지리학 박사 학위논문

비모수 공간모형과 앙상블 학습에 기초한 단독주택가격 추정

2015년 8월

서울대학교 대학원

지 리 학 과

이 창 로

비모수 공간모형과 앙상블 학습에 기초한 단독주택가격 추정

지도교수 박 기 호

이 논문을 지리학 박사학위논문으로 제출함
2015년 4월

서울대학교 대학원
지리학과
이 창 로

이창로의 박사학위논문을 인준함
2015년 7월

위 원 장 _____ (인)

부위원장 _____ (인)

위 원 _____ (인)

위 원 _____ (인)

위 원 _____ (인)

국문초록

부동산 가격추정 모형은 최근 부동산 자료의 공개 및 구득 가능성 증가로 과거 그 어느 때보다 다양한 분야에서 활용되고 있다. 즉 자산 포트폴리오의 구성, 금융기관의 담보물 가치 추정, 부동산 개발의 타당성 판단 등 여러 업무에서 활용되고 있으며, 특히 과세평가는 파급효과가 전 국민에게 미치는 등 중요성이 매우 큰 분야의 하나이다.

그러나 우리나라의 대표적 과세가치에 해당되는 공시가격의 경우 현실화율이 낮고 가격 균형성이 미흡하다는 지적은 과거부터 꾸준히 제기된 문제점이다. 이는 조세저항 등 정치적 요인에서도 그 원인을 찾을 수 있지만 세금 부과의 기본이 되는 과세평가 과정, 보다 구체적으로 가격추정 모형이 잘못된 것에 기인한다. 본 연구는 보다 정확한 부동산 가격추정 방법론의 탐색으로부터 시작되었다.

또한 수리 또는 계량적 모형을 사용하는 사회과학연구에서 지금까지 설명 중심의 모형(Explanatory Modeling)이 주류를 이루었으며, 설명력이 좋은 모형은 예측력 또한 좋을 것으로 암묵적 가정을 하여 왔다. 그러나 이러한 두 가지 성능이 항상 일치하는 것은 아니며, 본 연구에서는 모형의 해석 가능성 등을 희생하더라도 신균 관찰치의 예측력 향상을 강조하는, 예측 중심의 모형(Predictive Modeling)을 구축하였다.

부동산 가격을 추정하기 위해 전통적으로 사용된 모형은 대부분 모수 모형(Parametric Model)으로서 설명변수의 독립성, 자료의 정규성, 모형 설계(Model Specification) 오류의 부재 등 엄격한 가정이 많았다. 뿐만 아니라 가격함수를 모수 및 설명변수와의 선형결합 형태로 전제하는 등 자료 특성을 지나치게 단순화하는 단점이

있었다. 이러한 비현실적 통계적 가정과 사전에 설정된 가격함수 형태를 강제하지 않는 예측 중심의 모형들이 기계학습(Machine Learning) 분야에서 다양하게 제시되었으며, 이러한 모형들은 그 특징상 대부분 비모수 모형(Non-parametric Model)에 해당된다.

본 연구에서는 그간 모수 모형에 집중되었던 부동산 가격추정 방법론을 예측 중심의 비모수 모형으로 확대하고자 한다. 아울러 다양한 비모수 모형 중 가장 우수한 것으로 판명된 모형 하나를 선택하는 것이 아니라, 개별 모형들의 추정값을 적정하게 결합하는 앙상블 학습(Ensemble Learning) 개념을 가격결정 과정에 도입하고자 한다. 마지막으로 이와 같은 모형의 정교화 외에 사례지역에 대해 주택가격을 직접 추정함으로써 모형을 통해 산정된 가격과 실제 거래가격 및 현행 주택공시가격과의 차이점을 파악하고자 한다.

2011년부터 2014년 사이에 신고된 실거래가 자료를 투입자료로 사용하였으며 사례지역은 대도시, 중소도시 및 군 지역을 대표할 수 있도록 서울시 강남구, 전주시 덕진구, 전라남도 해남군을 선정하였으며 주요 결과는 다음과 같다.

기계학습 분야에서 제시된 여러 비모수 모형 중 SVM(Support Vector Machine)이나 MARS(Multivariate Adaptive Regression Splines) 등 최근에 개발된 모형들의 성능이 비교적 우수한 것으로 나타나 이러한 모형들의 확대 적용이 필요한 것으로 보인다. 또한 지역 측면에서 강남구보다는 덕진구가, 덕진구보다는 해남군이 가격추정의 정확성이 떨어졌는데, 이는 농촌지역으로 갈수록 주택집단의 이질성이 높아지기 때문인 것으로 풀이된다.

가격추정 모형이 특히 어떠한 부분에서 취약한지 효율적으로 파악하기 위해 회귀트리 알고리즘(Regression Tree Algorithm)에 기반한 국지적 모형성능 진단을 수행한 결과, 토지 면적(또는 주택규모)에 따른 자료 층화가 선행된 후 본격적인 모형 구축이 이루어질 경우 가격추정의 정확성이 높아질 것으로 파악되었다.

한편 기계학습 분야에서 제시된 이러한 비모수 모형들은 기본적으로 속성정보만 고려할 뿐, 공간사상의 특징인 공간적 종속성(Spatial Dependence)을 반영하는데 관심이 적다. 본 연구에서는 비모수 모형에 공간적 종속성을 추가로 반영하기 위해 SVM의 scale parameter를 공간적 종속성이 미치는 지리적 범위로 해석하여 모형을 정교화하였다. 또한 여러 비모수 모형에 모두 적용할 수 있도록 주변 주택가격의 평균적인 가격수준을 나타내는 공간차변수(Spatially Lagged Variable) **WY**를 구성하여 공간적 종속성을 모형의 한 요소로 반영하였다.

주택에 대한 최종 예측가격은 개별 모형들 중 가장 성능이 우수하게 나타난 모형의 예측치로 결정하는 대신, 개별 모형들에서 산출된 예측치를 가중평균하는 앙상블 평균(Ensemble Averaging)을 적용하여 결정하였다. 앙상블 평균은 해남군과 같이 개별 모형들에서 산출된 예측치 간의 상관성이 낮은 경우 탁월한 성과를 보였다.

마지막으로 본 연구에서 제시한 앙상블 예측가격과 실제 거래가격, 그리고 현행 공시가격을 비교하였으며 여러 측면에서 공시가격보다는 앙상블 예측가격이 실제 거래가격을 보다 가깝게 반영하였다. 그러나 공시가격의 특징 내지 품질은 표준주택의 선정 등 자료수집 단계, 이해관계자 의견청취 단계 등 여러 절차에서 발생한 오류가 집적된 것임을 감안하여 해석할 필요가 있다.

주요어: 기계학습, 비모수 모형, 공간적 종속성, 앙상블 학습, 공시가격
학 번: 2012-30841

목 차

표 목 차

그 림 목 차

제 1 장 서 론

제 1 절 연구 배경과 목적

1. 부동산 가격의 추정

자산(Asset)의 경제적 가치를 판정하고 이를 화폐액으로 표시하는 것을 가치평가(Valuation)라 하며(장희순・방경식, 2014), 자산 중에서도 금융자산을 제외한 부동산의 가치를 추정하는 경우 감정평가(鑑定評價, Appraisal)라는 용어를 일반적으로 사용한다. 부동산에 대한 감정평가는 다시 대량평가(Mass Appraisal)와 정밀평가(Single-Property Appraisal)로 구분할 수 있다.

부동산의 대량평가란 주어진 날짜에 공통된 자료를 이용하여 통계적 검증을 거친 표준화된 모형으로 모집단 전체의 부동산 가격을 일시에 산정하는 것을 말한다(IAAO, 2008). 반면, 정밀평가는 대상 부동산에 대해 개별적인 자료 수집과 현장조사를 거쳐 평가하는 것을 의미한다.

정밀평가에 대한 수요는 지속적으로 감소하고 있는 반면, 저렴한 평가비용, 신속한 처리, 가격 결정에 대한 자의성 개입 최소화 등을 비교 우위로 하여 대량평가는 여러 분야에서 중요한 도구로 부상하고 있다. 특히 부동산 가격 및 특성 자료의 공개 및 구득 가능성 증가로 대량평가는 과거 어느 때보다 다양한 분야에서 활용되고 있다.

먼저 경제학 측면에서 토지나 주택 등을 하나의 경제재로 보아 이들의 가격 변화에 따른 수요와 공급의 변화 양상 파악 및 예측, 토지시장이나 주택시장의 효율성 검증, 신축 주택의 시장에서의 배분과정 등을 살필 때 부동산의 가격 추정은 분석의 핵심 역할을 하게 된다.

재무관리 측면에서도 부동산을 포함한 여러 자산의 포트폴리오 구성, 부동산을 기초자산으로 한 파생상품의 설계업무 등에 있어 정확한 부동산 가격정보는 필수적이다.

뿐만 아니라 금융기관의 부동산 담보 대출시 담보물의 적정가치 추정, 개발예정 부동산의 투자 타당성 판단, 소음·일조 등으로 인한 자산가치 하락분의 산정 등 민간부문에서의 활용 용도는 계속하여 확대되고 있다.

또한, 공익사업 시행으로 인한 수용토지에 대한 적정 보상금액 산정, 재산세 등 각종 조세부과를 위한 과세가치(Assessed Value)의 산정 등 공공부문에서도 그 중요성이 날로 커지고 있다.

이러한 부동산 대량평가의 다양한 활용영역 중 가장 역사가 오래되고 그 파급효과가 전 국민에게 미치는 등 중요성이 큰 분야는 바로 과세평가(Property Assessment) 분야라 할 수 있다. 캐나다 등 북미의 경우 대량평가 기법을 과세평가에 활용한 지 80년이 넘었고(IAAO, 2008), 덴마크 등 유럽의 경우에도 100년이 넘었다(Gloudemans & Almy, 2011).

재산세, 종합부동산세, 상속세, 증여세 등 각종 조세는 원칙상 종가세(從價稅)이므로 과세대상인 부동산의 정확한 가치 추정이 전제되어야 과세관청은 최초의 의도한 효과를 기대할 수 있다. 즉 세수의 충분한 확보, 조세 부담의 형평성 등을 실현할 수 있게 된다. 그러나 우리나라의 대표적 과세가치에 해당되는 공시지가나 주택공시가격의 경우 현실화율이 낮고 가격 균형성이 미흡하다는 지적은 과거부터 꾸준히 제기된 문제점이다. 이는 세금 부과의 기본이 되는 과세평가 과정, 보다 구체적으로는 가격추정 모형이 잘못된 것에 기인한다(안정근, 2004).

일례로 [표 1-1]은 우리나라의 과세평가 체계를 보여 주는데, 시장에서의 실거래가를 기준으로 과세가치를 산정하는 것이 아니라, 간단한 몇 가지 수식을 적용하여 과세가치를 산출하고 있다. 즉 과세평가 체계가 다른 선진국에 비해 초보적 단계에 머무르고 있다.

본 연구에서는 전통적 가격모형의 문제점을 인식하고, 새로운 가격모형의 적용을 통해 현행 가격추정 모형의 기술적인 개선 가능성을 살펴보고자 한다. 그러나 부동산 세법, 세율, 조세체계 등 과세정책과 관련된 보다 상위의 영역은 다루지 않는다.

[표 1-1] 우리나라의 부동산 과세평가 체계

유 형		과세평가 산식	비 고
토지		표준지 공시지가 x 비준율	비준율: 선형회귀모형으로 산출
건물		건물신축가격기준액 x 구조지수 x 용도지수 x 위치지수 x 잔가율	건물시가표준액 건물기준시가
주택 (토지+건물)	단독주택	표준주택 공시가격 x 비준율	비준율: 선형회귀모형으로 산출
	공동주택	거래가격 x 층·향 등의 보정	전수 조사

2. 예측 중심의 모형(Predictive Modeling)[1)]

수리 또는 계량적 모형을 사용하는 사회과학연구에서 분석의 초점은 종속변수와 설명변수의 인과관계(Causal Relationship)를 추론하는 것이 대부분이었다. 즉 설명 중심의 모형(Explanatory Modeling)이 주류를 이루었다. 그러나 예측(Prediction) 능력이 없거나 미흡한 모형은 현실세계에서 그 유용성이 급격히 떨어질 수밖에 없다.

통상 설명력이 좋은 모형은 예측력 또한 좋을 것으로 암묵적 가정을 하지만 이러한 두 가지 성능이 항상 일치하는 것은 아니다(Shmueli, 2010). 본 연구에서는 이론 및 가설을 검증하거나 인과관계를 밝히는 설명 중심의 모형이 아니라 신규 관찰치의 예측력 향상을 강조하는, 예측 중심의 모형(Predictive Modeling)을 구축하고자 한다.

예측 중심의 모형은 설계에서부터 최종 진단에 이르는 모든 과정이 설명 중심의 모형과 차이점을 보이며, 특히 설명 중심의 모형은 말 그대로 설명, 즉 해석이 가능해야 하므로 단순한 형태의 함수를 선호한다. 기존의 많은 연구에서 선형 함수를 빈번하게 활용한 이유가 여기에 있다. 반면 예측 중심의 모형은 목적이 정확한 예측에 있으므로 모수 등의 '해석 불가능'이 분석의 걸림돌이 되지 않는다.

1) 경제학 등 사회과학에서 '예측'은 통상 미래에 발생할 상황을 추측할 때 사용하는 용어이지만, 본 논문에서는 미래 시점, 관측되지 않은 지점 등 신규 관찰치가 발생할 때 해당 관찰치의 값을 추측하는 의미로 사용하기로 한다.

[표 1-2]는 설명 중심의 모형과 예측 중심의 모형을 비교한 것인데, 사회과학 분야의 경우 설명 중심의 모형에 대한 연구 실적은 풍부한 반면, 예측 중심의 모형은 그렇지 않은 편이다. 그러나 이러한 공백을 현재 기계학습(Machine Learning) 분야에서 활발하게 메워가고 있다(Shmueli, 2010).

[표 1-2] 설명 중심의 모형과 예측 중심의 모형 비교

구 분	설명 중심의 모형	예측 중심의 모형
분석 목적	이론이나 가설의 검증	신규 관찰치 값의 예측
주된 변수	개념 수준의 변수 (Conceptual Level)	측정 수준의 변수 (Measurable Level)
최적 모형 결정 기준	편의의 최소화	편의 및 분산의 최소화
주된 위험	Type I, II 오류	과다적합(Over-Fitting)
모형에 대한 제약	해석 가능하고 모형의 형태나 변수의 선정 등이 이론과 부합해야 함	분석 당시 활용 가능한 변수만 사용해야 하며, 사후적으로 확보된 변수 사용 불가
모형 진단기준	결정계수 R^2, 계수의 통계적 유의성(p-value), 잔차, 다중공선성	검증 데이터(Test Data)를 기준한 예측치의 오차 정도(MSE 등)

* 출처: Shmueli & Koppius (2011)에서 인용 및 재정리

[표 1-2]에서 예측 중심의 모형은 과다적합(Over-Fitting)이 가장 신경을 써야 하는 위험이고, 따라서 이러한 과다적합 위험을 피하기 위해 검증 데이터(Test Data)를 기준으로 오차를 판단하고 있음을 알 수 있다. 또한 설명 중심의 모형에서는 주로 선형의 함수를 가정함으로써 모형에 대한 해석 가능성을 높이려는 반면, 예측 중심의 모형에서는 이러한 노력을 찾아볼 수 없다. 기술적인 측면에서 가장 뚜렷한 차이점은 최적 모형의 결정 기준이라 할 수 있다. 설명 중심의 모형에서는 편의를 최소화하는 모형이 가장 바람직한 모형이지만, 예측 중심의 모형은 편의와 함께 분산을 최소화하는 것이 목적이므로 경우에 따라 편의 추정량이라 하더라도 분산을 현격하게 줄일 수 있다면 그러한 추정량을 사용하기도 한다.

이와 같이 모형이 가지는 주된 위험이나 진단기준, 그리고 모형에 대한 제약 등을 고려할 때 예측 중심의 모형은 대부분 비모수 모형

(Non-parametric Model)에 해당되는 경우가 많다(Abbott, 2014, p.213). 비모수 모형은 설명변수와 모수의 결합형태가 사전에 정해진 형태(선형함수 등)를 취하지 않고, 데이터가 가진 정보로부터 함수 형태를 추출하게 된다. 부동산 가격함수는 경제학적 측면에서 보았을 때 다양한 소비자 기호와 생산자 기술수준을 나타내기 때문에 그 정확한 형태를 가늠하기 어려우며, 선형으로 근사화할 수 있다는 가정도 받아들이기 어렵다.

따라서 비모수 모형은 모수 모형(Parametric Model)보다 합리적인 가격추정 모형이 될 수 있다. 다만 비모수 모형은 모수 모형보다 더 많은 수의 샘플을 필요로 하는 단점이 있다. 왜냐하면 모수 추정뿐만 아니라 모형의 구조 자체도 데이터가 가진 정보로부터 유도하여야 하기 때문이다. 하지만 이러한 제약은 '빅 데이터의 시대'라 불릴 만큼 자료의 공개 및 구득 가능성이 높아진 상황에서 더 이상 걸림돌로 작용하기 어렵다.

부동산 가치를 추정하기 위하여 전통적으로 사용되었던 모형은 회귀모형으로서 초기의 OLS(Ordinary Least Squares) 모형에서부터 시작하여 이후 공간적 종속성(Spatial Dependence)을 계량화하여 모형의 구성요소로 반영한 공간회귀모형, 공간보간법의 일종인 크리깅(Kriging) 기법 등이 제시되었다. 더불어 공간적 이질성(Spatial Heterogeneity)을 반영하기 위한 시도로 지리적 가중회귀모형(Geographically Weighted Regression Model) 및 다수준 모형(Multi-Level Model) 등이 제안되기도 하였다.

그러나 이러한 모형들은 대부분 모수 모형으로서 설명변수의 독립성, 자료의 정규성, 종속변수와 설명변수 간의 선형성(Linearity) 등 엄격한 가정이 많고 자료 특성을 지나치게 단순화하는 등 추정가격의 신뢰성에 한계가 있다(Ekeland, 1988; Kummerow & Galfalvy, 2002; Gloudemans & Almy, 2011). 특히 종속변수인 부동산 가격과 설명변수인 부동산 특성(면적, 건물구조 등) 사이에 선형의 함수 관계가 성립되는 것은 매우 예외적인 경우에 해당되어 많은 경우 추정가격의 정확성을 떨어뜨린다(Weirick & Ingram, 1990; Hastie et al., 2009, p.139).

이러한 비현실적인 통계적 가정을 부과하지 않는 보다 유연한 모형들이 최근 데이터 마이닝(Data Mining) 또는 기계학습 분야에서 다양하게

개발되었으며, 트리기반 모형(Tree-Based Model), MARS(Multivariate Adaptive Regression Spline), SVM(Support Vector Machine) 등이 그 예라고 할 수 있다. 기계학습 분야에서 제시된 이러한 모형들은 대부분 비모수적 방법에 해당되어 종속변수와 설명변수 간 선형의 함수 형태를 고집하지 않는다. 따라서 부동산 가격 추정에 있어 보다 현실적인 대안이 될 수 있을 것으로 판단된다.

전통적인 부동산 가격추정 모형은 설명 중심의 모형 및 모수 모형이었다. 본 연구에서는 이러한 가격추정 모형의 문제점을 인식하고, 예측 중심의 비모수 모형(Predictive Non-parametric Model)으로 방법론을 확대하고자 한다. 아울러 다양한 비모수 모형 중 가장 우수한 모형 하나를 선택하는 것이 아니라, 개별 모형들의 추정값을 적정하게 결합하는 앙상블 학습(Ensemble Learning) 개념을 가격결정 과정에 도입하고자 한다. 개인보다는 다수의 지성이 더 큰 힘을 발휘하듯, 개별 모형이 아닌 여러 모형을 결합하여 최종값을 예측하는 앙상블 학습은 발표된 즉시 많은 관심을 받으며 다양한 분야에서 분석의 대상이 되어왔다(Schapire, 1999; Banfield, 2007; Wang, 2008). 여기에서는 이러한 앙상블 학습의 이점을 실증적으로 밝히고자 한다.

마지막으로 이와 같은 모형의 정교화 외에 사례지역에 대해 주택가격을 직접 추정함으로써 모형을 통해 산정된 가격과 실제 거래가격 및 현행 주택공시가격과의 차이점을 파악하고자 한다. 파악된 차이점은 주택공시가격의 특징과 품질을 이해하는데 도움이 될 것으로 보인다.

제 2 절 연구의 범위와 방법

본 연구의 분석 대상은 주택으로 한정하며, 분석에 투입할 기초 자료는 실거래가 신고 자료를 사용한다. 다양한 부동산 유형 중 분석 대상을 주택으로 한정한 이유는 자료의 신뢰성 문제 때문이다. 실거래가 신고가격은 거래 당사자가 자발적으로 신고한 가격으로 입력 오류, 계산 착오, 의도적 허위 신고, 급매 등 이상치(Outliers)와 잡음(Noise)이 매우 많다.

주택 실거래가 신고 자료라 하여 이러한 이상치와 잡음이 없는 것은 아니지만, 공장이나 상가 등 비주거용 부동산은 주거용 부동산과 비교할 때 이상치와 잡음의 종류가 더욱 다양하여 신고가격을 신뢰하기가 매우 어렵다. 예를 들어 기계나 설비의 거래가격 포함 여부, 영업권이나 권리금 등의 포함 여부, 소유주의 경영능력 등 부동산 가치를 구성하는 것으로 볼 수 없는 항목들이 비주거용 부동산 신고가격에는 상당 부분 존재한다. 주택의 실거래가 신고자료는 이러한 문제가 상대적으로 덜하기 때문이다.

이러한 주택은 다시 공동주택(아파트, 연립주택, 다세대주택)과 단독주택으로 나뉘며, 표준화가 잘 되어 있어 대량평가 적용 가능성이 이미 입증된 공동주택은 분석 대상에서 제외한다. 따라서 본 연구에서는 단독주택가격 추정에 주안점을 둔다.

본 연구에서 예측의 대상이 되는 가격은 시장가치(Market Value)[2]를 의미하며, 이러한 시장가치 추정을 위한 데이터로 실거래가 신고자료를 사용하였다. 시장가치는 현행 법령상 '통상적인 시장에서 충분한 기간 거래를 위하여 공개된 후 그 대상 물건의 내용에 정통한 당사자 사이에 신중하고 자발적인 거래가 있을 경우 성립될 가능성이 가장 높다고 인정되는 대상물건의 가액'[3]으로 정의되어 매우 이상적인 가상거래 조건하에서의 체결가격으로 볼 수 있다. 따라서 어떠한 가격자료를 사용하든 그 가격이 시장가치와 일치한다고 보기 어렵다.

이와 같이 시장가치는 실제로 관찰할 수 없는 현상이므로 적절한 대리변수를 확보하여 사용하여야 한다. 2006년 실거래가 신고제도 도입 이전에는 주로 전문가에 의한 감정평가 가격, 부동산 중개업체의 호가, 공익사업으로 인해 지출된 보상금액, 법원 경매에서의 낙찰가액 등을 사용하였다. 그러나 실거래가 신고제도 도입 이후 시장에서의 실제 거래가격에 기초한 연구가 늘고 있으며, 본 연구도 이러한 흐름을 따라 실거래가 신고자료를 시장가치의 대리변수로 활용한다.

---

2) 상반되는 가치기준이 비시장가치(Non-market Value)이며, 사용가치, 공익가치, 청산가치, 투자가치 등을 포함한다.

3) 감정평가에 관한 규칙 제2조 제1호. 미국의 Appraisal Institute(1996)에서도 이와 유사한 정의를 내리고 있다.

실거래가 신고자료는 2011년부터 2014년까지 최근 4년 동안의 자료를 사용하며[4], 가격수준 및 가격형성요인이 상이할 것으로 예상되는 대도시, 중소도시 및 군 지역별로 대표적인 지역(시군구)을 각각 선정하여 사례분석을 실시한다.

연구방법은 우선 문헌조사를 통해 가격 추정을 위한 다양한 가치평가기법을 살펴보고자 한다. 본 연구에서는 여러 가치평가기법 중 주관성 개입이 적고 실행하는데 적은 비용이 소요되는 헤도닉 가격 모형(Hedonic Pricing Model)에 초점을 맞춘다. 지금까지 헤도닉 가격 모형 적용시 모수 모형을 집중적으로 사용하였으나, 본 연구에서는 이러한 방법론을 확장하여 엄격한 통계적 가정을 부여하지 않는 등 유연성이 높은 비모수 모형을 적용하고자 한다. 각 모형의 성능은 검증 데이터(Test Data)의 실제 거래가격과 모형 추정가격을 비교하여 판단한다. 즉 실거래가 신고자료의 30%를 임의 분할(random split)하여 검증 데이터로 유보한 뒤, 모형 성능 검증에 활용한다.

제 3 절 연구의 내용과 구성

본 논문은 총 6개의 장으로 구성되어 있다. 2장에서는 가치평가기법에 대한 이론적 검토를 한다. 즉 다양한 가치평가기법을 살펴보고, 그 중 헤도닉 가격 모형의 특징 및 주요 이슈를 파악한다. 다음으로 가격 함수의 형태에 따른 선형 모형과 비선형 모형을 고찰하고, 가격 함수의 비선형성을 반영하기 위한 비모수 모형에 대해 살펴본다. 기계학습 분야에서 제시된 이러한 비모수 모형들은 주로 분류(Classification)의 문제를 다루어 왔다. 즉 목표변수(Target Variable)의 형태가 이진 종속변수(Binary Response)였으나 본 연구에서의 목표변수는 주택가격, 즉 연속형 변수이므로 이러한 맥락에 맞추어 비모수 모형들을 검토한다.

3장에서는 비모수 모형을 실제 데이터에 적용하고 그 성능을 진단한

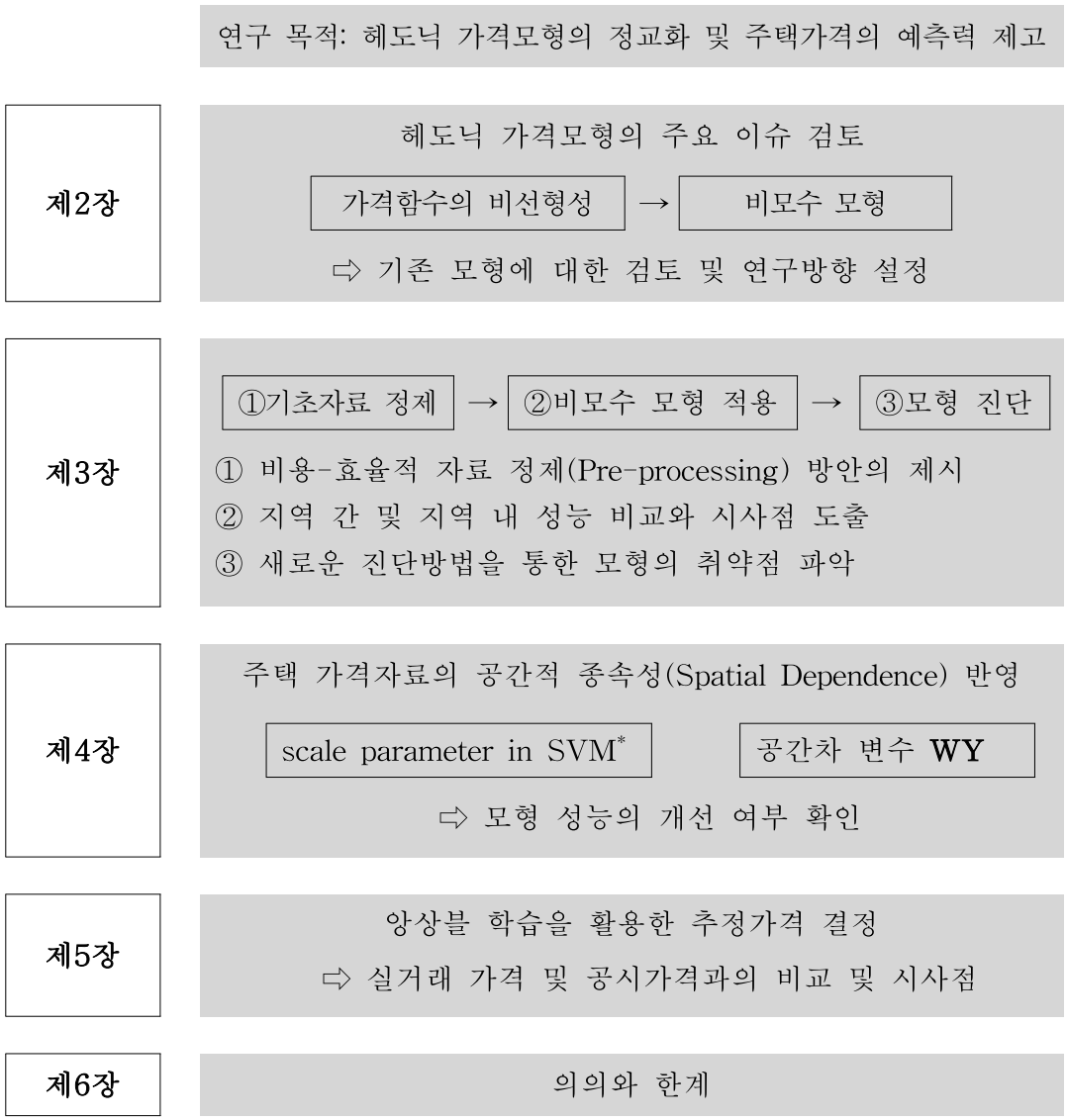
4) 실거래가 신고제도 도입 초기(2006년~2010년)에는 아직 제도가 정착되지 않아 신고가격의 신뢰성이 낮은 편이다. 이러한 이유로 2011년 자료부터 사용한다.

다. 즉 사례지역을 정하고 실거래가 데이터를 확보하되, 실거래가 데이터는 거래 당사자가 자발적으로 신고한 가격이므로 이상치와 잡음이 상당 부분 포함되어 있음을 감안한다. 즉 데이터에 포함된 이상치와 잡음을 체계적인 기준을 세워 사전에 제거하는 등 자료의 전처리(Pre-processing)를 비중 있게 실시한다. 이후 분석에 사용할 수 있는 수준까지 정제된 실거래가 데이터에 다양한 비모수 모형을 적용하고, 그 성능을 검증 데이터를 기준으로 하여 비교한다. 모형 성능의 진단은 RMSE(Root Mean Squared Error) 같은 전역적 지표(Global Index) 뿐 아니라 국지적 지표(Local Index)를 별도 개발하여 각 모형들이 어떠한 부분에서 특히 취약한지 파악한다.

4장에서는 공간자료의 특징인 공간적 종속성을 비모수 모형에 반영할 수 있는 방안에 대해 검토한다. 구체적으로 SVM(Support Vector Machine)에서의 scale parameter 조정, 시계열 분석에서의 시차 변수(Lagged Variable)와 유사한 공간차 변수(Spatially Lagged Variable)의 활용 등을 살펴본다.

5장에서는 이러한 공간적 종속성까지 반영된 비모수 모형들을 앙상블 평균으로 결합하여 추정가격을 결정한다. 또한 결정된 추정가격은 실제 거래가격 및 현행 주택공시가격과 비교하여 시사점을 도출한다.

마지막 6장에서는 본 연구의 의의와 한계를 정리한다. [그림 1-1]은 이와 같은 연구의 흐름을 보여준다.



[그림 1-1] 연구의 흐름도

* SVM: Support Vector Machine

제 2 장 이론적 고찰

제 1 절 헤도닉 가격 모형과 주요 이슈

1. 가치평가기법

가치평가에 관한 기법은 매우 다양한데, [표 2-1]은 시장에서 흔히 거래되는 일반적인 경제재[5] 뿐 아니라 환경재(Environmental Goods)[6]의 가치평가기법을 3개의 접근방식으로 나누어 정리한 것이다.

3개의 접근방식 중 첫 번째 직접관찰법(Direct Observation)은 시장에서 발생하는 가격 자료를 직접 살펴봄으로써 대상 재화의 가치를 추정하는 방법을 말한다. 이 방법은 다시 3가지 방식으로 나눌 수 있는데, 시장성(Marketability)에 근거한 시장접근법, 비용성(Cost)에 기반한 대체비용법, 회피가능손실법 및 생산함수법, 그리고 수익성(Profitability)에 근거한 순소득법이 그것이다. 시장성에 근거한 시장접근법은 대상 재화와 유사한 재화들이 시장에서 얼마에 거래되는지 파악하여 대상 재화의 가치를 추정하는 것으로, 거래가 빈번하고 관찰이 쉬운 주식, 채권 등에 수월하게 적용할 수 있다. 비용성에 기반한 대체비용법, 회피가능손실법 및 생산함수법은 재화의 공급자 입장에서 해당 재화를 대체 또는 재생산하는데 소요되는 원가를 계산하여 대상 재화의 가치를 추정한다. 마지막으로 수익성에 근거한 순소득법은 대상 재화를 소유함으로써 기대되는 미래 순편익(Future Net Benefits)을 적정한 할인율로 현재가치화(Discounting)하여 가치를 추정한다. 이러한 3가지 방식을 감정평가 3방식이라고도 하며(장희순・방경식, 2014), 시장이 균형 상태(Equilibrium State)에 있고 시장 참여자들이 모두 완전한 정보를 가지고 있다면 상기

5) 인간이 처분할 수 있어 매매의 대상이 되는 재화
6) 공기, 물, 산림, 하천, 호수 등 매매의 대상이 될 수 없으나 인간이 소비함으로써 효용을 느낄 수 있는 재화

3가지 방식에 의한 가격은 동일하거나 유사하여야 한다. 그러나 수요와 공급의 불균형, 정보의 비대칭, 거래 당사자의 협상력 차이, 정부의 규제 등 다양한 제약으로 3가지 가격은 일치하지 않는 것이 보다 일반적이다.

[표 2-1] 가치평가기법

접근방식	평가기법		내용
직접 관찰법 (Direct Observation)	시장접근법		시장에서의 거래가격을 직접 관찰(주식, 채권, 외환 등)
	비용성	대체비용법	해당 재화를 신규로 대체(신축)할 때 소요되는 비용을 추정
		회피가능손실법	해당 재화로 인해 회피할 수 있었던 손실을 추정
		생산함수법	판매 가능한 재화를 만드는데 필요한 투입물로서의 원가를 추정
	순소득법		해당 재화의 판매 수입에서부터 부수 비용을 차감한 후 현가화
현시선호법 (Revealed Preference)	여행경비법		해당 재화(통상 경관자원)까지 이동하는데 소요되는 비용을 추정(교통비, 숙박비 등)
	헤도닉 가격모형		재화를 구성하는 특성들의 가격을 통해 해당 재화의 가격을 추정
진술선호법 (Stated Preference)	가상조건부 평가 (Contingent Valuation)		해당 재화에 대한 지불용의액(willing to pay)을 설문 대상자에게 질의
	선택모형 (Choice Modeling)		해당 재화와 여타 재화 간의 교환 가능성 및 가능한 경우 교환 비율을 설문 대상자에게 질의

* 출처: Pagiola et al.(2004) 및 Graves et al.(2009)에서 발췌 및 재정리

3개의 접근방식 중 두 번째 현시선호법(Revealed Preference)은 대상 재화가 시장에서 직접 거래되는 것은 아니나, 시장 참여자들의 행위를 살펴 그들의 행동이나 의사결정을 통해 간접적으로 나타난(Revealed) 선호도를 파악하여 해당 재화의 가치를 추정하는 방법이다. 예를 들어 여행경비법은 특정 경관자원을 즐기기 위하여 관광객이 지출한 비용과 시간을 파악하여 해당 경관자원의 가치를 추정한다. 같은 맥락에서 헤도닉 가격모형은 모든 조건이 동일하지만 1가지 조건만 다른 주택들, 예를 들어 한강조망 여부만 차이가 있는 주택의 거래가격 차이를 살펴봄으로써

주택 매수자가 조망권에 대해 부여하는 가치를 추정할 수 있다.

마지막 세 번째 진술선호법(Stated Preference)은 시장 참여자들의 행위를 간접적으로 살펴보는 대신, 설문 등을 통해 시장 참여자들에게 지불용의액을 직접 질의하는 방법이다. 환경재 등 시장에서 거래되지 않는 재화의 지불용의액을 직접 질의하기 때문에 수요함수 도출 등의 복잡한 중간과정을 거치지 않고 힉스적 후생개념(Hicksian Welfare), 즉 동등잉여(Equivalent Surplus)나 보상잉여(Compensating Surplus)를 직접 이끌어 낼 수 있는 장점이 있다(곽승준・전영섭, 1995, p.52). 진술선호법 중 가상조건부 평가가 가장 일반적으로 활용되고 있으나 한 가지 속성에 대한 지불용의액만을 도출할 수 있다는 한계가 있다. 환경재의 다양한 속성들로 구성된 2개 이상의 대안을 제시하여 응답자가 선택하도록 하는 선택모형은 이러한 가상조건부 평가의 한계를 보완한 방법이라 할 수 있다.

이와 같은 다양한 가치평가기법 중 직접 관찰법은 일종의 정밀평가(Single-Property Appraisal)에 해당하는 것으로, 한 개 내지 두 개 정도의 특정 재화를 대상으로 시간과 비용을 집중적으로 투입하여 가치를 산정하는 방법으로 대량의 부동산 집단 가격을 일시에 산정하려는 본 연구의 내용과는 차이가 있다.

또한 진술선호법은 설문지의 구성, 가상 상황의 설정, 설문자와 응답자의 대상 재화에 대한 인지도 차이 등 여러 가지 요인으로 인해 다양한 편의가 발생할 수 있으며, 면접 또는 우편 방식에 의하므로 시간과 비용 또한 많이 소요된다.

반면 현시선호법 중 헤도닉 가격 모형(Hedonic Pricing Model, 이하 '헤도닉 모형')은 타 가치평가기법에 비해 실행하는데 상대적으로 적은 비용이 소요될 뿐 아니라 주관성 개입이 적어 부동산 가치추정에 폭 넓게 사용되는 방법이다.

헤도닉 모형은 19세기 말 등장한 한계 혁명(Marginal Revolution) 조류 속에 Lancaster(1971)가 제시한, 재화의 소비는 재화를 구성하는 특성에 의해 결정된다는 다속성 효용이론(Multi-attribute Utility Theory)에서 그 뿌리를 찾을 수 있다. 이후 Rosen(1974)이 헤도닉 모형을 통한 균형 시장가격 도출이 가능함을 이론적으로 증명함으로써 이 모형은 널리

활용되기 시작하였다.

헤도닉 모형은 '이질적인 재화의 가치는 해당 재화에 내포되어 있는 특성(Attributes)에 의해 결정된다'라는 가정을 전제하고 있다(Rosen, 1974). 그러나 특성들에 대한 가격은 관찰되지 않는데, 시장에서 개별적으로 거래되지 않기 때문이다. 시장에서 관찰되는 것은 이러한 특성들을 하나의 묶음으로 하여 거래된 재화의 가격이다. 따라서 재화의 가격을 특성들의 양(Quantity)에 대해 회귀(Regression)함으로써 특성 가격을 추정하는 것이다(이용만, 2008).

Rosen(1974)은 완전경쟁시장 하에서 회귀모형을 통해 구한 특성 가격이 특성에 대한 수요·공급에 의해 결정되는 균형가격과 같다는 것을 이론적으로 밝혔다. 또한 개별 수요자에 따라 헤도닉 모형의 함수 형태가 달라지지 않는다는 점을 증명하였다[7]. 그의 이러한 이론적 해명 덕분에 헤도닉 모형은 현재 광범위하게 활용되고 있다(이용만, 2008). 본 연구에서도 헤도닉 모형을 이용하여 주택 가격을 추정하고자 한다.

2. 헤도닉 모형

헤도닉 모형은 다양한 특성(속성)으로 구성된 재화, 예를 들어 자동차나 컴퓨터의 가격을 추정하기 위해 사용되다가 점차 토지나 주택의 가격 추정에 활용되기 시작하였다. 이후 토지나 주택처럼 시장에서 거래되어 그 가격을 비교적 쉽게 관찰할 수 있는 자산 뿐 아니라 시장에서 거래되기 어려운 자산, 예를 들어 산림의 공익가치, 습지의 생태가치 등까지 확대 적용되었다. 또한 관심의 범위를 보다 넓혀 특정 자산의 가치 뿐 아니라 특정 자산을 둘러싼 환경 내지 경제활동이 해당 자산의 가치 형성에 미치는 영향, 즉 외부효과(External Effect)를 측정하기 위한 연구가 늘어나기 시작하였다. 이 밖에도 최근 들어 부동산 시장상황 진단을 위한 각종 가격지수의 산정에 헤도닉 모형이 활용되는 등 적용 분야는 계속하여 확대되고 있다. 이하에서는 자산가치의 측정, 외부효과의 가치 측정, 가격지수의 산정 및 기타의 네 가지 활용분야로 나누어 살펴본다.

---

7) Rosen(1974)의 논문 발표 이후 헤도닉 함수의 설명변수에 수요자(거래 당사자) 특성(개인/법인, 성별, 소득, 인종 등)을 포함하는 관행이 상당 부분 사라졌다.

첫째, 자산가치의 측정은 시장에서 거래되어 가격 관찰이 용이한 토지 및 주택의 가치추정에서부터 초기 연구가 시작되었다. 이후 시장에서 거래되지 않는 것, 그래서 가치추정이 어렵거나 일반인들이 과소추정하기 쉬운 재화로 확대 적용되었다. 대표적인 예로 산림(Tyrvainen, 1997), 습지(Woodward & Wui, 2001), 학교(Downes & Zabel, 2002), 호수(Lansford & Jones, 1995), 해변(Gopalakrishnan et al., 2011), 공원(Morancho, 2003), 산책로(Parent & Hofe, 2013) 등이 있으며, 특히 동남아시아 망그로브 숲의 가치는 90년대 이후 매우 빈번하게 그리고 다양하게 연구가 이루어졌다(Spaninks & Beukering, 1997; Sathirathai et al., 2001; Williams, 2005; Zhao et al., 2007).

둘째는 외부효과의 가치측정으로서, 크게 긍정적 외부효과 및 부정적 외부효과의 측정으로 구분할 수 있다. 긍정적 외부효과로는 조망(Fraser & Spencer, 1998), 일조(Jim & Chen, 2009), 역사문화적 건물(Nijkamp et al., 2011), 전철역 개통(Bae et al., 2003) 등으로 인한 자산가치의 상승을 들 수 있다. 반면 부정적 외부효과에 관한 것으로는 자동차, 기차 및 비행기 등으로 인한 소음(Theebe, 2004; Nelson, 2008), 폐기물 매립지의 존재(Reichert, 1997), 원자력 발전소의 설치(Kinnard et al., 1991), 공기의 질(Kim et al., 2003), 수질(Leggett & Bockstael, 2000), 소수 계층의 밀집 거주지역이 주변 토지가격에 미치는 영향(Kiel & Zabel, 1996; Myers, 2004), 오염 부동산(Smolen et al., 1992; Simons, 1997)으로 인한 자산가치 변동 연구 등이 있다. 특히 부정적 외부효과에 관한 연구는 사회적 소외계층이나 경제적 약자, 또는 이들이 거주하는 주택이나 인근지역이 상기와 같은 부정적 외부효과에 집중적으로 노출되어 환경정의(Environmental Justice)가 지켜지지 않고 있다는 주장을 일관되게 제시하고 있다.

셋째는 가격지수의 산정으로서 헤도닉 모형을 활용하여 주택매매지수나 임대료 지수 등(Can & Megbolugbe, 1997; Wallace & Meese, 1997; Wu et al., 2014)을 작성·공표함으로써, 민간에는 거래나 투자의 지표, 정부에는 시장개입의 시기나 강도 등을 판단하는 기초자료로 사용되고 있다.

마지막으로 매수자-매도자의 정보 비대칭(Harding et al., 2003), 중개

인의 활용 등 매도자의 매도전략(Yavas & Yang, 1995), 법령 개정(예를 들어 대상 부동산과 관련된 모든 정보를 매도자가 공개해야하는 의무의 신설)에 따른 거래가격 변화(Pope, 2008) 등 매수자-매도자의 행동 변화를 연구한 분야(Behavioral Research)를 들 수 있다.

헤도닉 모형의 응용 분야는 상기의 내용처럼 확대되어 왔으며, 방법론도 선형회귀모형(Linear Regression Model)을 시작으로 하여 다음과 같이 정교화되어 왔다.

헤도닉 모형 중 선형회귀모형은 그 논리적 근거와 적용 절차가 확고하게 정립되어 실무자 및 학자들이 오랫동안 사용하여 온 전통적 모형이다(Zurada et al., 2011). 그러나 통계학 내지 경제학 분야에서 제시된 이러한 선형회귀모형은 공간자료가 갖는 가장 기본적인 특징, 즉 공간적 종속성(Spatial Dependence)과 이질성(Spatial Heterogeneity)을 고려하지 않아 추정가격이 부정확하다는 지적이 제기되었고, 이는 보다 정교한 모형을 개발하는 동기가 되었다.

방법론 측면에서 헤도닉 모형은 여러 관점에서 그 발전 과정을 추적할 수 있으나 본 연구에서는 전통적인 선형회귀모형이 충분히 고려하지 못한 두 가지 공간자료의 특징, 즉 공간적 종속성과 이질성을 모형에 반영하려는 연구 경향에 초점을 두어 설명한다.

자료의 독립성 가정을 완화하고 인근의 관찰치는 원거리 관찰치보다 유사한 값을 갖는다는 공간적 종속성을 명시적으로 고려한 대표적 모형이 바로 공간회귀모형(Spatial Regression Model)이다. 이 모형은 Anselin(1988) 이후 현재까지 부동산 가격추정에 꾸준히 사용되고 있다(Can, 1992; 김성우, 2010; 송용철・박헌수, 2012; Dube & Legros, 2013; Parent & Hofe, 2013). 그러나 공간회귀모형은 공간적 종속성을 수치화하는 공간가중행렬(Spatial Weight Matrix) 구성의 주관성이 단점으로 지적되고 있으며, 모형의 세부 종류(공간시차모형, 공간오차모형, 공간더빈모형 등)가 많아 모형 선택에 어려움이 존재한다.

점(Point) 자료를 다루는 공간 모델링 분야, 즉 Geostatistics에서 자료의 공간적 종속성을 반영하는 대표적 기법이 바로 크리깅(Kriging)이다. 크리깅은 공간보간을 위한 대표적인 기법으로 관찰되지 않은 지점의 예

측값을 주변 관찰지점 값의 가중선형조합으로 산출하는 방법이다(Isaaks & Srivastava, 1989). 크리깅은 환경과학 분야에서 가장 일반적으로 사용되는 방법임에도(Webster & Oliver, 2007) 불구하고, 부동산 가격추정 등 국내의 사회과학 분야에서는 잘 시도되지 않고 있다. 그러나 해외의 경우 부동산 가격추정을 위한 연구에서 자료의 공간적 종속성을 크리깅 기법으로 모형에 반영하고자 하는 노력을 찾아 볼 수 있다(Militino et al., 2004; Chica-Olmo, 2007; Montero & Larraz, 2011; Kuntz & Helbich, 2014).

크리깅 기법을 적용한 상기 연구들은 모두 일종의 회귀-크리깅 모형(Regression-kriging Model)을 적용한 예에 해당한다. 회귀-크리깅은 종속변수인 부동산 가격에 대해 직접 공간보간을 하는 것이 아니라, 부동산 가격을 예측할 수 있는 체계적 요인들(평균 구조 또는 전역적 경향)은 선형회귀모형을 통해 통제한 후 잔차, 즉 설명되지 않는 변이에 대해 크리깅 기법을 적용하는 방법이다. 따라서 회귀-크리깅은 선형회귀모형과 공간 보간기법인 크리깅을 결합한 모형이며 토양학, 기후학 등 자연과학 분야에서 시작하여(Hudson & Wackernagel, 1994; Li, 2010; Zhu & Lin, 2010) 부동산 가격추정을 비롯한 사회과학 분야로 점차 저변을 확대하고 있다.

부동산 가격은 본질적으로 속성정보에 의해 설명될 수 있는 구조적 측면과 위치정보에 의해 설명될 수 있는 공간적 측면으로 나눌 수 있는 바, 회귀-크리깅 모형은 이러한 두 가지 사안을 동시에 고려할 수 있어 매우 효율적인 가격추정 방법이 될 수 있다.

다음으로 공간적 이질성은 여러 가지 정의가 있을 수 있으나 관측치의 평균값이 세부지역에 따라 변하는 현상(Waller & Gotway, 2004, p.204)으로 이해할 수 있으며 부동산 시장의 경우 시장 내부에 가격 수준이 상이한 하부시장(Submarket)의 존재, 철로나 도로 등 인공 건조물에 의한 지가의 불연속 현상 등을 예로 들 수 있다.

공간적 이질성을 계량화하는 대표적인 모형이 지리적 가중회귀모형(Geographically Weighted Regression, GWR)이다. GWR은 계수 추정을 위해 전체 데이터의 일부분만 사용하는 일종의 국지모형(Local Model)이

며, 데이터의 반복 사용으로 인한 다중공선성 문제 등으로 탐색적 분석 단계에서 보다 많이 활용되고 있다. Fotheringham et al.(2002)에 의해 제시된 이후 부동산 가격추정 등 공간 사상의 지도화가 필요한 분야에서 폭 넓게 활용되고 있다.

최근에는 부동산 가격추정시 기본적인 GWR 모형을 좀더 발전시켜, 공간상에서 가변적인 값과 고정 값을 갖는 계수를 구분하여 GWR 모형을 적합시키려는 시도(Mixed GWR)가 이루어지고 있다(Geniaux & Napoleone, 2008; Helbich et al., 2013). 또한 공간상의 이질성을 파악하는 도구로서의 GWR 모형과 종속변수의 전체 분포 패턴을 파악하려는 분위회귀모형(Quantile Regression)을 결합한 GWQR(Geographically Weighted Quantile Regression) 모형의 적용(Chen et al., 2012), GWR 모형에 시간 요소를 추가하여 주택 가격을 추정한 GTWR(Geographically and Temporally Weighted Regression) 모형의 적용 사례(Huang et al., 2010) 등도 찾아 볼 수 있다.

자료의 이질성은 공간 측면 뿐 아니라 속성 측면에서도 파악할 수 있는데, 예를 들어 부동산의 이용상황(주거용, 상업용 등)에 따라 가격형성 요인이 상이한 경우 이는 속성 측면의 이질성이라 할 수 있다. 공간 및 속성 측면에서 자료의 이질성을 계량화하기 적합한 모형으로 다수준 모형(Multi-level Model)을 들 수 있다. 다수준 모형은 공간 및 속성 등 어떠한 측면에서든 포섭구조(Nested Structure)를 갖는 자료의 경우 적용할 수 있다. 예를 들어 가격자료가 2개 이상의 상이한 공간 수준(필지와 인근지역)에서 측정되었다면, 각 공간 수준에 따라 상이한 오차항 분산을 가정하는 것이 보다 합리적이다(Jones & Bullen, 1993). 필지는 인근지역이라는 상위 지역에 포섭되는 구조이므로, 동일 인근지역에 속한 필지들끼리는 여러 항목에서 유사하지만 다른 인근지역에 속한 필지들과는 상이한 점, 즉 이질성이 높을 것이다. 따라서 인근지역별로 뚜렷이 구분되는 자료의 이질성을 계량화할 때 다수준 모형이 효율적이다. 속성 측면에서도 예를 들어 필지는 자신이 속한 부동산 유형별로(주거용, 상업용, 농업용 등) 가격 수준이 형성되는 경우가 일반적이다. 이 경우 필지는 부동산 유형에 포섭되는 구조로 보아 다수준 모형을 적용할 수 있다. 또는 부동산 시장은 지역별로 그리고 유형별로 분화되어 있음이 일반적

이므로(이정전, 2013, p.180) 지역과 유형을 동시에 고려한 포섭구조를 적용할 수도 있다.

부동산 자료가 갖는 이러한 이질성을 반영하기 위해 공간이나 속성 측면에서 자료를 미리 층화하는 것은 추정가격의 정확성을 높이기 위한 효율적인 절차라고 할 수 있다(Mark & Goldberg, 1988).

