ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2017 September: 23(3): 95~118 http://dx.doi.org/10.13088/jiis.2017.23.3.095

K-Means Clustering 알고리즘과 헤도닉 모형을 활용한 서울시 연립 · 다세대 군집분류 방법에 관한 연구

권순재
대구대학교 경영학과
(*kwonsj72@gmail.com*)

김성현
한국정보화진흥원 빅데이터센터
(*kimcon@nia.or.kr*)

탁온식
케이앤컴퍼니 데이터연구팀
(*kntak1026@gmail.com*)

정현희
대구대학교 경영학과 박사과정
(*hu912@naver.com*)

최근 도심을 중심으로 연립·다세대의 거래가 활성화되고 직방, 다방등과 같은 플랫폼 서비스가 성장하고 있다. 연립·다세대는 수요 변화에 따른 시장 규모 확대와 함께 정보 비대칭으로 인해 사회적 문제가 발생 되는 등 부동산 정보의 사각지대이다. 또한, 서울특별시 또는 한국감정원에서 사용하는 5개 또는 25개의 권역 구분은 행정구역 내부를 중심으로 설정되었으며, 기존의 부동산 연구에서 사용되어 왔다. 이는 도시계획에 의한 권역 구분이기 때문에 부동산 연구를 위한 권역 구분이 아니다. 이에 본 연구에서는 기존 연구를 토대로 향후 주택가격추정에 있어 서울특별시의 공간구조를 재설정할 필요가 있다고 보았다. 이에 본 연구에서는 연립·다세대 실거래가 데이터를 기초로 하여 헤도닉 모형에 적용하였으며, 이를 K-Means Clustering 알고리즘을 사용해 서울특별시의 공간구조를 다시 군집하였다. 본 연구에서는 2014년 1월부터 2016년 12월까지 3년간 국토교통부의 서울시 연립·다세대 실거래가 데이터와 2016년 공시지가를 활용하였다. 실거래가 데이터에서 본 연구에서는 지하거래 제거, 면적당 가격 표준화 및 5이상 -5이하의 실거래 사례 제거와 같이 데이터 제거를 통한 데이터 전처리 작업을 수행하였다. 데이터전처리 후 고정된 초기값 설정으로 결정된 중심점이 매번 같은 결과로 나오게 K-means Clustering을 수행한 후 군집 별로 헤도닉 모형을 활용한 회귀분석을 하였으며, 코사인 유사도를 계산하여 유사성 분석을 진행하였다. 이에 본 연구의 결과는 모형 적합도가 평균 75% 이상으로, 헤도닉 모형에 사용된 변수는 유의미하였다. 즉, 기존 서울을 행정구역 25개 또는 5개의 권역으로 나뉘어 실거래가지수 등 부동산 가격 관련 통계지표를 작성하던 방식을 속성의 영향력이 유사한 영역을 묶어 16개의 구역으로 나누었다. 따라서 본 연구에서는 K-Means Clustering 알고리즘에 실거래가 데이터로 헤도닉 모형을 활용하여 연립·다세대 실거래가를 기반으로 한 군집분류방법을 도출하였다. 또한, 학문적 실무적 시사점을 제시하였고, 본 연구의 한계점과 향후 연구 방향에 대해 제시하였다.

주제어 : 군집분류, 다세대주택, 연립주택, 헤도닉 모형, K-Means Clustering 알고리즘

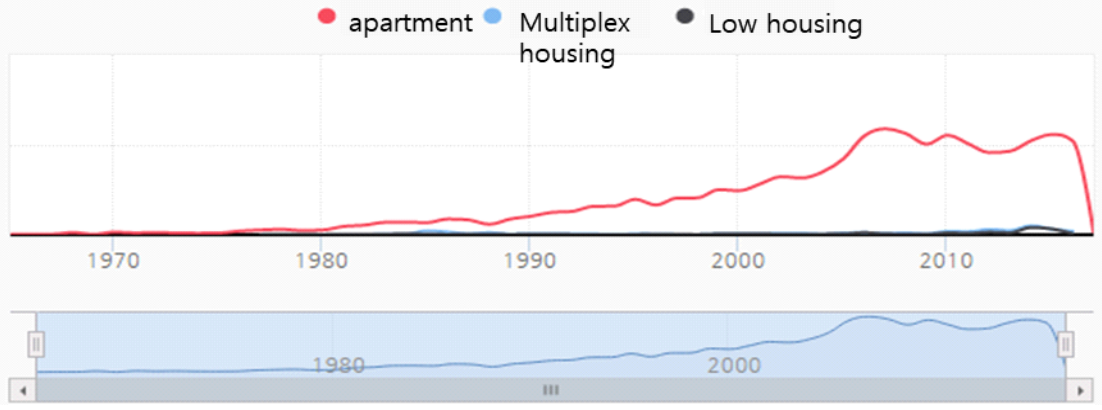
논문접수일 : 2017년 7월 31일 논문수정일 : 2017년 9월 17일 게재확정일 : 2017년 9월 20일
원고유형 : 일반논문 교신저자 : 정현희

1. 서론

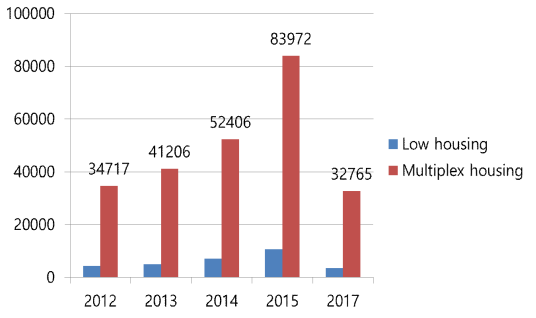
2016년 통계청 조사에 따르면, 1인 가구는 500만을 넘었으며 전체 가구의 27% 이상을 차지한다. 또한, Nam and Kim(2015)에 따르면 서울특별시는 1인 가구가 1985년에 비해 5.5배로 2010년 기준으로 1인 가구의 비율이 전체 가구의 24.4%를 차지하고, 주택 규모는 60m^2 이하의 소형 주택이 전체주택의 37.2%를 차지한다고 하였다. 이처럼 1인 가구의 증가로 인하여 주거형태도 변화하게 되어 연립·다세대주택에 관한 연구의 중요성이 대두되기 시작하였다. Jang and Kang(2014)에 따르면 연립·다세대는 아파트에 비해 낮은 가격으로 인하여 서민에게 효율적인 거주 공간을 제공했다. 연립·다세대 주택은 건축법 2조 2항에 따라 구분이 된다. 연립주택과 다세대 주택은 모두 4층 이하의 규모로 개별분양 또는 구분소유가 가능한 공동주택을 의미한다(Jang and Kang, 2014; Ryu et al., 2012; Lee et al., 2007).

기존의 부동산 시장의 연구는 아파트가 주를 이루었으며, 연립·다세대와 같은 비정형 부동산 시장의 연구가 부족하다(Ryu et al., 2012; Lee et al., 2007). 기존 연립·다세대 주택의 문헌은 경제성 분석, 계획 특성 연구, 시장 분석 분야에서 연구가 진행되어왔다. 한국감정원에서 조사 분석하여 발표하는 통계자료에 따르면, 아파트 관련 통계는 연간 371,973건인데 비해 연립·다세대 주택의 경우에는 연간 81,240건으로 전체 통계자료의 21.8% 수준에 그치고 있다. 또한, 국회도서관 소장자료 검색결과 <Figure 1>과 같이 아파트의 경우 10,114건의 학위논문과 학술기사가 검색되는 반면 연립, 다세대, 빌라의 경우 10,114건이 존재하는 등 각종 통계자료 및 학계 연구가 미흡하다.

또한, 최근 도심을 중심으로 연립·다세대의 거래가 활성화되고 직방, 다방 등과 같이 부동산 플랫폼 서비스가 성장하고 있다. <Figure 2>는 한국감정원에서 조사한 서울특별시의 연립·다세대 주택거래 현황으로 통계청을 통해 공공데이터로 제시하고 있다. 이는 최근 연립·다세대의 거래가 활성화되고 있음을 보여줌으로써, 연립·다세대 관련 연구의 필요성을 보여주고 있다.



〈Figure 1〉 Relevant Retrieval Infographic Chart in National Assembly Library



〈Figure 2〉 Transaction status of Low and Multiplex housing in Seoul (Korean Appraisal Board)

Lee and Kim(2013)의 연구에 따르면 기존의 부동산 연구가 개별공시지가와 주택 실거래가 사이의 공간적 불일치를 겪고 있다고 하였다. 이에 동 단위와 같은 공간적 변이에 대한 구체적인 분석이 이루어지지 않았고 상이한 공간 범위에 따라 개별공시지가와 주택 실거래가의 공간적 불일치를 실증적으로 확인하였다. 또한 Kim (2016)에 따르면 기존 연구는 일반적으로 전통적인 헤도닉모형을 활용하여 공간자료를 분석하였지만, 공간상의 관계가 공시지가, 매매시세 자료와 실질적인 가격의 변동 간의 괴리감을 준다고 설명하였다. 즉, 서울특별시 도시계획에 따른 5개 구역은 공간적 이질성이 있다고 하였다. 하지만, 기존 연구들은 헤도닉 모형에서 어떤 데이터를 사용하는 지와는 별개로 데이터가 수집되는 공간 범위에 따라 주택가격이 달리 산출되는 문제에 대한 연구가 다소 부족하다. 즉, 기존 연구를 통하여 상이한 공간 범위에 따라 주택 가격이 달리 산출되며, 공시가격과 실거래 가격의 불일치가 심화 될 수 있는 점에서 본 연구에서 제시하고자 하는 공간 범위의 재구성 필요성을 보여주고 있다.

