대한지리학회지 제51권 제2호 2016(219~233)

단독주택가격 추정을 위한 기계학습 모형의 응용

이창로* · 박기호**

Application of machine learning models for estimating house price

Chang Ro Lee* · Key Ho Park**

요약 : 수리 또는 계량적 모형을 사용하는 사회과학연구에서 분석의 초점은 종속변수와 설명변수의 관계를 밝히는 것, 즉 설명 중심의 모형(explanatory modeling)이 지금까지 주류를 이루었다. 반면 예측(prediction) 능력 제고에 초점을 맞춘 분석은 드물었다. 본 연구에서는 이론 및 가설을 검증하거나 변수 간의 관계를 밝히는 설명 중심의 모형이 아니라 신규 관찰치에 대한 예측 오차를 줄이는, 예측 중심의 비모수 모형(non-parametric model)을 검토하였다. 서울시 강남구를 사례지역으로 선정한 후, 2011년부터 2014년까지 신고된 단독주택 실거래가를 기초자료로 하여 주택가격을 추정하였다. 적용한 비모수 모형은 기계학습 분야에서 제시된 일반가산모형(generalized additive model), 랜덤 포리스트, MARS(multivariate adaptive regression splines), SVM(support vector machines) 등이며 비교적 최근에 개발된 MARS나 SVM의 예측력이 뛰어남을 확인할 수 있었다. 마지막으로 이러한 비모수 모형에 공간적 자기상관성을 추가적으로 반영한 결과, 모형의 가격 예측력이 보다 개선되었음을 알 수 있었다. 본 연구를 계기로 그간 모수 모형에 집중되었던 부동산 가격추정 방법론이 비모수 모형으로 확대 및 다양화되기를 기대한다.

주요어 : 기계학습, 예측 중심의 모형, 비모수 모형, 공간적 자기상관성, 주택가격

Abstract : In social science fields, statistical models are used almost exclusively for causal explanation, and explanatory modeling has been a mainstream until now. In contrast, predictive modeling has been rare in the fields. Hence, we focus on constructing the predictive non-parametric model, instead of the explanatory model. Gangnam-gu, Seoul was chosen as a study area and we collected single-family house sales data sold between 2011 and 2014. We applied non-parametric models proposed in machine learning area including generalized additive model(GAM), random forest, multivariate adaptive regression splines(MARS) and support vector machines(SVM). Models developed recently such as MARS and SVM were found to be superior in predictive power for house price estimation. Finally, spatial autocorrelation was accounted for in the non-parametric models additionally, and the result showed that their predictive power was enhanced further. We hope that this study will prompt methodology for property price estimation to be extended from traditional parametric models into non-parametric ones.

Key Words : machine learning, predictive modeling, non-parametric model, spatial autocorrelation, house price

---

본 논문은 제1저자의 박사학위 논문「비모수 공간모형과 앙상블 학습에 기초한 단독주택가격 추정」(2015.8) 중 일부를 요약·수정한 것임.

* 서울대학교 국토문제연구소 연구원(Researcher, Institute for Korean Regional Studies), spatialstat@naver.com

** 서울대학교 지리학과 교수 및 국토문제연구소 겸무 연구원(Professor, Department of Geography, Seoul National University, and Researcher, Institute for Korean Regional Studies), khp@snu.ac.kr

1. 서론

지금까지 수리 또는 계량적 모형을 사용하는 사회과학연구에서 분석의 초점은 종속변수와 설명변수의 관계를 추론하는 것이 대부분이었다. 즉 설명 중심의 모형(explanatory/exploratory modeling)이 주류를 이루었다. 반면 예측(prediction) 능력 제고에 초점을 맞춘 분석은 드물었다.[1)]

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 변수 간의 관계를 밝히는 설명 중심의 모형이 아니라 신규 관찰치에 대한 예측 오차를 줄이는, 예측 중심의 모형(predictive modeling)을 단독주택 가격 추정에 적용하고자 한다. 즉 주택가격 형성에 영향을 미치는 요소와 그 한계효과(marginal effects)를 측정하는 설명모형 대신, 단독주택 가격을 정확히 예측하여 실거래 가격과의 오차를 최소화하는 예측모형을 설계하고자 한다. 또한 이 과정에서 주택가격과 같은 공간사상(空間事象)의 특징인 공간적 자기상관성(spatial autocorrelation)을 표현한 변수가 예측모형의 변수로서 유용한지 살피고자 한다.

예측 중심의 모형은 목적에서부터 모형 진단에 이르기까지 여러 측면에서 설명 중심의 모형과 차이점을 보인다. 표 1은 설명 중심의 모형과 예측 중심의 모형을 비교한 것인데, 특히 설명 중심의 모형은 연구자의 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.

표 1에서 예측 중심의 모형은 과다적합(over-fitting)이 가장 신경을 써야 하는 위험이고, 따라서 이러한 과다적합 위험을 피하기 위해 검증 데이터(test data)를 기준으로 오차 정도를 가늠한다. 즉 모형 적합에 사용된 훈련 데이터(train data)를 기준으로 예측 오차를 계산할 경우 모형의 복잡도가 증가할수록(예를 들어 설명변수 추가) 예측 오차는 항상 감소하기 마련이며 이는 과다적합으로 이어진다. 그러나 검증 데이터를 기준으로 계산한 예측 오차("out-of-sample error")는 모형이 과도하게 복잡해질 경우 오히려 증가하는 경향이 나타나게 되어, 과다적합의 위험을 사전에 파악할 수 있다.

또한 최적 모형의 결정 기준에서도 설명 중심의 모형에서는 편의를 최소화하는 모형이 가장 바람직하지만, 예측 중심의 모형은 편의와 함께 분산을 최소화하는 것이 목적이므로 경우에 따라 편의 추정량이라 하더라도 분산을 현격하게 줄일 수 있다면 그러한 추정량을 사용하기도 한다. 능형회귀(ridge regression)나 LASSO(least absolute shrinkage and selection operator) 회귀가 편의 추정량을 사용하는 예라 할 수 있다.

