pISSN : 2465-9754 eISSN : 2508-1292 https://doi.org/10.30902/jrea.2022.8.3.1

Journal of Real Estate Analysis

http://www.ejrea.org *November 2022, Vol.8, No.3, pp. 1~29*

기계학습 방법론을 활용한 아파트 매매가격지수 연구*

A Study on Apartment Sales Price Index Using Machine Learning Methodology

김이환** · 김형준*** · 류두진**** · 조훈*****

Yihwan Kim · Hyeongjun Kim · Doojin Ryu · Hoon Cho

Abstract

This study proposes to calculate new housing price indices through machine learning techniques. Our research is conducted focusing on the random forest and artificial neural network methodologies that proved excellence in existing real estate studies. We use micro-level real estate transaction data and housing characteristics information to train our models. As a result, the forecasting powers of the machine learning based models are found to be much superior in terms of explanatory powers and estimation performances compared to the hedonic methodology-based model. And the random forest model shows the best explanatory power and performance. Our results show that the housing price indices based on the machine learning models have greater volatility than currently used indices at the time of the housing price increase. Considering the limitation that the existing indices have a smoothing problem, our results can be interpreted that the new machine learning based indices reflect the market trend successfully.

Keywords: Artificial neural network, Hedonic, Housing price index, Machine learning, Random forest

* 본 논문은 김이환의 학위논문 내용을 확장·보완한 연구임. 이 논문 또는 저서는 2021년 대한민국 교육부와 한국연구재단의 일반공동연구지원사업의 지원을 받아 수행된 연구임(NRF-2021S1A5A2A03063960).

** KB자산운용 ETF 솔루션 운용본부 사원(주저자) | Associate, ETF Solution Management Division, KB Asset Management | First Author | kimleehwan3@gmail.com |

*** 영남대학교 경영학과 조교수(교신저자) | Assistant Professor, Department of Business Administration, Yeungnam University | Corresponding Author | hkim@yu.ac.kr |

**** 성균관대학교 경제학과 교수(교신저자) | Professor, Department of Economics, Sungkyunkwan University | Corresponding Author | sharpjin@skku.edu |

***** 한국과학기술원 경영대학 부교수 | Associate Professor, College of Business, Korea Advanced Institute of Science and Technology | hooncho@kaist.ac.kr |

Ⅰ. 서론

1. 연구의 배경 및 목적

지난 2022년 5월, 제20대 대통령직인수위원회에서 발표한 '윤석열 정부 110대 국정과제'에 따르면, 정부는 부동산시장 안정화를 위하여 주택공급을 확대하고 관련 세제와 대출 규제를 완화할 것으로 전망된다. 부동산 정책은 국민 대다수의 삶에 직접적인 영향을 미치는 대표적인 분야로, 신한은행에서 발표한 '2021 보통사람 금융생활 보고서'에 따르면 지난 2020년 우리나라 가계의 전체 자산 규모 중 부동산이 차지하는 비중은 70%가 넘는다. 〈표 1〉은 2018년부터 3년간 국내 가계의 자산 비중을 보여주는데, 금융자산과 부동산의 격차는 점점 커지고 있다. 이렇듯 가계 자산 중 부동산이 차지하는 비중이 절대적이기에 일반 국민은 부동산 정책변화와 가격변동에 민감하게 반응한다(김형준 외, 2018a, 2018b).

주택은 자산으로서의 가치뿐만 아니라 거주 공간으로서 삶의 질과 직접적인 연관을 가진 재화이다. 따라서 정부는 주택시장의 지나친 변동성을 통제하기 위해 시장 안정화 또는 활성화를 목적으로 다양한 주택 관련 정책들을 발표한다. 지난 2017년 출범한 문재인 정부의 경우 여러 차례에 걸쳐 부동산 정책을 발표하며 주택시장 변동성을 줄이고 부동산 가격의 안정화를 위한 여러 노력과 시도를 하였다. 그러나 정책의 실효성에 대해서는 많은 의문이 제기되는데, 특히 지난 2020년 8월 한국경제학회에서 실시한 설문조사에 따르면 설문에 참여한 전문가 가운데 약 3/4이 정부 정책이 수도권 주택가격 폭등 현상의 주요 원인이라고 지적했다.

〈표 1〉 가계 평균 보유 자산

비중(%)	2018년	2019년	2020년
부동산	75.9	76.0	78.0
금융자산	16.8	16.5	14.7
기타 자산	7.3	7.5	7.3

자료: 신한은행(2021).

이러한 가운데, 실제 주택시장의 변동성을 확인하거나 정책의 방향성을 설정하기 위해서는 주택 실거래가의 변화를 정확하게 반영하는 주택가격지수의 개발이 요구된다. 현재 국내에서 발표되는 주택지수들은 그 값과 변동성이 실제 시장에서 체감하는 변화와 상이하여 통계자료로서의 신뢰성이 도전받고 있다. 일례로 지난 2021년에 경제정의실천시민연합(경실련)은 2017년 5월 이후 4년간 서울시 주택의 평당 가격이 93% 상승했다고 발표했지만, 정부 측에서는 동기간 내 서울시의 주택가격 상승률이 17%에 그쳤다고 밝혔다(최한종, 2021). 이러한 논란이 있었던 배경에는 상승률 계산에 중위 · 평균 매매가격을 사용함으로써 노후 아파트 영향이 제한되고 시장을 과잉해석하는데 그 원인이 있다.[1] 하지만 주택시장에서 소비자가 체감하는 가격 변화와 실제 주택가격지수의 변화율 간의 괴리가 커지면서 이를 보완할 방법론이 요구되고 있다. 정책을 진단하는 과정

1) 국토교통부 보도자료, "[설명] 정부는 실수요자의 내 집 마련 기회를 확대하기 위해 지속적으로 노력하고 있습니다(2020. 08. 04.)" 참고.

 https://doi.org/10.30902/jrea.2022.8.3.1

에서 소비자 체감에 가까운 지표를 활용하면 미온적 대응에서 벗어나 실효성 높은 정책이 검토될 것으로 기대할 수 있다.

이처럼 같은 기간 동안 서울의 주택가격 변화를 놓고 상반된 결과가 도출된 이유는 부동산이라는 자산이 가지는 특징으로 인하여 발생하는 현상이다. 부동산의 특징 중 부동성과 개별성은 주택 간의 이질성을 초래하여 표준화된 지수 산출을 어렵게 만든다. 또한, 실제 주택가격을 정확하게 반영하는 주택가격지수를 만들기 위해서는 실거래가 기반의 지수 산출이 이뤄져야 하는데, 주택의 경우 상대적으로 거래가 빈번하지 않은 재화이기에 실거래가 기반의 표본 산정이 어렵다. 마찬가지로 실제 주택거래가 주택지수에 반영되기까지 약 30일 내외의 시간이 소요된다는 점 또한 정확한 주택지수 산출에 장애물로 작용한다.[2)]

그럼에도 불구하고, 우리나라는 주택가격에 기초한 부동산지수를 만들기에 상대적으로 용이한 환경을 갖추고 있다. 우리나라는 2019년 기준 전체주택의 77.2%를 공동주택이 차지하고 있으며, 이 가운데 80.6%가 아파트로 구성되어 있어 소위 '아파트 공화국'이란 이름으로 불린다.[3)] 아파트는 다른 주택 유형보다 동질적인 특성을 공유하고 있으므로 주택 중 아파트의 비중이 높은 우리나라에서 아파트 매매가격을 활용하여 지수를 만든다면, 비교적 정확하고 신뢰할 수 있는 아파트 매매가격지수를 도출할 수 있다.

실거래가 기반의 가격지수를 만들기 어렵다는 한계는 가격추정을 통해 극복할 수 있다. 주택가격은 감정평가를 통해 추정하는데, 부동산 감정평가 방법으로는 비용접근법(cost approach), 수익접근법(income approach), 시장접근법(market approach), 대량평가모형(mass appraisal model) 등이 사용된다. 감정평가사가 실제 조사 후 비용접근법, 수익접근법, 시장접근법 등을 활용하여 가치를 산정하는 정밀 산정방식은 주택가격을 비교적 정확하게 추정하지만 큰 비용과 인력이 요구된다. 이에 비해 대량평가모형은 통계적 모형을 활용한 가격산정방식으로, 개별공시지가 또는 개별주택가격 산정에 활용되는 간이 산정방식이다.[4)] 최근 대량평가모형은 기계학습 방법론을 적용하여 효율적이면서도 정확도 높은 추정을 하는 방향으로 발전하고 있다(홍정의, 2021).

본 연구에서는 아파트 매매가격 추정의 성과를 높이는 방법으로써 기계학습(machine learning) 방법론을 적용했다. 기계학습은 주어진 표본을 바탕으로 스스로 훈련하여 모형을 구축하므로 입력변수의 변동성이 크거나 표본의 수가 적은 경우에도 반복 학습을 통해 오차를 줄여 매우 신뢰할 수 있는 추정성과를 유도한다(James et al., 2013). 본 논문에서는 여러 기계학습 방법론들 가운데 기존 선행연구에서 주택가격 추정에 높은 성과를 보인 인공신경망(artificial neural network)과 랜

2) 법제처 국가법령정보센터에 따르면, 지난 2019년 8월 20일 '부동산 거래신고 등에 관한 법률' 일부 개정으로 부동산 거래신고 기한이 과거 60일 이내에서 30일 이내로 단축되었으며, 2020년 2월 21부터 시행되고 있다.

3) 통계청에서 발표한 2019년 인구주택총조사를 기준으로 한다.

4) 본 내용을 지적해주신 익명의 심사위원께 감사드립니다.

덤포레스트(random forest) 방법론을 사용했다. 나아가 두 방법론을 통해 추정한 아파트 매매가격을 바탕으로 지수를 만들어 현재 통용되고 있는 아파트 매매가격 지수와 비교했다. 또한, 본 연구에서 사용한 것과 동일한 표본을 가지고 아파트의 매매가격을 추정하고, 이를 바탕으로 지수를 산출할 수 있는 벤치마크로서 헤도닉(hedonic) 방법론을 함께 활용한다.

본 연구는 부동산 시장 분석에서 기계학습 방법론을 활용하는 구체적 방안을 제시한다. 본 연구는 기계학습 방법론이 일반적으로 보이는 높은 설명력을 토대로 추정 주택가격을 계산하고, 이를 활용해 새로운 부동산지수 작성법을 제안한다. 거래되지 않은 주택의 적정 가격을 통계적 모형을 통해 추정하고 이에 기반을 둔 주택가격지수를 만든다면, 정책 수립과정에서 참고할 수 있을 뿐만 아니라 다양한 영역에서 활용할 수 있다. 주택을 담보로 대출을 시행해주는 금융기관들은 추정된 주택가격을 기준으로 대출을 시행해줄 수 있게 되고, 차주 또한 보유 주택에 상응하는 금액의 대출가액을 받을 수 있어 효율적인 자금 운용이 가능하다. 그리고 거래되지 않은 주택들에 대해서도 적정가에 대한 추정치를 제공함으로써 투기성 매매나 양도세 회피 목적으로 지나치게 낮은 가격에 매매되는 주택들을 식별하고 이를 규제할 수 있다.

본문의 구성은 다음과 같다. 2장에서는 국내 아파트 매매가격지수에 대한 소개와 관련된 선행연구를 살펴보고, 기계학습 방법론 및 헤도닉 방법론을 적용하여 국내 부동산시장을 분석한 연구를 조사한다. 그리고 기존의 선행연구들과 달리 본 논문이 가지는 차별점에 관해 서술한다. 3장에서는 본 연구에서 사용한 자료의 수집과정을 소개하고, 분석에 사용된 설명변수와 기술통계량을 제공한다. 4장은 연구방법론 및 모형 성과분석으로, 각 방법론에 대한 설명과 주택가격 추정을 위한 모형들을 비교하고 그 과정에서 추정성과가 높았던 모형들의 분석 결과에 관해 서술한다. 5장에서는 추정성과가 우수한 모형을 이용한 아파트 매매가격지수를 산출하고, 이를 다른 지수들과 비교하여 그 차이에 대해 논의한다. 마지막으로 6장에서는 지금까지의 논의를 정리하고, 결론과 한계점 등을 제시한다.

