딥러닝 모형을 활용한 서울 주택가격지수 예측에 관한 연구
다변량 시계열 자료를 중심으로

Prediction of Seoul House Price Index Using Deep Learning Algorithms with Multivariate Time Series Data

저자 (Authors)
이태형, 전명진
Lee, Tae Hyeong, Jun, Myung-Jin

출처 (Source)
주택도시연구 8(2), 2018.8, 39-56(18 pages)
SH Urban Research & Insight 8(2), 2018.8, 39-56(18 pages)

발행처 (Publisher)
SH도시연구원
SH Urban Research Institute

URL
http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE07530536

APA Style
이태형, 전명진 (2018). 딥러닝 모형을 활용한 서울 주택가격지수 예측에 관한 연구. 주택도시연구, 8(2), 39-56

주택도시연구 제8권 제2호
SH Urban Research & Insight Vol. 8, No. 2, pp. 39~56(2018. 8)
https://doi.org/10.26700/shuri.2018.08.8.2.39

딥러닝 모형을 활용한 서울 주택가격지수 예측에 관한 연구*

- 다변량 시계열 자료를 중심으로 -

Prediction of Seoul House Price Index Using Deep Learning Algorithms with Multivariate Time Series Data

이태형** · 전명진***
Lee, Tae Hyeong · Jun, Myung-Jin

Abstract

This study aims to evaluate the predictability of Deep Learning Neural Network algorithms (RNN and LSTM) in the forecast of the Seoul apartment price index. For the empirical analysis, we collect monthly housing price index data for the medium-sized and large-sized apartment units in Seoul during January 2006-October 2017 period. We also collect six macroeconomic variables that are known to affect housing price including expected inflation rate, rental price index, debt interest rate, stock price index, consumer price index, and unemployment rate. For the comparative purpose, we build Vector Autoregressive model (VAR) for multivariate time-series forecast. The analysis results indicate that the LSTM model best performed with the lowest RMSEs (0.826 and 1.038) for the medium-sized and large-sized apartment price indices, respectively, which is about 52 and 63 percent reductions from the VAR's RMSE (1.708 and 2.825). We also found that standard deviations of predicted values from the LSTM are substantially lower than those of simple RNN, indicating higher stability of predicted price index from the LSTM than simple RNN.

| Key Words | 주택가격지수, 벡터자기회귀모델, 인공신경망, 딥러닝
Deep Learning, LSTM, Vector Autoregressive Model, Housing Price Index

* 이 논문은 2016년도 중앙대학교 연구년 결과물로 제출됨.
** 중앙대학교 도시계획부동산학과 박사과정수료 (주저자)
*** 중앙대학교 도시계획부동산학과 교수 (교신저자: mjjun1@cau.ac.kr)

1. 서론

과거 50년 동안 우리나라 주택 가격은 내 · 외적 요인으로 인해 등락을 반복해 왔다. 1980년대 중반 3저 호황(저금리, 저유가, 저달러)과 수도권으로의 인구 유입에 따른 부동산 투기로 주택가격이 폭등하였다가 노태우 정부 시절 수도권 5개 신도시 건설 등 주택 200만호 건설을 통한 주택 공급 확대 정책으로 주택 가격이 안정화 되었다. 그러나 1997년 말 IMF 외환 위기 이후 주택 가격이 폭락하였다가 2000년대 초반에 주택 시장이 상승국면으로 전환되었다. 노무현 정부 들어 강남과 분당에서 시작된 부동산 투기 열풍이 강북과 수도권으로 확대되는 상황에서 2006년 3.30 부동산 종합대책 등의 강력한 부동산 시장 규제 정책을 통해 주택 가격 안정화를 도모하였다. 2008년 글로벌 금융 위기로 인해 부동산 시장은 다시 침체 국면에 들어갔다가 2010년대 중반 이후 서서히 상승국면으로 접어들었다.

주택 가격의 급격한 변동은 국가 거시경제뿐만 아니라 국민의 삶의 질에 부정적 영향을 미치는 중요한 사안이기 때문에 주택 가격 변동에 영향을 미치는 요인에 대한 연구와 미래 주택가격 변화를 예측하는 모형의 개발이 활발하게 진행되어 왔다. 주택 가격 변동 요인에 대한 정확한 이해와 예측력 높은 주택가격 예측 모형의 개발은 부동산 시장에서의 이해 당사자인 부동산 개발자 및 투자자, 공인중개사, 감정평가사, 주택 담보대출 금융기관뿐만 아니라 부동산 정책을 추진하는 공무원과 주택을 거래하는 일반 시민 모두에게 유익한 정보를 제공할 수 있기 때문에 이에 대한 심도 있는 연구가 요구되고 있다. 주택 가격을 평가하고 예측하는 전통적인 통계기법은 개별 주택의 물리적 구조와 입지 및 환경적 특성에 기반하여 특정 주택의 가격을 평가하는 헤도닉 가격 모형(hedonic price model)과 과거 주택 가격 변화 추이를 분석하여 미래 주택 가격을 예측하는 시계열 모형(time series model)으로 구분이 가능하다. 헤도닉 모형의 경우 전통적인 OLS(ordinary least square)추정법에서 최근 시간적 상관성(serial correlation)과 공간적 이질성(spatial heterogeneity)을 동시에 고려한 GWR-TS(geographically weighted regression-time series)모형으로 발전하고 있다(Fotheringham et al., 2015).

시계열 모형의 경우 주택가격 변동 추세 변수만을 이용하는 일변량 시계열 모형인 ARIMA모형(Auto Regressive Integrated Moving Average model)과 주택가격 추세뿐만 아니라 주택가격에 영향을 주는 거시경제 지표나 주택 공급 및 수요 변수를 포함하는 다변량 모형인 벡터자기회귀모형(Vector Auto Regressive Model: VAR)으로 구분할 수 있다. ARIMA모형의 경우 자기상관(autocorrelation), 이동평균(moving average), 추세관계(cointegration) 등을 고려하여 주택 가격의 미래를 예측하는 모형인 반면 벡터자기회귀모형(VAR)은 다변량 정상 시계열 변수로 구성된 연립방정식 체계를 가지며 특정 변수의 변화가 내생변수에 미치는 동태적 효과를 파악할 수

있는 충격반응분석(impulse response analysis)과 전체 변동에 대한 내생변수 변동의 상대적 크기를 분석할 수 있는 분산분해(variance decomposition)분석이 가능하다(문권순, 1997).

최근 빅데이터를 활용한 기계학습(machine learning: ML) 알고리즘의 비약적 발전으로 인해 도시 분야에서도 인공지능 기법을 이용한 시계열 예측 모형이 개발되고 있다. 인공신경망(Artificial Neural Networks: ANNs)을 이용한 시계열 예측 방법은 전통적인 시계열 통계분석법과는 달리 변수들 간의 상호 독립성 가정, 오차항의 분포에 대한 제약, 변수들 간의 선형성 가정, 식별의 문제 등으로부터 자유로운 상황에서 분석 및 예측이 가능해 모형의 적용 범위가 넓고 예측력이 높은 장점을 가지고 있다(Jain & Payal, 2011).

