적합한지 판단한 다음 최적 군집수를 결정하였다.

사용된 통계치를 간략하게 고찰해 보면 첫째, 군집수의 변화에 따라 Cubic Clustering Criterion(CCC)의 변화를 보고 결정하는 방법으로 군집수의 변화에 따라 CCC값이 급격한 상승을 보이다가 그 다음 완만한 감소를 보이는 국부적 최고점(local peak)에서 최적의 군

집수를 결정한다. 둘째, Pseudo F Statistics(PSF)가 군집수의 변화에 따라 국부적 최고점을 보이는 곳에서 최적의 군집수를 결정하는 방법으로 모든 군집들 간의 분리정도를 측정하여 값이 크면 더 이상 군집을 합치는 것이 의미가 없다고 본다. 셋째, 군집수의 변화에 따라 Pseudo Hotel ling's T2(PST2)의 변화가 전 구간 보다 급격히 상승하는 구간의 바로 전 구간을 최적의 군집수로 결정한다. PST2는 단일 변수에 대한 집단간 평균차이를 검정하는 Two-sample T-test의 확장된 형태로 여러 변수들의 집단간 차이가 있는지 검정하는 다변량 분산분석(MANOVA)과 비슷한 개념이다. 넷째, RMSSTD(Root Mean Square Standard deviation)는 새로 형성되는 군집의 RMS 표준편차로 값이 작을수록 군집의 균질성이 높다고 볼 수 있다. 다섯째, RSQ(R-squared)는 다중상관계수의 제곱으로 군집간 제곱합을 전체 제곱합으로 나눈 것이다. RSQ는 군집의 분리정도를 나타내는 하나의 척도로서 값이 크면 군집간의 이질성(군집내의 동질성)이 높음을 나타낸다. 끝으로, SPR(Semipartial R-squared)은 동질성 손실(통합 전후의 제곱합 차이) 정도를 전체 제곱합으로 나눈 비율로 값이 작을수록 군집의 동질성이 크다고 볼 수 있다.

다음으로는 K-means 클러스터링에 제충적 방법의 평균연결법에서 구한 최적 군집수와 군집 중심점 정보를 초기 군집의 개수 k와 초기 군집 중심점 initial seed에 제공해 준다. 일반적으로 초기 군집의 개수 k만을 넣어주는 경우에는 K-means 클러스터링 처리 과정에서 임의로 초기 중심점을 선정하고 군집의 개체를 분

표 2. 최적 군집수 선정을 위한 통계치 분류 및 솔루션

STATISTIC	COMMENTS
CCC(Cubic Clustering Criteria)	Local peak
PSF(Pseudo F)	Local peak
PST2(Pseudo Hotel ling's T2)	Δ PST2 should be high
RMSSTD(Root-mean-square STD)	Value should be small
RSQ(R-Squared)	Value should be small
SPR(Semipartial R-Squared)	Value should be high

류할 때 처음 선정한 군집 중심들의 영향을 많이 받으므로 부적절한 위치에 군집 중심이 위치한 경우 좋은 군집화 결과를 얻기 힘들다. 따라서 연구에서는 군집수 k와 초기 군집 중심점을 같이 입력하여 클러스터링을 수행하였다.

2) 결과 분석

표 3은 일반 상업지역과 일반 주거지역에 대한 산출 통계량을 나타낸 것으로 일반 상업지역의 경우, 총 327 필지를 대상으로 평균연결법을 적용하여 최적의 군집수를 선정하였다.

군집수는 지가 수준이 비슷한 지역적 범위의 공간적인 분할 수 즉, 비슷한 가격권역이 구획되는 수를 나타내는 것이다. 그런데 현재로서는 다양한 연구의 부재와 모든 토지특성 관련 데이터 적용의 한계성 등 여러 다양한 문제로 인하여 어떤 특정 군집수를 객관적인 기준에 의하여 정확하게 독립된 수치 하나로 정의하기에는 무리가 있다고 판단된다. 그렇기 때문에 본 연구의 방법론 역시 군집화 기법을 이용하고, 여러 통계치 분석을 통하여 가장 양호하다고 판단되는 군집수의 선정에 범위를 한정하고자 한다. 물론 군집수를 선정함에 있어 명확하게 통계량으로 구별되지 않는 부분에 대해서는 유사한 군집수 까지도 포함하여 분석하기로 한다.

분석결과, 군집수가 7 → 6 으로 줄어들 때 CCC와 PSF 값이 국부적 최고값(local peak)을 보이며 PST2 수치가 군집수가 6일 때의 166 과 비교했을 때 Δ PST2 수치가 다른 군집 구간보다 상대적으로 높은 변화를

보인다. 또한 RMSSTD, RSQ, SPRSQ 수치가 군집수가 7 → 6 으로 줄어들 때 급격한 변화를 보이는 것을 볼 수 있다. 따라서 연구의 적용 방법론상으로는 군집수 7개를 가장 최적의 군집수로 선정할 수 있을 것이다. 즉, 연구 대상 지역의 가장 객관적이고 최적의 유사가격 권역의 수는 7개로 해석할 수 있다.

그런데 군집수가 7개 이외에도 11개, 17개도 유사하게 국부적 최고점을 보이는 것으로 분석되었다. 따라서 이 연구의 대상지역에서는 특정 군집수 하나를 선정하고 방법론을 적용하기 보다는 다양한 군집수를 대상으로 실험하여 현행 제도와 가장 높은 부합률을 보이는 유사가격권역 수를 분석하는 것이 바람직하다고 판단하였다. 그래서 분석된 7개의 군집수 이외에도 유사하게 국부적 최고점을 보이는 군집수 11, 17개도 같이 실험을 하였다.

