한국경제지리학회지 제25권 제1호 2022(171~181)

머신러닝을 활용한 결측 부동산 매매 지수의 추정에 대한 연구*

김경민** · 김규석*** · 남대식****

A Study on the Index Estimation of Missing Real Estate Transaction Cases Using Machine Learning*

Kyung-Min Kim** · Kyuseok Kim*** · Daisik Nam****

요약: 부동산 시장 분석에 있어 기본이 되는 정량적 데이터는 부동산 가격 지수이다. OECD와 같은 국제기구에서는 국가별 부동산 가격 지수를 공표하고, 한국부동산원에서는 광역시 단위와 시군구 단위의 지수를 산출한다. 그런데 공간단위를 시군구보다 정교한 동단위, 아파트 단지 단위로 설정하는 경우, 여러 문제점을 맞이하게 된다. 대표적인 문제는 결측치이다. 공간적 범위를 좁힐수록 단위 기간에 따라 거래가 적거나 아예 존재하지 않는 경우가 존재하기에 이 경우에는 지수의 산출이 불가능한 결측치가 발생할 수 있다. 본 연구에서는 지도학습 기반의 머신러닝 기법을 활용하여 특정 범위와 기간에 거래가 존재하지 않아 발생할 수 있는 결측치를 보완하는 기법을 제안한다. 본 모형을 통해 부동산 매매 지수의 실제값이 존재하는 것들의 예측을 통해 그 정확도를 검증하고 결측치가 발생한 것들의 예측도 해 볼 수 있었다.

주요어: 부동산 매매가격 지수, 머신러닝, 결측치, MAPE(평균절대비오차), RNN(순환신경망), LSTM(장단기 메모리)

Abstract: The real estate price index plays key roles as quantitative data in real estate market analysis. International organizations including OECD publish the real estate price indexes by country, and the Korea Real Estate Board announces metropolitan-level and municipal-level indexes. However, when the index is set on the smaller spatial unit level than metropolitan and municipal-level, problems occur: missing values. As the spatial scope is narrowed down, there are cases where there are few or no transactions depending on the unit period, which lead index calculation difficult or even impossible. This study suggests a supervised learning-based machine learning model to compensate for missing values that may occur due to no transaction in a specific range and period. The models proposed in our research verify the accuracy of predicting the existing values and missing values.

Key Words : real estate price index, machine learning, missing value, mape(mean absolute percentage error), rnn(recurrent neural network), lstm(long short-term memory)

* 이 논문은 서울대학교 환경계획연구소와 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(2020-0-01389, 인공지능융합연구센터지원(인하대학교)) 의 지원을 받았습니다.

** 제1저자: 서울대학교 환경대학원 환경계획학과 교수 / 환경계획연구소 겸무연구원 (Professor, Department of Environmental Planning, Graduate School of Environmental Studies, Seoul National University / Adjunct Researcher, Environmental Planning Institute, kkim2@snu.ac.kr)

*** 공동저자: 서울대학교 환경대학원 환경계획학과 박사수료 / 한국폴리텍대학 데이터융합SW과 교수 (Ph.D. Candidate, Department of Environmental Planning, Graduate School of Environmental Studies, Seoul National University / Professor, Department of Data Convergence Software, Korea Polytechnics, kyuseokkim@kopo.ac.kr)

**** 교신저자: 인하대학교 아태물류학부 교수 (Professor, Asia Pacific School of Logistics, Inha University, namd@inha.ac.kr)

https://doi.org/10.23841/egsk.2022.25.1.171

1. 서론

주택 유형 중 아파트는 동질적 성격과 이질적 성격을 모두 갖춘 재화라는 특성이 있다. 즉, 거대 아파트 단지의 특정 평형대의 아파트는 동일한 구조를 갖고 있기도 하나, 주변 어메니티는 각 동마다 이질적 특징을 갖게 한다(박주연·박찬일·유선종, 2014). 동일 구조의 동일 평형대 아파트여도 지하철역까지의 거리 혹은 초등학교까지의 거리 등 주변 시설과의 접근성에 따라 가격의 차이가 존재하는 것이다.

또한, 아파트 세대의 평형대별로 참여하는 수요자들의 특징이 다를 수 있다. 예를 들어, 국민주택규모인 85m^2 이하 아파트와 50평형 이상 대형 아파트의 시장 참여자는 다르다.

시장 참여자들은 부동산 매매가격의 흐름을 알고자 하는 경향이 크다. 과거의 부동산 가격 흐름을 파악하여 미래의 부동산 가격을 예측하고, 이를 기초로 부동산 매입을 결정하기 때문이다(박천규·이영, 2010).

과거 부동산 가격 흐름을 파악하기 위해서는 적정한 공간 단위의 부동산 가격지수가 존재해야 한다. 그런데, 현존하는 부동산가격지수는 소비자가 원하는 세밀한 공간 단위의 정보를 제공하지 못하고 있다. 예를 들어, 부동산원과 KB가격지수의 공간단위는 시군구이다. 하지만, 시장참여자들은 서울시 강남구의 가격도 알고 싶지만, 압구정동의 가격 흐름과 특정 단지의 가격 흐름을 알고 싶어한다. 따라서 좀 더 세밀한 차원의 공간 단위와 시간 단위의 부동산 가격 지수가 필요하다.