기계학습 알고리즘은 크게 지도학습(supervised learning), 비지도학습(unsupervised learning), 강화학습(reinforcement learning)으로 구분되는데 실제 기계학습 응용 사례는 타깃 변수의 값이 존재하는 지도학습 예측 모형이 대부분을 차지한다. 지도학습 예측 모형의 경우 예측하는 결과 값이 연속형 변수인 경우 회귀모형(regression)을, 이산형 변수인 경우 분류(classification)모형을 사용한다.

인공지능을 활용한 시계열 예측모형으로 Deep Neural Network(DNN)의 일종인 Recurrent Neural Network(RNN)은 시간 연속(time sequence)을 고려할 수 있는 장점이 있어 기존의 multilayer perceptron(MLP) neural networks보다 더 강력한 인공지능 시계열 예측모형으로 알려져 있다(Brownlee, 2017). 그러나 RNN이 MLP보다 우수한 시계열 예측모형임에도 불구하고 시간 역전파(backpropagation through time: BPTT)과정에서 발생하는 기울기 값이 사라지는 문제(vanishing gradient problem)로 인한 한계를 가지고 있다. 이 문제를 해결하기 위한 대안으로 개발된 인공지능 시계열 예측 알고리즘이 Long Short-Term Memory(LSTM) 알고리즘이다(RNN과 LSTM 알고리즘의 구조는 제3장의 분석 모형과 자료에서 구체적으로 설명한다).

본 연구는 RNN과 LSTM 알고리즘을 이용하여 서울 주택가격지수 예측 모형을 구축하는 것을 목적으로 한다. 이를 위하여 2006년 1월 ~ 2017년 10월 기간 동안 월별 서울 중대형(85㎡초과 135㎡이하)과 대형(135㎡초과) 아파트 가격지수 자료를 타깃 변수(target variable)로, 선행연구에서 아파트 가격에 영향을 미치는 것으로 나타난 거시경제 지표 중 기대인플레이션율(김현재, 2011), 주택전세가격지수(김현재, 2011; 고종완, 2014), 대출금리(이훈자, 2017; 이석원 2018), 주가지수(이훈자, 2017), 소비자물가지수(김현재, 2011; 이훈자 2017), 실업률(김현재, 2011)등 6개의 거시경제 지표를 투입 변수로 활용한다. 또한 전통적 다변량 시계열 통계 모형인 벡터자기회귀모형(VAR)을 구축한 후 인공지능 시계열 모형과 VAR 모형의 예측력을 비교 분석한다.

본 연구의 구성은 다음과 같다. 제2장에서는 선행연구 검토 및 연구 차별성을 기술하고 제3장에서는 분석 모형과 자료에 대한 내용을 설명한다. 제 4장에서는 VAR, RNN, LSTM 모형에

대한 실증 분석 결과를 제시한다. 제5장에서는 모형의 비교 분석 결과 및 시사점을 제시한다.

2. 선행연구 검토 및 연구의 차별성

본 연구 주제와 관련한 연구 분야는 인공지능 알고리즘 개발 분야, 주가지수, 대기오염, 날씨 등 다양한 시계열 자료의 예측에 대한 인공지능 기법 활용 등을 포함하고 있지만 본 연구에서는 인공지능 알고리즘을 활용한 주택가격지수 예측 모형과 관련한 최근의 국내외 연구들을 중심으로 선행연구를 검토한다.

기계학습을 이용하여 주택가격을 예측한 최근 해외 선행연구들을 정리하면 다음과 같다. 카미스(Khamis, 2014)는 MLR(Multiple Linear Regression)과 Neural Network 모델을 이용하여 뉴욕의 주택가격을 추정하는 연구를 수행하였다. 인공신경망(ANN)분석의 경우 입력층과 은닉층, 그리고 출력층으로 구성하였고, 은닉층을 구성하는 뉴런과 층의 수를 조절하여 실제값과 예측값의 평균제곱오차(MSE)가 최소가 되는 값을 구하였다. 분석 결과 Neural Network 모델의 R2 값이 MLR에 비해 26.5% 높고, MSE는 26.3% 낮게 나왔다. 박과 배(Park & Bae, 2015)는 미국 버지니아 주 Fairfax county의 5359개 연립주택의 거래 자료를 이용하여 Ripper, Naïve Bayesian, AdaBoost와 같은 머신러닝 알고리즘으로 주택가격을 예측하는 연구를 수행 하였다. 주택 가격에 영향을 미치는 요인들 중 주택의 물리적 특성과 공립학교 순위, 주택담보 대출 금리 등 총 27가지 변수를 이용하여 분석한 결과 여러 머신러닝 알고리즘 중 Ripper 알고리즘의 예측력이 가장 높다는 것을 확인하였다.

카라파라흐(Khalafallah, 2008)는 다층퍼셉트론(MLP) 신경망 모형으로 미국 올랜드 지역의 9년 동안의 주택 거래 시점, 평균 이자율, 전년 대비 중간 주택 가격의 변화율 등 8개 변수를 입력 데이터로 이용하여 3개월 후의 주택 가격을 예측하였다. 최적의 MLP 모델을 찾기 위해 은닉층의 개수와 뉴론의 수, 전달 함수, 훈련 방법, validation과 test 샘플의 수를 조절하여 분석을 실시하였다. 연구 결과 최종 선택된 모델의 예측 오차 범위는 −2~2%로 나타났다. 라드찌(Radzi, 2012)등은 인공신경망(ANN)을 이용하여 말레이시아의 주택가격지수(HPI)를 예측하였다. 이 연구에서 주택가격지수를 종속변수로 하고 주택 시장의 구조를 반영할 수 있는 실업률, 인구 규모, 주택담보대출금리, 가구주의 수입을 독립변수로 사용하였다. ANN 분석을 위한 도구로는 역전파 방법인 NeuroShell 2를 사용하였고, 시행착오를 거쳐 입력층과 은닉층, 출력층의 뉴런 수를 결정하였다. 그 결과 은닉층의 뉴런 수가 5이고, training과 testing 자료 비율을 80% : 20%로 했을 때 상관계수가 0.9966으로 가장 높고, 평균 제곱오차(MSE)는 0.819로 가장 낮다는

분석결과를 제시하였다. 엔구옌과 크립스(Nguyen & Cripps, 2001)는 인공신경망(ANN)과 다중회귀분석을 이용하여 단독주택 가격에 대한 예측력을 비교하였다. 이 연구는 1993년 1월 1일부터 1994년 6월 30일까지 6 분기 동안 미국 테네시 주 Rutherford 카운티에서 판매된 총 3,906개의 단독주택에 대해 거실면적, 침실 수 등 주택의 물리적 속성들을 이용하여 주택가격을 분석하였다. 분석 결과 기계학습을 위한 표본의 수가 충분할 경우에는 변수에 대한 이론적 기초와 관계없이 ANN 모델을 추천하고, 그렇지 않을 경우에는 다중회귀분석 모형을 추천하였다.