시각적인 이해를 돕기 위해 표 2에 나타난 수치들의 변화를 군집수(NCL) 별 각각의 통계치를 매칭 시켜 산점도로 표현해 보았다. +로 표시된 부분은 최적 군집수로 선정된 구간을 의미한다.

CCC 값이 3 이상이며 급격한 상승을 보이다가 완만히 감소하는 구간을 보이는 국부적 최고점(local peak)은 군집수가 7, 11, 17일 때 나타나며 PSF 수치 또한 동일한 구간에서 국부적 최고점을 보이고 있다(그림 5).

두개의 군집을 하나의 새로운 군집으로 형성하는 과정에서 PST2 값의 차이가 뚜렷하게 나타나는 구간은 7, 11, 17 로 나타나며, 새롭게 형성되는 군집의 균질성(homogeneity) 정도를 보이는 RMSSTD 수치는 군집수가 7인 구간까지 어느 정도 완만한 증가를 보이다가 6인 구간에서 급격히 증가하는 경향을 보이므로 군집수가 7인 구간이 최적 군집수로 의미가 있다고 본다(그림 5).

RSQ 수치가 높다는 것은 군집간의 이질성(heterogeneity)이 높음을 나타내는데 군집수가 7인 구간 이후부터 값이 급격히 감소하는 경향을 보이며, 군집의 통합 과정에서 발생하는 동질성 손실(loss of homogeneity) 정도를 나타내는 SPR 수치가 군집수가 7인 구간까지 거의 변화가 없다가 그 이후부터 값이 급격히 증가하는 경향을 보이므로 군집수가 7인 구간이 최적 군집수로 의미가 있다고 본다(그림 5).

일반주거지역은 총 1976 필지를 대상으로 최적 군

표 3. 일반상업지역과 주거지역의 산출 통계량

일반상업지역							일반주거지역						
NCL	CCC	PSF	PST2	RMSSTD	RSQ	SPRSQ	NCL	CCC	PSF	PST2	RMSSTD	RSQ	SPRSQ
20	41.8	704	13.6	0.1894	0.978	0.0005	20	-2	682	35.5	0.3801	0.869	0.0005
19	42.3	723	22.7	0.1637	0.977	0.0007	19	-1.2	702	57.4	0.3537	0.866	0.0030
18	42.5	731	25.4	0.2118	0.976	0.0012	18	-1.8	706	169	0.4238	0.860	0.0062
17	42.6	739	17.1	0.2208	0.974	0.0013	17	-0.1	739	43.2	0.4091	0.858	0.0019
16	35.3	556	119	0.255	0.964	0.0104	16	1.65	775	47.4	0.4028	0.856	0.0021
15	35.1	556	40.6	0.2622	0.961	0.0026	15	2.6	802	85.4	0.4288	0.852	0.0044
14	33.4	525	64.4	0.2577	0.956	0.0053	14	2.71	819	234	0.4915	0.845	0.0069
13	31.8	499	55.8	0.3005	0.95	0.006	13	5.17	872	40.8	0.4986	0.843	0.0023
12	28.3	440	112	0.332	0.939	0.0113	12	-28	484	1405	0.5758	0.731	0.1113
11	29.3	463	29.5	0.3232	0.936	0.0028	11	-25	524	35.4	0.4869	0.728	0.0033
10	25.3	402	94.8	0.3883	0.919	0.0166	10	-24	548	139	0.5749	0.716	0.0124
9	21.8	358	90.5	0.429	0.9	0.0194	9	-21	592	66.6	0.6269	0.708	0.0083
8	21.8	364	60.5	0.3128	0.889	0.0113	8	-17	660	152	0.5846	0.702	0.0053
7	23.7	399	90	0.2848	0.882	0.0067	7	-17	689	220	0.7008	0.678	0.0238
6	16.5	314	166	0.4649	0.83	0.0518	6	-26	608	672	0.7427	0.608	0.0704
5	4.31	205	185	0.6129	0.718	0.1124	5	-32	552	460	0.6713	0.529	0.0787
4	3.86	211	248	0.6509	0.663	0.0552	4	-26	667	104	0.8338	0.505	0.0245
3	-7	127	188	0.816	0.44	0.2224	3	-47	300	1124	0.8817	0.234	0.2708
2	-13	48.3	174	0.9501	0.129	0.3108	2	-41	99.6	475	0.9771	0.048	0.1857
1	0		48.3	1	0	0.1294	1	0		99.6	1	0	0.0483

집수 선정을 위한 실험이 이루어졌는데 이에 대한 산출 통계량은 표 2와 같다. 이를 분석하여 보면, 군집수가 13 → 12로 줄어들 때 CCC와 PSF 값이 국부적 최고값(local peak)을 보이며 PST2 수치가 군집수가 12일 때의 1405와 비교해 Δ PST2 수치가 다른 군집 구간보다 상대적으로 높은 변화를 보인다. 또한 RMSSTD, RSQ, SPRSQ 수치가 군집수가 13 → 12로 줄어들 때 급격한 변화를 보이는 것을 볼 수 있다. 따라서 최적의 군집 수를 13개로 선정 할 수 있다.