그러나 공간과 시간 범위를 세밀하게 하는 경우, 기존 방법론에 기초한 가격지수는 제약사항이 존재한다. 공간 단위가 작아질수록 거래 건수는 자연스럽게 상대적으로 줄어들기 때문이다. 심지어 해당 기간 내 특정 지역의 부동산 거래가 존재하지 않을 수도 있다. 이는 해당 기간, 해당 지역의 부동산 매매가격 지수의 결측치(Missing value)를 발생시키는 것이다. 따라서 기존의 회귀분석 방식의 가격지수는 상당한 오차를 보여주거나, 계산을 못하는 경우도 나타날 수 있다.

본 연구는 부동산 매매가격지수 산출에 있어, 거래의 부재로 인해 발생하는 결측치를 예측하는 모형을 제안하고자 한다.

연구 데이터의 시간적 범위는 2006년 1월부터 2021년 12월까지 16년간의 월별 데이터이다. 공간적 범위는 서울특별시 강남 대표 지역구인 강남구 아파트 단지이다. 연구 방법론으로는 지도 학습(Supervised learning) 기반의 RNN(Recurrent Neural Network)와 LSTM(Long Short-Term Memory) 모형을 활용하였으며, 검증을 위해 MAPE(Mean Absolute Percentage Error)를 활용하였다.

본 논문의 구성은 다음과 같다. 2장에서는 본 연구에서 다루는 인공지능 분석 기법에 대한 이론과 부동산 매매 가격과 관련된 연구를 검토한다. 3장에서는 본 연구에서 활용하는 데이터와 연구 방법론인 지도 학습 기반의 RNN, LSTM 분석 기법과 MAPE를 설명한다. 4장에서는 연구 결과를 분석하며, 마지막으로 5장에서는 결론 및 향후 연구 과제를 설명한다.

2. 선행연구

1) 머신러닝

머신러닝은 컴퓨터가 데이터를 학습하고 학습된 데이터를 기반으로 분석하여 판단, 예측하는 기술이다. 머신러닝은 학습 방법에 따라 지도 학습, 비지도 습(Unsupervised Learning), 강화 학습(Reinforcement Learning)으로 나뉜다(Hihn·Braun, 2020).

지도 학습은 입력 값과 출력 값이 존재하는 자료를 학습하여 미래의 데이터 값을 예측하는 것으로서 분류(Classification) 또는 회귀(Regression)분석에 사

용된다. 지도 학습의 대표적 학습 알고리즘은 SVM (Supported Vector Machine), 의사결정나무(Decision Tree), 인공 신경망(Artificial Neural Networks, ANN), 릿지 회귀(Ridge Regression), 라쏘 회귀(Lasso Regression), RNN, LSTM, GRU(Gated Recurrent Unit) 등이 있다(Kumar·Kurmar, 2022; Grinberg·Orhobor, 2019; Khan·Sarfaraz, 2019).

비지도 학습은 출력값을 알 수 없는 데이터를 컴퓨터가 스스로 학습하여 데이터 내부의 패턴과 관계를 찾아내는 것으로 주성분 분석(Principal Component Analysis, PCA), 비음수 행렬 분해(Non-negative Matrix Factorization, NMF), k-평균 군집(k-means), DBSCAN (Density Based Spatial Clustering of Applications with Noise) 등이 있다. 지도 학습과 비지도 학습의 가장 큰 차이점은 결과값이 주어진 데이터를 이용하여 학습하는지 여부이다.

강화 학습은 행동 심리학에서 영감을 받아 고안된 것이다. 현재의 상태를 인식하여 선택 가능한 행동들 중 보상을 최대화 하는 행동 또는 행동 순서를 선택하는 것으로 DQN(Deep Q-Network), A3C(Asynchonous Advantage Actor-Citic) 등이 있다(Duan·Ma·Aggarwal·Sathe, 2020).

시계열 분석에 활용되는 대표적인 인공지능 기반의 방법론으로 RNN과 LSTM이 있다. 그림 1과 같이 RNN은 하나의 입출력 패턴을 가진 DNN(Deep Neural Network) 병렬 체인 구조로 연결되어 있어 과거 학습 결과를 현재 학습에 활용하는 시계열 분석 방법론이다(신동하·최광호·김창복, 2017).

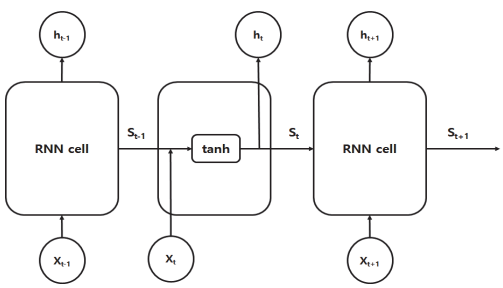


그림 1. RNN 구조

이러한 RNN 방법론은 짧은 시퀀스를 처리할 때 유리하며, 정보와 정보 사이의 거리가 멀어질 경우 학습 능력이 떨어지는 단점이 있다. 이러한 장기 의존성 문제를 해결하기 위하여 그림 2와 같은 구조의 LSTM이 있다. LSTM은 전체의 체인을 관통하여 일종의 컨베이어 벨트 역할을 하는 셀 스테이트(Cell State)를 통해 과거의 학습 결과를 큰 변함없이 전달하는 구조로 되어 있다(신동하·최광호·김창복, 2017).

