JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지 Vol. 25, No. 12: 1739~1748, Dec. 2021

Autoencoder 기법을 활용한 부동산 가격 이상치 분석

김윤서[1] · 박종찬[1] · 오하영[2*]

Analsis Of Outliers In Real Estate Prices Using Autoencoder

Yoonseo Kim[1] · Jongchan Park[1] · Hayoung Oh[2*]

[1]Undergraduate Student, Department of Business Administration, Sungkyunkwan University, Seoul, 03063 Korea
[1]Undergraduate Student, Department of Economics, Sungkyunkwan University, Seoul, 03063 Korea
[2*]Associate professor, College of Computing and Informatics, Sungkyunkwan University, Seoul, 03063 Korea

요 약

부동산 가격은 국가, 기업, 가계에 영향을 미치며 최근 급등하는 부동산 가격에 부동산 버블에 관한 연구가 많이 시행되고 있다. 하지만 부동산 버블 예측에서 단순히 부동산 가격만을 비교하거나, 부동산 매매에서 핵심적인 심리적 변수를 반영하지 못한다면 버블 예측 모형의 정확성이 떨어진다 판단할 수 있다. 본 연구는 오토인코더 기법을 사용하여 지역별 부동산 버블 상황을 설명할 수 있는 예측 모형을 설계하는 것이 목적이다. 기존의 부동산 버블 분석 연구들이 가격에 영향을 미치는 다양한 종류의 변수를 설정하지 못하였고 주로 선형 모형을 기반으로 연구를 진행했다는 부분에서, 본 연구는 기존 부동산 버블 연구에 사용되지 않았던 기법과 변수들의 도입 가능성을 시사한다.

ABSTRACT

Real estate prices affect countries, businesses, and households, and many studies have been conducted on the real estate bubble in recent soaring real estate prices. However, if the real estate bubble prediction simply compares the real estate price, or if it does not reflect key psychological variables in real estate sales, it can be judged that the accuracy of the bubble prediction model is poor. The purpose of this study is to design a predictive model that can explain the real estate bubble situation by region using the autoencoder technique. Existing real estate bubble analysis studies failed to set various types of variables that affect prices, and most of them were conducted based on linear models. Thus, this study suggests the possibility of introducing techniques and variables that have not been used in existing real estate bubble studies.

키워드 : 부동산, 머신러닝, 오토인코더, 버블, 이상치 선별

Keywords : Real estate, Machine learning, Autoencoder, Bubble, Anomaly detection

Received 16 September 2021, **Revised** 29 September 2021, **Accepted** 11 October 2021

* **Corresponding Author** Hayoung Oh (E-mail: hyoh79@gmail.com, Tel:+82-2-583-8585)
Associate professor, College of Computing and Informatics, Sungkyunkwan University, Seoul, 03063 Korea

Open Access http://doi.org/10.6109/jkiice.2021.25.12.1739 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

부동산은 민법 제99조 1항에 따른 토지 및 정착물을 의미하는 것으로 국가, 기업, 가계가 보유한 자산 중 가장 높은 비율을 차지하고 있다. 재무관리의 관점에서 부동산은 환금성이 낮고 양도세 등의 거래비용이 높으며 가격 변동의 관점에서 주식과의 상관관계가 높아 주식보다 열등한 투자자산으로 분류된다. 이 때문에 해외에서는 투자자산으로서의 인기가 주식과 채권에 밀리지만, 대한민국 부동산의 수익률은 2008년에서 2017년간 주식 수익률의 25.7%p나 상회하였고, 이후 2017년 11월의 매매가를 기준으로 한 '아파트 매매 실거래 가격지수'에서도, 2021년 2월 서울 기준 64.7%p의 상승률을 보이면서 같은 기간 코스피 상승률을 훨씬 뛰어넘는 변동성을 보이고 있다. 이러한 현상의 원인으로는 시중에 풀린 유동성이 가격이 뛰는 부동산으로 몰리고 있기 때문이라 전망되며 한국은행이 지난 6월 발표한 '2021년 4월 통화 및 유동성'의 M1, M2 비율은 역대 최대치로 사상 최대의 유동성이 시장에 유입되었다는 것을 시사하고 있다. 이러한 시장 환경 때문에 전문가들은 현재 부동산 시장의 급변하는 지수에 근거하여 부동산 매입에 적신호를 보내었고, 정부에서도 부동산 관련 규제를 강화하며 선제적인 대응에 힘썼다. 또한 최근 수도권 아파트 가격을 중심으로 부동산 가격이 급등하는 양상을 보여, 부동산 버블이 발생하였을 것이란 우려가 제기되고 있다.