최근 국내에서도 인공지능을 활용한 주택가격 예측 연구가 활발히 진행되고 있다. 정원구와 이상엽(2007)은 다중퍼셉트론(MLP)을 이용하여 1999년 9월부터 2005년 9월까지의 거시경제지표변수(실질 GDP, 회사채수익률 등 7개 변수)와 서울, 강북, 강남의 공동주택가격지수로 인공지능망 모형을 구축한 후 이들 지역의 공동주택가격지수를 예측 하였다. RMSE값으로 표현된 MLP 모형의 예측력은 각각 서울 2.54, 강남 3.53, 강북 1.02로 나타났다. 이 연구는 기계 학습에 사용된 자료가 72개로 적었고, 시간을 통한 역전파 알고리즘(BPTT)을 사용하지 않았다는 한계가 있다. 이형욱과 이호병(2009)은 ARIMA 모형과 ANN 모형을 이용하여 서울시 주택가격지수를 예측하였다. 분석결과 두 모형의 예측력에 있어 통계적으로 유의미한 차이를 발견하지 못했으나 주택가격 변동 폭이 크지 않은 특정 하위 주택 시장(4 군집 시장)의 경우 ANN 모형의 예측력이 ARIMA 모형보다 현저히 높다는 분석 결과를 제시하고 있다.

민성욱(2016)은 딥러닝(Deep Learing)을 이용한 주택가격 예측모형 연구를 통해 2006년 1월부터 2016년 2월까지의 서울 주택가격지수(매매가격, 전세가격, 지가)를 분석하고, 그 결과 인공신경망을 이용한 분석의 예측력이 높다는 것을 확인하였다. 배성완과 유정석(2017)은 2006년 1월부터 2015년 12월까지의 서울 지역 부동산가격지수를 이용하여 DNN과 LSTM 모형으로 부동산 가격지수를 예측하였고, ARIMA 모형을 비교대상으로 하여 평균제곱근 오차(RMSE)를 비교하였다. 그 결과 인공지능을 이용한 DNN과 LSTM의 예측력이 ARIMA 모형에 비해 우수하고, 그 중에서도 DNN 모형의 예측력이 상대적으로 높다는 것을 확인했다.

배성완과 유정석(2018)은 아파트 매매실거래가격지수를 이용하여 시계열분석 모형과 머신러닝 방법의 예측력을 비교 분석하는 연구를 수행하였다. 시계열분석 모형으로는 단변량의 경우 ARIMA, 다변량의 경우 VAR, BVAR를 이용하였고, 머신러닝 방법으로는 SVM, RF, GBRT, DNN, LSTM을 이용하였다. 분석결과 2006년 1월부터 2017년 8월까지의 전체자료를 10개 구간으로 나누어 교차검증을 실시하여 머신러닝 모형이 시계열분석 모형보다 예측력이 우수하고, 머신러닝 모형 간에는 다변량보다 단변량 변수의 예측력이 우수하다는 분석 결과를 제시하였다.

인공신경망 기법을 이용한 주택가격 예측에 대한 국내외 선행연구 검토 결과를 요약하면 다음과 같다. 첫째, 기계학습 알고리즘의 발전에 따라 연구에서 사용하는 알고리즘의 유형이 빠른

속도로 변화하고 있다. 2000~2015년에 사용된 예측 알고리즘은 MLP 신경망이나 Ripper, Naïve Bayesian, AdaBoost와 같은 머신러닝 알고리즘 등이 있으나 2016년 이후는 DNN, RNN, LSTM과 같은 신경망 알고리즘을 활용한 연구가 증가하고 있다. 둘째, 대부분의 국내외 연구에서 인공신경망 모형과 다중 회귀모형, ARIMA와 VAR 등 전통적인 통계 모형과의 예측력 비교 분석을 통해 인공신경망 모형의 우수성을 실증적으로 검정하고 있다. 셋째, 인공신경망 모형을 활용한 연구는 은닉층의 수, 은닉층 내 뉴런 수, 활성화 함수, 테스트 횟수 등을 결정하는 초모수(hyperparameters)의 반복적 조정 과정을 통해 예측력 높은 최적의 신경망 모형을 도출한다.

본 연구는 기존 선행연구와 다음과 같은 차별성을 가진다. 첫째, 본 연구는 최근 인공지능 시계열 예측분야에서 예측력을 인정받고 있는 RNN과 LSTM 알고리즘을 활용하여 다변량 주택 가격 예측 모형을 구축하였다. 국내에서는 유일하게 배성완과 유정석(2018)의 연구가 본 연구와 유사한 자료를 활용하여 다변량 LSTM 모형을 구축했는데 이들의 연구는 LSTM을 포함한 5가지 머신러닝 알고리즘을 활용하면서 초모수 설정과정에 대한 구체적 설명 없이 최종 선정 모형의 초모수값만 제시하고 있다. 반면에 본 연구는 RNN과 LSTM 모형의 설계, 모형 최적화 과정 등 예측 모형 구축 과정을 구체적으로 제시하였을 뿐만아니라 최종 선정된 초모수 값, 학습(train) 및 시험(test) 기간, RMSE 분석 결과 등에서 위 연구와 차이를 보이고 있다. 머신러닝의 경우 동일한 알고리즘을 사용하더라도 최적의 예측모형을 찾아가는 과정은 다를 수 있기 때문에 본 연구에서 제시하는 방법이 향후 유사한 연구 수행시 예측모형 구축의 지침과 방향을 제시할 수 있을 것이다.

둘째, 모수를 이용한 시계열 모형과 달리 인공신경망을 이용한 시계열 예측은 초기 모형의 무작위성(random state in an initial stage)으로 인해 매회 다른 예측값을 가진다. 따라서 모형의 안정성을 확보하기 위해서는 반복 구동을 통해 평균값과 표준편차를 구하여 예측 값의 범위를 파악하는 과정이 중요하다. 본 연구에서는 최적의 인공신경망 모형을 구축하기 위해 각각의 초모수 값에 대해 10번의 반복 구동을 통하여 예측 값에 대한 평균값과 표준편차를 구하고 이를 이용하여 초모수를 선정하였다.

3. 분석 모형과 자료

본 연구에서는 전통적인 벡터자기회귀모형(VAR)으로 서울의 주택가격지수를 예측하는 모형을 구축한 후 그 예측값을 인공지능을 활용한 시계열 모형의 예측값과 비교하여 각 모형의 예측력을 비교하고자 한다.

3.1 벡터회귀(Vector Autoregression, VAR) 모형

심스(Sims, 1980)에 의해 처음 소개된 벡터회귀(Vector Autoregressive: VAR)모형은 인과관계가 있는 k개의 현재 변수들을 종수변수로 하고, 이들 변수들의 과거 값들을 설명변수로 하는 선형 회귀방정식을 통해 시계열의 확률과정(stochastic process)을 추정하는 방법이다(박헌수 · 안지아, 2009). VAR 모형은 특별한 제약이나 경제 이론 없이도 다변량 시계열 자료를 분석할 수 있다는 장점 때문에 시계열 분석에 자주 이용되고 있다. VAR 모형의 기본적인 가정은 변수들 사이의 동적 관계가 단기적인 시간 지연(lag)에는 영향을 받지만 장기적으로 영향을 받지 않는다는 것이다.