그림 3은 RNN과 LSTM에서 시계열 예측을 하기 위한 시계열 데이터 구성의 4가지 방법이다. 그림과 같이 One-to-one은 한 가지 종류의 변수로 다음 시점의 변수값을 예측하는 것이고, One-to-many는 한 가지 종류의 변수로 다음 시점의 여러 가지 종류의 변수값을 예측하는 것이다. Many-to-one은 여러 가지 종류의 변수로 다음 시점의 한 가지 종류의 변수값

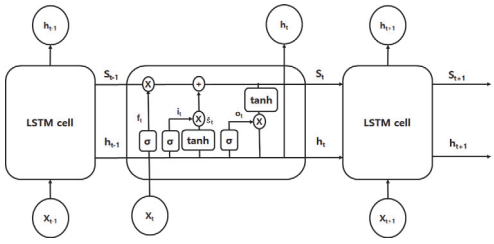
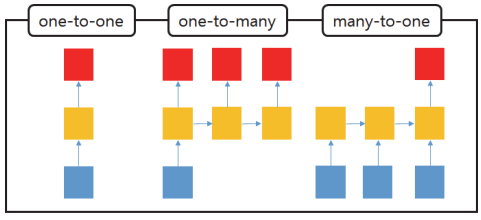


그림 2. LSTM 구조



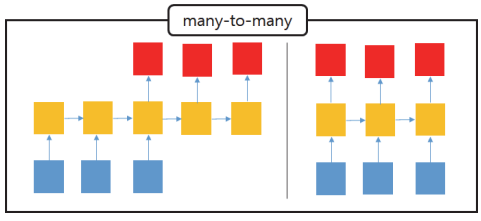


그림 3. 시계열 데이터의 구성

을 예측하는 것이며, Many-to-many는 여러 가지 종류의 변수로 다음 시점의 여러 가지 변수들의 값을 예측하는 것이다. 이때, 입력값이 되는 변수들의 시간적 범위를 Window size라고 한다.

2) 거시경제변수와 부동산 가격 간 연관성에 관한 연구

거시경제변수는 부동산 가격에 큰 영향을 준다. 특히 기준 금리와 장기 국고채 수익률과 물가 상승률의 부동산 가격에 대한 영향도 분석은 상당히 오랜 기간 연구된 분야이다.

Sivitanides et al.(2001)은 상업용과 주거용 부동산 유형의 수익률 결정 요인을 분석하였다. 미국 대도시 수익률 자료를 활용하여, 무위험 이자율과 소비자물가 지수(CPI) 상승률, 실질 임대료 수준과 10년 국채 금리를 포함하는 추정 모델을 구성하였다. 분석결과, 시장별 수익률의 움직임은 지역 임대 성장의 시간 경로와 거시경제변수 (이자율 및 CPI) 의 영향에 의해 영향을 받고 있음을 밝혔다.

Hendershott et al.(2005)은 부동산과 주식 시장 간 관계에 주목하였다. 주식 수익률(배당/가격 비율)이 영국 오피스와 리테일 부동산 수익률에 미치는 영향을 분석하였다.

Hollies(2007)는 JLL 데이터를 사용하여 단기 이자율, 연간 임대 기간, 인플레이션, 시장 투명성 및 유동성과 같은 설명 변수를 포함하는 글로벌 오피스 시장 데이터베이스를 구성하였다. 패널 모형 결과, 유동성은 오피스 수익률(임대수입 / 부동산 가격)을 하락시켜, 분모인 오피스 가격 상승에 일조함을 밝혔다.

Kim et al.(2019)는 2008년 글로벌 금융위기 이후 금융시장의 유동성과 오피스 시장의 관계를 연구하였다. 연구에 활용한 데이터는 아시아 6개국의 2007년 1분기부터 2015년 4분기까지의 분기별 데이터이다. 분석결과, 금리 인상은 오피스 수익률을 높이며, 실질 임대료 증가는 오피스 수익률에 부정적인 영향을 미치는 것으로 나타났다. 또한, M2를 비롯한 유동성 지표가 상승하면, 아시아 6개국의 오피스 수익률 하락(오피스 가격 상승)이 발생하는 것으로 나타났다.

국내의 최근 연구를 살펴보면, 배종찬·정재호(2021)는 거시경제변수와 부동산 정책이 서울 아파트 가격에 미치는 영향을 분석하였다. 2003년 1월부터 2021년 6월까지 한국부동산원의 아파트 매매가격 지수를 활용하였으며, 서울 아파트 매매가격 지수와 통화량, 금리 등의 거시경제 변수는 서로 반대의 흐름을 보이는 것을 나타냈다.

최문기·이성화(2021)는 거시경제변수와 정부의 부동산 대책이 강남 4구 아파트 시장에 미치는 영향을 분석하였다. 연구 결과, 실거래가격지수와 산업생산지수, 주가지수, 주택담보대출금리, 소비자물가상승률과 같은 거시경제변수들과는 시점별로 1개 이상의 변수에 영향을 받는 것으로 나타났다.