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, 213). 비모수 모형은 설명변수와 모수의 결합형태가 사전에 정해진 형태(선형 함수 등)를 취하지 않고, 데이터가 가진 정보로부터 함수 형태를 추출하게 된다.

예를 들어 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(Kummerow and Galfalvy, 2002). 이러한 경우 비모수 모형은 모수 모형(parametric model)의 대안으로 기능할 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 초기의 OLS(Ordinary Least Squares) 회귀모형에서부터 시작하여 이후 공간적 자기상관성을 계량화하여 모형의 구성요소로 반영한 공간회귀모형, 공간보간법의 일종인 크리깅(kriging) 기법 등

이 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.

이러한 모형들은 대부분 모수 모형으로서 설명변수의 독립성, 자료의 정규성, 종속변수와 설명변수 간의 선형성(linearity) 등 엄격한 가정이 많고 자료 특성을 단순화하는 특징이 있다(Ekeland, 1988; Kummerow and Galfalvy, 2002; Gloudemans and Almy, 2011). 이러한 접근법이 연구 맥락에 따라 유용한 경우도 있지만 자료가 정규성에서 크게 벗어나거나 종속변수와 설명변수 간의 관계가 선형이 아닌 복잡한 비선형 관계 등일 때에는 비모수 모형을 적용하는 것이 신규 관찰치 예측에 보다 유리할 수 있다.

특히 종속변수인 부동산 가격과 설명변수인 부동산 특성(면적, 건물구조 등) 사이에 선형의 함수 관계가 성립되는 것은 예외적인 경우에 해당되어 추정가격의 정확성을 떨어뜨릴 가능성이 높다(Weirick and Ingram, 1990; Hastie *et al.*, 2009, 139).

이러한 선형성 가정 등에서 자유로운 모형들이 최근 기계학습(machine learning) 분야에서 다양하게 개발되었으며, 트리기반 모형(tree-based model), MARS(multivariate adaptive regression spline), SVM(support vector machines) 등이 그 예라고 할 수 있다. 기계학습 분야에서 제시된 이러한 예측 중심의 모형들은 대부분 비모수적 방법에 해당되어 종속변수와 설명변수 간 선형의 함수 형태를 고집하지 않는다. 따라서 부동산 가격 추정에 있어 또 하나의 실질적인 대안이 될 수 있을 것으로 판단된다.

전통적인 부동산 가격추정 모형은 설명 중심의 모형 및 모수 모형이었다. 본 연구에서는 예측 중심의 비모수 모형(predictive non-parametric model)을 활용하여 단독주택가격을 추정하고자 한다. 이후 비모수 모형을 통해 산출된 단독주택 가격과 기존 모수 모형의 예측치를 비교하여 정확성 개선 여부를 확인하고, 마지막으로 단독주택가격이 갖는 공간적 자기상관성을 모형의 구성요소로 반영함으로써 가격의 정확성이 추가적으로 개선될 수 있는지 살핀다. 이를 통해 비모수 모형이 갖는 이점 및 공간적 자기상관성의 모형 반영 필요성을 실증적으로 밝히고자 한다.

2. 기계학습 관련 선행연구 검토

1) 기계학습과 비모수 모형

기계학습은 인공지능의 한 분야로 데이터로부터 학습하여 알고리즘을 구축하고, 개발된 알고리즘에 기초하여 관측되지 않은 자료에 대해 목표값(target value)을 예측한다. 기계학습 분야에서 제시된 이러한 예측 중심의 모형들은 사전에 정해진 엄격한 규칙을

표 1. 설명 중심의 모형과 예측 중심의 모형 비교

구분	설명 중심의 모형	예측 중심의 모형
분석 목적	이론이나 가설의 검증	신규 관찰치의 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 R2, 계수의 통계적 유의성(p-value), 잔차, 다중공선성	검증 데이터(Test Data)를 기준한 예측치의 오차 정도(MSE 등)

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

따르지 않는 특징이 있다.

예를 들어 부동산 가격 예측에 흔히 사용되는 헤도닉 가격 모형(hedonic pricing model)의 경우 가격과 부동산 특성 변수 간에 선형의 함수형태를 가정하는 경우가 많다. 즉 사전에 정해진 암묵적인 규칙이 존재한다. 그러나 경제학 측면에서 함수 형태가 선형이라는 것은 변수 값의 양에 관계없이 해당 특성(속성)이 발휘하는 한계가치(marginal utility)가 불변임을 의미한다. 하지만 이는 경제학의 한계효용 체감법칙에 배치될 뿐만 아니라(Maclennan, 1977) 일반적인 직관에도 반한다. 어떠한 재화이든 존재량이 많아지면 희소성이 떨어져 가치는 하락하기 때문이다. 토지 면적이 증가할수록 거래단가가 하락하는 것이 대표적인 예라할 수 있다.

또한, 이론적으로 가격함수는 수많은 주택가격 결정요인(거주지역에 대한 주민들의 애착심, 해당 지역의 장래 개발 가능성, 혐오시설이나 인구유입시설 등 신규시설의 설치 여부 등)과의 관계를 모두 포함하는 개념이므로 그 형태를 정확히 파악하는 것은 불가능에 가깝다(Mason and Quigley, 1996).

따라서 가격함수의 형태는 이론적 문제라기보다는 실증적 문제에 해당하며 가격함수 f를 온전히 데이터에 기반하여 찾으려는 비모수 모형들이 기계학습 분야에서 활발하게 시도되고 있다. 이하에서는 이러한 기계학습 분야에서 개발된 비모수 모형의 종류 중 본 논문에서 시도한 모형에 대해 간략하게 설명한다.