II. 선행연구

1. 국내 아파트 매매가격지수

현재 우리나라에는 아파트의 매매가격에 기반을 둔 아파트 매매가격지수가 이미 존재한다. 오늘날 국내 부동산시장의 가격 변화를 추종하기 위해 사용되는 아파트 매매가격지수로는 대표적으로 KB 국민은행에서 발표하는 국민은행 가격지수(이하 'KB 지수')[5]와 한국부동산원의 아파트 실거래가 지수(이하 '실거래가 지수')가 있다. 두 가격지수는 서로 같은 대상을 나타내고 있음에도 불구하고 상당히 다른 움직임을 보이는데, 이는

5) KB 지수는 과거 한국 주택은행에서 발표하던 주택가격지수로, 2001년 4월 한국 주택은행과 국민은행이 통합한 이후 국민은행에서 발표하고 있다.

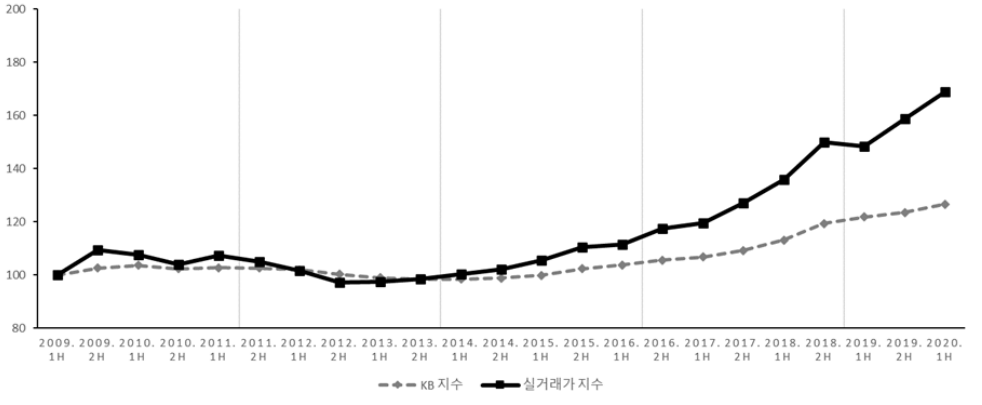
두 지수의 작성 과정에 사용되는 통계적 방법론이 다르기 때문이다. 〈그림 1〉은 KB 지수와 실거래가 지수가 서울시 아파트 가격 변화를 서로 다르게 설명하고 있음을 보여준다.

1) KB 지수

국민은행은 1986년부터 매주, 매월 주택가격지수를 발표한다. KB 지수는 종합주가지수 산정에 사용되는 라스파이레스(Laspeyres) 방식을 사용하는데, 이는 매 시점 거래가 이뤄지기 힘든 부동산 자산의 특성을 보완하는 지수산정방식이다(KB주택가격동향, 2022). 기본적으로 실거래가를 사용하되 실제 거래가 이루어지지 않은 주택에 대해서는 매매 사례 비교법을 이용한 감정가를 혼용하여 지수를 산출한다. 이때 감정가로는 '거래 가능 가격'을 사용하는데, 이는 지수 작성 과정에 참여하는 해당 지역 인근의 부동산 중개사들이 평가한 가격을 의미한다. 기간 내 많은 주택을 표본으로 설정하여 지수를 계산할 수 있다는 장점이 있지만, 실제 거래되지 않은 주택도 표본에 포함하기 때문에 여러 문제점을 가진다. 일례로 평가가격이 실제 매매가와 유사할 가능성도 있지만, 거래량이 많지 않으면 매각 희망자의 호가가 실거래가격으로 통용되고, 이것이 실제 거래가격으로 전이될 수 있다는 한계가 있다(노영훈, 2007). 또한, 조사자들의 성향이나 능력, 경험에 따라 발생할 수 있는 오류 또는 편의를 배제할 수 없으며, 평가가격에 기초하기 때문에 평활화(smoothing) 현상에 노출될 수 있음이 실증적으로 입증되었다(이용만 · 이상한, 2008).

2) 실거래가 지수

한국부동산원에서 공표하는 실거래가 지수는 반복 매매모형(repeated sales model)을 기반으로 주택재고량을 가중치로 한 제본스 지수 산식(Jevons index formula)으로 가격지수를 산출



주 : 본 그림은 KB 국민은행에서 발표하는 서울시 아파트 매매가격지수와 한국부동산원에서 발표하는 서울시 아파트 실거래가 지수를 2009.1H=100으로 재조정하여 비교한다.

〈그림 1〉 KB 지수와 실거래가 지수 비교

한다(한국감정원, 2017). 이는 기준이 되는 두 시점을 선정하고, 두 기간 모두에서 매매된 주택의 가격 변화를 이용하여 가격지수를 계산하는 방법이다(홍정의 외, 2022). 실제 실거래가 자료를 이용한다는 점과 두 기간 모두에서 거래된 주택을 대상으로 하므로 신뢰할 수 있고, 다른 특성이 주택가격에 미치는 영향을 고려하지 않아도 된다는 장점이 있다. 그러나 최소 두 번 이상 매매되는 주택만을 표본에 포함하고 있으므로 지수 산출에 필요한 표본의 숫자가 KB 지수 대비 현저히 적고 표본 선택과정에서 편의가 발생할 수 있다. 송영선 외(2020)는 반복 매매지수 모형을 사용하는 실거래가 지수의 경우 부족한 표본의 한계로 하부시장에서 지수 산출이 극히 불안해질 수 있다는 점을 지적했다. 일반적으로 거래가 적은 지역에서 반복 매매모형을 사용할 때 표본의 수가 급격하게 감소할 수 있기 때문인데, 서울시의 각 자치구에 대한 아파트 매매가격지수를 발표하는 KB 지수와 달리 실거래가 지수는 이런 한계로 서울시의 하위 지수로써 5개 권역별 지수만을 공표하고 있다(한국감정원, 2017). 특히 도심권의 경우 종로구, 중구, 용산구 3개 구를 포함하고 있음에도 거래량이 충분하지 못해 지수의 안정성을 보장받지 못한다는 지적을 받고 있다. 또한, 반복 매매모형의 경우 새롭게 발생한 거래 건이 과거 표본으로 활용되지 않았던 거래 건과 거래 쌍을 구성할 수 있기에 표본에 자료가 추가되는 과정에서 과거 지숫값이 변할 수 있다는 문제점도 내포하고 있다(이창무 외, 2007).

2. 관련 선행연구

1) 주택가격지수

부동산 시장의 움직임을 신속하고 정확하게 나타내는 주택가격지수에 대한 수요는 이전부터 꾸준히 존재했으며 관련 연구 또한 지속적으로 이루어졌다. 류강민 · 이상영(2010)은 부동산원에서 발표하는 실거래가 지수 산출에 사용되는 반복 매매비교법의 경우 가중치를 어떻게 부여하느냐에 따라 지수 간 차이가 유의미하게 나타남을 확인했다. 따라서 거주 안정을 목적으로 정책을 발표하는 정부 기관이나 주택을 담보로 대출을 진행하는 금융기관의 경우 동일 가중 방식을 활용하는 것이 적합하고, Case-Shiller home price 선물지수와 같이 개별 주택보다는 투자자산군의 하나로서 부동산 시장에 투자하는 경우 가치-가중 방식이 투자자 관점에서 더욱 적합할 것이라 주장했다. 이형욱 · 이호병(2009)은 주택가격지수를 예측할 때 인공신경망 모형이 ARIMA 모형보다 더 높은 예측력을 보여주는 것을 확인했다. 배성완 · 유정석(2018a)은 기존의 국민은행 가격지수는 평활화 문제에 노출되어 있고, 조사자의 다양한 특성에 따라 오류나 편의가 발생할 수 있음을 강조하며, 정확한 주택가격지수 산정을 위해서는 이 같은 오차를 제거하려는 방법이 필요하다고 주장하였다. 그리고 해결책으로 기계학습을 통한 공동주택 가격산정 방식을 제시했다.

2) 기계학습 방법론

최근 들어 기계학습 방법론을 부동산 연구에 적용한 사례가 증가하고 있다. Čeh et al.(2018)

은 슬로베니아의 수도인 류블랴나 지역의 주택시장을 대상으로 랜덤포레스트 모형의 가격 예측력이 헤도닉 방법론보다 우수하다는 것을 보였다. 국내에서는 배성완 · 유정석(2018c)이 2015년부터 2017년까지 서울 강남구 아파트를 대상으로 인공신경망 및 랜덤포레스트 모형을 사용한 가격지수를 산정하였다. 다만 연구에서 사용된 표본이 2015년부터 2017년까지의 강남구 아파트로 한정된다는 점에서 시간적, 공간적 한계가 있어 연구 성과를 일반화하기에 다소 어려움이 있다. 홍정의(2021)는 랜덤포레스트 모형을 통해 국내 시장의 주택가격추정을 진행하고, 추정성과를 기존의 헤도닉 모형과 비교하는 연구를 진행했다. 이 결과 하부 시장(sub market)에서의 비선형성과 입지 특성을 포착하는 데 랜덤포레스트 모형이 더욱더 효과적임을 입증했다. 박대현 외(2021)는 기존의 정형화된 로지스틱 방법론에 기계학습 방법론을 추가하여 주택시장의 조기경보체계를 구축해보았고, 미래 예측성과에서 로지스틱 분석보다 기계학습, 특히 랜덤포레스트 모형의 성과가 높았음을 실증적으로 확인했다. 또한, Hong et al.(2020)은 랜덤포레스트 모형을 사용한 주택가격 대량평가모형을 개발하였으며, Park and Ryu(2021)는 기계학습방법론을 사용하여 주택시장과 주식시장의 거품에 대한 조기경보시스템을 제안하였다.

3) 주택특성변수

아파트 매매가격에 영향을 미치는 주택특성변수들에 관하여 다양한 연구가 실시되었다. 먼저 배상영 외(2018)의 연구에서는 세대 수가 많을수록 관리비가 절감되고, 커뮤니티 시설 등의 인프라가 잘 구축되어 있어 거주 편의성이 높아져 주택가격에 유의미한 양의 영향을 준다는 것을 보였다. 한편 전용면적의 증가는 아파트의 면적당 가격에 부정적(negative)인 영향을 미쳤는데, 이는 최근 부동산시장에서 나타나는 소형평수에 대한 선호 현상으로 해석된다(금상수 외, 2014; 이옥자 · 최진배, 2015; 한다솜 · 최창규, 2018). 이번송 외(2002)와 하유정 · 이현석(2020)은 각각 난방방식과 아파트 복도구조를 가격변수로 설정하여 각 유형이 주택가격에 미치는 영향을 확인하였는데, 일반적으로 지역난방의 경우와 계단식 구조가 가격 프리미엄이 있는 것으로 관찰되었다. 그 밖에 강승우(2017)는 지하철까지의 거리를 교통편의 시설의 측정 지표로 추가하여 역세권에서의 가격 프리미엄을 확인했고, 윤병우 · 최경욱(2017)은 고등학교까지의 거리가 가까울 때 아파트의 면적당 가격에 긍정적인 효과를 줄 수 있음을 입증했다.