3. 헤도닉 모형의 주요 이슈

선형회귀모형이 제시된 이후 자료의 종속성과 이질성을 고려한 다양한 헤도닉 모형이 앞서 설명에서처럼 개발되어 이제 이러한 공간자료의 특징은 더 이상 특별한 문제거리로 취급되지 않게 되었다. 그러나 헤도닉 모형은 여전히 실행 단계별 중요한 부분에 대해서 뚜렷한 이론적 근거를 제공하지 못하고 있다. 헤도닉 모형이 이론적 근거를 제공하지 못하는 부분은 선행연구 결과 등을 참고하거나 연구자의 달관(達觀)으로 보완할 수밖에 없는데, 이는 헤도닉 모형의 신뢰성을 크게 약화시킨다. 이러한 대표적인 약점으로 모형 적용의 지역적 범위 문제, 설명변수의 선택 문제, 그리고 가격 함수의 형태 문제 등 세 가지를 들 수 있다(Goodmand & Thibodeau, 2003; Du Preez et al., 2013).

가. 모형 적용의 지역적 범위 문제

헤도닉 모형을 적용할 지역적 범위 문제는 지리학에서의 공간 단위 임의성 문제(Modifiable Areal Unit Problem, MAUP)와 관련이 깊다. 공간 단위 임의성 문제, 즉 MAUP는 자료의 수집단위에 따라 분석 결과가 달라질 수 있다는 사실을 의미하는 것으로, 오래 전부터 공간 데이터의 독특한 특징으로 인식되어 왔다(Holt et al., 1996).

이러한 MAUP는 부동산 가격 추정과 관련하여 '하부시장 구획'이라는 주제로 보다 빈번하게 다루어지고 있다. 즉 헤도닉 모형을 적용할 특정 지역 내에 하부시장이 존재함에도 이를 간과하거나 하부시장 구획이 부

정확한 경우 헤도닉 모형 적용 결과의 신뢰성은 저하될 수밖에 없기 때문이다. 고용 중심점이라든지 도시의 인프라(간선도로, 고속도로, 교량 등)는 부동산 가격의 차별 현상을 발생시키는 등 대표적인 하부시장 구획 요인으로 알려져 있다.

헤도닉 모형 적용의 지역적 범위는 주어진 행정구역을 그대로 사용하거나 인근지역의 질을 나타내는 몇몇 변수(주민 소득, 도심과의 거리 등)를 이용하여 하위 행정구역을 몇 개의 커다란 상위 행정구역으로 묶는 방법(원제무 외, 2009; Watkins, 2002), 또는 부동산 관련 전문가가 사전에 설정한 하부시장 범위를 그대로 습용하는 방법(Bourassa et al., 2007) 등이 지금까지 주로 사용되었다.

헤도닉 모형의 가격 예측력을 결정하는 가장 중요한 사안 중의 하나임에도 불구하고 모형 적용의 지역적 범위가 활발하게 다루어지지 않은 이유는 다음과 같은 두 가지 어려움 때문으로 풀이된다. 첫째는 명확한 경계를 가지는 지역 개념에서부터 그러한 경계는 사실상 존재하지 않는다는 견해(Paez et al., 2008)에까지 지역적 범위가 무엇인지에 대한 일치된 정의 자체가 존재하지 않는다. 둘째는 군집분석, 공간적 연접성을 고려한 Automatic Zoning Procedure 등 공간 구획을 위한 방법론이 매우 다양하게 제시되었지만 모두 일반화하기 어려운 경우가 대부분이다. 즉 정의와 방법론 모두에 대해 합의된 바가 없어 연구 실적이 미진한 분야라고 할 수 있다.

국내의 경우 최근 이건학・김감영(2013)은 주택공시가격과 실거래 가격 차이의 발생에 대해 헤도닉 모형(비준표) 작성에 내재된 MAUP를 검토하는 등 모형 적용의 범위에 대한 중요성 인식은 점차 증가하고 있다. 모형 적용의 지역 범위를 좁게 정할수록 해당 지역 내에는 비교적 균일한 부동산 집단만 소재하게 되는 이점이 있으나 표본 수가 적어지는 단점이 있다. 반면 지역 범위를 넓게 확대할수록 표본 수가 많아지는 이점은 있으나 이질적인 지역들을 포함하게 되어 가격 예측력은 오히려 약화될 수 있다.

이론적으로 명확한 가이드라인이 없으므로 헤도닉 모형의 성능이 가장 우수하게 나타날 수 있도록 상기 두 가지 경우의 상쇄관계

(Trade-off)를 고려하여 중간 정도 수준의 공간 범위를 대상지역으로 정하는 것이 일반적이다. 또한 실무상 별도의 지역을 구획하기보다는 시도 또는 시군구와 같은 이미 주어진 행정경계를 따르는 경우가 통상이다. 본 연구에서는 비교적 공간 범위가 좁은 시군구를 사례지역으로 선택하였으며 주어진 행정경계를 따라 분석을 수행하였다. 이는 본 연구의 초점이 비모수 모형의 적용 등에 있고, 지역 범위의 확장 또는 축소에 따른 결과의 변동성은 별도의 연구 주제라 판단하였기 때문이다.

나. 설명변수의 선택 문제

헤도닉 모형은 가격에 영향을 미치는 중요한 설명변수들이 모두 적정하게 선별되어 모형에 포함되었다고 가정한다. 그러나 '중요한' 설명변수들을 어떻게 선별할 것인지에 대해서 명확한 기준이나 절차가 정립되어 있지 않다.

따라서 설명변수의 선택 또한 주관적 요소가 개입될 수밖에 없다. 가장 손쉬운 선택방법은 전진 선택법(Forward Selection), 후진 선택법(Backward Selection), 단계적 회귀분석(Stepwise Regression) 등을 통해 획일적으로(p-value 기준) 설명변수를 선별하는 것이다. 많은 연구에서 활용되었으나(Conway & Lathrop, 2005; Dunse & Jones, 1998; Kong et al., 2007) 그 결과의 신뢰성이나 일반화 가능성은 그리 높지 않다.

p-value에 의존하는 이러한 방법과는 달리 선행연구 결과, 전문가 면담, 실무 경험 등을 토대로 중요 설명변수를 선별하려는 연구도 있었으나(Anderson, 2000) 단편적인 실무 지식을 활용하는데 그치거나 자의성 개입이라는 지적으로부터 자유롭지 못한 편이다.

최근 기계학습 분야에서는 전체 설명변수 집합 중 부분 집합(subset of features)을 반복적으로 생성한 후, 이들 부분 집합에 대해 일정한 평가 점수(evaluation scores)를 부여하여 최적 설명변수 집합을 선택하는 알고리즘이 활발하게 제시되고 있다('feature selection algorithm'이라 한다). 기계학습 분야의 이러한 알고리즘은 대개 설명변수 부분 집합을 반복적으로 선택한 후, 이들 중 사전에 정한 오류율 지표를 최소화하는 부

분 집합을 찾는 방식으로 구성되어 있다. 이 경우 어떻게 오류율 지표를 구성할지가 부분 집합 선택에 가장 큰 영향을 미치게 된다.

기계학습 분야의 이러한 설명변수 부분 집합 선택의 대표적인 접근방법에는 능형 회귀(Ridge Regression), LASSO(Least Absolute Shrinkage and Selection Operator), 랜덤 포리스트(Random Forest), 부스팅(Boosting) 등이 있다(Meinshausen & Bühlmann, 2006; Zhao & Yu, 2006; Tian et al., 2015). 특히 LASSO는 모형 구축과 설명변수 선택을 동시에 수행하는 방법으로, 회귀계수에 대해 일정한 벌점을 부여하고(penalizing), 중요하지 않은 설명변수들에 대해서는 계수값을 0으로 축소('shrinking')시킨다. 궁극적으로 0이 아닌 계수값을 갖는 설명변수들은 LASSO 알고리즘에 의해 선택된 '중요한' 변수들인 것이다.

기계학습 분야의 이러한 설명변수 선택 알고리즘은 관찰치(n)보다 변수의 수(p)가 많은 고차원 자료(p>n)의 경우 매우 효과적으로 설명변수를 선별할 수 있는 것으로 알려져 있다. 예를 들어 10명의 암 환자를 대상으로 한 100종의 임상실험 측정치를 분석하여야 하는 의학 분야, 희귀식물 5개에 대한 100종의 유전자 측정치를 검토하여야 하는 생물통계 등에서 활발하게 응용되고 있으나 고차원 자료를 찾기 힘든 부동산 분야에서는 잘 시도되지 않고 있다.

본 연구에서는 전진 선택이나 후진 선택 같은 획일적인 방법 대신 선형회귀모형의 반복적 적합을 통해 회귀계수의 p-value가 일관되게 유의하거나(5% 이하) p-value 값 자체는 유의하지 않더라도 계수의 부호가 일반적인 직관과 일치하는 변수들은 모두 설명변수에 포함시키고자 하였다. 이는 통계적 판단 기준과 부동산 실무 경험을 적절하게 결합한 방법으로 볼 수 있다. 반면 본 연구에서 활용한 실거래가 자료는 후보 설명변수의 개수 자체가 많지 않아 고차원 자료에 적합한 LASSO와 같은 알고리즘은 적용하지 않았다[8].

8) 탐색석 분석 단계에서 LASSO 알고리즘을 적용한 결과, 설명변수를 지나치게 과소 선정하는 경향이 있는 것으로 파악되었다.

다. 가격함수의 형태 문제

헤도닉 모형을 통한 가격 예측에 있어 또 하나의 중요한 관건은 가격 함수 형태의 결정이다. 종속변수인 가격과 이에 영향을 미치는 설명변수들 간에 어떠한 관계가 형성되어 있는지 파악하는 작업은 쉬운 일이 아니다. 잘 들어맞는 유용한 이론이 제시되어 있는 것도 아니며, 이론이 존재한다 하더라도 실제 측정하기 어려울 수 있다(Mason & Quigley, 1996).

따라서 가격과 설명변수들 간의 관계를 선형(Linear)으로 가정하고 이러한 가정이 현실과 크게 틀리지 않는다는 기대 하에 접근하는 것이 지금까지의 일반적인 해결책이었다. 다만 선형 관계를 가정하는 것이 명백하게 불합리한 경우, 다항변수(2차항, 3차항, 상호작용항 등)나 변수 변환(자연로그, 역수, 제곱근 등)을 통해 해결을 시도하였다(Cropper et al., 1988).

그러나 이러한 방법은 모두 모수적 접근에 해당되는 것으로 항상 설계 오류(model specification error)에 노출되어 있는 셈이다. 이러한 오류를 피하기 위해 최근에는 가격과 설명변수들 간의 함수 형태를 사전에 정하고 분석을 진행하기보다는 함수 f 자체를 찾으려는, 즉 비모수적 접근이 많이 시도되고 있다(Bajari & Kahn, 2005; Redfearn, 2009; McMillen, 2010 등).

비모수적 접근은 처음부터 잘못된 함수 형태를 설정하는 오류를 피할 수 있고, 다양한 형태의 함수를 탐색할 수 있어 기존의 모수적 접근보다 '일반화'된 방법이라고 할 수 있다. 또한 비모수 모형은 Pace(1993)가 제시한 이상적인 모형의 2가지 조건을 모두 충족시킬 수 있다. 즉 이상적인 모형은 첫째 설계 오류에 민감하지 않고, 둘째 이상치에 강건할 필요가 있는데, 비모수 모형은 이러한 두 가지 요건을 모두 충족시킬 수 있다. 비모수 모형은 설명변수의 변환 등 모형의 형태를 변화시켜도 그 예측값에 큰 변화가 없고, 이상치로 판단되는 관찰치를 포함 또는 제외하였는지 상관 없이 예측값이 비교적 일관성을 유지하기 때문이다(Pace, 1993).

그러나 비모수 모형은 가격 함수 f가 다양한 형태를 가질 수 있다는 가능성을 열어두게 되어, 그만큼 추정에 필요한 데이터가 많아야 하는

단점이 있다.

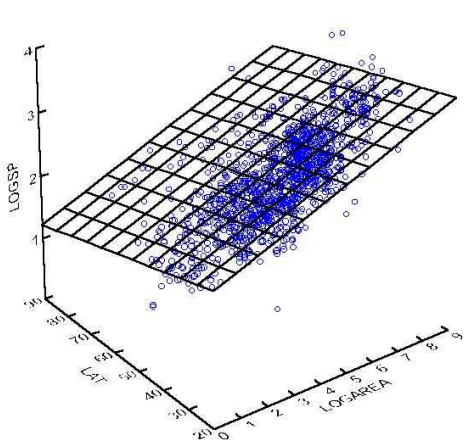
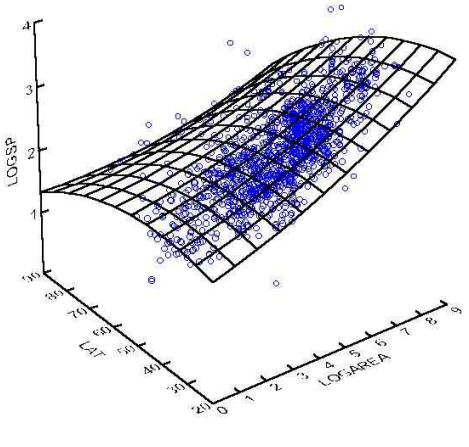
다음 절에서는 가격 함수의 형태는 선형으로 가정할 만큼 단순하지 않다는 전제하에, 선형 모형과 비선형 모형의 의미를 개관하며, 부동산 가격 추정에 있어 대표적인 비선형성 예를 살펴보고자 한다. 또한 이러한 비선형성을 반영할 수 있는 모형에 대해 검토한다.

제 2 절 헤도닉 가격 함수의 비선형성

1. 선형 모형과 비선형 모형

회귀분석과 같은 통계적 모형에 있어 선형성(Linearity)이란 설명변수의 값이 변할 때 해당 설명변수와 관련된 모수 값이 일정하게 유지되는 성질을 말한다. 예를 들어 종속변수가 토지가격, 설명변수가 토지면적일 때, 토지면적에 대한 모수가 100만원이었다면 토지면적이 1㎡에서 2㎡로 증가할 때 토지가격이 100만원 증가하고, 1,000㎡에서 1,001㎡로 증가할 때에도 역시 토지가격은 100만원 증가함을 의미한다.

즉 설명변수가 하나인 단순 회귀분석이라면 종속변수와 설명변수의 관계가 직선의 관계를 보인다는 의미이며, 설명변수가 2개 이상인 경우라면 종속변수와 설명변수의 관계가 휘어지지 않은 초평면(Hyperplane)의 형태임을 의미한다. [그림 2-1]은 설명변수가 2개인 경우의 선형 및 비선형 초평면을 보여 주는데, 유연한 형태의 비선형 초평면이 대부분의 현실 관계를 보다 잘 설명할 수 있음을 짐작할 수 있다.

선형의 초평면	비선형 초평면
	

[그림 2-1] 선형 및 비선형 초평면

* 출처: Ordination methods for ecologists, www.ordination.okstate.edu

함수 형태에 대해 선형성을 가정하는 것은 모형을 단순하게 구성할 수 있고, 해석이 용이해지며 실무적으로 실행하기에도 수월한 점 등 많은 장점을 가지고 있다. 그러나 현실에서 종속변수와 설명변수는 복잡다기한 형태의 비선형적인 관계를 맺고 있는 것이 통상이다. 진정한 함수 형태가 [그림 2-1]의 좌측처럼 선형인 경우는 오히려 극단적인 예에 해당된다(Hastie et al., 2009, p.139). 부동산의 경우에도 마찬가지여서 가격과 부동산 특성은 비선형으로 보는 것이 보다 현실에 부합하며(Weirick & Ingram, 1990), 선형성을 가정하는 것은 단지 편의를 위한 근사적 계산에 불과하다(Ekeland, 1988).

경제학 측면에서 함수 형태가 선형이라는 것은 변수 값의 양(Quantity)에 관계없이 해당 특성(속성)이 발휘하는 한계가치(Marginal Utility)가 불변임을 의미한다. 그러나 이는 경제학의 한계효용 체감법칙에 정면으로 배치될 뿐만 아니라(Maclennan, 1977) 일반적인 직관에도 반한다. 어떠한 재화 또는 용역이든 그 존재량이 많아지면 희소성이 떨어져 가치는 하락하기 때문이다. '물과 다이아몬드의 역설'에서 인간생활에 필수불가결한 물은 값이 싼 데 반해, 없어도 살 수 있는 다이아몬드의 값은 매우 비싸다. 이러한 역설은 물은 존재량이 많다보니 아무리 인간생활에 꼭 필요하더라도 한계효용이 작고(희소성 없음), 다이아몬드는 존재량이 적다보니 한계효용이 높아 가격이 비싼 것이다. 즉 모든 재화

의 한계가치는 그 양에 따라 변할 수밖에 없지만, 선형성 가정은 이러한 사실을 부정하는 것이다.

함수 형태가 선형성이 성립되는 것은 오히려 예외적이라는 주장은 Rosen(1974)이 헤도닉 모형 이론을 정립하기 이전(Lessinger, 1969)까지 거슬러 올라간다. 사실 헤도닉 모형 이론의 어느 부분을 살펴 보더라도 여러 설명변수들의 값이 변할 때 회귀분석에서의 계수 값이 일정하게 유지되리라는 보장은 없다(Kummerow & Galfalvy, 2002; Gloudemans & Almy, 2011).

선형모형은 지난 30년간 통계학의 주류였으며 현재에도 매우 중요한 방법론으로 남아 있다(Hastie et al., 2009, p.11). 선형모형은 설명변수와 모수와의 관계에 대하여 상당히 많은 가정을 하며, 그 예측결과는 비교적 '안정적'이지만 부정확할 수 있다. 반면, 비선형모형은 설명변수와 모수에 대해 특별한 가정을 하지 않으며, 그 예측결과는 '정확'할 수 있으나 불안정한 편이다(James et al., 2013, p.265).

기본적으로 선형모형에서 예측치 $\hat{Y}$은 다음과 같이 표현할 수 있다.

$$\hat{Y}= \hat{\beta}_0+ \sum_{j=1}^{p} X_j\hat{\beta}_j \tag{2-1}$$

반면 비선형모형의 대표적인 예로 knn(K-Nearest Neighbor Method)의 경우 예측치 $\hat{Y}$은 다음과 같이 표현할 수 있다.

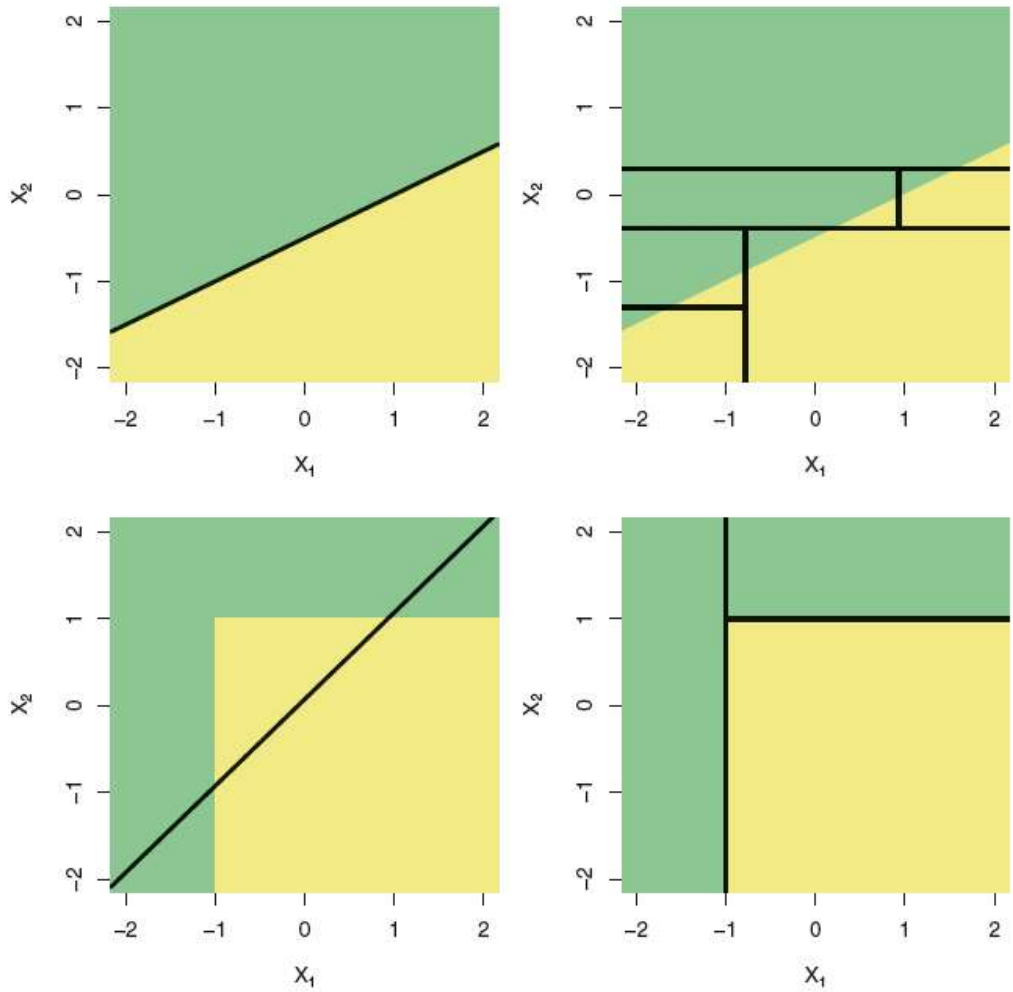
$$\hat{Y}(x)= \frac{1}{k} \sum_{x_i \in N_k(x)} y_i \tag{2-2}$$

$N_k(x)$는 설명변수 x와 가장 유사한 값을 갖는 k개 x_i의 인근집단을 말하며, 이렇게 선택된 인근집단의 종속변수값을 평균한 수치가 예측치가 된다.

현대 통계학에서 쓰이는 대부분의 모형은 상기 두 가지 방법을 각각 정교화시키거나, 두 가지 방법을 다양한 배합으로 병용한 것으로 해석할 수 있다.

그렇다면 두 가지 모형 중 어떠한 모형이 보다 우수하다고 할 수 있는가? 비선형 모형이 선형 모형보다 항상 우월하다고 단언하기는 어렵다. 경우에 따라 종속변수와 설명변수의 관계가 선형의 형태로 일정하거나 일정한 것으로 간주할 수 있다면 선형모형이 보다 우수할 것이다. 반면, 종속변수와 설명변수의 관계가 일정하지 않고 비선형의 형태를 보이거나, 상호작용 효과(Interaction Effects)가 크다면 비선형 모형이 보다 우수한 결과를 산출할 것이다. [그림 2-2]는 이러한 예를 보여준다. 상단의 두 그림은 두 변수 X_1, X_2의 진정한 관계가 선형일 때의 예측 결과(여기에서는 분류 결과)를 보여준다. 이 경우 좌측의 선형모형이 우측의 비선형 모형(knn 모형 등)보다 분류 결과가 보다 우수함을 알 수 있다. 반면 하단의 두 그림은 진정한 관계가 비선형일 때의 분류 결과로서, 우측 비선형 모형의 성능이 좌측 선형 모형보다 우수함을 알 수 있다.

그러나 앞서 언급하였듯 실제 현실에서 종속변수와 설명변수의 함수형태가 선형인 경우는 거의 없다고 할 수 있으며, 따라서 복잡다기한 현실 관계를 설명하고 예측하는데 있어서 비선형 모형이 보다 효과적이라고 할 수 있다.



[그림 2-2] 변수간의 선형관계 여부

* 출처: James et al.(2013) p.315에서 인용

2. 부동산 가격과 설명변수 간의 비선형성

부동산 가격과 설명변수와의 관계가 선형이 아니라는 사실, 즉 설명변수 값이 변동함에 따라 가격도 그에 따라 비례하여 변하지 않는다는 것은 여러 부동산 특성항목(설명변수)에서 찾아 볼 수 있다. 면적, 건물의 경과연수, 고급 부대시설(에어컨, 벽난로, 수영장 등)이 이러한 예에 해당한다.

가. 부동산 면적

토지 면적 또는 건물의 연면적은 가격과 선형성이 성립되지 않는 대표적인 특성항목이다. 즉 면적이 증가할수록 대응되는 단가(원/㎡)는 체감하기 마련이며 경제학에서의 한계효용 체감법칙과 유사한 현상이라 할 수 있다.

부동산 평가실무에서는 이러한 현상을 '광평수 감가(廣坪數 減價)'라 일컫는다. 광평수 토지는 인근지역의 표준적인 이용 규모를 훨씬 초과하는 토지로서, 환가성이 떨어지고 최고의 가치를 창출할 수 있는 용도로 사용하기 어려워 가격을 낮게 책정하는 것이 일반적인 관행이었다. 평가실무에서의 이러한 관행은 [표 2-2]와 같은 임야 가격 비준표[9]에서도 나타난다.

[표 2-2] 임야가격 비준표 (강원도 홍천군 농림지역, 2014년 기준)

비교토지＼대상토지	~ 3,300㎡	~16,500㎡	~33,000㎡	~66,000㎡	그 이상
~ 3,300㎡	1.00	0.98	0.97	0.96	0.95
~16,500㎡	1.02	1.00	0.99	0.98	0.97
~33,000㎡	1.03	1.01	1.00	0.99	0.98
~66,000㎡	1.04	1.02	1.01	1.00	0.99
그 이상	1.05	1.03	1.02	1.01	1.00

* 출처: 한국부동산연구원 비준표 열람사이트, (http://www.kreri.re.kr/lprt/view/index.asp)

상기 표의 첫 행을 보면 비교 토지가 3,300㎡ 이하의 토지이고 평가하고자 하는 대상 토지가 66,000㎡를 초과하는 경우 5% 감가하여 가격 배율 0.95가 적용됨을 알 수 있다.

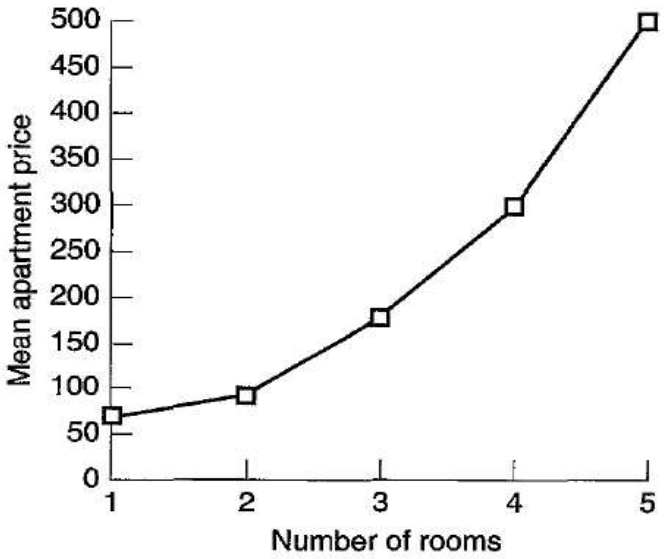
이러한 광평수 감가 경향은 외국에서도 동일하게 발견되는 현상이다. Keith(1991)는 농지를 할인현금흐름수지법(Discounted Cash Flow Analysis)으로 평가하면서 농지 면적 증가에 따른 거래단가 체감 현상이 모형 구성의 중요한 요소임을 실증적으로 밝혔다. 즉 작은 규모의 농지에 대해서는 구매 수요가 많아 단가가 높게 형성되지만, 이후 면적이 증가할수록 환가성이 떨어져 단가가 하락하며, 일정 면적을 넘어서면 더 이상 해당 농지에 대한 수요가 없어 수평선 형태를 유지한다.

9) 개별공시지가를 산정하기 위해 행정기관에서 사용하는 일종의 간이 가격배율표이다.

Keith(1991)는 이 때 단가 체감의 기울기는 해당 농지의 배후마을 크기(인구 수)에 따라 상당히 달라진다고 주장하였다.

광평수 감가가 일반적인 현상이긴 하나 최근에는 그 반대인 광평수 증가 현상도 발견할 수 있다. 도시화 및 고밀도 개발에 따라 대도시 지역에서는 오히려 광평수 토지가 없어 대규모 아파트 단지나 대형 할인마트 부지를 찾기 어려운 경우가 많다. 이 경우 면적의 증가는 오히려 가치 할증요인으로 작용하게 된다.

또한 Huh & Kwak(1997)의 연구에서도 광평수 증가와 비슷한 현상이 발견된다. 이들은 한국의 아파트 방 개수와 가격 간에는 선형 관계가 아니라 체증 관계가 성립하며([그림 2-3] 참조), 따라서 헤도닉 모형 설계 시 이러한 관계를 함수 형태에 반영하여야 한다고 주장하였다. 이들은 [그림 2-3]과 같은 체증 형태가 나타나는 이유에 대해, 한국 정부가 30평대 이상의 중대형 아파트 신축은 규제를 강화하고, 대신 소형 아파트의 신축을 일정 부분 의무화하여 중대형 아파트에 대해서는 초과 수요가 존재하기 때문이라고 해석하였다.



[그림 2-3] 아파트 가격(평균)과 방 개수와의 관계

* 출처: Huh & Kwak(1997), *The Choice of Functional Form and Variables in the Hedonic Price Model in Seoul*에서 인용

이와 같이 면적과 가격과의 관계는 수요와 공급의 균형 여부, 지역 특성, 정부의 규제정책 등에 따라 체증, 체감 등 다양한 비선형 형태를 보일 수 있다.

나. 경과연수

건물 신축 시점으로부터의 경과연수도 부동산 가격과 비선형 관계를 보이는 대표적인 항목이라 할 수 있다. 우리나라 대도시의 경우 아파트나 단독주택 모두 신축 이후 처음 20년 정도는 시간이 흐를수록 그 가치는 하락하지만 20년을 초과하는 시점부터 재건축이나 재개발에 대한 기대감이 커져 오히려 가치가 증가하는 현상이 나타난다. 이는 해외의 경우도 마찬가지여서 Goodman & Thibodeau(1995)는 Dallas 소재 단독주택을 대상으로 경과연수에 따른 건물감가의 비선형 패턴을 실증적으로 밝힌 바 있다.

최근에는 분위회귀모형(Quantile Regression Model)을 이용하여 부동산 가격형성에 영향을 미친다고 생각되는 항목들(면적, 건물구조, 경과연수 등)의 영향력이 모든 가격수준에서 일정하지 않으며 가격 분위별로 상이하다는 점을 많은 연구에서 제시하고 있다(Zietz et al., 2008; Kostov, 2009; Farmer & Lipscomb, 2010; Liao & Wang, 2012). 국내 연구의 경우 임재만(2010)은 서울시 아파트 가격에서 경과연수는 주택가격 분포의 하위 50% 분위까지는 주택가격에 부(-)의 영향을 미쳤으나 60% 분위부터는 정(+)의 영향을 미치는 등 그 관계가 단순한 선형이 아님을 설명하였다.

부동산 평가실무에서도 경과연수에 따른 가치 하락의 다양한 패턴을 반영하여 가격을 추정하고 있다. 예를 들어 그 관계가 선형에 가까운 경우, 즉 해마다 일정액이 감가되면서 가격이 하락한다고 볼 수 있는 경우에는 정액법을 주된 감가상각기법으로 사용하지만, 신축 이후 초기에는 감가액이 매우 크게 발생하고 경과기간이 장기화됨에 따라 감가액이 작아지는 패턴일 경우에는 정률법[10]을 감가상각기법으로 활용하고 있다.

---

10) 매년 남아 있는 가치(잔존가치)에 일정률을 곱하여 감가액을 산정하는 방법. 따라서 시간이 경과할수록 감가액은 작아진다.

다. 부대시설

부대시설 또한 가격과의 선형 관계가 성립되기 어려운 항목 중의 하나이다. Gloudemans & Almy(2011)는 미국 단독주택의 경우 에어컨, 벽난로, 수영장 같은 부대시설의 존재는 모든 주택에 대하여 동일한 가치 할증을 가져오는 것이 아니라 주택의 규모, 등급 등에 따라 가치 할증 정도가 매우 상이하다고 보았다. 즉 작고 조잡한 구조의 주택에 수영장이 존재하는 것은 오히려 과잉 투자에 따른 감가요인으로 작용하며, 일정 규모 이상의 고급 주택에 수영장이 존재할 때 가장 큰 가치 할증을 가져온다는 것이다.

그들은 에어컨의 예를 들면서 북미의 과세평가에서 흔히 볼 수 있는, 다음과 같은 함수 형태를 제시하였다(Gloudemans & Almy, 2011, p.323).

$$\mathrm{P}rice = \$300\,Aircon + \$20\,Area + \$0.5\,Aircon \times Area^{1.2} \quad (2\text{-}3)$$

위 식에서 Price는 주택가격 총액, Aircon은 에어컨 존재 유무를 나타내는 더미변수, Area는 건물면적을 의미한다. 위 식을 보면 에어컨이 존재하는 주택의 경우 $300의 가치 증가뿐 아니라 건물면적의 대소에 따라 추가적인 가치증분이 있음을 알 수 있다. 예를 들어 면적이 100㎡인 주택은 에어컨이 존재함으로써 $300 이외에 $0.5 × $100^{1.2}$ = $126의 가치증분이 있는 반면, 200㎡ 주택은 $300 이외에 $0.5 × $200^{1.2}$ = $289의 가치증분이 있어 $126의 두 배 이상임을 알 수 있다.

즉 북미의 과세평가 실무에서는 부대시설과 가격과의 비선형 관계를 포착하기 위해 다차항(Polynomial Term)과 상호작용항(Interaction Term)을 효과적으로 활용하고 있는 것이다.

3. 비선형성을 반영하기 위한 비모수 모형

이와 같이 부동산 가격과 특성은 비선형의 정형화되지 않은 관계를 가질 수 있는데, 사전에 특정 함수 형태를 가정하지 않는 비모수 모형은 이러한 특징을 반영하는 대안이 될 수 있다.

비모수 모형은 모수 모형과 대를 이루어 설명하는 것이 일반적이다. 유한한 모수(母數, Parameter)를 동원하여 설명이 가능한 확률분포군(群)을 모수 모형이라 한다. 이때 이러한 모수들이 k개라면 k 차원의 모수 벡터(Parameter Vector) $\theta=(\theta_1,\theta_2,...,\theta_k)$로 모수 모형을 표현하게 된다. 예를 들어 포아송 확률모형은 1 차원 모수 $\theta=\lambda>0$에 의해 모형의 형태가 결정되며, 정규 확률모형은 2차원 모수 $\theta=(\mu,\sigma)$에 의해 모형의 형태가 정해진다. 반면 무한 차원의 모수 공간을 가지는 경우 그러한 모형을 비모수 모형이라 부른다.

비모수 모형이 모수 모형과 구별되는 가장 뚜렷한 차이점은 설명변수와 모수의 결합 형태에 대한 관점에 있다. 모수 모형의 경우, 결합 형태를 사전에 어떠한 방식으로든 정형화하며, 해석의 편의를 위해 선형으로 정형화하는 것이 일반적이다[11]. 그러나 경우에 따라 비선형인 것이 분명하고 그러한 비선형 함수의 형태가 이론적으로 명확하게 제시되었을 때에는 비선형으로 정형화하기도 한다. 예를 들어 다음과 같은 수목의 비선형 성장함수는 명확한 이론적 근거를 가지고 있으며, 실제 실무에서도 널리 활용되고 있다.

$$Y=\frac{\phi_1}{1+\exp\left[-\left(X-\phi_2\right)/\phi_3\right]} \tag{2-4}$$

위 식에서 Y는 수목의 높이, X는 경과시간을 나타낸다. ϕ_1는 성장 가능한 최대 높이를 의미하며, ϕ_2는 최대 높이의 1/2에 도달하는데 소요되는 시간, 그리고 ϕ_3은 최대 높이의 1/2에서 출발하여 3/4까지 도달하는데 소요되는 시간을 나타낸다(Pinheiro & Bates, 2006, p.274).

반면 비모수 모형은 설명변수와 모수의 결합 형태에 대해 특별한 가정을 하지 않으며, 오로지 데이터에 포함된 정보에 기초해 결합 형태를 찾으려 노력한다.

따라서 모수 모형은 극단적인 경우 사전에 설정한 설명변수와 모수와의 결합 형태, 즉 헤도닉 가격 함수의 형태가 데이터에 포함된 정보보다 더 큰 영향을 추정 결과에 미칠 수 있다. 즉, 가격 함수의 형태를 사전에

11) 본 연구에서는 2차, 3차 등의 다항변수, 자연로그 변환 등도 넓은 의미의 선형 관계에 포함시켰다.

결정할 때 사용된 연구자의 사전지식 등이 데이터에 포함된 정보보다 더 큰 역할을 할 가능성이 높다. 반면 비모수 모형은 가격함수의 형태를 특정하지 않으므로 추정 결과는 온전히 데이터에 포함된 정보에만 의존하게 된다.

이와 같이 비모수 모형은 모수 모형이 주로 가정하는 선형의 가격 함수 형태를 주장하지 않는다는 점에서 보다 유연한 접근법이라 할 수 있다. 모수 모형도 경우에 따라 비선형 함수를 사용하지만, 이는 그 형태가 이론적으로 정립되어 있는 경우에 국한되며, 물리학, 생물학 등 자연과학 분야에서 흔하게 관찰할 수 있으나, 사회과학 분야에서는 비선형의 모수 모형 적용 예를 찾아보기 힘들다.

[표 2-3]은 $y_i = f(x_i) + \epsilon_i$라는 간단한 형태의 모형을 기준으로 모수 및 비모수 모형을 비교한 것이다. 표에서 보듯 양 모형의 가장 큰 차이점은 가격함수 $f(x_i)$를 어떻게 정형화하는가에 있다(노희상 et al., 2014). 모수 모형이 정형화된 함수 형태, 그 중에서도 주로 선형의 함수 형태를 받아들이는 반면, 비모수 모형은 비선형의 정형화되지 않은 함수 형태를 받아들이고 있다.

[표 2-3] 모수 모형 및 비모수 모형의 비교

구 분		모수 모형	비모수 모형
확률분포함수		유한개의 모수로 결정되는 확률분포함수 가정	특정한 확률분포함수를 가정하지 않음
헤도닉 가격함수	오차 ϵ_i의 분포	통상 정규분포 가정	특별한 가정 없음
	가격함수 $f(x_i)$의 형태	설명변수와 모수의 결합형태를 정형화 · 1차 선형 · 다항 함수 · 멱함수 등	특정한 형태로 정형화하지 않되, 기저함수를 이용한 확장(Basis Function Expansion)이 대표적인 접근방식
장점		· 계산의 편리함 · 높은 이해가능성	· 이상치 등에 강건 · 범주형 설명변수가 많을 경우 유리
단점		· 엄격한 통계적 가정 · 자료 특성의 지나친 단순화	· 대량의 표본 필요 · 계산량 많음 · 이해가능성 떨어짐

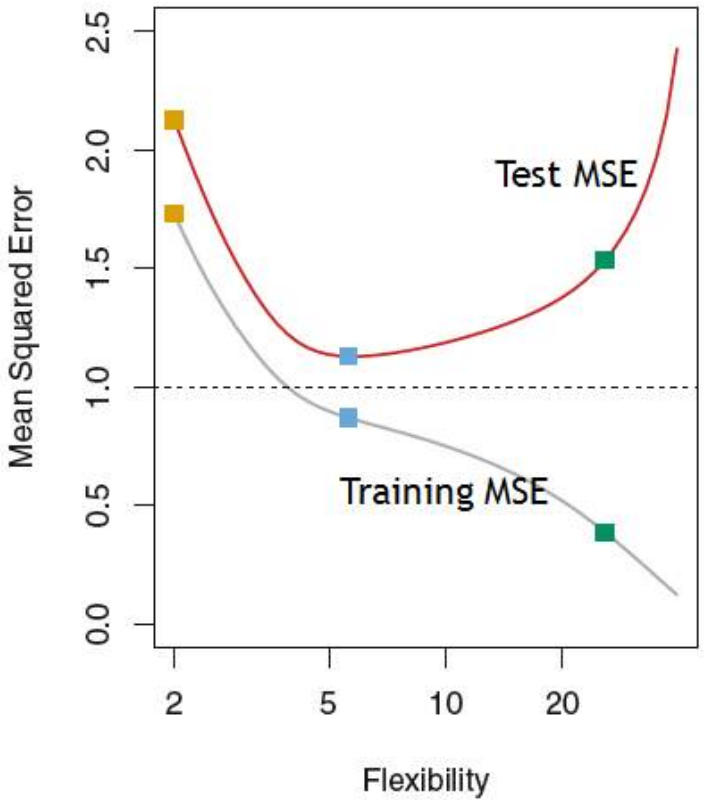
제 3 절 모형 성능의 진단 기준

신규 데이터를 잘 예측할 수 있는 최적의 모형을 선택하는 것은 가장 중요하고도 어려운 과업이다(James et al., 2013, p.29). 예측 중심의 모형에서 가장 일반적으로 쓰이는 모형 성능의 판단 지표는 다음과 같은 평균제곱오차(Mean Squared Error, MSE)라고 할 수 있다.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y_i}\right)^2 \quad (2\text{-}5)$$

위 식에서 y_i는 실제 관찰치, $\hat{y_i}$은 모형을 통해 산출된 추정치를 나타내며, MSE 값이 작을수록 해당 모형의 성능은 우수한 것으로 판단할 수 있다. MSE 계산시 모형 구축에 사용된 데이터(Training Data)를 활용한 경우 해당 MSE를 모형 MSE(Training MSE), 모형 구축에 사용되지 않은 새로운 데이터(Test Data)를 활용하여 계산한 경우 검증 MSE(Test MSE)라고 지칭한다. 모형 성능을 비교하는 관점에서 목표는 검증 MSE의 최소화에 있다.

모형 MSE와 검증 MSE는 상호 유사한 패턴을 보일 것 같지만 실제 그렇지 않다. 사용한 모형이 복잡해질수록 모형 MSE는 항상 감소하지만 검증 MSE는 U자 형태를 보인다. [그림 2-4]는 모형의 복잡성, 즉 Flexibility와 MSE의 관계를 보여 주는데, 검증 MSE의 경우 모형이 일정 수준 이상 복잡해지면 오히려 그 값이 증가하는 것을 알 수 있다. 따라서 검증 MSE 값이 최소인 모형을 선택하는 것이 곧 최적 모형을 선택하는 셈이 된다. 모형 MSE는 작지만 검증 MSE가 큰 경우 그 모형은 해당 데이터를 과다 적합(Over-Fitting)하고 있는 것으로 볼 수 있다.



[그림 2-4] 모형의 복잡성과 MSE와의 관계

* 출처: James et al., 2013, p.31

현실적으로 별도의 데이터를 확보하여 검증 MSE 등을 계산하기란 쉽지 않다. 따라서 본 연구에서는 데이터 중 일부를 처음부터 검증 데이터로 유보시키는 방법(Validation Dataset Approach)을 활용하였으며, 임의분할(random split)을 통해 데이터의 70%는 모형 데이터로, 나머지 30%는 검증 데이터로 사용하여 모형 성능을 비교한다.

가. 진단 지표

헤도닉 모형 성능의 진단은 언제나 중요한 이슈였으나(Moore, 2006) 합의된 진단 지표는 없는 것으로 보인다. 부동산 가격 추정과 관련된 선행연구를 살펴보면 AIC(Akaike Information Criterion) 및 BIC(Bayesian Information Criterion) 같은 지표를 활용한 연구(Militino et al., 2004), RMSE(Root Mean Squared Error)나 MAE(Mean Absolute Error) 또는

이와 유사한 지표를 사용한 연구(Nguyen & Cripps, 2001; Hoshino & Kuriyama, 2010; Wheeler et al., 2014), 그리고 COD(Coefficient Of Dispersion)를 활용한 연구(McCluskey et al., 2000; Pace et al., 2002; Moore, 2006) 등이 있다.

일치된 견해가 있는 것은 아니지만 부동산 과세평가(Assessment)를 위한 행정 목적의 헤도닉 모형 분야에서는 통상 세 가지 지표로 모형의 성능을 비교한다(Huang, 2002; IAAO, 2010; Gloudemans & Almy, 2011).

첫째는 추정가격(Estimated Price)이 얼마나 실제 가격을 가깝게 반영하는지 나타내는 정확성(Accuracy)으로, 평가수준(Appraisal Level) 또는 효율성(Efficiency)이라 칭하기도 한다. 추정가격의 정확성은 추정가격을 실제가격으로 나눈 현실화율(Sales Ratio, 이하 'SR')로 측정하는 것이 통상이며, IAAO(2013)는 모형을 통해 산출된 과세평가액의 평균 SR이 0.90 ~ 1.10 범위에 존재하여야 해당 모형을 수용할 수 있는 것으로 판단한다. 즉 평균 SR 1.00을 목표로 설정하고 있다.

둘째는 이러한 실제가격 대비 추정가격 비율, 즉 SR이 개별 부동산에 따라 큰 차이 없이 일관된 비율을 유지하는지 나타내는 형평성(Equity)으로, 평가 균일성(Appraisal Uniformity)이라 하기도 한다. 이러한 형평성을 측정하는 지표는 COD(Coefficient Of Dispersion), COV(Coefficient Of Variation) 등 여러 가지가 있으나 실무에서 흔히 활용하는 것은 COD이며 그 산식은 다음과 같다.

$$COD = \frac{\left[\frac{\Sigma|SR - median(SR)|}{n}\right]}{[median(SR)]} \times 100 \qquad (2\text{-}6)$$

과세평가 실무에서 COD가 널리 활용되는 이유는 위 식의 분모에서 알 수 있듯이 계산값을 최종적으로 SR의 중위수 값으로 나누어 주어 지역 간 또는 부동산 유형 간 통일된 비교가 가능하기 때문이다. 또한 미국이나 캐나다 등 북미에서 매우 오랫동안 사용되었으므로 지역 특성별(도시, 근교, 농촌 등), 부동산 유형별(주거용, 상업용, 공업용 등)로 벤치마크가 되는 COD 기준값이 대략적으로 정해져 있기 때문이다.

세부 지역별로 상이하지만 IAAO(2013)는 대체로 COD 25.0 이하인 경우 해당 모형을 수용할 수 있는 것으로 본다. 그러나 이러한 기준은 주로 미국이나 캐나다 등 북미의 지역 사정에 적합한 것으로 좁은 국토 공간에 이질적인 부동산이 밀집하여 소재하는 한국의 경우 상한선을 보다 상향조정하여야 할 것으로 예상된다.

마지막으로 헤도닉 모형의 성능 측정과 직접적인 연관성은 떨어지지만 과세정책상 중요한 의미를 갖는 지표가 역진성(Regressive) 지표이다. SR이 부동산 가격 수준에 따라 달라질 때, 특히 고가 부동산일수록 SR이 낮을 때 추정가격에 역진성이 존재하는 것으로 본다. 반대로 고가 부동산일수록 SR이 높아지면 추정가격에 누진성(Progressive)이 존재하는 것으로 본다. 민규식(1994), Kochin & Parks(1982), Clapp(1990), Cesare & Ruddock(1998) 등은 용어와 수식에 약간의 차이는 있지만 다음과 같은 식을 통해 가격의 역진성 여부를 파악하고자 하였다.

$$MV = \beta_0 (AP)^{\beta_1} e^{\epsilon} \tag{2-7}$$

위 식에서 MV는 시장가치(Market Value), AP는 모형을 통해 산출된 과세가격(Assessed Price)을 나타낸다. 양변에 자연로그를 취해 선형함수로 표현하면 다음과 같다.

$$\ln MV = \ln \beta_0 + \beta_1 \ln (AP) + \epsilon \tag{2-8}$$

위 식에서 $\beta_1 = 1.00$이면 추정가격에 아무런 편의가 없는 것으로, $\beta_1 > 1.00$이면 역진성, 흔하지는 않지만 $\beta_1 < 1.00$이면 누진성이 있음을 나타낸다.

본 연구에서는 추정가격의 정확성 지표로 SR의 평균 및 중위수를, 형평성 지표로 COD를 사용한다. 아울러 통계학 분야에서 모형의 적합 정도를 표현하는 대표적 지표인 RMSE[12]와 MAE를 추가적인 형평성 판단 지표로 사용한다. RMSE와 MAE의 수식은 다음과 같다.

12) 식(2-1)의 MSE에 제곱근을 취한 것으로 MAE와 비교가 용이하도록 하기 위해 선택하였다.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y_i}-y_i\right)^2} \quad (2\text{-}9)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y_i}-y_i\right| \quad (2\text{-}10)$$

요약하면 SR, COD, RMSE 및 MAE를 주요 진단지표로 활용하며, 모형 데이터보다는 검증 데이터에서 산출된 상기 지표들의 비교에 중점을 둔다. 다만 추정 가격의 과대평가나 과소평가 여부를 파악하고자 할 때에는 정확성 지표인 SR에 보다 많은 비중을 두고, 모형의 전반적인 적합도(Goodness Of Fit)를 가늠하고자 할 때에는 COD, RMSE 및 MAE 같은 균형성 지표를 중점적으로 사용하고자 한다.

나. 모형 성능의 국지적 진단

본 연구에서는 모형 성능을 진단할 수 있는 새로운 방법으로 현실화율(SR)을 활용한 국지적 접근방법(Local Approach)을 제시하고자 한다. 앞서 설명한 SR의 평균이나 중위수, COD 등은 모두 전역적 지표(Global Index)로서 모형의 전반적인 성능을 나타낸다. 반면 국지적 지표(Local Index)는 해당 모형이 어떤 지역에서(in geographic space), 또는 어떠한 유형에서(in feature space) 결함을 보이는지 지역별 또는 항목별 부분적인 성능을 나타낸다. 따라서 이러한 국지적 지표는 모형의 결점을 쉽게 파악함으로써 후속 모형 개선에 중요한 시사점을 제공할 것으로 기대된다. 모형 성능에 대한 이러한 국지적 접근방법의 제시는 기존 선행연구와 차별되는 점이기도 하다.

본 연구에서 제시한 국지적 진단은 다음과 같은 절차를 통해 모형의 미비점을 파악할 수 있다. 먼저, 실제가격 대비 모형을 통해 추정된 가격의 비율, 즉 SR을 모형 데이터 및 검증 데이터 모두에 대해 계산한다. 다음으로 SR을 종속변수로, 해당 모형에 동원된 공변량을 설명변수로 하여 회귀트리(Regression-tree) 알고리즘을 적용한다. 알고리즘은 모형 데이터 또는 검증 데이터 모두에 적용할 수 있다[13]. 마지막으로 회귀트

리 결과를 분석하여 모형의 개선점을 찾는다.

국지적 접근의 이점 중 하나는 국지 통계량(Local Statistic) 계산이 가능하므로 계산된 값을 공간상에 시각화할 수 있다는 점이다. 따라서 SR 값을 지도화하여 모형 성능이 미흡한 지역을 찾을 수 있다. 즉 SR 값을 살펴 가격을 일관되게 과소 또는 과대추정한 부분이 있는지 살피고, 그러한 부분이 존재하는 경우 모형의 세부 모수를 바꾸거나 모형 자체를 처음부터 다시 설계하는 등 필요한 조치를 취할 수 있다. 이러한 미비점이 나타나지 않았다면 모형 성능의 일관성이 있는 것으로 판단할 수 있다.

제 4 절 비모수 모형의 유형

함수가 비선형의 형태를 보일 때, 가장 단순한 조치는 변수를 변환하는 것이었다. 특히 설명변수 X를 변환하여 함수 관계를 직선화할 수 있다면 종속변수 Y의 분포 형태에 영향을 주지 않게 되므로 가장 손쉬운 해결 방법이 된다. 통상 변수값들의 산점도 패턴 등을 살핀 후 자연로그, 제곱근, 지수, 역수 등의 변환을 시도하게 된다. 이와 같은 접근을 보다 일반화하여 지수 변환 중 최적의 변환 형태를 찾아내는 것이 Box-Cox 변환이다(Box & Cox, 1964).