또한, <Figure 3>과 같이 아파트와 연립·다세대의 플랫폼 서비스의 정보 접근성을 비교할 수

		Apartment			Low & Multiplex housing	
		Naver Real Estate	Real Estate 114	Zigbang Apartment	Zigbang	MOLIT Real Transaction Price
Informa-tion Offer	Offerings	○	○	○	○	X
	Surveyed Market Price	○	○	○	X	X
	Real Transacti-on Price	○	○	○	X	○
	Site Informati-on	○	○	○	X	X
	Plan	○	○	○	X	X
	Communi-ty	○	○	○	X	X
	Sale	○	○	○	X	X

*More than 1 million apps downloaded(Information Availability Status: O, X)

〈Figure 3〉 Information Accessibility about the Service of the Real Estate Platform

있다. <Figure 3>은 부동산 플랫폼 서비스에서는 아파트 정보 서비스에 있어 시세를 포함한 다양한 관련 정보를 제공하지만, 연립·다세대의 경우 매물정보 또는 공공정보 중 실거래가를 제공하고 있다. 아파트 정보 서비스는 초기 매물거래 플랫폼을 지나 시세 정보를 구축하여 투자분석, 부동산 플랫폼, 금융상품 개발 등 다양한 서비스로 확산이 되고 있으나, 연립·다세대의 경우 건물마다 평형, 건축구조 등이 달라 건물의 군집화가 어려워 시세 정보의 구축이 어렵다. 즉, 부동산 플랫폼 서비스에서는 시세를 포함한 다양한 관련 정보를 제공하지만, 연립·다세대의 경우 매물정보만을 활용한 거래플랫폼 또는 실거래가의 공공정보 제공의 수준으로 아파트 정보서비스보다 정보제공 범주가 협소하여 사용자의 부동산 비교판단의 어려움이 있다. 이에 본 연구에서는 실거래가 기준으로 연립다세대의 시장움직임을 알고자 하였다.

이처럼 연립·다세대는 아파트와 달리 대형 단지를 조성하고 있지 않으므로, 동질성이 높은 특정 군집을 사전에 나눠서 시세 추정 모델을 구축해야 한다. 하지만, 기존의 법정동 혹은 행정동의 체계로 구분하여 접근하기에는 권역과 권역의 경계점에서 인위적인 가격 차이가 발생하는 등 데이터 오류가 커 부동산 가격을 예측하기 위해서는 부동산 가격특성을 반영하는 군집화 연구가 필요하다. 이에 본 연구에서는 기존에 사용되고 있는 국가의 획일적인 권역이 아닌 위경도 기반으로 하여 서울특별시의 새로운 권역을 재구성하고자 한다. 즉, 기존의 행정구역에 의한 단순한 구분으로 인한 비효율적 측면이 발생했던 문제를 보다 효율적인 부동산 분석을 위해 서울특별시를 새로운 권역으로 군집하고자 하는 것이다.

즉, 서울특별시와 한국감정원에서 사용하는 5개 또는 25개의 권역 구분은 도시계획에 의한 권역 구분으로 부동산 연구를 위한 권역 구분이 아니다. 따라서 주택가격 추정이나 현실 시세와의 불일치를 설명하는 데 있어 공간 범위의 재구성이 필요하다고 판단하였다. 이에 본 연구에서는 서울특별시의 공간정보로 새로운 공간 범위를 구축하고자 한다. 본 연구는 K-Means Clustering 알고리즘을 활용하여 25개의 권역에서 16개 권역으로 구분된 서울특별시에 대한 설명력을 설명하고 헤도닉 모형을 활용하여, 연립·다세대의 공간적 불일치를 극복하기 위해 서울특별시를 대상으로 권역을 최적의 군집을 제시하고자 한다.

2. 문헌 연구

2.1 데이터마이닝

데이터마이닝은 의미 있는 패턴이나 규칙들을 발견하기 위해 다량의 데이터를 필터링하여 처리하는 과정으로 정의할 수 있다(Koo, 2016, Park at el., 2011 etc). 즉, 데이터마이닝은 다량의 데이터 에서 유용한 정보를 탐색하고 분석하여 새롭고 의미 있는 패턴이나 규칙들을 찾아내는 과정이라 정의할 수 있다(Yong et al., 2007; Berry and Linoff, 2011). 이에 반해, 추가로 Kang et al(2006)은 대량의 데이터 내에 존재하는 정보를 탐색하여, 과거 행위의 분석을 기초로 미래 행위를 예측하는 의사결정 지원을 위한 모형을 만들어내기 위하여 정보기술을 적용하는 과정으로 정의하였다. 이처럼 데이터마이닝의 기본 개념은 미래를 예측하고 데이터를 설명하는 데 도

움이 되는 도구로써 데이터 속의 구조적 패턴을 찾아 서술하고 있다(Hall et al., 2011).

최근 디지털 정보 기술의 급속하게 발전이 이루어 졌고, 다양한 시장 공간을 창출시키고 있다. 기업은 업무의 효율적인 수행을 위해 데이터베이스를 단순히 활용하는 단계를 벗어나, 오늘날의 마케팅 분야에서 데이터마이닝은 큰 비중을 차지하고 있다. 이처럼 고객 관계 관리에서는 효과적인 고객관리 전략을 개발하고 지속해서 수행하고 있는 능력이 중요하다. 이를 위해 데이터마이닝은 고객 정보를 분석하는 도구로 사용되고 있다. 데이터마이닝은 기계학습방법론은 물론 다양한 분야의 데이터마이닝 기법에는 일반적으로 의사결정나무, 인공신경망과 연관규칙 패턴 분석, 군집분석, 신경망 등의 기법들과 통계학에서 사용되는 분석기법이 포함되어 있다(Brachman and Anand, 1996).

본 연구에서 사용되는 데이터마이닝 기법은 군집분석으로, 주어진 데이터 중에서 유사한 것들을 몇몇 집단으로 분류하여, 각 집단의 성격을 파악함으로써 데이터 전체와 구조에 대한 이해를 돕고자 하는 탐색적 데이터 분석방법이다.

2.1.1 군집분석

기존 연구에 따르면 군집분석은 경영학, 관광학, 환경공학, 심리학, 공학, 통계학, 신문방송학, 경제학, 교육학 등과 같은 다양한 학문 분야에서 활용되고 있으며, 각 표본의 유사성에 기초하여 한 집단에 분류시키고자 할 때 사용되고 있다. 또한, 외국의 경우 군집분석 방법은 의약학, 해양학 분야 등 점차 응용범위가 넓어지고 있다(Yun et al, 2017; Yeom and Kim, 2011; Meghani and Knafl, 2017; Pejman et al., 2017 etc). 또한, 군집분석은 빅데이터 처리를 처리하기 용이 하며, 수십 년간 활발히 연구되어 왔고, 많은 클러스터링 알고리즘이 기존 문헌을 통해 제안되어 왔다. 이와 같이 군집분석은 이질적인 요소가 섞여 있는 대상을 그것들의 유사도에 기초해서 유사성이 높은 것들을 몇 개의 군집으로 분류하는 방법으로는 계층적 군집방법과 비 계층적 군집방법으로 나누어진다고 하였다(Kim, 2007; Lee and Kim, 2003 etc). 계층적 군집방법은 보편적인 방법으로 사용되며, 군집 간의 거리를 어떻게 정의하느냐에 따라 사용된다. 군집 간의 거리는 단일연결법, 완전연결법, 평균연결법, 중심연결법, Ward 법을 사용하고 있다. 이에 반해, 비 계층적 군집분석으로는 대표적으로 K-Means Clustering 알고리즘이 있다. 비 계층적 군집분석은 계층적 군집분석보다 계산속도가 빠르고 대량의 군집을 발견하는데 효과적인 장점이 있지만, 초기 중심값을 임의의 k개의 군집 수로 사용하기 때문에 최적의 군집화 결과를 얻기 힘들다고 하였다. 또한, 군집분석은 군집의 개수가 아직 명확한 기준이 없어 연구자의 해석에 따라 차이를 가질 수 있으며, 판정 기준은 R-Square. Pseudo-F, Cubic Clustering Criterion 등이 있다(Anderberg, 2014; Romesburg, 2004 etc) 또한, 본 연구에서는 군집분석을 대상의 유사성을 측정하여, 유사성이 높은 대상집단을 분류하는 법으로 비 계층적 군집방법 중 K-Means Clustering 알고리즘을 사용한다.

2.1.2 K-Means Clustering 알고리즘

대량 데이터에 대한 클러스터링 기법 중 일반적으로 사용되는 비 계층적 군집방법은 K-Means Clustering 알고리즘이다. K-Means Clustering 알

고리즘은 패턴들과 그 패턴이 속하는 클러스터의 중심과의 평균 유클리디안(Euclidean) 거리를 최소화하는 것으로 주어진 데이터를 특정 성질에 기초하여 K개의 군집으로 나누는 방법이다(Lee, 2012; Jain, 2010; Lloyd, 1982; Macqeen, 1967 etc). 즉, 데이터의 묶음의 분산도를 최소화하는 알고리즘으로 K-Means의 성능은 무게중심 즉, 초기 중심 값을 어떻게 선정하는 가와 어떠한 데이터부터 처리하는가에 따라 달라진다고 하였다(Lee and Lee, 2011).

Macqeen(1967)이 제안한 K-Means Clustering 알고리즘은 입력 값을 k로 취하고 군집 내 유사성은 높고, 군집끼리 유사성이 낮게 되도록 n개 객체의 집합을 k개의 군집으로 분해하고, 유사성은 객체들의 평균값으로 측정한다고 한다. 즉, K-Means Clustering 알고리즘은 특정 성질의 데이터들이 유사성을 기초로 한 고정된 수인 k의 군집을 찾는 알고리즘을 의미한다. 이와 같은 작업을 통해 사용자가 설정한 임의의 임계치를 만족할 때까지 반복적으로 군집분류를 진행한다(Arthur and Vassilvitskii, 2006). K-Means clustering 알고리즘의 전체 분산은 <Eq. 1>과 같이 계산한다.

$$< \text{Eq. } 1 > \quad \text{V} = \sum_{i=1}^{k} \sum_{j \in S_i} |X_j - \mu_i|^2$$

K-Means는 초기 중심의 선정에 따라 성능이 크게 달라진다. 기존의 알고리즘은 초기 중심을 선정할 때 무작위로 설정되어 왔다. 하지만, 무작위로 선정된 초기 중심에서 진행되어온 클러스터링은 편차가 크므로, 선행 연구를 통해 초기 중심 설정의 문제점을 해결하고자 하였다(Lee, 2012).