CCC 값이 3 이상이며 급격한 상승을 보이다가 완만히 감소하는 구간을 보이는 국부적 최고점(local peak)은 군집수가 13, 22, 27일 때 나타나며 PSF 수치 또한 동일한 구간에서 국부적 최고점을 보이고 있다. 가장 뚜렷한 변화를 보이는 구간은 군집수가 13인 구간이다

(그림 6).

두개의 군집을 하나의 새로운 군집으로 형성하는 과정에서 PST2 값의 차이가 뚜렷하게 나타나는 구간은 앞서 CCC와 PSF에서 국부적 최고점을 보였던 13, 22, 27 구간 중 13인 반면, 새롭게 형성되는 군집의 균질성(homogeneity) 정도를 보이는 RMSSTD 수치는 군집수가 변할 때 급격히 증가하는 경향을 보이는 구간을 찾아 볼 수 없다(그림 6).

RSQ수치가 높다는 것은 군집간의 이질성(heterogeneity)이 높음을 나타내는데 군집수가 13인 구간 이후부터 값이 급격히 감소하는 경향을 보이며, 군집의 통합 과정에서 발생하는 동질성 손실(loss of homogeneity) 정도를 나타내는 SPR 수치가 군집수가 13인 구간 까지 거의 변화가 없다가 그 이후부터 값이

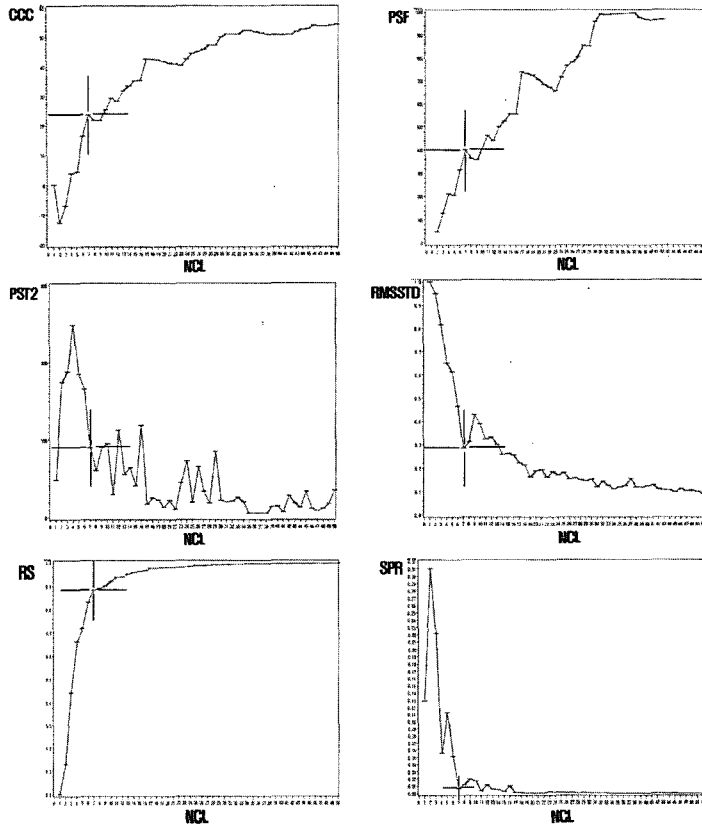


그림 5. 일반상업지역 군집구간별 산점도

급격히 증가하는 경향을 보이므로 군집수가 13인 구간이 최적 군집수로 의미가 있다고 본다(그림 6).

3) 클러스터링의 재현성 평가

본 연구에서는 군집화를 시행하고 이의 객관적인 정확도를 평가하기 위해 Rand Index를 이용하였다. 이와 더불어 구간별 군집화의 시행결과가 현행 제도하의 유사가격권을 어느 정도 반영하고 있는가를 분석하기 위해 군집내 개별필지의 비교표준지 일치정도를 분석하였다. 또한 군집화의 정확도와 현행 비교표준지와의 일치율이 가장 높다고 판단되는 군집 구간을 선별하여 구간별 각 군집이 포함하고 있는 필지 수, 지가 평균, 표준편차 등을 분석하였다.

표 4는 연구지역(일반 주거지역/일반 상업지역)에 대한 구간별 각 군집내 개별필지의 비교표준지 분포현황

과 Rand Index 수치를 나타낸 것이다. 결과를 분석해 보면, 일반 주거지역의 경우는 군집구간을 13으로 설정 하였을 때 비교표준지와의 일치율이 대략 96%로 가장 높게 나타났고, Rand Index 수치 역시 대략 0.86으로 가장 높게 나타났다. 비교표준지와의 일치율이 가장 낮은 구간은 27(91%)로 나타났고, Rand Index 수치가 가장 낮은 구간은 22(0.75)로 나타났다. 일반 상업지역의 경우는 군집 구간을 7로 설정 하였을 때 비교표준지의 일치율이 100%로 가장 높았고 Rand Index 수치 역시 0.96으로 가장 높게 나타났다. 비교표준지의 일치율이 가장 낮은 구간은 17(93.8%)이었고, Rand Index 수치가 가장 낮은 구간은 11(0.8)로 나타났다.