본 논문에서 사용한 자료들을 소비자물가지수(y_{1t}), 실업률(y_{2t}), 기대인플레이션율(y_{3t}), 대출금리(y_{4t}), 주가지수(y_{5t}), 주택전세가격지수(y_{6t}), 주택가격지수(y_{7t})로 표기하고, 이들로 구성된 다변량 시계열 벡터 y_t를 다음과 같이 정의하였다.

$$y_t = (y_{1t}, y_{2t}, y_{3t}, y_{4t}, y_{5t}, y_{6t}, y_{7t})^T \quad \text{식 (1)}$$

단위근 검정을 통해 이 변수들이 모두 단위근을 가지는 불안정한 시계열이라는 것을 확인하고 모두 차분을 취하여 사용하였다.[1] VAR의 기본식은 다음과 같다.

$$y_t = A_1 y_{t-1} + \cdots + A_p y_{t-p} + c + u_t \quad \text{식 (2)}$$

3.2 순환신경망(Recurrent Neural Network, RNN) 모형

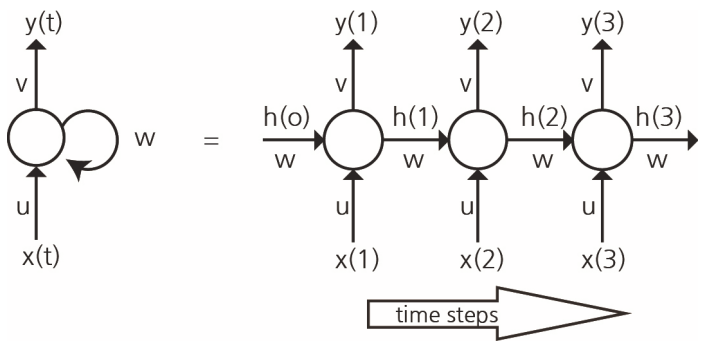
시계열 자료를 다루기 위해 개발된 인공신경망이 순환신경망(Recurrent Neural Network, RNN) 모형이다. 전통적인 다층 퍼셉트론(multilayer perceptron, MLP)이 입력 데이터들이 서로 독립적이라는 가정을 가지는 반면, RNN은 시계열 자료를 다루면서 시간에 따른 출력 자료와 입력 자료의 대응관계를 학습시킬 수 있는 신경망이다. RNN에서 시간 t에 대한 은닉 상태의 값은 시간 t의 입력 값과 시간 t-1의 은닉 상태의 값에 대한 함수이다.

1) 실증분석에서 단위근 검정을 통해 확인하였다.

3.2.1 simple RNN

전통적인 신경망처럼 RNN은 세 개의 층(입력, 출력, 은닉)으로 구성되어 있는데 이 층들은 각각에 해당되는 세 개의 가중 행렬(Weight Matrics) U, V, W를 가지고 있다. 〈그림 1〉에서 보는 것처럼 RNN 모델은 무한히 많은 층을 가지는 딥뉴럴네트워크(DNN)로 생각할 수 있다.

RNN은 시간을 통한 역전파알고리즘(Backpropagation through time, BPTT)이라고 불리는 과정을 통해 훈련을 실시한다. 전통적인 인공신경망에서의 역전파알고리즘(BP)에 비해 BPTT의 중요한 차이는 변수들이 모든 시간 단계를 공유하기 때문에 각각의 출력 값의 기울기가 현재의 시간 단계 뿐 아니라 이전 단계들에도 의존한다는 것이다(Gulli & Pal, 2017). 그런데 BPTT는 기울기 값이 소멸되거나 폭발하는 문제점이 있다. 즉, 이전 값에 대한 은닉층의 개별 기울기가 1보다 작을 경우에는 기울기가 소멸하는 문제가 발생하고, 반대로 개별 기울기가 1보다 클 경우에는 기울기가 폭발하는 문제점이 발생한다. 따라서 단순한 RNN은 장기적인 패턴을 학습하는 데 한계가 있다.



자료: Gulli and Pal(2017: 345)

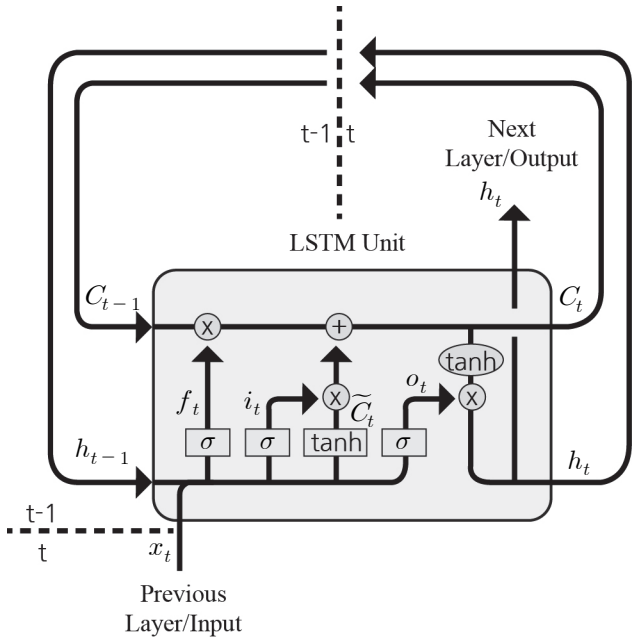
〈그림 1〉 RNN 모델

3.2.2 LSTM

LSTM(Long Short-Term Memory)은 장기적인 패턴을 학습할 때 기울기가 소멸하는 문제를 극복할 수 있는 RNN의 한 형태로 규모가 큰 순환 네트워크를 다룰 수 있고 어려운 시퀀스 문제들을 인공신경망에서 다룰 수 있게 해 준다. LSTM의 핵심 특징은 뉴런 대신에 다수의 층을 통해 연결된 메모리 블록이 있다는 것이다(Brownlee, 2017). LSTM 네트워크의 기본 모델은 〈그림 2〉와 같다.

〈그림 2〉에서 LSTM Unit의 맨 위에 보이는 선은 내부 메모리를 나타내는 셀 상태 C_t로 우리

몸에서 유전자를 전달하는 세포와 같은 역할을 하고, 아래에 보이는 선은 은닉층의 상태 h_t를 나타낸다. 하나의 메모리 블록은 그 상태와 출력을 조절하는 망각(f), 입력(i), 출력(o)의 세 개 게이트를 가진다. 각각의 게이트는 하나의 시그모이드(σ) 신경망층과 곱셉(×) 연산으로 구성되어 있다. 그 중 입력 게이트는 입력 값을 메모리에 업데이트 할 지를 결정하고, 출력 게이트는 입력과 메모리를 토대로 출력을 결정한다. 망각 게이트는 블록으로부터 버릴 정보를 결정하고, 메모리에 기억시킬 값의 범위를 조절한다(Brownlee, 2017).