본 연구에서의 K-Means Clustering 알고리즘은 주어진 데이터를 특정 성질에 기초하여 K개의 군집으로 나누는 방법으로 정의하였다. 또한, 기존의 초기 중심 선정방법인 무작위 추출방법이 아닌, 고정된 초기값으로 초기중심설정의 문제점을 해결하고자 하였다.

2.2 헤도닉 모형

헤도닉 모형은 Rosen(1974)에 의해 이론적으로 정의된 헤도닉 가격모형(Hedonic Price Model)으로 시장에서는 직접 거래되지 않는 요인이 특정재화의 가격에 영향을 미친다는 가정에 따라 주택 가격 추정에 있어 이론적 토대를 마련한 분석방법이다. 부동산의 가치를 평가하는 분석방법으로써 부동산학계에서 자주 사용하고 있으며 부동산의 가치뿐만 아니라 부동산 환경 가치, 주택가격지수 등 많은 연구에서 헤도닉 모형이 사용되고 있다(Lee, 2008; Kim and Chung, 2010). 헤도닉 모형은 기존 연구를 통해 분석대상을 다양하게 사용해 왔다. Kwon and Kim(2006)은 헤도닉 가격모형을 주택가격에 영향을 주는 주택특성변수들을 독립변수로 두고 회귀함수 식을 추정하여 주택가격을 추정하는 방식으로 사용하였다. 또한, 부동산공시지가는 헤도닉모형을 의해 토지가격비준표를 작성하기도 한다(Seo and Kwak, 2014). 그 외 선행연구로는 해수변의 조망가치 추정, 공원녹지의 가치 추정 등과 같이 근린공원을 대상으로 가치 평가를 한 연구가 있으며 부동산가격에 영향을 미치는 주변 요인분석 연구, 부동산학뿐만 아니라 경영학, 관광학, 광고학 등 다양한 분야에서 헤도닉 모형을 활용한 연구 등이 있다(Kim et al., 2017; Kim, 2017; Jung and Lee, 2017; Yang et al., 2016;

Kim and Chung, 2010; Leonard et al., 2016; Benfratello, 2009 etc)

Lee(2008)의 연구에 의하면 특정가격 추정 상의 문제점을 변수선정 문제, 모형설정 문제, 모형추정 문제로 나누어 제시하여, 더욱 깊은 이해의 필요성을 강조하였다. 또한, 헤도닉 가설에 근거하는 헤도닉 모형의 설정을 위해 선형함수, 반 로그함수, 이중 로그함수 등 다양한 형태의 함수들이 사용되고 있다(Yang et al., 2016; Lee, 2008; Melpezzi, 2003).

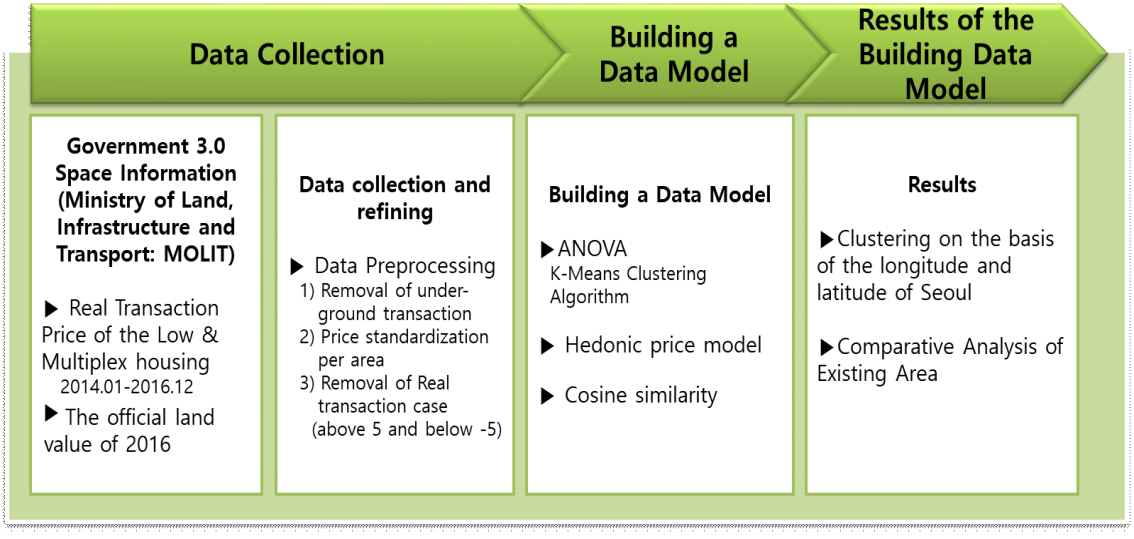
3. 연구설계 및 분석

본 연구는 국토교통부의 자료를 활용하여 기존의 서울특별시 권역을 연립·다세대 실거래가를 위해 군집분석 중 K-Means Clustering 알고리즘과 헤도닉모형을 통해 개선하기 위한 연구이다.

3.1 연구설계

지역유형을 구분하기 위한 연구방법은 단순통계기법과 다변량 분석기법 중 하나인 주성분 분석, 요인분석, 군집 분석 등이 있다. 본 연구에서는 K-Means Clustering 알고리즘과 헤도닉 모형을 통해 기존의 데이터 오류의 문제점을 극복하기 위해 서울특별시를 군집화하고자 한다. 군집분석에는 비계층 군집분석을 활용하고, 연립·다세대 의 주택가격모형을 위한 군집화로, 주택가격모형은 헤도닉모형을 활용하고자 한다. 이에 본 연구의 데이터 분석은 <Figure 4>와 같은 방법으로 진행되었다.

본 연구의 데이터는 정부 3.0 공공정보를 활용하였으며, 공간 범위는 서울특별시로 한정하였다. 데이터 전처리 후, K-Means Clustering 알고리즘을 통하여 최적의 군집을 도출하기 위해 기존의 초기 중심 선정방법인 무작위 추출방법이 아닌, 서울시 행정구역 25개를 고정된 초기값으



〈Figure 4〉 The framework of this study

로 선정하여 분석 및 헤도닉모형을 통한 회귀분석을 진행하였다. 또한, 데이터 분석에는 R program을 사용하였다.

앞서 연구 진행을 하기 전에, 본 연구에서 활용하는 본 연구의 공간 범위에 대해서 알아보고자 한다.

3.2 본 연구의 공간 범위

본 연구 공간 범위는 <Figure 5>로 서울특별시로 선정하였다. 서울특별시는 자치구가 25개로 동대문구, 성동구, 은평구, 마포구, 용산구, 서초구 외 20개가 있다. 기존 연구에 따르면, 서울특별시의 25개 자치구를 도시계획기준에 따라 5개의 권역으로 구분되어 부동산연구가 진행됐다.

〈Figure 5〉 The spatial extent of this study

이는 <Figure 6>과 같이 서울특별시와 한국감정원에서 사용되는 5개의 권역은 지리적인 위치를 포함한 자연적, 물리적 환경뿐만 아니라, 관련 계획, 교육 환경, 주거지와 거주인구의 특성 등 종합적인 고려를 통해 서울특별시의 도시계획 기준에 따라 나뉘었다. 이에 5개 권역은 기존의 대다수 연구에서 사용되어 왔다. 5개 권역은 도심권역, 동남권역, 동북권역, 서남권역, 서북권역으로 구분하며, 5개 권역에 대한 자치구는 다음과 같이 구분된다. 도심권역에는 종로구, 중구, 용산구이며, 동북권역에는 동대문구, 성동구, 중랑구, 광진구, 노원구, 성북구, 강북구, 도봉구가 있다. 또한, 서북권역에는 서대문구, 마포구, 은평구 서남권역에는 동작구, 관악구, 구로구, 영등포구, 금천구, 양천구, 강서구, 동남권역에는 서초구, 강남구, 송파구, 강동구로 구분할 수 있다. 이와 같이, 서울특별시에서는 5개 권역을 생활권역으로 구분하여 도시계획을 진행하고 있다.

〈Figure 6〉 Five Major Areas of Seoul according to the Urban Planning Criteria

이와 같이 기존 연구에서 활용해 온 5개 권역이 아닌 위경도를 기반으로 25개 자치구를 초기값으로 두고 연립·다세대의 실거래가 추정에 맞는 권역으로 구분하고자 한다. 즉, 변수설정에 있어서 연립·다세대의 실거래 특성을 위한 헤도닉모형을 제시하고 이에 대한 연구모형을 K-Means Clustering 알고리즘을 활용하여 구축하고자 한다. 본 연구에서 활용하고자 하는 주택가격모형에 대한 적절한 서울특별시 군집을 도출하고자 한다.

3.3 연구분석

3.3.1 데이터 수집

정부 3.0 정책에 따라 2015년 3월 건축물 대장을 시작으로 토지 대장을 비롯한 다양한 공간정보가 지속해서 민간 개방되어 활용할 수 있게 되었다. 즉, 지도에 표현 할 수 있도록 위치, 분포 등과 같은 기초정보와 기준을 제시하는 공간정보는, 최근 연립·다세대 주택 실거래가의 개방으로 인해 부동산정보를 더욱 쉽게 이용이 가능해졌다.

이와 같이 정부3.0에서 개방된 정보 중 국토교통부에서 조사한 2014년 1월부터 2016년 12월까지의 총 3년간의 서울특별시 연립·다세대 실거래가 데이터와 2016년 개별공시지가를 활용하였다. 실거래가 데이터의 개수는 132,707건이다.

이에 정부 3.0에서 개방된 정보를 기반으로 데이터를 수집, 정제하여 분석하였다. 데이터전처리로는 실거래가 데이터에서 지하거래 제거, 면적당 가격 표준화 및 5 이상 -5 이하 실거래 사례 제거를 진행하였다. 이와 같이 본 연구에서는 데이터 수집과정에 있어 데이터 전처리 과정을 통하여 데이터를 정제하여 추출하였다. 이를 통하여 본 연구에서는 132,707건에서 데이터 전처리를 통해 126,759건의 데이터로 분석을 실시하였다.