Ⅲ. 자료

1. 자료 수집 방법

본 연구는 서울시 25개 자치구의 아파트를 대상으로 진행한다. 해당 지역의 모집단에 가까운 자료를 선정하기 위하여 부동산114의 REPS 서비스에서 제공되는 서울시 법정동별 모든 아파트 자료를 사용하였다. 논문에 사용된 대부분의 아파트 특성 변수는 네이버 부동산 사이트(2022)에

서 크롤링 방식을 통해 개별 수집하였다. 네이버 부동산에서 확인할 수 있는 정보는 아파트별 세대수, 사용승인일, 세대당 주차대수, 전용면적, 방개수, 화장실 개수, 위도, 경도, 건설사, 공급면적, 용적률(FAR), 건폐율(BC), 난방방식, 복도구조를 포함한다. 다만 REPS는 건축물 대장상 아파트로 규정된 것만을 아파트로 정의하는 것에 비해, 네이버 부동산은 일부 도시형 생활주택을 아파트로 포함한다. 따라서 본 논문에서는 REPS의 정의를 따라 표본의 범위를 건축물 대장상 아파트로 한정하였다. 한편 네이버 부동산의 일부 아파트 정보의 경우 상기 언급한 변수 중 특정 변숫값을 제공하지 않는 경우가 있었는데, 이 경우 해당 아파트 데이터를 결측치로 분류 표본에서 제거하였다. 마찬가지로 난방방식을 혼용하고 있거나, 두 개 이상의 복도구조로 이루어진 아파트도 세대별로 어떤 변숫값을 가지는지 확인할 수 없으므로 표본에서 제외하였다.

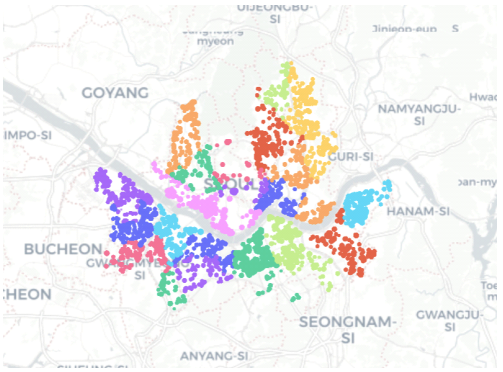
아파트매매실거래가 자료의 경우 공공데이터 포털에서 open API의 형태로 제공하는 국토교통부 아파트매매 실거래 상세 자료를 통해 취합하였으며, 2009년 1월부터 2020년 12월까지 매매된 실거래 자료를 사용하였다. 이후 수집된 아파트 정보를 면적 유형별로 분류하여 기존의 아파트별로 정렬되어 있던 정보들을 세대별로 재분류했다. 국토교통부에서 제공하는 실거래가 자료와 네이버 부동산에서 수집한 아파트 특성 변수를 세대별로 매칭시키기 위한 값(key value)으로는 '사용승인 시점+전용면적+법정동' 정보를 사용했다. 자료가 잘못 분류되는 것을 방지하기 위해 두 표본에서 함께 확인할 수 있는 아파트명 정보를 사용함으로써 오류를 최소화했다. 이후, 안정적인 지수 작성을 위하여 6개월 단위의 반기 자료로 재구성하였다.[6)]

이후 개별 아파트와 인접한 고등학교, 지하철, 공원까지의 거리를 측정하였다. 거리 측정에 사용한 학교의 좌표는 한국교원대학교 지방교육재정연구원에서 관리하는 전국 고등학교의 위치정보를 사용했고, 지하철 정보의 경우 서울교통공사에서 제공하는 노선별 지하철역 정보를 활용했다. 공원의 위치정보는 서울특별시 열린데이터 광장에서 제공하는 서울시의 공원 정보를 사용했다. 해당 정보를 사용해 거리를 측정하기 위해 개별 아파트의 좌푯값에서부터 개별 시설의 좌표까지 거리를 하버사인 공식(haversine formula)을 통해 도출하여 이 중 최솟값에 해당하는 거리정보를 각 고등학교, 지하철, 역까지의 거리로 입력했다.

〈그림 2〉는 상기 과정을 거쳐 산출한 자료의 좌표를 국내 지도 위에 표시한 값이며, 자치구별로 색을 달리하여 지도에 나타냈다. 이를 통해 본 연구에서 사용한 자료가 서울시 전역의 25개 자치구에 잘 분포되어 있음을 시각적으로 확인할 수 있다. 최종적으로 산출된 자료는 24개 반기 동안 서울시 25개 자치구에서 거래된 498,879개의 거래 정보로 세대별 개별특성과 거리변수, 시간 변수, 지역변수를 포함한다.

6) 더욱 실용적인 부동산지수의 설계를 위해서는 월간 혹은 주간 지수 작성이 요구된다. 본 연구에서도 월별 지수 계산을 시도하였으나, 도심권 등 표본 확충이 어려운 지역의 경우 기간 내 거래량이 충분하지 않은 경우가 존재하는 것으로 확인되었다. 오버샘플링(oversampling) 등을 활용하여 합성데이터를 생성함으로써 표본을 늘리는 방법도 있겠으나, 본 연구에서는 기계학습 방법론을 활용한 아파트 매매가격지수 작성법을 제안하는 것으로 논의의 범위를 제한하였다.

 https://doi.org/10.30902/jrea.2022.8.3.1



주 : 본 그림은 연구에서 사용된 아파트들의 좌표를 파이썬 plotly 패키지를 활용해 자치구별로 색을 달리하여 지도에 나타낸 것이다.

〈그림 2〉 분석에 사용된 주택의 분포

2. 변수소개 및 기술통계량 분석

본 연구에서 사용한 변수는 크게 세대별 특성에 기반을 둔 물리적 특성 변수, 거리로 표현되는 입지 특성 변수, 지역적 특성을 나타내는 자치구 더미 변수, 시간 특성을 나타내는 반기별 시간 더미 변수로 구분된다. 물리적 특성 변수들은 분석을 위해 선별한 변수로써 기존 선행연구들에서 아파트 가격에 대한 설명력이 높았던 전용면적, 세대수, 방 개수, 화장실 개수, 층, 난방방식, 구조방식, 도급순위 상위 건설사[7] 여부, 경과 연수, 용적률, 건폐율을 포함한다. 여기서 재건축 연한 도래에 따른 아파트 가격 상승효과를 확인하기 위해 경과 연수의 제곱 항을 추가했다. 주차 문제와 관련된 세대별 주차대수를 추가하였으며, 층수가 높아지면서 가격이 비선형적으로 영향받는 것을 확인하기 위하여 층 제곱항을 함께 사용했다. 또한, 1층 더미 변수를 추가하여 다른 층수에 비해 1층에 위치한 주택이 가격에 미치는 효과를 추가로 확인했다.

각 시설과의 거리를 통해 측정한 입지적 특성 변수에는 각 아파트로부터 가장 가까운 고등학교, 지하철, 공원까지의 거리변수를 사용했다. 또한, 지역적 특성과 시간 특성을 통제하기 위해 각 자치구에 해당하는 25개의 지역 더미와 2009년 상반기부터 2020년 하반기까지 24개의 시간 더미 변수를 설명변수로 추가했다. 또한, 두 더미 변수의 곱으로 이뤄진 상호작용항(interaction term)을 추가하여, 보다 세부적으로 각 변수의 영향력을 확인했다. 이 과정은 모형 분석과정에서 상세히 비교한다.

〈표 2〉는 각 더미 변숫값을 이용해 분석한 자료의 분포이다. 자료의 기간인 2009년부터 2020년까지 가장 많은 아파트가 거래된 자치구는 노원구로 총 53,992건의 거래가 있었으며 성북구, 강서구, 송파구, 양천구가 뒤를 이었다. 시점별로 가장 많은 아파트가 거래된 시점은 2015년 상반기로 33,899건의 거래가 있었으며 19년 하반기, 16년 하반기, 17년 상반기, 15년 하반기가 뒤를 이었다.

〈표 3〉은 상기 소개한 변수들에 대한 기술통계량 값을 포함하고 있다. 자료에서 연수의 평균값은 14.694년으로 이는 사용승인일부터 아파트매매 시점까지 평균적으로 14~15년 정도가 걸렸음을 의미한다. 세대수는 평균 1,076.976세대로 단지별 최소 6세대부터 최대 9,510세대로 다양하게 나타났다. 서울시의 표본에서는 평균적으로

7) 2021년 국토교통부 시공 능력 평가액 기준 상위 10개 사를 선정하였다. 대상은 삼성물산, 현대건설, GS건설, 포스코건설, 대우건설, 현대엔지니어링, 롯데건설, 디엘이앤씨(대림건설), HDC현대산업개발, SK에코플랜트이다.

〈표 2〉 분석에 사용된 자료의 분포

자치구	거래건수	비율(%)	시점	거래건수	비율(%)
용산구	7,891	1.58	2009. 1H	18,125	3.63
종로구	3,051	0.61	2009. 2H	17,875	3.58
중구	6,016	1.21	2010. 1H	10,659	2.14
강북구	11,681	2.34	2010. 2H	13,282	2.66
광진구	11,848	2.37	2011. 1H	16,461	3.30
노원구	53,992	10.82	2011. 2H	13,537	2.71
도봉구	13,107	2.63	2012. 1H	11,223	2.25
동대문구	15,936	3.19	2012. 2H	11,523	2.31
성동구	20,978	4.21	2013. 1H	17,757	3.56
성북구	32,088	6.43	2013. 2H	19,519	3.91
중랑구	16,638	3.34	2014. 1H	21,317	4.27
마포구	22,342	4.48	2014. 2H	24,394	4.89
서대문구	13,665	2.74	2015. 1H	33,899	6.80
은평구	16,158	3.24	2015. 2H	27,402	5.49
강서구	31,490	6.31	2016. 1H	27,310	5.47
관악구	18,148	3.64	2016. 2H	29,792	5.97
구로구	25,901	5.19	2017. 1H	28,288	5.67
금천구	7,906	1.58	2017. 2H	27,210	5.45
동작구	21,140	4.24	2018. 1H	24,951	5.00
양천구	26,535	5.32	2018. 2H	19,215	3.85
영등포구	19,572	3.92	2019. 1H	10,543	2.11
강남구	25,694	5.15	2019. 2H	29,928	6.00
강동구	26,296	5.27	2020. 1H	23,859	4.78
서초구	21,513	4.31	2020. 2H	20,810	4.17
송파구	29,293	5.87			

주 : 총 거래 건수: 498,879건.

1.133대의 차량을 세대별로 주차할 수 있는 것으로 나타났으며, 2대 이상의 차량을 보유한 가구가 많다는 점을 고려할 때 해당 수치는 아파트 주차 문제를 정량적으로 나타낸다고 볼 수 있다. 이외 물리적 특성 변수에 대한 평균값은 방 개수 2.984개, 화장실 개수 1.646개, 전용면적 79.885 등으로 나타났다. 다음으로 입지 특성 변수에 대한 기술통계량을 보면, 아파트부터 고등학교까지의 거

 https://doi.org/10.30902/jrea.2022.8.3.1

〈표 3〉 기술통계량 분석

변수	설명	평균	표준편차	최소	최대
Individual characteristics					
old	연수(년)	14.694	7.872	0.000	48.917
old_sq	연수 제곱	277.868	266.786	0.000	2,392.840
num	세대수	1,076.976	1,144.299	6.000	9,510.000
car	주차대수_총	1,182.204	1,437.841	6.000	12,456.000
car_per	주차대수_세대	1.133	0.453	0.120	6.530
area	전용면적(㎡)	79.885	26.945	23.340	268.580
room	방 개수	2.984	0.641	1.000	7.000
toilet	화장실 개수	1.646	0.493	1.000	5.000
floor	층	9.648	6.171	1.000	68.000
floor_sq	층 제곱	131.161	165.533	1.000	4,624.000
first	1층 여부	0.053	0.224	-	-
H1	난방방식(개별)	0.647	0.478	-	-
H2	난방방식(지역)	0.275	0.447	-	-
H3	난방방식(중앙)	0.077	0.267	-	-
T1	구조(계단식)	0.704	0.457	-	-
T2	구조(복도식)	0.270	0.444	-	-
T3	구조(복합식)	0.026	0.160	-	-
C1	건설사(도급순위 상위 10개)	0.352	0.478	-	-
FAR	용적률	283.036	122.351	12.000	1,249.000
BC	건폐율	23.058	9.722	2.000	93.000
Efficiency	전용률	76.870	4.794	54.000	100.000
Distance characteristics					
dist_high	고등학교까지 거리(km)	0.591	0.351	0.048	2.824
dist_sub	지하철까지 거리(km)	0.577	0.363	0.032	3.121
dist_park	공원까지 거리(km)	1.069	0.562	0.064	3.207
District characteristics					
GU1	용산구	0.016	0.125	-	-
GU2	종로구	0.006	0.078	-	-
GU3	중구	0.012	0.109	-	-
GU4	강북구	0.023	0.151	-	-