임성식(2014)은 정부의 부동산 정책이나 국내외 경기상황과 같은 외부충격요인에 따라 주택가격 지수 예측의 정확도를 비교하는 연구를 진행하였다. 연구 모형으로는 자기회귀오차모형, ARIMA 모형, 개입분석 모형을 활용하여 주택가격지수 예측을 수행하였다. 해당 모형 간 예측력을 비교한 결과 개입분석모형, ARIMA 모형, 자기회귀오차모형 순으로 예측력이 우수한 것을 확인하였다.

손정식·김관영·김용순(2003)은 ARIMA 모형과 VAR 모형을 이용하여 주택매매가격의 변동률, 전세가격의 변동률 그리고 지가 변동률에 대한 예측을 수행하였다. 분석결과, VAR 모형의 예측력이 ARIMA 모형보다 우수함을 알 수 있었다.

3) 머신러닝을 활용한 부동산 가격 예측에 관한 연구

부동산 가격, 지수 산출 및 예측과 관련하여 기존에는 전통적인 회귀분석, 시계열 분석 방법을 활용한

연구들이었으나 최근에는 머신러닝 기법을 활용하여 예측 정확도를 높이는 다수의 연구들이 존재한다.

Chanasit·Chuangsuwanich·Suchato·Punyabukkana(2021)은 미국 메사추세츠 주 보스턴의 주택과 관련하여 미국 인구조사국이 수집한 정보를 활용하여 ANN 기반의 머신러닝을 활용한 부동산 가치 추정을 위한 모형을 제안하였다. 연구결과, 이 연구 모형의 MAE(Mean Absolute Error)는 0.06대였음을 알 수 있었다.

Louati·Lahyani·Aldaej·Aldumaykhi·Otai(2021)는 사우디아라비아 리야드 북부 지역에 위치한 5,946개 토지 데이터를 기반으로 기계학습 기반의 Decision tree와 RF(Random forest) 방법론을 활용하여 가격 예측을 위한 모형을 구성하였다. 연구결과, RF를 활용한 연구 모형의 예측 오차율이 선형 회귀분석이나 Decision tree의 예측 오차율보다 낮았음을 알 수 있었다.

Tchuente·Nyawa(2021)는 Geocoding을 통한 지리적 위치 좌표를 포함하는 것과 하지 않는 것의 예측 정확도 차이에 대해 분석하였다. 연구 데이터는 프랑스 정부에서 제공하는 5년간의 부동산 거래 데이터였으며, 연구 방법론은 RF, AdaBoost, Gradient boosting이었다. 연구결과, Geocoding의 좌표를 포함하는 것이 포함하지 않는 것보다 예측 오차율이 개선됨을 알 수 있었고, 일부는 50% 이상 개선되었음을 알 수 있었다.

Huang(2019)은 미국 캘리포니아주 로스앤젤레스 3개 카운티의 최신 부동산 데이터를 기반으로 Decision tree, Boosting, RF, SVM 등 선형 및 비선형 기계학습 방법을 사용하여 주택 가치를 예측하는 연구를 하였다. 이 중에서 Tree 기반 머신러닝 기법의 MSE(Mean Squared Error)는 데이터양을 늘려감에 따라 0.21부터 그 값이 개선됨을 알 수 있었다.

이태형(2019)은 서울 아파트 가격 지수 예측을 위해 RNN과 LSTM을 통한 예측 가능성을 평가하였다. 2006년 1월부터 2017년 10월까지의 서울 중대형 아파트에 대한 월별 주택가격지수 자료를 수집하고 임대물가 지수, 부채 금리, 주가지수 등의 거시 경제 변수를 수집하여 수행하였다. 분석결과 중대형 아파트 가격 지수는 LSTM의 RSME는 0.826으로 가장 우수함을 알 수 있었다.

배성완·유정석(2018)은 부동산 가격 지수 예측을 위해 SVM, RF(Random Forest), Gradient Boosting Regression Tree, DNN(Deep Neural Network), LSTM, 자기회귀이동평균모형, 벡터자기회귀모형, 베이지언 벡터자기회귀모형을 활용하였다. 연구 결과, 머신러닝의 방법이 시계열 분석 보다 우수한 예측력을 보이는 것으로 나타났다. 또한, 시계열 분석 방법은 구조적인 변화나 외부 충격으로 급변하는 경우 시장 추세를 전혀 예측할 수 없는 것으로 나타났음을 알 수 있었다.

Park·Bae(2015)는 미국 버지니아주 Fairfax 카운티의 5,359개 타운하우스 주택 데이터를 기반으로 Naive Bayesian 및 AdaBoost와 같은 기계학습 방법론을 활용하여 주택 가격을 예측하였다. 실험결과, 이 연구 모형들의 예측 평균 오차율은 0.2대였음을 알 수 있었다.

3) 본 연구의 차별성

본 연구는 다음과 같은 차별성이 존재한다. 부동산 가격 추정에 관한 많은 연구들이 존재하나, 공간과 시간 범위를 좁혔을 때 발생하는 결측치 등에 대한 예외 상황에 대한 대책 연구는 매우 부족하다.

결측치에 대한 예측치가 적정하게 추정되는 경우, 예측치들의 과거 패턴은 사람들이 생각하는 과거 가격 흐름에 크게 벗어나지 않는 모습을 보일 수 있다. 이는 보다 작은 공간 단위의 시계열 트랜드(예를 들어, 특정 아파트 단지별 시계열 트랜드) 구축을 가능하게 할 것이다.