데이터 분석에 활용된 126,759건 데이터 중 7개 샘플은 다음 <Table 1>과 같다.

이와 같이 <Table 1>은 총 16개의 변수로 나뉘어 제공되었다.

〈Table 1〉 Data samples of this study

No	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	log_std_unit value
1	Gaepo-dong, Gangnam-gu, Seoul	169	11	(169-11)	32.13	18.58	1~10	12300	3	1995	Seolleung-ro 10-gil	201401	127.0608	37.48345	-0.21891
2	Gaepo-dong, Gangnam-gu, Seoul	169	11	(169-11)	32.19	18.62	1~10	12300	3	1995	Seolleung-ro 10-gil	201401	127.0608	37.48345	-0.22386
3	Gaepo-dong, Gangnam-gu, Seoul	169	11	(169-11)	34.02	19.68	1~10	12100	1	1995	Seolleung-ro 10-gil	201401	127.0608	37.48345	-0.41409
4	Gaepo-dong, Gangnam-gu, Seoul	169	11	(169-11)	38.21	22.1	1~10	14800	2	1995	Seolleung-ro 10-gil	201401	127.0608	37.48345	-0.1878
5	Gaepo-dong, Gangnam-gu, Seoul	169	11	(169-11)	40.02	23.14	1~10	15500	2	1995	Seolleung-ro 10-gil	201401	127.0608	37.48345	-0.18799
6	Gaepo-dong, Gangnam-gu, Seoul	169	11	(169-11)	40.8	23.6	1~10	13000	1	1995	Seolleung-ro 10-gil	201401	127.0608	37.48345	-0.70597
7	Gaepo-dong, Gangnam-gu, Seoul	170	5	(170-5)	56.7	26.78	11~20	21800	4	1988	Seolleung-ro 10-gil	201401	127.0607	37.48396	-0.20743

X1은 주소 구분으로, 예를 들어 서울특별시 강남구 개포동의 데이터임을 말해준다. 또한 X2는 본 번, X3는 부 번, X4은 단지명, X5는 전용면적(제곱미터), X6은 대지권 면적(제곱미터), X7은 거래일자, X8은 만원 단위로 실거래가, X9는 층, X10은 건축연도, X11은 도로명 주소, X12은 거래 연월, X14는 X 좌표로 위도, X15는 Y 좌표로 경도, log_std_unit value는 표준화된 ln(단위당 가격)이다.

예를 들어 1번째 데이터는 서울특별시 강남구 개포동 169-11번지(선릉로 10길)에 있는 연립·다세대 주택은 3층건물로 3층 32.13m^2의 전용면적, 18.58m^2의 대지권면적을 가지고 있으며, 1995년에 건축되었다. 2014년 1월 1일에서 10일 사이에 12,300만원에 실거래된 건축물으로써 표준화된 ln(단위당 가격)은 -0.21891라고 해석할 수 있다.

이에 <Table 2>는 기초통계 분석으로 서울특별시를 25개 자치구로 구분하여 실거래가 데이

〈Table 2〉 Analysis of basic statistics by 25 areas of this study

No	Borough	avg	min	max	median
1	Gangnam-gu	745.37	74.02	2308.05	669.94
2	Gangdong-gu	522.08	70.65	1554	502
3	Gangbuk-gu	340.21	86.57	1203.29	308.62
4	Gangseo-gu	361.78	111.43	1393.12	327.2
5	Gwanak-gu	449.01	94.49	1768.03	391.74
6	Gwangjin-gu	542.83	98.91	2481.23	501.18
7	Guro-gu	366.03	88.23	1025.06	336.08
8	Geumcheon-gu	380.79	78.5	1061.78	350.48
9	Nowon-gu	373.74	118.42	889.24	334.7
10	Dobong-gu	319.1	84.4	1117.15	297.03
11	Dongdaemun-gu	451.82	92.01	1481.48	430.73
12	Dongjak-gu	490.53	142.63	1762.98	437.93
13	Mapo-gu	514.78	63.21	1667.89	480.55
14	Seodaemun-gu	386.02	98.21	1404.15	366.88
15	Seocho-gu	692.85	121.47	2556.33	649.88
16	Seongdong-gu	568.84	108.55	2292.02	524.17
17	Seongbuk-gu	392.68	99.08	1878.79	366.78
18	Songpa-gu	543.19	148.83	1712.1	518.42
19	Yangcheon-gu	370.4	83.02	1296.36	334.45
20	Yeongdeungpo-gu	478.54	78.99	1442.38	425.54
21	Yongsan-gu	743.69	241.47	2723.97	647.92
22	Eunpyeong-gu	359.8	80.17	1100.72	326.02
23	Jongno-gu	437.7	147.29	1618.64	425.7
24	Jung-gu	467.32	139.01	1217.32	446.35
25	Jungnang-gu	425.03	139.03	1399.04	401.41

터를 전용면적으로 나눈 단위당 가격의 평균값, 최소값, 최대값, 중앙값을 나타내고 있다. 분석값은 소수점 둘째 자리까지 반올림하여 나타내었다.

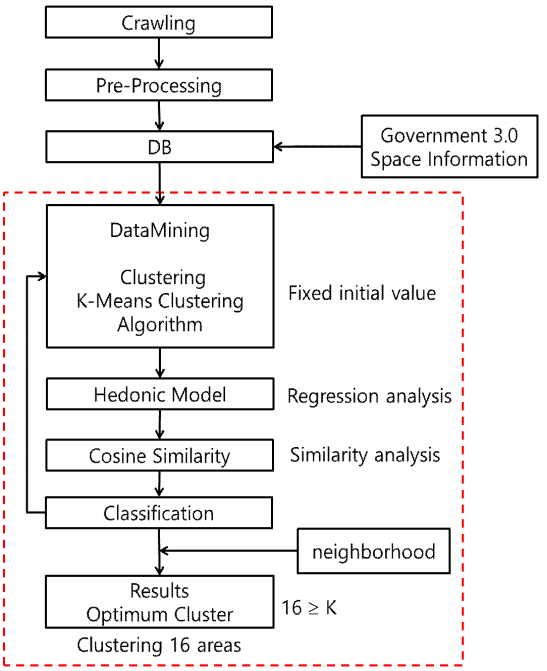
<Table 2>를 살펴보면 평균 단위당 면적 수치는 강남구, 용산구 서초구 순으로 강남구가 제일 높으나, 가장 높은 단위당 면적 수치는 용산구, 서초구, 광진구 순이다. 또한, 중앙값은 강남구, 서초구, 용산구 순으로 높다. 이와 같이 단위당 면적은 서초구, 용산구가 상위랭크에 있음을 보여준다. 평균 단위당 면적 수치가 낮은 순은 도봉구, 강북구, 은평구 순으로 동북권과 서북권의 북쪽이 가장 면적당 단위가 낮은 것으로 보인다.

즉, 본 연구에서는 본 연구에 필요한 자료를 정부 3.0 공공정보를 통해 건축물, 위치, 주변 환경, 매매가, 임대가 등의 공간정보를 수집하였다. 이를 분석데이터의 이상치 제거와 같이 데이터 전처리 과정을 거쳐 데이터를 추출하였으며, 추출된 공간정보를 활용하여 본 연구의 목적인 연립·다세대주택의 실거래가를 활용하여 부동산연구에 적절한 서울특별시 공간 범위를 군집 분류하고자 한다.

3.3.2 데이터 분석모형 구축

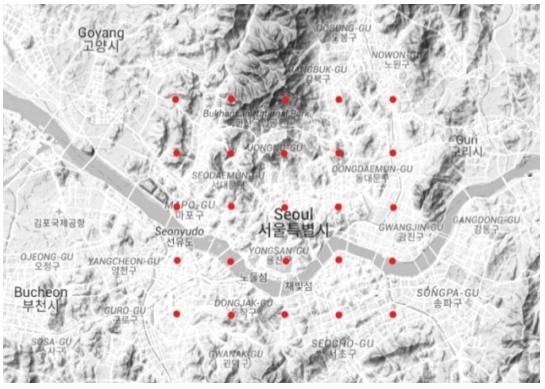
본 연구는 기존 부동산 연구의 문제점인 공간적 오류를 연립·다세대 주택 실거래 예측 모형을 활용하기 위해 최적의 군집화 방법을 제시하기 위해서 다음 아래 <Figure 7>과 같이 분석을 진행한다.

첫째, 데이터 전 처리된 실거래 데이터의 위경도 좌표로 K-Means 클러스터링을 수행한다. 초기값은 서울특별시 행정구 25개 구역에 따라, k값을 25로 임의로 지정하여 사용하였다.



〈Figure 7〉 Research Method of this study

본 연구에서 사용된 군집 분석 방법은 계산 속도가 빠르고 대량의 자료에서 군집을 발견하는데 있어 효과적인 K-Means Clustering 알고리즘이다(Brain, 1993). 본 연구에서는 대량의 공공정보를 활용하여, 군집을 발견하고자 하기 때문에 K-Means Clustering 알고리즘을 선정, 이에 연구를 진행하였다. K-Means Clustering 알고리즘은 초기값을 랜덤으로 설정하여 실행할 때 군집별로 중심점 결정 결과가 다르게 나오는 문제점을 고정된 초기값 설정(k=25)으로 인해 같은 결과로 나오게 하여 문제점을 해결하였다. 본 연구에서는 데이터를 실거래가에 기초하여 자치구 단위인 25개의 군집을 초기값으로 선정, 이에 대해 <Figure 8>은 우선적으로 본 연구의 초기 중심값을 도식화하여 산출했다.



〈Figure 8〉 Any initial center value of this study

<Figure 8>과 같이 본 연구에서는 초기값을 25로 두고 이를 통해 초기 중심점을 임의로 지정하였다.