〈표 3〉 Continued

변수	설명	평균	표준편차	최소	최대
GU5	광진구	0.024	0.152	-	-
GU6	노원구	0.108	0.311	-	-
GU7	도봉구	0.026	0.160	-	-
GU8	동대문구	0.032	0.176	-	-
GU9	성동구	0.042	0.201	-	-
GU10	성북구	0.064	0.245	-	-
GU11	중랑구	0.033	0.180	-	-
GU12	마포구	0.045	0.207	-	-
GU13	서대문구	0.027	0.163	-	-
GU14	은평구	0.032	0.177	-	-
GU15	강서구	0.063	0.243	-	-
GU16	관악구	0.036	0.187	-	-
GU17	구로구	0.052	0.222	-	-
GU18	금천구	0.016	0.125	-	-
GU19	동작구	0.042	0.201	-	-
GU20	양천구	0.053	0.224	-	-
GU21	영등포구	0.039	0.194	-	-
GU22	강남구	0.052	0.221	-	-
GU23	강동구	0.053	0.223	-	-
GU24	서초구	0.043	0.203	-	-
GU25	송파구	0.059	0.235	-	-
Time dummy					
Half 1	2009년 상반기	0.036	0.187	-	-
Half 2	2009년 하반기	0.036	0.186	-	-
Half 3	2010년 상반기	0.021	0.145	-	-
Half 4	2010년 하반기	0.027	0.161	-	-
Half 5	2011년 상반기	0.033	0.179	-	-
Half 6	2011년 하반기	0.027	0.162	-	-
Half 7	2012년 상반기	0.022	0.148	-	-
Half 8	2012년 하반기	0.023	0.150	-	-
Half 9	2013년 상반기	0.036	0.185	-	-

 https://doi.org/10.30902/jrea.2022.8.3.1

〈표 3〉 Continued

변수	설명	평균	표준편차	최소	최대
Half 10	2013년 하반기	0.039	0.194	-	-
Half 11	2014년 상반기	0.043	0.202	-	-
Half 12	2014년 하반기	0.049	0.216	-	-
Half 13	2015년 상반기	0.068	0.252	-	-
Half 14	2015년 하반기	0.055	0.228	-	-
Half 15	2016년 상반기	0.055	0.227	-	-
Half 16	2016년 하반기	0.060	0.237	-	-
Half 17	2017년 상반기	0.057	0.231	-	-
Half 18	2017년 하반기	0.055	0.227	-	-
Half 19	2018년 상반기	0.050	0.218	-	-
Half 20	2018년 하반기	0.039	0.192	-	-
Half 21	2019년 상반기	0.021	0.144	-	-
Half 22	2019년 하반기	0.060	0.237	-	-
Half 23	2020년 상반기	0.048	0.213	-	-
Half 24	2020년 하반기	0.042	0.200	-	-

주 : 1) 용적률: 대지면적 대비 지하층을 제외한 지상층 면적합계.
2) 건폐율: 대지면적 대비 건축 면적.
3) 전용률: 공급면적 대비 전용면적.
4) 표본수: 498,879.

리는 평균적으로 0.591km이며, 지하철과 공원까지의 거리는 각각 0.577km, 1.069km인 것으로 나타났다.

Ⅳ. 연구방법론 및 모형 성과분석

1. 연구방법론 소개

1) 인공신경망

인공신경망 모형은 실제 두뇌의 작동 방법에서 착안한 기계학습 알고리즘으로 두뇌의 구조를 모방하여 단순한 구조의 인공 뉴런(neuron)을 연결함으로써 복잡한 문제를 해결하고자 하는 알고리즘이다(김형준 외, 2019). 즉 인공신경망 모형은 뉴런과 뉴런이 연결되는 신경계의 구조를 수학적 모형으로 표현한 것이다. Rosenblatt(1958)의 퍼셉트론(perceptron) 학습 규칙이라는 개념으로 등장한 이후 Rumelhart et al.(1986)이 역전파 알고리즘(backpropagation algorithm)을 소개하며 은닉층을 가진 다층퍼셉트론을 활용할 수 있는 방법을 제시함으로써 다양한 연구에

적용, 발전해왔다(정성훈 · 진창하, 2020).

본문에서 사용된 인공신경망 모형의 경우 1개의 입력층과 4개의 은닉층 그리고 1개의 출력층으로 구성되며, 각 층이 은닉 노드를 통해 연결되어 있기에 총 5개의 은닉 노드를 가진다. 이때 모형에서 이뤄지는 선형결합과 비선형결합 처리를 (식 1)과 같이 표현할 수 있다. 식에서 j는 층과 층을 연결하는 은닉 노드의 개수로 총 5개가 되며, 각 X는 설명변수로서 총 n개의 값을 갖는다. 입력된 변수 X가 연결 가중치 w와 곱해진 후 별도의 선형결합 없이 다음 노드로 값이 전달되고, 그 후 활성함수 f를 통해 값이 변환되어 이를 신호로써 다음 층으로 전달한다. 이런 일련의 과정이 반복되면 최종적인 출력값 Y를 얻을 수 있다(정성훈 · 진창하, 2020; Kim et al., 2021, 2022).

$$H_j = f(b_j + \sum_{i=1}^{n} w_{ij} X_i)$$

$$Y = f(b_0 + w_{10}H_1 + w_{20}H_2 + w_{30}H_3 + w_{40}H_4 + w_{50}H_5) \quad \text{(식 1)}$$

기계학습 방법론의 경우 과적합(over-fitting) 문제에 노출될 수 있으므로 본 연구에서는 이를 해결하기 위해 일부 은닉층의 단계에서 값을 누락(drop out)시켜 모형을 전개했고, 훈련 표본 중 일부에 대해 훈련 중 검정(validation)이 이뤄지게 하여 과적합의 문제를 해소하고자 하였다(Kim et al., 2020).

2) 랜덤포레스트(random forest, RF)

Breiman(2001)은 랜덤포레스트 모형을 앙상블 기법을 적용하여 다수의 의사결정트리(decision tree)를 합쳐놓은 것으로 설명했다. 해당 과정을 상세히 살펴보면 다음과 같다. 우선 의사결정트리 모형을 연속형 종속변수에 적용하는 모형을 회귀 트리(regression tree) 모형이라 한다. 회귀트리 모형은 설명변수 n개를 J개의 지역 $R_1, R_2, \ldots, R_j$ 에 서로 겹치지 않게 할당하고, m번째($m \leq j$) 지역에 속하는 R_m의 관측치에 대해 R_m지역 관측치의 평균값을 추정치로 제시하게 된다. 이때 R_j 지역은 (식 2)의 잔차제곱합(residual sum of squares)이 최소가 되도록 분할한다.

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \widehat{y_{R_j}})^2 \quad \text{(식 2)}$$

잔차제곱합을 최소화하는 기준으로 전체 트리를 구성한다면 언제나 과적합 문제에 노출된다. 이와 같은 문제를 해결하기 일반적으로는 가능한 가장 큰 크기의 의사결정트리에서 시작하여 가지를 줄여가는 방식으로 적정규모의 트리를 결정하게 되는데, 이를 수식으로 표현하면 (식 2)를 최소화하는 과정이라 볼 수 있다(배성완 · 유정석, 2018b).

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

(식 3)에서 $|T|$는 트리 T의 가지 수를 의미하고, R_m은 m 번째 가지에 해당하는 분할지역을 의미한다. α는 동조 파라미터(tunning parameter)로 페널티를 뜻하는데, $\alpha = 0$인 경우 아무런 페널티가 없으므로 최대 트리 규모가 된다. 반대로 α의 수준이 커질 때 페널티가 커짐에 따라 트리 규

모는 작아지게 되는데, 이때 최적 α값은 교차 타당성 검증을 통해 이뤄진다(이창로, 2015). 본문에서는 검증 자료에 대한 오차 분석을 시행하고, 이 값을 교차 검증하는 과정에서 최적의 트리 규모를 도출한다.

본문에서 사용된 랜덤포레스트 모형은 다수의 의사결정트리를 합쳐 학습과 가격추정을 진행한다. 의사결정트리는 속성값의 전처리가 복잡하지 않다는 장점이 있지만, 과적합의 문제에 노출되는 한계가 있는데, 랜덤포레스트 모형에서는 의사결정트리 여러 개를 구성함으로써 이를 해소한다. 랜덤포레스트 모형의 또 다른 특징은 최종 모형을 제시하는 과정에서 중요 관계에 있는 변수를 선별해낼 수 있다는 것이다. 따라서 본 연구에서는 랜덤포레스트 방법론을 활용한 모형을 통해 도출한 추정성과를 설명하며 모형에서의 변수 중요도 또한 함께 고려한다.

랜덤포레스트 구조의 또 다른 특징은 단계별 노드에 따른 표본 분류에 있다. 모형에서는 노드에 따라 표본을 여러 개로 분류하고, 최종 단계에서 이를 다시 평균화하여 값을 추정한다. 이는 기존의 회귀분석 모형이라면 더미 변수로 각각 분리하여 입력해야 했던 변수를 수열화(indexing)하여 단 한 개의 변수로 표현할 수 있음을 의미한다. 이 과정이 중요한 이유는 설명변수 중 더미 변수가 있으면 변수의 범주 수만큼 모형 내 변수 개수가 급증하여 연산 효율성뿐만 아니라 과적합 가능성이 늘어나 결과적으로 모형의 설명력 또한 낮출 수 있기 때문이다. 따라서 각각 시간 더미 변수 24개와 지역 더미 변수 25개를 시간 변수와 지역변수로 수열화하여 2개의 변수로 수정하여 가격추정에 사용했고, 이 과정을 통해 모형의 효율성을 증대시킬 수 있다.

3) 헤도닉 방법론(hedonic methodology)

헤도닉 모형은 변수의 형태에 따라 선형회귀모형과 준-로그(semi-log) 모형이 일반적으로 사용된다. 그리고 본 연구에서는 두 형태 중 준-로그 모형을 사용했다. 준-로그 모형의 경우 개별 설명변수의 변화로 인한 부동산 가격의 변화량을 파악할 수 있다는 것과 함께 회귀계수의 해석이 선형회귀모형보다 더 간단하다는 장점이 있다. 본문에서 사용한 헤도닉 회귀모형은 크게 두 가지로 아래와 같은 식을 따른다.

$$\ln price = c + \sum_{i=1}^{20} \alpha_i X_i + \sum_{i=1}^{3} \beta_i Dist_i + \sum_{i=1}^{25}\sum_{j=1}^{24} \gamma_{ij} District_i * Time_j + \epsilon \quad \text{(식 4)}$$

$$\ln price = c + \sum_{i=1}^{20} \alpha_i X_i + \sum_{i=1}^{3} \beta_i Dist_i + \sum_{i=1}^{25} \gamma_i District_i + \sum_{i=1}^{24} \delta_i Time_i + \epsilon \quad \text{(식 5)}$$

두 가지 식의 차이점은 다음과 같다. (식 4)에서는 상호작용항을 사용하여 총 600개의 설명변수를 추가했고, 이를 통해 회귀분석 과정에서 시점별로 각 자치구가 주택가격에 미치는 영향을 확인할 수 있다. 해당 모형을 통해 주택가격추정이 이뤄지면 자치구별 주택가격 지수를 직접적으로 확인할 수 있다는 장점이 있다. (식 5)에서는 고정 효과를 주기 위해 자치구 더미 변수와 시점 더미 변수를 각각 추가하여 회귀식을 구성했다. 김명준 외(2008)의 연구에서는 해당 회귀식에서 시점별

더미 변수의 계수 값을 이용해 주택가격지수를 도출한 바 있다. 본 연구에서는 이 방법론을 활용한 모형과 상호작용항을 통해 구현한 모형의 설명력과 추정성과를 비교하여 지수 산출을 위한 최적 모형을 선정한다.