그러나 이 같은 변수 변환은 설명변수가 1개이거나 비교적 적을 때 적용하기 용이하며 설명변수의 개수가 많아질 경우 실제 활용이 어렵다. 또한 설명변수의 개수가 적다하더라도 단순한 변환만으로 실제의 복잡한 함수 형태를 포착하기 어려운 경우가 대부분이다. 본질적으로 가격함수는 경제학적 측면에서 다양한 소비자 기호와 생산자 기술을 나타내기 때문에 그 형태를 정확히 가늠하기 어렵다. 또한, 가격함수는 실제 측정하기 어려운 항목들(소음이나 공기오염의 정도와 같은 자연환경의 양부,

13) 그러나 모형의 미비점을 파악한다는 측면에서 검증 데이터에 적용하는 것이 보다 유용할 것으로 보이며, 본 연구에서도 검증 데이터를 기준으로 국지적 접근을 시도하였다.

인근지역에 대한 주민들의 애착심 등)과의 관계도 모두 포함하는 개념이므로 그 형태를 정확히 파악하는 것은 불가능에 가깝다(Mason & Quigley, 1996).

따라서 가격함수의 형태는 이론적 문제라기보다는 실증적 문제에 해당하며 가격함수 f를 온전히 데이터에 기반하여 찾으려는 비모수적 접근들이 여러 분야에서 시도되고 있다. 이하에서는 기계학습 분야에서 개발된 비모수 모형의 유형과 원리 등에 대해 순차적으로 설명한다.

1. 다항회귀(Polynomial Regression) 모형

선형회귀모형을 비선형 모형으로 확장하는 전통적 방법은 다항함수를 사용하는 것이었다. 즉, 단순 선형회귀모형의 경우 아래처럼 식 (2-11)의 선형회귀모형을 식 (2-12)의 다항함수로 대체하는 것이다.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{2-11}$$

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + ... + \beta_d x_i^d + \epsilon_i \tag{2-12}$$

이 때 다항함수의 차수 d는 2 이하를 사용하는 것이 통상이다. 3 또는 4 이상을 사용할 경우 데이터를 과다적합(Over-Fitting)할 가능성이 커지며, 추세선(Fitted Line) 자체도 매우 불안정한 모습을 보이게 된다.

다항회귀모형은 Cubbin(1974) 등 초기에 활발하게 활용되었으며, 90년대 들어서도 Goodman & Thibodeau(1997)가 건물 경과연수를 3차항(age, age^2, age^3)으로 구성하여 모형 성능의 개선을 보이는 등 꾸준하게 활용되고 있다. 최근에는 경과연수나 방 개수 등 전형적인 설명변수 뿐 아니라 지리좌표(x, y) 자체를 2차항, 3차항, 상호작용항 등 다항 형태로 구성한 후 일반적인 속성 변수와 함께 헤도닉 모형에 투입하는 사례(Fik et al., 2003)도 보고되고 있다.

이러한 다항회귀는 다음과 같은 두 가지 한계를 단점으로 지적할 수 있다. 첫째는, 현실세계에서 설명변수와 종속변수의 복잡한 관계를 차수

(Order)나 상호작용항 정도로 표현하기 어렵다는 것이다. 둘째는 설명변수 X의 함수형태에 대해 하나의 전역적인 구조(a single global structure)를 부과한다는 점이다. 예를 들어 2차항 함수 $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$를 선택한 경우, 이러한 2차식의 함수 관계가 X 값의 전 범위에 걸쳐 동일하게 작용한다고 보는 것이다. 그러나 현실에서는 X 값의 범위에 따라 상이한 함수 관계가 형성될 가능성이 더 높다[14].

이러한 다항회귀의 한계를 극복하기 위해 여러 가지 방법이 제시되었는데, 특히 한 개의 전역적 구조 대신 설명변수 X의 범위를 세분화하여 여러 개의 국지적 구조를 모델링하려는 유연한 방법이 다수 개발되었다. 이하에서 설명할 비모수 모형들은 모두 이러한 부류에 속하는 것들이다.

2. 일반가산모형(Generalized Additive Model, GAM)

통상적인 선형회귀모형을 아래와 같이 표현한다면,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i \tag{2-13}$$

일반가산모형(이하 'GAM')에서는 설명변수와 종속변수의 비선형 관계를 반영하기 위해 선형결합 $\beta_j x_{ij}$를 비선형 함수 $f_j(x_{ij})$로 대체한다. 따라서 GAM은 다음과 같이 표현할 수 있다.

$$\begin{aligned} y_i &= \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \ldots + f_p(x_{ip}) + \epsilon_i \\ &= \beta_0 + \sum_{j=1}^{p} f_j(x_{ij}) + \epsilon_i \end{aligned} \tag{2-14}$$

즉, 각각의 설명변수 x_j에 대하여 함수 f_j를 계산하고 이를 합산(Additive)한다. 함수 f_j는 다양한 방법으로 계산될 수 있는데, Natural Spline, Smoothing Spline, 국지회귀(Local Regression), 다항회귀

14) 예를 들어 [-∞, 0] 구간에서는 X를 2차식으로, [0, ∞] 구간에서는 X를 3차식으로 표현하는 것이 더 타당할 수 있다.

(Polynomial Regression) 등 여러 가지 방법을 동원할 수 있고, 또한 한 가지 방법에 국한하지 않고 서로 다른 방법을 설명변수별로 각각 사용할 수도 있다. 함수 f_j를 다양한 방법으로 추정할 수 있다는 것은 GAM의 유연성을 보여주는 장점이라고 할 수 있다.

모수 추정은 최소자승법으로 추정이 가능한 경우에는 최소자승법으로 수행할 수 있으나, 그렇지 않은 경우가 대부분이며 이때에는 Backfitting 알고리즘을 사용한다. 다음의 PRSS(Penalizing Residual Sum-of-Squares)를 최소화하는 것이 알고리즘의 목적이 된다(Hastie et al., 2009, p.297).

$$PRSS(\beta_0, f_1, f_2, \ldots, f_p) = \sum_{i=1}^{N}\left[y_i - \beta_0 - \sum_{j=1}^{p} f_j(x_{ij})\right]^2 + \sum_{j=1}^{p}\lambda_j \int f_j''(t_j)^2 dt_j \quad (2\text{-}15)$$

위 식에서 $\lambda_j \geq 0$는 동조 파라미터이며 통상 해(Solution)가 하나로 정해지지 않으므로 제약식 $\Sigma_1^N f_j(x_{ij}) = 0 \; \forall j$을 부과하는 것이 일반적이다.

GAM은 Pace(1998)가 주택 가격 추정에 적용하여 변수 변환이나 다항회귀모형보다 가격 예측력이 우수함을 보인 이후 지속적으로 부동산 가격추정에 사용되고 있다. Bao & Wan(2004)은 Smoothing Spline 방법을 사용하여 홍콩의 주택가격을 예측하였으며 기존의 Box-Cox 변환보다 모형의 가격 예측력이 전반적으로 개선되었음을 실증적으로 밝혔다. Karato et al.(2010)은 주택가격의 설명변수 중 경과연수와 코호트(Cohort)[15]에 초점을 두어 이 두 가지 특성이 주택가격 형성에 미치는 영향이 선형이 아님을 보였다.

GAM을 활용할 경우, 설명변수 각각에 대해 일일이 변수 변환을 시도하는 등 개별적으로 함수 f_j를 찾을 필요가 없다. 즉 모형 적합의 효율성이 높다. 또한 GAM은 본질적으로 가산(Additive) 형태의 함수이므로 다른 설명변수는 고정되었다고 가정하고 관심의 대상이 되는 설명변수 X_j의 효과만을 분석하는 것이 가능하며 논리적으로도 무리가 없다. 즉 예측이 아닌 종속변수와 설명변수 간의 인과관계 추론에 주안점이 있는

15) 경과연수는 주택의 물리적 노후화 정도를, 코호트는 주택 신축 당시의 건축공법이나 스타일을 나타낸다.

경우에도 GAM은 유용한 분석 틀이 될 수 있다.

이와 같은 장점에도 불구하고 GAM은 가산 모형이라는 한계가 있어 랜덤 포리스트(Random Forest) 같은 완전한 비선형 또는 비모수 모형에 속한다고 볼 수 없다. 따라서 GAM은 모수 모형과 완전한 비모수 모형의 중간 정도에 해당된다고 볼 수 있다(James et al., 2013, p.286).

3. 트리기반 모형(Tree-based Methods)

트리기반 모형은 최근에 개발된 것으로 인식되고 있으나 Morgan & Sonquist(1963)까지 거슬러 올라가며, 보다 현대 통계학적인 접근은 Breiman et al.(1984) 및 Quinlan(1993)에서 찾아볼 수 있다. 트리기반 모형을 비롯한, 이하에서 설명할 MARS(Multivariate Adaptive Regression Splines), SVM(Support Vector Machine) 등은 모두 기계학습 분야에서 제시된 방법론이다. 이들 방법론은 전통적인 선형회귀모형과 달리, 비모수 모형에 해당하며 종속변수와 설명변수 간의 비선형 관계나 상호작용 효과를 사전에 설정할 필요 없이 데이터로부터 학습해 나가는 알고리즘이라 할 수 있다(Grömping, 2009).

종속변수가 연속형일 경우에 적용되는 트리기반 모형을 회귀트리 모형(Regression Tree Model)이라고 한다. 회귀트리 모형은 설명변수 공간(Predictor Space 또는 Feature Space)을 분할하는 것으로부터 시작한다. 즉, 설명변수 $X_1, X_2, ..., X_p$를 J개의 지역(Region) $R_1, R_2, ..., R_J$로 서로 겹치지 않게 분할한다. 다음으로 R_j 지역에 속하는 관찰치에 대해서는 R_j 지역 관찰치 평균값을 예측치로 제시하게 된다. R_j 지역은 다음과 같은 잔차제곱합(Residual Sum of Squares, RSS)이 최소가 되도록 분할한다(이하 식 2-16에서 2-19는 James et al.(2013) pp.306-309에서 인용).

$$\sum_{j=1}^{J} \sum_{i \in R_j} \left(y_i - \widehat{y_{R_j}} \right)^2 \tag{2-16}$$

다시 말해, 설명변수 j 및 임계치 s에 대하여 다음과 같은 두 개의

지역으로 구분한 후,

$$R_1(j,s)=\{X|X_j<s\} \text{ and } R_2(j,s)=\{X|X_j\geq s\} \quad (2\text{-}17)$$

다음 식을 최소화하는 j 및 s를 탐색하게 된다.

$$\sum_{i:x_i\in R_1(j,s)}\left(y_i-\widehat{y_{R_1}}\right)^2+\sum_{i:x_i\in R_2(j,s)}\left(y_i-\widehat{y_{R_2}}\right)^2 \quad (2\text{-}18)$$

RSS 값을 최소화하는 기준으로 트리를 구성할 경우 언제나 과다적합할 가능성이 높아진다. 이와 같은 문제점을 해결하기 위해 통상 트리를 최대한 키워 놓고, 해당 트리의 가지를 쳐가면서('*Pruning Tree*') 적정 규모의 트리를 결정하게 된다. 트리의 규모를 줄이는 기준은 다음 식을 최소화하는 것이다.

$$\sum_{m=1}^{|T|}\sum_{x_i\in R_m}\left(y_i-\widehat{y_{R_m}}\right)^2+\alpha|T| \quad (2\text{-}19)$$

위 식에서 $|T|$는 트리 T의 가지(Terminal Node) 수를, R_m은 m번째 가지에 해당하는 분할지역을 의미한다. α는 동조 파라미터(Tuning Parameter)로서 $\alpha=0$인 경우 아무런 패널티가 없으므로 최대 트리가 되며, α가 커질수록 트리 규모는 작아진다. α값은 교차 타당성 검증(Cross-Validation)이나 검증 자료(Test Data)의 적합 등을 통해 정한다.

트리기반 모형은 개념이 단순하고 시각적으로 표현하기 수월하며, 따라서 해석하기도 용이하다. 또한 비전문가에게도 매우 쉽게 설명할 수 있다. 반면 다른 비선형 모형(MARS 등)에 비해 추정가격의 정확성이 떨어지는 경우가 많다. 그러나 하나의 트리가 아닌 수백, 수천개의 트리 결과를 종합하는 앙상블 접근(Ensemble Approach)을 취할 경우 현격한 모형 성능의 개선을 가져오기도 한다. 여러 개 트리의 결과를 종합하는 앙상블 접근에는 *Bagging, Random Forest, Boosting* 등이 있다.

이러한 앙상블 접근은 트리 기반 모형이 아닌 다른 모형에 대해서도 적용할 수 있으나 트리 기반 모형에 특히 자주 활용된다. 그 이유는 앙상블 접근은 개별 모형들이 편의는 작으나 분산이 큰 경우 가장 큰 성능의 개선을 가져올 수 있기 때문이다(Kuhn & Johnson, 2013, p.390). 트

리 기반 모형은 편의가 작고 분산이 큰 대표적인 모형이라 할 수 있다. 따라서 다수의 트리들을 결합함으로써 편의도 작으면서 분산도 작은 모형을 만들 수 있다.

하나의 단일 트리에 기초한 예측 결과는 안정성이 떨어진다. 분석 데이터에 대해 다시 한번 트리 모형을 적용할 경우, 그 결과는 직전의 결과와 매우 상이할 수 있다. 즉 예측치의 분산이 높은 편이다. 이러한 문제를 해결하기 위해 데이터로부터 일부 데이터를 복원 추출하여, 즉 부트스트랩(Bootstrap)을 통해 B개의 데이터 집합(Dataset)을 확보하고, B개의 회귀트리 결과를 각각 계산한 후, 마지막으로 이를 평균하여 최종 예측치를 정할 수 있다. 이러한 방법을 *Bagging*이라고 하며, 아래와 같은 수식으로 표현할 수 있다(James et al., 2013, p.316).

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}_b(x) \qquad (2\text{-}20)$$

반면, *Random Forest*는 *Bagging*과 유사하지만 회귀트리 모형 구성 시 데이터에 포함된 p개의 설명변수 모두를 활용하지 않고 일부만 활용하는 차이점이 있다. 즉 *Bagging*은 p개의 설명변수 모두를 동원하여 개별 트리를 구성하는데 반해, *Random Forest*는 m개(m<p)의 설명변수를 동원하여 개별 트리를 구성한다. 존재하는 설명변수 p보다 적은 설명변수를 사용하는 것이 얼핏 보기에는 상식에 반하는 것으로 보이나, p보다 적은 m개의 설명변수를 사용함으로써(통상 $m \approx \sqrt{p}$) 개별 트리 간의 상관성(Correlation)을 제거하거나 완화할 수 있는 이점이 있다. 따라서 *Random Forest*를 통해 B개의 회귀트리 모형 결과를 평균하게 되면, 예측치의 분산은 *Bagging*보다 줄어들게 된다.

마지막으로 *Boosting*은 *Bagging* 및 *Random Forest*와 달리 부트스트랩에 기초한 여러 개의 독립적 개별 트리를 만들지 않는다. 대신 최초의 원데이터를 계속하여 수정하면서 트리를 연속적으로 키워 나간다. 즉 최초 트리를 구성한 후, 이후 종속변수 Y가 아닌 잔차를 업데이트하는 방식으로 트리를 수정하면서 최종 트리 모형에 이르게 된다. 따라서 이 방법은 이전 단계에서 구성된 트리 모양에 많은 영향을 받게 된다. 이러한 *Boosting*을 실행하는 알고리즘은 다양하며 본 연구에서는 Gradient

Boosting Machine 알고리즘(Friedman, 2001)을 활용하여 실증분석을 하였는 바, 동 알고리즘을 간략히 설명하면 다음과 같다(이하 수식의 표현은 Friedman(2001)을 따랐다).

먼저 상수항만으로 구성된 초기 모델을 다음과 같이 구현한다.

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma) \tag{2-21}$$

위 식에서 x는 설명변수, y는 종속변수를 나타내며 $L(y, F(x))$는 미분가능한 손실함수(Loss Function)를 의미한다. 다음으로 아래와 같이 유사 잔차(Pseudo-Residuals)를 M(1, 2, ..., M)번 반복하여 계산한다.

$$\gamma_{im} = - \left[\frac{\delta L(y_i, F(x_i))}{\delta F(x_i)} \right]_{F(x) = F_{m-1}(x)} \tag{2-22}$$

위와 같이 계산된 유사잔차에 대해 기본 학습자(Base Learner) $h_m(x)$를 적합한 후, 다음과 같은 1차 최적화 방정식을 풀어 γ_m을 계산한다.

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)) \tag{2-23}$$

마지막으로 모형을 업데이트하게 되며, 이러한 과정은 M번 반복된다.

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \tag{2-24}$$

트리 모형은 부동산 가격 추정 분야에서 활발하게 사용되지는 않았다. Feldman & Gross(2005)가 최초로 트리 모형을 이용, 부동산 담보대출에 대한 채무 불이행율을 예측하였고, 이후 Fan et al.(2006)이 단일 트리 모형을 이용하여 싱가폴 주택가격을 추정한 바 있다.

단일 트리 모형은 앞서 설명하였듯 예측 결과의 안정성이 떨어질 뿐 아니라 다른 비선형 모형에 비해서도 예측 정확성이 미흡한 것으로 알려져 있다. 따라서 여러 개 트리를 결합한 앙상블 접근의 추정 결과가 보다 의미가 있을 것이다. Kagie & Wezel(2007)이 *Boosting*을 이용하여 네덜란드 주택가격을 추정하였고, 가장 최근에는 Lasota et al.(2011)이

Random Forest 모형을 사용하여 폴란드 주택가격을 예측한 바 있다. 상기 연구 모두 선형회귀모형을 기본 모형(Null Model)으로 하여 트리 기반 모형의 성능이 상대적으로 우수함을 밝혔다.

4. MARS(Multivariate Adaptive Regression Splines)

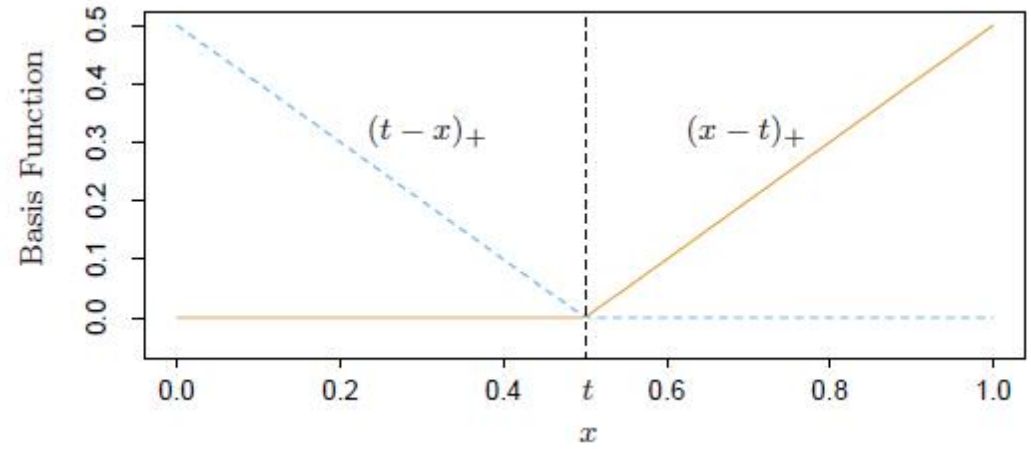
앞서 설명한 트리 기반 모형의 가장 뚜렷한 특징은 이분화 트리(Binary Tree)에 있다. 즉 데이터를 분할할 때 한번에 3개 또는 그 이상의 하위 그룹으로 분할하지 않고 정확히 2개의 하위 그룹으로만 분할한다는 점이다. 이러한 특징은 성급하게 최종 분할 결과에 도달하지 않게 하는 등 장점으로 작용하는 동시에 단점으로 작용하기도 한다. 대표적인 단점이 바로 함수 형태를 시각화하게 되면 불연속적인 형태가 된다는 점이다(Hastie et al., 2009, p.312). 즉 분할 지점(Splitting Point)에서 함수가 매끄럽게 이어지는 것이 아니라 불연속면이 발생하게 된다. 이러한 현상은 종속변수가 이진변수일 경우, 즉 분류(Classification)의 문제일 경우 큰 문제는 아니지만, 종속변수가 연속형일 경우에는 심각한 편의가 발생할 수 있다. 왜냐하면 종속변수가 연속형이라면 매끄러운 형태의 함수를 가정하는 것이 보다 자연스럽기 때문이다. 이러한 측면에서 MARS는 매끄러운 함수 형태를 반영할 수 있도록 트리 기반 모형을 수정한 방법으로 볼 수도 있다.

MARS의 기본적인 형태는 다음과 같다(Friedman, 1991).

$$\hat{f}(x)=\sum_{i=1}^{k} c_i B_i(x) \tag{2-25}$$

c_i는 계수, $B_i(x)$는 기저함수(Basis Function)를 나타내며, $B_i(x)$는 일종의 경첩함수(Hinge Function)로서 $\max(0, x-t)$ 또는 $\max(0, t-x)$로 표현된다[16]. 경첩함수는 MARS의 핵심 요소로서 예를 들어 $t=0.5$인 경우 그 형태는 [그림 2-5]와 같다[17].

16) t는 상수이며 'knot'라고 부른다.
17) $\max(0, x-t)$를 $(x-t)_+$로 표현하기도 한다.



[그림 2-5] $t=0.5$일 때 베이시스 함수의 형태

* 출처: Hastie et al.(2009) p.322에서 인용

MARS는 상수항, 즉 종속변수 값의 평균만으로 구성된 간단한 모형에서부터 출발하여 잔차제곱합(RSS)이 최소가 되도록 기저함수를 추가해 나간다. 기저함수를 추가할 때는 이미 모형에 포함된 설명변수, 앞으로 포함시킬 설명변수와 그 knot 값 등을 모두 고려하게 된다. 이러한 과정을 거쳐 완성된 MARS 모형은 통상 과다적합된 결과를 가져오게 되며, 따라서 과다적합된 모형의 규모를 줄여 나가야 하는데('Pruning') 그 기준은 다음과 같은 GCV(Generalized Cross-Validation) 값을 최소화하는 것이다(Hastie et al., 2009, p.325).

$$GCV(\lambda)=\frac{\sum_{i=1}^{N}\left[y_i-\hat{f}_{\lambda}(x_i)\right]^2}{[1-M(\lambda)/N]^2} \qquad (2\text{-}26)$$

위 식에서 분자는 잔차제곱합(RSS)을 의미하며, 분모의 $M(\lambda)$는 모형에 포함된 모수의 유효 개수(Effective Number of Parameters)를 나타낸다.

MARS는 변수 간 상호작용 효과를 포착하는데 적합한 특징을 가지고 있다[18]. 또한 트리 기반 모형이 범주형 설명변수를 비교적 잘 다룰 수 있다면, MARS는 연속형 설명변수를 처리하는데 보다 효과적이다. 경첩함수가 연속형 변수를 다루는데 적합한 성질을 가지고 있기 때문이다.

18) 모형에 포함시킬 상호작용항은 사용자가 그 최대값을 미리 정해야 한다.

이와 같은 MARS는 신용등급의 추정(Lee et al., 2006), 파산 확률의 예측(De Andrés et al., 2011) 등 일부 사회과학 분야에서 적용된 사례는 있으나 부동산 가격 추정 분야에서 활용된 예는 없는 것으로 보인다.

5. SVM(Support Vector Machines)

SVM은 이미지 분류나 패턴 인식처럼 분류(Classification)의 문제를 처리하기 위해 1990년대에 제시된(Vapnik, 1996) 기계학습 알고리즘 중의 하나이다. 분류의 문제가 아닌 연속형 종속변수의 예측에 적용할 경우 별도로 Support Vector Regression(SVR)으로 칭하기도 한다. 연속형 종속변수에 SVM을 적용하는 경우에도 이진 종속변수에 SVM을 적용하는 경우와 그 개념이나 논리는 동일하다(Hyper-Plane, Maximal Margin 등)[19].

선형회귀모형 $f(x)=x^T\beta+\beta_0$에서 β를 추정하기 위해 SVM은 다음의 식 (2-27)을 최소화하며, 이때의 V는 식 (2-28)과 같이 정의된다(Hastie et al., 2009, p.434).

$$H(\beta,\beta_0)=\sum_{i=1}^{N} V(y_i-f(x_i))+\frac{\lambda}{2}\|\beta\|^2 \qquad (2\text{-}27)$$

$$V_\epsilon(r)=\begin{cases}0 & |r|<\epsilon \\ |r|-\epsilon & o/w\end{cases} \qquad (2\text{-}28)$$

즉, ϵ보다 작은 오차는 무시한다는 의미로서('ϵ-insensitive error'), 분류의 문제를 다루는 SVM의 논리와 유사한 측면이 있다. 다시 말해 판별 경계선(Decision Boundary)을 기준으로 올바르게 분류된 데이터나 경계선으로부터 멀리 떨어진 데이터는 최적화 계산과정에서 무시되는 것과 마찬가지로 연속형 종속변수를 다루는 SVM에서도 오차가 작은 데이터는 계산과정에서 무시된다. λ는 일반적인 패널티 또는 동조 파라미터로

19) 자세한 사항은 James et al.(2013) Chapter 9, Hastie et al.(2009) Chapter 12 등 참조.

서 교차 검증(Cross-Validation)통해 산출된다.

식 (2-27)의 함수 H를 최소화하는 값 $\hat{\beta}$와 함수 $\hat{f}(x)$는 다음과 같다 (Hastie et al., 2009, p.435).

$$\hat{\beta}=\sum_{i=1}^{N}\left(\hat{\alpha}_i^*-\hat{\alpha}_i\right)x_i \tag{2-29}$$

$$\hat{f}(x)=\sum_{i=1}^{N}\left(\hat{\alpha}_i^*-\hat{\alpha}_i\right)\langle x,x_i\rangle+\hat{\beta}_0 \tag{2-30}$$

위 식에서 $\hat{\alpha}_i^*,\hat{\alpha}_i$은 다음과 같은 비선형 계획법(2차 계획법, Quadratic Programming)으로 해를 찾는다(Burges, 1998).

$$\min_{\alpha_i,\alpha_i^*}\epsilon\sum_{i=1}^{N}\left(\alpha_i^*+\alpha_i\right)-\sum_{i=1}^{N}y_i\left(\alpha_i^*-\alpha_i\right)+\frac{1}{2}\sum_{i,i'=1}^{N}\left(\alpha_i^*-\alpha_i\right)\left(\alpha_{i'}^*-\alpha_{i'}\right)\langle x_i,x_{i'}\rangle$$

$$\begin{aligned}\text{subject to }\ & 0\le\alpha_i,\alpha_i^*\le 1/\lambda,\\ & \sum_{i=1}^{N}\left(\alpha_i^*-\alpha_i\right)=0,\\ & \alpha_i\alpha_i^*=0\end{aligned} \tag{2-31}$$

위와 같은 제약식으로 인해 $\left(\hat{\alpha}_i^*-\hat{\alpha}_i\right)$ 중 일부만이 0이 아닌 값을 갖게 되며, 이때의 관찰치들을 서포트 벡터(Support Vector)라고 칭한다.

위 식의 해(Solution)는 내적(Inner Product) $\langle x,x_i\rangle$을 통해 계산된 설명변수 값에 따라 달라진다. 통상 선형뿐만 아니라 비선형적 관계를 포착하기 위해 내적을 좀더 발전시킨 커널(Kernel) 함수를 활용하는데, 흔히 쓰이는 커널의 형태는 다음과 같다(Basak & Patranabis, 2007; Kavousi-Fard, 2014 등).

$$\text{Polynomial: }\ k(x,x')=(1+\langle x,x'\rangle)^d \tag{2-32}$$

$$\text{Radial Basis: }k(x,x')=\exp\left(-\gamma\parallel x-x'\parallel^2\right) \tag{2-33}$$

$$\text{Neural Network: }\ k(x,x')=\tanh\left(\kappa_1\langle x,x'\rangle+\kappa_2\right) \tag{2-34}$$

SVM은 블랙박스 유형에 속하는 기계학습 알고리즘으로 선형회귀모형, Naive Bayes 등과 달리 모든 관찰치가 최적화 과정에 영향을 주지는 않는다. 오직 판별 경계선에 가까운 관찰치, 즉 예측하기 어려운 관찰치만이 최적화 과정에 영향을 미친다는 특징이 있다. 또한 트리기반 모형 등 다른 비선형 모형의 경우 사용한 소프트웨어나 데이터 분할 방식에 따라 그 결과 값이 불안정하기도 한데, 이는 모형이 전역적 최소값(Global Minimum)이 아닌 국지적 최소값(Local Minima)에 수렴할 가능성이 있음을 의미한다. 반면 SVM은 이러한 문제가 비교적 덜한 것으로 알려져 있다(Byvatov et al., 2003).

SVM이 최초로 개발·제시된 패턴 인식 분야 외에 응용 사례를 마케팅(Cui & Curry, 2005)이나 임상의학(Guyon et al., 2002) 분야에서 쉽게 찾아볼 수 있다. 하지만 부동산 분야에서는 토지피복(土地被覆)의 추정(Walton, 2008) 외에는 찾아보기 힘든 편이다. 이는 SVM이 비교적 최근(1990년대 이후)에 제시되었고 활용의 초점이 범주형 종속변수에 집중되었기 때문인 것으로 풀이된다.

제 3 장 비모수 모형의 적용 및 모형 성능 진단

이번 장에서는 앞서 설명한 모수 모형의 한계를 극복하기 위해 비모수 모형을 적용하여 주택가격을 추정하고자 한다. 즉 선형회귀모형을 기본 모형(Null Model)으로 하여 선형모형에서 비선형 모형으로 진전되는 중간 단계의 GAM을 적용하여 모형의 개선 정도를 살핀다. 이후 완전한 형태의 비선형 모형인 트리 기반 모형, MARS 및 SVM을 순차적으로 적용하여 모형 성능의 개선 정도를 검토한다. 다만 트리 기반 모형 중에서 하나의 트리에 기초한 회귀트리 모형은 결과값의 불안정성으로 인해 적용하지 않으며, Bagging 또한 Random Forest와 중복되는 측면이 있어 적용하지 않는다. 따라서 최종적으로 GAM, Random Forest, Boosting, MARS 및 SVM을 적용하여 그 결과를 선형회귀모형과 비교한 후 시사점을 도출하고자 한다.

제 1 절 실거래가 자료의 정제

1. 사례지역의 선정

비모수 모형을 적용할 사례지역으로 도시 규모별로 대도시, 중소도시 및 군 지역에서 각각 하나의 지역을 선정하였다. 대도시와 군 지역은 주택의 가격수준이나 가격형성요인 등의 측면에서 뚜렷이 구분되는 패턴을 보일 것이라 판단하였기 때문이다.

[표 3-1]은 전국의 평균 주택공시가격 수준을 100으로 보았을 때, 각 시도의 가격지수를 보여준다. 2014년 기준 서울은 전국 평균의 4배가 넘는 483인데 비해 전라남도는 전국 평균의 1/4 수준인 25에 불과함을 알 수 있다. 건물을 신축하는데 소요되는 건축비용은 지역에 관계없이 어느

정도 일정하다고 본다면, 이와 같은 주택공시가격 수준의 차이는 대부분 토지가격에서 기인하는 것으로 추론할 수 있다. 따라서 서울과 같은 대도시는 건물가격에 비해 상대적으로 토지가격 비중이, 전라남도와 같은 농어촌지역은 토지가격에 비해 상대적으로 건물가격 비중이 높을 것이다. 이에 따라 주택의 실제 거래가격 형성에 있어서도 대도시에서는 토지와 관련된 특성이, 농어촌지역에서는 건물과 관련된 특성이 보다 큰 영향력을 미칠 것으로 판단된다. 본 연구에서는 이러한 예상에 따라 대도시, 중소도시 및 군 지역에서 총 3개의 사례지역을 선정하였다([표 3-2]).

[표 3-1] 시도별 주택공시가격 지수(2014년 기준)

전국	서울	부산	대구	인천	광주	대전	울산	세종
100	483	142	266	209	114	193	205	124
경기	강원	충북	충남	전북	전남	경북	경남	제주
256	72	75	63	42	25	49	76	97

* 출처: 단독주택 공시가격 통계 e-book에서 인용(p.20)

[표 3-2] 사례지역 현황

시군구	면적(㎢)	인구(만명)	인구밀도(명/㎢)	1인당 지방세 부담액(원/명)
서울 강남구	39.54	57.7	14,593	443,000
전주 덕진구	111.23	28.8	2,589	370,000
전남 해남군	907.24	7.7	85	220,000

* 출처: 2014년 안전행정 통계 연보에서 발췌

[표 3-2]에서 서울 강남구는 우리나라의 대표적인 고급 주택지대(신사동, 청담동, 압구정동 등)가 위치한 지역이며, 전남 해남군은 한반도 최남단에 위치한 전형적인 농어촌 지역이라 할 수 있다. 전주 덕진구는 지리적 위치 및 가격수준에서 이들 두 지역의 중간 정도에 해당되는 도시로서 중심부의 도심지역과 외곽의 비도시지역으로 이루어진 대표적인 도농복합지역이라 할 수 있다.

본 연구에서는 상기 지역에 소재한 단독주택을 대상으로 다양한 비모수 모형을 적용하고 그 결과를 해석하고자 한다.

2. 자료의 성격 및 한계

부동산의 시장가치(Market Value)는 실제로 관찰할 수 있는 변량이 아니다. 따라서 차선책으로 시장가치를 가장 잘 반영할 수 있는 대리변수를 사용하여야 하는데, 대표적인 것이 실거래가 신고가격, 전문가가 추정한 감정평가가격, 부동산 중개업체의 시세(호가), 법원에서의 경매가격, 정부가 발표하는 공시지가 등이다. 실거래가 신고제도가 도입된 2006년 이전에는 주로 부동산 중개업체의 시세 자료 등을 활용하여 부동산 가격을 분석한 사례가 많으며, 2006년 이후부터는 실거래가 신고자료에 기초한 연구가 증가하는 추세이다.

본 연구에서 활용한 자료는 2011년부터 2014년까지의 실거래가 신고자료이다. 실거래가 신고제도는 부동산 거래시 실제의 거래가격을 확보하여 공평과세를 실현하고 부동산 투기를 미연에 방지하기 위하여 2006년 1월 도입된 것으로, 거래당사자 또는 부동산 중개업자에게 실거래 가격 신고를 의무화하고 있다.

실거래가 신고가격은 부동산 중개업체가 제공하는 시세나 정부의 주택공시가격보다 실제의 시장가치에 근접할 것으로 통상 인식되고 있다. 이와 같은 인식이 틀리다고는 할 수 없으나, 실거래가 신고가격에는 상당히 많은 이상치(Outlier)와 잡음(Noise)이 포함되어 있다. 따라서 분석에 사용하기 전 충분한 시간과 노력을 투입하여 이러한 이상치와 잡음을 제거할 필요가 있다.

본 연구에서처럼 정제되지 않은 데이터를 사용하려는 경우 모형의 정교한 구축보다 데이터 전처리(Pre-processing) 과정이 가격 예측력 제고를 위해 더욱 중요하다(Han & Kamber, 2006). 실거래가에 기초한 과세평가 업무가 일반화된 미국・캐나다 등의 업무 성과 보고서를 보면 과세평가 업무에 투입된 자원(Resources)의 80% 이상을 실거래가 자료 정제에 할당하고 있으며, 나머지 20%를 헤도닉 모형의 설계 및 검증에 배분하고 있다(Gloudemans & Almy, 2011).

실거래가 신고가격에서 이상치란 신고과정에서 신고자의 입력 오류[20],

계산과정에서의 착오[21], 양도소득세를 줄이기 위한 의도적 과소신고 등 여러 가지 이유로 발생할 수 있다.

반면 잡음은 당사자 사이에 합의된 실제의 거래가격을 오류나 허위 없이 그대로 신고하였으나 거래 상황, 거래 당사자의 입장 등으로 미루어 보아 정상적인 거래로 보기 어려운 가격을 말한다[22].

[표 3-3]은 실거래가 신고자료에서 '적정한 거래'로 보기 어려운 사례들을 일부 제시한 것이다. 이러한 사례들은 분석에 앞서 제거하거나, 시장가치를 제대로 반영할 수 있도록 합리적인 방법으로 보정한 후 분석에 포함시켜야 할 것이다.

[표 3-3] 적정한 거래로 보기 어려운 사례들

구 분	내 용
거래당사자	· 인접 부동산 소유자간 거래 · 정부 또는 공공기관이 관계된 거래 · 자선단체 또는 종교단체 등이 관계된 거래 · 관계회사(자회사 등) 간 거래
물적 특성	· 최소 분할 토지면적 이하의 거래 · 건폐율, 용적률 등이 상식적인 범위를 넘어서는 경우 (용적률 1,000% 등)
거래조건	· 공유지분의 거래 · 특약사항(건축허가의 확보 등)이 기재된 거래 · 교환 거래 · 다수 필지 거래 · 임차조건부 매도(Leaseback)
거래금액	· 비전형적인 대출조건이 수반된 거래 · 동산항목(시설물 등) 등이 거래금액에 포함된 경우 · 영업권 등 무형자산이 거래금액에 포함된 경우 · 공시가격 대비 저가 및 고가거래로 의심되는 경우

20) 현행 실거래가 신고 시스템은 수기 입력 방식으로 '3억원'에 거래된 부동산을 '3원'으로 잘못 입력하는 경우가 흔히 발생한다.
21) 예를 들어 거래금액 3억원 중 1억원을 금융기관에서 대출받은 경우, 착오로 3억원이 아닌 자신의 현금 지급액 2억원을 기재하는 경우도 흔히 발생한다.
22) 예를 들어 채권자의 강제경매 집행을 피하기 위한 급매, 부자(父子) 간의 거래, 건축허가를 득하는 조건으로 매매된 임야의 거래 등을 들 수 있다.

따라서 실거래가 신고가격의 이러한 성격과 한계를 정확히 파악하고 있어야 추후 분석 결과에 대한 해석에서도 올바른 추론을 할 수 있을 것이다.

3. 적정 실거래가 자료의 선별

본 연구에서는 실거래가 신고자료의 이상치와 잡음을 최대한 제거하기 위해 다음과 같은 절차를 거쳐 자료를 정제하였다[23]. 먼저 실거래가 신고가격과 신고된 연도의 주택공시가격과의 관계를 다음과 같이 설정하였다.

$$\text{신고가격}_i = \beta_0 + \beta_1 \times \text{공시가격}_i + \epsilon_i \quad (3\text{-}1)$$

즉, 실거래가 신고가격은 개략적으로 공시가격과 일정한 관계를 갖는 것으로 설정하였다. 위 식에서 i번째 주택의 신고가격과 공시가격은 다른 모든 조건은 동일하면서도(용도지역, 토지면적, 건물 연면적, 건물구조 등) 가격만 상이한 상태이다. 즉 주택가격 형성에 영향을 미칠 수 있는 대부분의 공변량들이 통제된 상태라고 볼 수 있으며, 가격만이 왼쪽은 거래 당사자 사이에서 자유로이 합의된 거래가격, 오른쪽은 과세표준 등으로 활용하기 위한 정부의 평가가격인 것이다. 이러한 가격 차이는 β_1에 의해 개략적으로 조정되며, 따라서 오차항 ϵ_i는 여러 가지 주택가격 형성요인(공변량) 및 실거래가 가격과 공시가격 간에 존재하는 일반적인 격차율 등을 모두 감안한 이후에도 남아 있는 가격의 변동량을 의미한다.

따라서 위 식을 실거래가 자료에 적용한 후, 잔차를 살펴 잔차가 지나치게 크거나 작은 경우에는 이상치 내지 잡음이 많이 포함된 신고 건으로 보아 자료에서 제거하였다. '지나치게 크거나 작은' 기준은 잔차 분포의 상하위 5%를 기준으로 하였다[24].

---

23) 외국의 경우 실거래가 신고자료의 정제는 본 연구에서와 같은 통계적 선별정도에 그치지 않는다. 필요한 경우 전문가가 직접 현장답사와 거래당사자 면담을 실시하기도 한다.

24) 상하위 5%에 대한 이론적 근거는 없으나 IQR(Inter-Quartile Range) 기준으로 부동산 가격 자료를 제거할 경우 자료 중심치에 근접한 관찰치에 대해서도 이상

이와 같은 과정은 현장답사나 거래당사자 인터뷰 등을 직접 실시할 수 없는 연구자 입장에서 가장 비용-효율적으로 신고자료를 정제할 수 있는 방법이라 생각되며, 실거래가 신고자료를 활용하는 향후 유사한 연구에서도 이와 같은 절차를 거칠 경우 분석의 신뢰성이 높아질 것으로 기대된다.

[표 3-4]는 이와 같은 과정을 거쳐 정제된 실거래가 신고자료의 현황을 보여주며, [그림 3-1]은 강남구에서 제거된 상하위 5% 잔차 현황을 보여준다. [그림 3-1] (a)를 보면 대략 +20억 이상 또는 -15억 이상 차이나는 사례들이 분석에서 제거되었음을 알 수 있다. 강남구에서 잔차값이 가장 큰 사례는 논현동에 위치한 토지면적 444.4㎡, 건물 연면적 2,087.96㎡의 주택으로 잔차는 무려 99.2억에 이른다. 본 주택은 토지면적 및 건물 연면적 모두 통상적인 주택의 규모를 상당히 상회하는 바, 항공사진 등을 검토한 결과, 5층 규모의 주상용 주택임을 확인할 수 있었다. 즉, 1층~4층은 상가, 사무실 등으로 이용 중이며 5층만 주택으로 이용 중인 건물이었다. 따라서 본 건물은 엄밀한 의미에서 '주택'이 아니며 오히려 상가 등으로 분류하는 것이 보다 합당하다. 그러나 실거래가 신고과정에서는 주택으로 신고 및 접수된 사례라 할 수 있다. 이러한 거래사례는 분석에서 제거하는 것이 합리적이다.

반면 강남구에서 잔차값이 가장 작은 사례는 역삼동에 위치한 토지면적 170.5㎡, 건물 연면적 216.1㎡의 주택으로 잔차는 -40.6억에 이른다. 세부 검토 결과 본 주택의 실제 토지면적은 588.1㎡로서 약 30%에 해당하는 170.5㎡의 토지만 소유권이 이전된 지분거래에 해당되는 사례였다. 특별한 사유가 없는 한 이러한 지분거래 역시 분석에서 제거하는 것이 합리적이다.

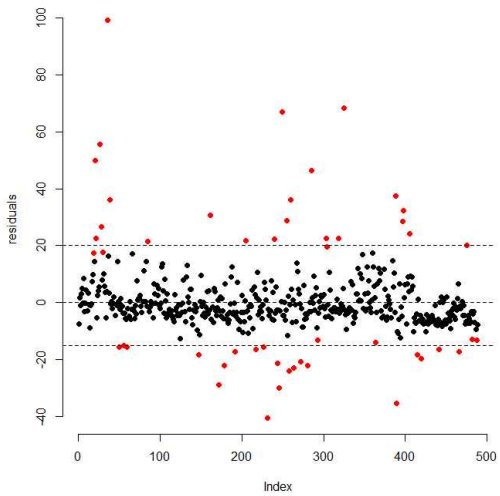
[그림 3-1]의 (b)~(d)를 보면 제거된 사례들에서 분포의 비대칭성이나 시공간상에서의 편중 현상은 없는 것으로 판단된다.

치 또는 극이상치 판단을 내릴 위험이 있으며, 자료가 정규분포를 따른다는 가정하에 관찰값의 10%를 초과하여 이상치를 제거하는 것은 바람직하지 않기 때문이다(Gloudemans & Almy, 2011, p.304).

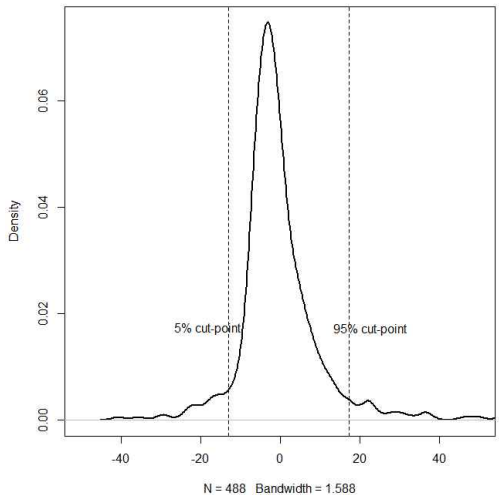
[표 3-4] 실거래가 신고자료의 정제

구 분	최초 건수	상하위 5%	정제 후 건수
서울 강남구	488건	50건	438건
전주 덕진구	2,202건	222건	1,980건
전남 해남군	697건	70건	627건

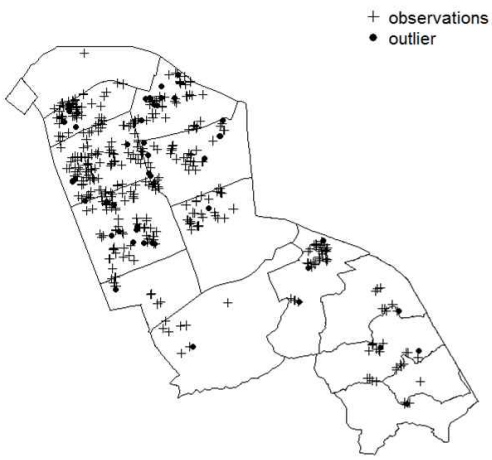
(a) 잔차 분포 I(적색: 이상치)



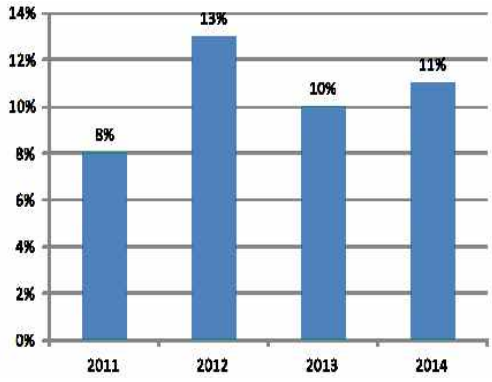
(b) 잔차 분포 II



(c) 제거된 거래사례(outlier)의 공간 분포

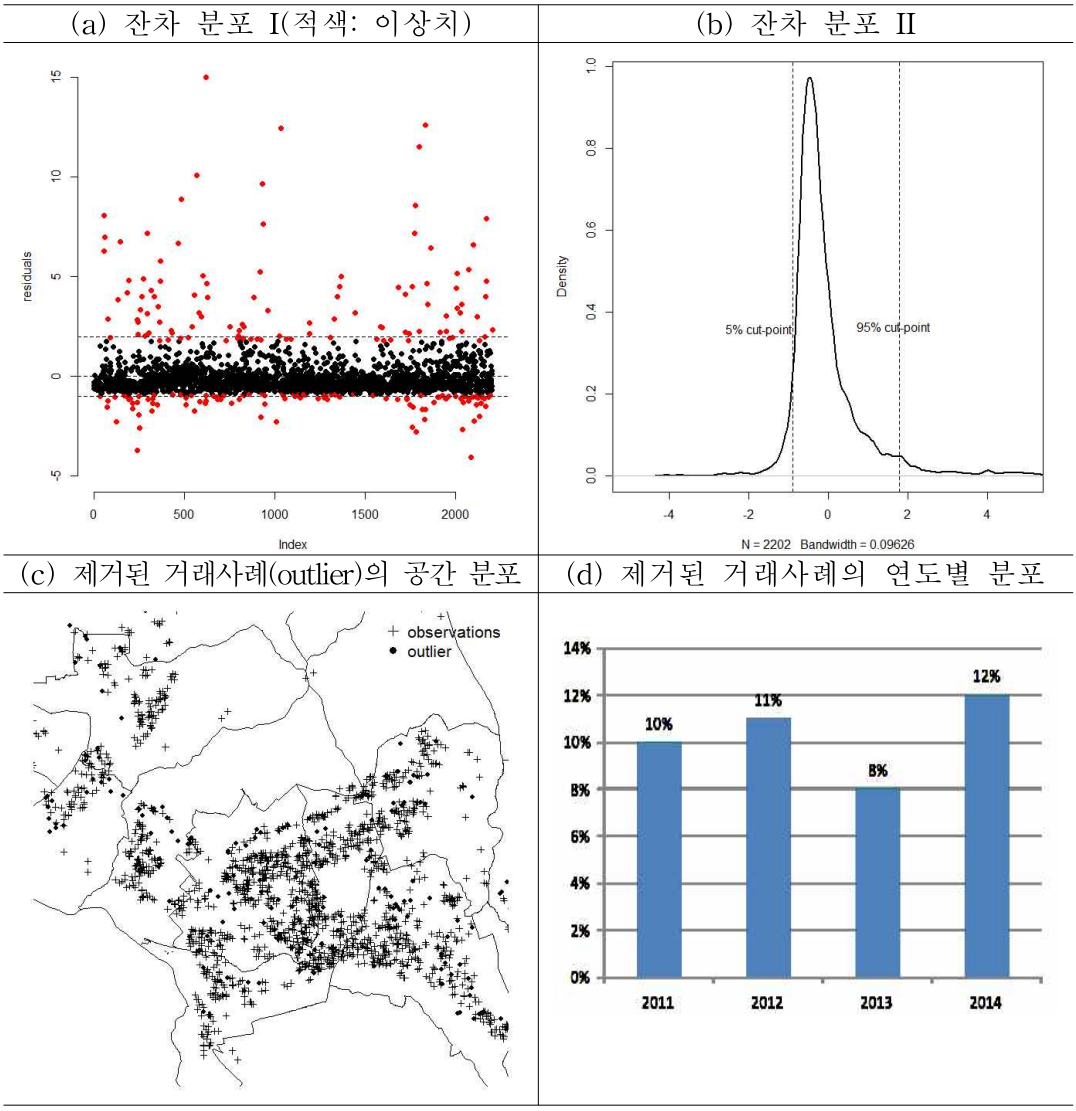


(d) 제거된 거래사례의 연도별 분포



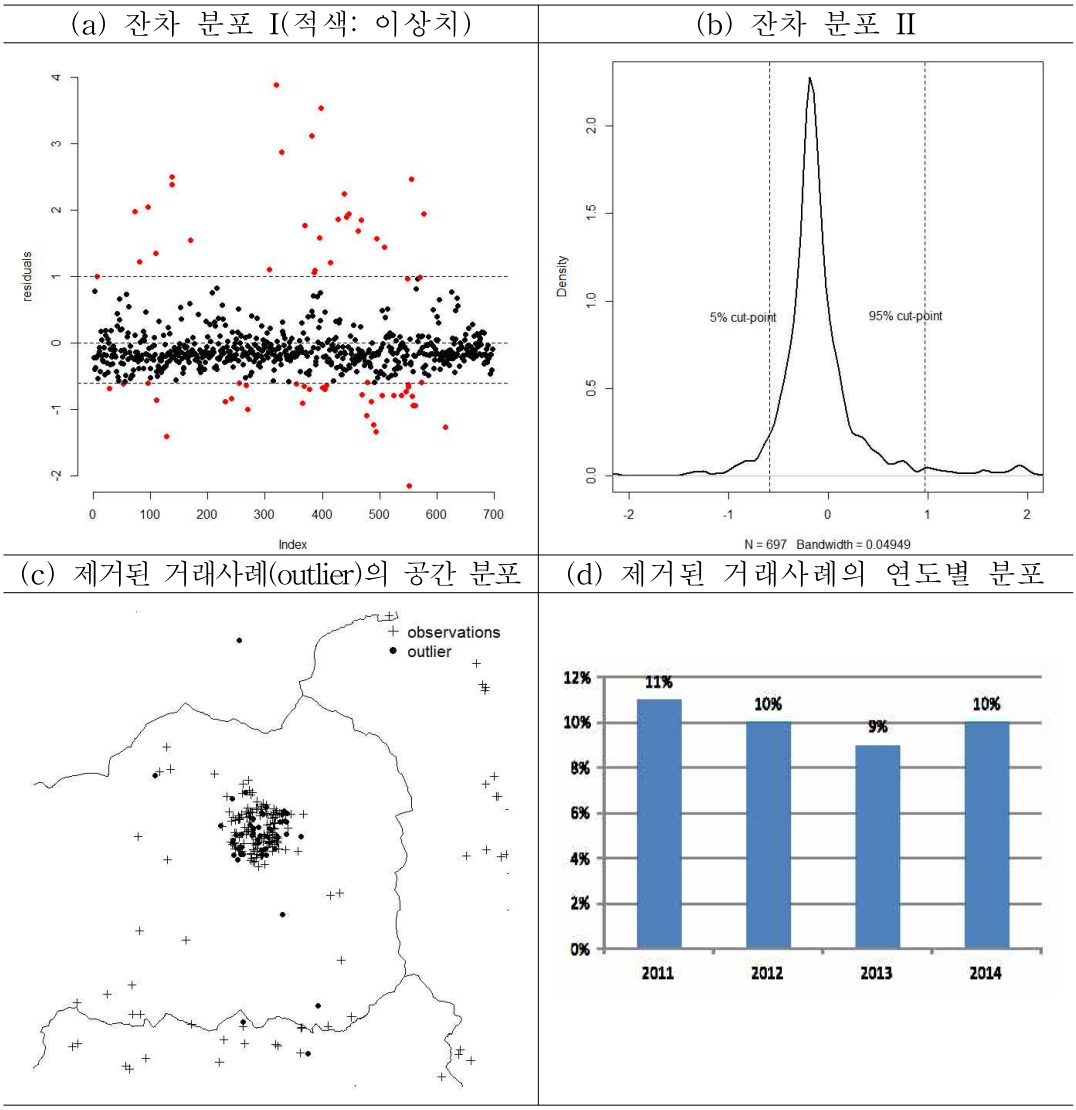
[그림 3-1] 강남구 잔차 분포

[그림 3-2] (a)를 보면, 전주시 덕진구의 경우 잔차 기준 대략 +2억 이상 또는 -1억 이상 차이나는 사례들이 분석에서 제거되었음을 알 수 있다. [그림 3-2] (b)의 경우 식 (3-1)에 의해 과대추정된 사례들, 즉 잔차가 음(-)이면서 값 자체가 크지 않은 사례들이 상대적으로 많이 제거된 것으로 보이나 분포의 비대칭성이 큰 편은 아니다. 강남구와 마찬가지로 시공간상에서의 이상치 편중 현상은 없는 것으로 판단된다.