둘째, 서울특별시 25개 군집은 연립·다세대의 가장 대표적인 네 가지 변수(전용면적, 사용연수, 대지권 면적, 개별공시지가)의 영향력의 유사성을 파악하기 위해 헤도닉 모형을 활용하여 회귀분석을 수행하였다. 즉, 주요변수의 영향력의 정도를 확인하고, intercept 항을 포함한 변수의 표준화 계수가 서로 유사하다고 판단하는 지역끼리 묶기 위해 헤도닉 모형을 사용하였다. 본 연구에서 헤도닉 모형은 연립다세대 특성에 맞는 실거래가 예측모형으로, 기존의 문헌에서 전무하였으나, 이에 본 연구에서는 기존의 아파트 실거래가 예측모형을 변수 4개를 통해 도출하였다.

일반적으로 헤도닉 가격모형의 회귀방정식은 다음의 <Eq. 2>와 같다.

$$< \text{Eq. } 2 > \quad P = h(S, N, L)$$

<Eq. 2>에서의 P는 재화의 가격이며 S, N, L은 개별 특성, h( )는 회귀식의 함수 형식을 나타낸다. 본 연구에서는 헤도닉 가격 모형을 연립·다세대주택 실거래가에 영향을 주는 특성 변수들을 독립변수로 두고 회귀함수 식을 추정하여 실거래가를 추정하는 방식으로 활용했다.

본 연구에서 <Eq. 2>를 활용한 헤도닉 모형은 다음 변수를 활용하였다.

종속변수는 ln(실거래가)로 두고, 독립변수는 실거래가에 영향을 주는 특성변수로써 ln(전용면적), ln(대지권 면적), ln(개별공시지가), ln(경과연수) 총 4개의 변수가 있다.

본 모형에서 사용된 개별 공시지가는 국가 공간정보 포털에서 2016년 개별공시지가를 주소별 Key 값으로 맵핑하여 사용하였다. 또한, 경과 연수는 거래연도에서 건축연도를 뺀 연수로 계산되며 이와 같이 헤도닉모형변수를 설정하였다.

셋째, 표준화 계수의 군집별 면적당 가격의 합산 평균을 통해 코사인 유사도를 계산하였다. 즉, 군집에 대한 intercept 항과 4개의 표준화 베타를 코사인 유사도 공식에서 사용한다. 본 연구에서 사용한 코사인 유사도는 <Eq. 3>과 같이 정의할 수 있다. 코사인 유사도는 군집 간 유사성을 비교하려는 방법으로써, 내적 공간의 두 벡터 간 각도의 코사인값을 이용하여 측정된 벡터 간의 유사한 정도를 의미한다(Recardo and Berthier, 1999).

$$< \text{Eq. } 3 > \quad \frac{\sum_{i=1}^{4} A_i \times B_i}{\sqrt{\sum_{i=1}^{4}(A_i)^2} \times \sqrt{\sum_{i=1}^{4}(B_i)^2}}$$

넷째, 이웃하면서 코사인 유사도 점수가 임의의 기준값 0.999 이상이면 결합하였다. 즉, 같은 행정구역에 속하는 거래사례를 공유하는 군집을 이웃한 것으로 본다. 예를 들어, 1번 군집 내에

행정구역이 강남구인 거래사례가 있으며, 4번 군집 내에도 강남구인 거래사례가 있을 경우 1번과 4번의 군집은 이웃한 것으로 보았다.

마지막으로, 군집이 16개 이하가 될 때까지 1-4번의 과정을 반복하였다.

즉, 패턴이 속하는 클러스터의 중심과의 평균 유클리디안(Euclidean) 거리를 최소화하는 것으로 본 연구는 활용된 정부 3.0 공공정보를 실거래가에 기초하여 군집이 1000개의 실거래 사례가 나오고 유사성을 확인하는 작업을 반복하였다. 이때 군집의 개수는 16개 이하로 16개 이상인 경우에는 유사성이 높지 않은데 묶이는 경우가 발생하여 집의 수가 16개 이하로 나올 때까지 작업을 반복한 것이다

3.3.3 구축된 데이터 모형

본 연구에서는 K-Means Clustering 알고리즘과 헤도닉 모형을 활용하여 서울시를 군집 분석하였다. Kim et al(2015)는 군집분석에서 가장 먼저 결정되어야 할 것은 군집의 개수라고 설명하였다. 이는 군집의 개수를 몇 개로 정하냐에 따라 분석결과를 연구자 임의대로 유도할 수 있는 위험성이 있기 때문이며, 개수를 정하는 과정에서는 객관적인 방법을 사용하는 것이 중요하다고 하였다. 각 단위 지역은 유형화 군집 중심에 가까운 군집으로 할당되게 된다. 본 연구에서는 서울특별시와 한국감정원에서 사용한 5개 구역보다 K-Means Clustering 알고리즘을 사용하여 25개의 군집에서 연립·다세대 실거래가 데이터를 활용해 도출한 16개 구역이 최적의 군집 수인지 비교 분석하고자 한다. 본 연구의 적합성을 판단하기 위해 R^2과 mse(Mean Square error)를 활용하였다. 본 데이터 분석모형을 활용하여 초기값 25개에 대한 Cluster 결합이 있기 전의 각각의 클러스터에 대한 R^2값과 mse값, 최종 16개 군집에 대한 R^2값과 mse값, 서울특별시와 한국감정원에서 사용된 기존의 5개 또는 25개의 권역에 대한 R^2값과 mse값을 비교 분석하였다. 이는 <Table 3>와 같이 나타낼 수 있다. 분석 값은 소수점 넷 째 자리까지 반올림하여 나타내었다.

<Table 3>를 살펴보면, 평균 R^2는 n개(집단 수)의 평균값으로 최종 군집된 16개의 클러스터가 서울특별시와 한국감정원에서 구분한 5개 또는 자치구 25개 권역에 대해 설명력이 약 1%에서 2%가량이 증가하는 것으로 나타났다. 또한, 평균 mse는 (실제값-예측값)^2 를 평균하여, 루트를 씌운 값으로, n개의 mse 평균값을 의미한다. 서울특별시, 한국감정원에서 사용된 5개 권역에 대한 헤도닉모형에 비해 본 연구의 헤도닉모형은 0.5% 정도 개선되었지만, 자치구 기준 25개의 권역과 초기에 군집한 25개의 군집에 비해 0.4% 정도 평균값이 높아졌음을 알 수 있다. 즉, K-Means Clustering 알고리즘과 헤도닉 모형을 활용한 연립다세대 분류 방법은 기존 5개 또는 25개 권역과 초기 군집보다는 설명력은 높지만, mse 평균은 5개 권역에보다는 개선이 되었지만,

〈Table 3〉 Comparative analysis result table

	Clustering 25 areas	Clustering 16 areas	Existing 25 areas	Existing 5 areas
Average R^2	0.7375	0.7446	0.731	0.7248
Average mse	0.0427	0.0466	0.042	0.0517

25개 권역과 초기 군집에 대해서는 개선이 되지 않았다.

이는 헤도닉 모형을 통하여, 실제값이 이에 5개 또는 25개의 권역에 비해 K-Means 알고리즘과 헤도닉 모형을 활용한 본 연구방법으로 도출된 권역이 평균 74.5%의 설명력을 갖고 있으며, 16개 권역으로 나누어진 서울특별시 권역은 연립·다세대 주택의 가격 예측 및 측정이 용이해 졌다고 보인다.

본 연구에서는 헤도닉 모형은, 주요변수의 영향력의 정도를 확인하고, intercept 항을 포함한 변수의 표준화 계수가 서로 유사하다고 판단하는 지역끼리 묶기 위해 사용되었다. 변수의 수를 늘릴수록, 코사인 유사도가 1에 가까워져 유사정도를 판단하기 어려워져 헤도닉 모형의 변수를 4개로 한정하여 연구를 진행하였다. 또한, 본 연구에서는 유사성의 척도인 코사인 유사도 공식을 사용하여 유사성 0.999 이상의 유사성을 갖고 이웃하는 군집 수가 최종적으로 16개의 구역으로 나눌 수 있었다. 즉, 일반 클러스터링을 통해 축소했을 경우 유사성이 떨어지는 그룹끼리 묶이게 되었던 그룹을 본 연구방법을 통하여 더욱 유사성이 높은 그룹끼리 묶을 수 있었다.

이처럼 본 연구는 서울특별시 연립·다세대 실거래가 데이터를 활용하여 25개의 자치구로 1차 군집화를 통하여 22개, 2차 군집화를 통하여 20개, 3차 군집화를 통하여 16개로 축소되었으며, 분석결과는 <Table 4>으로 정리할 수 있다.

이와 같이 <Table 4>은 K-Means Clustering 알고리즘과 헤도닉 모형을 사용하여 나타난 분석

〈Table 4〉 Analysis results of this study using K-Means Clustering algorithm and Hedonic Price Model

	New cluster 1	New cluster 2	New cluster 3	New cluster 4	New cluster 5	New cluster 6	New cluster 7	New cluster 8
Min.	121.47	74.02	70.65	84.4	98.91	83.02	142.63	63.21
1st Qu.	535.21	569.55	436.21	267.49	343.62	284.52	417.39	359.43
Median	608.26	693.83	512.82	303.65	424.84	329.43	546.36	437.15
Mean	652.27	764.17	533.97	332.31	459.5	364.41	589.52	485.49
3rd Qu.	724.76	916.85	607.03	363.19	544.62	413.08	710.98	578.24
Max	1700.81	2308.05	1712.1	1203.29	2481.23	1393.12	2556.33	1768.03
std.div	179.79	276.06	143.63	106.69	161.57	122.47	221.44	181.45
n	1892	3672	13502	11570	16899	16917	5745	13080
	New cluster 9	New cluster 10	New cluster 11	New cluster 12	New cluster 13	New cluster 14	New cluster 15	New cluster 16
Min.	78.5	92.01	136.71	80.17	92.59	98.21	139.01	99.08
1st Qu.	292.45	391.69	437.38	293.4	281.37	323.84	491.71	296.94
Median	348.88	470.18	523.72	354.08	323.77	408.09	710.42	378.41
Mean	384.53	502.37	579.98	376.37	357.3	414.44	789.34	401.17
3rd Qu.	447.95	582.2	656.24	435.03	410.44	484.4	1009.19	472.6
Max	1530.32	1967.01	2133.33	1240.99	1100.72	1262.06	2723.97	1878.79
std.div	131.01	167.85	216.92	119.38	114.43	122.94	383.94	146.01
n	11351	4546	1731	4967	13425	1796	1799	3349