2) 비모수 모형의 종류

① 일반가산모형(generalized additive model, GAM)
통상적인 선형회귀모형을 아래와 같이 표현한다면,

$$y_i=\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+...+\beta_p x_{ip}+\varepsilon_i \quad (1)$$

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.

$$\begin{aligned} y_i&=\beta_0+f_1(x_{i1})+f_2(x_{i2})+...+f_p(x_{ip})+\varepsilon_i \\ &=\beta_0+\sum_{j=1}^{p} f_j(x_{ij})+\varepsilon_i \end{aligned} \quad (2)$$

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(polynomial regression) 등 여러 가지 방법을 동원할 수 있고, 또한 한 가지 방법에 국한하지 않고 서로 다른 방법을 설명변수별로 각각 사용할 수도 있다.

GAM은 Pace(1998)가 주택 가격 추정에 적용하여 다항회귀모형보다 가격 예측력이 우수함을 보인 이후 지속적으로 부동산 가격추정에 사용되고 있다. Bao and 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)[2]에 초점을 두어 이 두 가지 특성이 주택가격 형성에 미치는 영향이 선형이 아님을 보였다.

② 랜덤 포리스트(random forest)

랜덤 포리스트는 회귀트리 모형(regression tree model)을 기초로 한다. 회귀트리 모형은 설명변수 공간(predictor space)을 분할하는 것으로부터 시작한다. 즉, 설명변수 X_1, X_2, ..., X_p를 J개의 지역(Region) R_1, R_2, ..., R_j로 서로 겹치지 않게 분할한다. 다음으로 R_j 지역에 속하는 관찰치에 대해서는 R_j 지역 관찰치 평균값을 예측치로 제시하게 된다. R_j 지역은 잔차제곱합(Residual Sum of Squares, RSS)이 최소가 되도록 분할한다.

그러나 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.

$$\sum_{m=1}^{|T|}\sum_{x_i\in R_m}(y_i-\hat{y_{R_m}})^2+\alpha|T| \quad (3)$$

위 식에서 $|T|$는 트리 T의 가지(terminal node) 수를, R_m은 m번째 가지에 해당하는 분할지역을 의미한다. α는 동조 파라미터(tuning parameter)로서 α=0인 경우 아무런 패널티가 없으므로 최대 트리가 되며, α가 커질수록 트리 규모는 작아진다. α값은 교차 검증(cross-validation) 등을 통해 정한다.

트리기반 모형은 개념이 단순하고 시각적으로 표현하기 수월하며, 따라서 해석하기도 용이하다. 또한 비전문가에게도 매우 쉽게 설명할 수 있다. 반면 다른 비선형 모형에 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 접근 중 하나가 랜덤 포리스트이다.

랜덤 포리스트는 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 정한다. 랜덤 포리스트는 최초 설명변수의 수 p보다 적은 m개의 설명변수를 사용하는 특징이 있다(통상 $m\approx\sqrt{p}$).

랜덤 포리스트를 적용한 최근의 연구로 폴란드 주택가격을 예측한 Lasota *et al.*(2011)을 들 수 있다. 동 연구는 선형회귀모형을 기본 모형(null model)으로 하여 랜덤 포리스트의 성능이 상대적으로 우수함을 밝혔다.

③ MARS(multivariate adaptive regression splines)

MARS의 기본적인 형태는 다음과 같다(Friedman, 1991).

$$\hat{f}(x)=\sum_{i=1}^{k} c_i B_i(x) \tag{4}$$

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.[3] 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, 325).

$$\mathrm{GCV}(\lambda)=\frac{\sum_{i=1}^{N}[y_i-\hat{f}_\lambda(x_i)]^2}{[1-M(\lambda)/N]^2} \tag{5}$$

위 식에서 분자는 잔차제곱합을 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. 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.

④ SVM(support vector machines)

SVM은 이미지 분류나 패턴 인식처럼 분류(classification)의 문제를 처리하기 위해 1990년대에 제시된 기계학습 알고리즘 중의 하나이다(Vapnik, 1996). 분류, 즉 이진 종속변수가 아닌 연속형 종속변수에 SVM을 적용하는 경우에도 hyper-plane, maximal margin 등 그 개념이나 논리는 동일하다.[4]

선형회귀모형 $f(x)=x^T\beta+\beta_0$에서 β를 추정하기 위해 SVM은 다음의 식(6)을 최소화하며, 이때의 V는 식(7)과 같이 정의된다(Hastie *et al.*, 2009, 434).

$$\mathrm{H}(\beta,\beta_0)=\sum_{i=1}^{N} V(y_i-f(x_i))+\frac{\lambda}{2}\|\beta\|^2 \tag{6}$$

$$\mathrm{V}_\varepsilon(r)=\begin{cases}0 & |r|<\varepsilon\\ |r|-\varepsilon & o/w\end{cases} \tag{7}$$

즉, ε보다 작은 오차는 무시한다는 의미로서('ε-insensitive error'), 분류의 문제를 다루는 SVM의 논리와 유사하다. λ는 일반적인 패널티 또는 동조 파라미터로서 교차 검증을 통해 산출된다.

SVM이 최초로 개발·제시된 패턴 인식 분야 외에 응용 사례를 마케팅(Cui and Curry, 2005)이나 임상의학(Guyon *et al.*, 2002) 분야에서 쉽게 찾아볼 수 있다. 하지만 부동산 분야에서는 쉽게 찾아보기 힘든데, 김종수·이성근(2012)의 경우 SVM을 활용하여 공업용 부동산의 가격을 추정한 후 기존 회귀모형보다 가격 예측력이 우수함을 보인 바 있다.

표 2는 지금까지 설명한 비모수 모형의 특징과 장단점을 정리하여 제시한 것인데, 대체로 비선형 관계의 포착, 고차원 데이터 분석 등에 장점이 있는 반면 계산량 과다, 함수형태의 복잡성 등이 단점이라고 말할 수 있다.

3) 공간적 자기상관성의 반영

① 공간가중행렬 W

기계학습 분야에서 제시된 예측 중심의 비모수 모형은 주택가격과 같은 공간사상이 가지는 독특한 특징, 즉 공간적 자기상관성에 대해 관심이 상대적으로 적은 편이다. 그러나 자료에 공간적 자기상관성이 존재함에도 불구하고 이를 고려하지 않을 경우 예측 결과는 부정확해질 수밖에 없다. 본 절에서는 앞절에서 설명한 비모수 모형에 공간적 자기상관성을 명시적으로 반영할 수 있는 방법에 대해 검토한다.