자료: Olah(2015)

〈그림 2〉 LSTM 네트워크 모델

3.3 분석자료

본 연구에서는 한국감정원이 공개한 규모별 아파트 실거래가격지수(2006년 1월 ~ 2017년 10월)에서 중대형(85㎡초과 135㎡이하)과 대형(135㎡초과) 아파트 가격지수를 사용하였다.[2] 또

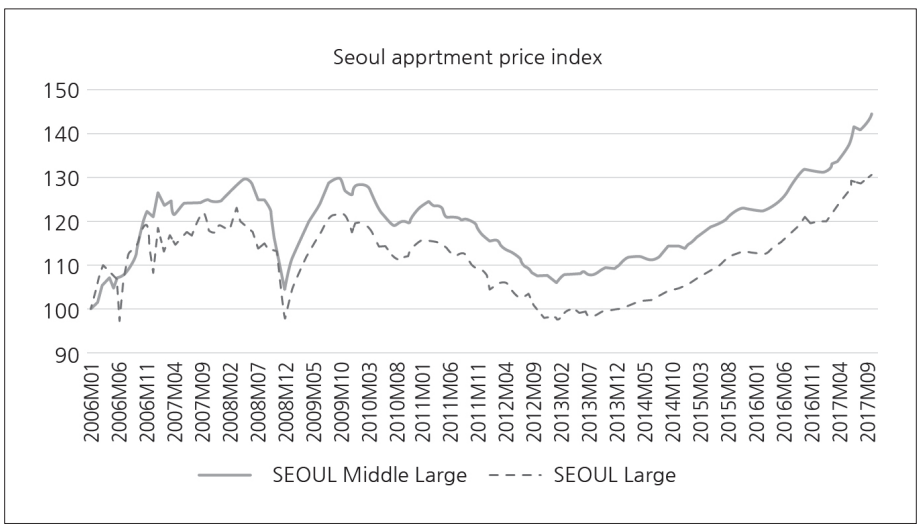
2) 다변량 시계열 모형인 VAR로 사전 분석한 결과 아파트가 대형으로 갈수록, 즉 가격이 높을수록 예측력이 떨어지는 것을 확인하였다. 중소형 RMSE 1.2, 중대형 RMSE 1.7, 대형 RMSE 2.8. 따라서 본 연구에서는 다변량 시계열 모형에서 가장 변동성이 큰 중대형과 대형 아파트의 가격지수를 타겟 변수로 선정하여 인공지능망을 이용한 예측 결과와 비교 분석을 시도하였다

한 선행연구에서 아파트 가격 변수에 영향을 주는 것으로 나타난 거시경제지표 중 소비자물가지수, 실업률, 기대 인플레이션율, 대출금리, 주가지수, 주택전세가격지수에 대하여 한국은행이 공개한 2006년 1월부터 2017년 10월까지 자료를 사용하였다.

거시경제지표와 규모별 아파트 실거래가격지수를 벡터회귀모형(VAR)으로 분석하였고, 이 결과를 RNN과 LSTM 모형으로 분석한 결과와 비교하였다.

3.3.1 주택가격지수

2006년 1월부터 20017년 10월까지 서울의 중대형, 대형 아파트 가격지수를 그래프로 표시하면 〈그림 3〉과 같다. 전체적으로 중대형아파트의 가격지수가 대형 아파트에 비해 큰 폭으로 증가하였음을 알 수 있다. 시기적으로 살펴보면 2006년 여름 대형 아파트가격이 중대형 아파트 가격에 비해 큰 폭으로 하락한 이후 2007년 이후에는 중대형아파트와 대형 아파트와의 가격지수 폭이 거의 일정하게 유지되고 있다. 중대형 아파트와 대형 아파트 모두 2008년의 글로벌 금융위기 때 가장 큰 폭으로 하락했으며, 이명박 정부 동안 줄곧 하락하던 아파트 가격지수가 2013년부터 꾸준하게 상승하고 있음을 알 수 있다.

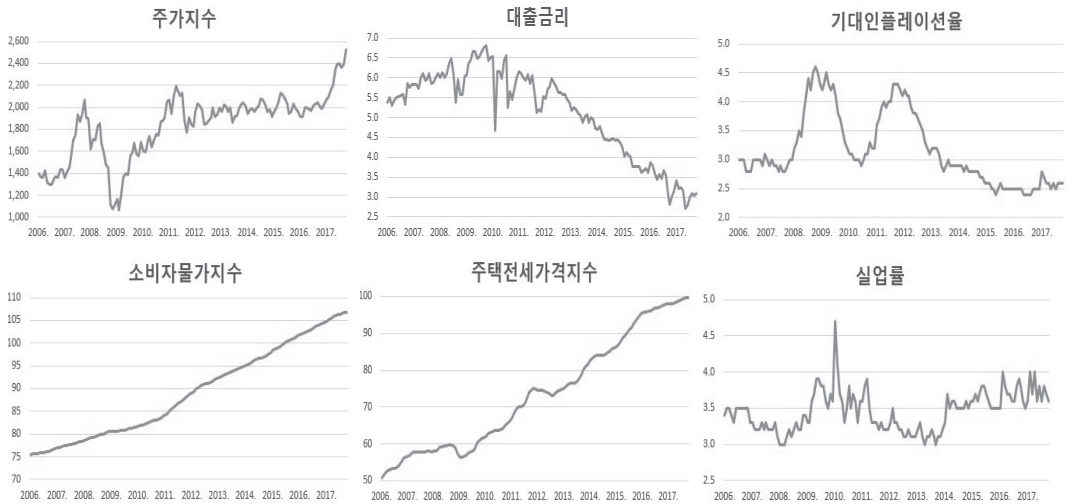


자료: 한국감정원(2017)

〈그림 3〉 서울 아파트 가격지수(2006.1-2017.10)

3.3.2 거시경제지표

〈그림 4〉에서 보는 것처럼 소비자 물가지수와 전세가격지수는 꾸준히 상승하고 있지만, 기대인플레이션율은 2008년 글로벌 금융위기 때와 2011년 말에 큰 폭으로 상승한 후 안정되었다. 주가지수의 경우 2008년 큰 폭으로 하락한 이후 꾸준히 상승하고 있으며, 대출금리는 2009년을 정점으로 지속적으로 하락하고 있다. 실업률은 2008년과 2013년 가장 낮았다가 이후 상승했다.



자료: 한국은행(2017)

〈그림 4〉 거시경제지표(2006.1-2017.10)

4. 실증분석

국내외 연구에서 전통적인 다변량 변수의 예측 모형으로 가장 많이 사용되는 VAR모형과 인공신경망 모형의 예측력을 비교하여 인공신경망 모형의 예측력을 비교하였다.