2. 모형 성과분석

신뢰할 수 있는 주택가격 추정모형을 만들기 위해서는 표본에서 모형의 설명력도 중요하지만, 실제 가격 대비 모형 추정값의 오차가 작고 상관성이 높아야 한다. 나아가 실제 가격을 추정모형이 잘 설명할 수 있어야 한다. 따라서 본 연구에서는 기존의 전체 표본을 훈련 표본(train sample)과 실험 표본(test sample)으로 구분하여 실거래가에 대한 가격추정을 진행했고, 모형의 설명력을 측정하기 위한 지표로 결정계수(R-squared) 값을 사용했다. 그리고 추정 오차와의 상관관계 측정을 위해 RMSE(root mean square error)와 MAPE(mean absolute percentage error), Corr.(correlation) 지표를 사용하였으며 실제 값에 대한 모형의 설명력을 점검하기 위해 두 값 간의 결정계수를 비교했다. 또한, 각 지표는 면적당 주택가격의 로그값을 사용하여 계산되었으며, MAPE의 경우 다른 연구와의 비교를 위하여 면적당 주택가격을 기준으로 계산되었다.[8] 훈련 표본에서 모형의 설명력을 측정하기 위한 결정계수는 R-squared로 표현했으며, 실험 표본에서 사용된 결정계수 값은 앞선 R-squared와 구분하기 위해 Score로 표시했다. 전체 표본을 무작위 추출 방법(random sampling)을 사용하여 훈련과 실험 표본으로 분류했으며 비율은 각각 8대 2로 조정하였다.

1) 인공신경망 분석

인공신경망 모형을 구성하는 데 있어 상호작용항 여부에 따른 모형의 설명력과 추정성과를 비교한다. 특히 기계학습의 경우 훈련 횟수(epoch)를 몇 회로 설정하는지가 훈련 성과를 표현하는 데 매우 중요하기 때문에 같은 훈련 횟수를 상정하고 결과를 논의한다. 훈련에 따른 성과를 지속해서 확인해야 하므로 기존 훈련 표본의 일부를 검증 표본(validation sample)으로 분류하는 과정이 필요한데, 이때 검증 표본의 개수가 훈련 표본보다 너무 적다면 검증이 올바르게 이뤄지지 않을 수 있어 충분한 수의 표본을 확보해야 한다(Zur et al., 2009). 또한, 훈련 성과가 개선되지 않을 때 성급히 훈련을 종료한다면 이후 훈련 성과가 개선될 여지 자체를 차단하는 것이므로 20회의 구간을 설정하여 훈련 성과 개선이 구간 내에서 이뤄지지 않으면 훈련을 종료하는 것으로 지정했다. 〈표 4〉는 인공신경망 모형의 훈련 성과를 나타낸다. 이 중 with interaction은 상호작용항을 포함하는 모형이고, without interaction은 상호작용항을 포함하지 않고 더미변수의 고정 효과로 시간과 지역을 표현한 모형을 나타낸다. 앞서 언급했듯 과적합을 막기 위해 마지막 은닉층 단계에서 드롭아웃(dropout)을 적용함으로써 일부

8) 본 내용을 지적해주신 익명의 심사위원께 감사드립니다.

〈표 4〉 인공신경망 모형의 훈련 성과 비교

분류		인공신경망	
훈련횟수		500	500
변수		With interaction	Without interaction
설명력	R-squared	0.946	0.930
성과	RMSE	0.107	0.115
	MAPE	7.763	8.552
	Corr.	0.970	0.964
	Score	0.937	0.927

주 : RMSE, root mean square error; MAPE, mean absolute percentage error; Corr., correlation.

값을 누락시켜 최종 추정값을 도출했다.

훈련 결과, 상호작용항을 포함한 모형이 포함하지 않는 모형에 비해 더 많은 변수를 포함하고 있으므로 역시 더 높은 설명력을 보인다. 일반적으로 모형의 훈련 횟수가 증가하면 설명력의 개선과 함께 과적합 문제가 발생하기 마련인데, 500회까지의 훈련 결과로는 훈련 횟수가 증가할 때 추정성과가 지속해서 개선되는 결과를 확인할 수 있다.

2) 랜덤포레스트 분석

랜덤포레스트 모형을 통해 아파트 매매가격을 추정하는 과정에서는 앞서 설명한 것과 같이 일부 더미 변수를 수열화하여 표현한다. 홍정의(2021)는 랜덤포레스트 모형을 사용하면 입지적 측면에서 동질적이라 판단되는 가장 하위 그룹의 평균값으로 결과를 도출하기 때문에, 특정 모형에 구속되지 않고 그룹 간의 차이를 명확하게 드러낼 수 있다고 주장했다. 이는 결과적으로 랜덤포레스트 모형에서는 입지적 차이를 나타내기 위하여 거리 변수와 자치구 변수를 추가하는 대신, 위도와 경도 자료를 사용함으로써 이를 대체할 수 있음을 의미한다. 따라서 본 연구에서는 선행연구에서처럼 더 적은 변수를 통해 충분한 설명력과 추정성과를 보여줄 수 있는지 확인한다. 위도와 경도 자료만을 활용하여 입지 특성 변수와 자치구 변수를 대체할 수 있다면, 변수의 수를 대폭 줄임으로써 모형의 효율성을 개선할 수 있기 때문이다. 따라서 모형에서의 변수 구성은 위도와 경도를 설명변수로 포함하는 훈련 모형(1)과 위도와 경도에 자치구 변수를 추가하여 전개한 모형(2), 그리고 다른 방법론의 모형에서처럼 위도, 경도를 제외하고 모든 설명변수를 추가한 모형(3)으로 나누어 훈련을 진행했다. 각 표본에서의 결괏값은 〈표 5〉와 같다. 본 연구에서의 결과 또한 선행연구와 유사하게 위도, 경도만을 포함한 모형의 설명력과 성과가 다른 입지 특성 변수나 지역변수를 추가한 모형과 큰 차이가 없음을 확인했다.

〈표 5〉 변수 조정에 따른 모형의 훈련 성과 비교

분류		랜덤포레스트		
변수		모형 (1)	모형 (2)	모형 (3)
설명력	R-squared	0.997	0.996	0.996
성과	RMSE	0.066	0.067	0.067
	MAPE	4.432	4.557	4.526
	Corr.	0.988	0.702	0.697
	Score	0.976	0.976	0.976

주 : 1) 모형 (1): 위도, 경도 포함.
2) 모형 (2): 위도, 경도, 자치구 포함.
3) 모형 (3): 모든 설명변수 포함.
4) RMSE, root mean square error; MAPE, mean absolute percentage error; Corr., correlation.

3) 헤도닉 분석

헤도닉 모형에서 역시 상호작용항의 추가 여부에 따라 모형의 설명력과 추정성과를 비교했다. 블랙박스(black box)의 한계를 가지는 기계학습 과정과 달리 헤도닉 모형의 경우 각 설명변수가 종속변수인 아파트의 면적당 가격에 어떤 영향을 주는지 확인할 수 있다는 장점이 있다.

〈표 6〉은 헤도닉 모형을 통해 추정한 모형의 설명력과 성과를 비교한 결괏값이다. 인공신경망 모형에서와 유사하게 헤도닉 모형을 통한 분석에서도 상호작용항을 포함한 모형이 설명력과 성과 모두에서 우월한 모습을 보였으며, 이는 더미 변수로 고정 효과를 주어 자료를 분석했을 때보다 상호작용항을 통해 자료를 분석함으로써 평활(smoothing)의 문제를 개선함에 따라 나타나는 성과로 해석할 수 있다. 따라서 헤도닉 모형을 통한 아파트 매매가격 추정에서는 상호작용항을 포함한 모형을 최종 모형으로 선정하고, 이를 통해 아파트 매매가격지수를 도출한다.

〈표 6〉 헤도닉 모형의 훈련 성과 비교

분류		헤도닉	
변수		With interaction	Without interaction
설명력	R-squared	0.846	0.838
성과	RMSE	0.167	0.171
	MAPE	12.852	13.254
	Corr.	0.921	0.916
	Score	0.848	0.840

주 : RMSE, root mean square error; MAPE, mean absolute percentage error; Corr., correlation.

Ⅴ. 아파트 매매가격지수 추정 및 비교

1. 아파트 매매가격지수

이번 장에서는 각 모형을 통해 아파트 매매가격지수를 추정하고 그 활용 방안에 대하여 논의한다. 〈표 7〉은 아파트 매매가격지수를 산출하기 위해 최종적으로 선정한 아파트 매매가격 추정모형의 성과를 비교한 값이다.

3가지 방법론에 기초한 아파트 매매가격 추정모형은 기본적으로 2009년부터 2020년까지 서울시 25개 자치구의 498,879건의 실거래가 자료를 표본으로 한다. 따라서 이용 가능한 모든 아파트 실거래가 정보를 전체 표본(full sample)으로 활용해 아파트의 잠재적인 가격을 추정하고, 이를 바탕으로 서울시를 대표하는 아파트 매매가격지수를 도출한다.

기계학습 방법론을 통해 작성된 아파트 매매가격지수는 기존 지수가 담지 못한 시장 정보를 나타낸다. 반복매매법을 사용하는 실거래가 지수의 경우 지수 계산에 사용되는 표본이 반복 거래된 아파트만을 대상으로 하기 때문에 전체 아파트 시

〈표 7〉 아파트 매매가격지수 추정모형별 성과

분류	인공신경망	랜덤포레스트	헤도닉
RMSE	0.107	0.066	0.167
MAPE	7.763	4.432	12.852
Corr.	0.970	0.988	0.921
Score	0.937	0.976	0.848

주 : RMSE, root mean square error; MAPE, mean absolute percentage error; Corr., correlation.

 https://doi.org/10.30902/jrea.2022.8.3.1

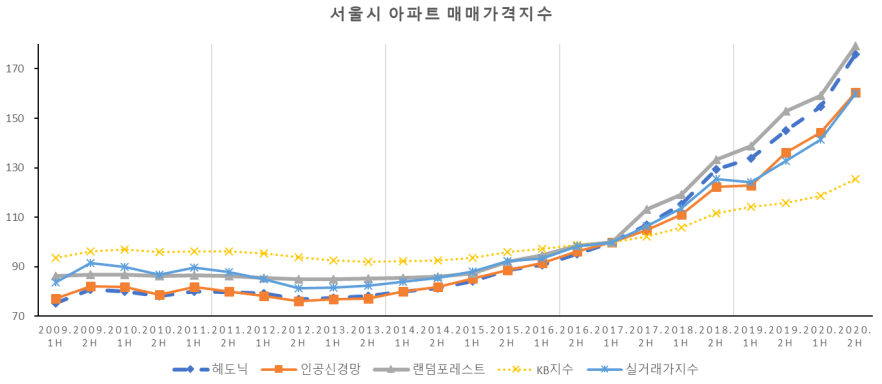
장에 대한 정보를 놓칠 수 있다. KB 부동산 지수의 경우 거래되지 않은 아파트들의 가격을 인근 부동산 중개업자들이 제공하는 거래 가능 가격을 통해 도출하므로 앞서 언급한 일반 호가가 실거래가로 통용되거나, 지수 평활화 문제에 크게 노출된다는 점에서 한계를 지닌다. 따라서 본 논문에서 제시하는 아파트 매매가격지수는 이용할 수 있는 모든 표본을 사용해 매매가격을 추정하여 이를 지수에 반영하므로 기존 지수의 한계를 보완하는 역할을 한다.