부동산 버블의 예시로는 미국의 서브프라임 모기지 사태(Subprime mortagage crisis)와 일본의 거품 경제(バブル景気) 등을 들 수 있는데, 수요와 공급의 관점에서 가격이 결정되는 경제의 논리에 투자자의 과도한 기대심리에 따른 추가 수요가 종국에는 버블의 발생과 소멸(Boom-bust price cycle)을 야기하여 사회 경제 전반에 부정적인 영향을 미치게 된다. 버블이 가져오는 부정적 효과로는 소득 및 자산의 불균형 확대, 금융 안정성 저하, 소비심리 위축 등이 있으며 그 예시로 거품 경제 이후 붕괴한 부동산 자산 가격으로 인하여 일본은 대규모의 경제 침체를 겪었고 현재까지 그 여파를 수습하고 있다. 이렇듯 부동산 가격은 자산에 있어 국가, 기관, 민간에 걸쳐 가장 중요한 지표2이자 그 자체로 국가의 경제 상황을 대표한다.

버블의 사전적 정의는 시장가치가 합리적으로 기대할 수 있는 예상 소득의 현재가치를 뛰어넘어 거품처럼 팽창하는 현상을 의미한다. 하지만 정책과 연구에서 버블의 정의는 명확한 것에 비해 버블의 기준은 모호한데 이는 같은 자료를 바탕으로도 연구자 간의 의견 충돌을 보이는 것으로도 확인할 수 있다. 버블의 발생은 주로 사후적으로 평가되며 현재 시점에서는 현재 상황이 버블인지 아닌지의 판단이 어렵다. 하지만 LTV, DTI 완화 또는 규제 정책 등 부동산 정책은 현재 시점에서 이루어지고 사후적인 영향을 미치기 때문에 현재 시점에서 버블을 측정하는 것은 매우 중요한 의미를 가진다는 것을 알 수 있다.

현재까지의 부동산 버블 관련 연구는 주로 거시 경제 변수를 바탕으로 현재의 경기를 판단할 수 있는 수식을 제안하는 방식으로 이루어졌다. 하지만 이런 연구 방식은 사용할 수 있는 데이터의 변수가 많지 않고 선형적 모델을 바탕으로 하여, 특정 변수가 측정에 미치는 영향이 과대 해석되는 경향이 있다. 또한 단순한 지표의 양상과 추세만을 보여줄 뿐 현재 경기가 과거 데이터를 기반으로 할 때 버블인지 아닌지를 판단하지는 못하였다. 따라서 본 연구에서는 딥러닝을 활용하여 부동산 가격에 영향을 미치는 다양한 변수들을 설정하고 이를 통해 데이터를 분석한다.

1.1. 선행연구 분석

본 연구를 위한 선행 연구 분석은 다음과 같다.

연구[1]에서는 부동산 가격 분석을 위하여 통화량, 소비자물가지수, 회사채 수익률, 광공업지수를 이용하여 아파트 매매 실거래 가격지수를 예측하였다. [1] 연구는 다양한 머신러닝 기법들의 활용이 부동산 가격지수 예측에 의미가 있음을 증명했다. 실험에 사용한 머신러닝 기법으로는 SVM, RF, GBRT, ARIMA, VAR, 베이지언 VAR 등이 있으며 데이터를 급변기 시장, 안정기 시장으로 구분하여 모델들의 예측력을 비교하였다.