4.1 VAR 분석

4.1.1 단위근(unit root) 검정

다수의 관찰된 시계열 자료로 통계적 추정을 하기 위해서는 표본이 정상적(stationary)이라는 가정이 필요하다. 정상적이라는 것은 시계열 모형의 확률적 성질이 시간에 따라 변하지 않는다는

것으로 자료의 평균값과 분산이 일정하고, 자료 값들의 차이가 시점과 무관하게 단지 시차에만 의존해야 한다는 의미이다.

본 연구에서는 정상성을 확인하기 위해 ADF(Augmented Dickey-Fuller) 검정법을 사용하여 단위근 검정을 실시하였고, 정상성이 확보되지 않은 경우 차분을 통해 자료의 정상성을 확보하였다. 단위근 검정결과 7개 변수 중 실업률(UNEM)을 제외한 6개 변수의 검정통계량은 각 수준의 임계값보다 모두 크기 때문에 비정상 시계열이라는 것을 알 수 있다. 따라서 각 변수들의 정상성을 확보하기 위해 차분안정화과정(Difference-Stationary Process)을 통해 차분한 후 ADF 검정을 실시한 결과 모든 변수의 검정통계량이 유의수준 1%보다 작아서 정상성을 만족하는 것으로 나타났다.

4.1.2 적정시차 분석

VAR 모형의 적정 시차를 결정하기 위해 최대 시차를 5로 설정한 후 다섯 가지 기준으로 각 시차를 비교하였다. 분석결과 서울 중대형 아파트와 대형 아파트 모두 시차 2에서 3가지(FPE, AIC, HQ)기준에서 최소값이 나왔다.

최종적으로 각 시차별로 잔차들 간의 상관관계가 없는 지를 확인하는 계열 상관성 분석을 위해 Lagrange multiplier(LM) 테스트를 하였다. 계열 상관성 분석 결과 시차 2에서 계열 상관성이 없다는 귀무가설을 채택할 수 있는 것으로 확인되어 VAR 모형의 적정시차를 2로 결정하였다.

4.1.3 VAR 모형을 활용한 예측 결과

2006년 1월부터 2015년 10월까지 118개월 동안의 6개 거시경제지표와 아파트 가격지수를 이용하여 앞에서 선정한 Lag=2, 외생변수=c 로 VAR 모형의 모수를 추정하여 2015년 11월부터 2017년 10월까지 24개월 동안 매달의 서울 중대형 아파트 가격지수와 대형 아파트 가격지수를 계산하고, 계산한 예측값과 실제 관측값을 비교하여 각각의 오차(관측값-예측값)에 대한 평균제곱급오차(RMSE)를 구한 결과 〈표 1〉에서 보는 바와 같이 중대형 아파트의 경우 RMSE가 1.708, 대형 아파트의 경우 RMSE가 2.825로 분석되었다.

〈표 1〉 24개월 동안의 예측값에 대한 RMSE(2015.11~2017.10)

VAR(log)	RMSE	표준편차 (Actual-Predicted)
서울중대형(seoul middle large)	1.708	0.978
서울 대형(seoul large)	2.825	1.151

4.2 DNN 분석

VAR 모델과 마찬가지로 RNN 모델에서도 2006년 1월부터 2015년 10월까지의 118개 데이터를 훈련시켰고, 나머지 24개월(2015년 11월부터 2017년 10월) 데이터를 이용하여 예측력을 테스트하였다. 분석을 위한 RNN은 Keras deep learning 라이브러리의 SimpleRNN(이하 RNN)과 LSTM 네트워크를 이용하였고, 데이터들은 sklearn의 MinMaxScaler를 사용하여 0에서 1 사이의 값으로 정규화(normalization)을 하였다. 활성화함수는 RNN과 LSTM 모두 hyperbolic tangent (tanh)를 사용하였고, 최적화를 위한 프로그램은 adam 알고리즘을 사용하였다.

simple RNN과 LSTM에서 초모수를 찾는 공인된 방법은 없고 다양한 조합을 가지고 여러 번의 시행착오를 거쳐 최적의 초모수를 찾는 것이 일반적인 방법이다. 본 연구에서는 먼저 layer와 lag을 정하기 위해 사전 테스트로서 layer 1에 대해 lag step 1, 2로 각각 Epochs, Batchs, Neurons 값을 테스트하였고, 선정된 lag step에 대해 layer2, layer3에 대해서도 동일한 테스트를 하였다. 최종적으로 선정된 layer와 lag에 대해 Epochs, Batchs, Neurons를 변경하면서 10번 이상의 반복 구동을 통해 모수값을 확정하였다. 확정된 모수값들에 대해 Dropout을 0.0, 0.1, 0.2, 0.3로 테스트 하여 dropout값을 결정하였다. 마지막으로 선정된 모수에 대해 다시 한 번 lag step을 바꾸어 최종적으로 RMSE값을 비교하였다. layer와 lag step은 경우의 수가 많지 않아 비교적 빨리 RMSE가 최소가 되는 모수를 선정할 수 있었고 Epochs, Batchs, Neurons은 경우의 수가 많아 다양한 조합에 대한 반복 구동으로 최적의 모수를 찾았다.[3)]

〈표 2〉는 모형 설계에 대한 내용을 정리한 것이다. 사전 테스트 결과 Layer는 1, lag stesps는 1 또는 2로 결정하였으며, Epochs, Batchs, Neurons은 다음과 같은 과정을 거쳐 최종 모수값을

〈표 2〉 Deep Learning 모형 설계

	Simple RNN	LSTM
Lag steps	1~2	1~2
Hidden layers	1~3	1~3
Epochs	100~2000	400~1500
Neurons	50~150	10~50
Batchs	50~200	90~110
Dropout	0~0.3	0~0.3

3) RNN과 LSTM의 경우 초모수(hyperparameters)를 어떻게 설정하느냐가 모형을 최적화 하는데 가장 중요한 요소이다. 최적의 초모수를 찾는 과정을 Snoek 등(2012)은 '전문가의 경험이 필요한 마술'이라고 했고, 엄지 법칙(rules of thumb), 완전 탐색(Brute-force search) 등 다양한 방법이 제시되고 있지만 대부분의 연구에서는 시행착오를 거쳐 초모수를 찾는 방법이 제시되고 있다. 최근들어 초모수를 찾는 다양한 방법들이 연구되고 있지만 각각의 모수들이 서로 영향을 주기 때문에 모든 영역에서 최적인 모수를 찾는 것은 힘든 일이다(Reimers & Gurevych, 2017).

결정하였고, 마지막으로 테스트한 Dropout 값은 0이다.

4.2.1 서울 중대형 아파트 가격지수

〈표 3〉은 최종적으로 채택된 서울 중대형 아파트의 RNN 및 LSTM 모형에 대한 RMSE 값을 나타내고 있다. RNN 모형은 EPOCH 2000, NEURON 80, BATCH 60, LAYER 1, lags 2, dropout 0.0으로 최종 결정되었고, 이때 RMSE값은 0.839, 표준편차 값은 0.026이다. LSTM의 경우는 EPOCH 410, NEOURON 10, BATCH 90, LAYER 1, lag 1, dropout 0.0으로 결정되었고, 이때 RMSE 값은 0.826, 표준편차는 0.004이다. 전체적으로 LSTM과 RNN의 RMSE값 차이는 크지 않으나 표준편차는 LSTM이 6.5배 우수해서 예측값의 안정성이 높은 것으로 나타났다.