3. 연구 방법

1) 연구데이터

본 연구에서 활용한 데이터의 시간적 범위와 공간적 범위는 표 1과 같다. 시간적 범위는 2006년 1월부터 2021년 12월까지이며, 공간적 범위는 20~30평대(전용면적 49~102m^2) 세대를 포함하고 있는 서울특별시 강남구 소재 337개의 모든 아파트 단지이다.

본 연구에서의 변수들은 선행연구들에서 확인된 것과 같이 CPI, 금리, 주가지수 등과 같은 거시경제 변수와 강남구 아파트 단지별로 구분되는 매매가격 지수, 단지 구분 인덱스를 활용하였다.

표 2는 본 연구에서 활용하는 데이터의 변수별 정의와 출처를 나타낸 것이다. RPI는 부동산 매매가격 지수로 MOLIT의 부동산 매매 실거래가 데이터를 기반으로 식 (1)을 활용하여 자체적으로 산출된 아파트 단지별 매매가격 지수이다. AI는 본 연구에서 활용하는 데이터의 377개의 아파트 단지를 구분하기 위한 값으로 0부터 366까지의 값을 가진다. 그리고 거시경제 변수로서 CPI는 소비자 물가지수, BR는 3년 만기 국채 금리, BIR은 기준 금리, 주가지수로서 KSP는 KOSPI 지수 그리고 KSD는 KOSDAQ 지수이다. 또한, AI 변수를 제외한 나머지는 상대적 크기에 따른 영향을 최소화하기 위하여 0~1로 정규화를 하였다.

표 1. 데이터의 범위

구분	내용
시간적 범위	2006년 1월~2021년 12월
공간적 범위	서울특별시 강남구에 소재한 20~30평대 세대를 포함한 모든 아파트 단지

표 2. 변수 및 출처

변수	내용	출처
RPI	Real Estate Price Index by Apartment Complex	국토교통부
AI	Aptment Index	
CPI	Consumer Price Index	통계청
BR	3-year Bond Rates	한국은행 경제통계시스템
BIR	Base Interest Rate	
KSP	KOSPI Index	통계청
KSD	KOSDAQ Index	

식 (1)은 저자가 기존 연구에서 개발한 아파트 매매가격 지수 모형이다(송의현·김경민, 2019). 이는 아파트 매매가격 지수를 산출함에 있어 해당 세대의 층수와 전용면적 특성을 반영하였다.

$$\ln P = \alpha + \beta_1 X_{Floor} + \beta_2 X_{Area} + \mu_i + \lambda_i + \epsilon \quad (1)$$

P : 아파트 매매가격

X_{Floor} : 층

X_{Area} : 전용면적

2) 연구 방법론

본 연구에서는 RNN과 LSTM을 활용하여 결측 부동산 매매 지수 예측을 하는데 목적을 둔다. 본 연구에서 활용하는 RNN과 LSTM의 기본 모형은 Many-to-one으로 표 2에서의 t-1 시점부터 t-n 시점까지의 변수들을 입력값으로 하여 t 시점에서의 RPI 값을 예측하는 것이다.

이렇게 예측된 값은 식 (2)와 같이 예측 절대 오차율인 MAPE를 통해 모형의 예측 정확도가 산출될 수 있다.

$$MAPE_i = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{A_i - F_i}{A_i}\right| \quad (2)$$

A_i : $Actual\ Value$
F_i : $Forcasted\ Value$

본 연구의 흐름은 다음과 같다. 첫째, 설정된 모형을 기반으로 Epoch, Window size를 다양하게 조합하여 RNN과 LSTM을 수행한다. 이 실험은 다양한 설정 조합으로 수행하여 MAPE가 상대적으로 낮은 최적의

표 3. 최초의 RNN 설정값

변수	값
Batch size	1
Epoch	100 or 200
Window size	1～5

설정값과 모형을 찾는데 목적이 있다.

본 작업을 수행함에 있어 학습 데이터와 검증 데이터는 동일하게 결측치를 제외한 전체 데이터로 한다. 설정값은 표 3과 같이 Batch size는 1로 고정하고 Epoch는 100 또는 200으로 하며, Window size는 1~5로 설정한다. 따라서 설정 조합의 총 개수는 10가지가 된다.

둘째, 앞에서 찾아진 최적의 설정값을 기반으로 20회 반복 수행하여 MAPE 값을 산출한다. 이 실험은 동일한 설정값 기반의 반복 수행을 통하여 실험 오차를 추정하는데 있다.

셋째, 학습 데이터와 검증 데이터를 7:3으로 무작위 추출하여 반복 검증한 후의 MAPE 값을 앞의 결과와 비교한다. 이 실험은 모델이 학습 데이터들에 대해 과하게 학습되는 현상인 Over-fitting과 부족하게 학습되는 현상인 Under-fitting을 피하기 위하여 수행한다.

마지막으로, 결측치가 있는 아파트 단지를 선정하여 시각화한다.

4. 연구 결과

1) 기술 통계량

표 4는 본 연구에서 활용하는 데이터의 기술 통계이다. 해당 단지의 해당 월에 거래가 없어 RPI가 0으로 나타나는 결측치는 총 56,681(78.31%)개이며, 이 데이터를 제외하고 본 연구에서 학습과 검증에 사용하는 데이터는 15,708개였다.