공간적 자기상관성을 정량화하기 위한 대표적인 도구로 공간가중행렬을 들 수 있다. 공간가중행렬은 n개의 관찰치에 대한 n×n 양행렬(Positive Matrix)로서, 구성요소 w_{ij}는 관찰치 i 및 j 간에 부여된 가중치를 의미한다. 공간가중행렬은 인접성(contiguity) 척도 또는 거리(distance) 척도에 따라 다양하게 구성할 수 있으며 분석의 목적이나 상황에 따라 바람직한 공간가중행렬의 종류는 상이하다. 다만 다른 조건이 동일하다면 각 지점의 좌표나 지점 간 거리를 알 수 있는 경우 인접성 척도보다 거리 척도를 이용한 가중행렬이 보다 바람직할 수 있다(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. 거리조락함수는 다음과 같은 형태로 표현할 수 있다.

$$w_{ij}=f(d_{ij}, b) \tag{8}$$

여기서 d_{ij}와 w_{ij}는 관찰치 i 및 관찰치 j간의 거리 및

표 2. 비모수 모형의 특징과 장단점

모형의 종류	특징	장점	단점
GAM	선형결합 $X\beta$를 spline 함수 $f(x)$로 대체한 후 각 설명변수를 가산하여 모형을 구성	매우 복잡한 비선형 관계를 포착하는데 유리	span, knot 등 사전에 동조 파라미터값을 정해야 함
RF	회귀트리 모형의 변동성(variance)을 줄이기 위한 앙상블 기법의 하나	– 회귀트리 모형의 단순 반복수행에 기반하므로 계산속도가 빠르고 대량 데이터에도 적용 가능 – GAM 등 다른 모형처럼 동조 파라미터값을 정할 필요 없음	연속형 종속변수 예측에 사용될 경우, 자료에 나타난 범위값을 벗어난 예측은 할 수 없음
MARS	기저함수의 구성을 통해 자료를 분류하거나 예측	기저함수의 모형 기여도를 측정할 수 있으므로 변수들의 상호작용 효과를 포착하는데 효율적	계산량이 많아 대량 데이터 적용에 한계
SVM	최대 여백(max. margin)을 가진 선형평면(linear plane)에 기반하여 자료를 분류하거나 예측	– 고차원 데이터(자료수 대비 변수가 많은 데이터) 취급 용이 – 커널함수 사용시 변수 간 비선형 관계 처리 용이	함수형태가 단순하지 않아 결과 해석에 어려움 존재

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 아래와 같은 식으로 나타낼 수 있다(이창로 · 박기호, 2013).

$$w_{ij}=1/(d_{ij})^{\alpha} \quad (9)$$

$$w_{ij}=\exp(-\beta d_{ij}) \quad (10)$$

최적의 함수 형태 $f(\bullet)$를 결정하기 위한 연구가 여럿 있었으나 널리 받아들여지는 의견이나 결론은 없는 것으로 보인다. 공간가중행렬의 구성과 관련된 최근의 연구 결과를 보면 국내의 경우 대부분 거리 또는 거리 제곱에 반비례하도록 가중치를 부여하였으며, 해외의 경우 거리의 멱함수, 커널함수 등의 시도가 이루어졌다(안지아 · 박헌수, 2005; Guo *et al.*, 2008).

② 공간차 변수 WY

위에서 구성한 W에 종속변수(본 연구에서는 주택가격) Y를 곱하면 공간차 변수(Spatially Lagged Variable) WY가 되며, 이는 예측하려는 대상 주택의 인근에 위치한 주택들의 평균적인 가격수준을 의미한다. 특정 지점 i에서의 주택가격은 건물구조, 신축연도 등과 같은 대상 주택의 일반적인 속성 뿐 아니라 인근에 위치한 주택가격의 영향도 크게 받을 수밖에 없다. 따라서 공간차 변수 WY를 구성하여 추가적인 설명변수로 동원할 경우, 이는 주택가격의 공간적 자기상관성을 명시적으로 고려하는 것이 된다. 이러한 접근은 아래 식의 공간시차모형(spatially lagged model)의 논리와 그 맥락을 같이 한다고 볼 수 있다.

$$y=\rho Wy+\beta X+\varepsilon$$
$$\varepsilon\sim N(0,\sigma^2 I) \quad (11)$$

상기 식을 보면 주택가격 Y를 예측하기 위해 일반적인 속성변수 X뿐 아니라 주변의 주택가격 수준 Wy도 함께 고려함을 알 수 있다.

공간적 의존성을 고려하는 소위 'spatial model'에는 상기의 공간시차모형(SLM) 외에도 SEM(spatial error model), hierarchical model, non-stationary model 등 광범위하다. 공간모형 전체를 아울러 검토하는 것은 본 연구의 범위에서 벗어난다. 다만 공간적 자기상관성을 표현한 WY를 설명변수로 동원하는 것이 예측모형의 오차를 줄일 수 있는지 제한적으로 검토하고자 한다.

3. 사례 분석

1) 데이터 및 기초 통계량

본 연구에서는 서울시 강남구에 소재한 단독주택을 대상으로 가격을 추정하였으며 가격 추정을 위한 투입자료로 2011년부터 2014년까지의 실거래가 신고자료를 사용하였다. 단독주택은 아파트에 비해 가격추정모형이 시도된 사례가 극히 적어 감정평가나 공시가격 추정시 더 많은 어려움을 초래하는 부동산 유형이라 할 수 있다. 이러한 단독주택을 대상으로 가격을 추정한 점이 본 연구의 차별성 중 하나라 할 수 있다. 최초의 단독주택 신고건수는 488건이었으나 지분거래, 중복신고건 등을 제외한 438건을 대상으로 분석을 실시하였다.[5)]

주택과 관련된 구득 가능한 설명변수는 13개였으나 OLS의 반복적 적합과정을 통해 9개 항목(용도지역, 도로접면, 방위, 경과연수, 건물 연면적, 토지면적, 건물구조, 지붕구조, 인근지역 특징)을 유의한 설명변수로 최종 결정하였다. 나머지 4개 항목(형상, 경사도, 空家 여부, 거래연도)은 통계적으로 유의하지 않거나 직관과 반하는 결과가 나와 제외하였다. 종속변수는 거래금액(단위:억원)을 사용하였으며, 기초통계량은 표 3과 같고 최종 OLS 적합 결과는 표 4와 같다(조정 결정계수 R^2=61.6%)[6)].