결과 분석에 대하여 전반적으로 고찰할 경우, 군집 구간을 일반주거지 13과 일반상업지 7로 정의하였을 경우, 개별필지의 비교표준지가 가장 많이 분포하고 있다. 그리고 군집수를 증가시켜 정의할 경우, 상대적

으로 비교 표준지가 감소하는 것으로 나타났다. 이것은 군집 수에 따라 상대적으로 포함할 수 있는 필지 수가 감소하기 때문으로 분석된다. 또한, 군집화 결과의 군집별 일치도 측정지표인 Rand Index도 군집수가 상대적으로 적을 때 비교적 높은 수치가 나타났다.

결과적으로 군집화의 정확도를 나타낸다고 볼 수 있는 Rand Index 수치와 현행 비교표준지와의 일치율을 고려하여 유사가격권을 설정할 경우 연구 대상 지역의 일반 주거지역은 군집구간을 13(유사가격 권역 수)으로 일반 상업지역은 군집구간을 7로 설정하는 것이 가장 합리적이고 타당할 것으로 사료된다. 이렇게 혼합 클러스터링 기법을 이용하여 유사가격권을 구획한다면 대상 지역별로 객관적인 유사가격 권역 수를 결정할 수 있고, 또한 유사가격권의 구획이 가능하다.

그림 7은 대상지역 일반상업지역과 일반주거지역의 군집화 결과를 나타내고 있는 것으로 각각 군집수 K를 7과 13으로 설정한 결과이다. 모두 주어진 군집수일 경우 현행 비교표준지와의 일치율과 Rand Index의 수치가 가장 높게 나타났다. 군집화 된 결과를 보면 적정 거리별로 가격이 유사한 필지들이 하나의 군집을 이루고 있는 것을 확인할 수 있다.

표 5는 Rand Index 수치와 비교표준지와의 일치율이 가장 높았던 군집구간(일반주거지 : k=13, 일반상업지 : k=7)을 선별하여 군집별 포함 필지 수, 평균지가, 표준편차를 산정한 결과이다. 이를 분석해 보면, 일반 주거 지역에서 13군집의 평균지가가 가장 높고 상대적으로 4군집의 평균지가가 가장 낮은 것으로 나타났다. 표준편차의 경우는 5군집이 가장 높게 나타났고, 7군집이 가장 낮게 나타났다. 일반 상업지역은 1군

집의 평균지가가 가장 낮았고 5군집의 평균지가가 가장 높게 나타났다. 표준편차의 경우 1군집이 가장 낮게 나타났다.

종합적으로 결과를 고찰해 보면 다음과 같다. 군집화의 재현성을 평가하는 지표로 사용된 Rand Index 수치를 통해 평균연결법에 의해 구한 최적군집수 K=13(일반주거지역), K=7(일반상업지역) 일 때 군집 내의 동질성과 군집 간의 이질성이 높음을 검증할 수 있었으며 평균연결법과 K-means 클러스터링 기법을 이용하여 유사가격권을 설정할 경우 현행 비교표준지와의 일치율이 일반 주거지역에서 95.8% 일반 상업지역에서 100%로 나타나 높은 부합률을 보이는 것으로 분석되었다.

이러한 결과는 현행 수작업된 결과와의 비교이므로 절대적인 수치라고는 할 수 없으나 현행 제도의 반영 정도로서 해석할 경우 상당히 높은 수치라고 판단된다. 그리고 이러한 혼합 클러스터링 방법론을 적용할 경우 지가의 분포 정도를 파악할 수 있어 지가 불균형 초래를 방지할 수 있고 향후 비교표준지를 자동으로 선정할 경우보다 정확하고 객관적인 선정이 이루어질 수 있을 것으로 판단된다.

5. 결론

본 연구에서는 지리학적인 공간법칙이나 공간이론의 탐구와 다양한 공간의사결정 문제를 해결하는데 사용되고 있는 GIS 공간분석과 다량의 공간 데이터를 가지고 효율적으로 의미 있는 정보들을 추출 할 수 있는

표 4. 구간별 각 군집내 개별필지의 비교표준지 분포 현황

구분	군집구간	필지 수	표준지 수	비교표준지 수	일치율(%)	Rand Index
일반주거 지역	13	1,967	79	1,891	95.8	0.861
	22	1,967	79	1,867	94.4	0.752
	27	1,967	79	1,802	90.8	0.851
일반상업 지역	7	327	32	327	100	0.963
	11	327	32	324	98.8	0.799
	17	327	32	311	93.8	0.909