표 4. 결측치를 제외한 후 산출한 기술 통계량 (n=15,708)

변수	평균값	표준편차	최소값	최대값
RPI	153.22	72.88	8.25	1620.11
CPI	1.88	1.22	-0.4	5.9
BY	2.80	1.36	0.83	5.96
BIR	2.26	1.20	0.5	5.25
SKP	1981.77	386.66	1063.03	3296.68
KSD	633.50	134.45	307.48	1038.33

결측치를 제외한 후, 모든 변수들의 표준편차는 평균값보다는 작아 그 값들이 크게 변하여 최소값과 최대값의 범위가 큰 변수는 없는 것으로 파악되었다.

RPI 값의 기준값은 데이터의 시간적 범위의 첫 시점인 2006년 1월이며, 100.0으로 설정하였다. 표 4와 같이 RPI의 최소값은 8.25로 기준 시점대비 92.75% 감소한 가격의 거래도 있었으며, 최대값은 1620.11로 기준 시점대비 1620.11% 증가한 거래도 존재함을 알 수 있었다.

2) 최적의 ANN 설정값 탐색 결과

표 5와 그림 4는 MAPE가 최소값을 보이는 설정값을 찾기 위하여 학습 데이터와 검증 데이터를 동일하게 한 후, Epoch는 100 또는 200, Window size는 1에서 5까지 구성하여 RNN과 LSTM 수행을 한 결과이다.

Epoch가 100일 때, RNN과 LSTM의 수행 후, MAPE 평균값은 각 12.69, 18.54였다. Epoch가 200일 때, RNN 수행 후 MAPE 평균값은 12.13이며, LSTM 수행 후 MAPE 평균값은 16.48이었다. 이를 통해, LSTM 보다 RNN의 예측 오차율이 낮았으며, Epoch의 값이 100 보다 200일 때의 평균 예측 오차율이 낮았음을 알 수 있었다. 이 실험을 통해 MAPE가 가장 낮은 11.16의 값을 보인 Epoch가 200이면서 Window size가 2인 RNN 모형을 최적의 설정값으로 판단하였다.

표 5. Epoch와 Window size 별 RNN, LSTM 수행 후 MAPE 측정 결과

		1	2	3	4	5
RNN	100	13.07	13.07	12.26	12.51	12.56
	200	13.82	11.16	12.98	11.22	11.45
LSTM	100	22.85	16.43	13.62	25.35	14.47
	200	17.94	19.45	14.69	18.69	11.62

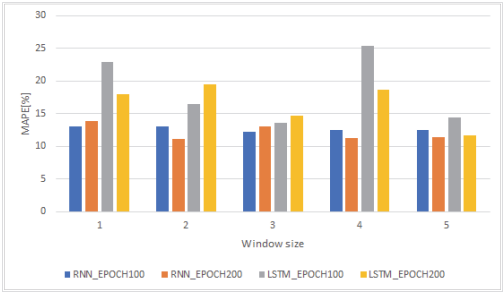


그림 4. Epoch와 Window size 별 RNN, LSTM 수행 후 MAPE 측정 결과

3) 최적의 RNN 설정값으로 반복 수행 결과

본 절에서는 Epoch가 200이면서 Window size가 2의 RNN 모델을 반복 수행한 결과를 설명하고자 한다. 학습 데이터와 실험 데이터를 동일하게 한 경우(Case A)의 MAPE 값과 과적합(Over-fitting)을 고려하여 매번 학습 데이터와 실험 데이터를 임의의 7대 3으로 나눈 경우(Case B)의 MAPE 값을 비교하였다.

표 6은 Case A와 B를 각 20회씩 수행하였을 때의 MAPE 측정값이다. 학습 데이터와 실험 데이터를 동일하게 전체 데이터로 한 Case A의 평균 MAPE 값은 12.12였으나 데이터를 임의로 7대 3으로 나눠 수행한 Case B의 평균 MAPE 값은 39.49로 평균 오차율이 상승하는 것으로 나타났다.

표 6과 같이 Case A와 Case B의 편차가 생기는 이유는 단지별 결측치 개수에 따른 차이가 발생하기 때문이다. 이는 조건이 비슷하나 결측치 개수에 차이가 발생하는 아파트 단지 비교를 통해 알 수 있다.

압구정동 현대5차와 한양1차 아파트 단지 결과값

표 6. Epoch 200, Window size가 2일 때의 RNN의 반복수행 결과

	평균값	표준편차	최소값	최대값
Case A	12.12	0.91	11.28	15.19
Case B	39.49	1.66	36.77	42.84

표 7. 압구정동 현대5차, 한양1차 아파트 단지의 MAPE 비교

	평균값	표준편차	최소값	최대값
현대5차	16.03	14.88	1.78	86.98
한양1차	14.57	13.43	0.15	83.53

을 통해 비교하고자 한다. 이 두 아파트 단지는 모두 압구정동에 위치하고, 두 단지 간의 직선거리는 1km 이내이며, 준공년도는 동일하게 1977년이다. 압구정동 현대5차와 한양1차 아파트 단지의 예측 결과는 표 7과 같다. 현대5차의 MAPE 값이 16.03, 한양1차의 MAPE 값은 14.57로 한양1차 아파트에 대한 예측력이 더 우수함을 알 수 있다. 이는 현대5차와 한양1차 아파트의 존재하는 부동산 매매가격 지수의 데이터의 수가 각 31개, 98개인 것으로 미루어보아 결측치가 적을수록 예측의 정확도가 더 높아지는 것으로 판단된다.