〈표 3〉 서울 중대형 아파트 가격지수 예측에 대한 RMSE

Lags	1	2
Simple RNN	0.931	0.839(0.026)
LSTM	0.826(0.004)	1.693
RNN Epoch 2000, Neuron 80, Batch 60, Layer 1 LSTM Epoch 410, Neuron 10, Batch 90, Layer 1		
10차례 예측값 평균. ( )은 표준편차, VAR : RMSE 1.708		

4.2.2 서울 대형 아파트 가격지수

〈표 4〉는 최종적으로 채택된 서울 대형 아파트의 RNN 및 LSTM 모형에 대한 RMSE 값을 나타내고 있다. 최종적으로 서울 대형 아파트의 RNN 모형은 EPOCH 2000, NEURON 100, BATCH 150, LAYER 1, lags 2, dropout 0.0으로 결정되었고, RMSE값은 1.042, 표준편차는 0.030이다. LSTM의 경우는 EPOCH 800, NEOURON 20, BATCH 100, LAYER 1, lag 1, dropout 0.0으로 결정되었고, RMSE 값은 1.038, 표준편차는 0.003이다. 중대형 아파트의 경우와 마찬가지로 LSTM과 RNN의 RMSE 차이는 크지 않으나 표준편차는 LSTM이 10배 정도

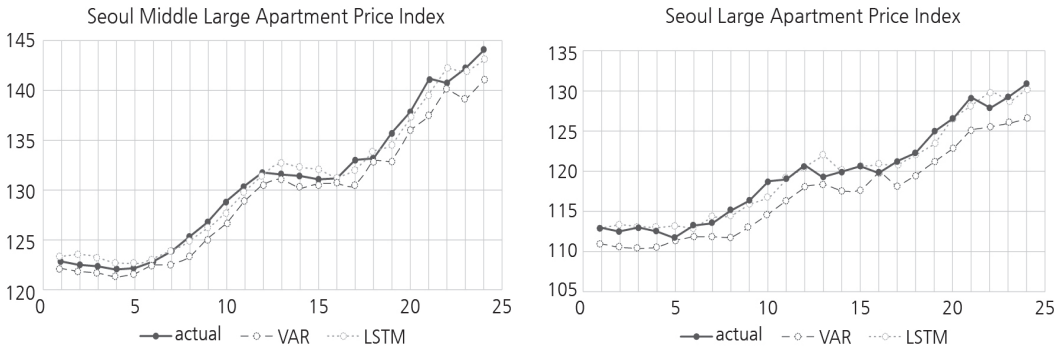
〈표 4〉 서울 대형 아파트 가격지수 예측에 대한 RMSE

Lags	1	2
Simple RNN	2.053	1.042(0.030)
LSTM	1.038(0.003)	1.462
RNN Epoch 2000, Neuron 100, Batch 150, Layer 1 LSTM Epoch 800, Neuron 20, Batch 100, Layer 1		
10차례 예측값 평균. ( )은 표준편차, VAR : RMSE 2.825		

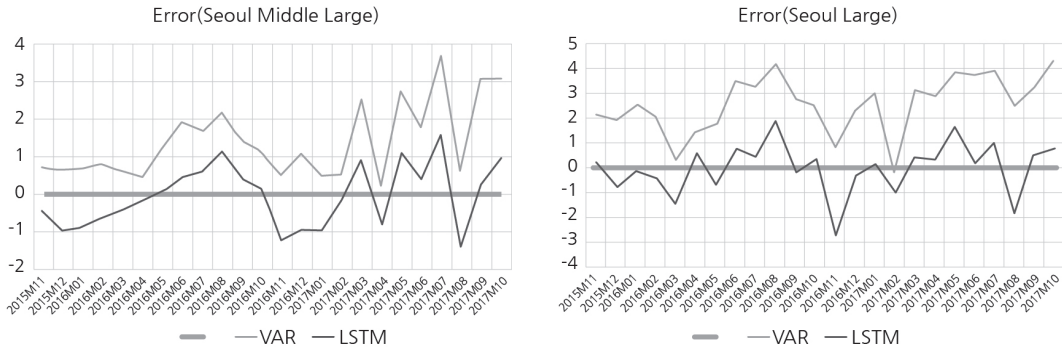
우수해서 중대형 아파트의 경우보다 예측값의 안정성이 더 높은 것으로 나타났다.

4.3.3 VAR모형과의 예측값 비교

〈그림 5〉는 VAR 모형과 LSTM 예측 결과를 실제값과 비교한 것이다. 분석 결과 LSTM 모형이 VAR 모형에 비해 중대형의 경우 52%(0.826 : 1.708), 대형 아파트의 경우 63%(1.038 : 2.825) 평균 제곱근 오차(root mean square error: RMSE)를 줄이는 것으로 분석되었다. 또한 〈그림 6〉은 각 모형의 예측값과 실제값의 차이를 나타내는 오차 값을 그래프로 나타낸 것이다. 중대형의 경우 예측기간의 전반부는 두 모형간 오차가 크지 않지만 후반부로 가면 VAR 모형의 예측 오차가 LSTM보다 훨씬 높은 것으로 나타났다. 반면 대형의 경우는 대부분의 예측 기간 동안 LSTM의 오차가 VAR 모형의 오차보다 현저히 낮은 것으로 나타났다.



〈그림 5〉 VAR와 LSTM의 예측값 비교



〈그림 6〉 VAR와 LSTM의 오차 비교

5. 결론

본 연구는 서울 중대형 아파트와 대형 아파트 가격지수와 6가지 거시경제지표를 이용하여 VAR 모형과 DNN(Simple RNN, LSTM) 모형의 주택가격지수 예측력을 비교하였다. 분석 결과 VAR모형이 LSTM 모형에 비해 중대형의 경우 2.07배(1.708/0.826=2.068), 대형 아파트의 경우 2.72배(2.825/1.038=2.722) RMSE가 더 높은 것으로 분석되었다. 이러한 분석 결과는 인공신경망 알고리즘인 LSTM이 전통적인 다변량 시계열 통계분석인 VAR 모형보다 훨씬 높은 주택가격 예측력을 가지고 있다는 것을 실증적으로 보여주고 있다.

또한 simple RNN과 LSTM의 예측력 차이는 1% 정도로 크지 않지만, 표준편차의 차이는 8~10배 정도로 LSTM이 작았기 때문에 simple RNN에 비해 LSTM 예측치의 재현성이 좋고 안정적이라는 것을 확인할 수 있었다. LSTM과 simple RNN의 예측값 차이가 크지 않은 것은 본 연구의 경우 샘플의 수가 많지 않고 훈련 기간이 상대적으로 짧아 RNN의 시간 역전파 과정에서의 기울기 소멸 문제가 크지 않았기 때문으로 판단된다.