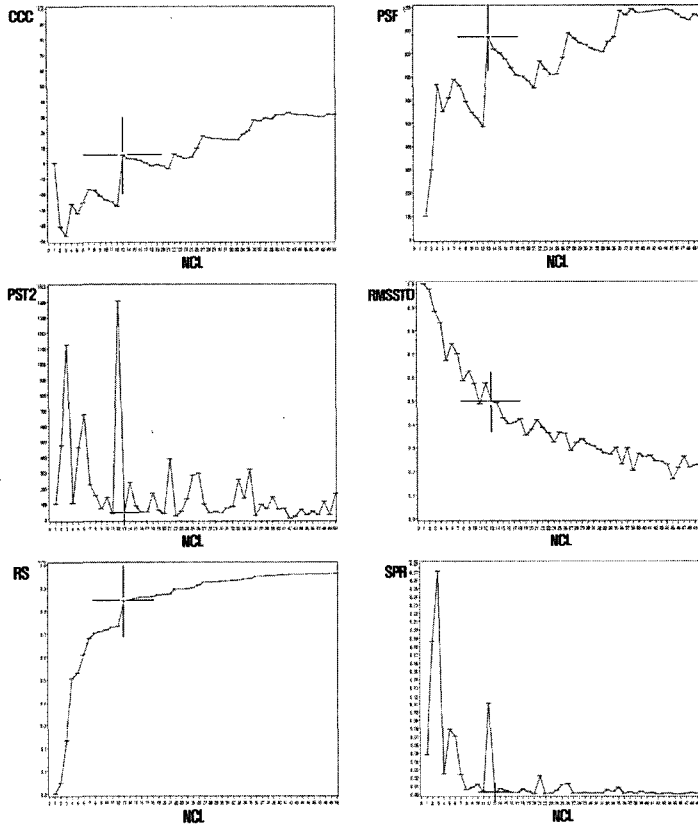


그림 6. 일반주거지역 군집구간별 산점도

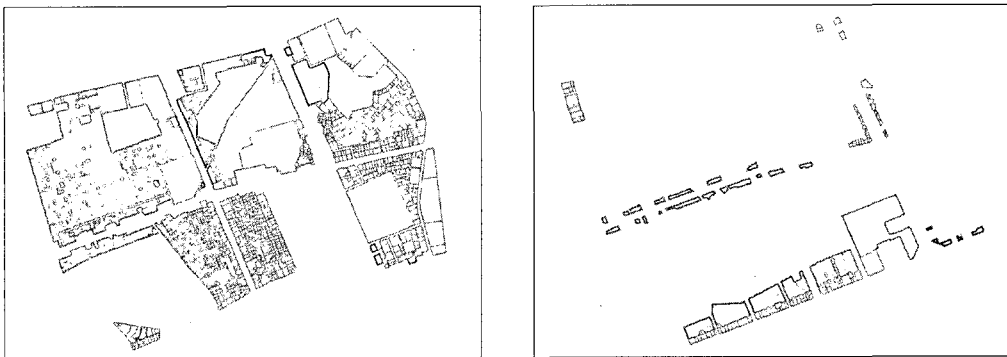


그림 7. 군집화된 유사가격권역 (상 : 일반상업지역 K=7, 하 : 일반주거지역 K=13)

군집화 기법 등을 이용하여 개별공시지가 산정 과정에서 객관적인 기준 부재로 문제점이 발생되고 있는 유사가격권역을 객관적이고 효율적으로 구획하여 보고자 하였다. 구체적으로는 혼합 클러스터링 모형을 설

계하여 객관적으로 유사가격권역을 구획하여 보았다. 연구 결과 다음과 같은 결론을 얻을 수 있었다. 먼저 유사가격권에 대하여 연구대상 지역인 강남구 삼성동 일부 용도지역별로 가격 수준에 따른 군집화가 가능하

표 5. 구간별 군집내 필지 수, 평균, 표준편차 산출 결과 (단위 : 원)

구분	군 집	필지수	평균	표준편차
일반 주거지역 (k=13)	1군집	567	1,697,460	349,975
	2군집	75	2,110,800	416,012
	3군집	103	4,234,854	474,409
	4군집	243	1,558,753	243,036
	5군집	113	3,020,885	608,569
	6군집	677	1,623,693	262,027
	7군집	33	3,348,788	151,112
	8군집	28	3,499,643	437,268
	9군집	11	1,604,545	270,130
	10군집	56	3,620,714	314,528
	11군집	45	3,752,667	507,903
	12군집	9	4,964,444	207,210
	13군집	7	7,832,857	565,097
일반 상업지역 (k=7)	1군집	39	3,747,179	489,442
	2군집	58	4,066,379	523,597
	3군집	13	8,109,231	1,015,984
	4군집	100	4,002,800	1,150,923
	5군집	56	12,703,571	933,467
	6군집	48	5,497,917	1,749,587
	7군집	13	6,024,615	1,727,860

였다. 그리고 군집내 개별필지와 일치하는 비교표준지를 조사한 결과 일반 주거지역의 경우 군집수를 대략 13으로 설정할 경우 95.8%(Rand Index : 0.861), 일반 상업지역은 군집수를 대략 7로 설정할 경우 100%(Rand Index : 0.963)가 일치하여 군집화의 타당성과 적용 가능성을 제시할 수 있었다.

이렇게 유사가격권역의 구획에 있어 혼합 클러스터링 기법을 이용한다면 지역별로 객관적이고 합리적인 유사가격권역 수를 결정할 수 있다. 이것은 현재 모호한 개념으로만 규정하고 있는 유사가격권을 정량화된 수치화 값으로 해석이 가능함으로 보다 정확하게 비교표준지를 선정할 수 있을 것이다.

또한 공시지가 산정 일련의 과정을 자동화하기 위해서는 중간 과정에서 비교표준지를 자동으로 선정하여야 하는 과정이 필수적이다. 그러나 현재까지의 연구

성과는 유사가격권을 수치적으로 해석할 수 있는 합리적인 방법론을 제시하지 못하고 있어 완전자동화에 한 계성을 드러내고 있다. 따라서 비교표준지 자동 선정과정에서 본 연구의 방법론을 적용할 경우 보다 더 정확하고 합리적으로 비교표준지의 선정이 가능할 것이다.