표 4 설명변수에서 p-value 기준으로 유의하지 않더라도 계수의 부호가 일반적인 직관과 일치하는 경우 설명변수로 포함시켰다. 예를 들어 방위에서 "남향 외" 주택의 회귀계수는 −0.42로서 통계적 측면에서 유의하지 않지만(p-value 0.57)은 남향 주택 대비

표 3. 강남구 단독주택 실거래가 기초 통계량(438건)

구분	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)	

표 4. OLS 모형 적합 결과

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	
경과연수2		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	

가격 수준이 낮게 형성된다는 것은 일반적인 상식에 부합하므로 설명변수로 포함시켰다. 주택가격 결정요인은 가격시점 당시의 금리, 물가 등과 같은 외생변수, 구매자 심리 변수, 경매물건인지 여부 등 맥락과 시각에 따라 다양한 변수와 형태의 모형 구성이 가능하지만 본 연구에서는 단독주택가격 감정평가시 현실적 기준으로 삼는 변수에 국한하여 모형을 구성하였다.[7)]

2) 비모수 모형의 적합

표 4에서 제시된 설명변수를 투입변수로 하여 GAM, 랜덤 포리스트, MARS, SVM을 순차적으로 적용하였으며, 표 5는 GAM의 적합 결과를 보여준다. 설명변수 중 경과연수와 건물 연면적은 선형이 아닌 비선형으로 모형을 구성하였음을 알 수 있다. 그림 1은 이러한 비선형 효과를 시각적으로 보여주는데, 경과연수(age)의 경우 신축 후 약 30년까지는 주택가격이 완만하게 하락하다가 이후에는 재건축·재개발

표 5. GAM 적합 결과

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

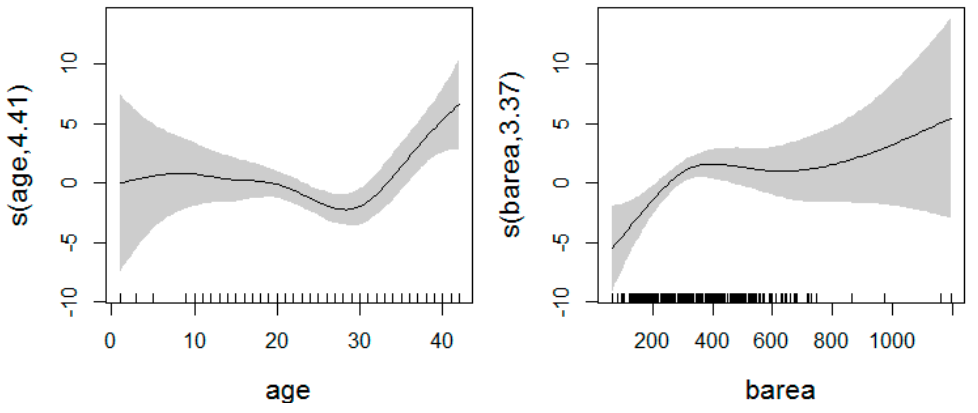

그림 1. 경과연수(age) 및 건물 연면적(barea)의 비선형 효과

* 그림에서 수직축은 경과연수 및 건물 연면적 변수에 대한 평활화 함수(smoothing function)를 보여주며, ( )안의 수치는 각 변수에 대한 추정 자유도를 나타냄

기대감 등으로 다시 상승함을 알 수 있다. 건축 연면적(barea)의 경우에도 약 400㎡를 초과하면 주거용 건물로서의 효용증가가 거의 없는 것으로 보인다.

랜덤 포리스트의 경우 적합 결과를 표의 형태로 표현하기에 적합하지 않다. 수백 개의 회귀트리 결과를 종합하는 앙상블 접근을 취하므로 중간 계산과정을 명확하게 파악할 수 없기 때문이다. 그림 2는 수백 개의 회귀트리 중 하나의 적합 결과를 보여 주는데 토지면적(larea), 건물면적(barea), 경과연수(age), 인근지역 특징(land.neighbor) 등이 예측의 주요 변수로 사

용되었음을 알 수 있다. 예측치는 최소 8.377억원에서 53.86억원까지 이르고 있다.

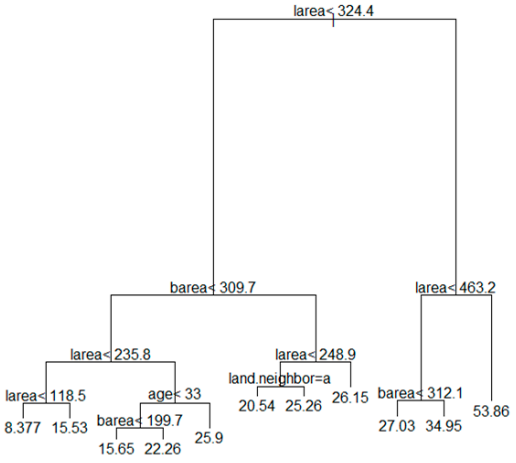

그림 2. 회귀트리 적합 결과(예시)

본 연구에서는 이와 같은 회귀트리 500개를 생성하였으나 그림 3을 보면 약 200개 회귀트리 이후부터는 Error 값이 일정 수준으로 수렴함을 알 수 있다. 그림에서 Error 값은 잔차제곱합의 평균을 나타낸다. 따라서 500개 트리에 기초한 본 연구의 추론에 문제가 없는 것으로 보인다.