[그림 3-2] 덕진구 잔차 분포

마찬가지로 [그림 3-3] (a)를 보면, 해남군의 경우 잔차 기준 대략 +1억 이상 또는 -0.5억 이상 차이나는 사례들이 분석에서 제거되었음을 알 수 있다. [그림 3-3] (b)의 경우 식 (3-1)에 의해 과대추정된 사례들, 즉 잔차가 음(-)이면서 값 자체가 크지 않은 사례들이 상대적으로 많이 제거된 것으로 보이나 분포의 비대칭성이 큰 편은 아니다. 강남구 및 덕진구와 마찬가지로 시공간상에서의 이상치 편중 현상은 없는 것으로 판단된다.



[그림 3-3] 해남군 잔차 분포

마지막으로 이와 같은 자료 정제 외에, 범주형 항목 중 빈도 수가 낮은 것은 통계적 유의성을 높이기 위해 가격형성 측면에서 유사한 범주끼리 통합하였다. [표 3-5]는 전체 설명변수 및 통합한 설명변수를 보여준다.

[표 3-5] 설명변수 목록

설명변수	통합 설명변수	설명변수	통합 설명변수
토지면적	(단위: ㎡)	인근지역 특징	6개에서 4개 범주로 통합 (주거/상업/주상/기타지대)
용도지역	7개 범주로 통합 (주거/상업/공업/녹지/관리/농림/자연환경보전)	건물 연면적	(단위: ㎡)
경사도	4개에서 2개 범주로 통합 (평지/완경사)	건물 경과연수	(단위: 年)
형상	8개에서 3개 범주로 통합 (정형/부정형/자루형)	건물 구조	25개에서 6개 범주로 통합 (철근콘크리트/연와/목조/블럭/판넬/기타)
방위	8개에서 2개 범주로 통합 (남향/남향외)	지붕 구조	16개에서 4개 범주로 통합 (슬라브/기와/슁글/기타)
도로접면	12개에서 6개 범주로 통합 (광로/중로/소로/세로/세로불/맹지)[25]	거래연도	2011년 ~ 2014년
공가	공가(空家) 여부	-	-

[표 3-5]를 보면 범주형 변수는 세부 항목을 모두 통합하였음을 알 수 있다. 예를 들어 용도지역의 경우 세분류 용도지역(주거지역인 경우 전용주거지역, 제1종일반주거지역, 제2종일반주거지역 등)은 대분류 용도지역인 주거지역으로 통합하였다. 이는 최초 범주형 변수의 세부 항목이 지나치게 많아 항목을 통합하지 않고 그대로 분석에 활용할 경우 대부분의 항목이 유의하지 않게 나올 가능성이 높기 때문이다. 예를 들어 건물구조의 경우 [표 3-5]에서는 가격수준 등을 참작하여 6개의 유사한 범주로 통합하였지만 최초의 범주는 [표 3-6]과 같이 건물구조가 지나치게 세분화되어 있음을 알 수 있다(25개 항목).

25)광로(廣路)가 폭이 가장 넓은 도로이며 세로불(細路不)이 폭이 가장 좁은 도로를 의미한다(광로>중로>소로>세로>세로불). 접하는 도로가 없는 필지를 맹지(盲地)라 하며, 기준범주(reference)는 중로(中路)로 정하였다.

[표 3-6] 건물구조의 최초 분류 현황

철골철근	통나무	철근콘크리트	철골	석조
Precast Concrete	목구조	라멘조	스틸하우스	연와
보강콘크리트	보강블럭	황토	시멘트벽돌	목조
시멘트블럭	경량철골	조립릭판넬	석회	토담
컨테이너	철파이프	ALC	와이어패널	기타

4. 기초 통계량

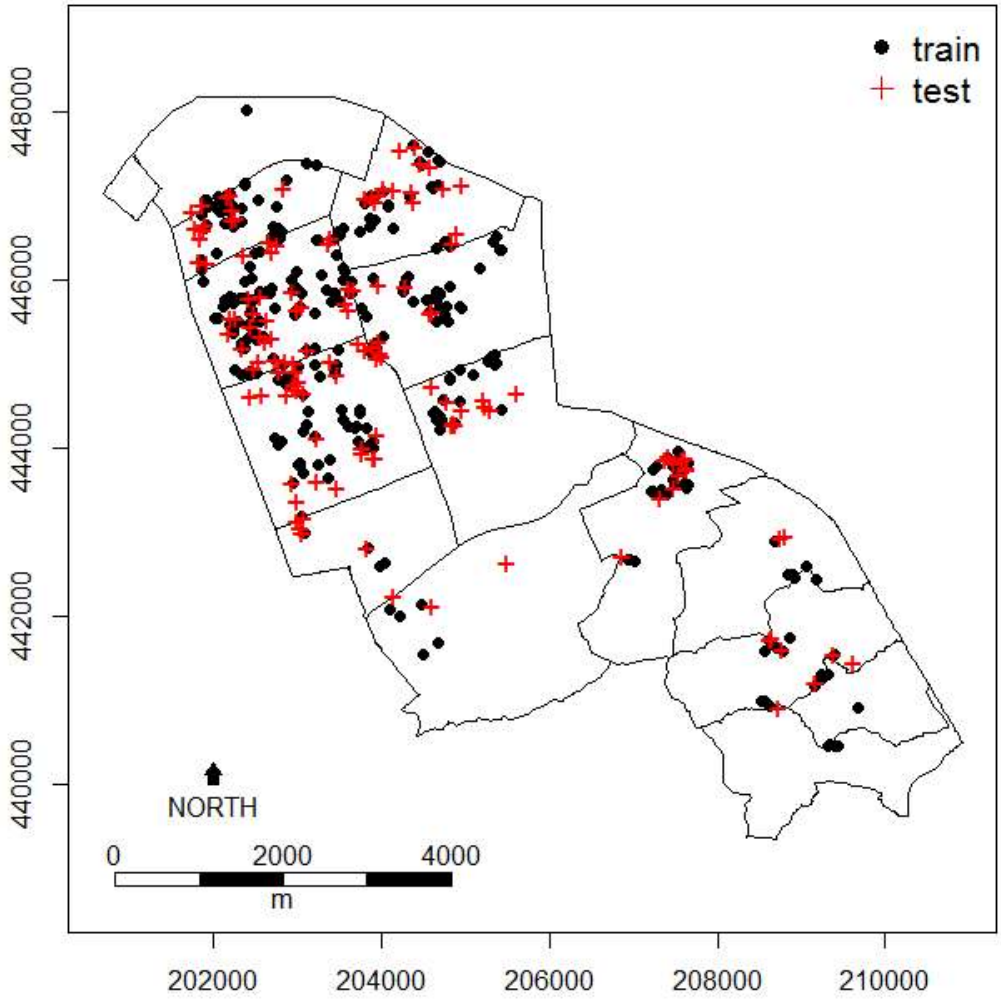
[표 3-7]은 분석에 활용된 강남구 주택 실거래가 438건의 기본적인 특징을 보여준다. 강남구의 전형적인 주택은 토지면적 약 220~240㎡에 건축된 지 24년 정도 경과된 건물로 거래금액은 20억원 내외임을 알 수 있다. 이는 토지면적 기준 약 900만원/㎡(3,000만원/3.3㎡)으로 총액이나 단가 모두 매우 고가임을 알 수 있다.

강남구라는 대도시 특성상 대부분의 주택이 주거지역에 위치하며, 철근콘크리트조 내지 연와조의 비교적 견고한 건물구조가 대다수임을 알 수 있다.

주택 실거래가의 공간적 분포는 [그림 3-4]와 같으며, 주로 북쪽의 신사동, 청담동, 논현동 등 주거지대에 밀집 분포하며, 남동쪽의 일원동 및 세곡동에도 간헐적으로 소재하고 있다.

[표 3-7] 강남구 주택 실거래가 기초 통계량(438건)

구 분	최소	최대	평균	중위수	표준편차
거래금액(억원)	2	82	21.07	19.84	9.23
토지면적(㎡)	41.92	908.30	241.78	224.60	90.76
건물 연면적(㎡)	65.79	1,193.42	344.95	304.47	179.92
경과연수(年)	1	42	24.45	24.00	8.10
용도지역	주거지역 (430건)			녹지지역 (8건)	
건물구조	철근콘크리트조 (188건)			연와조 (250건)	



[그림 3-4] 강남구 주택 거래사례 분포 현황

* 임의분할(random split)을 통해 전체 데이터의 70%는 모형 데이터(・train), 30%는 검증 데이터(+ test)로 활용(이하 타 지역 동일)

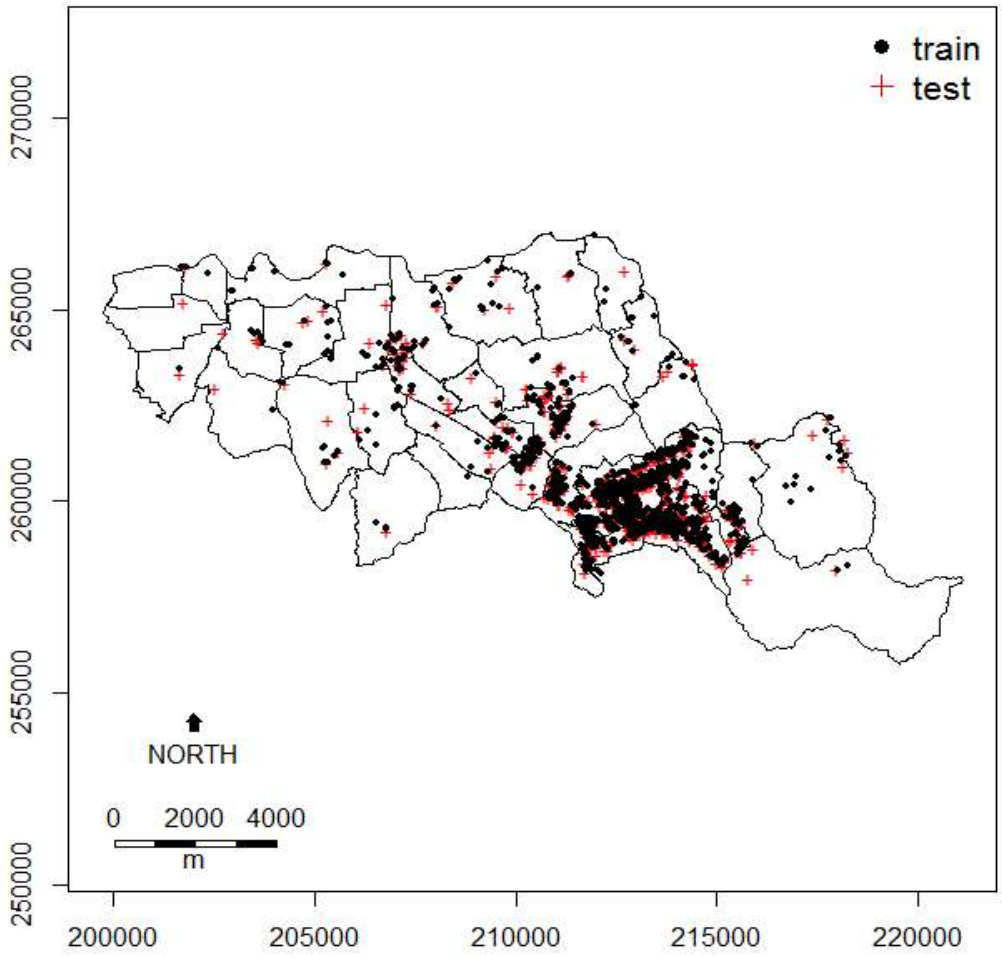
[표 3-8]은 분석에 활용된 덕진구 주택 실거래가 1,980건의 기본적인 특징을 보여준다. 덕진구의 전형적인 주택은 토지면적 약 200㎡ 내외에 건축된 지 30년 정도 경과된 건물로 거래금액은 1.0~1.5억원 수준임을 알 수 있다. 이는 토지면적 기준 약 60만원/㎡(200만원/3.3㎡) 정도에 해당된다.

도농복합지역의 특성상 주거지역에서부터 관리지역에 이르기까지 다양한 용도지역에 걸쳐 주택이 분포하고 있다. 건물구조 또한 철근콘크리트조 및 연와조 이외에도 목조, 블럭조 등 강남구에 비해 상대적으로 다양한 종류의 구조가 존재함을 알 수 있다. 가장 오래된 주택은 114년 경과된 목조 주택으로 1900년대에 신축되었다.

주택 실거래가의 공간적 분포는 [그림 3-5]와 같으며, 주로 도심 부분에 주택 거래사례가 밀집 분포하고 있음을 알 수 있다. 덕진구의 도심은 금암동, 인후동, 진북동 일대로서 통상 2층 규모의 30년 이상 경과된 연와조 건물이 주를 이룬다.

[표 3-8] 덕진구 주택 실거래가 기초 통계량(1,980건)

구 분	최소	최대	평균	중위수	표준편차
거래금액(억원)	0.06	9.40	1.57	1.09	1.48
토지면적(㎡)	12.0	14,918.0	225.8	187.1	349.83
건물 연면적(㎡)	22.2	1,166.6	164.0	106.9	148.10
경과연수(年)	1	114	31.49	33	14.50
용도지역	주거지역 (1,638건)	상업지역 (118건)	공업지역 (63건)	녹지지역 (149건)	관리지역 (12건)
건물구조	철근콘크리트 (334건)	목조 (172건)	블럭조 (128건)	연와조 (1,317건)	기타 (29건)



[그림 3-5] 덕진구 주택 거래사례 분포 현황

[표 3-9]는 분석에 활용된 해남군 주택 실거래가 627건의 기본적인 특징을 보여준다. 해남군의 전형적인 주택은 토지면적 약 400㎡ 내외에 건축된 지 35년 정도 경과된 건물로 거래금액은 3천만원~5천만원 수준임을 알 수 있다. 이는 토지면적 기준 약 10만원/㎡(33만원/3.3㎡) 정도에 해당된다.

해남군은 전형적인 농어촌 지역으로 관리지역에 전체 거래사례의 약 2/3 정도가 분포(407건, 65%)한다. 그러나 농림지역이나 자연환경보전지역에 위치한 주택도 상당 수 존재함을 알 수 있다. 건물구조 중 가장 많은 비중을 차지하는 것은 목조(260건, 41%)로서 대부분 경과연수 50년 이상이며 가장 오래된 주택은 109년 경과된 목조 주택으로 1900년대에 신축되었다.

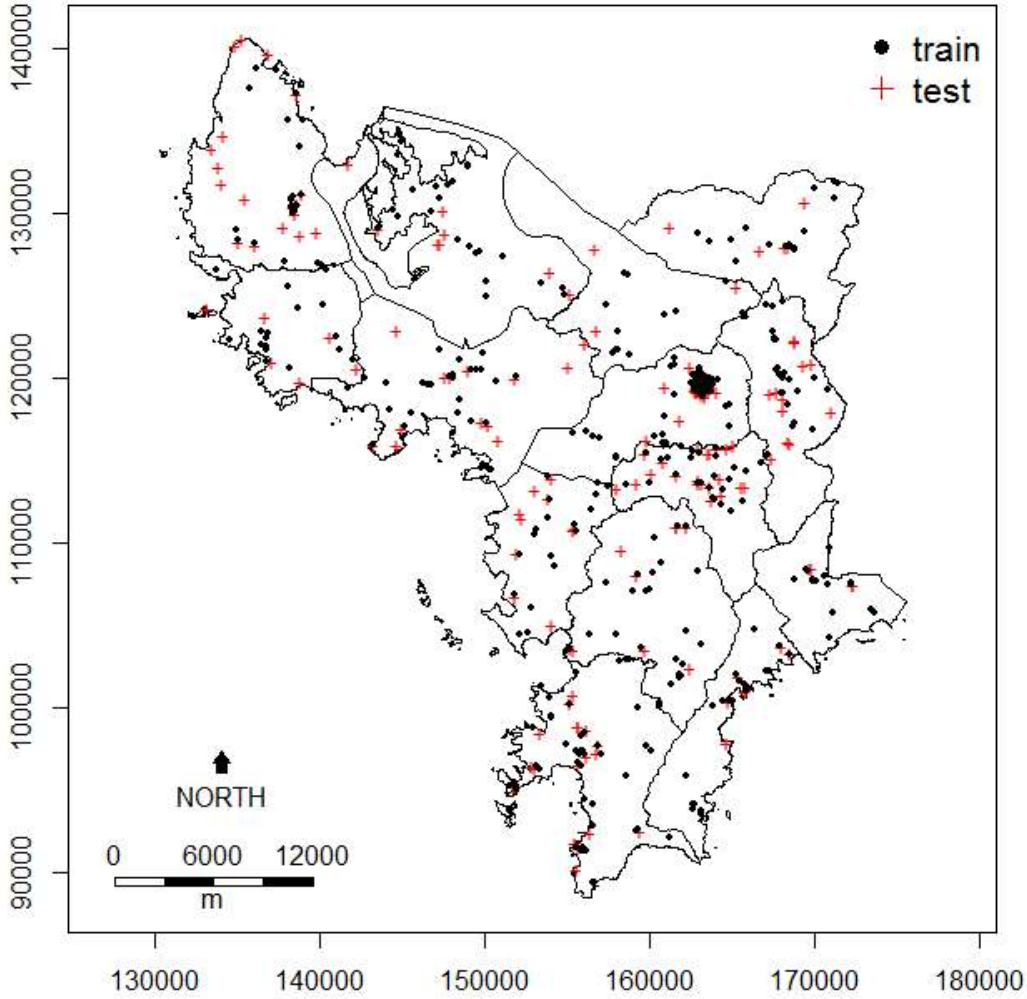
주택 실거래가의 공간적 분포는 [그림 3-6]과 같으며, 중심부(해남읍)에 일부 주택 거래사례가 밀집 분포하는 것을 제외하고는 군 전체에 걸쳐 고르게 분포하고 있다. 해남군의 주택은 통상 1층 규모의 30년 이상 경과된 목조 건물이 주를 이룬다.

[표 3-9] 해남군 주택 실거래가 기초 통계량(627건)

구 분	최소	최대	평균	중위수	표준편차
거래금액(억원)	0.006	4.0	0.47	0.30	0.49
토지면적(㎡)	18.0	4,693.0	456.6	367.0	399.20
건물 연면적(㎡)	15.60	479.84	96.79	84.69	57.92
경과연수(年)	2	109	36.38	34	20.89

용도지역	주거지역 (164건)	상업지역 (19건)	녹지지역 (8건)	관리지역 (407건)	농림지역 (21건)	자연지역* (8건)
건물구조	철근** (40건)	목조 (260건)	블럭조 (101건)	연와조 (193건)	판넬조 (8건)	기타 (25건)

* 자연환경보전지역, ** 철근콘크리트조



[그림 3-6] 해남군 주택 거래사례 분포 현황

제 2 절 선형회귀모형(OLS)의 적용

본 절에서는 기본 모형(Null Model)에 해당하는 선형회귀모형의 적합에 대해 논의한다. 선형회귀모형의 적합을 통해 주택 가격 예측에 도움이 되는 설명변수를 선별하고자 하며, 이후 논의될 비모수 모형과 성능을 비교할 때 준거 모형으로 삼고자 한다.[26]

선형회귀모형(Ordinary Least Squares Model)의 적합을 통해, 통계적 유의성이 있다고 인정되는 변수 또는 통계적 유의성은 없다 하더라도 주택시장에서의 거래관행 내지 일반적인 기대와 부합하는 부호(+, -)가 산출된 변수를 설명변수로 선정하였다.

[표 3-10] ~ [표 3-12]는 3개 사례지역의 OLS 모형 적합 결과를 보여준다[27]. 앞서 언급하였듯 설명변수는 통계적으로 유의하지 않더라도 계수의 부호가 직관과 일치한다면 모형에 포함시켰다. 예를 들어 [표 3-10]에서 방위 변수는 통계적 측면에서 유의하지 않지만(p-value 0.57), 기준범주인 남향 대비 남향이 아닌 주택은 약 4,200만원(-0.42) 낮게 가격이 형성되는 등 부호는 일반적인 기대와 일치한다. 지붕구조의 경우에도 마찬가지여서 통계적으로는 유의하지 않지만(p-value 0.15) 기준범주인 슬라브 지붕 대비 기와 지붕으로 지어진 주택의 가격이 낮게 형성되는 등 일반적인 기대와 일치한다. 따라서 이러한 설명변수들은 모두 모형에 포함시켰다. 예측 중심의 모형 구축에서 설명변수의 p-value가 유의하지 않더라도 부호(+, -)가 통상적인 기대와 일치한다면 모형에 포함시키는 것이 일반적이다(Kuhn & Johnson, 2013). 이와 같이 처리하는 것은 조금이라도 모형의 예측력을 향상시키는데 도움이 되기 때문이다(Gelman & Hill, 2007, p.69).

[표 3-10]에 나타난 서울시 강남구의 OLS 모형 적합 결과를 보면, 용도지역, 도로접면, 방위, 경과연수, 건물 연면적, 토지면적, 건물구조, 지붕구조 및 인근지역 특징이 비교적 유의한 설명변수로 도출되었다. 반면

26) 모든 모형의 적합은 통계 패키지 R(version 3.0.2)을 사용하였다.
27) 종속변수는 주택 거래가격 총액(단위: 억원)을 사용하였으며, 자연로그를 취한 값보다 모형의 적합 결과가 양호하여 별도의 종속변수 변환은 시도하지 않았다.

거래연도(2011년~2014년)는 유의하지 않게 산출되어 설명변수에서 제외하였다.

[표 3-10] 서울시 강남구 OLS 모형(Adj. R^2 = 61.6%)

설명변수		계수	표준오차	t-value	p-value	기준범주
상수항		10.60	4.12	2.57	0.01	
용도지역	녹지지역	-5.01	2.75	-1.83	0.07	주거지역
도로접면	소로	-3.42	2.47	-1.39	0.17	중로
	세로	-2.91	2.35	-1.24	0.22	
	세로불	-3.50	3.00	-1.17	0.24	
방위	남향 외	-0.42	0.73	-0.57	0.57	남향
경과연수		-0.50	0.22	-2.24	0.03	
경과연수^2		0.01	0.00	2.77	0.01	
건물 연면적		0.01	0.00	2.39	0.02	
토지 면적		0.07	0.00	16.61	0.00	
건물구조	연와조	-3.67	1.13	-3.25	0.00	철근콘크리트
지붕구조	기와	-4.38	3.03	-1.45	0.15	슬라브
인근지역 특징	상업지대	8.17	2.55	3.20	0.00	주거지대
	주상지대	4.26	1.27	3.36	0.00	

[표 3-11]은 전주시 덕진구의 OLS 모형 적합 결과를 보여준다. 서울시 강남구와 비교할 때 경사도, 거래연도 및 공가(空家)가 유의한 설명변수로 추가되었다. 즉 평지에 건축된 주택에 비해 완경사지에 건축된 주택은 약 100만원 정도(-0.01) 가격이 하락하며, 기준범주인 2011년 거래대비 이후 연도의 거래는 상승 추세에 있는 것으로 해석할 수 있다. 또한 사람이 거주하는 정상적 주택에 비해 공가는 약 4,400만원(-0.44) 가격이 하락하는 것으로 산출되었다[28].

28) 서울 강남구의 경우 공가주택은 없었다.

[표 3-11] 전주시 덕진구 OLS 모형(Adj. R^2 = 83.5%)

설명변수		계수	표준오차	t-value	p-value	기준범주
상수항		1.60	0.14	11.70	0.00	
용도지역	상업지역	0.03	0.08	0.41	0.68	주거지역
	공업지역	-0.36	0.10	-3.65	0.00	
	녹지지역	-0.39	0.14	-2.88	0.00	
	관리지역	-0.87	0.26	-3.31	0.00	
도로접면	광로	0.19	0.13	1.43	0.15	중로
	소로	-0.13	0.08	-1.69	0.09	
	세로	-0.22	0.08	-2.89	0.00	
	세로불	-0.21	0.08	-2.55	0.01	
방위	남향 외	-0.01	0.03	-0.35	0.73	남향
경사도	완경사	-0.01	0.06	-0.24	0.81	평지
경과연수		-0.07	0.00	-13.37	0.00	
경과연수^2		0.00	0.00	11.65	0.00	
건물 연면적		0.01	0.00	33.28	0.00	
토지 면적		0.00	0.00	8.50	0.00	
건물구조	기타구조	-0.25	0.15	-1.72	0.09	철근콘크리트
	목조	0.28	0.11	2.53	0.01	
	블럭조	0.32	0.11	2.96	0.00	
	연와조	0.12	0.08	1.61	0.11	
인근지역 특징	상업지대	0.28	0.12	2.28	0.02	주거지대
	주상지대	0.03	0.06	0.45	0.65	
	기타지대	-0.17	0.13	-1.28	0.20	
거래연도	2012년	0.18	0.04	4.52	0.00	2011년
	2013년	0.12	0.04	2.85	0.00	
	2014년	0.15	0.06	2.73	0.01	
공가	공가(空家)	-0.44	0.17	-2.57	0.01	非空家

마지막으로 [표 3-12] 해남군의 결과를 보면 용도지역, 도로접면, 방위, 경과연수, 건물 연면적, 토지 면적 및 건물구조가 비교적 유의한 설명변수로 도출되었다. 특히 건물구조의 경우 모든 세부 항목에서 통계적 유의성이 높게 산출되었는데 기준이 되는 철근콘크리트 구조 대비 다른 구조의 건물은 모두 음(-)의 부호가 산출되어 일반적인 직관과 일치한다. 특히 건물구조 중 가장 조잡한 것으로 여겨지는 블럭조나 판넬조의 계수가 가장 낮게 산출되어 건물구조에 따른 서열도 시장에서의 일반적인 선호도와 일치한다.

[표 3-12] 전라남도 해남군 OLS 모형(Adj. R^2 = 50.2%)

설명변수		계수	표준오차	t-value	p-value	기준범주
상수항		1.56	0.22	6.94	0.00	
용도지역	상업지역	-0.02	0.11	-0.14	0.89	주거지역
	녹지지역	-0.40	0.13	-2.95	0.00	
	관리지역	-0.37	0.04	-8.95	0.00	
	농림지역	-0.44	0.10	-4.26	0.00	
	자보지역*	-0.56	0.13	-4.26	0.00	
도로접면	소로	-0.04	0.21	-0.20	0.84	중로
	세로	-0.07	0.21	-0.32	0.75	
	세로불	-0.05	0.21	-0.23	0.82	
방위	남향 외	-0.04	0.03	-1.30	0.19	남향
경과연수		-0.02	0.00	-6.13	0.00	
경과연수^2		0.00	0.00	4.98	0.00	
건물 연면적		0.00	0.00	5.85	0.00	
토지 면적		0.00	0.00	3.34	0.00	
건물구조	기타구조	-0.38	0.11	-3.55	0.00	철근콘크리트
	목조	-0.46	0.09	-5.02	0.00	
	블럭조	-0.66	0.09	-7.51	0.00	
	연와조	-0.53	0.07	-7.12	0.00	
	판넬조	-0.63	0.18	-3.44	0.00	

* 자연환경보전지역

[표 3-13]은 [표 3-10] ~ [표 3-12]에서 제시된 OLS 모형의 가격 예측 성능을 보여준다. 검증 데이터를 기준한 결과가 보다 신뢰성이 있다고 볼 수 있으므로 이를 기준으로 해석하면 다음과 같다.

가격 정확성을 나타내는 SR의 경우 매우 과소평가하거나 과대평가한 관찰치가 많아 이러한 극단치에 비교적 덜 민감한 중위수 SR을 기준할 경우 1.01 ~ 1.15 정도의 범위를 보이고 있어 모형에 약간의 상향 편의가 존재함을 알 수 있다. 이러한 현상은 SR을 기준한 분석에서 종종 나타나는 것인데, SR의 하한은 0.0으로 제한되어 있으나 상한은 그렇지 않아 나타나는 현상으로 해석된다.

가격 형평성을 나타내는 RMSE, MAE 및 COD 중에서 RMSE 및 MAE는 강남구→덕진구→해남군으로 갈수록 그 값이 낮아지는데, 이는

모형 성능보다는 주택가격의 크기와 관련이 있는 것으로 보인다. 즉 강남구는 10억원대 이상의 주택이 흔한 편이나, 덕진구는 10억원대 미만의 주택이 대부분이며 해남군의 경우 1억원 이하의 주택이 많은 비중을 차지하고 있다. 따라서 주택가격의 크기가 작을수록 RMSE 및 MAE는 작게 나올 수밖에 없는 바, 지역 간 유효한 비교 지표가 될 수 없는 것으로 판단된다.

따라서 COD를 기준으로 가격 형평성 또는 모형의 적합도(Goodness Of Fit)를 판단하면 강남구의 경우가 가장 우수하고 다음이 덕진구, 해남군의 순임을 알 수 있다. 특히 강남구는 현재의 OLS 모형을 그대로 사용하여도 될 만큼 가격 형평성이 우수한 것으로 판단되는 반면, 해남군과 같은 농촌지역의 주택가격 예측은 상대적으로 어렵다는 것을 추론할 수 있다. 강남구 소재 단독주택 부지는 대부분 200㎡ 내외의 평탄한 사각형 토지이며, 건물구조 또한 철근콘크리트조 내지 연와조 정도로 단순하여 대량평가 모형으로 접근하기가 수월한 편이다. 반면, 해남군은 20㎡~30㎡ 규모의 작은 부지에서부터 4,000㎡ 이상의 대규모 부지에 이르기까지 면적도 다양하며 사각형을 비롯한 부정형, 자루형 등 토지의 형상 또한 그 종류가 매우 많다. 건물구조도 철근콘크리트조, 연와조 뿐만 아니라 목조, 블록조, 판넬조 등 다양한 편이어서 해남군 소재 단독주택은 대량평가 모형으로 접근하기가 수월하지 않다. 전주시 덕진구는 양 지역의 중간 정도에 해당된다고 할 수 있다. 이러한 지역 특성으로 인해 강남구 주택은 헤도닉 모형으로 접근하기가 비교적 수월하고, 해남군은 그렇지 않은 것으로 풀이된다.

[표 3-13] OLS 모형 성능

지역	구분	평균 SR	중위수 SR	RMSE	MAE	COD
강남구	모형 데이터	1.07	1.02	5.82	4.33	22.14
	검증 데이터	**1.16**	**1.06**	**5.83**	**4.52**	**28.41**
덕진구	모형 데이터	1.17	1.03	0.60	0.42	40.14
	검증 데이터	**1.13**	**1.01**	**0.65**	**0.42**	**37.64**
해남군	모형 데이터	1.59	1.16	0.32	0.22	79.46
	검증 데이터	**1.52**	**1.15**	**0.34**	**0.21**	**83.86**

제 3 절 비모수 모형의 적용

1. 일반가산모형(Generalized Additive Model)

GAM은 앞 절의 선형회귀모형과 다음에서 설명할 비모수 모형(랜덤 포리스트 등)의 중간 정도에 해당되는 모형이다. 각 설명변수마다 비선형 함수의 형태를 가정하지만 최종 단계에서는 이를 모두 합산하는, 즉 '선형결합하는' 형태를 취하므로 완전한 형태의 비모수 모형은 아니다. [표 3-14] ~ [표 3-16]은 3개 사례지역의 GAM 적합 결과를 보여준다. 선형 변수(Linear Term)들의 부호 등은 OLS 결과와 유사하며 강남구는 경과연수와 건물 연면적, 덕진구는 경과연수와 토지면적, 마지막으로 해남군은 경과연수만이 비선형 변수(Non-linear Term)로 분류되었다[29].

[표 3-14] 서울시 강남구 GAM 모형

설명변수		계수	표준오차	t-value	p-value	기준범주
상수항		6.67	2.72	2.45	0.01	
용도지역	녹지지역	-4.10	2.71	-1.51	0.13	주거지역
도로접면	소로	-2.36	2.44	-0.97	0.33	중로
	세로	-2.22	2.31	-0.96	0.34	
	세로불	-2.54	2.95	-0.86	0.39	
방위	남향 외	-0.07	0.72	-0.10	0.92	남향
토지 면적		0.07	0.00	16.96	0.00	
건물구조	연와조	-2.64	1.19	-2.23	0.03	철근콘크리트
지붕구조	기와	-3.82	2.95	-1.29	0.20	슬라브
인근지역 특징	상업지대	8.08	2.49	3.25	0.00	주거지대
	주상지대	4.98	1.24	4.01	0.00	
비선형 변수(smooth terms)						
설명변수		추정 자유도(estimated d.f.)		F-value	p-value	
경과연수		4.41		5.09	0.00	
건물 연면적		3.37		3.28	0.01	

29) 설명변수별 비선형 함수, 즉 기저함수(Basis Function)는 Cubic Regression Spline을 사용하였으며, 강남구의 경우 토지면적은 주택가격과 선형에 가까운 관계가 있는 것으로 산출되어(추정 자유도 2.35) 선형 변수로 분류하였다. 마찬가지로 덕진구의 경우 건물 연면적은 선형 변수(추정 자유도 4.21)로 처리하는 것이 더 적정한 것으로 분석되었다. 해남군의 경우 토지면적(추정 자유도 0.97) 및 건물 연면적(추정 자유도 1.58) 모두 선형변수로 처리하는 것이 보다 적정한 것으로 판단되었다.

[표 3-15] 전주시 덕진구 GAM 모형

설명변수		계수	표준오차	t-value	p-value	기준범주
상수항		0.62	0.10	5.90	0.00	
용도지역	상업지역	-0.04	0.07	-0.49	0.63	주거지역
	공업지역	-0.40	0.09	-4.60	0.00	
	녹지지역	-0.40	0.12	-3.27	0.00	
	관리지역	-0.95	0.23	-4.05	0.00	
도로접면	광로	0.21	0.12	1.76	0.08	중로
	소로	-0.05	0.07	-0.69	0.49	
	세로	-0.13	0.07	-1.85	0.06	
	세로불	-0.14	0.07	-2.00	0.05	
방위	남향 외	-0.02	0.03	-0.74	0.46	남향
경사도	완경사	0.01	0.05	0.23	0.82	평지
건물 연면적		0.01	0.00	29.65	0.00	
건물구조	기타구조	-0.36	0.13	-2.82	0.00	철근콘크리트
	목조	0.18	0.10	1.77	0.08	
	블럭조	0.18	0.10	1.80	0.07	
	연와조	0.17	0.07	2.36	0.02	
인근지역 특징	상업지대	0.42	0.11	3.89	0.00	주거지대
	주상지대	0.16	0.05	3.09	0.00	
	기타지대	-0.11	0.12	-0.94	0.35	
거래연도	2012년	0.14	0.04	3.97	0.00	2011년
	2013년	0.07	0.04	1.78	0.08	
	2014년	0.09	0.05	1.88	0.06	
공가	공가(空家)	-0.18	0.15	-1.17	0.24	非空家

비선형 변수(smooth terms)			
설명변수	추정 자유도(estimated d.f.)	F-value	p-value
경과연수	8.59	59.84	0.00
토지면적	8.13	22.75	0.00

[표 3-16] 전라남도 해남군 GAM 모형

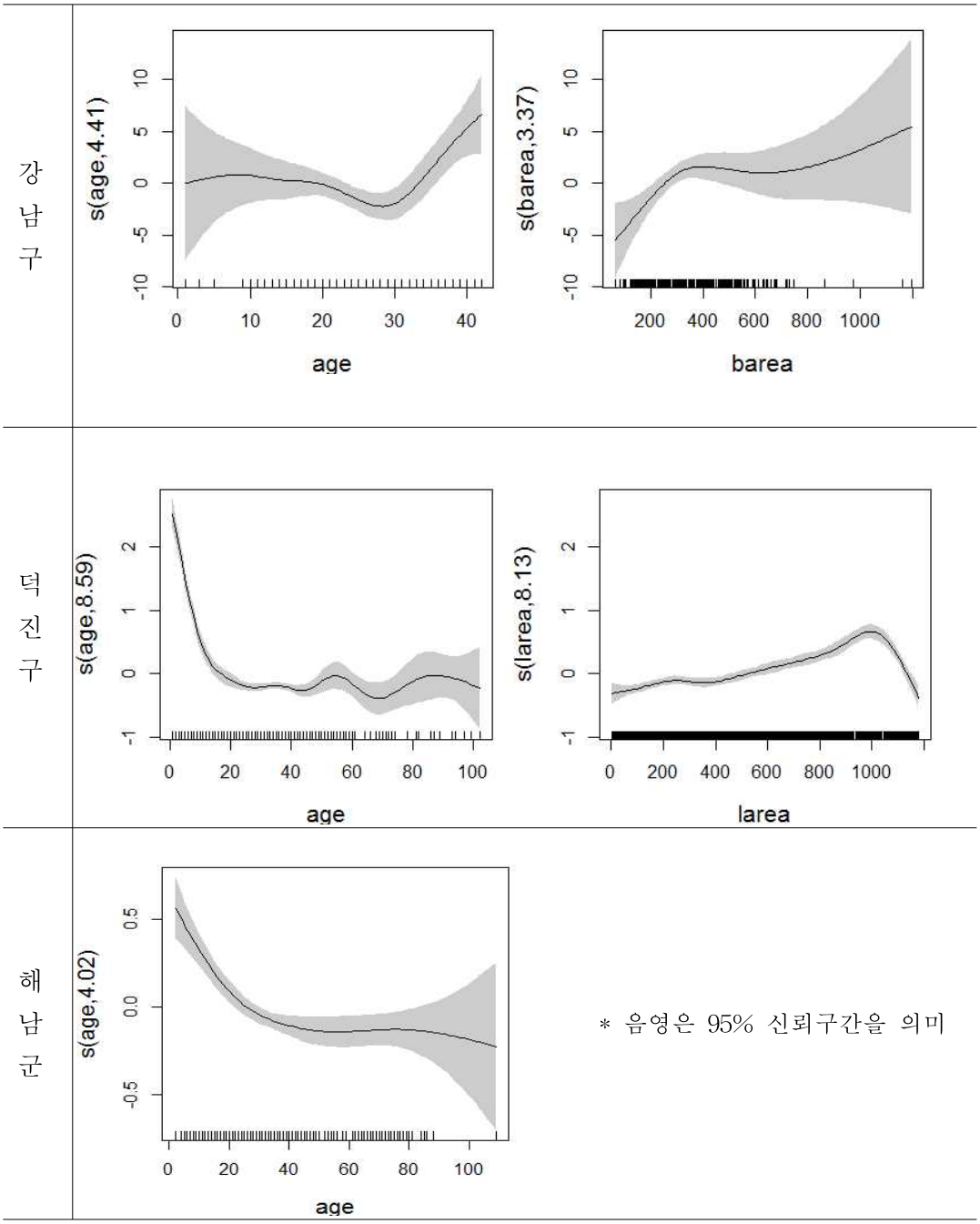
설명변수		계수	표준오차	t-value	p-value	기준범주
상수항		1.06	0.21	5.01	0.00	
용도지역	상업지역	-0.02	0.11	-0.23	0.82	주거지역
	녹지지역	-0.40	0.13	-3.00	0.00	
	관리지역	-0.38	0.04	-9.19	0.00	
	농림지역	-0.45	0.10	-4.38	0.00	
	자보지역*	-0.57	0.13	-4.40	0.00	
도로접면	소로	-0.10	0.21	-0.47	0.64	중로
	세로	-0.12	0.21	-0.57	0.57	
	세로불	-0.09	0.21	-0.45	0.65	
방위	남향 외	-0.04	0.03	-1.13	0.26	남향
토지 면적		0.00	0.00	3.14	0.00	
건물 연면적		0.00	0.00	6.25	0.00	
건물구조	기타구조	-0.38	0.11	-3.58	0.00	철근콘크리트
	목조	-0.45	0.09	-4.95	0.00	
	블럭조	-0.62	0.09	-6.99	0.00	
	연와조	-0.49	0.08	-6.41	0.00	
	판넬조	-0.59	0.18	-3.21	0.00	

비선형 변수(smooth terms)			
설명변수	추정 자유도(estimated d.f.)	F-value	p-value
경과연수	4.02	10.4	0.00

* 자연환경보전지역

[그림 3-7]은 지역별 비선형 변수들의 패턴을 보여준다. 3개 사례지역 모두에서 경과연수가 비선형 변수로 채택되었음을 알 수 있는데, 그 패턴은 지역마다 상이한 것으로 보인다. 경과연수 이외에 강남구는 건물면적(*barea*), 덕진구는 토지면적(*larea*)이 비선형 변수로 채택되었고 그 패턴 또한 지역마다 상이한 것으로 판단된다.

[그림 3-7] 강남구에서 경과연수의 경우, 신축 이후 약 10년까지는 가격의 증감이 거의 없다가 10년~30년 구간에서 시간이 지날수록 주택 가격이 점진적으로 하락한다. 그러나 30년 이상을 경과하는 시점부터는 가격이 다시 급속히 오르기 시작하는데, 재건축·재개발 등으로 인한 기대감 때문인 것으로 풀이된다.



[그림 3-7] 지역별 비선형 변수의 패턴(GAM 모형)

강남구 건물 연면적의 경우 약 400㎡까지는 면적이 증가할수록 주택 가격도 비례하여 상승하지만 그 이상 면적이 증가하는 경우에는 주택 가격의 증감이 거의 없는 것을 알 수 있다. 이는 건물 연면적 400㎡, 즉 거주 공간 120평까지는 주택 수요자들이 정상적인 주거효용을 느끼는 규모이지만, 이를 초과하는 경우 관리비 증가, 재산세 증가 등으로 인해 주택으로서의 효용 증가를 거의 느끼지 못하는 것으로 해석할 수 있다.

전주시 덕진구의 경우 경과연수와 토지면적이 주택 가격과 비선형 관계가 있는 것으로 파악되었다. 경과연수의 경우 대도시인 강남구와는 다른 패턴이 산출되었는데, 신축 이후 약 20년까지는 주택 가치가 급속하게 하락하다가 그 이후부터는 가치 하락이 거의 없거나 완만하게 하락하는 형태를 보이며, 이러한 기간은 최고 100년까지 지속되는 것으로 보인다. 덕진구는 강남구와 달리 재개발·재건축 기대감으로 인한 가격상승 현상은 없는 것으로 해석된다.

반면 덕진구 토지면적 변수의 경우 면적이 증가할수록 주택가격도 증가하지만, 일정 면적(여기에서는 약 1,000㎡)을 넘어서면 오히려 주택가격이 하락하고 있다. 이러한 현상은 일반적인 직관에 반한다. 건물과 달리 토지는 면적이 과도하게 크더라도 분할하여 매각·이용하는 등 토지의 가치를 유지시킬 수 있는 여러 가지 방법이 존재하기 때문이다.

덕진구에서 토지면적과 주택가격의 이러한 비선형성은 덕진구가 전형적인 도농복합도시이며 따라서 주택 또한 도시형 주택과 농가형 주택으로 구성되어 있다는 사실을 감안하면 충분히 설명될 수 있는 현상이다. 다시 말해 덕진구 도심(인후동, 금암동 등)에 위치한 주택부지의 평균 면적은 200 ~ 400㎡ 정도이며, 외곽의 농촌지대(금상동, 호성동, 전미동 등)에 위치한 주택부지의 평균 면적은 400 ~ 600㎡ 정도이다. 경우에 따라 순수 농경지대에 위치한 농가주택은 주택부지가 1,000㎡를 상회하기도 한다. 주택부지가 1,000㎡를 상회하는 경우 해당 토지 전부를 주택부지로 사용하는 것이 아니라 약 30% 정도만 실제 주택의 부지로 사용하고, 나머지 약 70%는 전·답(텃밭)이나 임야(입목도가 매우 낮은 구릉지 같은 경사지)로 사용 중인 사례가 많다. 부지 규모가 1,000㎡ 이상인 주택은 토지의 지목도 대(垈)가 아닌 전(田)인 경우가 흔하다.

요약하면 덕진구의 경우 일정 면적까지는 일반적인 상식과 부합되게 토지면적이 증가할수록 주택가격도 증가하지만, 1,000㎡를 초과하는 시점부터는 주택부지에 전·답과 같은 타 용도의 토지가 더해짐으로써 전체적인 주택가격은 오히려 하락하는 패턴을 보이고 있는 것이다. 이와 같이 토지면적과 주택가격이 일반적으로 생각하는 선형의 관계를 보이지 않는 경우, 이러한 복잡한 관계를 반영할 수 있는 비선형 모형을 적용하는 것은 필수적이라 할 수 있다.

마지막으로 해남군은 경과연수만이 유일하게 뚜렷한 비선형 관계를 보여주었다. 주택가격과 경과연수의 관계는 강남구 및 덕진구와 상이한데, 신축 이후 약 40년까지 지속적으로 주택가격이 하락하다가 이후부터는 주택가격 변화에 큰 영향을 주지 못하는 것으로 보인다. 따라서 해남군 주택의 경우 강남구 및 덕진구 주택과 비교할 때 상대적으로 장기간(약 40년) 주택가격의 하락이 지속되는 것으로 볼 수 있다. 이는 한번 신축하면 중간에 철거하기보다는 물리적 내용연수가 다할 때까지 주택으로 사용하는 농가주택의 특징을 보여주는 것이다.

[표 3-17]은 GAM 모형의 성능을 보여준다. 가격 형평성 지표인 RMSE, MAE 및 COD를 기준으로 할 때, 3개 지역 모두 OLS 모형 대비 개선되었음을 알 수 있다. 즉 일부 설명변수에 비선형 효과를 반영함으로써 전반적인 모형 성능이 개선된 것으로 보인다.

[표 3-17] GAM 모형 성능

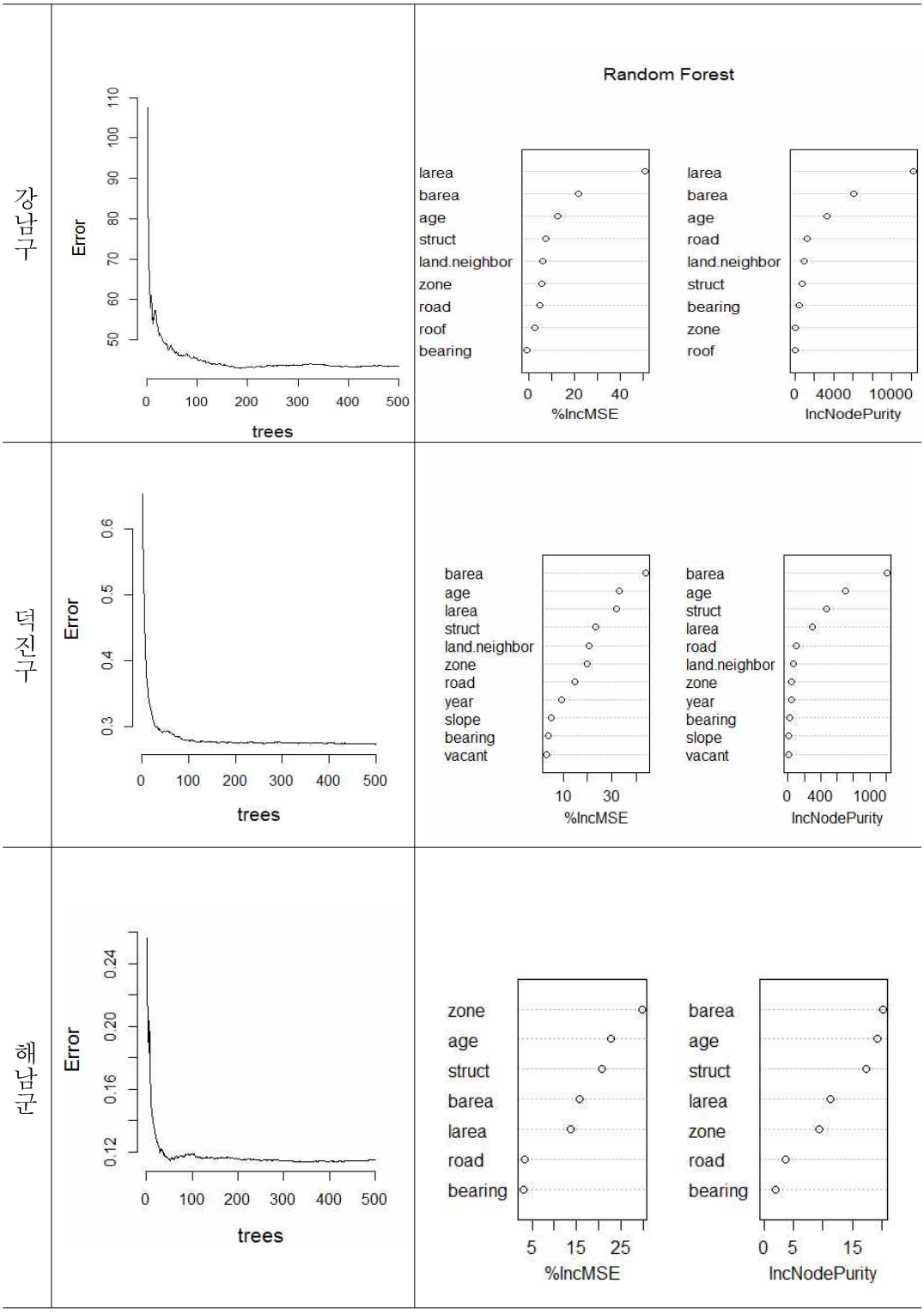
지역	구분	평균 SR	중위수 SR	RMSE	MAE	COD
강남구	모형 데이터	1.06	1.03	5.58	4.28	21.89
	검증 데이터	**1.12**	**1.04**	**5.58**	**4.43**	**27.02**
덕진구	모형 데이터	1.16	1.05	0.52	0.36	34.07
	검증 데이터	**1.13**	**1.03**	**0.58**	**0.37**	**35.13**
해남군	모형 데이터	1.58	1.19	0.32	0.22	75.36
	검증 데이터	**1.60**	**1.15**	**0.33**	**0.21**	**82.93**

2. 랜덤 포리스트(Random Forest)

회귀트리 500개를 생성하여 랜덤 포리스트를 적용한 결과는 [그림 3-8]과 같다. [그림 3-8]의 좌측은 회귀트리 개수의 적절성을 보여주는 것으로 수직축의 *Error*는 잔차제곱합의 평균을 나타낸다. 3개 지역 모두 200개 트리 이후부터는 일정한 *Error* 값으로 수렴한 것으로 보이며 따라서 500개 트리 결과에 기반한 추론에 무리가 없는 것으로 판단된다.