그림 5는 이 두 아파트 단지의 결측치가 많이 존재했던 2018년 4월부터 2019년 3월까지의 실제값과 예측값을 그래프로 시각화한 것이다. 해당 기간, 현대5차 아파트 단지는 2019년 3월에만 데이터가 존재하였고, 한양1차 아파트 단지는 2018년 5월부터 11월까지

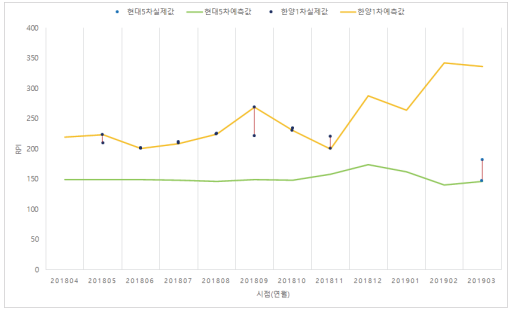


그림 5. 압구정동 현대5차, 한양1차 아파트 단지의 실제값과 예측값

존재하였다. 해당 기간내에 존재하는 값들에 대한 MAPE 값은 평균 7.59였다.

그림 5에서 실제값에 해당하는 선은 회색과 파란색이며 결측치는 끊어져서 표시되었다. 그리고 예측값에 해당하는 선은 빨간색과 노란색으로 결측치까지도 예측을 한 것이다. 약간의 오차가 존재함에도, 실제값과 예측값의 흐름이 비슷하게 진행됨을 알 수 있었다.

5. 결론

본 연구는 부동산 매매가격 지수를 산출함에 있어 거래가 존재하지 않아 발생하는 결측치를 예측하기 위한 인공지능 기반의 방법론을 제안하였다.

연구 데이터의 시간적 범위는 2006년 1월부터 2021년 12월까지이며, 공간적 범위는 20~30평대(전용면적 49~102m^2) 세대를 포함하고 있는 서울특별시 강남구 337개의 모든 아파트 단지이다. 연구 방법론은 머신러닝 기반의 시계열 분석 기법인 RNN과 LSTM을 활용하여 결과 예측 정확도가 높은 최적의 설정값을 찾고 이를 반복 수행하였다.

분석 결과, LSTM 모형에 비해 RNN 모형의 예측 정확도가 더 높았다. RNN이 단기적 기억을 LSTM이 장기적 기억을 고려하는 모형임을 상기할 때, 서울 강남구 부동산 시장 참여자들은 먼 과거 시점의 가격 대신 가까운 과거 시점 가격에 민감한 것으로 보인다.

RNN 모형의 Window size 2가 가장 좋은 모델이라는 부분은 시장 참여자들은 t-2 (두 달 전) 가격에 민감하다는 것이다. 다만, 본 논문은 강남구 20/30평형대 샘플에 기반한 분석 결과이기에, 서울시 혹은 다른 시장에도 일반화하여 설명하기는 어려울 수 있다. 따라서, 부동산 시장 참여자들이 어느 정도 과거 가격에 민감한지의 지역별 차이는 추가적으로 살펴볼 과제이다.

RNN을 20회 반복 수행한 결과, MAPE 평균값은 39.49이었으나, 자료 내부를 자세히 살펴보면 결측치가 많이 존재하는 단지와 적게 존재하는 단지 사이 상당한 편차가 존재했다. 학습 데이터와 검증 데이터를 매번 7대 3의 비율로 임의 배분하더라도 거래의 결측치가 많았던 단지들은 RNN 모델의 학습 가능성과 데이터의 수가 적을 수밖에 없기 때문이다.

본 논문은 결측치 예측을 통해 아주 작은 공간 단위의 가격 흐름의 예측 가능성을 보여주었다. 또한, 시장 참여자들이 가까운 과거 시점 가격에 기반하여 움직인다는 점(Backward-looking)도 확인하였다. 그래서 거래가 이루어진 부동산 실거래 데이터를 인공지능 모듈이 학습하고 이를 기반으로 미래의 가격 지수 예측의 가능성도 보여주었다.

추후 연구에서는 부동산 매매가격 지수에 영향을 미치는 미시적인 변수 추가를 통해 예측 오차를 줄일 수 있을 것으로 판단된다. 또한, 정확도를 더 높이면 결측치 뿐만 아니라 미래의 지수를 예측하는 도구로서도 활용될 수 있을 것으로 기대한다.

참고문헌

국토교통부 실거래가 공개시스템, https://rt.molit.go.kr/, 2021년 2월 1일 최종열람.

박주연·박찬일·유선종, 2014, “아파트 특성이 분양 후 가격 변동에 미치는 요인: 판교 신도시를 대상으로,” 부동산연구, 24(4), pp.25-37.

박천규·이영, 2010, “주택시장 체감지표의 주택시장지표 예측력 분석,” 부동산학연구, 16(1), pp.131-146.