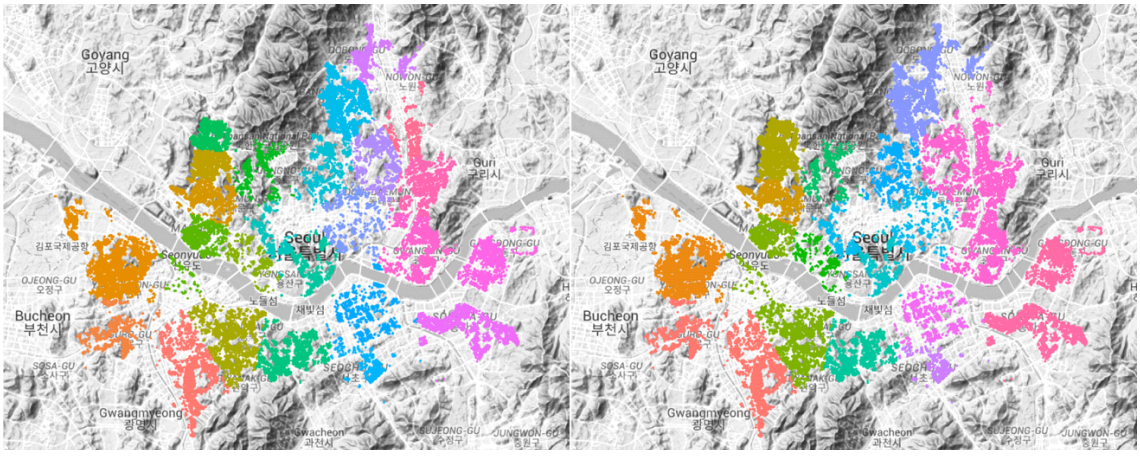
결과이다. 이는 소수점 둘째 자리까지 반올림하여 나타내었다. 군집 결과를 살펴보면 실거래 수는 각각 1000개 이상으로 군집화되었으며 군집에 따른 실거래가(단위당 가격)의 평균값, 최대값, 최소값, 중앙값, 1분위 값, 3분위 값, 표준편차 값, 집단 수를 의미한다.

<Table 4>과 마찬가지로 서울특별시와 한국감정원에서 사용된 5개 권역에 대해 본 데이터 분석모형에 따라 계산하면 <Table 5>로 나타낼 수 있다. 분석 값은 소수점 둘째 자리까지 반올림하여 나타내었다. <Table 4>과 <Table 5>를 본 연구방법으로 적용하여 분석결과치를 비교하면 <Table 4>에 비해 <Table 5>가 표준편차가 적다. 즉, 단위당 가격의 범위가 작게 생겨, 클러스터간에 이질적인 결과가 있음을 보여준다.

이와 같이 본 연구에서는 서울 특별시를 연립·다세대 실거래가 데이터를 기반으로 K-Means Clustering 알고리즘과 헤도닉 모형을 활용하여 16개 구역으로 구분하였다. 이는 <Figure 9>와 같이 나타낼 수 있다. <Figure 9>는

〈Table 5〉 Analysis Results of Existing Research

	The Heart of the City Zone	The Northwest Zone	The Southwest Zone	The Southeast Zone	The Northeast Zone
Min.	139.01	63.21	78.5	70.65	84.4
1st Qu.	406.38	297.12	298.99	469.59	299.49
Median	504.31	366.51	361.06	558.2	373.96
Mean	594.51	400.57	403.14	603.03	413.95
3rd Qu.	686.5	473.03	466.03	689.34	490.34
Max.	2723.97	1667.89	1768.03	2556.33	2481.23
std.div	295.66	143.81	154.53	205.83	159.83
n	5824	25851	39870	21448	33255



〈Figure 9〉 Existing 25 areas(left) and Clustering 16 areas(right)

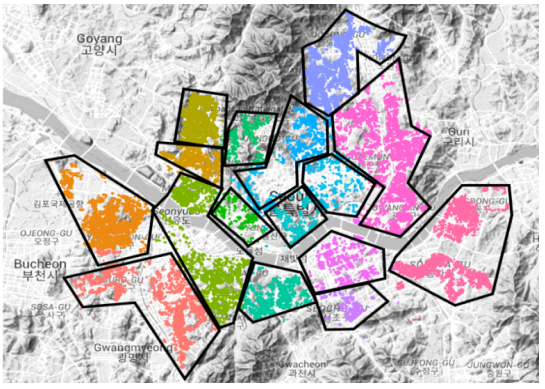
왼쪽 그림인 25개 구역의 군집 분류가 데이터 분석을 통해 오른쪽의 16개 구역 군집분류와 같이 유사성이 높은 군집끼리 분류된 그림이다.

세부적인 결과로는 기존의 구역 구분과 달리, 용산구는 한남동을 포함하는 동부지역과 서부지역을 구분하게 되었다. 또한, 중구와 가까운 지역은 중구지역의 실거래가 특성과 유사한 경향이 있어 군집화되었으며, 서초구는 서쪽 지역과 동작구의 동쪽 지역이 실거래가 특성이 유사하여 군집을 형성하였다. 동작구 서쪽 지역은 관악구, 영등포구, 마포구가 실거래와 유사한 구역으로 군집되었다.

3.4 분석결과 요약

서울특별시에서는 정책 수행을 위해 5개 권역으로 나누었으며, 공간계획, 주택계획 등 정책계획, 기존의 부동산연구 역시 5개 권역 분류를 바탕으로 한다. 이는 부동산 시장의 공간적인 실체를 고려하지 않았으며 공간적 이질성이 발생하고 부동산 실거래가 데이터와 공간 간의 괴리감으로 인해 부동산연구에 최적화된 권역 분류가 아니다. 즉, 상이한 공간 범위에 따라 주택 가격이 달리 산출되며, 공시가격과 실거래 가격의 불일치가 심화 될 수 있는 점에서 공간 범위의 재구성이 필요하였다.

이에 본 연구에서는 연립·다세대 부동산 가격특성을 반영하여 공간적 이질성이 발생하지 않는 권역 분류를 하고자 하였다. 이에 본 연구에서는 부동산 가격특성이 고려되지 않은 기존의 연구에서 벗어나, <Figure 10>과 같이 16개 구역으로 부동산 가격특성과 공간적 실체를 반영하여 군집화하였다.



〈Figure 10〉 Clustering 16 areas in Seoul

4. 시사점

4.1 학문적 시사점

본 연구는 앞서 제시한 연구결과를 토대로 다음과 같은 학문적 시사점을 도출하였다.

첫째, 서울특별시, 한국감정원 및 기존의 부동산연구에서 사용되었던 권역의 문제점을 개선하고자 연립·다세대 부동산 가격특성을 반영하여 군집화를 했던 점이다. 서울특별시와 한국감정원에서 사용한 5개 또는 25개의 권역은 도시기본계획에 따른 분류로 부동산가격을 예측하기 위한 최적의 군집이 아니다. 상이한 공간 범위에 따라 주택 가격이 달리 산출되며, 공시가격과 실거래 가격의 불일치가 심화 될 수 있는 점에서 탈피하고자 본 연구에서는 K-Means Clustering 알고리즘과 헤도닉모형을 활용하여 주택가격모형 연구에 적합한 군집을 제시하였다. 이는 연립·다세대뿐만 아니라, 기존의 부동산 연구에서 부동산 가격을 예측하기 위한 최적의 군집을 제시하여 향후 주택가격 모형에 대해 기초자료로 의미가 있을 것으로 보인다.

둘째, 아파트에 관한 연구가 주로 이루어졌던 기존 연구에서 탈피하여, 본 연구에서는 연립·다세대 연구 영역으로 확장했다. 대다수 기존의 연구들은 연립·다세대 연구에 대한 연구가 미흡했다. 하지만, 본 연구에서는 연립·다세대 연구를 진행하였다는 점에 학술적 시사점이 있다.

마지막으로, 정부 3.0 공공정보 중 연립·다세대의 실거래가 데이터를 활용하고 서울특별시의 공간정보를 K-Means Clustering 알고리즘과 헤도닉 모형을 적용한 방식으로 군집 분류를 제시하였다. 이처럼 본 연구의 시도는 후속연구자들에게 방대한 양의 데이터를 통한 공간정보의 군집분류 방식을 제안하고 연립·다세대의 시세제공에 대한 정보를 구축하는 데에 있어 의미가 있을 것이다. 이는 연립다세대 실거래가 데이터를 포함한 공간정보를 통해 연립·다세대의 정보서비스 구축에 필요한 기초자료가 되는데 의미가 있을 것으로 보인다.

4.2 실무적 시사점

본 연구를 토대로 학술적 시사점 외에도 기존의 연립·다세대의 공공정보를 활용하는 실무자인 금융기관, 회계사, 부동산 컨설팅 혹은 개발사 및 개별사용자를 위한 실무적인 시사점을 도출할 수 있다.

첫째, 본 연구는 기존의 연립·다세대의 공공정보 데이터를 토대로 서울시 공간 범위 분류를 재구성하여 연립·다세대 부동산 관련 연구에 대한 기초자료로 활용할 수 있다는 점이다. 이로 인해 부동산 시세에서 사각지대였던 연립·다세대 실거래가 데이터와 공간 범위에 대한 이해를 가능하게 하였다. 이에 본 연구에서는 이를 통하여 정책 및 연구자료 활용이 가능할 것으로 보인다. 이는 연립·다세대와 같은 비정형 부동산에 대한 정책을 수립할 시 근거자료로 활용을 할 수 있으며, 연립·다세대 부동산 관련 연구를 위한 기초 자료로 활용될 것으로 기대된다.