[그림 3-8]의 우측은 랜덤 포리스트를 통해 파악된 설명변수의 중요도를 보여준다. [그림 3-8] 우측에 제시된 두 가지 지표[30]에 따르면 강남구는 토지면적(*larea*), 덕진구는 건물 연면적(*barea*)이 가장 중요한 변수임을 알 수 있다. 해남군은 두 가지 지표의 결과가 상이한데, 용도지역(*zone*) 및 건물 연면적(*barea*)이 상대적으로 중요한 변수임을 추론할 수 있다.

30) %IncMSE라고 표기된 좌측은 MSE(Mean Squared Error), IncNodePurity라교 표기된 우측은 Node Impurity의 감소 정도[여기에서는 RSS(Residual Sum of Squares)의 감소 정도]를 기준한 변수의 중요성을 보여준다.



[그림 3-8] 트리개수의 적정성(좌) 및 설명변수의 중요도(우)

[표 3-18]은 지역별 랜덤 포리스트 모형의 성능을 보여준다. 가격 균형성 측면에서 전반적으로 강남구 및 덕진구는 OLS 모형보다 개선되었고, 해남군은 OLS 모형과 비슷한 수준임을 알 수 있다.

[표 3-18] 랜덤 포리스트 모형 성능

지역	구분	평균 SR	중위수 SR	RMSE	MAE	COD
강남구	모형 데이터	1.10	1.06	6.60	4.74	23.32
	검증 데이터	**1.21**	**1.09**	**5.50**	**4.15**	**28.19**
덕진구	모형 데이터	1.19	1.07	0.52	0.36	33.50
	검증 데이터	**1.16**	**1.02**	**0.56**	**0.37**	**34.06**
해남군	모형 데이터	1.78	1.29	0.34	0.23	77.05
	검증 데이터	**2.03**	**1.31**	**0.32**	**0.22**	**92.07**

3. 부스팅(Boosting)

부스팅은 앞 장에서 설명하였듯 최초 트리를 구성한 후, 이후 종속변수 Y가 아닌 잔차를 업데이트하는 방식으로 트리를 계속하여 수정하는 방법이다. 3개 사례지역 모두 *Gaussian Loss* 함수를 적용하였고, 5,000번의 반복 수정을 통해 최종 모형을 산출하였다[31].

[표 3-19]는 부스팅 적합 과정에서 나타난 설명변수의 상대적 영향력(Relative Influence)[32]을 나타낸다. 강남구는 토지면적이 영향력이 가장 큰 변수로 나타났고, 덕진구는 건물 연면적이 가장 중요한 변수로 산출되었다. 반면 해남군은 특별히 우월한 영향력을 보이는 변수는 없고, 건물구조 및 건물 연면적이 비슷한 비중으로 가격형성에 영향을 주는 것으로 나타났다. 앞에서의 랜덤 포리스트 결과와 유사하게 주요 변수가 산출된 것으로 보인다.

---

31) 부스팅을 실행하는 알고리즘은 다양하며(AdaBoost, C5 Boosting, Gradient Boosting 등) 본 연구에서는 Friedman(2001)의 Gradient Boosting Machine 알고리즘을 따랐다.

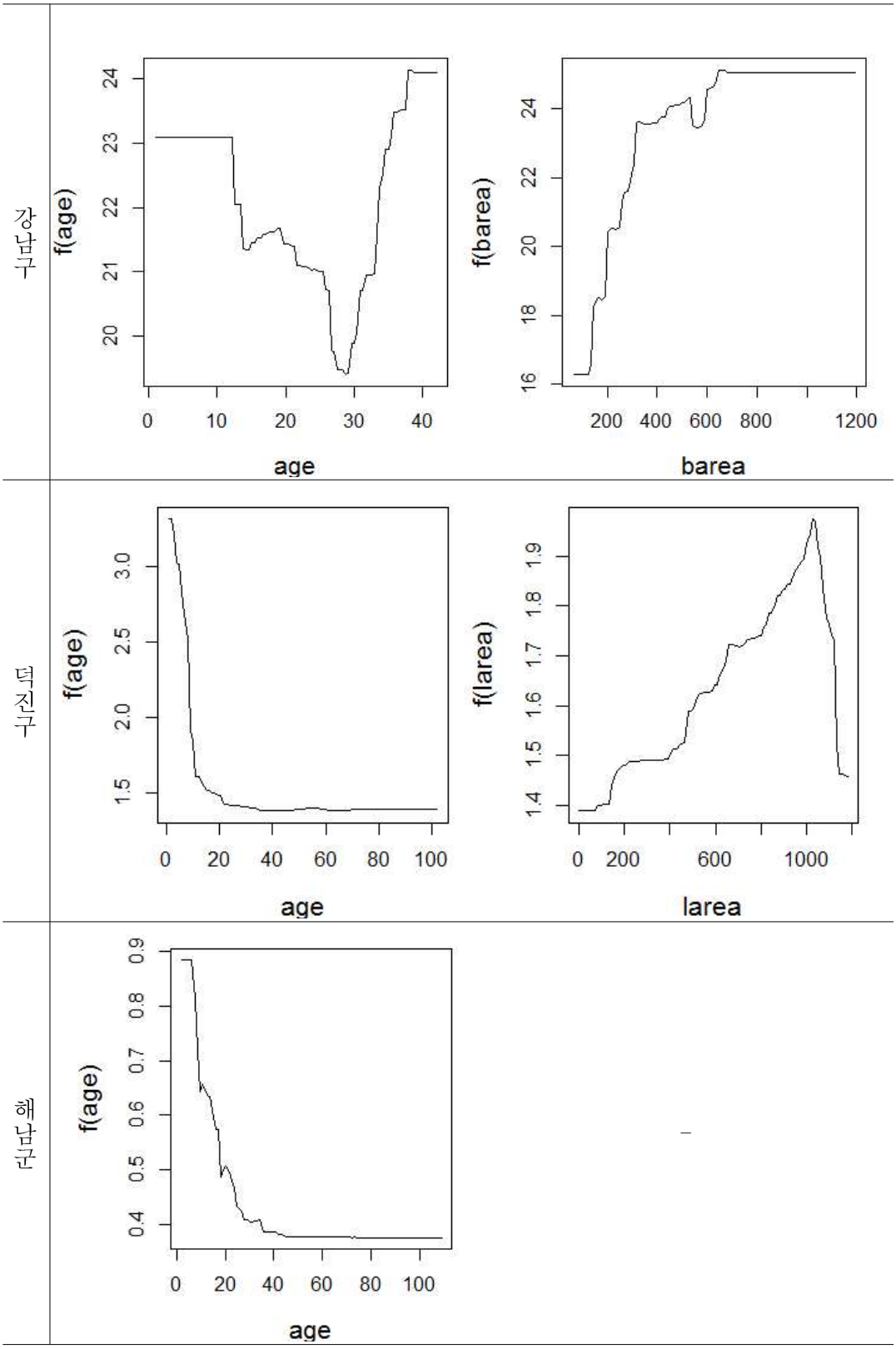
32) 오차제곱(Squared Error) 최소화 기여분을 계량화한 것으로 전체 합이 100이 되도록 조정된 수치[자세한 사항은 Friedman(2001) 참조].

[표 3-19] 설명변수의 상대적 영향력

설명변수		상대적 영향력
강남구	**토지면적**	**61.57**
	건물 연면적	20.04
	경과연수	9.93
	인근지역 특징	4.84
	도로접면	2.27
	건물구조	0.99
	방위	0.36
	용도지역	0.00
	지붕구조	0.00
덕진구	**건물 연면적**	**73.18**
	경과연수	18.99
	토지면적	4.29
	인근지역 특징	0.97
	도로접면	0.93
	용도지역	0.75
	건물구조	0.56
	거래연도	0.28
	경사도	0.04
	방위	0.01
	공가	0.01
해남군	**건물구조**	**31.35**
	건물 연면적	**25.09**
	경과연수	19.76
	용도지역	12.85
	토지면적	9.18
	도로접면	1.21
	방위	0.56

[그림 3-9]는 이전 모형들에서 뚜렷한 비선형 관계가 있는 것으로 파악된 설명변수들의 *Partial Dependence Plot*[33]이다. 강남구는 경과연수 30년 이후에는 주택가격이 다시 상승하고, 건물 연면적은 대략 400㎡을 초과할 경우 주태가격 상승에 미치는 영향력이 미미하다. 덕진구는 경과연수 20년까지 지속적으로 주택가격이 하락하고, 토지면적이 1,000㎡를 상회할 경우 전·답 등 타 용도 토지의 추가로 주택가격이 오히려 하락하고 있다. 마지막으로 해남군은 경과연수 40년까지 주택가격이 하락하는 등 GAM 모형과 유사한 결과가 산출되었다.

33) 다른 설명변수들을 통제한 상태에서 해당 설명변수가 종속변수에 미치는 효과(Marginal Effect)를 보여주는 그래프.



[그림 3-9] 지역별 비선형 변수의 패턴(Boosting 모형)

[표 3-20]은 부스팅 모형의 성능을 보여준다. 가격 균형성 측면에서 OLS 모형보다 대부분 개선되었으나 해남군은 OLS 모형과 비슷한 수준의 성능임을 알 수 있다.

[표 3-20] 부스팅 모형 성능

지역	구분	평균 SR	중위수 SR	RMSE	MAE	COD
강남구	모형 데이터	1.07	1.03	5.35	3.86	19.19
	검증 데이터	**1.18**	**1.07**	**5.39**	**4.17**	**27.24**
덕진구	모형 데이터	1.19	1.07	0.47	0.34	32.03
	검증 데이터	**1.18**	**1.04**	**0.59**	**0.37**	**34.35**
해남군	모형 데이터	1.63	1.18	0.27	0.19	71.55
	검증 데이터	**1.90**	**1.19**	**0.33**	**0.21**	**91.73**

4. MARS

앞서 설명한 랜덤 포리스트나 부스팅은 모두 트리를 기초로 한 방법이다. 이와 같이 트리에 기반한 모형의 함수를 시각적으로 표현하면 불연속적인 형태로 표현된다. 본 연구에서처럼 종속변수가 범주형이 아닌 연속형인 경우 함수의 불연속성은 모형의 단점이 될 가능성이 높다. 연속형 종속변수에 대해서는 대개 매끄러운 연속 함수를 가정하는 것이 일반적이기 때문이다.

MARS는 트리에 기반한 모형과 달리, 함수를 연속적인 형태로 표현하므로 연속형 종속변수의 예측에 보다 효율적이다. 또한 MARS는 GCV(Generalized Cross-Validation)를 최소화하는 과정에서 자연스럽게 중요한 설명변수를 선별해 주는 이점이 있다. 이때 선별되는 변수는 상호작용항을 포함하는 것으로, 다양한 상호작용항을 연구자가 일일이 설정하지 않더라도 자동적으로 선별해 주기 때문에 MARS는 종속변수와 설명변수 간 비선형 관계를 파악하는데 매우 효율적인 도구라 할 수 있다.

[표 3-21]은 이러한 MARS의 적합 결과를 보여준다. 강남구의 경우 경과연수는 29년을 기준으로, 건물 연면적은 365.79㎡를 기준으로 주택가격과의 관계가 변하고 있음을 확인할 수 있다. 이는 앞 절에서의 분석(경과연수 30년, 건물 연면적 400㎡ 기준)과 일치하는 것이다.

덕진구의 경우에도 경과연수는 17년을 기준으로, 토지면적은 1,038㎡를 기준으로 주택가격과의 관계가 달라지고 있다. 이 역시 앞 절에서의 분석(경과연수 20년, 토지면적 1,000㎡ 기준)과 일맥상통한다.

마지막으로 해남군의 경우 경과연수 18년을 기준으로 주택가격과의 관계가 변하고 있다. 이는 앞 절에서 해남군의 경우 경과연수 40년까지 주택가격이 지속적으로 하락한다는 분석과 차이가 있다. 그러나 MARS에서 경과연수는 해당 변수 자체뿐 아니라 용도지역, 방위 및 건물 연면적과의 상호작용항으로도 표현되고 있어 단순히 경첩함수의 변곡점 위치를 가지고 이전 모형과 비교할 수는 없다.

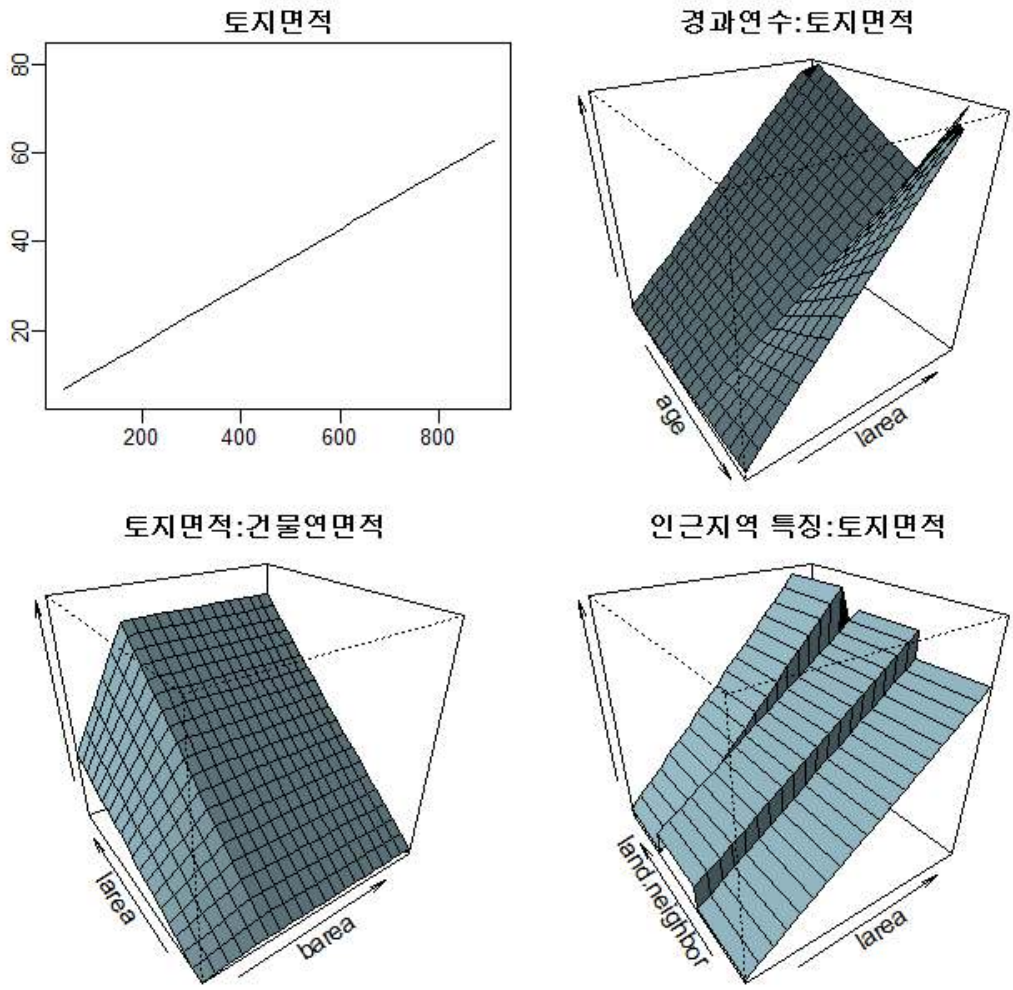
[그림 3-10]은 강남구의 MARS 적합 과정에서 나타난 주요 비선형 변수의 패턴을 시각적으로 표현한 것이다. 그림에서처럼 토지면적은 선형의 형태로 표현하는 것이 합리적이다. 하지만 경과연수는 약 29년을 기준으로 하여 이전까지는 주택가격을 하락시키다 이후에는 다시 상승시키고 있다. 건물 연면적 또한 365.79㎡까지만 주택가격에 양(+)의 영향을 미치고 그 이상의 초과 면적은 주택가격에 별다른 영향을 주지 못하고 있다. 이러한 현상은 이전 모형에서도 밝혀진 사항들이다.

MARS가 밝혀낸 새로운 사실은 인근지역 특징 변수가 주택가격에 미치는 영향이다. 인근지역 특징의 마지막 세 번째 범주(주상용 지대)의 경우 토지 면적이 증가할수록 주택가격에 보다 커다란 양(+)의 영향을 미치고 있다. 반면 첫 번째 범주(주거지대)와 두 번째 범주(상업지대)는 가격 수준의 차이만 다른 뿐 기울기는 유사하다. 이는 강남구 주택시장의 경우, 주거지대나 상업지대에 위치한 주택은 토지 규모에 관계 없이 그 공급량이 일정하게 유지되고 있으나 주상용 지대에 위치한 주택, 특히 토지 규모가 어느 정도 이상인 주택은 매우 희소한 것으로 해석할 수 있다.

[표 3-21] MARS 적합 결과

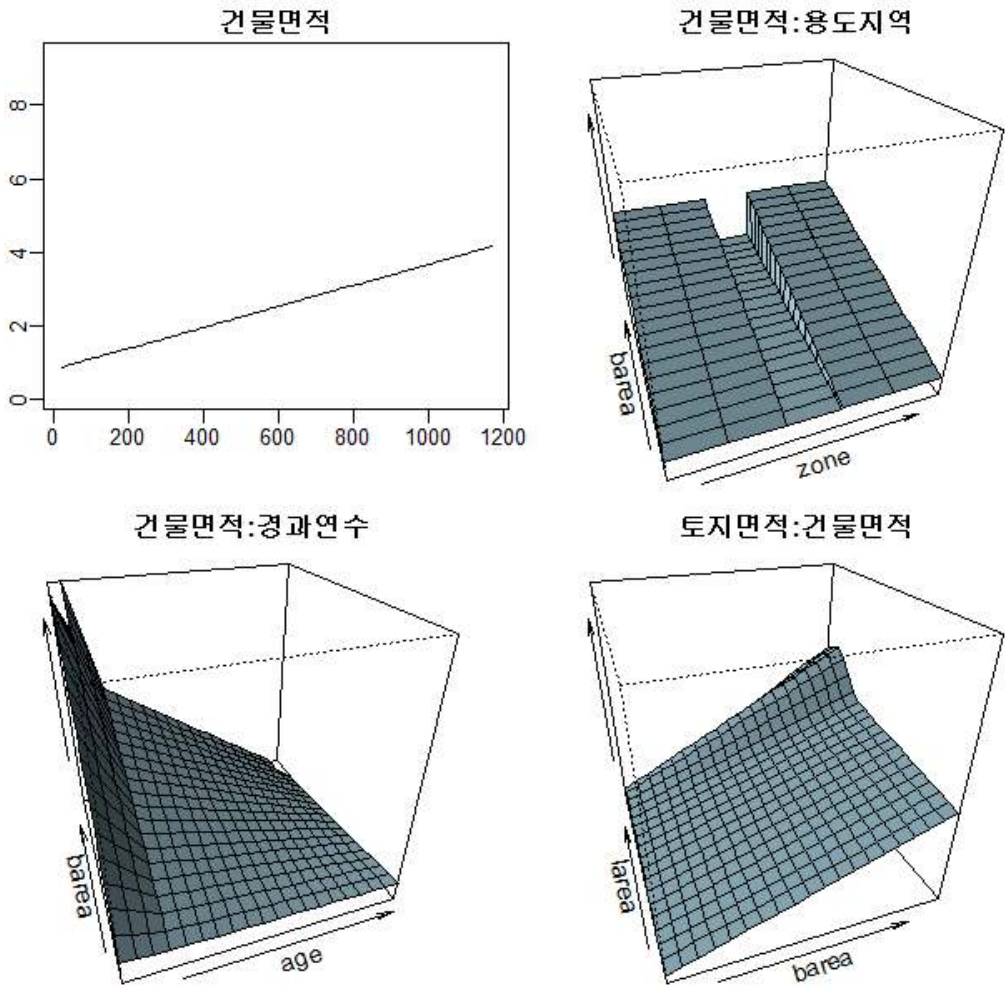
설명변수		계수
강남구 (GCV 32.57)	상수항	4.0344
	토지면적	0.0690
	인근지역 특징[상업지대]	8.4797
	토지면적 x 인근지역 특징[주상지대]	0.0274
	h*(29-경과연수) x 토지면적	0.0009
	h(경과연수-29) x 토지면적	0.0034
	h(365.79-건물 연면적) x 토지면적	-0.0001
덕진구 (GCV 0.26)	상수항	1.4629
	건물 연면적	0.0037
	인근지역 특징[상업지대]	0.5648
	인근지역 특징[기타지대]	-1.0464
	h(1,038-토지면적)	-0.0010
	h(토지면적-1,038)	-0.0099
	용도지역[공업지역] x 건물 연면적	-0.0018
	h(17-경과연수) x 건물 연면적	0.0003
	h(경과연수-17) x 건물 연면적	-0.0001
	건물 연면적 x h(토지면적-843)	0.0000
	h(1,038-토지면적) x 인근지역 특징[주상지대]	0.0005
	h(981-토지면적) x 인근지역 특징[기타지대]	0.0013
	h(토지면적-981) x 인근지역 특징[기타지대]	0.0060
해남군 (GCV 0.09)	상수항	0.2915
	용도지역[녹지지역]	-0.3791
	용도지역[농림지역]	-0.3787
	건물 연면적	0.0049
	건물구조[블럭조]	-0.1424
	h(18-경과연수)	-0.0300
	h(경과연수-18)	-0.0035
	용도지역[관리지역] x 건물 연면적	-0.0047
	용도지역[관리지역] x 토지면적	0.0001
	용도지역[자연환경보전지역] x 건물 연면적	-0.0050
	건물 연면적 x 건물구조[기타구조]	0.0035
	용도지역[상업지역] x h(18-경과연수)	-0.1429
	방위[남향 외] x h(18-경과연수)	-0.0272
	h(18-경과연수) x 건물 연면적	0.0009

* Hinge Function



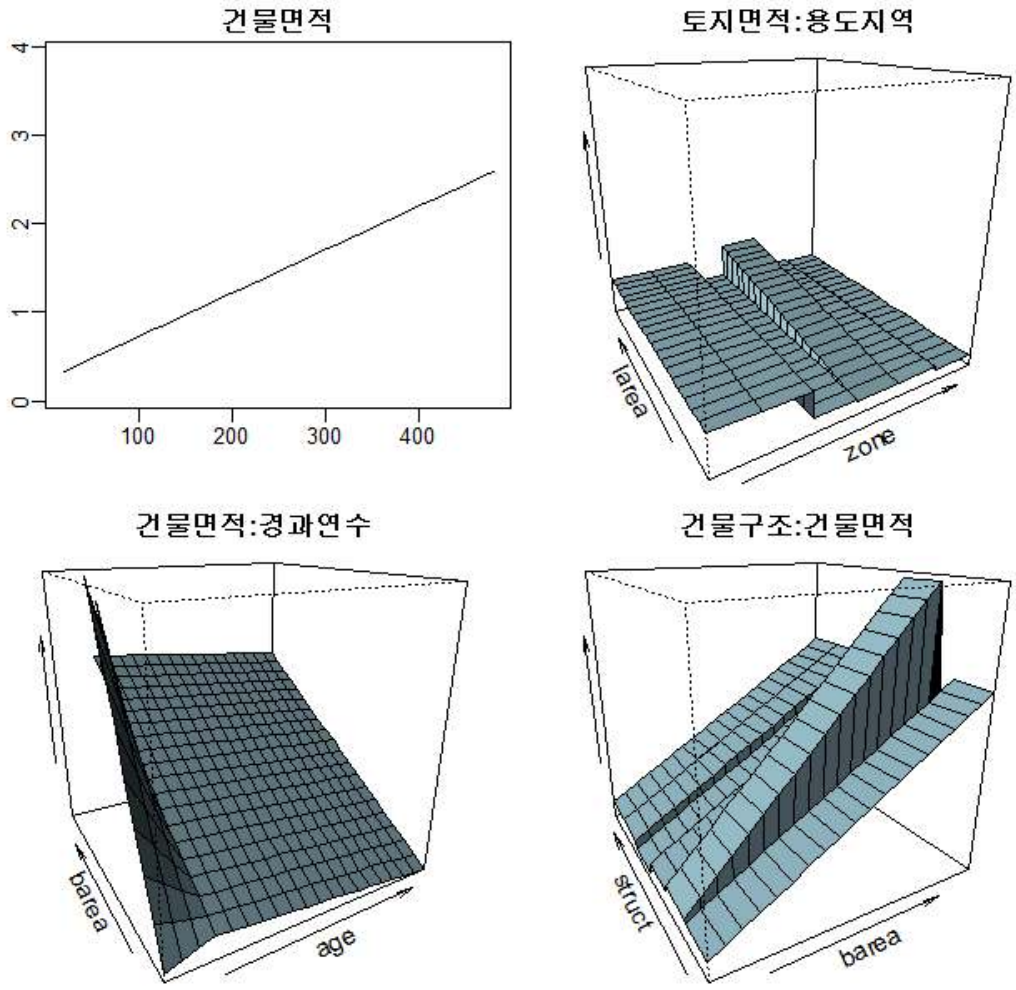
[그림 3-10] 강남구 비선형 변수의 패턴

[그림 3-11]은 덕진구의 비선형 변수 패턴을 보여준다. 건물 면적은 선형으로 표현하는 것이 합리적임을 알 수 있다. 그러나 용도지역의 경우 건물면적 증가에 따라 세부 용도지역별로 가격상승 기울기가 상이함을 알 수 있다. 특히 좌측으로부터 세 번째 용도지역('공업지역')은 건물면적 증가에 따른 주택가격 상승이 거의 없음을 알 수 있다. 경과연수는 17년까지 가격이 급속하게 하락하다가 이후 평탄해지는 것을 확인할 수 있다. 토지면적의 경우 1,038㎡을 초과하는 경우 주택가격이 하락하는 경향으로 반전되는 것을 알 수 있다.



[그림 3-11] 덕진구 비선형 변수의 패턴

[그림 3-12]는 해남군의 비선형 변수 패턴을 보여준다. 건물 면적은 선형으로 표현하는 것이 합리적임을 알 수 있다. 그러나 용도지역의 경우 토지면적 증가에 따라 세부 용도지역별로 가격상승 기울기가 상이함을 알 수 있다. 특히 좌측으로부터 네 번째 용도지역('관리지역')은 토지면적 증가에 따라 주택가격이 매우 가파르게 상승함을 알 수 있다. 경과연수는 18년까지 가격이 급속하게 하락하다가 이후 평탄해지는 것을 확인할 수 있다. 또한 건물구조의 경우 두 번째 범주('기타구조')는 건물면적이 증가할수록 주택가격이 매우 가파르게 상승하는 것을 알 수 있다. 해남군의 경우 기타구조로 분류된 건물구조는 주로 석조, 흙벽돌조 등으로 건축면적이 커질 경우 단가가 체감하기보다는 체증하는 경향이 강한 구조들이라 할 수 있다.



[그림 3-12] 해남군 비선형 변수의 패턴

이와 같이 상호작용 효과와 인근지역 특징 변수를 추가적으로 고려함에 따라 모형의 성능은 [표 3-22]에서 보는 바와 같이 강남구와 덕진구는 OLS 모형 대비 현격하게 향상되었고, 해남군은 OLS 모형과 유사한 수준의 성능을 보이고 있다.

[표 3-22] MARS 성능

지역	구분	평균 SR	중위수 SR	RMSE	MAE	COD
강남구	모형 데이터	1.07	1.05	5.41	4.23	20.67
	검증 데이터	**1.12**	**1.06**	**5.22**	**4.12**	**23.07**
덕진구	모형 데이터	1.20	1.06	0.50	0.36	34.79
	검증 데이터	**1.17**	**1.05**	**0.55**	**0.35**	**32.74**
해남군	모형 데이터	1.70	1.18	0.28	0.20	79.37
	검증 데이터	**1.90**	**1.27**	**0.40**	**0.24**	**87.54**

5. SVM(Support Vector Machine)

SVM은 분류(Classification)의 문제를 다루기 위해 1990년대에 제시된 (Vapnik, 1996) 기계학습 알고리즘이지만, 연속형 종속변수의 예측에도 동일하게 적용할 수 있다. 본 연구에서는 SVM 모형에 [표 3-23]과 같은 파라미터를 적용하였다.

[표 3-23] SVM 모형 상세

구 분	강남구	덕진구	해남군
SVM Type	ϵ-regression[34]	ϵ-regression	ϵ-regression
Kernel Function	Radial Basis	Radial Basis	Linear
Cost Parameter	2.35	17.5	6.25
# of Support Vectors	251	912	350

34) 연속형 종속변수를 다루는 Regression 맥락에서 흔히 사용되는 SVM은 ϵ-regression과 ν-regression이다. 두 가지 유형의 SVM은 패널티 파라미터로 ϵ 또는 ν를 사용하는 정도의 차이만 있다. 결과도 대부분 유사한 편이어서 어떠한 유형의 SVM을 사용하였는지는 그리 중요한 사안이 아니다. 자세한 사항은 Chang & Lin(2001) 참조.

커널 함수의 경우 Linear, Polynomial 및 Radial Basis를 모두 적용하였고, 모형의 성능이 가장 우수하게 나타나는 커널함수를 지역별로 각각 선택하였다. 즉 강남구 및 덕진구는 Radial Basis, 해남군은 Linear 커널 함수를 사용하는 것이 모형 적합도 측면에서 가장 우수하였다.

가장 중요한 cost 파라미터는 다음과 같은 유사 결정계수(pseudo R^2) 값이 최대가 될 때의 값으로 설정하였다(Smola & Schölkopf, 2004; Rakotomalala, 2005).

$$pseudo\,R^2 = 1 - \frac{RSS_{model}}{RSS_{null}} \tag{3-2}$$

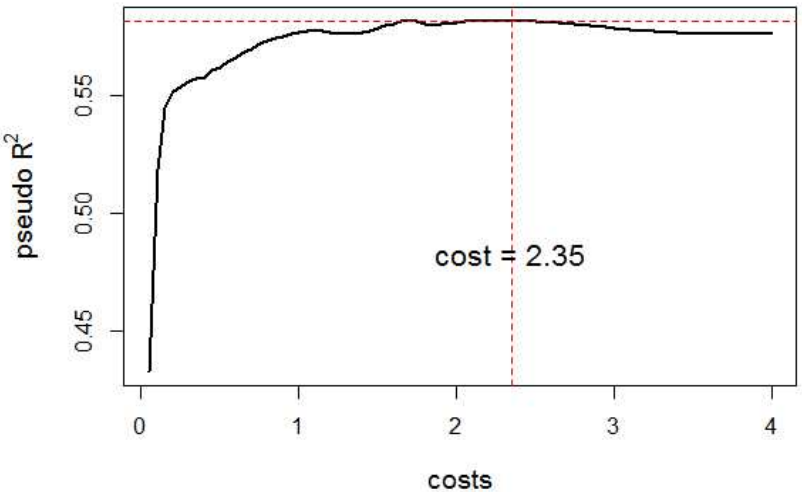
위 식에서 RSS_{null}은 아래와 같다.

$$RSS_{null} = \sum_{i=1}^{N} \left(y_i - \bar{y}_{train}\right)^2 \tag{3-3}$$

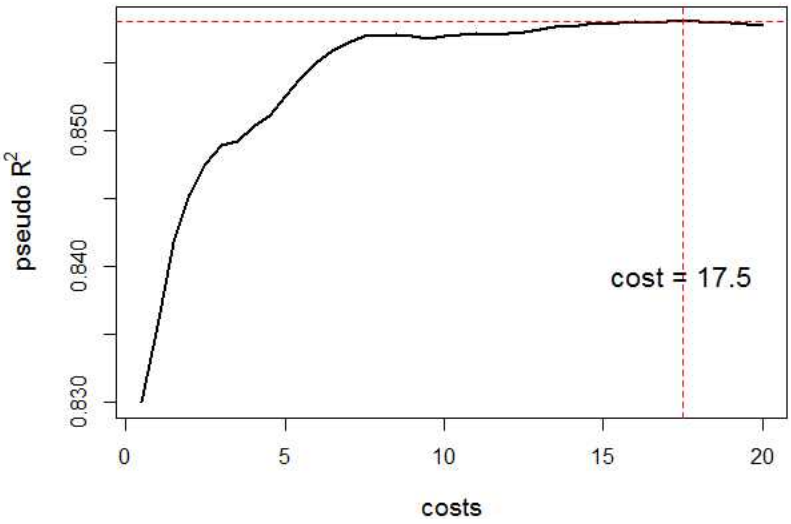
즉 검증 데이터 종속변수 값과 모형 데이터 종속변수 평균값(Mean)의 차이를 제곱하여 합산한 수치를 나타낸다. RSS_{model}은 cost 값 변화에 따른 해당 SVM 모형의 RSS를 의미한다.

[그림 3-13]은 이러한 유사 결정계수를 기준하여 선택된 cost 값을 보여준다. 예를 들어 강남구는 cost = 2.35일 때 유사 결정계수값이 최대(0.581)가 됨을 알 수 있다. 덕진구는 17.5, 해남군은 6.25로 cost 값이 결정되었는 바, 데이터에 따라 상당한 차이가 있음을 알 수 있다. cost 값이 크다는 것은 많은 관찰치를 서포트 벡터(Support Vector)로 선택할 수 있음을 의미한다. 마찬가지 의미로 cost 값이 작다는 것은 서포트 벡터의 역할을 하는 관찰치가 적음을 의미한다.

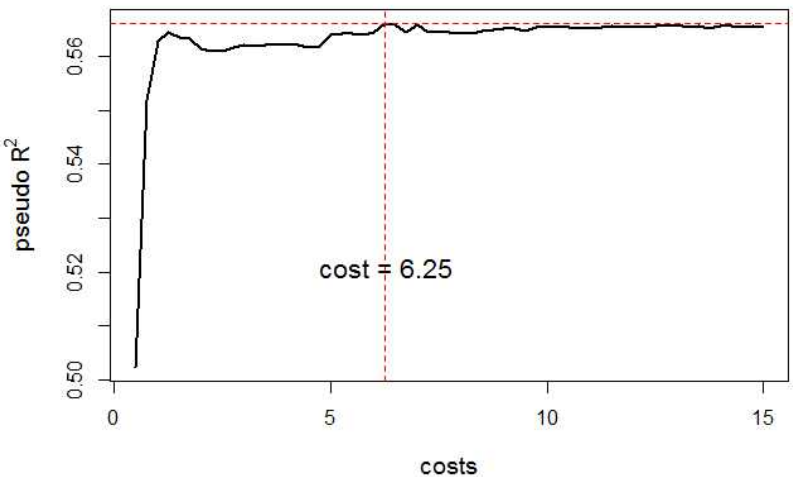
(강남구)



(덕진구)



(해남군)



[그림 3-13] 적정 cost 값의 결정

이와 같은 SVM의 성능은 [표 3-24]와 같다. OLS 모형과 비교하였을 때 덕진구는 뚜렷한 성능 개선을 보이고 있다. 반면 강남구나 해남군은 OLS 모형 대비 근소한 성능 개선을 보여주고 있다.

[표 3-24] SVM 성능

지역	구분	평균 SR	중위수 SR	RMSE	MAE	COD
강남구	모형 데이터	1.03	1.00	5.57	3.88	19.78
	검증 데이터	**1.14**	**1.01**	**5.41**	**4.19**	**29.77**
덕진구	모형 데이터	1.07	1.02	0.42	0.28	24.98
	검증 데이터	**1.09**	**1.00**	**0.56**	**0.36**	**32.38**
해남군	모형 데이터	1.37	1.01	0.34	0.21	75.19
	검증 데이터	**1.45**	**1.03**	**0.37**	**0.21**	**76.82**

제 4 절 모형 성능의 비교

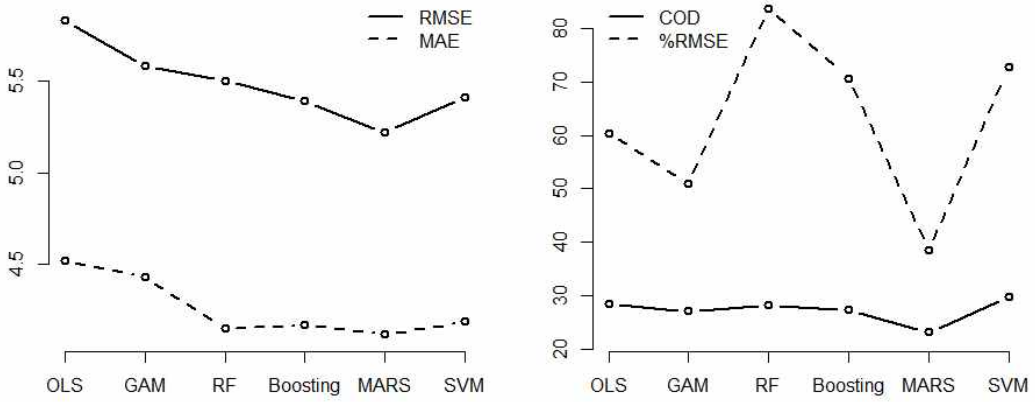
1. 지역 간 모형 성능의 비교

본 절에서는 가격 정확성 지표인 SR보다는 가격 형평성 지표인 RMSE, MAE 및 COD에 중점을 두어 모형 간 성능을 비교한다. 일반적으로 잔차의 합은 0이 되듯이 SR의 평균값(또는 중위수)도 대부분 1.00에 수렴되는 특징이 있어 모형 간 성능 비교에 큰 의미가 없는 경우가 많다. 반면 RMSE, MAE 및 COD는 일종의 모형 적합도를 보여주는 지표로서 성능 비교의 기준으로 유용한 편이다.

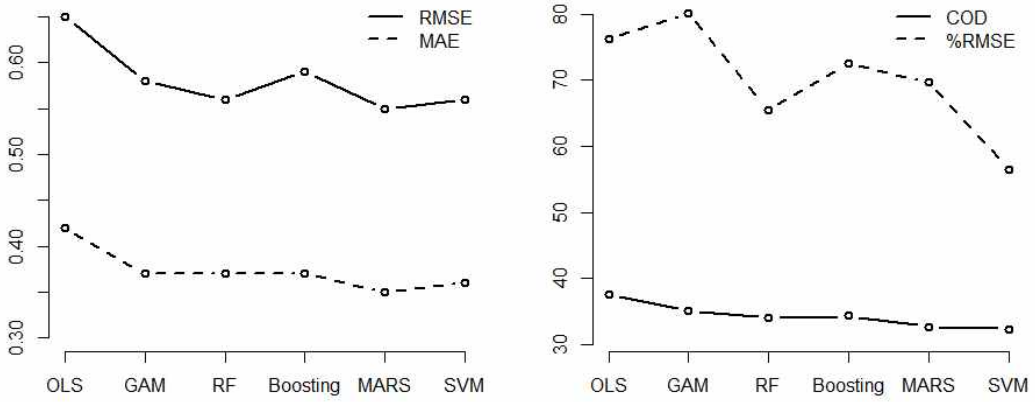
먼저 지역 간 비교에서는 RMSE 및 MAE를 비교 지표로 사용하지 않는다. 앞서 설명하였듯 RMSE 및 MAE는 주택가격의 절대적 크기에 영향을 받기 때문이다. 대신 이러한 단위에 영향을 받지 않는 지표인 COD, 그리고 RMSE를 변형한 퍼센트 RMSE(%RMSE)[35]를 비교 지표로 하여 검토한다.

35) $\%RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\hat{y_i} - y_i}{y_i}\right)^2}$

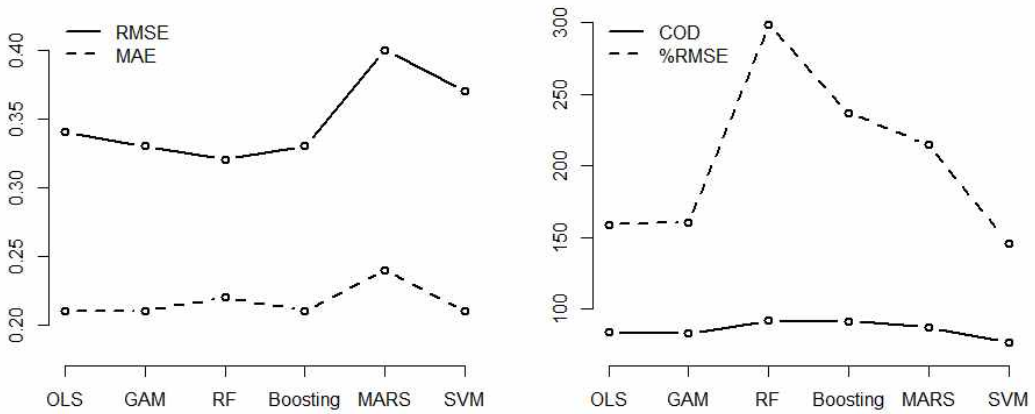
[그림 3-14]~[그림 3-16]은 검증 데이터를 기준으로 계산한 지역 및 모형별 RMSE, MAE, COD 및 %RMSE 현황을 보여준다.



[그림 3-14] 서울시 강남구 모형 성능 비교



[그림 3-15] 전주시 덕진구 모형 성능 비교



[그림 3-16] 전라남도 해남군 모형 성능 비교

COD를 기준으로 지역 간 모형 성능을 살펴보면, 사용한 모형에 관계없이 강남구는 COD가 약 23~30의 범위에 있고, 덕진구가 30~40, 해남군이 75~95 수준에 있음을 알 수 있다. %RMSE도 강남구는 대략 60 내외, 덕진구 70 내외, 그리고 해남군이 200 내외 수준에 있다.

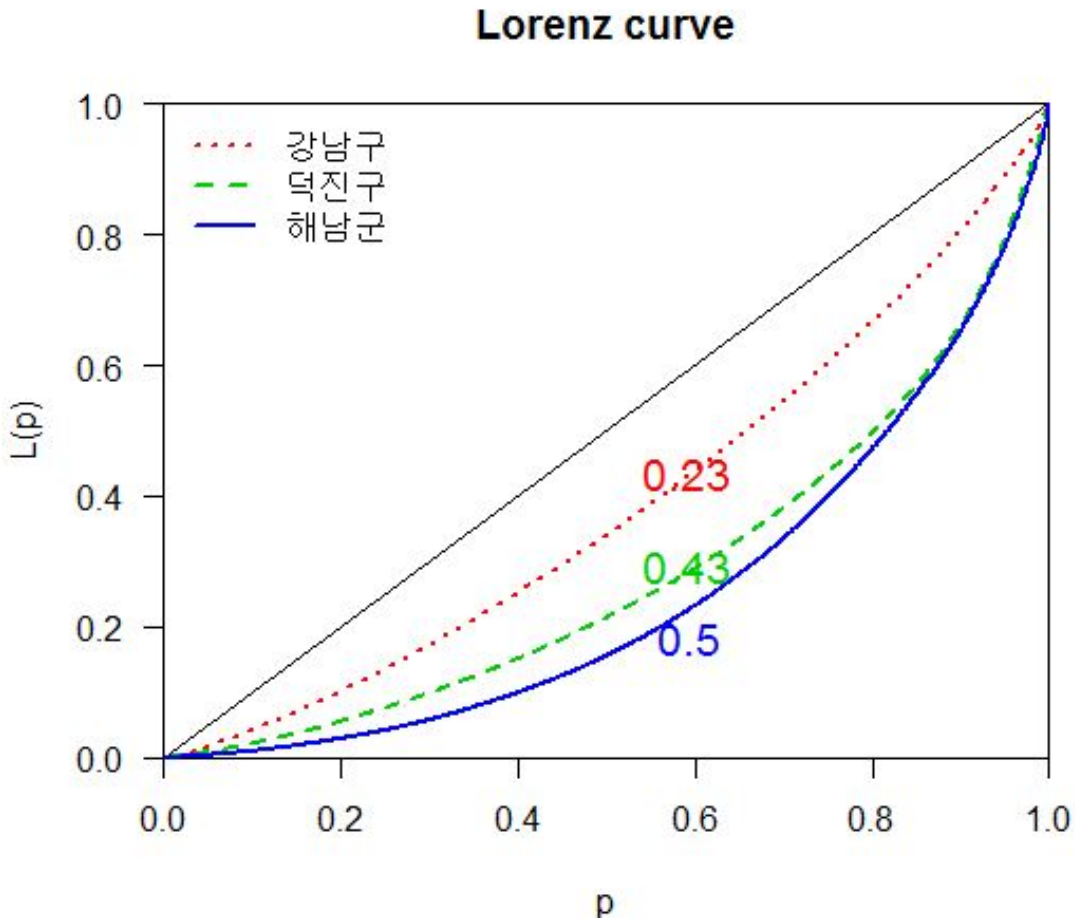
이는 단독주택의 경우 농촌지역일수록 가격을 모형화하기 어려움을 의미한다. 강남구와 같은 대도시는 지역의 면적 자체가 작을 뿐 아니라 지역 내 주택 특징이 비교적 균일하여 모형을 통한 가격 예측이 상대적으로 수월하다. 예를 들어 강남구에 소재하는 주택은 대부분 1970년대~80년대에 택지개발사업의 일환으로 지어진 철근콘크리트조 내지 연와조 2~4층 규모의 건물이다. 토지 또한 대부분 주거지역에 속하고, 잘 정비된 가로망을 따라 바둑판 형태로 조밀하게 위치하고 있다. 이처럼 비교적 동질적인 주택 집단이 존재하는 강남구는 가격 예측이 수월할 수밖에 없다.

반면 덕진구 및 해남군으로 갈수록 COD 및 %RMSE 지표가 급격하게 악화됨을 알 수 있는데, 이는 농촌지역일수록 이질적인 주택이 많아 가격을 특정한 모형에 의해 일괄적으로 예측하기 어려움을 의미한다. 예를 들어 해남군에 소재하는 주택은 개화기 때(1905년) 지어진 목조 건물에서부터 최근에 지어진 철근콘크리트조의 고급 별장까지 신축 연대와 건물 구조가 매우 다양하다. 토지 또한 주거지역과 같은 도시지역 뿐 아니라 관리지역, 농림지역 그리고 자연환경보전지역에 이르기까지 매우 다양한 용도지역에 걸쳐 분포하고 있다.

[그림 3-17]은 동일한 축척(1:10,000)으로 표현된 강남구와 해남군의 주택 분포 경관을 보여준다. 강남구는 규칙적인 간선도로와 이면도로망을 따라 사각형 형태의 주택이 조밀하게 분포하고 있다. 반면 해남군은 농지와 임지 사이에 간헐적으로 주택이 분포함을 알 수 있다. 특히 해남군 지도에서 서측에 표시된 주택은 해안마을, 북측에 표시된 주택은 산간마을, 그리고 동측에 표시된 주택은 농촌마을의 특징을 갖는 등 인근 지역의 특징도 서로 상이한 것을 알 수 있다.

[그림 3-17] 강남구(위) 및 해남군(아래) 주택의 분포 경관(축척: 1:10,000)
* 출처: map.naver.com

강남구보다는 덕진구가, 덕진구보다는 해남군의 주택가격 이질성이 크다는 사실은 [그림 3-18]의 로렌츠 곡선(Lorenz Curve)을 통해서도 쉽게 확인할 수 있다.



[그림 3-18] 주택가격에 대한 로렌츠 곡선(숫자는 지니계수)

상기와 같은 이유들로 인해 대도시를 대상으로 한 모형의 성능이 가장 우수하게 나왔고, 중소도시 및 군지역으로 갈수록 모형의 성능이 좋지 않게 나온 것으로 풀이된다.

모형 간 비교 결과는 다음과 같다. 강남구는 세 가지 지표(RMSE, MAE, COD) 전부에서 MARS의 성능이 가장 우수한 것으로 나타났다. 덕진구는 세 가지 지표 중 두 가지 지표(RMSE, MAE)에서 MARS 성능이 가장 우수한 것으로 산출되었다. 마지막으로 해남군은 세 가지 지표 중 두 가지 지표(MAE, COD)에서 SVM 성능이 가장 우수한 것으로 산출되었다.

즉 OLS, GAM 및 트리 기반 모형(랜덤 포리스트, 부스팅)보다는

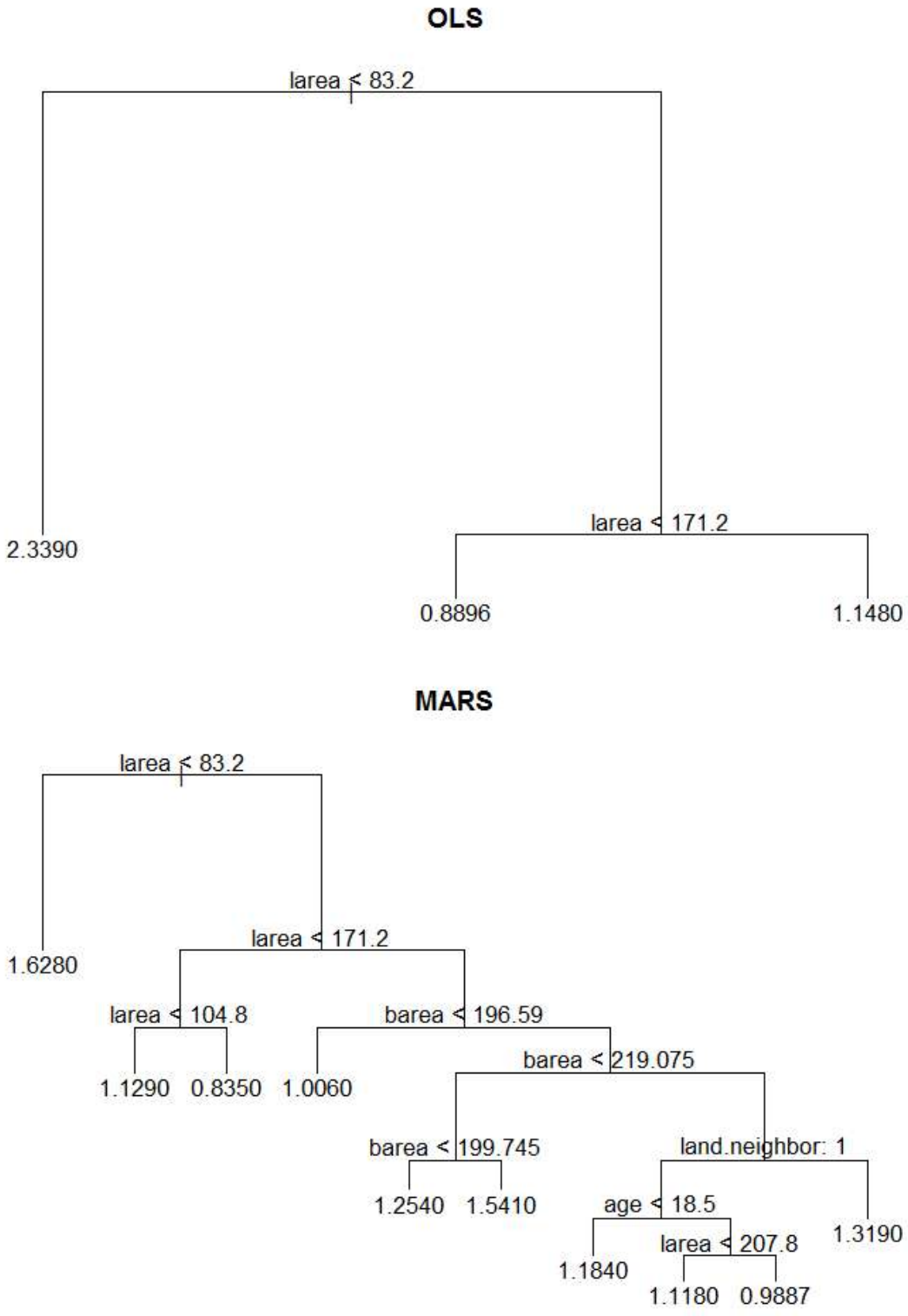
MARS 내지 SVM 같은 최근에 개발된 모형의 성능이 전반적으로 우수하게 나타났다. 랜덤 포리스트나 부스팅 같은 트리 기반 모형은 연속형 종속변수에 적용한 실례가 많고, 랜덤 포리스트가 부스팅보다 그 성능이 우월하다는 등(Abbott, 2014, p.321) 비교 연구도 풍부한 편이다. 그러나 MARS나 SVM은 비교적 최근에 제시된 방법으로 부동산 가격추정에 적용된 예가 없는 바, 이러한 방법들의 적극적인 활용이 필요해 보인다.