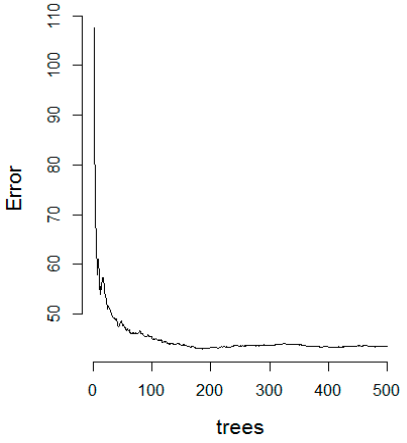

그림 3. 랜덤 포리스트 내 회귀트리 개수의 적정성

표 6은 MARS의 적합 결과를 보여준다. 강남구 단독주택의 경우 29년을 기준으로 경과연수가 주택가격에 미치는 효과가 상이하고, 365.79㎡를 기준으로 건물 연면적이 주택가격에 미치는 효과가 상이함을 알 수 있다.[8)]

표 6. MARS 적합 결과

설명변수	계수
상수항	4.0344
토지면적	0.0690
인근지역 특징[상업지대]	8.4797
토지면적×인근지역 특징[주상지대]	0.0274
h*(29−경과연수)×토지면적	0.0009
h(경과연수−29)×토지면적	0.0034
h(365.79−건물 연면적)×토지면적	−0.0001

* GCV: 32.57

마지막으로 SVM의 경우 랜덤 포리스트와 마찬가지로 적합 결과를 표의 형태로 표현하기가 수월하지 않다. SVM은 중간 과정을 쉽게 이해할 수 없는 일종의 블랙박스 모형으로, 산출된 예측 결과의 정확성에만 의미를 두는 특징이 있다. 표 7은 본 연구에서 적합시킨 SVM 모형의 상세 내역을 보여준다.[9)]

표 7. SVM 상세 내역

SVM Type	ε-regression
Kernel Function	Radial Basis
Cost Parameter	2.35
# of Support Vectors	251

연속형 종속변수를 다루는 경우 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 and Lin (2001)을 참조할 수 있다.

이와 같이 비모수 모형은 적합 결과를 비교하는 것보다는 예측 결과의 정확성을 비교하는 접근이 더욱 의미가 있다. 본 연구의 경우 예측 결과의 정확성은 검증 데이터의 실제 거래가격과 추정가격을 비교하여 판단하였다. 즉 실거래가 신고자료 438건을 7:3으로 임의분할(random split)하여 70%(306건)는 모형

적합에, 나머지 30%(132건)는 검증 데이터로 유보하여 예측 결과의 정확성을 비교하는데 사용하였다.[10] 정확성 비교 지표는 아래 식과 같은 RMSE(Root Mean Squared Error)와 MAE(Mean Absolute Error)를 사용하였다.

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-y_i)^2} \quad (12)$$

$$MAE=\frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i-y_i| \quad (13)$$

표 8 및 그림 4는 30%의 검증 데이터에 대해 분석한 각 모형의 성능을 보여준다.

표 8. 모형별 RMSE 및 MAE

구분	OLS	GAM	RF	MARS	SVM
RMSE	5.83	5.58	5.50	5.22	5.41
MAE	4.52	4.43	4.15	4.12	4.19

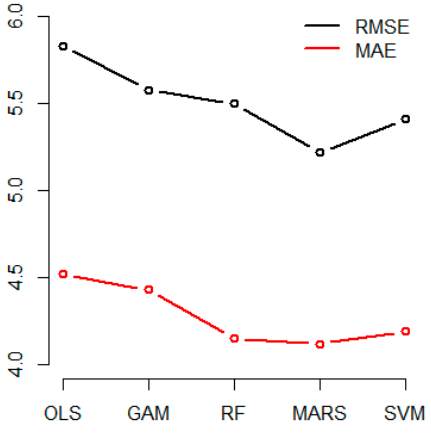

그림 4. 모형 성능의 비교

모수 모형 OLS 대비 비모수 모형의 가격 예측 성능이 전반적으로 우수한 것으로 나타났으며, 특히 최근에 제시된 MARS의 성능이 비교적 뛰어난 것으로 분석되었다.

3) 공간적 자기상관성의 반영

상기 모형 성능은 공간적 자기상관성을 반영하지 않은 상태에서의 비교 결과이다. 그러나 자료에 공간적 자기상관성이 존재하는 경우 그러한 특징을 반영하는 것이 합리적일 것이다. 그림 5는 모형 성능이 가장 우수하게 산출된 MARS의 주택 예측가격에 대해 비율값을 표시한 것이다. 비율값이란 MARS에서 산출된 예측가격 $\hat{y}$을 실제 주택 거래가격 y으로 나눈 것($\hat{y}/y$)으로 이 값이 1.00보다 크면 과대추정 경향, 1.00보다 작으면 과소추정 경향을 나타낸다. 이러한 비율값은 모형 적합에 사용하지 않은 30%의 검증 데이터를 대상으로 계산하였다.

공간적 자기상관성이 존재하지 않는다면 비율값이 지역 전체에 걸쳐 무작위하게 분포하여야 하나, 그림 5를 보면 일관되게 과대 및 과소평가하는 지역이 있음을 알 수 있다. 공간적 자기상관성 여부를 판단할 때 흔히 활용되는 검정 통계량 Moran's I값은 0.11(p-value 0.02)로서 비교적 유의하게 나타나 비율값에 공간적 자기상관성이 존재함을 추론할 수 있다.[11]

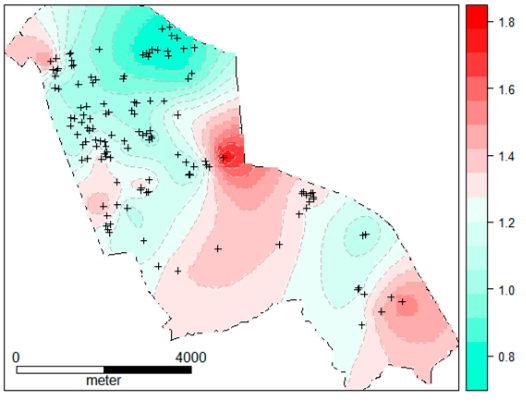

그림 5. 비율값 연속표면
+ : 검증 데이터에 해당하는 단독주택

W는 여러 가지 방법으로 구성할 수 있으나, 본 연구에서는 $1/d^2$를 가중치로 한 W를 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.[12] 공간차 변수 WY에서 Y는 거래금액(단위: 억원)을 토지면적으로 나눈 단가를 사용하였다.