한편, 〈표 7〉을 보면 기존의 지수 산출 방법론인 헤도닉 방법론 모형과 비교해 기계학습을 적용한 인공신경망과 랜덤포레스트 방법론의 모형에서 주택가격 추정성과가 훨씬 우수함을 확인할 수 있다. 세부적으로 살펴보면 헤도닉 모형이 면적당주택가격의 로그값을 84.8% 수준에서 설명하는 반면, 인공신경망 모형은 93.7% 수준에서 설명한다. 가장 모형의 설명력이 높았던 랜덤포레스트 모형의 경우 97.6% 수준에서 실거래가를 설명하며, 상관계수 값 또한 0.988로 1에 매우 가까운 수치를 보여준다.

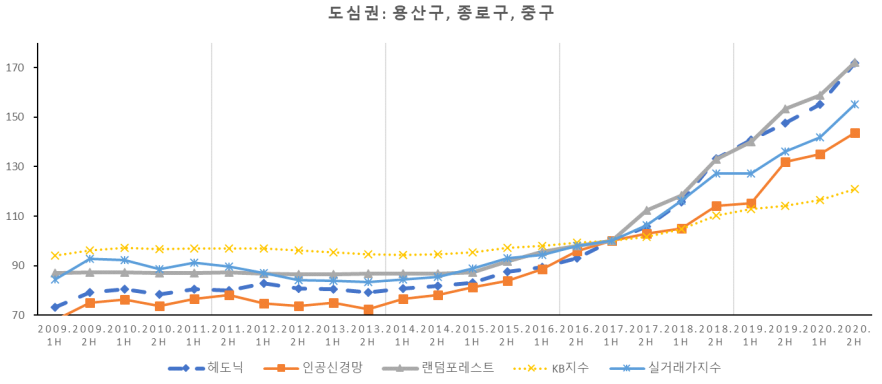
추적 오차 또한 기계학습 기반의 모형에서 훨씬 작은 오차를 보였다. 추적 오차 역시 헤도닉 모형에서 가장 컸으며, 인공신경망 모형과 비교해 랜덤포레스트 모형의 성과가 가장 우수하게 나타났다. 헤도닉 방법론과 반복 매매모형의 경우 주택가격을 설명할 때 주택을 이루는 여러 변수의 선형관계로써 표현한다. 그러나 주택가격은 일반적으로 주택을 구성하는 요인들의 비선형적 결합으로 이루어져 있어 변수 간의 선형관계를 가정한다면 큰 추적 오차가 발생한다. 반면, 기계학습에 기초한 방법론의 경우 주택가격과 변수 간의 비선형적 결합을 여러 차례의 훈련 과정을 통해 식별할 수 있고, 앞장에서 다룬 모형 간 교차검증을 통해 추정성과를 극대화하는 모형을 선별할 수 있다. 비선형적 결합으로 이루어진 주택가격을 기계학습의 메커니즘(mechanism)으로 설명한다면, 변수 간 선형관계로 설명하는 모형에 비해 우수한 추정성과를 보이게 된다. 그리고 해당 모형을 기반으로 주택가격지수를 도출한다면, 다른 변수들이 미치는 영향을 제거했을 때 순수한 시간적, 공간적 변화가 주택가격의 변동에 영향을 주는 정도를 더 잘 파악할 수 있다.

본 연구의 지수 산출 과정은 다음과 같다. 〈표 7〉에서 훈련한 아파트 매매가격 추정모형을 바탕으로 아파트 특성변수의 영향을 배제하고 지역별 시점별 가격을 추정한다. 이를 위해 아파트 특성변수에는 〈표 3〉에서 계산한 각 변수의 평균값을 대입한다. 이후 각 시점 및 지역 더미에 0과 1값을 대입하여 해당하는 시점 및 지역의 주택가격을 추정한다. 예를 들어, 도심권 지역의 2009년 상반기 주택가격을 추정하고자 하는 경우 도심권 지역 더미에 1을, 2009년 상반기 더미에 1을 입력하고 나머지 지역 및 시점 더미에 0을 대입한다. 이를 반복하여 각 지역 및 시점별 주택가격추정치를 계산한 후 기준시점을 정하고 지숫값을 계산한다. 본 연구에서는 지난 2017년 이후의 주택가격 급등기간에서 각 지수의 상승률을 비교하기 위하여 2017년 상반기의 지숫값을 100으로 설정하였다.

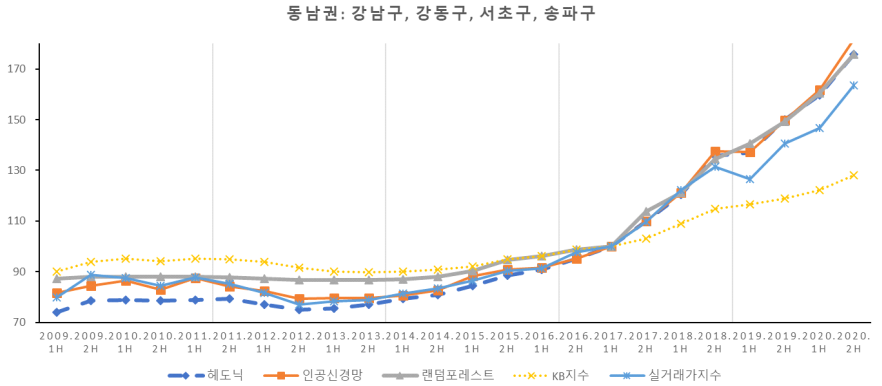
〈그림 3〉은 각 모형을 통해 도출한 아파트 매매가격지수를 보여준다. Panel A는 서울시 아파트 매매가격지수이고 Panel B부터 F는 권역별 아파



Panel A: 서울시 아파트 매매가격지수



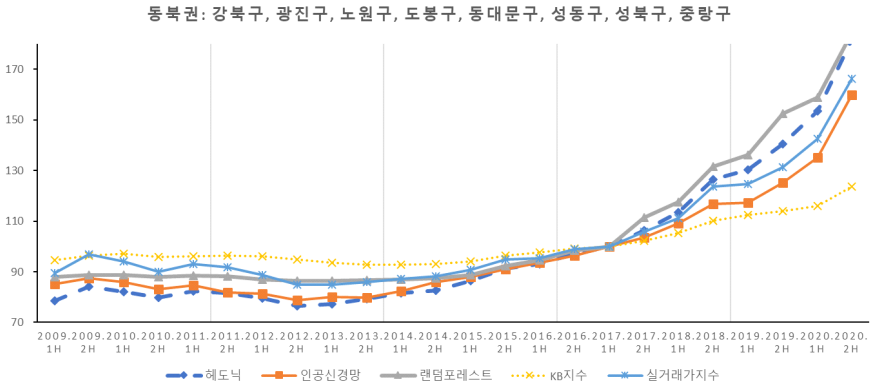
Panel B: 도심권 아파트 매매가격지수



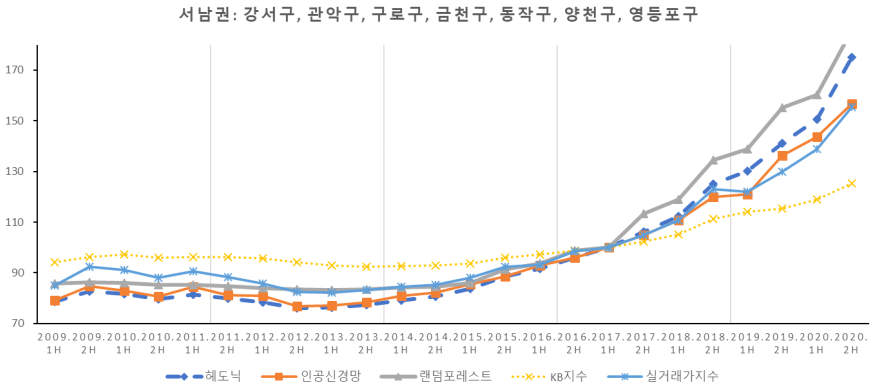
Panel C: 동남권 아파트 매매가격지수

〈그림 3〉 아파트 매매가격지수

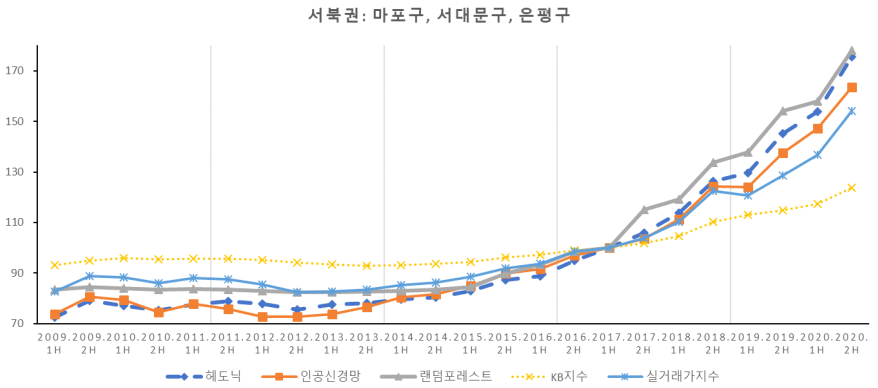
 https://doi.org/10.30902/jrea.2022.8.3.1



Panel D: 동북권 아파트 매매가격지수



Panel E: 서남권 아파트 매매가격지수



Panel F: 서북권 아파트 매매가격지수

〈그림 3〉 Continued

트 매매가격지수이다. 서울시의 하위시장(sub-market)으로서 도심권, 동남권, 동북권, 서남권, 서북권[9]을 선정하여 권역별 지수를 추정하였는데, 하위시장을 대표하는 기준으로 5개 권역을 선택하는 것에는 크게 2가지 장점이 있다. 첫 번째, 권역별 분류기준은 서울 도시기본계획에 근거한 분류기준이므로 하위 영역의 동질성에 대한 충분한 설명이 가능하다. 두 번째, 한국감정원에서 발표하는 실거래가의 경우 자치구별 지수를 공표하지 않고, 서울시의 하위 집단으로서 권역별 실거래가를 발표하기 때문에 권역별 지수를 추정한다면 기존의 지수와 비교할 수 있다.

2. 아파트 매매가격의 누적 상승률

본 절에서는 〈그림 3〉의 결과를 바탕으로 인공신경망, 랜덤포레스트, 그리고 헤도닉 방법론으로 계산한 아파트 매매가격지수와 현재 사용되는 KB지수 및 실거래가 지수를 비교한다. 우선 KB지수의 경우 다른 지수들에 비해 아파트 시장의 변동성을 많이 축소하는 경향이 관찰되었다. 이는 다수의 선행연구에서 언급한 것처럼 추정 가격을 사용함에 따라 나타나는 지수의 평활 현상 때문으로 해석할 수 있다. 나머지 추정 지수들과 실거래가 지수는 유사한 추세를 보인다. 2017년 상반기 대비 2020년 하반기의 누적 상승률은 서울시의 경우 헤도닉 모형에서 75.96% 상승하였고, 인공신경망 모형에서는 60.50% 그리고 랜덤포레스트 모형은 79.18%의 상승을 보여주었다. 동기간 실거래가 지수는 60.02%, KB 지수는 25.47% 상승했다.

이를 권역별로 분류하여 살펴보면, 용산과 종로, 중구로 구성된 도심권의 동기간 누적 상승률은 헤도닉 모형의 경우 72.00%, 인공신경망 모형 43.63%, 랜덤포레스트 모형 72.02%로 관찰되었다. 한편 실거래가 지수 55.27%의 상승률로 인공신경망 지수와 랜덤포레스트 모형 사이의 상승률을 보였지만, KB 지수의 누적 상승률은 21.02%에 그쳤다.