부수적으로는 비슷한 가격권역을 형성하는 지역별로 군집화하여 시각적, 수치적으로 나타내어 주기 때문에 현재 수작업에 의해 이루어지고 있는 비교표준지 선정에 따른 오류를 최소화할 수 있을 것이다. 또한 현재 지가업무와 관련하여 다량의 공간데이터가 구축되어 있는 환경에서 요구하고 있는 공간데이터의 다양한 활용성 강구, 방대한 데이터에 대한 효율적인 처리방법 강구, 합리적인 의사결정지원 도구의 필요성 요구 등에 부합할 수 있도록 GIS와 군집화 분석 기법을 적용시킨 연구로 향후 자가관련 다양한 응용시스템을 설계하거나 개

발할때 기반 방법론으로 효율적인 활용이 기대된다.

연구의 한계성은 다음과 같다. 군집화된 결과를 비교하는 과정에서 각 필지별 비교표준지들이 해당 소관청의 실무자들에 의해 작업된 결과이므로 본 연구의 결과를 이와 비교한다는데 있어 한계성이 있을 수 있다. 따라서 구획된 유사가격권역에서 실제 비교표준지를 선정하여 개별지의 가격을 산정하였을 경우, 가격에 대한 변화량 또는 실거래가격과의 차이 등을 적용한 방법론의 타당성 분석이 요구된다.

그리고 연구에서는 유사가격권을 구획하는데 있어 모든 토지특성을 고려하기에 한계가 있어, 지가와 거리라는 두 가지 인자를 사용하였다. 그러나 이외에 영향을 줄 수 있는 다른 토지특성 인자도 많을 것으로 사료되기 때문에 이러한 부분에 대한 연구의 보완이 필요하다.

文 獻

- 건설교통부, 2004, 2004년도 적용 개별공시지가 조사·산정 지침.
- 국토연구원, 2004, GIS기반 공간분석방법론 적용 연구.
- 김성호·조성빈·백승익, 2002, “자료융합방법의 성과에 대체수준이 미치는 영향에 관한 연구 : 몬테카를로 시뮬레이션 접근방법,” *경영과학*, 16(1), 129-141.
- 김윤식·모경주·윤인섭, 2000, “클러스터링 기법을 이용한 공정 데이터의 압축 저장 기법에 관한 연구,” *한국가스학회지*, 4(4), 58-64.
- 마상진, 1997, *고등학교 학생의 학습방식에 관한 군집분석*, 서울대학교 석사학위논문.
- 문태현, 2000, “GIS 기반 지가산정 및 시뮬레이션 시스템,” *한국지리정보학회지*, 3(2), 1-10.
- 박수홍·홍성언·김현석·김정엽, 2003, “공간 다기준 의사결정 방법을 이용한 개별공시지가 비교표준지 선정,” *한국GIS학회지*, 11(1), 1-11.
- 박성규, 1999, *토지 평가의 자동화를 위한 GIS의 적용에 관한 연구*, 조선대학교 박사학위논문.
- 아우름솔루션, 2003, *개별공시지가 자동 산정을 위한 방법론 개발*.
- 이건학, 2004, “GIS와 공간 데이터마이닝을 이용한 교통사
- 고의 공간적 패턴 분석 : 서울시 강남구를 사례로,” *대한지리학회지*, 39(4), 457-472.
- 이성규·홍성언, 2004, “K-평균 군집화 기법을 이용한 공시지가 유사가격권의 설정,” *한국지리교육학회, 지리학 연구*, 38(3), 319-327.
- 조형기·민준영·최종욱, 1996, “클러스터링 방법을 이용한 차종인식 모형,” *한국정보처리학회 논문지*, 3(2), 369-380.
- 허명희·이용구, 2003, *클레멘타인을 활용한 K-평균 군집화 결과의 재현성 평가*, SPSS White Paper.
- Brian S. E., 1993, *Cluster analysis*(3rd ed.), Halsted Press.
- Kaufman, L. and Rousseeuw, P. J., 1990, *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York.
- 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, 281-297.
- 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, 87 -63.
- Rand, W. M., 1971, Objective Criteria for the Evaluation of Clustering Methods, *Journal of American Statistical Association*, 66, 846-850.
- 교신: 홍성언, 360-764, 충청북도 청주시 상당구 내덕동 36, 청주대학교 복지·토지정보학부(이메일 : hongsu2005@hanmail.net)
- Correspondence: Sungeon Hong, Department of Social welfare·Land Information, Cheongju University, 36, Naedok-Dong, Sangdang-Gu, Cheongju, Chungbuk 360-764, Korea(hongsu2005@hanmail.net)

최초투고일 06. 1. 20
최종접수일 06. 3. 20

국한문대역 택리지

(관계 : 한국 최초의 인문지리서)

이중환 지음, 이민수 옮김, 박용수 논문과 사진, 이찬 해설

평화출판사, 2005년 12월 발행, ISBN 89-367-8085-9, 46배판 양장, 327쪽

청담(淸潭) 이중환(李重煥)의 택리지(擇里志)를 한글로 새로 번역한 『국한문대역 택리지』가 나왔다. 물론 택리지의 출판은 이번이 처음은 아니다. 20세기 초 택리지 한자본이 조선광문회(朝鮮光文會, 1912)에서 발간된 이래 한 세기에 걸쳐 한글 번역본이 여럿 출판된 적이 있다. 또한 택리지의 내용은 우리 지리학계 뿐 아니라 일반 독자들에게도 비교적 널리 알려져 있다. 그럼에도, 대부분의 책이 출판되지 오래되어 절판된 때문에 가까이 내 서가에 두고 볼만한 것을 구하기가 쉽지 않다. 가령 지리학자가 번역하고 해설한 책으로는 이영택(1975), 노도양(1972) 선생의 택리지와 있으며, 노도양 선생의 번역본은 출판사를 달리하여 몇 번 더 출간(1976, 1985)되기도 하였다. 그러나 출판연도도 미루어 짐작할 수 있듯이, 이 번역본들은 이제 절판되어 서점에서는 구할 수 없으며 도서관이나 가야 볼 수 있다. 이런 점에서 새 번역본 『국한문대역 택리지』는 반가운 일이 아닐 수 없기에, 그리고 책의 체제와 편집이 독자에 대한 배려가 뛰어나기에 이 서평을 쓰게 되었다.