연구[2]에서는 지역별 주택시장의 버블을 추정하였으며 2003년부터 2019년까지의 임대료 / 주택매매가격 비율의 안정성을 Rolling ADF test를 이용하여 판별하였다. 연구에 활용한 변수로는 지역별 전세가격지수와 매매가격지수, CD 금리 등이 있다. 분석 결과 2017년 10월에서 2018년 4월까지의 강남 지역의 버블 발생을 판별하였으며 버블의 크기의 차이는 있으나 버블의 발생이 주기성을 가진다는 것을 확인하였다.

연구[3]에서는 버블을 평가 대상 자산의 시장가격과 발생하는 수익의 현재가치 차이로 정의하였다. 연구의 활용한 변수는 아파트 매매 가격지수와 전세지수이며 전세 지수의 수익 흐름을 매매가격지수로 환원시키고 이를 현금 할인 모형을 통해 버블률을 지역별, 시기별로 측정하고자 하였다. 이를 통해 서울, 인천 지역이 지방 광역시에 비하여 매매가격지수는 낮지만 버블률을 높은 것을 확인하며 매매가격 지수와 버블률이 정비례 관계를 보이지는 않는다는 것을 밝혀내었다.

연구[4]에서는 글로벌 주택 가격의 추이를 분석하여 각국의 부동산 가격의 상승, 하락 요인을 파악하였다. 또한 주요국들의 버블 지수를 나타내었고 이를 실증 분석에서의 버블 유무를 판단하였다. 이를 한국의 주택 가격 분석에 적용하고자 하였으며 주택 가격에 영향을 미치는 요인들을 정리하여 각 요인의 의의를 파악하여 부동산 버블 의사 결정 트리를 제작하여 가격 하락 이벤트가 발생할지, 상승 이벤트가 발생할지 예측하였다.

연구[5]에서는 한국의 부동산 가격의 버블의 존재 유무를 추정하고 경제정책에 대한 시사점을 제시하였다. 이를 위하여 칼만 필터(kalman filter)를 활용하였으며 외환위기 이전과 이후의 한국의 부동산 버블의 비중이 안정화되었는지 판단하였다. 또한 충격 반응함수 분석을 통하여 버블이 이자율과 음의 상관관계이며 가계대출과 양의 상관관계를 이룬다는 것을 밝혀내고 이를 정책 제안에 활용하고자 하였다.

연구[6]에서는 2003년 11월부터 2013년 8월까지 각 지역별 아파트 매매가격의 버블을 금융 위기 전후로 구분하고 칼만 필터와 상태 공간 모형을 이용하여 합리적 버블율을 추정하였다. 전체 기간에서 강남의 합리적 버블율이 25.4%로 가장 높았고 그다음에 강북, 수도권, 전국, 비수도권이 뒤를 이었다.

연구[7]에서는 부동산은 크기의 방대함으로 다양한 변수에 영향을 받는다는 것을 가정하고 연구를 진행하였다. 연구에서는 아파트 실거래가격을 활용하여 합리적인 아파트 가격 지수를 산출하고자 하였으며 이를 위해 트리 회귀 모형 주성분 회귀 모형, 그라디언트 부스팅 모형을 활용하여 다양한 내, 외생 변수를 분석하였다. 그 밖에 공급면적, 아파트의 층수, 전용면적, 방수, 완공 연도, 주차 공간, 화장실 수 등의 내생 변수와 지하철, 쇼핑시설, 병원, 학군 등의 10가지 외생 변수를 추가로 설정하여 아파트 가격지수를 산정하고 검토하여 기존 KAB 지수의 매매 모형의 한계점으로 지적되던 편의성 문제를 해결하고자 하였다. 하지만 주택시장에서 투기적 행태가 나타나는 것과 그것이 가지는 경제적 함의를 고려했음에도 불구하고 주식시장과 거시경제 변수를 외적 요인으로 고려하여 계량화하지 않은 것은 한계점으로 보인다.