2. 지역 내 모형 성능의 국지적 비교(Local Approach)

앞 절에서의 분석이 전역적 분석이었다면 본 절에서는 국지적 진단방법을 사용하여 지역 내 모형의 성능을 세부적으로 검토한다. 즉 준거가 되는 모형(선형회귀모형)과 가장 성능이 우수한 것으로 나타난 모형(강남구 및 덕진구: MARS, 해남군: SVM)을 중심으로 검토하되, 선형회귀모형의 취약점이 어떻게 개선되었는지 살펴본다. 모형 성능에 대한 이러한 국지적 진단은 전역적 진단에 그친 대부분의 선행연구와 차별되는 점이기도 하다.

국지적 진단은 제2장에서 설명하였듯, SR을 종속변수로 정한 후 해당 모형에 동원된 공변량을 설명변수로 하여 회귀트리(Regression-tree) 알고리즘을 적용한다. 연속형 종속변수를 기준으로 노드를 분할하는 이 알고리즘은 특정 노드에서 하위 노드로 이진 분할을 반복적으로 수행하며, 분산의 감소량을 기준으로 트리를 생성하게 된다.

[그림 3-19]는 검증 데이터를 기준한 강남구의 국지적 진단 결과를 보여준다. 강남구의 경우 전반적으로 모형 성능은 우수한 편이나(COD 30.0 이내) [그림 3-19] 상단의 OLS 결과를 보면 토지면적(*larea*) 83.2㎡ 이하의 소규모 주택을 과다평가하고 있음을 알 수 있다(평균 2.3배 과다평가). 그러나 MARS는 토지면적 83.2㎡ 이하의 소규모 주택에 대해서 뚜렷한 개선 실적을 보이고 있는 바, 여전히 과다평가하는 경향이 남아 있기는 하나 그 편의를 상당히 축소하였다(평균 1.6배 과다평가). 소규모 토지면적 이외에 심각한 추정 편의를 보이는 부분은 없는 것으로 판단된다.



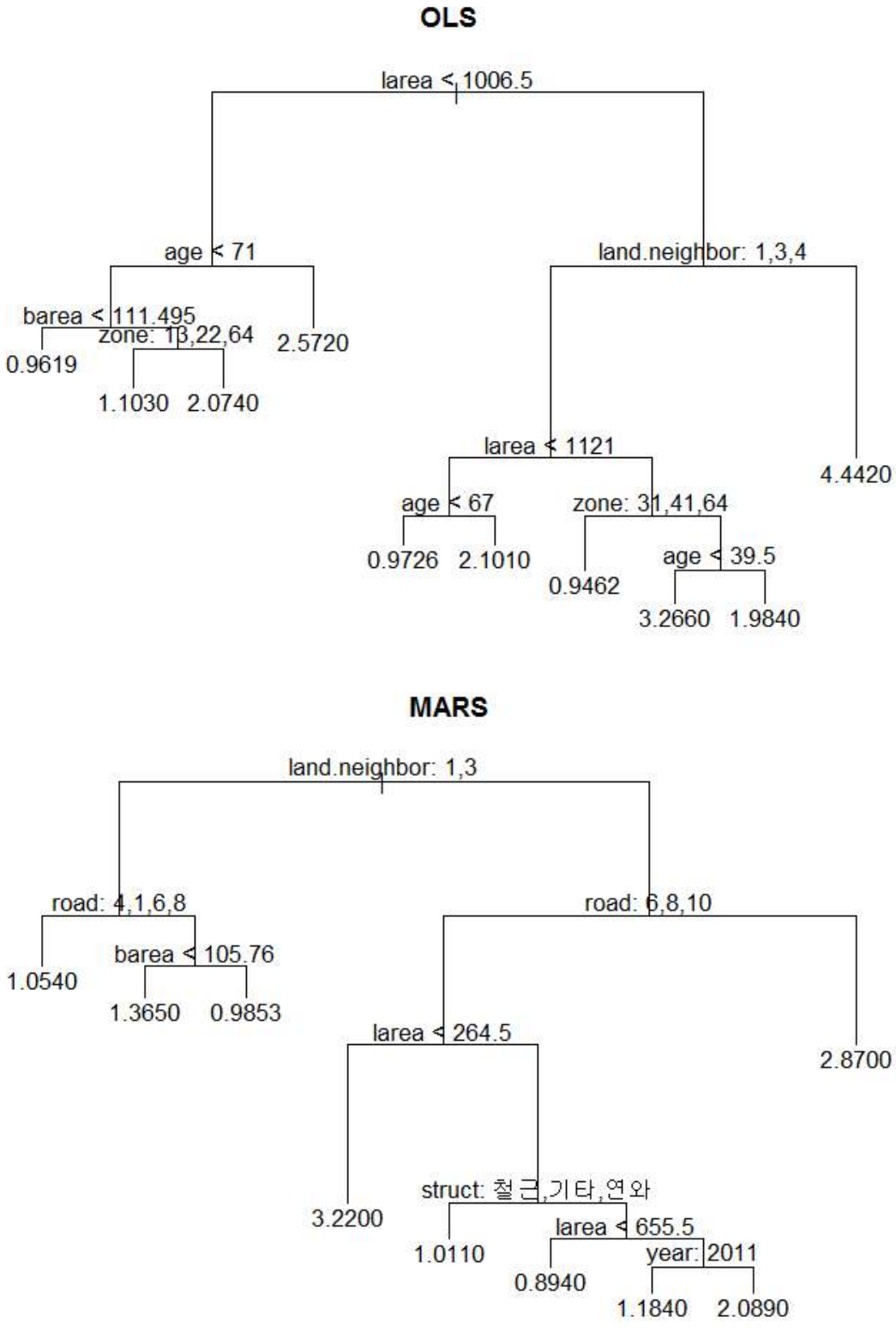
[그림 3-19] 강남구 국지적 진단 결과

[그림 3-20]은 덕진구의 국지적 진단 결과를 보여 준다. 상단의 OLS 모형을 보면 맨 우측 범주에서 평균 4.4배의 과다평가 경향을 보여주고 있다. 이 범주에 속하는 주택 유형은 토지면적(*larea*)이 1,006.5㎡ 이상의 대규모 주택이면서 인근지역 특징(*land.neighbor*)이 코드 2(상업지대)에 해당하는 사례들이다. 즉 상업지대에 위치한 비교적 큰 주택들을 과다평가하고 있음을 알 수 있다.

반면 [그림 3-20]의 하단 MARS 모형을 보면 토지면적을 기준한 이러한 과다평가 경향이 거의 사라졌음을 알 수 있다. 즉 MARS는 토지면적 h=1,038㎡을 기준으로 경첩함수를 설정하는 등 대규모 주택의 가격이 체감하는 현상을 적정하게 반영하여 과다평가 경향을 제거한 것으로 풀이할 수 있다.

다만 MARS의 경우 여전히 3.2배 정도의 과다평가 경향을 보여주는 범주가 존재하는데, 이 범주에 속하는 주택 유형은 인근지역 특징(*land.neighbor*)이 코드 2(상업지대) 및 4(기타지대)에 해당하면서 도로접면은 코드 6, 8, 10(소로, 세로, 세로불)이고 토지면적(*larea*)이 264.5㎡ 이하인 주택들이다. 즉 상업지대나 농경지대[36]에 도로 폭이 비교적 협소한 곳에 위치한 중소 규모의 주택이라 할 수 있다. 이러한 유형에 속하는 주택들은 누락된 설명변수가 없는지 등 추가적인 검토가 필요한 부분이라 할 수 있다.

36) 덕진구의 기타지대는 대부분 농경지대에 해당한다.



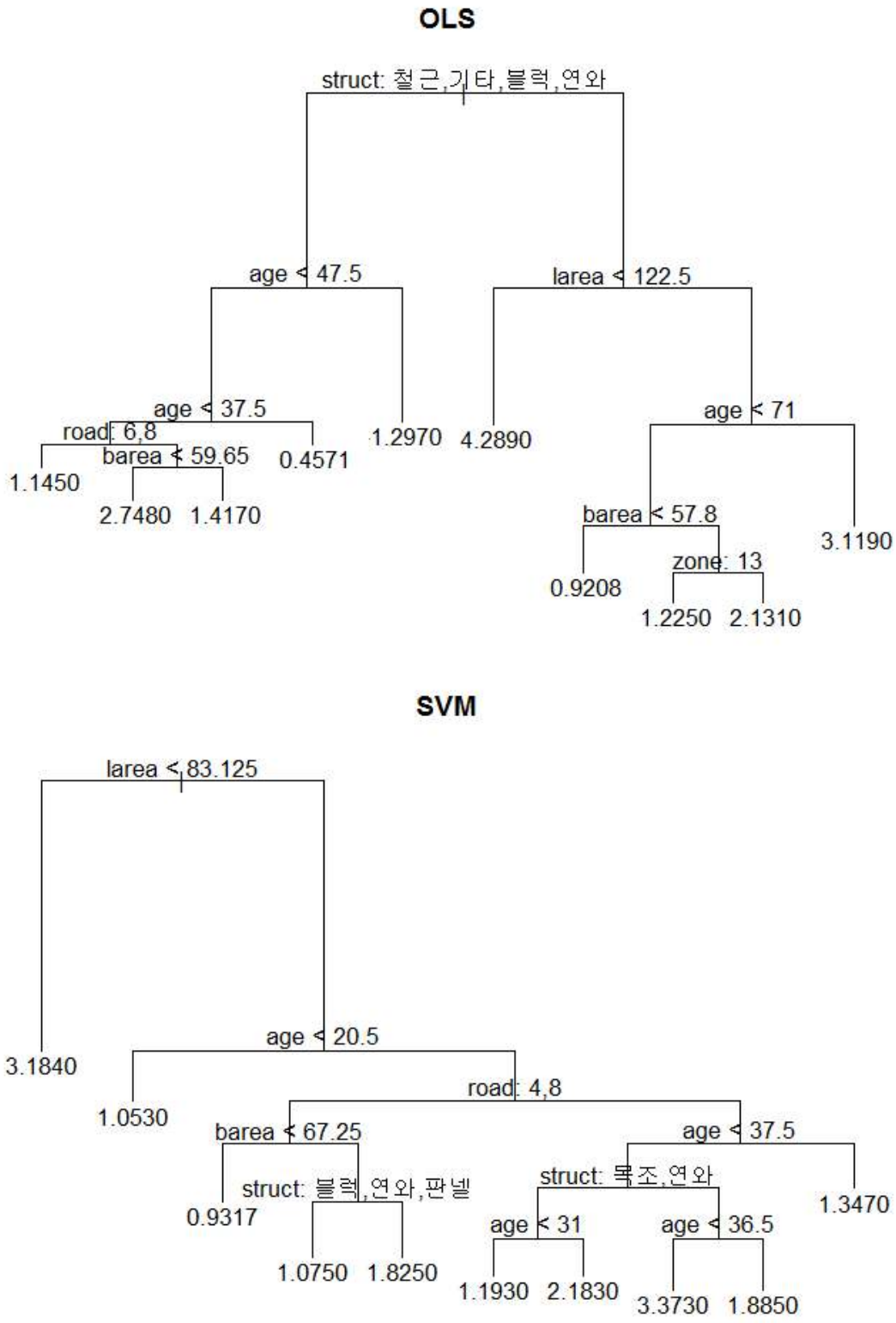
[그림 3-20] 덕진구 국지적 진단 결과

[그림 3-21]은 해남군의 국지적 진단 결과를 보여준다. 상단의 OLS 모형 결과를 보면 건물구조가 철근콘크리트조, 기타구조, 블록조 및 연와조가 아니면서 토지면적(*larea*)이 122.5㎡ 이하인 주택을 평균 4.3배 과다평가하고 있음을 알 수 있다. 해남군에서 이러한 유형에 속하는 주택은 주로 소규모 목조 주택에 해당한다. 즉 OLS 모형은 오래 전에 지어진 소규모 목조 주택에 대해 실제 거래가격 대비 약 4배 이상 과다평가하는 편의를 보이고 있다. 이러한 주택은 실제 공가일 가능성이 높다.

그러나 하단의 SVM은 과다평가될 수 있는 주택의 범위를 토지면적 122.5㎡에서 83.125㎡로 현격하게 줄였고(48개 주택에서 23개 주택으로 축소), 또한 과다평가 편의도 평균 4.3배에서 3.2배로 상당히 줄였음을 알 수 있다.

다만 SVM의 경우 여전히 3.4배 정도의 과다평가 경향을 보여주는 범주가 존재하는데, 이 범주에 속하는 주택 유형[37]은 개수가 적어(12개) 전체적인 모형 성능에 미치는 영향은 적다고 할 수 있다. 그러나 주택 수가 적다고 하여도 이러한 범주에 속하는 주택들이 왜 과다평가 경향을 보이는지 별도의 검토는 필요한 것으로 보인다.

37) 토지면적(larea) 83.125㎡ 이상, 도로접면 4(중로) 및 8(세로)에 해당하지 않고 건물구조가 목조 및 연와조가 아니면서 경과연수 20.5년 ~ 36.5년 사이인 주택



[그림 3-21] 해남군 국지적 진단 결과

본 절에서는 다양한 비선형 모형들의 성능을 비교 및 검토하였다. 지역 간 비교의 경우 강남구와 같은 대도시에서 모형의 성능이 비교적 우수하게 나타났고, 덕진구 및 해남군처럼 농촌지역으로 갈수록 모형 성능이 미흡하게 산출되었다(COD 및 %RMSE 기준). 이는 강남구는 좁은 지역에 비교적 균일한 주택 집단이 밀집 분포하고(높은 동질성), 해남군은 상이한 주택 집단이 넓은 지역에 산재하여(높은 이질성) 이러한 현상이 나타난 것으로 해석할 수 있다.

모형 간 비교에 있어서는 기계학습 분야에서 최근에 제시된 MARS나 SVM의 성능이 뛰어난 것으로 분석되어, 새로운 기법의 적극적인 확대 적용이 필요한 것으로 보인다.

모형 성능에 대한 지역 내 국지적 진단에 있어서는 OLS 모형에서 나타난 심각한 과다평가 편의를 MARS나 SVM 모형이 상당히 축소하였음을 확인할 수 있었다. 국지적 진단 결과에서 특기할만한 것은 과다평가 경향을 구분짓는 기준 변수로 토지면적이 자주 등장하였다는 점이다. 즉 3개 사례지역 모두 토지면적의 특정 값을 기준으로 과다 또는 과소평가 주택으로 대략적인 경향이 구분되었다. 토지면적은 주택의 규모와 직접적인 연관이 있고, 특히 강남구와 같은 대도시라면 주택금액(고가주택과 저가주택)과 깊은 관련이 있다. 따라서 주택가격을 예측하는 모형 구축에 있어서 주택의 규모나 금액을 기준한 자료의 층화가 선행되는 경우 모형의 예측력 개선에 큰 도움이 될 것으로 예상된다.

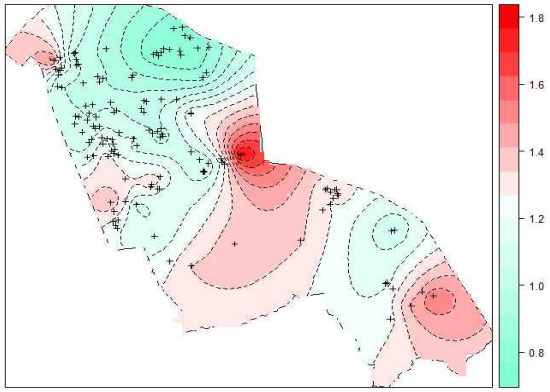
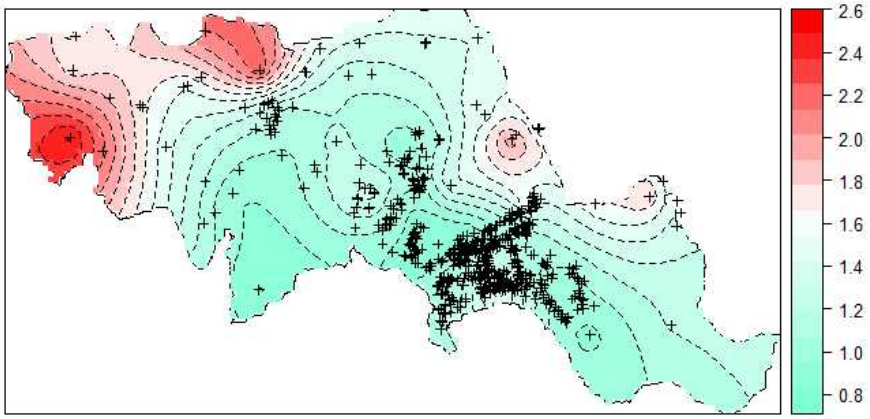
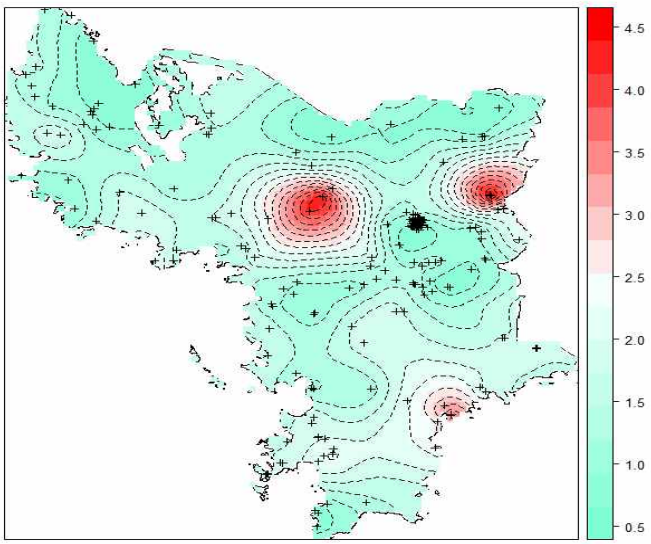
제 4 장 공간적 종속성을 반영한 비모수 모형

본 장에서는 공간사상의 가장 큰 특징인 공간적 종속성(Spatial Dependence)을 앞 장에서 적용한 비모수 모형에 반영할 수 있는 방법을 살펴보고, 이를 통해 실제 모형 성능이 개선되었는지 확인한다.

제 1 절 베리오그램을 활용한 SVM 모형의 적용

앞 장에서 설명한 비모수 모형은 주택가격과 같은 공간사상이 가지는 독특한 특징, 즉 공간적 종속성을 별도로 고려하지 않았다. 지역에 따라 공간적 종속성이 그리 강하지 않은 경우 공간효과를 추가하여 고려할 필요성은 적어진다. 복잡성을 줄이고 모형의 단순함을 유지하는 것이 오히려 더 효율적인 경우가 많기 때문이다. 그러나 공간적 종속성이 무시할 만한 수준을 넘어서는 경우 그러한 효과는 모형에 명시적으로 반영하는 것이 바람직하다.

[그림 4-1]은 지역별로 성능이 가장 우수하게 나타난 모형을 대상으로 검증 데이터를 기준한 SR 표면 및 Moran's I 값을 표시한 것이다. 지역별로 정도의 차이는 있으나 SR이 상대적으로 높은 주택(과대평가된 주택)과 낮은 주택(과소평가된 주택)이 공간상에 무작위로 분포하기보다는 특정 지역에 집중되어 있음을 알 수 있다. 즉 지역별로 성능이 가장 우수하게 나타난 모형에서도 공간적 종속성이 여전히 남아 있다고 볼 수 있다.

강남구 (MARS)	 Moran's I = 0.11 (p-value 0.02)
덕진구 (MARS)	 Moran's I = 0.02 (p-value 0.15)
해남군 (SVM)	 Moran's I = 0.24 (p-value 0.00)

[그림 4-1] 지역별 SR 표면 및 Moran's I

본 장에서는 앞 장에서 제시한 일반적인 비모수 모형에 공간적 종속성을 추가로 고려하여 모형 성능을 개선하고자 한다. 공간적 종속성을 고려할 수 있는 방법은 다양하나 본 연구에서는 첫째 SVM 모형 적합과정에서 베리오그램(Variogram)을 활용하는 방법, 둘째 공간가중행렬(Spatial Weight Matrix, **W**)을 사용하여 계산된 공간차 변수(Spatially Lagged Variable, **WY**)를 설명변수로 동원하는 방법을 제안하고자 한다.

SVM에서 가장 흔히 쓰이는 커널의 종류는 Radial Basis 함수라 할 수 있다. 본 연구의 실증분석에서도 서울 강남구 및 전주 덕진구는 Radial Basis 함수를 사용하였을 때 그 결과가 가장 양호하였는데, 식 (4-1)과 같이 표현하는 것이 일반적이다.

$$k(x, x') = \exp\left(-\gamma \| x - x' \|^2\right) \tag{4-1}$$

위 식은 아래와 같이 표현할 수도 있다.

$$k(x, x') = \exp\left(-\frac{\| x - x' \|^2}{\sigma^2}\right) \tag{4-2}$$

즉 γ를 $1/\sigma^2$로 대체하였다. 식 (4-1) 및 (4-2)에서 $\| x - x' \|^2$는 유클리디언(Euclidean) 거리제곱에 해당하며, σ는 일종의 스케일 파라미터로서 커널의 대역폭(Bandwidth)으로 해석할 수 있다.

Radial Basis 함수를 커널로 사용한 SVM 모형 적합에서 가장 중요한 과정은 동조 파라미터 cost 값과 γ 값의 결정이다. cost 값이 클수록 더 많은 수의 서포트 벡터(Support Vector)를 선택할 수 있음을, cost 값이 작으면 매우 작은 수의 서포트 벡터가 선택됨을 의미한다. 이러한 cost 값은 앞 장에서 유사 결정계수(Pseudo R^2)값이 최대가 될 때의 값으로 설정하였다. 여기에서는 γ값의 최적 결정에 대해 검토한다.

식 (4-1)을 보면 γ는 서포트 벡터로 선택된 관찰치들의 영향력이 미치는 범위를 의미한다. 즉 γ는 스케일 파라미터 σ의 역수이므로 γ 값이 작으면 매우 먼 거리까지, γ 값이 크면 매우 가까운 거리까지만 서포트 벡터의 영향력이 미침을 나타낸다. 따라서 γ 값에 의해 모형의 적합 결과가 민감하게 변할 가능성이 있다.

즉 γ 값이 지나치게 크면 서포트 벡터가 영향을 미치는 범위가 자신 하나일 수 있고, 지나치게 작으면 영향 범위가 데이터 전체가 되어 극단적인 경우 일반적인 선형 모형과 그 예측 행태가 유사해질 수 있다.

본 연구에서는 γ, 즉 σ 값이 주택가격이 상호 영향을 미치는 지리적 범위로 해석할 수 있다는 사실에 착안하여 다음과 같은 방법으로 주택가격의 공간적 종속성을 SVM 적합 과정에 반영하였다.

① OLS 모형을 통해[38] 잔차(Residuals)를 계산한다.

② 0.0 ~ 1.0의 범위로 스케일링(Scaling)한 지리좌표값(x,y)을 설명변수로, ① 단계에서 계산한 잔차를 종속변수로 하여 베리오그램을 적합시키고 레인지(Range) 값(σ)을 찾는다.

③ 스케일링한 지리좌표값(x,y)을 설명변수로, 잔차를 종속변수로 하여 SVM을 적합한다. 이때 γ 값은 ② 단계에서 찾은 레인지 값(σ)을 사용한다($\gamma = 1/\sigma^2$).

④ OLS 모형을 검증 데이터(Test Data)에 적용하여 가격을 예측한다($\hat{Y}$).

⑤ SVM을 검증 데이터에 적용하여 잔차를 예측한다($\hat{\epsilon}$).

⑥ $\hat{Y}+\hat{\epsilon}$을 예측가격으로 결정한다.

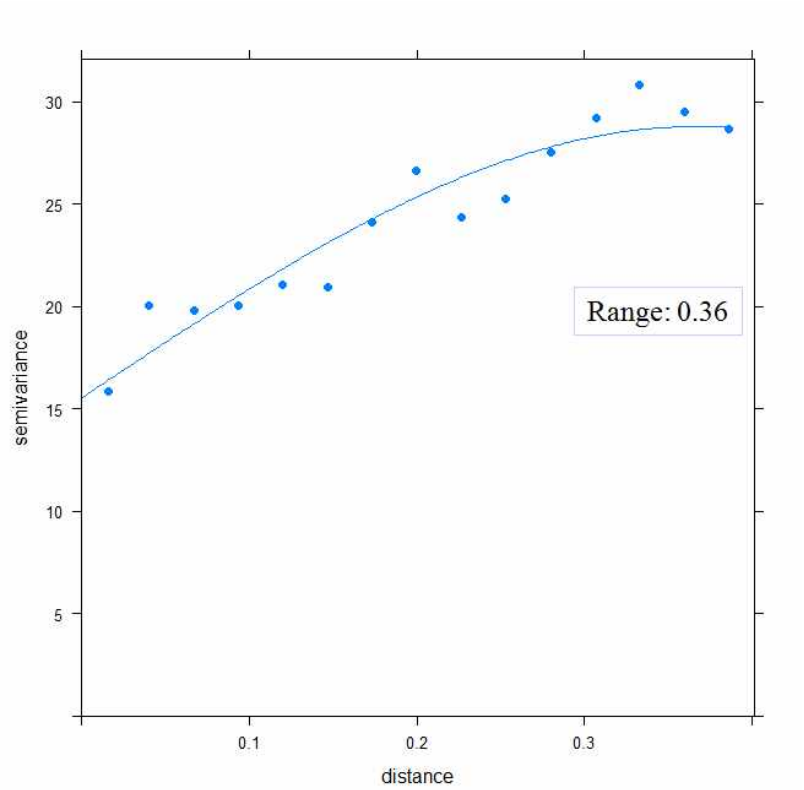
즉, 주택가격을 예측할 수 있는 체계적 요인들은 OLS 모형 등을 통해 통제한 후 잔차, 즉 설명되지 않는 변이에 대해 크리깅 기법을 적용하는 회귀-크리깅(Regression-kriging) 접근과 유사한 논리이다.

지리좌표값(x,y)을 0.0 ~ 1.0의 범위로 스케일링하였으므로 예를 들어 $\sigma = 0.3$인 경우 대상 주택이 위치하는 지점을 기준으로 전체 지리적 범위의 약 30% 이내에 소재하는 주택들만 대상 주택의 가격 예측에 활용됨을 의미한다.

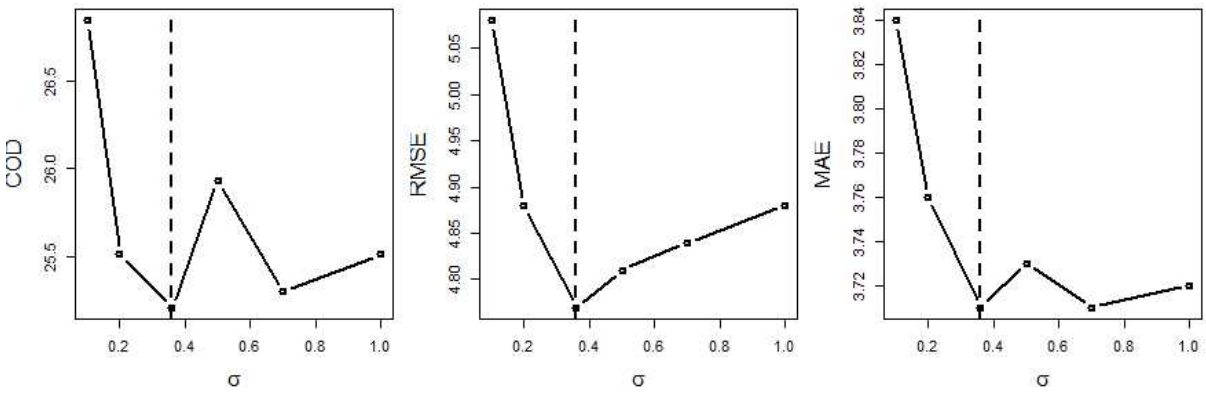
[그림 4-2]는 서울시 강남구를 대상으로 OLS 모형에서 산출된 잔차

38) 다른 모형을 사용하여도 무방하다.

에 대해 베리오그램을 적합시킨 결과를 보여준다. 지수형(Exponential) 베리오그램을 사용하였으며, 주택가격이 상호 영향을 미치는 지리적 범위는 $\sigma = 0.36$으로 산출되어 전체 지리적 범위 중 약 36% 이내의 주택들이 가격 예측에 활용됨을 의미한다. [그림 4-3]을 보면 $\sigma = 0.36$인 경우 검증 데이터를 기준한 SVM 모형의 예측 결과가 가장 우수함을 알 수 있다. 즉 $\sigma = 0.36$보다 작거나 큰 범위의 스케일 파라미터를 사용할 경우 예측 결과가 악화되고 있다.

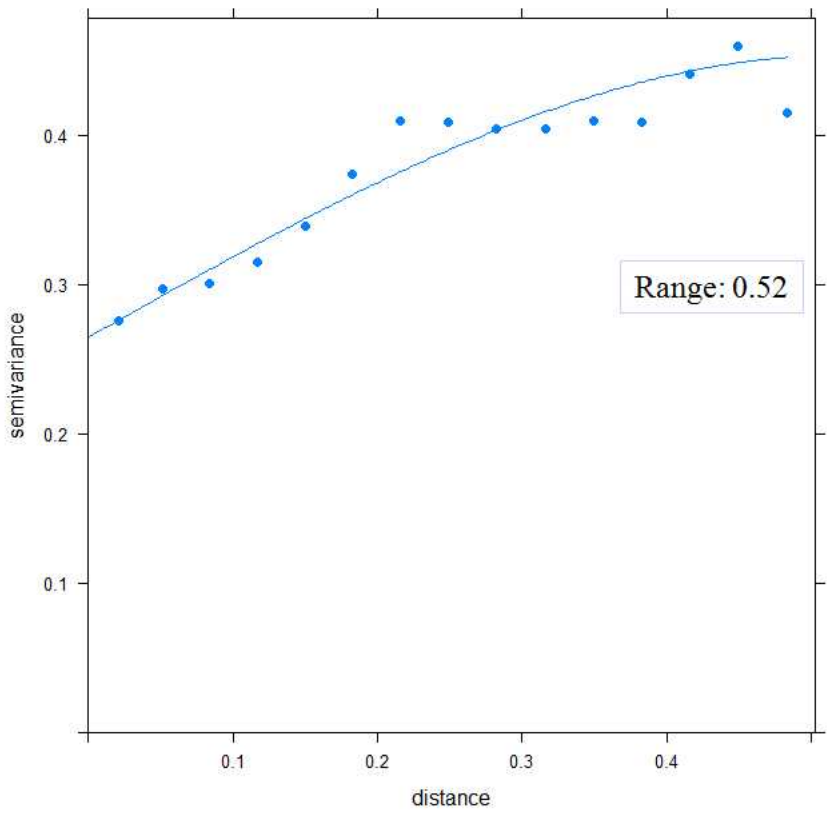


[그림 4-2] 서울시 강남구 베리오그램 적합 결과

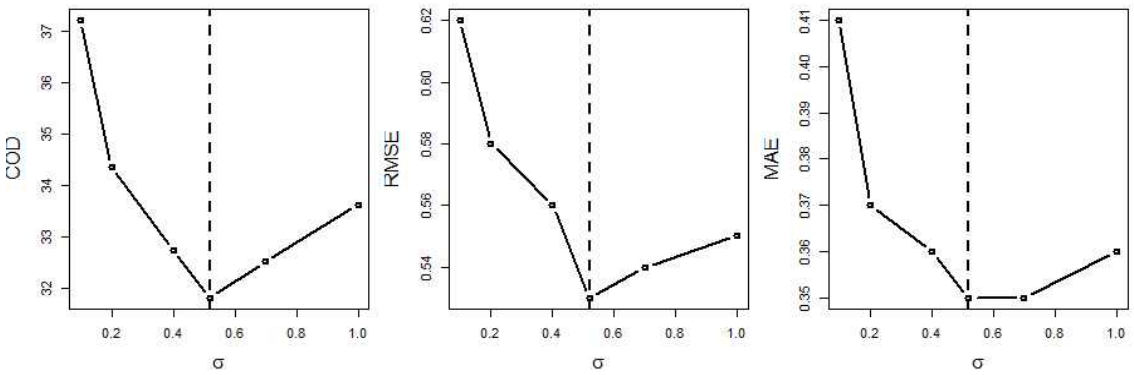


[그림 4-3] 서울시 강남구 SVM 성능(수직 점선 $\sigma = 0.36$)

[그림 4-4] 및 [그림 4-5]는 전주시 덕진구의 경우를 보여준다. 서울시 강남구와 유사한 결과가 산출되었으며, 다만 주택가격이 상호 영향을 미치는 지리적 범위는 $\sigma = 0.52$로 산출되어 강남구의 경우보다 넓다는 것을 알 수 있다.



[그림 4-4] 전주시 덕진구 베리오그램 적합 결과



[그림 4-5] 전주시 덕진구 SVM 성능(수직 점선 $\sigma = 0.52$)

전라남도 해남군은 최초 SVM 적용시 Radial Basis 함수가 아닌 Linear 커널을 사용하였으므로, 스케일 파라미터의 적용 여지가 없어 분석에서 제외하였다.

[표 4-1]은 검증 데이터를 기준한 모형 성능의 개선 정도를 보여준다. 두 개 지역 모두 최초의 SVM 모형보다 베리오그램을 활용한 SVM 모형의 성능이 보다 우수하게 나타났음을 알 수 있다.

[표 4-1] 모형 성능의 개선 정도(검증 데이터 기준)

지역	구분	RMSE	MAE	COD
강남구	SVM	5.41→4.77 (▼)	4.19→3.71 (▼)	29.77→25.21 (▼)
덕진구	SVM	0.56→0.53 (▼)	0.36→0.35 (▼)	32.38→31.81 (▼)

본 절에서는 Radial Basis 함수를 사용한 SVM 모형의 성능을 개선하기 위해 스케일 파라미터 γ(또는 σ)를 보다 정교하게 구성하였다. 즉 주택가격이 상호 영향을 미치는 지리적 범위로 γ 값을 해석하여 가격의 공간적 종속성을 SVM 적합 과정에 반영하였다.

그러나 본 절에서 시도한 접근은 SVM, 특히 Radial Basis 함수를 사용한 SVM 모형에 국한되는 한계가 있다. 따라서 다음 절에서는 SVM 뿐 아니라 본 연구에서 활용한 모든 모형에 대해 폭 넓게 시도할 수 있는 접근방법에 대해 살펴본다.

제 2 절 공간차 변수를 활용한 모형의 적용

1. 공간가중행렬의 구성

공간사상의 가장 큰 특징인 공간적 종속성을 정량화하기 위한 대표적인 도구가 공간가중행렬이다. 공간가중행렬은 n개의 관찰치에 대한 n x n 대칭양행렬(Positive Symmetric Matrix)로서, 구성요소 w_{ij}는 관찰치 i 및 j 간에 부여된 가중치를 의미한다. 가중치 w_{ij}는 각 관찰치 쌍의 공간관계를 잘 표현할 수 있는 규칙에 의해 결정되며 일반적으로 대각선 요소는 $w_{ii}=0$을 부여한다.

w_{ij}, 즉 공간가중행렬이 다양하게 정의되고 수치화될 수 있다는 것은 잘 알려진 사실이며, 따라서 어떤 공간가중행렬이 '올바른' 것이냐 하는 논의에 정답은 없다. 다만 어떤 공간가중행렬이 보다 '쓸모 있는지'를 찾아내는 것이 공간통계기법을 사용하는 연구자들의 몫이고 그들의 능력이다(박기호, 2004).

공간가중행렬은 인접성(Contiguity) 척도 또는 거리(Distance) 척도에 따라 구성할 수 있으며, 각 지점의 좌표나 지점 간 거리를 알 수 있는 경우 거리 척도를 이용한 가중행렬이 보다 바람직하다(Anselin, 1988). 거리 척도를 활용하는 경우, 거리가 멀어짐에 따라 관찰치들 간의 영향력이 감소한다는 공간현상을 정량화할 필요가 있는데, 거리조락함수(Distance-Decay Function)로 표현하는 것이 일반적이다(이창로・박기호, 2013). 거리조락함수는 다음과 같은 형태로 표현할 수 있다.

$$w_{ij}=f(d_{ij},b) \tag{4-3}$$

여기서 d_{ij}와 w_{ij}는 관찰치 i 및 관찰치 j간의 거리 및 부여된 가중치를 의미하며, b는 임계치(또는 대역폭)를 나타낸다.

함수 $f(\bullet)$는 다양한 형태로 표현할 수 있으나 주로 멱함수(Power Function)나 지수함수(Exponential Function)가 많이 활용되며 식 (4-4) 및 식 (4-5)로 나타낼 수 있다(이창로・박기호, 2013).

$$w_{ij} = 1/(d_{ij})^{\alpha} \tag{4-4}$$

$$w_{ij} = \exp(-\beta d_{ij}) \tag{4-5}$$

최적의 함수 형태 $f(\bullet)$를 결정하기 위한 연구가 여럿 있었으나 널리 받아들여지는 의견이나 결론은 없는 것으로 보인다. [표 4-2]는 공간가중행렬의 구성과 관련된 최근의 연구 결과를 정리한 것인데, 국내의 경우 대부분 거리 또는 거리 제곱에 반비례하도록 가중치를 부여하였으며, 해외의 경우 거리의 멱함수, 커널함수 등의 시도가 이루어졌다.

앞서 언급하였듯 **W**는 여러 가지 방법으로 구성할 수 있으나, 본 연구에서는 3개 지역 모두 $1/d^2$를 가중치로 한 **W**를 적용하였다. 본 연구에서 선정한 사례지역의 경우 $1/d, 1/\sqrt{d}$ 등 다른 가중치를 사용하는 것보다 $1/d^2$를 사용하는 것이 모형 성능 개선에 보다 유리하였기 때문이다[39].

39) 모형 성능은 RMSE, MAE 및 COD를 기준으로 검토하였으며, $1/d^2$이 보다 유리하다는 사실은 주택의 경우 가격이 상호 영향을 주는 지역적 범위를 좁게 해석하는 것이 바람직함을 의미한다.

[표 4-2] 공간가중행렬의 구성

연구자	대상지역	거리조락함수	임계치 설정
서경천 외 (2001)	부산시 서측 7개구	$\frac{1}{d_{ij}^2}$	0.5km
박헌수 외 (2003)	서울시 광진구	$\frac{1}{d_{ij}}$	임계치 설정하지 않음
안지아 외 (2005)	서울시 한강 이남 11개구	$\frac{1}{d_{ij}^2}$	5km
허윤경(2007)	서울시	$\frac{1}{d_{ij}^2}$	1km에서 6km까지 1km 단위로 행렬구성 후, 5km 임계치를 중심으로 설명
	부산시		1km에서 6km까지 1km 단위로 행렬구성 후, 3km 임계치를 중심으로 설명
김성우(2010)	부산시	$\frac{1}{d_{ij}}$	임계치 설정하지 않음
송용철 외 (2012)	경기도 광주시	$\frac{1}{d_{ij}}$	임계치 설정하지 않음
Dube et al. (2013)	Quebec City, Canada	binary weight	들로네 삼각형(DeLaunay triangle)에 기반
		$\frac{1}{d_{ij}^2}$	0.5km, 1.0km
Parent et al. (2013)	Hamilton County, City of Cincinnati, USA	상관계수(ρ)	공분산함수(covariance function)로부터 공간상관행렬(spatial correlation matrix) 구성 후 상관계수 값 이용 (임계치: ρ = 0.05 미만은 가중치 0 부여)
Getis et al. (2004)	Simulation data	$\frac{1}{d_{ij}}, \frac{1}{d_{ij}^2}, \frac{1}{d_{ij}^5}$	임계치 설정하지 않음
		상관계수(ρ)	베리오그램(variogram)으로부터 공간상관행렬 구성 후 상관계수 값 이용 (임계치: 베리오그램의 range를 초과하는 경우 가중치 0 부여)
Guo et al. (2008)	Sault Ste. Marie, Ontario, USA	$\left[1-\left(\frac{d_{ij}}{h_i}\right)^2\right]^2$	Bisquare 커널함수 사용 임계치(h_i): 5m, 10m, 15m

* 출처: 이창로 · 박기호(2013)에서 인용

본 연구에서는 공간차 변수 **WY**를 생성하는 것이 최종 목적이므로 **W**의 구성 뿐 아니라 **Y**의 계산도 필요하다. **Y**의 경우 주택가격 총액이나 단가를 사용할 수 있는데, 토지나 건물의 규모가 미치는 영향력을 제거하기 위해 단가로 환산하여 계산하는 것이 바람직하다. 이 때 단가는 토지면적당 단가를 사용할 수도 있고, 건물 연면적당 단가를 사용할 수도 있다. 어떠한 기준의 단가를 사용할지는 사례 지역의 주택가격 형성 과정을 살펴 결정할 필요가 있다.

예를 들어 토지가격이 상대적으로 고가인 대도시에서는 토지면적당 단가가 가격의 예측이나 하부시장의 구획 등에 주로 활용되고, 토지가격보다는 건물가격의 비중이 큰 농촌지역의 경우 건물 연면적당 단가가 보다 적정할 수 있다. 본 연구에서는 강남구의 경우 토지면적당 단가, 덕진구 및 해남군은 건물 연면적당 단가를 활용하여 **Y** 및 **WY**를 계산하였다. 이러한 결정은 부동산 실무 관행을 따른 것이기도 하지만[40], 설명변수의 영향력 정도를 보여 준 랜덤 포리스트 및 부스팅의 결과를 인용한 결과이기도 하다. 랜덤 포리스트를 적용한 결과([그림 3-8])에서 강남구는 토지면적, 덕진구와 해남군은 건물 연면적이 상대적으로 중요한 변수인 것으로 산출되었으며 부스팅을 적용한 결과([표 3-19])에서도 역시 이와 유사한 결과가 도출되었다. 따라서 이러한 결과를 토대로 강남구는 토지면적당 단가, 덕진구 및 해남군은 건물연면적당 단가를 **Y** 계산에 활용하였다.

공간차 변수 **WY**를 또 하나의 설명변수로 동원하여 모형의 예측력을 높이려는 시도는 공간통계의 가장 본질적인 특징에 해당한다(Chun & Griffith, 2013, p.68). 즉 특정 지점 i에서의 주택가격은 건물구조, 신축연도 등과 같은 해당 주택의 일반적인 속성 뿐 아니라 인근에 위치한 주택가격의 영향도 크게 받기 마련이다.

앞 장에서 주택가격 예측에 주로 활용한 설명변수(토지면적, 건물 연면적, 경과연수 등) 외에 **WY**를 추가한 모형은 식 (4-6)의 공간시차모형(Spatially Lagged Model)의 논리와 그 맥락을 같이 한다. 즉 종속변수

40) 주택의 경우 대도시에서 토지:건물의 비중은 약 80:20 정도로 보는 것이 일반적이다. 반면 농촌지역에서 토지가 차지하는 비중은 50 이하로 간주하는 것이 통상이다.

Y를 추정하기 위해 일반적인 속성변수 X뿐 아니라 주변의 주택가격 수준 Wy도 함께 고려하고 있다.

$$y = \rho Wy + \beta X + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I) \tag{4-6}$$

다만 통상적인 공간시차모형과 달리, 본 장에서 적용한 비모수 모형은 종속변수와 설명변수(X, Wy) 간의 선형 관계를 더 이상 고집하지 않는다는 유연성이 있다.

2. 모형의 개선 정도

[표 4-3]은 공간차 변수 WY를 모형에 추가하였을 때 나타난 성능의 개선 정도를 보여준다. 모형 성능의 개선 정도는 모형 데이터가 아닌 검증 데이터를 기준으로 계산하였다. WY를 추가한 것 이외에 모형의 세부내역은 앞 장(제3장)에서와 동일하다.

[표 4-3] 모형 성능의 개선 정도(검증 데이터 기준)

지역	구분	RMSE	MAE	COD
강남구	OLS	5.83→5.11 (▼)	4.52→3.82 (▼)	28.41→27.01 (▼)
	GAM	5.58→4.72 (▼)	4.43→3.57 (▼)	27.02→21.68 (▼)
	Random Forest	5.50→4.92 (▼)	4.15→3.58 (▼)	28.19→26.99 (▼)
	Boosting	5.39→4.84 (▼)	4.17→3.60 (▼)	27.24→25.13 (▼)
	MARS	5.22→4.87 (▼)	4.12→3.72 (▼)	23.07→21.55 (▼)
	SVM	5.41→4.95 (▼)	4.19→3.60 (▼)	29.77→25.17 (▼)
덕진구	OLS	0.65→0.60 (▼)	0.42→0.39 (▼)	37.64→35.15 (▼)
	GAM	0.58→0.56 (▼)	0.37→0.35 (▼)	35.13→33.64 (▼)
	Random Forest	0.56→0.55 (▼)	0.37→0.36 (▼)	34.06→33.22 (▼)
	Boosting	0.59→0.57 (▼)	0.37→0.36 (▼)	34.35→34.33 (▼)
	MARS	0.55→0.53 (▼)	0.35→0.33 (▼)	32.74→30.88 (▼)
	SVM	0.56→0.54 (▼)	0.36→0.35 (▼)	32.38→32.35 (▼)
해남군	OLS	0.34→0.32 (▼)	0.21→0.20 (▼)	83.86→74.66 (▼)
	GAM	0.33→0.32 (▼)	0.21→0.20 (▼)	82.93→72.88 (▼)
	Random Forest	0.32→0.30 (▼)	0.22→0.20 (▼)	92.07→77.25 (▼)
	Boosting	0.33→0.30 (▼)	0.21→0.19 (▼)	91.73→76.42 (▼)
	MARS	0.40→0.32 (▼)	0.24→0.20 (▼)	87.54→74.41 (▼)
	SVM	0.37→0.36 (▼)	0.21→0.20 (▼)	76.82→69.55 (▼)

[표 4-3]에서 알 수 있듯이, 공간차 변수 **WY**를 동원한 경우 모든 사례지역 및 모든 모형에 대해서, 그리고 모든 적합도 지표에 대해서 성능이 개선된 것을 확인할 수 있다. 이와 같은 결과는 주택가격을 예측하기 위한 모형 구축시 속성변수 뿐 아니라 공간적 관계를 나타내는 변수도 함께 활용하는 것이 필수적임을 의미한다.

마지막으로 [표 4-4]는 각 사례지역별로 가장 우수하게 나타난 모형에 대해 Moran's I 값을 계산한 결과를 보여준다. 즉 최초의 모형과 **WY**를 추가한 모형에서 산출된 SR을 기준으로 Moran's I 값과 그 통계적 유의성을 비교하였다.

강남구는 Moran's I값이 0.11에서 0.08로 떨어졌고 p-value도 0.02에서 0.08로 증가하여 0.05를 기준할 경우 '공간적 종속성이 제거되었다'고 주장하는 것이 가능해졌다.

덕진구는 최초 모형에서도 SR의 공간적 종속성은 거의 없는 것으로 보이나(p-value 0.15), **WY**를 추가한 경우 그러한 결론의 강도가 더욱 강해진 것을 알 수 있다(p-value 0.46).

마지막으로 해남군은 최초 모형이나 **WY**를 추가한 모형 모두 여전히 SR에 대한 공간적 종속성은 남아 있으나 Moran's I 값 자체가 0.24에서 0.21로 떨어져 공간적 종속성의 정도는 완화된 것을 알 수 있다.

[표 4-4] Moran's I 값의 변화

지 역	최초 모형	공간차 변수(**WY**) 포함 모형
강남구(MARS)	0.11 (0.02)	0.08 (0.08)
덕진구(MARS)	0.02 (0.15)	0.00 (0.46)
해남군(SVM)	0.24 (0.00)	0.21 (0.00)

* ( )안의 수치는 p-value를 나타냄

요약하면 시계열 분석에서의 시차변수(Lagged Variable)와 유사한 공간차 변수를 비모수 모형에 활용함으로써, 검증 데이터에 대한 가격 예측력이 뚜렷하게 개선되었고, SR에 남아 있는 공간적 종속성도 완화시킬 수 있었다.

제 5 장 앙상블 학습을 활용한 추정가격의 결정

본 장에서는 앞 장에서 **WY**를 추가하여 적합시킨 비모수 모형들에 대해 앙상블 학습 개념을 적용하여 추정가격을 최종적으로 결정한다. 결정된 가격은 실제 거래가격 및 현행 공시가격과 비교하여 시사점을 제시한다.

제 1 절 앙상블 평균(Ensemble Averaging)의 적용

가격을 예측하고자 할 때 가장 우수한 것으로 파악된 모형의 단일 추정값을 사용하는 것이 아니라, 여러 개의 모형을 구축하고 이들 모형을 결합하여 최적의 추정값을 산출하는 과정을 기계학습 분야에서 '앙상블(Ensemble)'이라고 한다. James Surowiecki는 그의 저서 *The Wisdom of Crowds*(2005)에서 비전문가라 할지라도 여러 사람들의 추측이나 의견을 합치면 소수 전문가 그룹의 예측 결과보다 나을 수 있음을 강조하였다. 그는 어느 농촌 카운티(County) 우시장(牛市場) 경매에 붙여진 소의 무게를 추측하는 데 있어 장터 사람들의 예측치 평균이 다른 어떤 가축 전문가의 예측치보다 정확하다는 것을 예로 들면서 집단 지성(Collective Intelligence)의 힘을 강조하였다. 이와 유사하게 가장 우수한 모형 하나만 활용하는 것이 아니라 다양한 모형들을 함께 사용하려는 것이 앙상블 접근의 기본적인 아이디어이다.

이와 같은 앙상블 접근은 여러 가지 방법으로 실행할 수 있다. 가장 일반적인 실행방법이 앞 장에서 다루었던 *Bagging*이나 *Boosting*이다. *Bagging*은 데이터의 복원추출(Bootstrap Sampling)에 기반한 것이고, *Boosting*은 이전 단계 모형에서의 예측오차, 즉 잔차에 기초하여 관찰치들의 가중치를 달리 부여하는 방법이다. 그러나 *Bagging*이나 *Boosting*은 주로 트리 기반 모형들을 결합하는데 사용되는 경향이 있으며, MARS나 SVM 같은 이질적인 모형들까지 아우르지 못하고 있다.

접근방법 자체가 상이한, 매우 이질적인 모형들까지 결합하려는 시도를 '이질적 앙상블(Heterogeneous Ensemble)'이라 하며(Abbott, 2014, p.321) 특히 신경망 모형 분야에서는 앙상블 평균(Ensemble Averaging)이라고 지칭한다. 즉 다수의 이질적인 학습 모형들(Heterogeneous Learners)을 합하여 하나의 강한 학습 모형(a strong learner)을 만들자는 것이 앙상블 평균의 기본적인 접근이다(Rokach, 2010). 앙상블 평균은 각 모형이 가지고 있는 상이한 경향들(과대추정, 과소추정)을 서로 상쇄시킬 수 있어('average out') 효율적으로 실행할 경우 개별 모형의 예측치보다 정확한 것으로 알려져 있다(Hashem, 1997).

앙상블 평균의 실행 절차는 비교적 간단하다. 먼저 편의가 적은 대신 분산이 높은 개별 모형들을 각각 구축한다. 이후 이들 모형을 '결합'함으로써 편의도 작고 분산도 낮은 예측치를 산출한다. 이때의 결합은 각 모형들에서 추정된 값의 평균, 중위수 등을 사용하는 것을 의미한다. 중위수가 평균보다 극단치에 덜 민감하고 치우친 자료 분포에도 강건(robust)하므로 보다 바람직한 결합방법으로 인정된다(Monteiro et al., 2013).