RNN과 LSTM 모형이 기존 시계열 모형보다 높은 예측력을 가지고 있다는 장점에도 불구하고 다음과 같은 모형의 한계를 가지고 있다. 첫째, VAR 모형의 경우 충격반응분석(impulse response analysis)과 분산분해(variance decomposition)분석 등을 통해 투입 변수들 간의 동태적 인과 관계나 전체 변동에 대한 내생변수 변동의 상대적 크기를 측정할 수 있는 반면 인공신경망 알고리즘은 이러한 분석이 어렵다. 이러한 이유로 RNN이나 LSTM 모형을 통해 내생변수들간의 직접적 관계를 파악하기 어렵다는 한계가 있다. 둘째, 인공신경망 모형의 예측력이 은닉층의 개수, 뉴런의 수, dropout 비율, 활성화 함수 선정 등 초모수 값에 민감하게 반응하기 때문에 최적 모형을 찾기 위한 객관적 기준이 없다는 단점이 있다. 결국 초모수 값의 변경을 통한 반복 훈련으로 최적의 모형을 찾아야 하기 때문에 최적 모델을 찾는데 시간이 많이 걸린다는 단점이 있고 최적 모형 선택에 작위성이 존재할 가능성이 높다. 셋째, 미래의 주택가격에 대한 정확한 예측을 위해서는 보다 많은 시계열 자료가 필요한데 우리나라의 경우 2006년부터 월별 주택가격자료가 공개되기 때문에 기계학습을 위한 충분한 자료를 확보하기 어려운 문제가 있다.

본 연구의 결과로부터 다음과 같은 정책적 시사점을 도출할 수 있다. 첫째, 예측력 높은 주택가격 예측 모형이 구축된다면 중앙 및 지방정부가 급격한 미래 주택 시장 변동에 선제적으로 대응할 수 있는 부동산 정책을 개발하는데 이를 유용하게 활용할 수 있다. 이를 통하여 부동산 시장의 안정화를 도모하고 부동산 시장 변동에 대한 대응력이 취약한 저소득층이나 청년층의 주거 안정을 도모하는데 기여할 수 있다. 둘째, 본 연구에서 제안하는 주택가격 예측모형의 경우 다변량 시계열 모형이기 때문에 이를 활용한 다양한 주택 정책 평가 시뮬레이션이 가능하다. 즉, 딥러닝

모형에 활용된 변수의 추가 혹은 삭제에 따른 예측력의 변화를 분석하여 특정 변수가 주택가격지수에 미치는 효과를 파악하는 것이 가능하다.

참고문헌

1. 고종완(2014), "서울시 아파트 매매시장 유형별 가격변동 영향요인 분석", 부동산학보, 58, pp.116~127.
2. 김현재(2011), "주택가격의 변동성 결정요인 분석", 부동산학보, 47, pp.255~269.
3. 문권순(1997), "벡터자기 (VAR)모형의 이해", 통계분석연구, 2(1), pp.23~56.
4. 민성욱(2016), "딥러닝을 이용한 주택가격 예측모형 연구", 강남대학교 대학원 박사학위 논문.
5. 박헌수 · 안지아(2009), "VAR 모형을 이용한 부동산가격 변동요인에 관한 연구", 부동산연구, 19(1), pp. 27~49.
6. 배성완 · 유정석(2017), "딥러닝을 이용한 부동산가격지수 예측", 부동산연구, 27(3), pp.71~86.
7. _____(2018), "머신러닝 방법과 시계열 분석 모형을 이용한 부동산 가격지수 예측", 주택연구, 26(1), pp.107~133.
8. 이석원(2018), "거시경제요인이 아파트가격 변동에 미치는 영향 연구", 목원대학교 박사학위 논문.
9. 이형욱 · 이호병(2009), "서울시 주택가격지수의 모형별 예측력 비교 분석," 부동산학보, 38, pp.215~235.
10. 이훈자(2017), "아파트매매가격지수와 거시경제변수에 관한 시계열 모형 연구", 한국데이터정보과학회지, 28(6), pp.1471~1479.
11. 정원구 · 이상엽(2007), "인공신경망을 이용항 공동주택 가격지수 예측에 관한 연구 -서울지역을 중심으로-", 주택연구, 15(3), pp.39~64.
12. Brownlee, J.(2017), "Long Short-Term Memory Networks With Python", *Machine Learning Mastery* (Ebook Edition: v1.2).
13. Olah, Christopher(2015), "Understanding LSTM Networks", colah's blog, http://colah.github.io/posts/2015-08-Understanding-LSTMs/
14. Fotheringham, A. S., Crespo, R. and Yao, J.(2015), "Exploring, modelling and predicting spatiotemporal variations in house prices", *The Annals of Regional Science,* 54(2), pp. 417~436.
15. Gulli, A. and Pal, Sujit(2017), "Deep Learning with Keras", *Packt Publishing, Birmingham - Mumbai.*
16. Jain, K. and Payal(2011), "A Review Study On Urban Planning & Artificial Intelligence", International Journal of Soft Computing and Engineering (IJSCE), 1(5), pp.101~104.
17. Snoek, Jasper, Larochelle, Hugo, and Adams, Ryan P.(2012), Practical Bayesian Optimization of Machine Learning Algorithms. *Advances in Neural Information Processing Systems,* 25, pp.2951~2959. Curran Associates, Inc.

18. Khalafallah, ahmed(2008), “Neural Network Based Model for Predicting Housing Market Performance”, TSINGHUA SCIENCE AND TECHNOLOGY, 13(S1), ISSN 1007-0214 52/67: pp.325~328.
19. Khamis, Azme Bin and Kamarudin, Nur Khalidah Khalilah Binti(2014), “Comparative Study On Estimate House Price Using Statistical And Neural Network Model”, *INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH,* 3(12), pp.126~131.
20. Nguyen, N. and Cripps, Al(2001), “Predicting housing value: a comparison of multiple regression analysis and artificial neural networks”, *Journal of Real Estate Research*, 22(3).
21. Park, Byeonghwa and Bae, Jae Kwon(2015), “Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data”, *Expert Systems with Applications,* 42, pp.2928~2934.
22. Radzi, Mohamad Shukry b. Mohd, Muthuveerappan, Chitrakala, Kamarudin, Norhaya and Mohmmad, Izran Sarrazin b.(2012), “Forecasting house price index using artificial neural network”, *International Journal of Real Estate Studies*, 7(1).
23. Reimers, Nils and Gurevych, Iryna(2017), Optimal Hyperparameters for Deep LSTM-Networks for Sequence Labeling Tasks, arXiv:1707.06799
24. Sims, Christopher A.(1980), “Macroeconomics and Reality.” *Econometrica,* 48(1)(January), pp.1~48.
25. Kumar, P.(2017), https://www.techleer.com/articles/185-backpropagation-through-time-recurrent-neural-network-training-technique/

논문접수 : 2018.05.04.
1차 심사 : 2018.06.18.
게재확정 : 2018.08.03.