표 9는 공간적 자기상관성을 모형의 구성요소로 추

가 반영한 결과를 보여준다. OLS를 포함한 모든 모형에서 모형 성능이 개선되었음을 알 수 있다.

표 9. 모형 성능의 개선 정도(검증 데이터 기준)

구분	RMSE	MAE
OLS	5.83→5.11 (▼)	4.52→3.82 (▼)
GAM	5.58→4.72 (▼)	4.43→3.57 (▼)
Random Forest	5.50→4.92 (▼)	4.15→3.58 (▼)
MARS	5.22→4.87 (▼)	4.12→3.72 (▼)
SVM	5.41→4.95 (▼)	4.19→3.60 (▼)

또한 Moran's I 값의 경우 최초 0.11에서 0.08로 떨어졌고 p-value도 0.02에서 0.08로 증가하였다. 따라서 미약하지만 공간적 자기상관성을 0.11에서 0.08로 다소 줄일 수 있었다.

표 9는 자료의 7:3 임의분할(1회)에 따른 모형 성능 비교치를 보여주는 것으로 동일한 비중의 임의분할을 100회 시행한 결과는 그림 6과 같다. 표 9의 결과와 유사하게 OLS보다는 여타의 비모수 모형이 우수하게 나온 것을 확인할 수 있다.

요약하면 모수 모형인 OLS 대비 예측능력 제고에 초점을 맞춘 비모수 모형의 가격 정확성이 보다 높았고, 이러한 비모수 모형에 시계열 분석에서의 시차변수(Lagged Variable)와 유사한 공간차 변수를 활용함으로써 가격 정확성을 추가적으로 개선시킬 수 있었다. 아울러 비율값에 남아 있는 공간적 자기상관성도 완화시킬 수 있었다.

4. 결론

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 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.

서울시 강남구를 사례지역으로 하여 기계학습 분야에서 제시된 다양한 비모수 모형들을 주택가격 추정에 적용한 결과, SVM이나 MARS 등 최근에 개발된 모형들의 성능이 비교적 우수한 것으로 나타나 이러한 모형들의 확대 적용이 필요한 것으로 보인다.

한편 기계학습 분야에서 제시된 이러한 비모수 모형들은 기본적으로 속성정보만을 고려할 뿐, 공간사상의 가장 큰 특징인 공간적 자기상관성을 반영하는 데 관심이 적다. 본 연구에서는 비모수 모형에 공간적 자기상관성을 추가로 반영하기 위해 주변 주택가격의 평균적인 가격수준을 나타내는 공간차 변수 WY

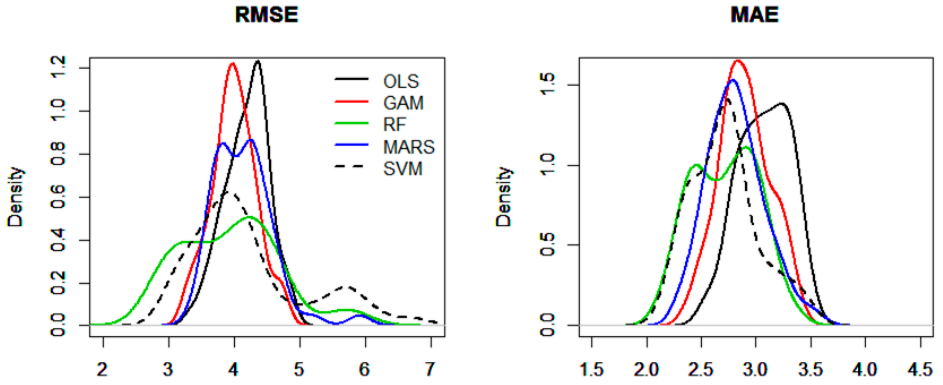

그림 6. 모형 성능의 개선 정도(100회 simulation 결과)

를 동원하여 공간적 자기상관성을 모형의 한 요소로 반영하였다. 그 결과 모형 성능이 개선됨을 확인할 수 있었다.

강남구는 건물가격 비중에 비해 토지가격 비중이 높은 지역으로 본 연구 결과는 이와 비슷한 토지:건물 비중을 보이는 대도시 지역에 무리 없이 적용 가능할 것으로 보이나 토지가격보다 건물가격 비중이 높은 군 지역에서는 설명변수의 선정 등에 있어 본 연구와는 다른 접근이 필요할 것으로 보인다.

모수 모형이 갖는 한계점을 설명하고 비모수 모형이 갖는 이점을 중심으로 살핀 본 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. 따라서 OLS 모형과 같은 모수 모형의 정교화에 상대적으로 노력을 덜 경주한 한계가 존재한다.

최근 공공자료의 개방과 이에 따른 구득성 증가로 데이터에 기반한 주택가격 추정이 어느 때보다 수월해졌다. 전문가에 의한 주택가격 추정을 정밀평가(a single-property appraisal)라 한다면 데이터에 기초하여 많은 수의 주택을 일시에 추정하는 것을 대량평가(mass appraisal)라 한다. 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.

따라서 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 기계학습 모형은 현행 공시가격과 달리 사람에 의한 편의가 적고 실제 거래가격에 근접한 가격을 제공할 수 있는 바, 현재 정부가 추진하는 실거래가 기반 공시제도 도입 업무 등에 의미 있는 시사점을 제공할 수 있을 것으로 예상된다.