이보다 좀더 정교한 접근은 다음과 같다. 단순히 개별 모형들이 추정한 값의 중위수 내지 평균을 최종 가격으로 할 것이 아니라 각 모형에서 산출한 추정값에 대해 적정한 가중치를 부여하는 것이다. 즉 각 모형의 특징이나 장단점, 신뢰성 등을 고려하여 모형별로 알맞은 가중치를 부여한 후, 이들 가중평균값을 최종 가격으로 결정하는 것이다.

각 모형에서 산출된 추정가격을 y_i라 하고, 최종 가격을 $\tilde{y}$라 할 경우 이러한 가중평균값은 다음과 같은 식으로 표현할 수 있다.

$$\tilde{y}(x;\alpha)=\sum_{j=1}^{p}\alpha_j y_j(x) \tag{5-1}$$

위 식에서 α는 일련의 가중치를 의미한다. 최적 α값은 신경망 모형의 적합을 통해 산출한다. 이때의 신경망 모형은 하부의 개별 모형들로 구성된 일종의 '메타 신경망 모형'이 되는 셈이다. 이러한 방식의 최종가격 결정을 개별 모형의 선형 결합(a linear combination of experts) 방식이라고 한다(Hashem, 1997).

최적이라고 판단되는 특정 모형 하나만(M_k) 선택하고, 나머지 모형들은 가격 예측에 활용하지 않은 경우 $\alpha_k = 1$이고 나머지 모형의 가중치는 $\alpha_j = 0$임을 의미한다. 또한 개별 모형들의 평균값을 최종가격으로 정하였다면, 모형들 수의 역수를 각 모형의 가중치로 부여한 셈이 된다.

이러한 앙상블 평균은 부동산 감정평가에서의 '시산가격 조정(試算價格 調整, Reconciliation of Value)'과 동일한 개념으로 해석할 수 있다. 부동산 가격을 추정할 때 한 가지 평가기법이 아닌 복수의 평가기법을 사용하였다면 도출된 가격 또한 평가기법에 따라 상이할 것이다. 예를 들어 3가지 서로 다른 평가기법을 활용하여 3가지 시산가격이 도출되었다면, 이 가격들을 잘 조정하여 최종 가격에 이르는 과정이 시산가격의 조정인 것이다.

감정평가 실무상 시산가격들의 산술평균치를 최종가격으로 정하기도 하지만, 각 평가기법의 논리적 타당성, 각 평가기법에 사용된 자료의 신뢰성 등을 종합적으로 참작하여 특정 시산가격에 더 많은 가중치를 부여하여 최종 가격을 정하기도 한다. 따라서 앙상블 평균과 같은 접근방식은 감정평가 분야에서 주관성 개입 여지가 많은 시산가격 조정 절차에 객관적인 기준을 제공할 수 있을 것으로 기대된다. 이하에서는 앙상블 평균 계산에 직접적으로 활용된 신경망 모형의 논리에 대해 설명한다.

신경망 모형은 설명변수를 다양한 방법으로 선형결합한 후, 이러한 선형결합을 비선형 함수 형태로 하여 종속변수를 예측하는 방법이다. 여기에서는 비교적 단순한 single-hidden-layer, feedforward 신경망 모형에 대해 설명한다. 그러나 이러한 설명은 multiple-hidden-layer 등 보다 복잡한 형태의 신경망 모형에도 동일하게 확장하여 적용할 수 있다.

신경망 모형에서 종속변수 Y_i는 비선형 함수 g_Y로 표현되며, 이 때 g_Y는 m개의 유도 설명변수(Derived Predictors) $H_{i0}, H_{i1}, \ldots, H_{i,m-1}$로 구성되며, 형태는 다음과 같다[41].

$$Y_i = g_Y\left(\beta_0 H_{i0} + \beta_1 H_{i1} + \ldots + \beta_{i,m-1} H_{i,m-1}\right) + \epsilon_i = g_Y\left(\boldsymbol{H_i'\beta}\right) + \epsilon_i \qquad (5\text{-}2)$$

41) 이하 notation 등은 Kutner et al.(2005)를 따랐다.

통상 $H_{i0}=1$로 할당하게 되며, 나머지 H_{ij}(i번째 관찰치의 j번째 유도 설명변수)는 비선형 함수 g_j로 표현하되, 이 때 g_j는 최초 설명변수 $\boldsymbol{X_i}$의 선형결합 형태로 이루어진다. 수식은 다음과 같다.

$$H_{ij}=g_j(\boldsymbol{X_i'\alpha_j}) \qquad j=1,...,m-1 \tag{5-3}$$

식(5-2)와 식(5-3)을 합해서 표현하면 다음과 같다.

$$Y_i=g_Y(\boldsymbol{H_i'\beta})+\epsilon_i=g_Y\left[\beta_0+\sum_{j=1}^{m-1}\beta_j g_j(\boldsymbol{X_i'\alpha_j})\right]+\epsilon_i \tag{5-4}$$

위 식에서 m개의 함수 $g_Y, g_1, ..., g_{m-1}$를 활성화 함수(Activation Function)라고 하며, 대표적인 형태는 다음과 같은 로지스틱 함수이다[42]. 이러한 함수 형태는 여러 가지 상황에 비교적 잘 들어맞는 유연성을 가지고 있는 것으로 알려져 있다(Kutner et al., 2005, p.538).

$$g(Z)=\frac{1}{1+e^{-Z}}=[1+e^{-Z}]^{-1} \tag{5-5}$$

활성화 함수를 이용하여 식(5-4)를 다시 표현하면 다음과 같다.

$$\begin{aligned} Y_i &= [1+\exp(-\boldsymbol{H_i'\beta})]^{-1}+\epsilon_i \\ &= \left[1+\exp\left[-\beta_0-\sum_{j=1}^{m-1}\beta_j[1+\exp(-\boldsymbol{X_i'\alpha_j})]^{-1}\right]\right]^{-1}+\epsilon_i \\ &= f(X_i, \boldsymbol{\alpha_1}, ..., \boldsymbol{\alpha_{m-1}}, \boldsymbol{\beta})+\epsilon_i \end{aligned} \tag{5-6}$$

$\boldsymbol{\beta, \alpha_1, ..., \alpha_{m-1}}$: 모수

$\boldsymbol{X_i}$: 설명변수

ϵ_i : 잔차

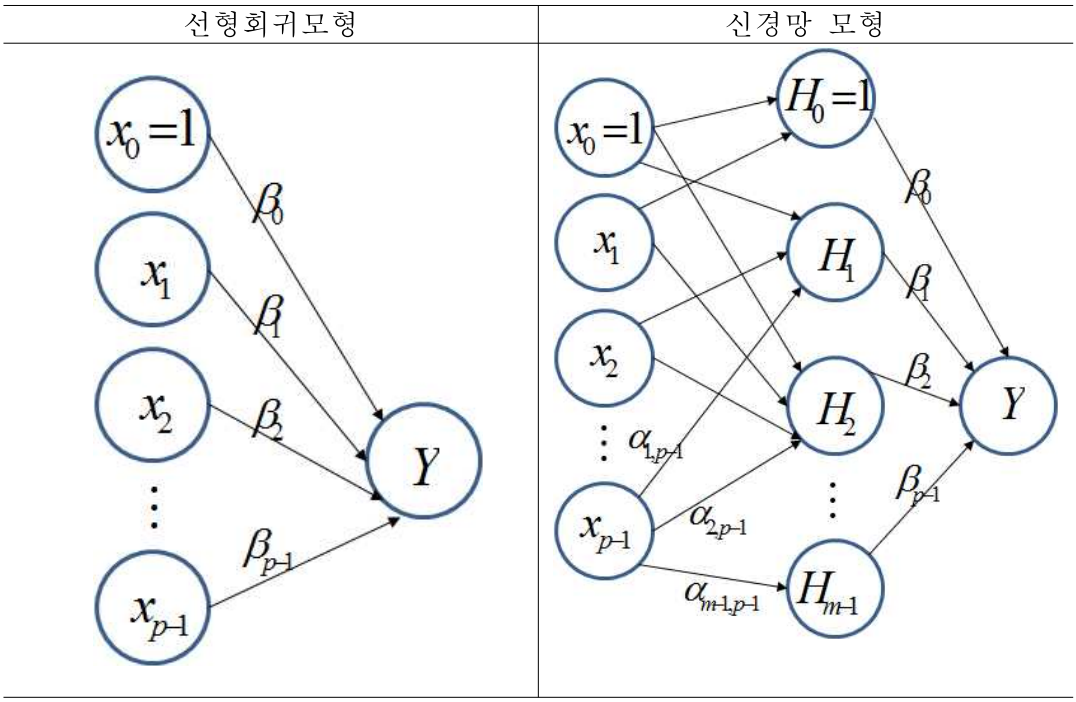
로지스틱 활성화 함수는 0과 1 사이의 값을 가지므로, 본 연구에서처럼 종속변수가 연속형인 경우 Y_i가 0과 1 사이의 값을 취하도록 다음과 같이 변환해 줄 필요가 있다[43].

---

42) 이 밖에 Radial Basis 함수 등을 사용하기도 한다.
43) 설명변수 또한 평균 0, 표준편차 1을 갖도록 변환(Scaling)하는 것이 일반적이다.

$$Y_i' = \frac{Y_i - Y_{\min}}{Y_{\max} - Y_{\min}} \tag{5-7}$$

신경망 모형은 네트워크 형태로 그 구조를 표현하는 경우가 일반적이다. 통상의 선형회귀모형과 식(5-6)과 같은 신경망 모형을 네트워크 형태로 표현하면 [그림 5-1]과 같다.



[그림 5-1] 선형회귀모형 및 신경망 모형의 네트워크 표현

[그림 5-1]의 오른쪽 신경망 모형에서 가운데 H는 m개의 은닉 노드(Hidden Node)를 나타낸다. 활성화 함수 $g_Y, g_1, \cdots, g_{m-1}$에 대해 로지스틱 함수가 아닌 항등함수(Identify Function) $g(Z)=Z$를 가정할 경우 신경망 모형은 일반적인 선형회귀모형으로 단순해진다. 다시 말해 신경망 모형은 선형회귀모형을 일반화시킨 것으로 해석할 수 있다. 이와 같은 신경망 모형에서 모수 값을 찾는 과정은 종종 'back-propagation'이라는 절차를 통해 이루어지며[44] 본 연구에서도 이러한 방법을 따라 모수를 추정하였다.

44) 국내에서는 '역전파'라고 번역하기도 한다. 자세한 사항은 Hastie et al.(2009) pp.395-397 참조.

신경망 모형을 활용하여 부동산 가격을 추정한 선행연구를 살펴보면, 먼저 함수의 전반적 형태를 비선형으로 가정하고 선형회귀모형과 비교하여 신경망 모형의 성능이 우수함을 밝힌 연구가 있다(Curry et al., 2001; Selim, 2009). 이와는 달리 신경망 모형을 모형 설정 오류(Model Specification Error)의 검증 수단으로 활용한 사례도 흔히 발견된다(Currey et al., 2002; Landajo, 2012). Currey et al.(2002)은 선형 함수 형태로 헤도닉 모형을 구성한 후, 신경망 모형을 이용하여 이러한 선형 근사화가 적정한지 검증하였다. 마찬가지로 Landajo et al.(2012)은 스페인 주택을 대상으로 선형 및 준로그(semi-log) 헤도닉 모형에 대해 모형 설정 오류를 신경망 모형을 이용하여 검증하였다. 또한 신경망 모형이 어떠한 상황에서 보다 우수한 성능을 보이는지 초점을 맞춘 연구도 상당수 있는 바, 예를 들어 Peterson & Flanagan(2009)은 설명변수 대부분이 더미변수일 때 신경망 모형이 선형회귀모형보다 가격 예측력이 상대적으로 우수함을 실증적으로 밝혔다.

이와 같이 선행연구 상당 수가 기존 모형에 비해 신경망 모형이 우수하다는 결론을 내렸으나, 신경망 모형이 특별히 우월하지 않다는 결론을 내린 사례도 보고되고 있다(Rossini, 1997, Limsombunchai et al., 2004). 특히 Limsombunchai et al.(2004)은 블랙박스와 같은 계산과정의 불투명성, 일관성의 결여, 동일한 결과의 재생산(Reproduction) 곤란 등을 지적하며 신경망 모형의 사용에 매우 회의적인 의견을 제시하기도 하였다.

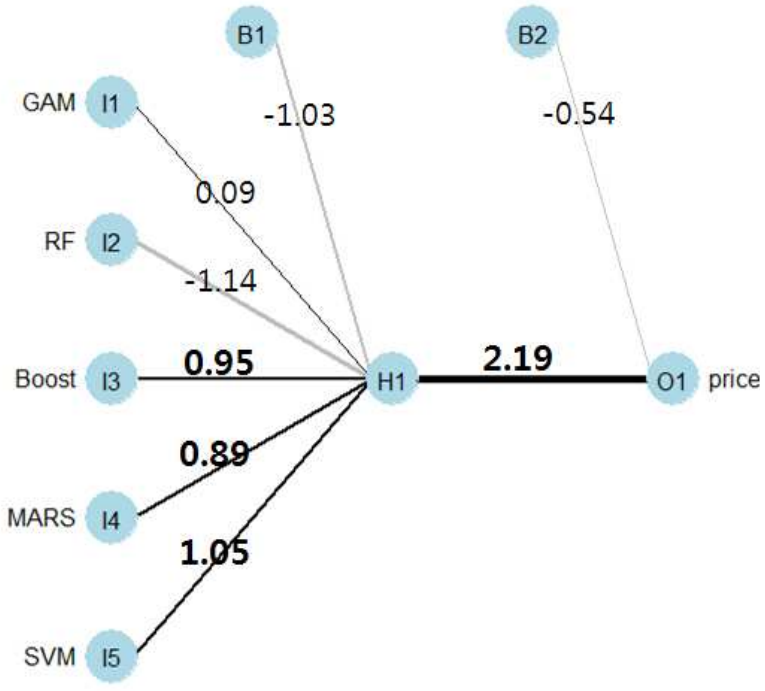
전반적으로 신경망 모형은 매우 유연한 특징을 가지고 있어 복잡한 함수 형태를 효과적으로 표현할 수 있다. 또한 선형회귀모형에서의 통상적인 가정들(잔차의 독립성, 정규성, 등분산 등)을 필요로 하지 않는 장점도 있다. 반면 종속변수 값의 예측에 주로 활용될 뿐, 모수 값을 해석하기가 수월하지 않고, 대용량의 데이터를 필요로 하는 단점이 있다(Kutner et al., 2005, p.547).

제 2 절 앙상블 평균가격의 해석

본 절에서는 제4장 제2절에서 적합시킨 모형, 즉 공간차 변수 **WY**까

지 동원된 [표 4-3] 모형에서 산출된 추정치들을 투입자료로 하여 앙상블 평균가격을 정하였다. 여기에서는 해석의 편의를 위해 은닉노드의 수는 하나로 정하고, 이에 따른 감쇠계수(Decay Parameter) 값은 부트스트랩 시뮬레이션(Bootstrapped Simulation)을 통해 RMSE가 최소화되는 시점에서의 값으로 정하였다[45]. 마지막으로 이 수치를 동조 파라미터(Tuning Parameter)로 하여 모형에 투입하였으며, [그림 5-2] ~ [그림 5-4]는 분석 결과를 보여준다.

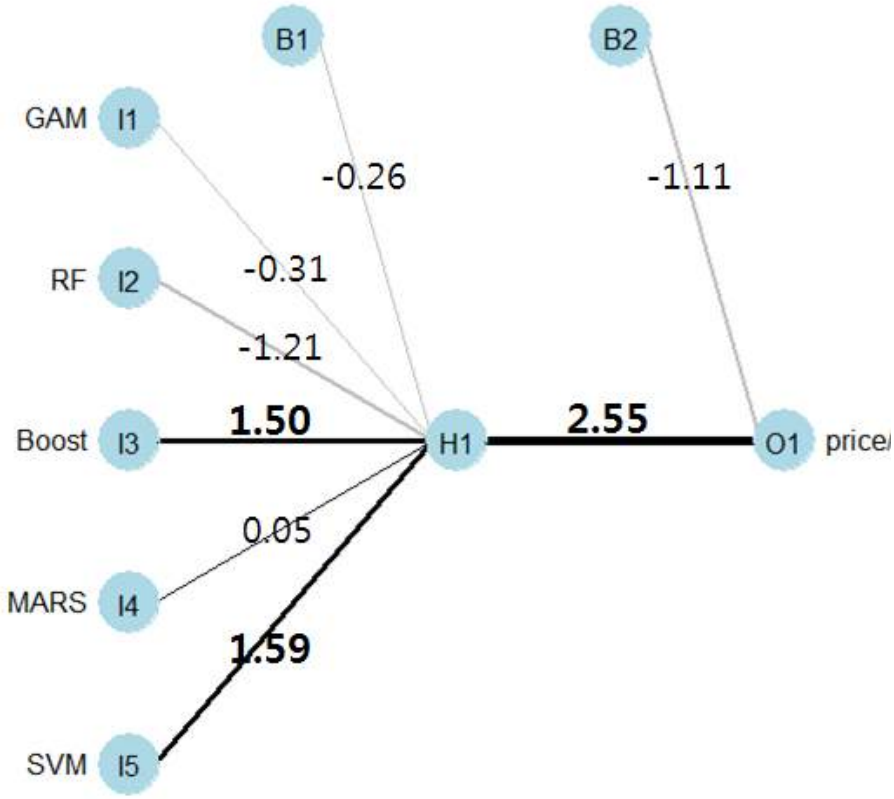
[그림 5-2]는 I(Input) → H(Hidden node) → O(Output)에 이르는 경로와 가중치를 표시한 것이다. 가중치 값이 클수록 최종 결과에 미치는 영향력이 큰 것으로 해석할 수 있으며[46], 그림에서는 가중치 값이 큰 경로를 볼드체와 굵은 실선으로 표시하였다. 그림 상단의 B는 일반적인 회귀모형에서의 상수항(Intercept)을 나타내며, 신경망 모형에서는 별도로 '편의(Bias)'라고 칭하기도 한다. [그림 5-2]를 보면 Boosting, MARS 및 SVM 모형의 결과값이 은닉 노드 H값 결정에 큰 영향을 미쳤으며, 이렇게 결정된 H값은 최종 결과 O값에 거의 그대로 전달되었음을 알 수 있다.



[그림 5-2] 신경망 모형 가중치 (강남구)

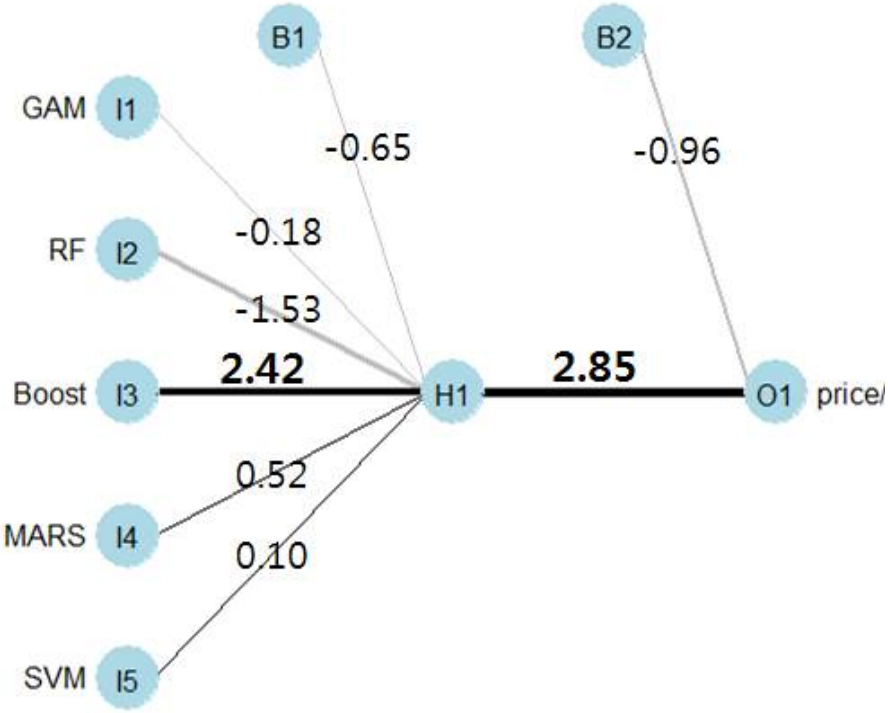
45) 3개 사례지역 모두 감쇠계수는 0.01로 정하였으며 민감도 분석 결과, 은닉노드의 수나 감쇠계수 값에 상관없이 모형 성능(COD 등)은 일정한 범위를 유지하였다.
46) 음(-)의 가중치는 최종값의 결정에 역방향으로 영향을 미치고 있음을 의미한다.

[그림 5-3]의 덕진구를 보면 최종가격 결정에 Boosting과 SVM이 거의 동등한 비중으로 큰 영향을 미쳤음을 알 수 있다. 이는 시산가격 조정 절차 측면에서 보면 다른 평가방법들(GAM, RF, MARS)에 의해 산출된 가격은 배제하고 Boosting과 SVM에 의해 산출된 가격을 산술평균한 것과 유사하다.



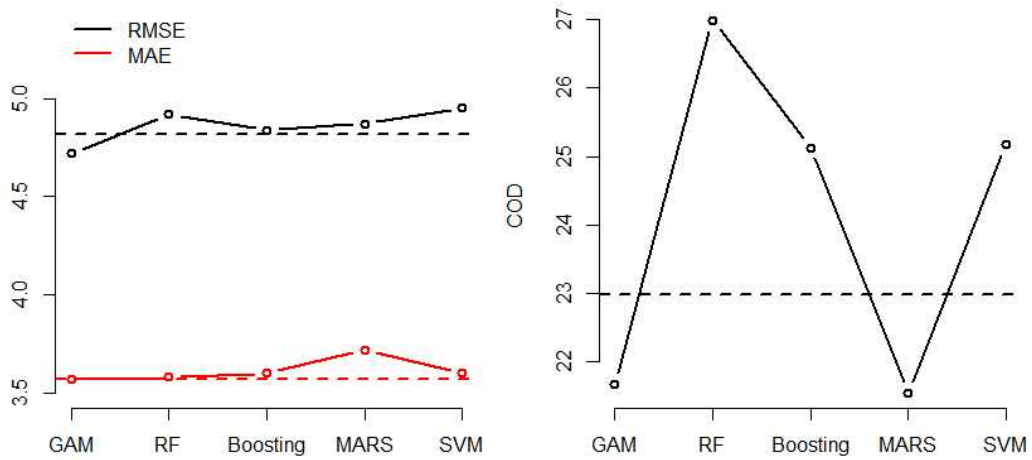
[그림 5-3] 신경망 모형 가중치 (덕진구)

마지막으로 [그림 5-4]는 해남군의 최종가격 결정을 보여 주는데, 앞의 사례지역과 달리 Boosting에 의한 가격이 최종가격 결정에 거의 독점적으로 영향을 주었음을 알 수 있다.



[그림 5-4] 신경망 모형 가중치 (해남군)

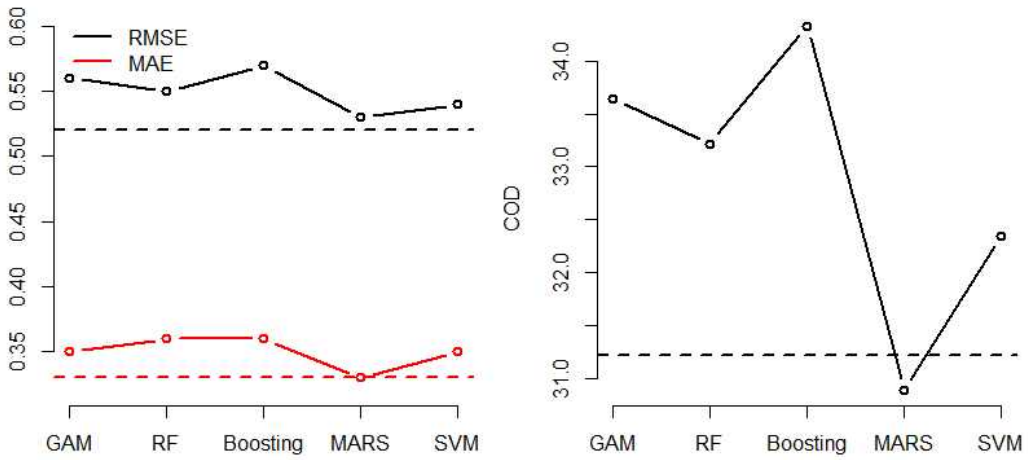
이와 같이 신경망 모형을 통해, 시산가격 조정과 유사한 절차를 거쳐 결정된 최종 가격은 실제 가격을 얼마나 잘 예측하였는가? [그림 5-5]는 검증 데이터를 기준으로 하여 계산한, 강남구 앙상블 평균 가격과 앞 장에서 공간적 종속성을 고려한 비모수 모형(**WY** 활용)의 결과를 보여주고 있다.



[그림 5-5] 강남구 앙상블 평균가격(수평 점선)의 정확성

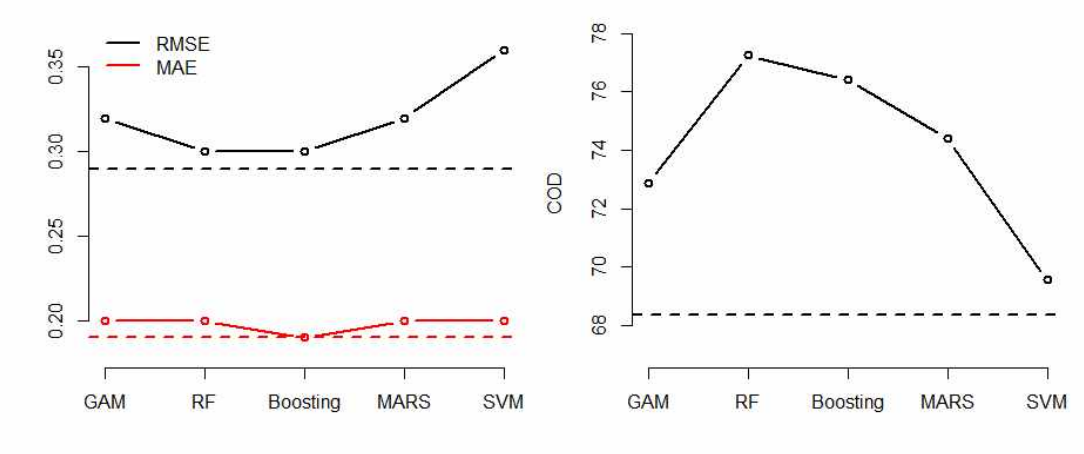
[그림 5-5]에서 점선으로 표시된 수평선이 앙상블 평균가격의 성능을 나타내며, 5개 비모수 모형과 비교할 때 비교적 정확한 가격 예측력을 보여주고 있다. 그러나 지표별 가장 우수한 모형(RMSE: GAM 모형, MAE: GAM 모형, COD: MARS)보다는 그 성능이 떨어짐을 알 수 있다. 이는 강남구의 경우 이전 모형들의 성능이 이미 상당히 우수한 수준에 이르렀기 때문에(COD 기준 약 20~30 수준) 더 이상 성능 개선의 여지가 거의 없어 나타난 현상으로 해석할 수 있다.

[그림 5-6]은 덕진구 앙상블 평균가격의 성능을 보여준다. RMSE 및 MAE 기준으로는 가장 우수한 성능을 보이고 있으며 COD 기준으로는 가장 우수한 모형인 MARS에 약간 미치지 못하고 있다. 그러나 전반적으로 앙상블 평균가격은 여타 개별 모형들보다 정확한 가격 예측력을 보이는 것으로 볼 수 있다.



[그림 5-6] 덕진구 앙상블 평균가격(수평 점선)의 정확성

마지막으로 [그림 5-7]은 해남군 앙상블 평균가격의 성능을 보여 준다. 해남군의 경우 3개 지표 모두에서 앙상블 평균가격이 근소하나마 개별 모형들보다 우수한 것으로 나타났다. 해남군은 COD 수준이 70~80에 이르는 등 전반적으로 모형 성능이 좋지 않은 지역이었다. 이러한 지역에서 앙상블 평균가격의 성능이 우수하게 나타난 것은 매우 긍정적인 현상이라 할 수 있다. 즉 개별 모형들에서 산출된 가격이 비교적 정확하지 못할 때, 이들 모형들을 적정하게 결합한 가중평균값은 그러한 부정확성을 어느 정도 상쇄할 수 있음을 보여주는 것이다.



[그림 5-7] 해남군 앙상블 평균가격(수평 점선)의 정확성

해남군의 결과는 또 다른 측면에서도 해석이 가능하다. 앙상블 접근이 탁월한 효과를 발휘하려면 개별 모형들의 예측치가 다양할 필요가 있다(Abbott, 2014, p.313). 즉 유사한 성격의 개별 모형들을 결합하는 것은 큰 의미가 없으며, 다양한 특징을 보이는 이질적인 모형들을 결합함으로써 앙상블 접근은 뚜렷한 성과를 보이게 된다. 본 연구는 주택가격의 예측이 목적이므로 부동산 가격을 과다 또는 과소 예측하는 패턴이 서로 상이한 개별 모형들로 앙상블을 구성하였을 때 성과가 크게 개선될 수 있다. 예를 들어 특정 개별 모형이 주택가격을 과다추정하더라도 다른 개별 모형이 과소추정하였다면 최종적인 결과는 오히려 정확할 수 있다.

개별 모형들의 다양성을 측정하는 방법 중 하나는 개별 모형들에서 산출된 예측치 간의 상관성(Association)을 계산하는 것이다. 예를 들어 예측치 간의 상관계수가 0.95를 상회한다면 이러한 개별 모형들을 결합하더라도, 얻을 수 있는 특별한 이점이 없을 것이라 쉽게 예상할 수 있다. [표 5-1]은 개별 모형들에서 산출된 예측치 간의 상관계수 행렬을 보여주며, 해남군이 상관성 정도가 가장 낮음을 알 수 있다.

[표 5-1] 상관계수 행렬

① 서울 강남구

	GAM	RF	Boosting	MARS	SVM
GAM	1.00				
RF	0.91	1.00			
Boosting	0.94	0.95	1.00		
MARS	0.97	0.92	0.95	1.00	
SVM	0.95	0.95	0.97	0.97	1.00

② 전주 덕진구

	GAM	RF	Boosting	MARS	SVM
GAM	1.00				
RF	0.98	1.00			
Boosting	0.97	0.99	1.00		
MARS	0.98	0.98	0.99	1.00	
SVM	0.95	0.98	0.98	0.97	1.00

③ 전남 해남군

	GAM	RF	Boosting	MARS	SVM
GAM	1.00				
RF	0.93	1.00			
Boosting	0.93	0.94	1.00		
MARS	0.87	0.85	0.93	1.00	
SVM	0.94	0.91	0.85	0.84	1.00

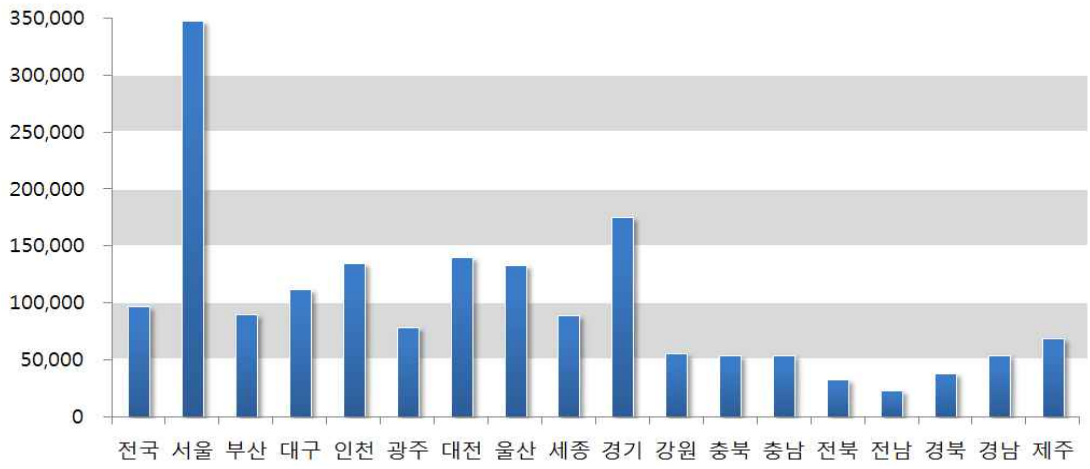
즉 해남군은 개별 모형들의 예측치 상관성이 0.90을 하회하는 경우가 다수 있음을 알 수 있다. 이러한 특징으로 인해 해남군에서 앙상블 평균의 적용 결과가 가장 우수하게 나타난 것으로 볼 수 있다.

제 3 절 단독주택 공시가격과의 비교 및 함의

1. 단독주택가격 공시제도

본 연구의 분석대상인 단독주택은 2005년부터 공시가격 제도가 도입되었으며, 이러한 공시가격은 일반적인 주택 거래의 지표가 될 뿐 아니라 주택 시장의 가격 정보를 제공하며, 국가·지방자치단체 등이 과세 등의 업무를 수행할 때 가격 기준의 역할을 하고 있다[47].

[그림 5-8]은 2014년 기준 전국 단독주택 공시가격의 평균 현황을 보여주고 있다. 서울의 평균 공시가격이 약 3억 5천만원으로 가장 높고, 전라남도의 평균 공시가격이 약 2천만원으로 가장 낮은 것을 알 수 있다. 이러한 공시가격은 정부가 2005년도에 제도를 도입하여 시행 11년차에 접어들었음에도, 시장에서의 실제 거래가격 대비 현실화율이 낮고 지역 및 유형 간 가격 균형도 미흡하다는 지적이 계속하여 제기되고 있다(이우진·방경식, 2006; 김종수, 2012).



[그림 5-8] 단독주택 공시가격 평균(2014년 기준, 단위:천원)

* 출처: 2014년 단독주택 공시가격 통계 e-book(한국감정원)

47) 부동산 가격공시 및 감정평가에 관한 법률 제18조.

공시가격의 산정은 크게 3가지 절차를 통해 이루어진다. 첫 번째는 일종의 표본에 해당하는 표준주택의 선정 단계이다. 2014년 기준 전국의 단독주택은 약 400만호이며 이 중 19만호는 표준주택에 해당한다(표본비율 약 4.75%). 표준주택은 부동산 가격 전문가인 감정평가사가 서류 및 현장조사를 통해 선정한 인근지역의 대표적인 주택으로 시장에서의 '거래사례'를 대신하는 역할을 한다. 감정평가사가 선정하고 조사한 표준주택의 특성과 평가가격은 이후 가격추정모형 구축의 투입자료로 사용된다.

두 번째는 가격추정모형을 구축하는 단계로, 헤도닉 모형(OLS 모형)에 기반하여 비준표를 작성하게 된다. 비준표는 표준주택이 아닌 개별주택 약 381만호의 가격을 빠르고 손쉽게 산정하기 위하여 고안된 일종의 간이 가격배율표이며, 이 표는 OLS 모형의 회귀계수를 지수화하여 작성된다.

마지막 세 번째는 비준표를 통해 산정된 개별주택 가격에 대해 여러 이해관계자들의 의견을 듣고 가격을 조정하는 단계이다. 담당 감정평가사의 검증을 통한 가격 조정, 토지 소유자 의견제출·이의신청 제기를 통한 가격 조정, 공익사업 시행자의 의견을 반영한 가격 조정, 중앙정부 및 지방자치단체장의 의견 수렴 등 이 단계에서는 모형에 기반하지 않은 정성적 성격의 가격 조정이 무수히 발생하게 된다. 따라서 세 번째 단계를 거쳐 최종적으로 발표되는 공시가격은 헤도닉 모형에 의해 1차적으로 산출된 가격과 매우 다른 양상을 보일 수 있다.

이러한 공시가격 산정절차를 고려할 때, 최종적으로 공표되는 공시가격에는 여러 단계에서 발생한 오류가 집적되어 있다. 우선 첫 번째 단계에서 선정된 표준주택 자체가 인근지역을 대표하지 못하는, 즉 대표성이 결여된 주택일 수 있다. 또는 표준주택이 올바르게 선정되었다 하더라도 표준주택의 다양한 특성(토지 형상, 도로조건, 리모델링 여부 등)을 감정평가사가 조사하는 단계에서 판단 오류 및 측정 오류를 일으킬 수 있다. 아울러 가격의 결정과정에 있어서도 감정평가사의 주관적 인지성향에서 발생하는 가격 평활화 현상(appraisal smoothing) 등이 발생할 수 있다(김용창, 2008).

두 번째 단계에서 발생 가능한 오류는 모형 구축과 관련된 오류로서 MAUP, 설명변수의 선별, 가격함수의 형태, 오차항의 구조 등 제2장에서 살펴 본 사안들이 대부분 여기에 해당된다. 현행 비준표는 모형 설계 측면에서 볼 때 매우 초보적인 단계의 OLS 모형으로 이러한 오류들을 상당 부분 포함하고 있을 가능성이 높다.

세 번째 단계에서 발생 가능한 오류는 이해 관계자의 의견청취 및 불복절차를 통해 발생하는 것들로 대부분 조세저항의 성격을 띠고 있어 최초의 가격안보다 하향시켜 달라는 요구가 주를 이룬다. 따라서 이러한 의견청취 절차를 거쳐 최종 결정된 공시가격은 최초 가격보다 낮은 수준의 가격이며, 이는 실제 시세와의 괴리를 심화시킨다.

따라서 현행 공시가격의 낮은 정확성은 가격추정 모형 자체에서 기인하는 것도 있지만 이 밖의 다양한 절차에서 발생한 오류들도 함께 포함되어 있다. 이하에서는 이러한 점을 감안하여 본 연구에서 제시한 앙상블 예측가격과 현행 단독주택 공시가격을 비교하고, 그 시사점을 제시하고자 한다.

2. 공시가격과의 비교

앙상블 예측가격과 공시가격은 어떠한 가격을 준거가격(reference price)으로 설정했는지에 따라 비교 결과가 상이해질 수 있다. 앙상블 예측가격은 실거래가 자료를 기초로 산출된 가격이며, 공시가격은 감정평가사가 추정한 가격(감정평가가액)을 기초로 산출된 가격이기 때문이다. 따라서 준거가격을 실거래가 자료로 정할 경우, 비교 결과는 앙상블 예측가격에 유리하게 나올 수밖에 없다. 반면 준거가격을 감정평가가액으로 한 경우, 비교 결과는 공시가격에 유리할 것이다.

본 연구에서는 상기와 같은 내용을 감안하되, 자료가 확보된 실거래가 자료를 준거가격으로 하여 비교하였다. 보다 구체적으로는 모형 적합에 사용하지 않고 별도로 유보한(30%) 실거래가 검증 데이터를 기준으로 분석하였다.

[그림 5-9]는 검증 데이터의 앙상블 예측가격, 공시가격 및 실제 거래

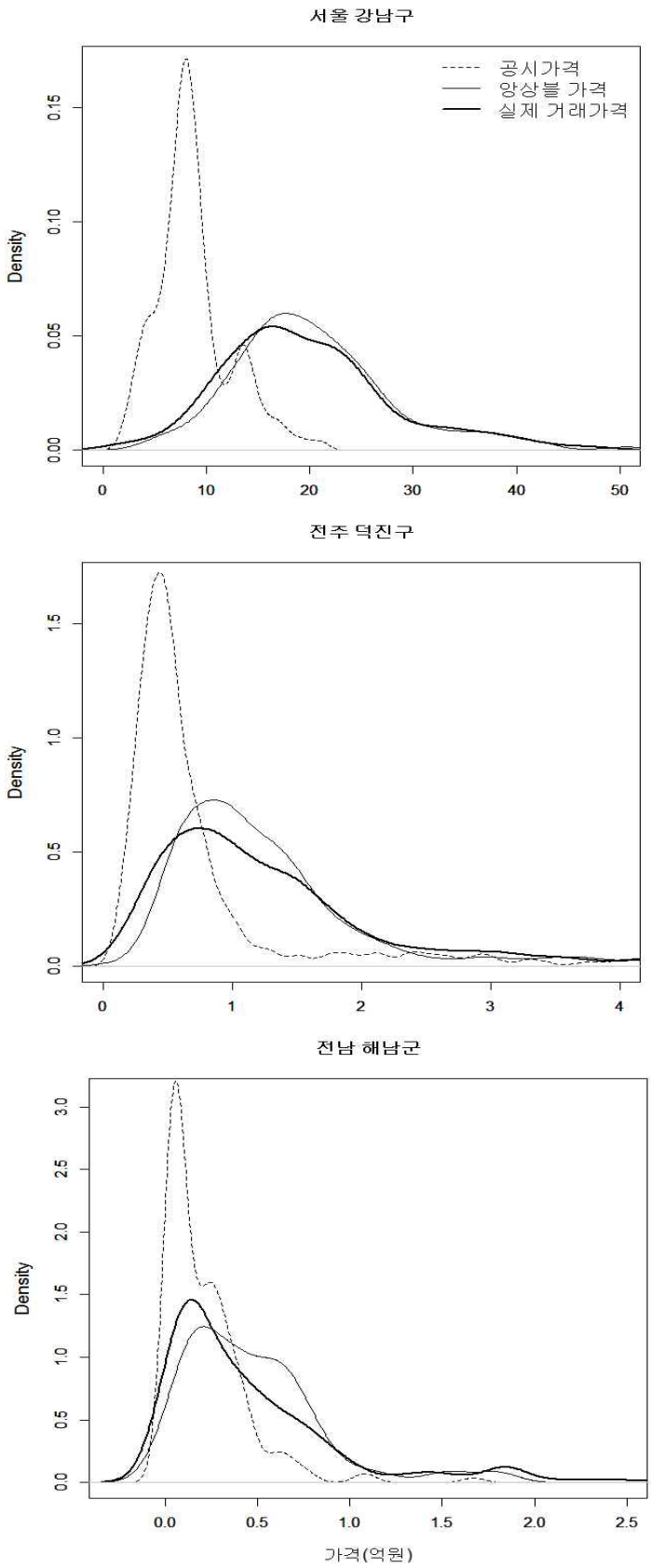
가격의 분포 현황을 보여 준다. 3개 사례지역 모두 공시가격은 대표적인 가격수준을 중심으로 매우 조밀하게 분포함을 알 수 있다. 앙상블 예측가격도 대표적인 가격수준을 중심으로 정규 분포의 형태를 보이고 있으나 비교적 넓은 범위에 걸쳐 다양하게 분포함을 알 수 있다. 전반적으로 앙상블 예측가격이 실제 거래가격과 유사한 형태를 보이고 있다.

이렇게 앙상블 예측가격이 공시가격과 상이한 패턴을 보이는 이유는 공시가격의 산정 단계를 따라 설명될 수 있다. 먼저 첫 번째 절차인 표준주택의 선정 및 가격평가 단계에서 사람에 의한 가격 평활화 현상을 들 수 있다. 표준주택가격은 사람이 판단하여 제시한 의견가격이므로 비록 그 사람이 전문가라 하더라도 자의성 개입을 완전히 배제할 수 없다. 이 경우 전문가는 자신이 생각하는 해당 지역의 대표적인 가격수준을 기준으로 대부분의 주택 가격을 추정하는, 가격 평활화 현상을 발생시킬 수 있다.

두 번째 절차인 가격추정모형 구축단계에서도 앙상블 예측가격과 공시가격은 그 격차가 더욱 벌어질 수 있다. 본 연구에서 제시한 앙상블 예측가격은 본질적으로 실거래 사례에 기초한 시장가치 추정치(Estimate of Market Value)이다. 반면 현행 비준표 모형은 엄밀한 의미에서의 시장가치 추정치는 아니며, 오히려 법령에 명시된 절차에 따라 기계적으로 산정된 과세가치(Assessed Value)일 가능성이 높다. 현행 비준표는 토지가격 비준표와 건물가격 비준표로 이원화되어 있는데, 토지가격 비준표에 의해 산정된 토지분 가격과 건물가격 비준표에 의해 산정된 건물분 가격을 합하여 공시가격으로 발표하고 있다. 이는 토지가격과 건물가격을 별도로 구분하지 않은 앙상블 예측가격의 추정과정과 대조를 이룬다. 시장에서의 거래관행도 역시 토지가격과 건물가격을 명시적으로 구분하지 않는다. 따라서 현행 비준표 모형은 시장에서의 일반전인 거래관행을 따르지 않고 토지가격과 건물가격을 별도로 산정하여 합산하는, 일종의 원가법(cost method)을 따르고 있어 앙상블 예측가격 및 실제 거래가격과 괴리를 보이고 있다.

마지막으로 이해관계자 의견청취 단계에서는 조세 부담을 우려한 가격 하향 의견을 받아들여, 공시가격 수준이 최초의 수준보다 낮아지게 된다. 따라서 공시가격의 대표적 가격수준은 앙상블 예측가격보다 매우

낮은 것을 그림에서 확인할 수 있다. 공시가격의 낮은 현실화율은 국민들로 하여금 공시가격은 과표일 뿐 시장가치를 대리할 수 없다는 인식을 심어주어 그 유용성을 떨어뜨리는 원인을 제공하기도 한다. 공시가격 수준과 실제의 거래가격 수준을 일치시킬 수 있는 방안은 가격 결정권자(감정평가사, 시군구 공무원 등)의 독립성을 보장하고, 정책 과정을 통해 과표 현실화율 100%라는 국민적 합의를 도출함으로써 가능하다(민규식, 1994).


서울 강남구
공시가격
앙상블 가격
실제 거래가격
Density
0.00
0.05
0.10
0.15
0
10
20
30
40
50
전주 덕진구
Density
0.0
0.5
1.0
1.5
0
1
2
3
4
전남 해남군
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0.0
0.5
1.0
1.5
2.0
2.5
가격(억원)

[그림 5-9] 가격 분포 현황

마지막으로 현행 공시가격과 앙상블 예측가격, 그리고 실제 거래가격의 공간적 분포 패턴을 비교하면 [그림 5-10] ~ [그림 5-12]와 같다. 강남구의 공시가격이 전반적으로 낮은 가격수준을 보이는 반면, 앙상블 예측가격은 실제 거래가격과 유사한 가격수준을 보이고 있다. 또한 강남구 북쪽의 대표적 고급주택인 신사동, 논현동 및 청담동 일대의 높은 가격수준이 앙상블 예측가격과 실제 거래가격에서는 뚜렷하게 나타나고 있다.

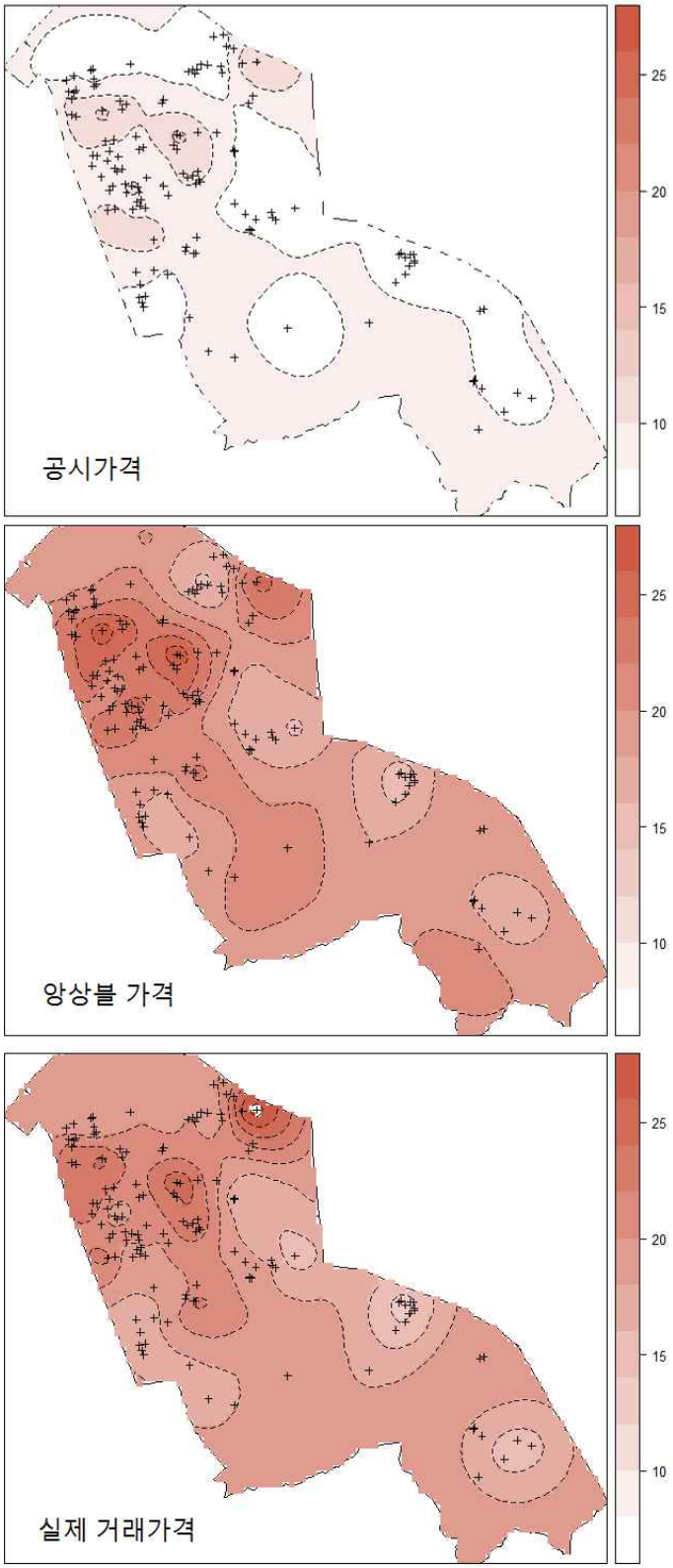
[그림 5-11]의 전주시 덕진구를 보면 세 가지 종류의 가격 모두 두 개의 지가 중심점을 명확하게 보여주고 있다. 보다 북쪽의 지가 중심점은 행정구역상 송천 1·2동을 아우르는 지역으로 최근 혁신도시 개발과 더불어 고급 아파트 단지가 형성되는 등 매우 빠르게 개발이 진행되고 있는 주택지대라 할 수 있다. 보다 남쪽의 지가 중심점은 행정구역상 우아3동에 해당되는 아중택지개발지구로서 2000년대 초에 개발되어 현재는 성숙단계에 이른 중상위 수준의 주택지대라 할 수 있다. 공시가격 및 앙상블 예측가격 모두 이러한 지가 중심점을 잘 보여주고 있으나, 지가 중심점을 기준으로 외곽으로 나갈수록 가격 수준이 체감하는 층화의 정도를 보다 정확하게 보여주는 것은 앙상블 예측가격임을 알 수 있다.

[그림 5-12] 전라남도 해남군의 경우 공시가격은 북측의 두 개 지가 중심점을 보여주는 반면, 앙상블 예측가격과 실제 거래가격은 대략 세 개 정도의 지가 중심점을 보여주고 있다. 북동측의 지가 중심점은 화원면에 소재한 기업형 혁신도시로서 레저 및 조선업 중심으로 활성화되고 있는 지역이다. 동측의 지가 중심점은 해남읍에 소재한 군청 소재지를 나타낸다. 마지막으로 남서측의 지가 중심점은 행정구역상 화산면 방축리에 해당하는 곳으로 전통적으로 군청 소재지 다음으로 높은 지가 및 주택가격 수준을 보여준 지역이다.