둘째, 연립·다세대 연구의 활성화 및 실거래가 모형의 정확성 증대가 기대되는 점이다. 공간단위 창출을 기반으로 연립·다세대 실거래가 모형은 연립·다세대 거주자의 금융지원 범위 확장, 부동산허위거래 예방과 중개 소통기반 마련, 금융의 안정적인 금융조달 인프라 구축 등 금융서비스에 이바지 할 수 있을 것으로 기대된다. 즉, 금융기관에 있어 참고 및 활용을 할 수 있을 것이다. 연립·다세대 시세정보의 부재로 금융기관과 대출자간 상호비용에 대한 소통접점을 이루기 수월하지 않았었지만, 본 연구를 통한 연립·다세대 시세정보는 시세를 활용한 금융서비스가 확대 될 것으로 보여진다. 또한, 헤도닉모형을 활용한 시세 산정은 연립·다세대 매매 기준가를 활용할 수 있으며, 임대차 계약시 기준가로 활용이 가능할 것으로 기대된다. 즉 주택 구매자의 투자과정 전반에 효율성을 제공이 가능할 것으로 보여진다.

5. 결론 및 한계점

5.1 결론

최근 도심을 중심으로 연립·다세대의 거래가 활성화됨으로써, 비정형 부동산에 대한 시세 정보제공이 필요성이 대두되었다. 기존 연구에 따르면 연립·다세대는 거래량의 비중이 상대적으로 낮아 해석상의 어려움을 발생시켰으며, 또한, 데이터가 수집되는 공간 범위에 따라 주택가격

이 달리 산출되는 등 데이터의 오류의 문제점이 발생하였다고 한다(Kim, 2016; Lee and Kim, 2013) 이에 본 연구에서는 연립·다세대의 시세정보 구축이 어려웠던 문제를 헤도닉모형을 통해 산출하고 서울특별시의 연립·다세대 군집분류체계를 연구하고자 하였다. 이에 본 연구는 정부3.0에서 개방한 정보를 기반으로 수집, 정제하여 K-Means Clustering 알고리즘과 헤도닉 모형을 활용한 서울특별시의 연립·다세대 군집분류 방법에 대해 연구하였다. 따라서, 연립·다세대의 시세정보를 산출할 수 있었으며, 기존의 행정구역 25개 또는 5개의 권역으로 나누어 관련 통계지표를 작성하던 방식을 본 연구를 통하여 속성이 유사한 16개의 구역으로 군집 분류할 수 있었다. 즉, 정부3.0의 공공정보 중 연립·다세대 실거래가 데이터를 본 연구에서 K-Means Clustering 알고리즘과 헤도닉모형을 통해 실증분석을 하여 서울특별시 내의 연립·다세대를 16개의 군집으로 최종 분류하였다.

5.2 한계점 및 향후 연구 방향

본 연구의 한계점으로는 행정구역을 구분했을 때 생기는 경계점에서의 거래 사례 구분이 위·경도 기반의 클러스터링에서도 발생한다는 점으로 이에 대한 연구가 충분하지 않았던 점에 대해 한계점을 갖는다. 즉, 각 방법론마다 군집을 판정하는 기준이 다르기 때문에 하나의 방법론보다는 여러 가지 방법 결과를 비교 분석하여 최적의 군집을 판정해야 한다는 점에서 한계점을 갖고 있다.

또한, 유사도의 임계치에 만족하는 군집형성이 아닌 최종적으로 임의로 정한 N개의 클러스터를 목표로 분석을 시작한 점이다. 군집분석은 군집 개수에 대한 명확한 기준이 없어 연구자의 해석에 의해 차이를 갖게 되는 위험성이 있다. 이에 임의적인 N개의 클러스터를 목표로 진행된 본 연구는 연구자에 따른 해석의 차이점을 갖게 되어 한계점을 갖고 있다.

또 다른 한계점으로는 헤도닉 모형의 독립변수들의 개수 문제가 있다. 더 많은 변수을 고려할 수도 있지만, 변수의 수를 늘릴수록 유사성의 척도인 코사인 유사도의 평균이 점점 증가, 판별력이 떨어지는 것으로 볼 수 있기 때문에 본 연구에서는 4가지 변수를 채택하였다. 향후 연구에서는 더 많은 독립변수를 사용하여 제시하고자 한다.

이에 본 연구의 향후 연구 방향은 한계점을 극복하기 위해 다양한 분석을 진행하여 비교 분석을 해야 하며, 더욱 심층적인 연구의 필요성을 나타내고 있다. 차기 연구에서는 다양한 최적 군집분석방법을 활용하여 비교 분석하고자 한다.

또한, 향후 연구가 진행된다면, 서울특별시뿐만 아니라, 전국단위로 확장하여 군집분류를 하여 분석을 제시할 수 있고, 연립·다세대 부동산 관련 연구, 시세정보를 제공하지 않는 소형단지 아파트에 관한 연구가 추가로 연구할 수 있을 것으로 기대된다.

참고문헌(References)

Adriaans, P. and D. Zantinge, *Data Mining*, Addision-Wesley Harlow, 1996

Arthur, D. and S. Vassilvitskii. "How Slow is the K-means Method?." *Proceedings of the Twenty-Second Annual Symposium on Computational Geometry*. ACM(2006), 144-153.

Anderberg, M. R., *Cluster Analysis for Applications. Monographs and Textbooks on Probability and 15 Mathematical Statistics.*, in Academic Press, Inc., New York, 1973.

Berry, M. J. A and G. S. Linoff, *Data Mining Techniques for Marketing, Sales and Customer Relationship Management*, Third Edition, John Wiley & Sons Inc, 2011.

Benfratello, L., M. Piacenza, and S. Sacchetto. "Taste or Reputation: What Drives Market Prices in the Wine Industry? Estimation of a Hedonic Model for Italian Premium Wines." *Applied Economics*, Vol. 41, No. 17(2009), 2197-2209.

Brachman, R. J., and T. Anand., *The Process of Knowledge Discovery in Databases*., Advances in Knowledge Discovery and Data Mining, 1996

Chen, M. S., J. Han, and P. S. Yu, "Data Mining: an Overview from a Database Perspective." *IEEE Transactions on Knowledge and data Engineering*, Vo1. 8, No. 6(1996), 866-883.

Fayyad, U. M. "Data Mining and Knowledge Discovery: Making Sense out of Data." *IEEE Expert: Intelligent Systems and Their Applications*, Vol. 11, No. 5(1996), 20-25.

Hall, M., I. Witten, and E. Frank., *Data Mining: Practical Machine Learning Tools and Techniques*., Kaufmann, Burlington, 2011

Jain, A. K., "Data Clustering: 50 years beyond K-means." *Pattern Recognition Letters*, Vol. 31, No. 8 (2010), 651-666.

Jang, M., and C. Kang., "A Study on the Spatial Structure of Row-House and Multi-Family House and Its Policy Implications in Seoul," *Journal of the Korea Real Estate Analysts Association,* Vol. 24, No. 2 (2014), 87-96.

Jang, N. S., S. W. Hong and J. H. Jang, *Data mining*, Seoul: Daechung Media, 1999

Jung, U. B. and H. R. Lee, "Core Attributes Influencing the Room Rate of Deluxe Hotels in Seoul: Focused on a Hedonic Price Model", *Journal of Tourism Sciences*, Vol. 41, No. 3(2017), 131-149.

Kang, H. C., S. T. Han, J. H. Choi, S. G. Lee., E. S. Kim, I. H. Eom., and M. G. Kim*., Data Mining Methodology.*, Seoul: Free Academy, 2006.

Kim, B. R., Y. I. Yoon, and M. S. Chung., "A Hedonic Model Effects for Consumer-oriented Retargeting Advertising Based on Internet of Things." *Journal of the Korea Society of Computer and Information,* Vol. 22, No. 2(2017), 75-80.

Kim, H. H., T. S. Lee., J. M. Kim., and T. H. Ahn., “Small Area Categorization by Socioeconomic Characteristics for Local Government Policy Development.”, *The Geographical Journal of Korea,* Vol. 49, No. 2(2015), 229-240.

Kim, S. W. and K. S. Chung, "Comparative Study of the Fitness between Traditional OLS Models and Spatial Econometrics Models Using the Real Transaction Housing Price in the Busan.", *Journal of the Korea Real Estate Analysts Association,* Vol. 16, No. 3(2010), 41-55.

Kim, J. H., "An Analysis on the Spatio-temporal Heterogeneity of Real Transaction Price of Apartment in Seoul Using the Geostatistical Methods", *Journal of the Korean Society for Geospatial Information Science*, Vol. 24, No. 4(2016), 75-81.

Kim, J. I., "The Comparison of Local Housing

Price Determinants by Housing Type", *Housing Studies Review*, Vol. 25, No. 2(2017), 175-195.

Kim, J. M., "New Optimization Algorithm for Data Clustering", *Journal of Intelligence and Information Systems*, Vol. 13, No. 3 (2007), 31-45.

Koo, W. Y. “Understanding Data Mining and Utilizing the Mechanical Field ” *Magazine of the SAREK,* Vol. 45, No. 1 (2016), 38-43.

Kwon, J. W., and H.C. Kim. "Estimation of Housing Price Index using a Varying Parameter Model." *Journal of the Korean Urban Management Association,* Vol. 19, No. 1(2006), 175-200.

Lee, C., J. Lee, and S. Lim, "The Non-Apartment Rental Housing Market Analysis," *Journal of the Korea Real Estate Analysts Association,* Vol. 13, No. 1(2007), 25-47.

Lee, S. W., and J. Y. Kim, "Transactions Clustering based on Item Similarity", *Journal of Intelligence and Information Systems*, Vol. 9, No. 1 (2003), 179-193.

Lee, S. W. and W. H. Lee, "Refining Initial Seeds using Max Average Distance for K-Means Clustering." *Journal of Korean Society for Internet Information,* Vo.12, No. 2(2011), 103-111.

Lee, S. W., "Comparison of Initial Seeds Methods for K-Means Clustering." *Journal of Korean Society for Internet Information*, Vol. 13, No. 6(2012), 1-8.

Lee, G., and K. Kim, "A Study on the Spatial Mismatch between the Assessed Land Value and Housing Market Price: Exploring the Scale Effect of the MAUP." *Journal of the Korean Geographical Society,* Vol. 48, No. 6(2013), 879-896.