배종찬·정재호, 2021, “거시경제와 부동산정책이 서울 아파트가격에 미치는 영향 연구,” LHI Journal, 12(4), pp.41-59.

배성완·유정석, 2018, “머신 러닝 방법과 시계열 분석 모형을 이용한 부동산 가격지수 예측”, 주택연구, 26(1), pp.107-133.

손정식·김관영·김용순, 2003, “부동산가격 예측 모형에 관한 연구,” 주택연구, 11(1), pp.49-76.

송의현·김경민, 2019, “제2기 수도권신도시 및 주변지역

아파트가격지수 추정," 부동산분석, 5(2), pp.17-41.

신동하·최광호·김창복, 2017, "RNN과 LSTM을 이용한 주가 예측율 향상을 위한 딥러닝 모델," 한국정보기술학회논문지, 15(10), pp.9-16.

이금숙·김경민·송예나, 2010, "복합용도개발과 교통이 아파트가격에 미치는 영향," 한국경제지리학회지, 13(4), pp.515-528.

이우민·김경민·김진석, 2019, "주택정책에 따른 서울 자치구별 주택시장 반응에 대한 연구," 한국경제지리학회지, 22(4), pp.555-575.

전해정, 2012, "유동성 관련 변수가 주택가격에 미치는 영향 및 정책적 시사점에 관한 연구," 한국경제지리학회지, 15(4), pp.585-600.

이태형, 2019, "인공신경망을 활용한 주택가격지수 예측에 관한 연구 : 서울 주택가격지수를 중심으로," 중앙대학교 대학원 박사학위 논문.

임성식, 2014, "주택가격지수 예측모형에 관한 비교연구," 한국데이터정보과학회지, 25(1), pp.65-76.

통계청, https://kostat.go.kr/, 2021년 2월 1일 최종열람.

최문기·이성화, 2021, "거시경제변수 및 부동산 대책이 강남4구 아파트 시장에 미치는 영향," 한국지적학회, 37(3), pp.105-114.

한국은행 경제통계시스템, https://ecos.bok.or.kr/, 2021년 2월 1일 최종열람.

Chanasit, K., Chuangsuwanich, E., Suchato, A. and Punyabukkana, P., 2021, "A Real Estate Valuaton Model Using Boosted Feature Selection," *IEEE Access*, 9, pp.86938-86953.

Duan, L., Ma, S. and Aggarwal, C., and Sathe, S., 2020, "Improving spectral clustering with deep embedding, cluster estimation and metric learning," *Knowledge and Information Systems*, 63(3), pp.675-694.

Grinberg, N. F., Orhobor, O. I. and King, R. D., 2019, "An evaluation of machine-learning for predicting phenotype: studies in yeast, rice, and wheat," *Machine Learning*, 109(2), pp.251-277.

Hihn. H. and Braun, D. A., 2020, "Specialization in Hierachical Learning Systems: A Unified Information-theoretic Approach for Supervised, Unsupervised and Reinforcement Learning," *Neural Processing Letters*, 52(3), pp.2319-2352.

Hendershott, P. H. and MacGregor, B. D., 2005, "Investor rationality: evidence from UK property capitalization rates," *Real Estate Economics*, 33(2), pp.299-322.

Hollies, R., 2007, "International variation in office yields: a panel approach," *Journal of Property Investment and Finance*, 25(4), pp.370-387.

Huang, Y., 2019, "Predicting Home Value in California, United States via Machine Learning Modeling," *Statistics, Optimization & Information Computing*, 7(1), pp.66-74.

Khan, A. and Sarfaraz, A., 2019, "RNN-LSTM-GRU based language transformation," *Soft Computing*, 23(24), pp.13007-13024.

Kim, K., Kim, G. and Tsolacos, S., 2019, "How does liquidity in the financial market affect the real estate market yields?," *Journal of Property Investment & Finance*, 37(10), pp.2-19.

Kumar, N. and Kumar, U., 2022, "Artificial intelligence for classification and regression tree based feature selection method for network intrusion detection system in various telecommunication technologies," *Computational Intelligence*, pp.1-13.

Lauati, A., Lahyani, R., Aldaej, A., Aldumaykhi, A. and Otai, S., 2021, "Price forecasting for real estate using machine learning: A case study on Riyadh city," pp.1-16.

Park, B. and Bae, J. W., 2015, "Using machine learning algorithms for housing pice prediction: The case of Fairfax County, Virginia housing data," *Expert Systems with Applications*, 42(6), pp.2928-2934.

Tchuente, D. and Nyawa, S., 2021, "Real estate price estimation in French cities using geocoding and machine learning," *Annals of Operations Research*, 308(1-2), pp.571-608.

Sivitanides, P., Southard, J., Torto, R. G. and Wheaton, W. C., 2001, "The determinants of appraisal-based capitalization rates," *Real Estate Finance*, 18(2), pp.27-38.

교신: 남대식, 인하대학교 아태물류학부 교수, 이메일: namd@inha.ac.kr
Correspondence: Daisik Nam, Professor, Asia Pacific School of Logistics, Inha University, E-mail: namd@inha.ac.kr

최초투고일 2022년 03월 07일
수 정 일 2022년 03월 11일
최종접수일 2022년 03월 24일