그간 모수 모형에 집중되었던 방법론이 비모수 모형으로 확대되는 등, 본 연구를 계기로 부동산 가격추정에 있어 기계학습 모형의 논리와 응용에 대한 관심이 제고되기를 기대한다.

주

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 관찰치를 추측하는 의미로 사용하기로 한다.
2) 경과연수는 주택의 물리적 노후화 정도를, 코호트는 주택 신축 당시의 건축공법이나 스타일을 나타낸다.
3) t는 상수이며 'knot'라고 부른다.
4) 자세한 사항은 James *et al.*(2013) Chapter 9, Hastie *et al.*(2009) Chapter 12 등 참조.
5) 국토교통부 실거래가 공개시스템(www.molit.go.kr) 및 건축데이터 민간개방 시스템(http://open.eais.go.kr)에서 자료를 확보하였다.
6) 이하 모든 분석은 통계 소프트웨어 R을 사용하였으며 활용한 패키지는 mgcv, randomForest, earth, e1071 등이다.
7) 단독주택을 대상으로 한 담보평가, 경매평가 및 공시가격 산정시 기준으로 삼는 변수들을 의미한다.
8) 경과연수 29년을 전후한 효과의 차이는 재개발·재건축에 따른 가격상승 기대감에 기인한 것으로 보이며, 단독주택의 건물 규모가 365.79㎡를 초과할 경우 초과분에 대해서는 주거용 건물로서의 추가적 효용 증가분이 거의 없다고 해석할 수 있다. 29년 또는 365.79㎡와 같은 knot point는 잔차제곱합(RSS)이 가장 크게 감소하는 모형을 찾는 과정에서 얻어진다. 다시 말해 MARS의 적합은 종속변수 평균값으로부터 출발하여, 경첩함수를 추가할 새로운 설명변수의 선정 및 선정된 설명변수의 knot point를 찾는 과정의 연속이라 할 수 있는데, 본 연구의 경우 경과연수는 29년, 건물면적은 365.79㎡에서 knot point를 설정하여 해당 값을 경계로 각각 다른 함수(piecewise functions)를 적용하는 것이 RSS를 가장 크게 줄일 수 있는 것으로 산출되었다.
9) 자세한 사항은 James *et al.*(2013) Chapter 9, Hastie *et al.*(2009) Chapter 12 등 참조.
10) 검증 데이터 비중이 높을수록 모형의 예측 오류를 과대평가하는 반면, 검증 데이터 비중이 낮을수록 모형의 예측 오

류를 과소평가할 가능성이 있다(Fortmann-Roe, 2015). 따라서 검증 데이터셋을 어떠한 비중으로 나눌 것인지는 이러한 상쇄관계를 고려해야 한다. 아래 표는 논문에서 활용한 설명변수를 동원하여 OLS 모형을 적용한 후, 검증 데이터에 대해 결정계수 값(Cross-validated R^2, 즉 MSE)을 계산한 결과이다(100회 랜덤 분할 시뮬레이션 결과).

① 5:5분할(100회 simulation)	② 7:3분할(100회 simulation)
Cross-validated R^2 49.5%	Cross-validated R^2 56.0%
③ 8:2분할(100회 simulation)	④ 9:1분할(100회 simulation)
Cross-validated R^2 55.8%	Cross-validated R^2 56.9%

검증 데이터 비중을 높여 50%로 분할한 경우 Cross-validated R^2이 49.5%로 산출되는 등 모형의 예측 오류를 상대적으로 과대평가하고 있다. 반면, 검증 데이터 비중을 30%, 20%, 10%로 낮출 경우 Cross-validated R^2은 56% 내외로서 큰 차이가 없음을 알 수 있다. 따라서 검증 데이터 비중을 5:5 정도로 크게 높이지 않는 한 7:3, 8:2, 9:1 분할은 실질적인 차이가 없는 것으로 보인다. 본 연구에서는 부동산 분야 논문에서 주로 쓰이는 비율(7:3)을 따랐다.

11) 여기에서 Moran's I 값은 수치 그 자체의 의미보다는 WY를 활용하여 예측모형의 오차를 줄이고 공간적 자기상관성을 완화시켰다는데 의미가 있다.

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

참고문헌

김종수 · 이성근, 2012, "헤도닉가격모형과 서포트 벡터 회귀분석모형을 이용한 공업용 부동산의 가격추정," 감정평가학 논집, 11(1), 71-89.

안지아 · 박헌수, 2005, "공간종속성을 이용한 아파트 가격의 공간효과에 관한 연구," 대한국토도시계획학회 정기학술대회, 957-965.

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

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

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

Bao, H. X. and 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.

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

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

De Andrés, J., Lorca, P., de Cos Juez, F. J. and 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.

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

Fortmann-Roe, S., 2015, Consistent and Clear Reporting of Results from Diverse Modeling Techniques: The A3 Method, *Journal of Statistical Software*, 66(1), 1-23.

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

Gloudemans, R. and Almy, R., 2011, *Fundamentals of Mass Appraisal*, IAAO, Kansas City.

Guo, L., Ma, Z. and 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. and Vapnik, V, 2002, Gene selection for cancer classification using support vector machines, *Machine learning,* 46, 389-422.

Hastie, T., Friedman, J. and Tibshirani, R., 2009, *The elements of statistical learning*, Springer, New York.

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

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

University of Toyama.

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

Lasota, T., Łuczak, T. and 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), Springer, Berlin and Heidelberg.

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

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

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

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

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

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

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

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

교신: 박기호, 151-742, 서울시 관악구 관악로 599, 서울대학교 지리학과(이메일: khp@snu.ac.kr)

Correspondence: Key Ho Park, Department of Geography, Seoul National University, 599 Gwanangno, Gwanak-gu, Seoul 151-742, Korea (e-mail: khp@snu.ac.kr)

최초투고일 2016. 1. 28
수정일 2016. 3. 15
최종접수일 2016. 4. 18