Lee, Y. M, "A Review of the Hedonic Price Model." *Journal of the Korea Real Estate Analysis Association,* Vol. 14, No. 1(2008), 81-87.

Leonard, T., T. M. Powell-Wiley., C. Ayers., J. C. Murdoch, W. Yin, and S. L. Pruitt, "Property Values as a Measure of Neighborhoods: An Application of Hedonic Price Theory." *Epidemiology,* Vol. 27, No. 4(2016), 518-524.

Lloyd, S., "Least Squares Quantization in PCM," *IEEE Transactions on Information Theory,* Vol. 28, No. 2 (1982), 129-137.

MacQueen, J., "Some Methods for Classification and Analysis of Multivariate Observations," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability,* Vol. 1. No. 14(1967), 281-297.

Malpezzi, S., "Hedonic Pricing Models: a Selective and Applied Review." In: O’Sullivan, T., Gibb, K. (Eds.), *Housing Economics and Public Policy.,* Blackwell, Oxford, UK, 2002, 67-89.

Meghani, S. H., and G. J. Knafl., "Salient Concerns in Using Analgesia for Cancer Pain among Outpatients: A Cluster Analysis Study." *World Journal of Clinical Oncology*, Vol. 8, No. 1(2017), 75.

Na, M. Y., *A Technique to Extract Useful Knowledge from a Large Knowledge Database.*, Data Base World, 1997.

Nam, J. and J. H.Kim. "An Analysis of Factor Influencing on the Choice of Housing Types and Tenure by Income Bracket in Seoul" *Journal of the Korean Urban Management Association*, Vol. 28, No. 2(2015), 199-222.

Pejman, A., G. N. Bidhendi, M. Ardestani, M. Saeedi, and A. Baghvand, “ Fractionation of Heavy Metals in Sediments and Assessment of their Availability Risk: A Case Study in the Northwestern of Persian Gulf.” *Marine Pollution Bulletin*, Vol. 114 No. 2(2017), 881-887.

Park, D. H., H. K. Kim, I. Y. Choi, and J. K. Kim, "A Literature Review and Classification of Recommender Systems on Academic Journals", *Journal of Intelligence and Information Systems*, Vol. 17, No. 1 (2011), 139-152.

Park, W. S. and B. J. Rhlm. “A Study on the Factors Affection Apartment Price by Using Hedonic Price Model”. *Korea Real Estate Society,* Vol. 28, No. 2 (2010). 245-271.

Romesburg, C., *Cluster Analysis for Researchers*., North Carolina: Lulu Press. 2004.

Redmond, S. J., and. H. Conor, "A Method for Initialising the K-means Clustering Algorithm Using Kd-trees,", *Pattern recognition letters* , Vol. 28, No. 8 (2007), 965-973.

Rosen, S., "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition.", *Journal of Political Economy,* Vol. 82, No. 1(1974), 34-55.

Ricardo, B. Y., and R. N. Berthier., *Modern Information Retrieval*., New York: ACM press, 1999.

Ryu, K., S. Choi, and S. Lee, "Median Price Index for Single-family housing and Multi-family housing in Seoul," *Journal of the Korea Real Estate Analysts Association*, Vol. 18, No. 2(2012), 57-72.

Seo, S. B., and S. N. Kwak., "A Study on the Adequacy of Standard Comparison Table of Land Price by Hedonic Price Model.", *Journal of Korea Planning Association,* Vol. 49, No. 5(2014), 187-204

Yang, M., Y. Lee., and J. S. Song., "Application of Hedonic Price Model to Korean Antique Art Data.", *Journal of Information Technology Applications & Management,* Vol. 23, No. 4 (2016), 41-53.

Yeom. M. B., and K. M. Kim., " Deriving the Causes of Low Fertility and Policy Demand through Cluster Analysis."., *Journal of Economy*, Vol. 29, No. 1(2011), 163-190.

Yong, H. S., Y. M. Na., J. S. Park., H. W. Seung., M. S. Lee., and R. Choi., *Data Mining*., Seoul: Infiniti Books, 2007.

Yun, H. Y., Y. S. Koo and D. R. Choi. "A Development of Ensemble Model Based on Cluster Analysis to improve PM10 Forecasting Accuracy : Focus on the Weighted Average Ensemble by Weather Cluster." *Journal of Korean Society of Urban Environment*, Vol. 17, No. 1(2017), 33-42.

Abstract

A Study on the Clustering Method of Row and Multiplex Housing in Seoul Using K-Means Clustering Algorithm and Hedonic Model

Soonjae Kwon* · Seonghyeon Kim** · Onsik Tak*** · Hyeonhee Jeong****

Recent centrally the downtown area, the transaction between the row housing and multiplex housing is activated and platform services such as Zigbang and Dabang are growing. The row housing and multiplex housing is a blind spot for real estate information. Because there is a social problem, due to the change in market size and information asymmetry due to changes in demand. Also, the 5 or 25 districts used by the Seoul Metropolitan Government or the Korean Appraisal Board(hereafter, KAB) were established within the administrative boundaries and used in existing real estate studies. This is not a district classification for real estate researches because it is zoned urban planning. Based on the existing study, this study found that the city needs to reset the Seoul Metropolitan Government's spatial structure in estimating future housing prices. So, This study attempted to classify the area without spatial heterogeneity by the reflected the property price characteristics of row housing and Multiplex housing. In other words, There has been a problem that an inefficient side has arisen due to the simple division by the existing administrative district. Therefore, this study aims to cluster Seoul as a new area for more efficient real estate analysis. This study was applied to the hedonic model based on the real transactions price data of row housing and multiplex housing. And the K-Means Clustering algorithm was used to cluster the spatial structure of Seoul. In this study, data onto real transactions price of the Seoul Row housing and Multiplex Housing from January 2014 to December 2016, and the official land value of 2016 was used and it provided by Ministry of Land, Infrastructure and Transport(hereafter, MOLIT). Data preprocessing was followed by the following processing procedures: Removal of underground transaction, Price standardization per area, Removal of

* Business Administration, Daegu University
** National Information Society Agency(NIA)
*** KN Company
**** Corresponding Author: Hyeonhee Jeong
Business Administration, Daegu University
201, Daegudae-ro, Jillyang-eup, Gyeongsan-si, Gyeongsangbuk-do, 38453, Korea
Tel: +82-10-3790-9091, Fax: +82-53-850-6239, E-mail: hu912@naver.com

Real transaction case(above 5 and below -5). In this study, we analyzed data from 132,707 cases to 126,759 data through data preprocessing. The data analysis tool used the R program. After data preprocessing, data model was constructed. Priority, the K-means Clustering was performed. In addition, a regression analysis was conducted using Hedonic model and it was conducted a cosine similarity analysis. Based on the constructed data model, we clustered on the basis of the longitude and latitude of Seoul and conducted comparative analysis of existing area. The results of this study indicated that the goodness of fit of the model was above 75 % and the variables used for the Hedonic model were significant. In other words, 5 or 25 districts that is the area of the existing administrative area are divided into 16 districts. So, this study derived a clustering method of row housing and multiplex housing in Seoul using K-Means Clustering algorithm and hedonic model by the reflected the property price characteristics. Moreover, they presented academic and practical implications and presented the limitations of this study and the direction of future research. Academic implication has clustered by reflecting the property price characteristics in order to improve the problems of the areas used in the Seoul Metropolitan Government, KAB, and Existing Real Estate Research. Another academic implications are that apartments were the main study of existing real estate research, and has proposed a method of classifying area in Seoul using public information(i.e., real-data of MOLIT) of government 3.0. Practical implication is that it can be used as a basic data for real estate related research on row housing and multiplex housing. Another practical implications are that is expected the activation of row housing and multiplex housing research and, that is expected to increase the accuracy of the model of the actual transaction. The future research direction of this study involves conducting various analyses to overcome the limitations of the threshold and indicates the need for deeper research.

Key Words : Classification of Clusters, Row Housing, Multiplex Housing, Hedonic Model, K-Means Clustering Algorithm

Received : July 31, 2017 Revised : September 17, 2017 Accepted : September 20, 2017
Publication Type : Regular Paper Corresponding Author : Hyeonhee Jeong

저 자 소 개

권 순 재

현재 대구대학교 경영학과 교수로 재직 중이다. 성균관대학교 경영학부를 졸업하고 성균관대학교에서 경영정보시스템 전공으로 석사 및 박사를 취득하였다. Journal of MIS, Information and Management, Decision Support Systems, Journal of Computer Information Systems, Behavior and Information Technology, Cyber Psychology and Behavior, Electronic Commerce and Research Application, Expert Systems with Applications 등에 논문을 게재하였으며, 국내에도 50여편의 연구가 있다. 주요 관심분야는 SNS에서 재미, 인터넷 및 모바일에서 소비자행동, 온라인 커뮤니티에서의 집단지성 등이다.

김 성 현

고려대학교에서 경영학석사, 성균관대학교에서 MIS 전공으로 경영학 박사학위를 취득하였다. 삼성SDS IT컨설팅본부에서 CRM 및 IT컨설팅 프로젝트를 수행하였으며, 현재 한국정보화진흥원 빅데이터센터에서 빅데이터 정책 기획 및 플래그십 시범사업을 담당하고 있다. Journal of Small Business Management 등 국내외 학술지에 14편의 논문을 게재하였다. 관심분야는 빅데이터, 클라우드컴퓨팅, 중소기업, IT 성과관리 연구이다.

탁 온 식

중앙대학교에서 통계학으로 석사학위를 취득하고, 현재 케이앤컴퍼니 데이터연구 부서에서 연구원으로 재직 중이다. 주요 관심분야는 구조방정식모형분석에 관한 연구와 빅데이터를 활용한 부동산 시세 예측 연구이다.

정 현 희

대구대학교에서 이학사와 회계학사를 취득하였으며, 대구대학교 일반대학원에서 MIS를 전공하여 경영학 석사학위를 취득하였다. 현재 케이앤컴퍼니 기업부설연구소에서 연구원으로 재직 중이며 대구대학교 일반대학원에서 MIS 전공으로 박사과정 재학 중이다. 주요 관심분야는 빅데이터, 데이터마이닝기법 등이다.