『국한문대역 택리지』의 본문 내용은 다른 번역본과 마찬가지로 사민총론(四民總論), 팔도총론(八道總論)과 평안, 함경, 황해, 강원, 경상, 전라, 충청, 경기도, 복거총론(卜居總論)과 지리, 생리, 인심, 산수, 마지막으로 총론(總論)의 순으로 배열되어 있다. 책의 앞머리에는 이찬(李燦) 교수께서 쓰신 택리지 해설과 문인(文人) 박용수(朴龍洙)님의 글이 실려 있다. 이찬 선생이 작고하신지 몇 해가 지났으니, 아마도 이 해설은 선생의 유고(遺稿)로는 가장 마지막으로 출판되는 것일 듯하다. 책의 말미에는 이중환 자신을 비롯하여 정약용,

이익, 정언유 등이 쓴 모두 7편의 택리지 서문(序文)과 발문(跋文)을 국한문 대역하였고, 이어서 옮긴이의 해제(解題)가 실려 있으며, 부록에는 택리지 관련 문헌목록을 담았다. 『국한문대역 택리지』는 조선광문회본을 저본(底本)으로 삼았다고 하였으나, 조선광문회본에는 정언유(鄭彦儒)의 서문 하나만 실려있으므로, 새 번역본에 실린 발문들은 이런 저런 사본들에서 가져온 듯하다. 이토록 많은 택리지 서문과 발문이 있다는 것을 통해, 조선 후기 택리지의 인기를 미루어 짐작할 수 있다.

『국한문대역 택리지』는 현대 독자에게 가까이 다가가도록 만들어졌다. 우선 한글 번역이 기존 번역본들에 비해 보다 매끄럽게 잘 읽히고, 주석도 풍부하고 자세하다. 옮긴이의 해박한 지식과 번역에 기울인 정성이 엿보인다. 또한 번역한 문단마다 해당 원문을 옆에 나란히 실어 참조하기 쉽도록 하였다. 국한문 대역(對譯)이라지만, 한글 번역문을 크게 한자 원문은 그 옆에 작게 배열시켜 한글에 더 많은 무게를 두는 체제를 취하고 있다. 이러한 국한문 대역 체제는 다른 번역본에서는 볼 수 없는 특징이자 장점이다. 기존 번역본에서는 원문을 책의 뒷부분 또는 각 장의 마지막에 모아 놓고 있어, 원문을 찾아보기가 여간 번거롭지 않고 결국 원문 보기를 포기한 채 한글 번역문만 읽고마는 결과로 이어진다. 비교적 근래에 나온 번역본들, 가령 『한국 풍수지리학의 원전 택리지』(이익성 옮김, 1993, 개정판 2002)나 『한글세대를 위한 우리 옛글 택리지』(허경진 옮김, 1996)도 한자 원문이 책의 뒤로 모아졌고, 『청소년을 위한 택리지』(한국과학문화재단 편, 1997)는 한자 원문을 아예 실지 않았을 뿐 아니라 편집적인 면에서도 청소년에 대한 배려를 찾아보기 어렵다.

『국한문대역 택리지』는 볼거리도 풍부하여, ‘읽는 책’ 보다는 ‘보는 책’에 마음이 더 기울어지는 현대 독자들의 취향을 잘 맞추고 있다. 즉, 책 곳곳에 컬러 사진을 해설을 곁들여 실고 있어 심심하지 않고, 조선팔도를 다룬 부분에서는 해당도(道)의 고지도를 담아, 보는 재미를 더하였다. 번역본에 따라 고지도를 실은 경우가 없는 것은 아니지만, 이처럼 화려한 사진을 곁들인 책은 일찍이 없었다. 이 책에 실린 110여매의 사진은 북한 지방 등 일부를 빼고는 박용수님이 손수 찍고 해설하였기에, 사진이 들어간 부분의 택리지 내용과 호흡이 잘 맞는다.

택리지를 ‘풍수지리학의 원전’ 또는 ‘실학서’로 특징 지은 다른 번역본에 비하면, 『국한문대역 택리지』의 관제를 “한국 최초의 문헌지리서”로 붙인 것도 호감이 가는 대목이다. 서평자(書評者)가 지리학을 전공하는 탓에 응축한 생각을 하고 있는 것일까. 책 앞머리에 실린 두 편의 해설은 택리지를 이해하는 데 큰 도움을 준다. 이찬 선생이 쓰신 6쪽 분량의 택리지 해설(“택리지고”)은 선생이 생전 연구하셨던 내용이 잘 정리되어 있다. 이 해설에서 선생은 택리지의 사상적 배경을 살피고, 택리지에 나타난 지리적 개념으로 인간과 자연, 문화·역사적 지역, 현지(玄地) 개념을 들어 설명하고 있다. 박용수님의 해설 “택리지와 청화산인 이종환”에서는 청담의 생애를 자세히 고증하여, 그동안 모호했거나 잘못 알려진 점들을 바로잡고 있다.