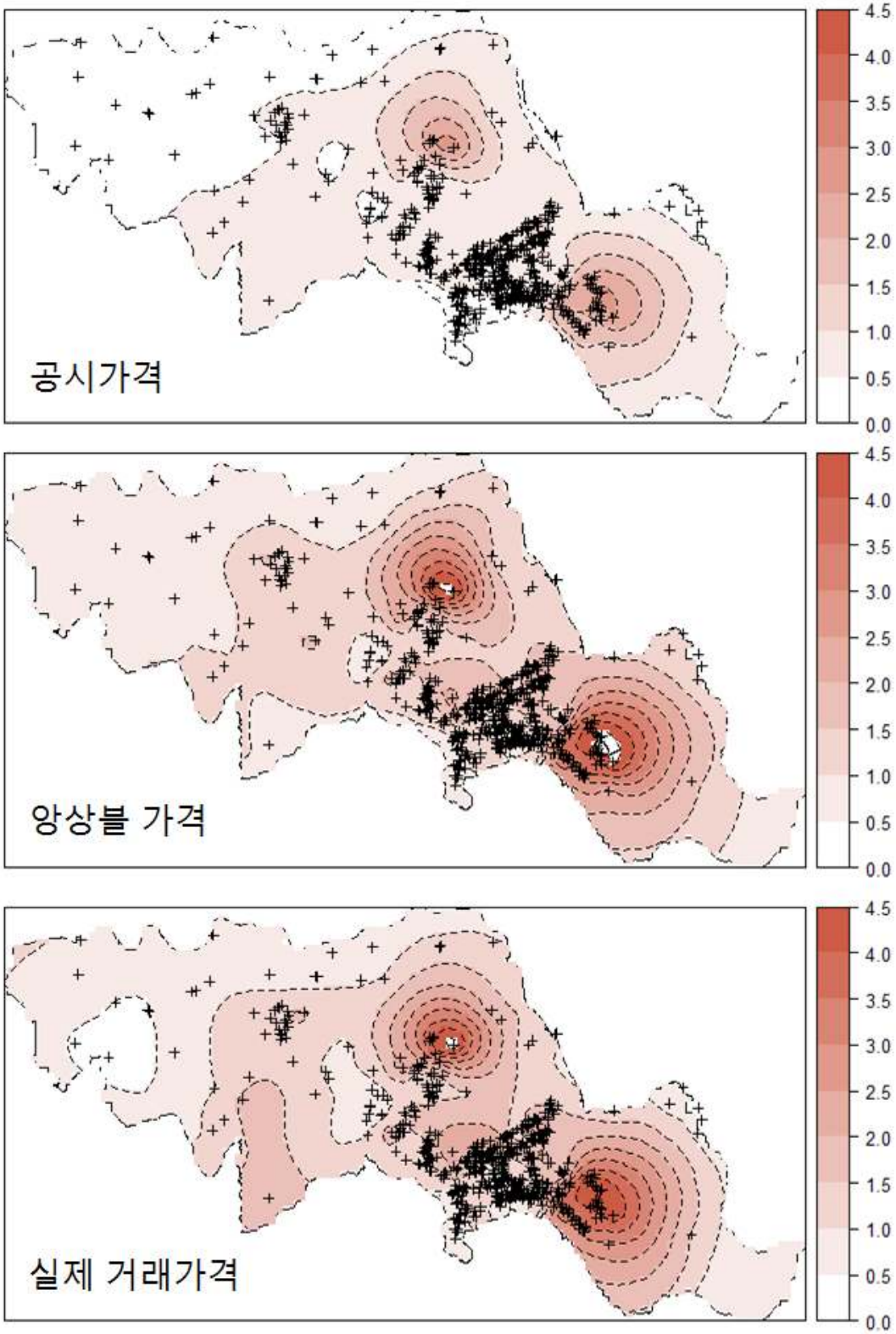
요약하면 공시가격은 앙상블 예측가격 대비 상당히 낮은 수준에서 형성되어 있으며, 그 분포의 모양 또한 특정 중심값을 중심으로 지나치게 집중된 경향을 보여주고 있다. 그러나 앙상블 예측가격은 평균가격을 상당히 넘어서는 구간에서도 계속하여 분포하는 것으로 나타나 가격 범위의 다양성을 잘 보여준다. 이는 공시가격이 일정 금액 이상의 고가 주택가격을 반영하는데 한계가 있음을 의미한다.

또한 공간적 분포와 관련하여 공시가격은 앙상블 예측가격이 보여주는 지역별 다양한 지가 중심점을 반영하는데 미흡함을 알 수 있다. 또한 지가 중심점을 기준으로 외곽으로 갈수록 가격이 낮아지는 체감 현상을 공시가격은 앙상블 예측가격만큼 정교하게 반영하지 못하고 있음을 확인할 수 있었다.

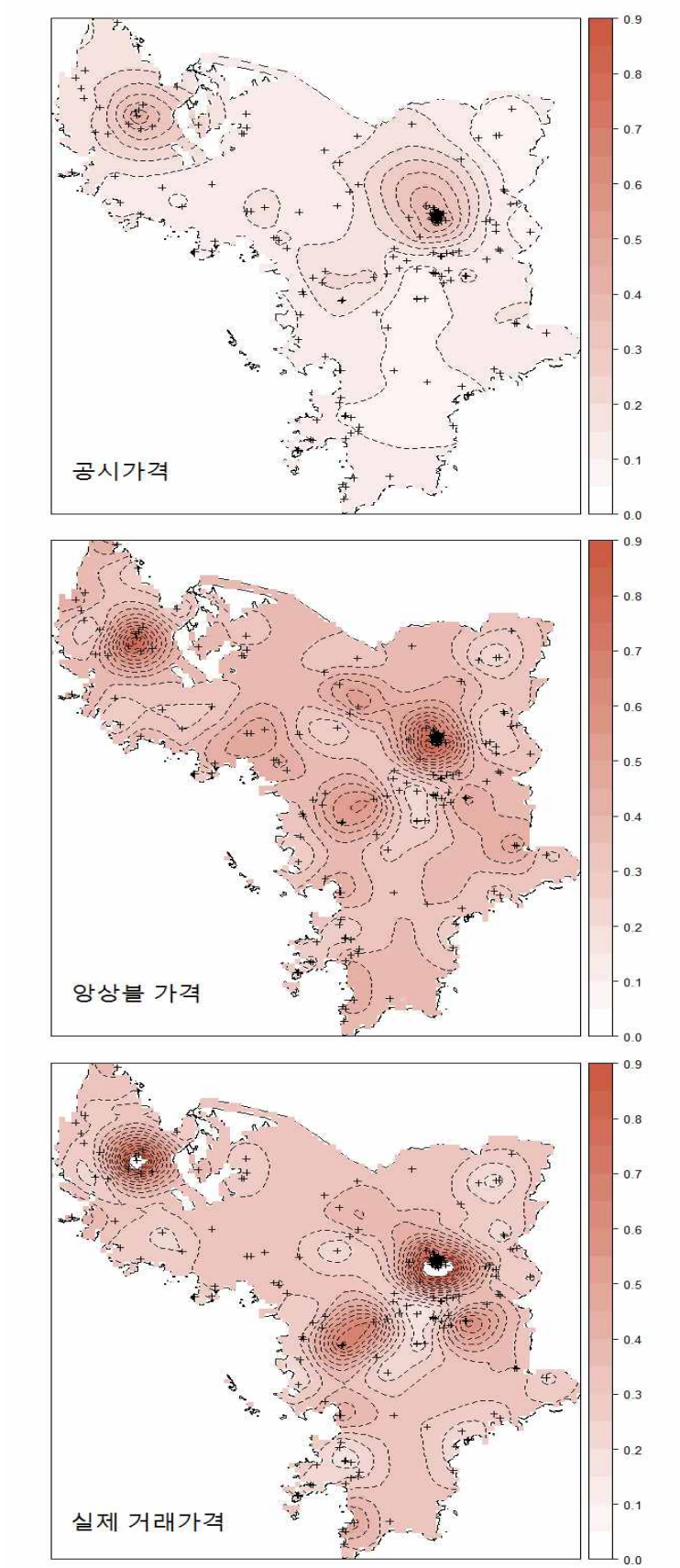
그러나 상기와 같은 공시가격의 특징 내지 품질은 오로지 가격추정모형 자체로부터 기인하는 것은 아니며, 표준주택의 선정이나 평가 같은 사람에 의한 자료수집 단계, 이의신청과 같은 이해관계자 의견청취 단계 등의 영향도 받아 형성되는 것이다. 따라서 본 절에서의 앙상블 예측가격과 현행 공시가격과의 비교 결과는 이러한 공시가격 산정 각 단계의 성격과 맥락을 고려하여 신중하게 해석할 필요가 있다.



[그림 5-10] 강남구 주택가격의 공간적 분포(검증 데이터 기준, 단위: 억원)



[그림 5-11] 덕진구 주택가격의 공간적 분포(검증 데이터 기준, 단위: 억원)



[그림 5-12] 해남군 주택가격의 공간적 분포(검증 데이터 기준, 단위: 억원)

제 6 장 결 론

1. 연구결과의 요약

최근 공공자료의 개방과 이에 따른 구득성 증가로 대량평가는 그 어느 때보다 적용하기 수월해졌다. 또한 대량평가는 정밀평가와 달리 저렴한 비용, 신속한 처리, 자의성 개입 최소화 등을 경쟁력으로 사회 각 분야에서 저변을 확대하고 있다. 특히 대량평가모형을 활용한 과세평가 영역은 그 역사가 오래되었을 뿐 아니라 국민의 재산권에 직접적인 영향을 미치는 등 파급효과가 매우 큰 분야라 할 수 있다. 본 연구는 부동산 대량평가에서의 주요 목표, 즉 부동산 가격을 보다 정확하게 추정하기 위한 방법론의 탐색으로부터 시작되었다.

부동산 가격을 추정하기 위한 기존의 연구들은 대부분 모수 모형을 활용하였다. 모수 모형은 자료의 양이 적어도 비교적 정확하게 원하는 모수값을 찾아낼 수 있고, 특히 선형의 함수 형태를 가정하는 경우 해석이 수월하여 매우 광범위하게 활용된 모형이라 할 수 있다. 그러나 이러한 모수 모형은 설명변수의 독립성, 자료의 정규성, 종속변수와 설명변수 간의 선형성 등 엄격한 가정이 많아 추정가격의 신뢰성에 한계가 있었다. 본 연구에서는 이러한 비현실적인 통계적 가정을 부과하지 않는 보다 유연한 비모수 모형들을 적용하여 주택가격을 추정하였다. 또한 모형의 해석 가능성 등을 희생하더라도 추정된 주택가격의 정확성을 높일 수 있는, 예측 중심의 모형을 구축하고자 하였다.

대도시, 중소도시 및 군 지역을 대표할 수 있도록 서울시 강남구, 전주시 덕진구, 전라남도 해남군을 선정하여 단독주택 가격을 추정한 과정을 살펴보면 먼저, 기초자료 정제의 중요성을 간과할 수 없다. 분석에 활용한 실거래가 신고자료는 거래 당사자가 자발적으로 신고한 가격으로 많은 이상치와 잡음을 내포하고 있다. 본 연구에서는 신고가격과 공시가격과의 관계를 살펴 비정상적 신고로 의심되는 자료들을 제외하였다.

기계학습 분야에서 제시된 다양한 비모수 모형들을 주택가격 추정에 적용한 결과, SVM이나 MARS 등 최근에 개발된 모형들의 성능이 비교

적 우수한 것으로 나타나 이러한 모형들의 확대 적용이 필요한 것으로 보인다. 또한 지역 측면에서 강남구보다는 덕진구가, 덕진구보다는 해남군이 가격추정의 정확성이 떨어졌는데, 이는 농촌지역으로 갈수록 주택 집단의 동질성이 낮아지고 대신 이질성이 높아지기 때문인 것으로 풀이된다. 또한 모형이 어떠한 부분에서 특히 취약한지 효율적으로 파악하기 위해 본 연구에서 제시한 국지적 성능 진단을 수행하였으며, 토지 면적(또는 주택 규모)에 따른 자료 층화가 선행된 후 본격적인 모형 구축이 이루어질 경우 가격추정의 정확성이 높아질 것으로 파악되었다.

한편 기계학습 분야에서 제시된 이러한 비모수 모형들은 기본적으로 속성정보만을 고려할 뿐, 공간사상의 가장 큰 특징인 공간적 종속성을 반영하는데 관심이 적다. 본 연구에서는 비모수 모형에 공간적 종속성을 추가로 반영하기 위해 SVM의 scale parameter를 공간적 종속성이 미치는 한계 범위로 해석하여 모형을 정교화하였다. 또한 여러 비모수 모형에 일관되게 적용할 수 있도록 주변 주택가격의 평균적인 가격수준을 나타내는 공간차 변수 **WY**를 동원하여 공간적 종속성을 모형의 한 요소로 반영하였다.

주택에 대한 최종가격은 개별 모형들 중 가장 성능이 우수하게 나타난 모형의 예측치로 결정하는 대신, 개별 모형들에서 산출된 예측치를 가중평균하는 앙상블 평균을 적용하여 결정하였다. 앙상블 평균은 해남군과 같이 개별 모형들에서 산출된 예측치 간의 상관성이 낮은 경우 보다 탁월한 성과를 보였다.

마지막으로 본 연구에서 제시한 앙상블 예측가격과 현행 과세가치에 해당되는 공시가격을 비교하였고, 앙상블 예측가격이 전반적인 가격의 분포 측면이나 공간적 패턴에서 실제 거래가격과 유사함을 확인하였다. 특히 현행 공시가격은 일정 금액 이상의 고가 주택가격을 제대로 반영하지 못해, 공시가격에 역진성이 존재하는 것으로 추론할 수 있었다. 그러나 이와 같은 공시가격의 특징 내지 품질은 오로지 모형 자체로부터 기인하는 것은 아니므로 공시가격 산정 각 단계의 성격과 맥락을 고려하여 신중하게 해석할 필요가 있다.

2. 연구의 의의와 한계

이 연구가 가지는 의의는 다음과 같다. 첫째, 기계학습 분야에서 제시된 다양한 비모수 모형들을 주택가격 추정에 접목시키고자 하였고, 모형의 성능을 국지적으로 진단하여 특정 모형이 어떠한 부분에서 특히 취약한지 파악하고자 하였다. 지금까지 부동산 가격 추정에 관한 연구는 많이 이루어졌지만, MARS 등 비모수 모형의 논리에 근거한 연구는 거의 없는 것으로 보인다. 또한 모형 성능의 국지적 진단은 특정 모형이 가지는 약점을 손쉽고 빠르게 파악할 수 있게 하여 후속모형 개선에 상당한 기여를 할 것으로 보인다.

둘째, 일반적인 비모수 모형들을 적용하였을 때 발생할 수 있는 문제점, 즉 공간적 종속성의 간과와 같은 모형 설정의 오류를 피하기 위해 공간차 변수 등을 제안하는 등 이러한 문제를 보완하고자 하였다. 본 연구의 경우 공간적 종속성을 추가로 반영했을 때 모든 모형에서 성능이 개선된 것을 확인할 수 있었다.

셋째, 하나의 강한 모형보다는 약하더라도 다수의 모형을 결합한 앙상블 학습 개념을 부동산 가격추정에 적용하고자 하였다. 모든 부동산 유형 그리고 모든 지역에서 항상 우월한 성과를 보이는 수퍼 모형은 존재하지 않으며, 각 모형은 나름대로의 장단점을 가지고 있다. 어떤 모형은 비교적 편의가 작은 대신 분산이 클 수 있고, 또는 그 반대일 수도 있다. 또한 어떤 모형은 가격을 일관되게 과소추정할 수 있고, 또 다른 모형은 과대추정할 수 있다. 앙상블 학습은 개별 모형들이 갖는 단점을 상쇄시킬 수 있다는 점에서 최종가격 결정과정의 훌륭한 도구가 될 수 있다.

마지막으로 본 연구의 결과는 부동산 가격을 필요로 하는 다양한 사회 분야에 적용할 수 있는데, 특히 과세평가 분야에서 정책적 의의가 크다고 볼 수 있다. 즉 본 연구에서 제시한 앙상블 예측가격은 현행 공시가격과 달리 사람에 의한 편의가 적고 실제 거래가격에 근접한 가격을 제공할 수 있는 바, 공시제도 도입의 원 취지를 살릴 수 있을 것으로 보인다.

모수 모형이 갖는 한계점을 지적하고 비모수 모형이 갖는 이점을 중심으로 살펴 본 이 연구는 엄밀한 의미에서 모수 모형과 비모수 모형을

비교한 것이 아니라는 한계가 존재한다. 즉 모수 모형의 경우 OLS라는 모형을 기본 모형으로 제시한 것에 그치고, 연구의 대부분은 비모수 모형의 적용 및 정교화에 초점을 맞추었다. 그러나 OLS 모형을 개선하여 보다 우월한 모수 모형을 구축할 수도 있고, 이러한 모수 모형과 비모수 모형을 비교하여야 각 접근법의 장단점이 정확히 파악될 수 있을 것이다. 비모수 모형을 확대, 정교화할 수 있는 방법이 많은 만큼 모수 모형의 개선 방법 또한 다양하며, 본 연구에서는 OLS 모형과 같은 모수 모형의 정교화에 상대적으로 노력을 덜 경주하였다.

모수 모형은 데이터 양이 적어도 비교적 신뢰성 있게 수행될 수 있으며 통계적 추론이 수월하고 해석 또한 용이하다. 그러나 엄격한 통계적 가정이 많고 자료 특성을 지나치게 단순화하는 단점이 있다. 반면 비모수 모형은 설명변수의 독립성, 선형의 함수 형태 등 비현실적인 가정을 부과하지 않는 유연성이 있다. 그러나 비모수 모형 또한 여러 가지 단점을 가지고 있는데, 대량의 표본을 필요로 하는 점, 계산량이 방대한 점, 이해 가능성이 떨어지는 점 등을 들 수 있다. 따라서 각 접근법이 갖는 장단점과 과세평가 등 현장 실무에서 어떠한 방법이 실제 이용하기 수월한지 등에 대한 다각적 고려가 있어야 양 접근법의 적부를 판단할 수 있을 것으로 보인다. 그러나 그간 모수 모형에 집중되었던 부동산 가격추정 방법론을 비모수 모형으로 확대하였다는 점에서 본 연구의 의의를 찾을 수 있다고 본다.

참 고 문 헌

곽승준・전영섭. (1995). *환경의 경제적 가치*, 서울: 학현사.

김성우. (2010). 공간계량모형에 따른 주택 가격 추정에 관한 연구: 부산시 아파트 실거래가를 중심으로, *공공관리학보*, 24(3), 119-137.

김용창. (2008). 전사적 품질관리 접근에 의한 지가변동률통계의 품질평가 연구, *한국지역지리학회지*, 14(5), 553-572.

김종수. (2012). 실거래가격을 활용한 개별주택가격의 적정성 분석, *부동산연구*, 22(2), 29-56.

노희상, 박진수, 심규석, 유재은, 정연승. (2014). 생물/보건/의학 연구를 위한 비모수 베이지안 통계모형, *응용통계연구*, 27(6), 867-889.

민규식. (1994). 과세평가행정의 개선에 관한 연구: 부동산관련 지방세를 중심으로, 건국대학교 박사학위 논문.

박기호. (2004). 근린가중치행렬이 공간적 자기상관 추정에 미치는 영향: 서울시를 사례로, *서울도시연구*, 5(3), 67-83.

박헌수・정수연・노태욱. (2003). 공간계량경제모형을 이용한 아파트가격과 공간효과분석, *국토계획*, 38(5), 115-125.

서경천・이성호. (2001). 공간적 자기회귀모델과 토지시장분할에 의한 효율적 지가추정에 관한 연구, *국토계획*, 36(4), 1-18.

송용철・박헌수. (2012). 공간계량경제 접근방법을 이용한 농지가격추정에 관한 연구, *국토연구*, 72, 121-140.

안정근. (2004). 대량평가시스템에 의한 과세가치의 추계방법, *부동산학연구*, 10(2), 1-16.

안지아・박헌수. (2005). 공간종속성을 이용한 아파트 가격의 공간효과에 관한 연구, *대한국토・도시계획학회 정기학술대회*, 957-965.

원제무, 정광섭, 김상원, 백진호, 백기형. (2009). 시장세분화를 통한 주상복합주택 가격 결정요인 특성에 관한 연구, *국토계획*, 2009, 44(3), 137-147.

이건학・김감영. (2013). 개별공시지가와 주택실거래가의 공간적 불일치에 관한 연구, *대한지리학회지*, 48(6), 879-896.

이용만. (2008). 헤도닉 가격 모형에 대한 소고, *부동산학연구*, 14(1), 81-87.

이우진・방경식. (2006). 단독주택 과세의 수직 공평성 실증분석 및 불공평성 완화방안, *부동산연구*, 16(1), 121-145.

이정전. (2013). *토지경제학*. 서울: 박영사.

이창로・박기호. (2013). 인근지역 범위 설정이 공간회귀모형 적합에 미치는 영향, *대한지리학회지*, 48(6), 978-993.

임재만. (2010). 서울시 아파트 가격분위별 가격결정요인의 변동 추이에 관한 연구, *국토연구*, 41-56.

장희순・방경식. (2014). *부동산 용어사전*. 서울:부연사.

한국감정원. (2014). *단독주택 공시가격 통계 e-book*.

행정자치부. (2014). *안전행정 통계연보*.

허윤경. (2007). 도시별 주택가격의 공간적 영향력 검증: 서울과 부산의 아파트 가격을 중심으로, *주택연구*, 15(4), 5-23.

Abbott, D. (2014). *Applied Predictive Analytics: principles and techniques for the professional data analyst*, Wiley.

Anderson, D. E. (2000). Hypothesis testing in hedonic price estimation: on the selection of independent variables. *The Annals of Regional Science*, 34(2), 293-304.

Anselin, L. (1988). *Spatial econometrics: methods and models*. Dordrecht: Kluwer Academic Publishers.

Anselin, L., & Lozano-Gracia, N. (2008). Errors in variables and spatial effects in hedonic house price models of ambient air quality. *Empirical Economics*, 34(1), 5-34.

Appraisal Institute. (1996). *The appraisal of real estate*. Appraisal Institute, US.

Bae, C. H. C., Jun, M. J., & Park, H. (2003). The impact of Seoul's subway Line 5 on residential property values. *Transport Policy*, 10(2), 85-94.

Bajari, P., & Kahn, M. E. (2005). Estimating Housing Demand with an Application to Explaining Racial Segregation in Cities, *Journal of*

Business and Economic Statistics, 23, 20-33.

Banfield, R. E. et al. (2007). A comparison of decision tree ensemble creation techniques, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1), 173-180.

Bao, H. X., & Wan, A. T. (2004). On the use of spline smoothing in estimating hedonic housing price models: empirical evidence using Hong Kong data. *Real estate economics*, 32(3), 487-507.

Basak, D., Pal, S., & Patranabis, D. C. (2007). Support vector regression. *Neural Information Processing-Letters and Reviews*, 11(10), 203-224.

Basu, S., & Thibodeau, T. G. (1998). Analysis of spatial autocorrelation in house prices. *The Journal of Real Estate Finance and Economics*, 17(1), 61-85.

Bishop, K. (2008). *A dynamic model of location choice and hedonic valuation*. Unpublished, Washington University in St. Louis.

Bourassa, S. C., Cantoni, E., & Hoesli, M. (2007). Spatial Dependence, Housing Submarkets, and House Price Prediction, *Journal of Real Estate Finance and Economics*, 35(2), 143-160.

Box, G. E. P., & Cox, D. R. (1964). An Analysis of Transformations, *Journal of the Royal Statistical Society B* 26, 211-243.

Breiman, L. (2001). Random forests, *Machine Learning*, 45(1), 5-32.

Breiman, L., & Ihaka, R. (1984). *Nonlinear discriminant analysis via scaling and ACE*. Department of Statistics, University of California, Berkeley.

Brunauer, W., Lang, S., & Umlauf, N. (2013). Modelling house prices using multilevel structured additive regression. *Statistical Modelling*, 13(2), 95-123.

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2), 121-167.

Byvatov, E., Fechner, U., Sadowski, J., & Schneider, G. (2003). Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *Journal of Chemical Information and Computer Sciences*, 43(6), 1882-1889.

Can, A. (1992). Specification and estimation of hedonic housing price models. *Regional science and urban economics*, 22(3), 453-474.

Can, A., & Megbolugbe, I. (1997). Spatial dependence and house price index construction. *The Journal of Real Estate Finance and Economics*, 14(1-2), 203-222.

Cesare, C. M., & Ruddock, L. (1998). A new approach to the analysis of assessment equity. *Assessment Journal*, 5, 57-69.

Chang, C. C., & Lin, C. J. (2001). Training ν-support vector classifiers: theory and algorithms. *Neural computation*, 13(9), 2119-2147.

Chen, V. Y. J., Deng, W. S., Yang, T. C., & Matthews, S. A. (2012). Geographically weighted quantile regression (GWQR): An application to US mortality data. *Geographical Analysis*, 44(2), 134-150.

Chica-Olmo, J. (2007). Prediction of housing location price by a multivariate spatial method: cokriging. *Journal of Real Estate Research*, 29, 91–114.

Chica-Olmo, J., Cano-Guervos, R., & Chica-Olmo, M. (2013). A coregionalized model to predict housing prices. *Urban Geography*, 34(3), 395-412.

Chun, Y., & Griffith, D. A. (2013). *Spatial Statistics and Geostatistics: Theory and Applications for Geographic Information Science and Technology*. SAGE.

Clapp, J. M. (1990). A new test for equitable real estate tax assessment. *The Journal of Real Estate Finance and Economics*, 3(3), 233-249.

Colwell, P. F., Cannaday, R. E., & Wu, C. (1983). The analytical foundations of adjustment grid methods. *Real Estate Economics*, 11(1), 11-29.

Conway, T. M., & Lathrop, R. G. (2005). Modeling the ecological consequences of land-use policies in an urbanizing region. *Environmental Management*, 35(3), 278–.291.

Cressie, N. & Wikle, C. K. (2011). *Statistics for Spatio-temporal data*. New Jersey: John Wiley & Sons.

Crome, F. H. J., Thomas, M. R., & Moore, L. A. (1996). A novel Bayesian

approach to assessing impacts of rain forest logging. *Ecological Applications*, 1104-1123.

Cropper, M. L., Deck, L. B., & McConnell, K. E. (1988). On the choice of funtional form for hedonic price functions. *The Review of Economics and Statistics*, 668-675.

Crosby, N. (2000). Valuation accuracy, variation and bias in the context of standards and expectations. *Journal of Property Investment & Finance*, 18(2), 130-161.

Cubbin, J. (1974). Price, quality, and selling time in the housing market. *Applied Economics*, 6(3), 171-187.

Cui, D., & Curry, D. (2005). Prediction in marketing using the support vector machine. *Marketing Science*, 24(4), 595-615.

Curry, B., Morgan, P., &Silver, M. (2001). Hedonic regressions: mis-specification and neural networks. *Applied Economics*, 33(5), 659-671.

Curry, B., Morgan, P., &Silver, M. (2002). Neural networks and non-linear statistical methods: an application to the modelling of price‐quality relationships. *Computers & Operations Research*, 29(8), 951-969.

De Andrés, J., Lorca, P., de Cos Juez, F. J., & Sánchez-Lasheras, F. (2011). Bankruptcy forecasting: A hybrid approach using Fuzzy c-means clustering and Multivariate Adaptive Regression Splines (MARS). *Expert Systems with Applications*, 38(3), 1866-1875.

Downes, T. A., & Zabel, J. E. (2002). The impact of school characteristics on house prices: Chicago 1987‐1991. *Journal of Urban Economics*, 52(1), 1-25.

Dubé, J., & Legros, D. (2013). A spatio temporal measure of spatial dependence: An example using real estate data. *Papers in Regional Science*, 92(1), 19-30.

Dunse, N., & Jones, C. (1998). A hedonic price model of office rent. *Journal of Property Valuation and Investment*, 16(3), 297‐.312.

Du Preez, M., Lee, D. E., & Sale, M. (2013). Nonparametric estimation of a hedonic price model: A South African case study (No. 379).

Ekeland, I. (1988). *Mathematics of the Unexpected*. Chicago: University of Chicago Press

Fan, G. Z., Ong, S. E., & Koh, H. C. (2006). Determinants of House Price: A Decision Tree Approach. *Urban Studies*, 43(12), 2301-2315.

Farmer, M. C., & Lipscomb, C. A. (2010). Using quantile regression in hedonic analysis to reveal submarket competition. *Journal of Real Estate Research*, 32(4), 435-460.

Feldman, D., & Gross, S. (2005). Mortgage default: classification trees analysis. *The Journal of Real Estate Finance and Economics*, 30(4), 369-396.

Fik, T. J., Ling, D. C., & Mulligan, G. F. (2003). Modeling spatial variation in housing prices: a variable interaction approach. *Real Estate Economics*, 31(4), 623-646.

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2002). *Geographically weighted regression: the analysis of spatially varying relationships*. John Wiley & Sons.

Fraser, R., & Spencer, G. (1998). The value of an ocean view: an example of hedonic property amenity valuation. *Australian Geographical Studies*, 36(1), 94-98.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *Aannals of Statistics*, 1-67.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 1189-1232.

Gallimore, P. (2002). *The components of appraisal accuracy. In Real Estate Valuation Theory* (pp. 45-59). Dordrecht: Kluwer Academic Publishers.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.

Geniaux, G., & Napoléone, C. (2008). *Semi-parametric tools for spatial hedonic models: an introduction to mixed geographically weighted regression and geoadditive models*. In Hedonic Methods in Housing Markets (pp. 101-127). New York: Springer.

Getis, A., & Aldstadt, J. (2004). Constructing the Spatial Weights Matrix Using a Local Statistic, *Geographical Analysis*, 36(2), 90-104.

Gill, J., & Walker, L. D. (2005). Elicited priors for Bayesian model specifications in political science research. *Journal of Politics*, 67(3), 841-872.

Gloudemans, R. & Almy, R. (2011). *Fundamentals of Mass Appraisal*. Kansas City: IAAO

Goodman, A. C., & Thibodeau, T. G. (1995). Age-related heteroskedasticity in hedonic house price equations. *Journal of Housing Research*, 6(1), 25-42.

Goodman, A. C., & Thibodeau, T. G. (1997). Dwelling-age-related heteroskedasticity in hedonic house price equations: An extension. *Journal of Housing Research*, 8, 299-317.

Goodman, A. C., & Thibodeau, T. G. (2003). Housing market segmentation and hedonic prediction accuracy. *Journal of Housing Economics*, 12(3), 181-201.

Gopalakrishnan, S., Smith, M. D., Slott, J. M., & Murray, A. B. (2011). The value of disappearing beaches: a hedonic pricing model with endogenous beach width. *Journal of Environmental Economics and Management*, 61(3), 297-310.

Graves, A., Morris, J., Chatterton, J., Angus, A., Harris, J., Potschin, M., & Haines-Young, R. (2009). *Valuation of Natural Resources*. A NERC Scoping Study Final Report.

Grömping, U. (2009). Variable importance assessment in regression: linear regression versus random forest. *The American Statistician*, 63(4).

Gujarati, D., & Porter, D. C. (2004). *Basic econometrics* (ed.) McGraw-Hill.

Guo, L., Ma, Z., & Zhang, L. (2008). Comparison of bandwidth selection in application of geographically weighted regression: a case study, *Canadian Journal of Forest Research*, 38, 2526-2534.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*,

46(1-3), 389-422.

Haining, R., & Law, J. (2007). Combining police perceptions with police records of serious crime areas: a modelling approach. *Journal of the Royal Statistical Society: Series A* (Statistics in Society), 170(4), 1019-1034.

Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Techniques*, San Francisco: Morgan Kaufmann

Harding, J. P., Knight, J. R., & Sirmans, C. F. (2003). Estimating bargaining effects in hedonic models: evidence from the housing market. *Real estate economics*, 31(4), 601-622.

Hashem, S. (1997). Optimal linear combinations of neural networks. *Neural networks*, 10(4), 599-614.

Hastie, T., Friedman, J., & Tibshirani, R. (2009). *The elements of statistical learning*, Springer

Helbich, M., Brunauer, W., Hagenauer, J., & Leitner, M. (2013). Data-driven regionalization of housing markets. *Annals of the Association of American Geographers*, 103(4), 871-889.

Holt, D., Steel, D. G., Tranmer, M., & Wrigley, N. (1996). Aggregation and ecological effects in geographically based data. Geographical analysis, 28(3), 244-261.

Huang, B., Wu, B., & Barry, M. (2010). Geographically and temporally weighted regression for modeling spatio-temporal variation in house prices. *International Journal of Geographical Information Science*, 24(3), 383-401.

Huang, S. (2002). *Grid-Adjustment Approach - Modern Appraisal Technique. In Real Estate Valuation Theory* (pp. 341-354). Dordrecht: Kluwer Academic Publishers.

Hudson, G., & Wackernagel, H. (1994). Mapping temperature using kriging with external drift: theory and an example from Scotland. *International Journal of Climatology*, 14(1), 77-91.

Huh, S., & Kwak, S. J. (1997). The choice of functional form and variables

in the hedonic price model in Seoul. *Urban Studies*, 34(7), 989-998.

International Association of Assessing Officers (2008). *Standard on Mass Appraisal of Real Property*. Kansas City: IAAO.

International Association of Assessing Officers (2013). *Standard on Ratio Studies*. Kansas City: IAAO.

Isaaks, E. H., & Srivastava, R.M. (1989). *An Introduction to Applied Geostatistics*, New York: Oxford University Press.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. New York: Springer.

James, S., (2005). *The wisdom of crowds*. Random House LLC.

Jim, C. Y., & Chen, W. Y. (2009). Value of scenic views: Hedonic assessment of private housing in Hong Kong. *Landscape and urban planning*, 91(4), 226-234.

Jones, K., & Bullen, N. (1993). A Multi-level Analysis of the Variations in Domestic Property Prices: Southern England. *Urban Studies*, 30(8), 1409-1426.

Kagie, M., & Wezel, M. V. (2007). Hedonic price models and indices based on boosting applied to the Dutch housing market. Intelligent Systems in Accounting, *Finance and Management*, 15(3 4), 85-106.

Karato, K., Movshuk, O., & Shimizu, C. (2010). *Semiparametric Estimation of Time, Age and Cohort Effects in An Hedonic Model of House Prices*. Faculty of Economics, University of Toyama.

Kavousi-Fard, A., Samet, H., & Marzbani, F. (2014). A new hybrid Modified Firefly Algorithm and Support Vector Regression model for accurate Short Term Load Forecasting. *Expert Systems with Applications*, 41(13), 6047-6056.

Keith, T. J. (1991). Applying discounted cash flow analyses to land in transition. *Appraisal Journal*, 59(4), 458-470.

Kiel, K. A., & Zabel, J. E. (1996). House price differentials in US cities: Household and neighborhood racial effects. *Journal of housing economics*, 5(2), 143-165.

Kim, W. C., Phipps, T. T., & Anselin, L. (2003). Measuring the benefits of air quality improvement: a spatial hedonic approach. *Journal of environmental economics and management*, 45(1), 24-39.

Kinnard Jr, W. N., Mitchell, P. S., Beron, G. L., & Webb, J. R. (1991). Market Reactions to an Announced Release of Radioactive Materials: The Impact on Assessable Value. *Property Tax Journal*, 10(3), 283-297.

Kochin, L. A., & Parks, R. W. (1982). Vertical equity in real estate assessment: a fair appraisal. *Economic Inquiry*, 20(4), 511-532.

Kong, F., Yin, H., & Nakagoshi, N. (2007). Using GIS and landscape metrics in the hedonic price modeling of the amenity value of urban green space: A case study in Jinan City, China. *Landscape and Urban Planning*, 79, 240–.252

Kontrimas, V., & Verikas, A. (2011). The mass appraisal of the real estate by computational intelligence. *Applied Soft Computing*, 11(1), 443–448

Kostov, P. (2009). A spatial quantile regression hedonic model of agricultural land prices. *Spatial Economic Analysis*, 4(1), 53-72.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. New York: Springer.

Kummerow M., & Galfalvy, H. (2002). *Error Trade-offs in Regression Appraisal Methods. In Real Estate Valuation Theory* (pp. 105-131). Dordrecht: Kluwer Academic Publishers.

Kuntz, M., & Helbich, M. (2014). Geostatistical mapping of real estate prices: an empirical comparison of kriging and cokriging. International *Journal of Geographical Information Science*, DOI: 10.1080/13658816.2014.906041.

Kutner, M. H. Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models*, Singapore: McGraw Hill.

Lancaster, K. (1971). *Consumer demand: A new approach*. New York: Columbia University Press.

Landajo, M., Bilbao, C., & Bilbao, A. (2012). Nonparametric neural network modeling of hedonic prices in the housing market. *Empirical*

Economics, 42(3), 987-1009.

Lansford Jr, N. H., & Jones, L. L. (1995). Recreational and aesthetic value of water using hedonic price analysis. *Journal of Agricultural and Resource Economics*, 341-355.

Lasota, T., Łuczak, T., & Trawiński, B. (2011). *Investigation of random subspace and random forest methods applied to property valuation data*. In Computational Collective Intelligence. Technologies and Applications(pp. 142-151), Berlin Heidelberg: Springer.

Leesinger, J. (1969). Econometrics and Appraisal. *Appraisal Journal*, 37, 501-512.

Lee, T. S., Chiu, C. C., Chou, Y. C., & Lu, C. J. (2006). Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. *Computational Statistics & Data Analysis*, 50(4), 1113-1130.

Leggett, C. G., & Bockstael, N. E. (2000). Evidence of the effects of water quality on residential land prices. *Journal of Environmental Economics and Management*, 39(2), 121-144.

Levy, R., & Crawford, A. V. (2009). Incorporating Substantive Knowledge Into Regression Via A Bayesian Approach to Modeling. *Multiple Linear Regression Viewpoints*, 35(2), 4-9.

Li, Y. (2010). Can the spatial prediction of soil organic matter contents at various sampling scales be improved by using regression kriging with auxiliary information? *Geoderma*, 159(1), 63-75.

Liao, W. C., & Wang, X. (2012). Hedonic house prices and spatial quantile regression. *Journal of Housing Economics*, 21(1), 16-27.

Limsombunchai, V., C. Gan, & M. Lee. (2004). House Price Prediction: Hedonic Price Model Vs. Artificial Neural Network. *American Journal of Applied Sciences*, 1(3), 193-201.

Maclennan, D. (1977). Some Thoughts on the Nature and Purpose of House Price Studies. *Urban Studies*, 14, 5-71.

Mark, J. & Goldberg, M. A. (1988). Multiple Regression Analysis and Mass

Assessment: A Review of the Issues. *Appraisal Journal*, 56, 89-109.

Martin, T. G., Kuhnert, P. M., Mengersen, K., & Possingham, H. P. (2005). The power of expert opinion in ecological models using Bayesian methods: impact of grazing on birds. *Ecological Applications*, 15(1), 266-280.

Mason, C., & Quigley, J. M. (1996). Non parametric hedonic housing prices. *Housing studies*, 11(3), 373-385.

McCluskey, W. J., Deddis, W. G., Lamont, I. G., & Borst, R. A. (2000). The application of surface generated interpolation models for the prediction of residential property values. *Journal of Property Investment & Finance*, 18(2), 162-176.

McGreal, S., Adair, A., McBurney, D., & Patterson, D. (1998). Neural Networks: The Prediction of Residential Values, *Journal of Property Valuation and Investment*, 16, 57-70.

McMillen, D. P. (2010), Issues in Spatial Data Analysis, *Journal of Regional Science* 50, 119-141

Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 1436-1462.

Militino, A. F., Ugarte, M. D., & Garcia-Reinaldos, L. (2004). Alternative Models for Describing Spatial Dependence among Dwelling Selling Prices. *Journal of Real Estate Finance and Economics*, 29(2), 193-209.

Monteiro, A., Ribeiro, I., Tchepel, O., Carvalho, A., Martins, H., Sá, E., & Borrego, C. (2013). Ensemble Techniques to Improve Air Quality Assessment: Focus on O3 and PM. *Environmental Modeling & Assessment*, 18(3), 249-257.

Montero, J., & Larraz, B. (2011). Interpolation methods for geographical data: housing and commercial establishment markets. *Journal of Real Estate Research*, 33, 233-244.

Moore, J. W. (2006). Performance comparison of automated valuation models. *Journal of Property Tax Assessment and Administration*, 3(1), 43-60.

Morancho, A. B. (2003). A hedonic valuation of urban green areas. *Landscape and urban planning*, 66(1), 35-41.

Morgan, J. N., &Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American statistical association*, 58(302), 415-434.

Myers, C. K. (2004). Discrimination and neighborhood effects: Understanding racial differentials in US housing prices. *Journal of urban economics*, 56(2), 279-302.

Nelson, J. P. (2008). *Hedonic property value studies of transportation noise: aircraft and road traffic*. In Hedonic methods in housing markets (pp. 57-82). New York: Springer.

Nguyen, N., & Cripps, A. (2001). Predicting housing value: a comparison of multiple regression analysis and artificial neural networks. *Journal of Real Estate Research*, 22(3), 313-336.

Nijkamp, P., Rievald, P., & Rouwendal, J. (2011). *The market value of listed heritage: An urban economic application of spatial hedonic pricing*. Vrije Universiteit, Faculty of Economics and Business Administration.

Pace, R. K. (1993). Nonparametric methods with applications to hedonic models. *The Journal of Real Estate Finance and Economics*, 7(3), 185-204.

Pace, R. K. (1998). Appraisal using generalized additive models. *Journal of Real Estate Research*, 15(1), 77-99.

Pace, R. K., Sirmans, C. F., & Slawson Jr, V. C. (2002). *Automated valuation models*. In Real Estate Valuation Theory (pp. 133-156). Dordrecht: Kluwer Academic Publishers.

Paez, A., Fei, L. and Farber, S., 2008, Moving window approaches for hedonic price estimation: An empirical comparison of modeling techniques, *Urban Studies,* 45(8), 1565-1581.

Pagiola, S., Ritter, K., & Bishop, J. (2004). *Assessing the Economic value of Ecosystem Conservation.* Environmental department paper No. 101,

World Bank.

Parent, O., & vom Hofe, R. (2013). Understanding the impact of trails on residential property values in the presence of spatial dependence. *The Annals of Regional Science*, 51(2), 355-375.

Peterson, S., & Flanagan, A. B. (2009). Neural network hedonic pricing models in mass real estate appraisal. *Journal of Real Estate Research*, 31(2), 147-164.

Pinheiro, J., & Bates, D. (2006). Mixed-effects models in S and S-PLUS. Springer Science & Business Media.

Pope, J. C. (2008). Do seller disclosures affect property values? Buyer information and the hedonic model. *Land Economics*, 84(4), 551-572.

Quagrainie, K. K., McCluskey, J. J., & Loureiro, M. L. (2003). A latent structure approach to measuring reputation. *Southern Economic Journal*, 966-977.

Quinlan, J. R. (1993). *C4.5: programs for machine learning* (Vol. 1). San Mateo: Morgan Kaufmann.

Rakotomalala, R. (2005). *TANAGRA: a free software for research and academic purposes*, EGC 2005.

Raymond, Y. C. (2002). Estimating neighbourhood effects in house prices: towards a new hedonic model approach. *Urban studies*, 39(7), 1165-1180.

Redfearn, C. L. (2009). How Informative Are Average Effects? Hedonic Regression and Amenity Capitalization in Complex Urban Housing Markets, *Regional Science and Urban Economics,* 39, 297-306.

Reichert, A. K. (1997). Impact of a toxic waste superfund site on property values. *Appraisal Journal*, 65, 381-392.

Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review*, 33(1-2), 1-39.

Rosen, S. (1974). Hedonic prices and implicit markets: product differentiation in pure competition. *The journal of political economy*, 34-55.

Rossini, P. (1997). Artificial Neural Networks Versus Multiple Regression in the Valuation of Residential Property, *Australian Land Economics Review*, 3(1), 1-12.

Sathirathai, S., & Barbier, E. B. (2001). Valuing mangrove conservation in southern Thailand. *Contemporary Economic Policy*, 19(2), 109-122.

Schapire, R. E. (1999). Theoretical views of boosting. Computational Learning Theory, *Fourth European Conference*, EuroCOLT, 1-10.

Schulz, R., & Werwatz, A. (2004). A state space model for Berlin house prices: Estimation and economic interpretation. *The Journal of Real Estate Finance and Economics*, 28(1), 37-57.

Selim, H. (2009). Determinants of house prices in Turkey: Hedonic regression versus artificial neural network. *Expert Systems with Applications*, 36(2), 2843-2852.

Shmueli, G., & Koppius, O. R. (2011). Predictive analytics in information systems research. *Mis Quarterly*, 35(3), 553-572.

Shmueli, G. (2010). To Explain or to Predict?, *Statistical Science,* 25(3), 289-310.

Simons, R. A., Levin, W. B., & Sementelli, A. (1997). The effect of underground storage tanks on residential property values in Cuyahoga County, Ohio. *Journal of Real Estate Research*, 14(1), 29-42.

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and computing*, 14(3), 199-222.

Smolen, G. E., Moore, G., & Conway, L. V. (1992). Economic effects of hazardous chemical and proposed radioactive waste landfills on surrounding real estate values. *Journal of Real Estate Research*, 7(3), 283-295.

Spaninks, F. & Van Beukering, P. (1997). E*conomic valuation of mangrove ecosystems: potential and limitations*.

Speyrer, J. F., & Ragas, W. R. (1991). Housing prices and flood risk: an examination using spline regression. *The Journal of Real Estate Finance and Economics*, 4(4), 395-407.

Stewart, T. R., Roebber, P. J., & Bosart L. F. (1997). The importance of the task in analyzing expert judgment. *Organization Behavior and Human Decision Processes*, 69(3), 205-219.

Theebe, M. A. (2004). Planes, trains, and automobiles: the impact of traffic noise on house prices. *The Journal of Real Estate Finance and Economics*, 28(2-3), 209-234.

Tian, S., Yu, Y., & Guo, H. (2015). Variable selection and corporate bankruptcy forecasts. *Journal of Banking & Finance*, 52, 89-100.

Turnbull, G. K., & Sirmans, C. F. (1993). Information, search, and house prices. *Regional Science and Urban Economics*, 23(4), 545-557

Tyrväinen, L. (1997). The amenity value of the urban forest: an application of the hedonic pricing method. *Landscape and Urban planning*, 37(3), 211-222.

Vandell, K. D. (1991). Optimal Comparable Selection and Weighting. *Journal of the American Real Estate and Urban Economics Association*, 19(2), 213-239.

Vapnik, V. (1996). *The nature of statistical learning theory*. New York: Springer.

Wallace, N. E., & Meese, R. A. (1997). The construction of residential housing price indices: a comparison of repeat-sales, hedonic-regression, and hybrid approaches. *The Journal of Real Estate Finance and Economics*, 14(1-2), 51-73.

Waller, L. A. & Gotway, C. A. (2004). *Applied Spatial Statistics for Public Health Data*. Hoboken, NJ: John Wiley & Sons.

Walton, J. T. (2008). Subpixel Urban Land Cover Estimation. *Photogrammetric Engineering & Remote Sensing*, 74(10), 1213-1222.

Wang, Y. Q. (2008). Building credit scoring systems based on support-based support vector machine ensemble, *Fourth International Conference on Natural Computation*, 323-326.

Watkins, C. A. (2002). The Definition and Identification of Housing Submarket, *Environment and Planning A,* 33(12), 2235-2253.

Webster, R. & Oliver, M.A. (2007). *Geostatistics for Environmental Scientists, Statistics in Practice*, Chichester: John Wiley & Sons.

Weirick, W. N. & Ingram, F. J. (1990). Functional Form Choice in Applied Real Estate Analysis. *Appraisal Journal*, January, 57-73.

Wheeler, D. C., Páez, A., Spinney, J., & Waller, L. A. (2014). A Bayesian approach to hedonic price analysis. *Papers in Regional Science*. 93(3), 663-683.

Williams, N. (2005). Tsunami insight to mangrove value. *Current Biology,* 15(3), R73.

Woodward, R. T., & Wui, Y. S. (2001). The economic value of wetland services: a meta-analysis. *Ecological economics*, 37(2), 257-270.

Wu, J., Deng, Y., & Liu, H. (2014). House price index construction in the nascent housing market: the case of China. *The Journal of Real Estate Finance and Economics*, 48(3), 522-545.

Yavas, A., & Yang, S. (1995). The strategic role of listing price in marketing real estate: theory and evidence. *Real Estate Economics,* 23(3), 347-368.

Yoo, S., Jungho, I., & Wagner, J. E. (2012). Variable selection for hedonic model using machine learning approaches: A case study in Onondaga County, NY. *Landscape and Urban Planning*, 107(3), 293-306.

Zhao, P., & Yu, B. (2006). On model selection consistency of Lasso. T*he Journal of Machine Learning Research*, 7, 2541-2563.

Zhao, S., Hong, H. S., Zhang, L. P., & Chen, W. Q. (2007). Energy Value of Mangrove Ecosystem Services in China *[J]. Resources Science*, 1, 147-154.

Zhu, Q., & Lin, H. S. (2010). Comparing ordinary kriging and regression kriging for soil properties in contrasting landscapes. *Pedosphere*, 20(5), 594-606.

Zietz, J., Zietz, E. N., & Sirmans, G. S. (2008). Determinants of house prices: a quantile regression approach. *The Journal of Real Estate Finance and Economics*, 37(4), 317-333.

Zurada, J., Levitan, A. S., & Guan, J. (2011). A Comparison of Regression and Artificial Intelligence Methods in a Mass Appraisal Context. *Journal of Real Estate Research*, 33(3), 349-387.

Abstract

Estimating single-family house prices using non-parametric spatial models and an ensemble learning approach

Lee, Changro

Department of Geography, The Graduate School

Seoul National University

Estimation models of property price are being used in a variety of fields due to increasing openness and availability of property data. For example, they are employed for asset portfolio, estimation of collateral value, feasibility analysis of a proposed property development, etc. In particular, property assessment is one of the important application areas in that its impact is spread over the whole taxpayers.

Property assessment in S. Korea, however, has been blamed for low price level and poor price equitability from the past. The reason for this blame lies in the fact that assessment process, that is, the price estimation models are not appropriate, though tax payers' appeals and other political factors contributed also to this problem. This study was initiated from an effort to investigate more precise methodology for price prediction.

In quantitative social science studies, explanatory modeling was main trend and explanatory power was considered to be identical to predictive power, which is not true. This study attempts to construct predictive models focusing on predictive accuracy of new observations.

Most traditional models such as a linear regression belong to parametric approach and impose rigid assumptions including independence of explanatory variables, normality of data, and linear functional form.

Predictive models have been developed in machine learning field, which do not impose the unrealistic assumptions above. Most of these models belong to non-parametric approach.

This study tries to expand prediction methodology from parametric models to non-parametric ones. In addition, we try to introduce the concept of ensemble learning while determining final estimate of property price. Lastly, we intend to predict house prices in case study areas and compare them with actual sales prices and assessed ones.

Actual sales data between 2011 and 2014 were used in the study. Study areas were chosen as followings to represent a large city, a middle-sized city and a rural area respectively: Gangnam-gu in Seoul, Dukjin-gu in Jeonju, and Haenam-gun in Jeonnam. Then non-parametric models are applied to the dataset sequentially, and the main findings are as followings: First, newly developed models such as MARS(multivariate adaptive regression splines) or SVM(support vector machine) show excellent performance of prediction, and thus active application of these algorithms seems to be urgent. Second, house prices in rural area were found to be more hard to predict because of heterogeneity in rural house groups.

This study also proposed local diagnostic approach for model performance. This local approach can identify at which parts a model performs poorly or competently. The result from the case study reveals that houses with land size being far below or much above average size are difficult to predict. This implies that house data should be stratified based on land area before going to main analysis.

All the non-parametric models utilized so far do not take into account spatial dependence, an essential feature of spatial data. Thus, this study incorporates spatial dependence into model specification through the use of a scale parameter in SVM, and a spatially lagged variable, $\boldsymbol{WY}$.

An ensemble averaging technique was used to decide a final estimate of price. The ensemble averaging was implemented through the use of a neural network model. The result shows that the ensemble averaging performs more

and more accurately as it predicts from in large city to rural area. The price estimates from individual models are more diverse in rural area than those in large city, and this leads to excellent performance of prediction in the rural area.

Finally, the ensemble averaging was compared with current assessed price, and the result shows that the ensemble averaging was more consistent with actual sales price than assessed price in terms of price distribution and spatial pattern. However, the quality of assessed price should be interpreted with due care because the deficiency of assessed price can be attributed to various assessment procedures including field survey and tax payers' appeals.

keywords : machine learning, non-parametric model, spatial dependence, ensemble learning, assessed price
Student Number **: 2012-30841**