연구[8]에서는 베이지안 네트워크를 활용하여 기존 연구에서 덜 다루어졌던 상업용 부동산을 연구하고자 하였다. 종속 변수로는 상가 낙찰가율을 설명변수로는 낙찰률, 산업생산지수, 상업용 건축물, 회사채 수익률, 경제심리지수, 착공현황으로 하여 수도권의 상업용 부동산을 다중 회귀로 분석하였다. 변수 간의 상관관계를 분석하기 위하여 베이지안 네트워크를 사용하였으며 이를 통해 수도권의 지역별 상가 낙찰가율을 각 설명변수들을 통해 설명하였다. 또한 설명 변수 간의 인과성과 시계열적 선행성을 나타내었다. 하지만 다양한 모형을 통해 예측력을 계량화하지 않았으며 변수를 거시경제 변수로 한정함으로 상업용 부동산에서 중요한 부동산 요인을 변수로 제시하여 측정하지 못했기 때문에 한계성을 지닌다.

결과, 본 연구에서는 아파트실거래가 지수를 포함한 주택시장을 대표하는 변수를 대상으로 연구를 진행했다. 본 연구에서 관심 있는 아파트 실거래가 지수를 선형회귀모형, SVM(Support Vector Machine), RF(Random Forest), 인공신경망을 통해 분석하였으며 모형 별 변수는 상이하나 인공신경망에서는 선행연구중 가장 다양한 65가지 변수를 활용하여 이를 분석하고자 했다. 또한 경기변동기의 시장에서의 모델 적합성 또한 고려대상으로 삼았고 변수의 형태(이산형, 연속형)와 다양성을 모형에 적용시키기 위한 방법론을 구상했기 때문에 의의를 지닌다. 허나 모형의 과적합 문제와 부동산 시장의 지역적 요인을 고려하지 못하였다는 것은 한계를 가진다.

Ⅱ. 분석 알고리즘

2.1. 이상치 선별

이상치(outlier)란 데이터의 대다수의 관측치의 정상 범주에 유의미하게 벗어난 값을 의미한다. 대부분의 실증 연구에서 이상치는 전처리 과정에서 제거되는 것이 일반적이며 머신러닝 모형의 성능을 저하시킬 수 있기

때문에 이를 판별하는 것이 중요하다. 하지만 이상치를 전처리하여 제거하는 것뿐만이 아니라, 이상치 탐지 자체를 목적으로 하는 경우도 존재하며, 이러한 사례에서 이상치는 단순히 오류나 우연으로 생성된 데이터가 아닌 정상 범주의 데이터와 다른 형성 요인을 가지고 만들어진 것으로 판단한다.

위와 같은 이유로 이상치 선별은 다양한 분야에서 연구되고 있고, 이상치를 생성하는 변수에 대한 연구도 시도되고 있다. 또한 이상치를 분석하는 기법으로 최근 머신러닝이 각광받고 있는데, 기존에 분석하지 못하였던 다양한 변수의 데이터를 단시간에 분석하여, 일부 분야에서는 기존의 방식에 비해 탁월한 성과를 내고 있기 때문이다.

본 연구에서는 부동산 버블을 이상치로 규정하고 이를 부동산 가격에 영향을 미치는 변수들을 통하여 정상 데이터와 분류하는 것에 목적이 있다, 이상치 선별 방법으로는 차원 축소(Dimensionality reduction)를 사용하였다, 차원 축소는 SVD와 PCA 등으로 세분화되며 주로 데이터에서 핵심적인 정보를 추출하여 이를 바탕으로 데이터를 재구성하는 것에 기반한다. 차원 축소에서 이상치 판단은 최초 주어진 데이터와 재구성된 데이터를 분석하여 얼마나 유의미한 차이를 보이는지 확인하는 것이다. 만약 반복적으로 발생하는 정상 데이터로 학습된 머신러닝에 정상 데이터를 입력하면 재구성된 데이터의 차이가 크지 않겠지만, 이상치를 입력한다면 기존에 학습된 데이터와 다른 양상을 보이기 때문에 정상 데이터에 비하여 큰 차이가 발생할 것이다. 여기서 차이는 재구성 오류(Reconstruction error)로 정의되며 주어진 데이터의 재구성 오류가 일정한 임계치 이상의 값을 가진다면 이를 이상치로 판단할 수 있다.