『국한문대역 택리지』에 아쉬운 부분이 없는 것은 아니다. 택리지에서는 지명이 여럿 나오는 데, 그 위치를 찾아보기에 알맞은 지도가 없어 아쉽다. 각 도의 고지도 실은 것만으로는 지명 찾기에 충분하지 않아 보인다. 요즘처럼 지도가 범람하는 시대에도 우리 나라의 주요 도시의 위치마저도 제대로 모르는 이가 많음을 발견하곤 무척 놀라는 때가 많다. 더구나 한글 세대인 요즘 독자들이 고지도에서 한자 지명 찾기를 아예 포기할까 걱정된다. 이런 현실을 감안할 때, 현대식 지도가 고지도와 나란히 제시되었으면 좋았을 듯하다. 택리지의 영어 번역본(Yoon, 1998)에서는 현대식 한반도 지도를 실어, 자주 나오는 지명을 찾아볼 수 있도록 배려하였다. 한자를 모르는 서양인을 대상으로 한 책이니 그럴 수 있겠다고 하겠지만, 한글 세대의 사정

도 서양인의 경우와 크게 다르지 않을 것이라는 생각이 든다.

택리지의 이름을 빌린 책들도 일부 있기에, 내친 김에 짧게나마 짚어보기로 한다. 『다시 쓰는 택리지』(신정일, 2004, 2006)는 2004년에 첫권 ‘경기·충청편’이 나온 이래 지금까지 모두 다섯 권이 발간되었다. 이 책들의 내용은 글쓰기가 전국 각지를 언급한 답사기이며, 책 곳곳에 청담 이종환의 택리지 내용을 다른 고전과 함께 인용하고 있다. 사진과 고지도가 풍부하여 『국한문대역 택리지』에 맞먹고, 필자의 화려한 글쓰기는 독자를 사로잡는다. 그러나, 이 책은 읽다보면 어디까지가 이종환의 생각이고 어디가 글쓴이의 답사기인지 구분이 어려워, 택리지의 원 내용을 보려는 목적에는 썩 부합되지 않는다. 수필집으로, 기행문으로, 재미삼아 술술 읽기에 알맞은 책이다. 『소설로 읽는 택리지, 재미있는 우리 땅 이야기』(최범서, 2003)는 청담이 전국을 주유하며 보고 겪는 내용으로 된 소설로, 지금까지 3권이 나와 있다. 이 책은 청담이 정치적으로 실각한 다음 세상을 마칠 때까지 30여년 동안의 그의 행적이 거의 알려져 있지 않다는 점을 소설의 모멘텀으로 잘 활용한 듯하다.

요즘 우리는, 날센 교통과 통신 덕분에, 전국의 정보를 손 안에 쥐락펴락하고 서울과 부산 사이의 천리길을 하룻만에 아무렇지도 않게 오간다. 게다가 살아가는 모습도 하루가 다르게 변하니, ‘자기가 사는 마을을 가리는 일’(擇里)이 청담이 살던 조선 후기와는 분명 다를 수 밖에 없다. 그러나, ‘살만한 곳’을 화두(話頭) 삼아 펼친 선조의 지리관은 현대인도 음미할 가치가 충분하지 않은가. 『국한문대역 택리지』의 발간이 우리 전통지리학의 부흥에 큰 보탬이 될 것으로 기대한다.

인용한 책

- 擇里志·北學議, 노도양·이석호 역, 1972, 대양서적, 한국명저대전집 3, 서울.
- 擇里志, 노도양 역, 1976, 상, 하. 명지대학 출판부, 명지대학문고 24~25, 서울.
- 擇里志, 노도양 역, 1985, 사단법인 한국자유교양추진회 세계고전전집 5. 삼경당, 서울.

하우금

擇里志, 이영택 역주, 1975, 삼중당, 서울, 삼중당문고 152.

擇里志, 崔南善 校, 1912. 朝鮮光文會, 서울.

청소년을 위한 택리지, 한국과학문화재단 편, 1997, 서해문
집, 서울.

한국 풍수지리학의 원전 擇里志, 이익성 역, 초판 1993년,
개정판 2002년, 을유문화사, 서울.

한글세대를 위한 우리 옛글 택리지, 허경진 역, 1996, 한양
출판, 서울, 한양고전산책 9.

*Yi Chung-Hwan's Taengniiji: The Korean Classic for
Choosing Settlements. Translation with introduc-
tion by Inshil Choe Yoon, 1998, University of
Sydney East Asian Series No.12. Wild Peony,
Sydney, Australia.*

신정일, 다시쓰는 택리지 1-경기·충청편(2004). 2-전
라·경상편(2004). 3-강원·함경·평안·황해편
(2004). 4-북거총론-어디에서 살 것인가(2004).
5-우리에게 산하는 무엇인가?(2006), 휴머니스트,
서울.

최범서, 2003, 재미있는 우리 땅 이야기 1, 2, 3(판제 : 소설
로 읽는 택리지), 기린원, 서울.

하우금

서울대학교 사회과학대학 지리학과 교수 (wkhuh@snu.ac.kr)

최초투고일 06. 2. 25

최종접수일 06. 3. 20