강남, 강동, 서초, 송파구로 이뤄진 동남권은 유일하게 전 지역이 2017년 8월 최초 투기지역으로 지정된 지역으로서 다른 지역에 비해 동기간 상승률이 높을 것이라 예상되었다.[10] 실제 관찰된 상승률은 헤도닉 지수 75.71%, 인공신경망 지수 81.70%, 랜덤포레스트 지수 75.77%, KB 지수 28.06%, 실거래가 지수 63.63%의 상승률이 확인되었다. 해당 수치를 도심권의 상승률과 비교해볼 때 모든 방법론이 동남권의 상승률을 높게 측정했다. 특히 인공신경망 지수에서는 그 차이가 38%에 이를 정도로 크게 나타났다. 하지만 서울시 전체 상승률과 비교 시, 그 값에 확연한 차이가 관찰되지는 않았는데, 이는 동남권이 물론 최

9) 도심권에는 용산구, 종로구, 중구가 포함되며, 동남권은 강남구, 강동구, 서초구, 송파구로 이루어져 있다. 동북권은 강북구, 광진구, 노원구, 도봉구, 동대문구, 성동구, 성북구, 중랑구 총 8개의 자치구로 이루어져 있으며, 서남권은 강서구, 관악구, 구로구, 금천구, 동작구, 양천구, 영등포구 총 7개로 구성된다. 마지막으로 서북권은 마포구, 서대문구, 은평구로 이루어져 있다.

10) 당시 정부는 부동산 대출 및 세금 규정에 차등을 두기 위해 주택 상승률이 높은 지역을 기점으로 투기지역, 투기과열지구, 조정지역으로 분류하였다. 서울시는 2017년 8월 31일 전 지역이 투기과열지구로 지정되었고, 이 중 강남, 서초, 송파, 강동, 용산, 노원, 마포, 양천, 영등포, 강서구가 투기지역으로 지정되었다. 이후 2018년 8월 28일 종로, 동대문, 동작, 중구를 투기지역에 추가 편입되었다.

초 투기지역으로 선정된 지역이지만 서울시 전체가 동 시점에 투기과열지구로 선정된 바 있고, 이후 2018년 8월 일부 지역이 역시 투기지역으로 추가 편입되었기에 최초 투기지역으로 선정되었다고 해서 확연한 누적 상승률이 관찰되지는 않은 것으로 해석할 수 있다.

동북권은 가장 많은 8개 자치구로 이루어져 있으며, 표본 내 포함된 거래량도 총 176,268건으로 전체 대비 35% 정도 규모로 다른 권역 대비 많은 거래가 관찰된 지역이다. 동북권의 누적 상승률에 대해 헤도닉 지수는 81.92%로 헤도닉 모형은 기간 내 동북권의 아파트 매매가격 상승률이 가장 높게 관찰하였다. 인공신경망 지수는 동 기간 상승률을 59.96%로 추정했고, 랜덤포레스트 지수는 84.36%, KB 지수는 23.60%, 실거래가 지수는 66.20% 상승률을 보여주었다.

서남권은 7개 구로 이뤄져 있으며 동 기간 누적 상승률이 헤도닉 지수 75.11%, 인공신경망 지수 56.79%, 랜덤포레스트 85.69%, KB 지수 25.35%, 실거래가 지수 55.47%로 관찰되었다. 특히 랜덤포레스트 모형의 경우 다른 지역에 비해 서남권의 가격 상승률을 가장 높게 추정했다는 것이 주목할만하다.

마지막으로 서북권의 동 기간 누적 상승률은 헤도닉 지수 75.42%, 인공신경망 지수 63.60%, 랜덤포레스트 지수 78.05%, KB 지수 23.77%, 실거래가 지수 54.14%로 확인되었다. 〈표 8〉은 2017년부터 2020년까지의 서울시 및 하위시장의 누적 상승률을 지수별로 요약한 표이다.

주목할만한 점은 헤도닉 모형과 랜덤포레스트 모형을 통해 산출한 지수에서 모든 권역의 누적

〈표 8〉 지수별 누적 상승률(%) 비교

분류	헤도닉	ANN	RF	KB 지수	실거래가 지수
서울시	75.96	60.50	79.18	25.47	60.02
도심권	72.00	43.63	72.02	21.02	55.27
동남권	75.71	**81.70**	75.77	**28.06**	63.63
동북권	**81.92**	59.96	84.36	23.60	**66.20**
서남권	75.11	56.79	**85.69**	25.35	55.47
서북권	75.42	63.60	78.05	23.77	54.14

주 : 1) 기준시점은 2017년 상반기, 비교시점은 2020년 하반기.
2) ANN, artificial neural network; RF, random forest.

상승률을 기존의 KB 지수와 실거래가 지수 대비 훨씬 높게 산출했다는 점이다. 앞서 헤도닉 모형과 반복 매매모형을 통해 산출한 지수와 기존의 KB 지수를 비교했던 김명준 외(2008)의 연구에서도 이와 유사한 결과가 관찰되었다. 그리고 헤도닉 모형의 변동성이 반복 매매모형에 기반을 둔 지수에 비해 큰 이유로 반복 매매지수 모형에서 같은 매물로 취급하는 거래 건은 표본 산정의 한계로 인해 완전 동일한 매물이 아닌 동일 아파트와 동일 평형에 그치게 되고, 지수를 산출할 시 동일 아파트의 동일 평형 매물 거래 건이 유독 많은 시점이 발생하면 이런 자료들의 평균값이 해당 시점의 거래가격으로 집계되기 때문에 가격 변화율의 변동성을 축소하는 경향이 있을 수 있음을 지적했다. 해당 연구는 한국부동산원의 실거래가 지수가 공표되기 전 이뤄진 연구로, 실거래가 지수에서 사용하는 반복 매매모형과 차이가 있을 수 있으나 오늘날 부동산원이 공표하는 실거래가 지수에서 동일 주택을 가정하는 기준이 아파트 단지, 면적, 동, 층 그룹(1, 2층/최상층/중간층)으로

되어 있어 현재 기준 역시 표본의 동일 주택 기준이 완전하지 못함을 알 수 있다. 따라서 해당 모형의 한계는 여전히 유효할 것이다. 한편, 다양한 선행연구에서는 KB 지수가 다른 지수에 비해 변동성을 작게 추정한다는 것을 감정가 사용에 따른 평활화 현상 때문으로 해석한 바 있다. 이 역시 여전히 유효한 것으로 관찰되었으며, 이는 추정모형에 기반한 다른 지수들은 평활의 문제로부터 다소 떨어져 있음을 의미한다.

추정모형 중 주택가격 추정성과가 가장 좋았던 랜덤포레스트 모형의 경우 기간 내 상승률을 79.18%로 나타내는데, 이는 가장 낮은 변동성을 보인 KB 지수의 상승률 25.47%에 비해 3배 이상 높은 수치다. 물론 단순히 지수의 변동성이 크다고 해서 더 정확한 지수라고 볼 수는 없다. 하지만 기존의 주택가격지수의 상승률이 소비자 체감보다 낮았음을 고려할 때 본 연구에서 제시하는 매매가격지수에는 다른 주택가격지수에서 누락된 정보를 포함하고 있음을 시사한다.

3. 토론

본 연구는 기계학습 방법론에 기반한 아파트 매매가격지수 설계 방법을 제안한다. 인공신경망과 랜덤포레스트 방법론을 사용함으로써 주택특성요인과 매매가격 사이의 관계를 선형적으로 해석하는 헤도닉 모형의 한계를 극복하였다. 이렇게 계산된 정확한 추정매매가격을 바탕으로 아파트 매매가격지수를 작성하였으며, 이를 부동산원에서 발표하는 실거래가 지수, KB 부동산지수, 그리고 헤도닉 방법론으로 계산한 지수와 비교한다. 다만 정확하게 관측되지 않는 시장의 흐름을 포착하려는 지수 설계의 특징을 고려할 때, 각 지수 간의 우월성 비교가 용이하지 않다는 한계가 있다. Case and Shiller(1987, 1989) 역시 반복매매법을 새로이 제시하였으나 기존 지수와의 명확한 통계적 비교는 제한적이다. 이러한 환경에서, 주택매매가격지수 간의 우월성을 통계적 방법론으로 비교하는 데에는 어려움이 따르며, 따라서 각 지수의 장단점을 비교하고 본 연구에서의 매매가격지수가 지니는 차별성을 확인할 필요가 있다.

주택에 대한 가격지수를 설계할 때, 개별 주택가격에 대한 예측 정확성보다는 표본에 대한 예측평균의 불편향성이 더 중요하다.[11] 이러한 관점에서 볼 때, 본 연구의 방법론은 기존의 반복매매법 등에 비해 모집단에 가까운 표본을 구성하기 용이하다는 장점이 있다. 본 연구에서 제시하는 기계학습 기반 방법론은 매매가격지수 산정 과정에서 모집단 전수에 가까운 주택의 가격을 정확도 높게 추정하고 이를 바탕으로 지수를 계산하므로, 모집단의 일부만 표본으로 사용하는 방법론에 비해 모집단 유사성을 높게 유지할 수 있다. 만약 표본집단이 모집단을 대표하지 못하는 경우, 가중치 조정 등을 통해 대표성을 높이는 방법 또한 사용할 수 있다.

다만 본 연구에서는 네이버 부동산에서 제공하는 자료를 수집하였으므로, 연구에 사용된 표본

11) 본 내용을 지적해주신 익명의 심사위원께 감사드립니다.

과 실제 모집단 간의 차이가 발생할 수 있다. 특히 네이버 부동산에 일부 주택특성정보가 누락되어 있는 경우 표본에서 제외되었는데, 이 과정에서 표본 편향이 발생하였을 가능성이 있다. 다만 본 연구에서는 기계학습을 활용한 주택가격지수 작성방법론을 제안하는 것이 주목적이며, 실제로 본 방법론을 이용하여 관련 지수를 작성할 때는 표본 집단이 모집단을 정확하게 대표할 수 있도록 가공할 필요가 있다.

본 연구의 방법론은 반년 단위의 지수 발표로 인해 실용성이 제한적이라는 문제를 지닌다. 기계학습 방법론의 특성상 충분한 거래량 확보를 위해 기간을 늘렸으나, 실용적 측면에서 부족함이 있다. 따라서 본 연구에서 제시하는 방법론은 기존 지수의 대체재가 아닌 보완재로서의 성격을 지닌다. 후속 연구에서는 오버샘플링 등을 활용하여 합성데이터를 생성함으로써 표본 수를 늘리고 지수 발표 기간을 줄이는 방안을 제시할 수 있다. 종합하여 볼 때, 본 연구에서 제안하는 기계학습 기반 매매가격지수는 기존 지수의 대체재가 아니라, 정확도 높은 아파트 매매가격 추정을 기반으로 기존 지수에서 누락된 정보를 보완하는 역할을 한다.

Ⅵ. 결론

본 연구는 부동산시장의 변동성을 확인하기 위한 수단으로서 정확한 아파트 매매가격지수의 필요성과 함께 기존의 아파트 매매가격지수가 가지는 한계를 확인하였다. 그리고 기존 한계를 극복할 수 있는 대안으로 기계학습 방법론을 통한 주택가격 추정과 아파트 매매가격지수 산출을 제안한다. 이 과정에서 기존 선행연구를 참고하여 부동산시장을 분석하는 데 높은 성과를 보인 인공신경망 방법론과 랜덤포레스트 방법론을 활용했으며, 기존의 주택가격지수 추정 방법론인 헤도닉 방법론을 벤치마크로 활용해 지수를 비교, 분석하였다.

본 연구의 결과는 크게 3가지로 요약할 수 있다. 첫째, 기계학습에 기초한 매매가격지수는 이용 가능한 모든 실거래 정보를 표본으로 사용함으로써 분석의 정확성을 높인다. 기계학습 방법론은 아파트의 매매가격과 특성 변수 간의 비선형성을 포착함으로써 실제 거래되지 않은 아파트에 대한 신뢰할 수 있는 추정 매매가격과 매매가격지수를 도출할 수 있다. 따라서 본 연구에서 제시하는 매매가격지수 작성방법은 기존 지수의 방법론과 표본 상의 한계로 인해 아파트 시장의 변동성을 정확하게 반영하지 못하는 문제를 극복하는 방안을 제시한다.

둘째, 기계학습 방법론을 사용한 매매가격 추정모형 설명력은 헤도닉 방법론 기반의 모형 대비 훨씬 우수한 것으로 나타났다. 이 중 랜덤포레스트 방법론에 기초한 추정모형의 설명력과 성과가 가장 우수했다. 랜덤포레스트 모형에서는 추정값이 실제 아파트 실거래가를 98% 수준에서 설명할 수 있었다. 그리고 인공신경망 모형에서는 96% 수준의 설명력을 보였다. 반면, 기존의 지수 산출 방법론으로서 벤치마크로 사용한 헤도닉 방법론 모형은 아파트 실거래가를 85% 수준에서 설명했다.

셋째, 기계학습 방법론을 사용한 매매가격 추

정모형에 기초한 매매가격지수는 변동성이 커지는 가격 상승 시점에서 기존 지수보다 더 큰 변동성을 가지는 것으로 확인되었다. 본 연구에서는 모든 지수에서 공통으로 아파트 매매가격 상승이 관찰된 2017년 상반기부터 2020년 하반기까지를 비교했고, 기계학습 방법론을 사용한 매매가격 추정모형에 기반을 둔 새로운 지수가 기존의 KB 지수나 실거래가 지수보다 더 큰 상승 폭을 보이는 것을 확인했다. 기존 지수가 평활의 문제를 갖고 있다는 한계에 비추어 볼 때 이는 본 연구에서 제시하는 새로운 매매가격지수가 시장 흐름을 반영하고 있는 것으로 해석할 수 있다.

기계학습 방법론을 사용한 매매가격 추정모형은 높은 설명력과 성과를 바탕으로 아파트 매매가격지수 산출에 설득력을 부여한다. 또한, 본 연구에서 제시한 방법론으로 훈련된 주택매매가격 추정모형을 활용하면 실제 거래되지 않은 아파트 매매가격을 추정할 수 있어 주택담보대출 과정에서 담보가액을 평가하거나 주택 보유세 부과를 위한 보다 합리적인 공시지가를 산출하는 데 기여한다. 추정모형의 높은 설명력을 바탕으로 아파트 매매가격지수를 산출한다면 시장의 움직임을 적시에 파악해 정부 정책 방향성을 설정하는 데 도움을 줄 수 있고, 주택의 수요자나 공급자가 이후 자산계획을 세울 수 있도록 정보를 제공한다. 나아가 법적 · 제도적 정비를 통해 부동산지수 상품화가 이뤄진다면, 실거래가에 기초한 해당 아파트 매매가격지수는 단순 투자상품으로서 가치뿐만 아니라 위험의 헤지(risk hedging)를 통해 부동산시장과 금융시장의 안정화에 이바지할 수 있을 것으로 기대된다.

부동산에 대한 국민의 관심이 점점 높아지고 있는 상황에서 정부 대책의 중요성이 커지고 있으며, 이를 뒷받침할 통계의 역할이 강조되고 있다. 특히 실제 거래되지 않는 주택의 가격을 추정할 수 있는 통계학적 방법론이 계속해서 개선되고 있다는 점은 실거래가에 기반을 둔 주택가격지수에 대한 기대감을 증대시킨다. 더욱 정확하고 시의성 있는 부동산지수 개발이 요구되는 시점이다.

ORCID

김이환 https://orcid.org/0000-0001-9333-3450
김형준 https://orcid.org/0000-0003-0386-005X
류두진 https://orcid.org/0000-0002-0059-4887
조훈 https://orcid.org/0000-0003-2322-320X

참고문헌

1. 강승우, 2017,「서울시 아파트 전매 프리미엄 결정요인에 관한 연구」,『부동산 · 도시연구』, 10(1): 121-144.
2. 금상수 · 한광호 · 백민석, 2014,「천안시 아파트시장의 특징과 가격형성요인」,『감정평가학 논집』, 13(2): 31-41.
3. 김명준 · 박광우 · 신용현 · 조훈 · 현정순, 2008,「주택가격지수 산정: 서울 아파트 실거래가격을 이용한 실증연구」,『금융경제연구』, 348:1-68.
4. 김형준 · 류두진 · 조훈, 2018a,「주택담보대출의 조기상환행태 분석: 안심전환대출 출시 이후의 이상현상을 중심으로」,『경영학연구』, 47(4):865-887.
5. _____, 2018b,「주택연금 유동화

증권에 관한 연구: 구조설계 및 예상현금흐름 분석을 중심으로」, 『한국증권학회지』, 47(2):327-347.

6. _____, 2019, 「기업부도예측과 기계학습」, 『금융공학연구』, 18(3):131-152.
7. 네이버 부동산, 2022, 네이버, Accessed September 1, 2022, https://land.naver.com/
8. 노영훈, 2007, 「부동산시장과 부동산 조세정책과제」, 세종: 한국조세재정연구원.
9. 류강민 · 이상영, 2010, 「S&P/Case-Shiller 반복매매모형을 이용한 주택가격지수 산정에 관한 연구」, 『주택연구』, 18(2):183-204.
10. 박대현 · 김정환 · 류두진, 2021, 「기계학습 기반 주택시장의 조기경보체계」, 『부동산분석』, 7(1):29-45.
11. 배상영 · 정의철 · 이상엽, 2018, 「도시철도 교통서비스가 주택가격에 미치는 영향」, 『부동산학연구』, 24(3):85-98.
12. 배성완 · 유정석, 2018a, 「기계 학습을 이용한 공동주택가격추정: 서울 강남구 사례로」, 『부동산학연구』, 24(1):69-85.
13. _____, 2018b, 「머신 러닝 방법과 시계열 분석 모형을 이용한 부동산 가격지수 예측」, 『주택연구』, 26(1):107-133.
14. _____, 2018c, 「표본 주택 가격 기반 부동산 가격지수 산정: 머신 러닝 방법의 활용을 중심으로」, 『주택연구』, 26(4):53-74.
15. 송영선 · 윤명탁 · 이창무, 2020, 「아파트 하위시장 실거래가 지수 산정방식 비교 연구」, 『부동산분석』, 6(3):1-19.
16. 신한은행, 2021, 「2021 보통사람 금융생활 보고서」, 서울: 신한은행.
17. 이번송 · 정의철 · 김용현, 2002, 「아파트 단지특성이 아파트 가격에 미치는 영향 분석」, 『국제경제연구』, 8(2):21-45.
18. 이옥자 · 최진배, 2015, 「부산지역의 아파트가격 결정요인에 관한 연구: 동 · 서쪽을 중심으로」, 『주거환경』, 13(2):53-66.
19. 이용만 · 이상한, 2008, 「국민은행 주택가격지수의 평활화 현상에 관한 연구」, 『주택연구』, 16(4):27-47.
20. 이창로, 2015, 「비모수 공간모형과 앙상블 학습에 기초한 단독주택가격 추정」, 서울대학교 박사학위논문.
21. 이창무 · 김용경 · 배익민, 2007, 「반복매매모형을 이용한 아파트 실거래지수 운영특성 분석」, 『부동산학연구』, 13(2):21-40.
22. 윤병우 · 최경욱, 2017, 「교육환경과 아파트 전세가격간의 관계 분석」, 『부동산학보』, 47:23-38.
23. 이형욱 · 이호병, 2009, 「서울시 주택가격지수의 모형별 예측력 비교 분석」, 『부동산학보』, 38:215-235.
24. 정성훈 · 진창하, 2020, 「머신 러닝 방법을 이용한 오피스 임대료 산정: 랜덤 포레스트, 인공 신경망, 서포트 벡터 머신 활용을 중심으로」, 『부동산학연구』, 26(2):23-53.
25. 최한종, 2021, 문재인 정부 들어 서울 아파트값 93%↑…17% 올랐다는 정부 통계는 거짓, 6월 23일, 한국경제신문.
26. 하유정 · 이현석, 2020, 「교육환경이 아파트 가격에 미치는 영향: 부산시를 중심으로」, 『부동산 · 도시연구』, 13(1):47-61.
27. 한국감정원, 2017, 「전국주택가격동향조사 통계정보 보고서」, 대구: 한국감정원.
28. 한다솜 · 최창규, 2018, 「우이신설선 건설이 주변 아파트 가격에 미치는 영향에 관한 연구」, 『한국지역개발학회 학술대회』, 792-819.
29. 홍정의, 2021, 「랜덤 포레스트 알고리즘을 통한 주택 대량평가모형 연구」, 『부동산분석』, 7(1):1-28.
30. 홍정의 · 김형준 · 안세륭, 2022, 「서울 아파트 가격은

거품인가?」, 『부동산분석』, 8(1):1-21.

31. Breiman, L., 2001, "Random forest", *Machine Learning*, 45(1):5-32.
32. Case, K. E. and R. J. Shiller, 1987, "Prices of single-family homes since 1970: New indexes for four cities," *New England Economic Review*, Sep.:45-56.
33. _____, 1989, "The efficiency of the market for single family homes," *American Economic Review*, 79(1):125-137.
34. Čeh, M., M. Kilibarda, A. Lisec, and B. Bajat, 2018, "Estimating the performance of random forest versus multiple regression for predicting prices of the apartments," *ISPRS International Journal of Geo-Information*, 7(5):168.
35. Hong, J., H. Choi, and W. S. Kim, 2020, "A house price valuation based on the random forest approach: The mass appraisal of residential property in South Korea," *International Journal of Strategic Property Management*, 24(3):140-152.
36. James, G., D. Witten, T. Hastie, and R. Tibshirani, 2013, *An Introduction to Statistical Learning: With Applications in R*, New York, NY: Springer.
37. KB주택가격동향, 2022, Accessed September 1, 2022, http://kbland.kr
38. Kim, H., H. Cho, and D. Ryu, 2020, "Forecasting consumer credit recovery failure: Classification approaches," *Journal of Credit Risk*, 17(3): 117-140.
39. _____, 2021, "Predicting corporate defaults using machine learning with geometric-lag variables," *Investment Analysts Journal*, 50(3):161-175.
40. _____, 2022, "Corporate bankruptcy prediction using machine learning methodologies with a focus on sequential data," *Computational Economics*, 59(3):1231-1249.
41. Park, D. and D. Ryu, 2021, "A machine learning-based early warning system for the housing and stock markets," *IEEE Access*, 9:85566-85572.
42. Rosenblatt, F., 1958, "The perceptron: A probabilistic model for information storage and organization in the brain", *Psychological Review*, 65(6):386-408.
43. Rumelhart, D. E., G. E. Hinton, and R. J. Williams, 1986, "Learning representations by back-propagating errors", *Nature*, 323(6088): 533-536.
44. Zur, R. M., Y. Jiang, L. L. Pesce, and K. Drukker, 2009, "Noise injection for training artificial neural networks: A comparison with weight decay and early stopping", *Medical Physics*, 36(10):4810-4818.

논문접수일 : 2022년 5월 28일
심사(수정)일 : 2022년 10월 29일
게재확정일 : 2022년 11월 15일

국문초록

본 연구는 기계학습 방법론을 통한 새로운 주택가격지수 산출 방법을 제안한다. 기존 연구에서 우수성이 입증된 랜덤포레스트와 인공신경망 방법론을 적용하였다. 훈련 과정에는 주택 실거래 자료와 개별 주택 정보를 매칭하여 사용하였다. 연구 결과, 기계학습 방법론을 사용하는 주택가격 추정모형은 헤도닉 방법론에 비해 설명력과 추정성과 측면에서 보다 우수한 것으로 나타났으며, 이 중 랜덤포레스트 방법론에 기초한 주택가격 추정모형의 설명력과 성과가 가장 우수했다. 또한, 기계학습방법론을 활용한 주택가격 추정모형을 기반으로 작성된 주택매매가격지수는 변동성이 커지는 가격 상승 시점에서 기존 지수에 비해 더 큰 변동성을 가지는 것으로 확인되었다. 기존 지수가 평활의 문제를 갖고 있다는 한계에 비추어 볼 때 이는 기계학습 방법론은 활용한 새로운 추정 지수가 시장 흐름을 보다 잘 반영하는 하나의 대안이 될 수 있음을 시사한다.

주제어 : 기계학습, 랜덤포레스트, 아파트 매매가격지수, 인공신경망, 헤도닉