2.2. 분석 알고리즘

이러한 재구성 오류를 분석하기 위하여 본 연구에서는 신경망 모형을 사용한다. 신경망 모형이란 입력층(Input layer)과 은닉층(Hidden layer), 출력층(Output layer)으로 이루어져 있으며 입력층에 주어진 데이터를 은닉층을 통해 학습하고 최종적으로 출력층을 통하여 산출하는 구조이다. 신경망 모형은 크게 지도학습(Supervised learning)과 비지도 학습(Unsupervised learning)으로 나눌 수 있는데 대표적인 비지도 학습 방법으로 오토인코더(Autoencoder)를 들 수 있다.

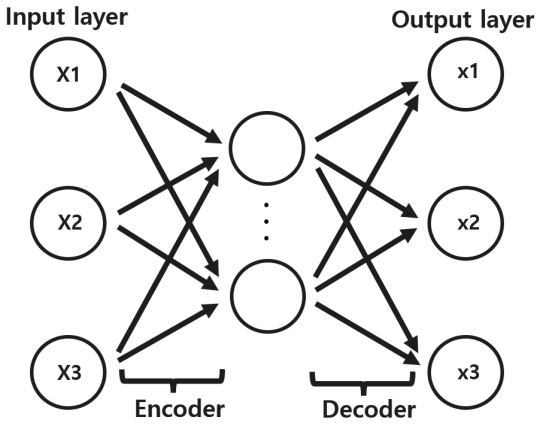

Fig. 1 Structure of Autoencoder

오토인코더는 위의 <그림.1> 과 같이 input 데이터를 내부 표현으로 변환하는 Encoder 부분하고, 내부 표현을 출력으로 변환하는 Decoder 부분으로 이루어져 있다. 오토인코더의 특징은 입력과 출력층의 뉴런의 수가 동일하며, 히든 레이어의 뉴런이 입력층과 출력층 보다 적으므로 input 데이터가 저차원으로 표현된다는 것이다. 데이터를 다시 출력하기 위하여 오토인코더는 input data의 특성을 학습하게 된다. 이러한 데이터의 가장 중요한 특성을 feature라고 정의한다. 오토인코더는 feature를 이용하여 input 데이터를 재구성하기 때문에 오토인코더의 출력은 재구성(Reconstruction)이라 정의한다. 오토인코더는 재구성 과정에서 파라미터 (Θ,ϕ)를 최소화하는 방향으로 구성하는데, 이를 위해 Encoder에 입력되는 input 데이터와 이를 Decoder 통하여 재구성된 데이터가 같아지도록 학습한다.

오토인코더는 데이터를 축약하고 이를 재구성하여 최초의 데이터와 유사하게 복원하는 것에서 다른 차원 축소 기법들보다 우수한 성능을 발휘한다. 이를 통하여 입력된 데이터 중 복원에 애로 사항이 있는 데이터는 독자적인 특징을 지니는 이상치로 분류할 수 있다. 오토인코더를 사용한다면 기존의 선형적인 함수식을 이용한 연구에서 벗어나 비선형적인 형태로 주성분 분석을 이루어 낼 수 있으며, 이러한 점을 이유로 본 연구에서는 오토인코더를 활용 머신러닝 모델로 선정하였다.

Ⅲ. 분석 데이터 및 분석 방법

3.1. 부동산 가격 관련 변수

본 연구에서는 머신러닝 학습을 위하여 2011년 7월부터 2021년 3월까지의 서울, 경기, 인천, 부산, 대구, 광주, 대전, 울산, 강원, 충북, 충남, 전북, 전남, 경북, 경남의 월별 부동산 관련 변수 데이터를 활용하였다. 주요 시도 중 세종시와 제주도의 경우에는 몇몇 데이터에 결측치가 존재하여 연구에 포함하지 않았다. 수집한 통계 데이터의 분류는 한국토지주택공사 SEE:REAL의 부동산 통계 중 주요 통계에 활용되고 있는 분류를 참고하였으며 각 <표>는 학습에 사용된 변수의 종류이다. <표.1>은 아파트와 주택의 매매가와 전세 가격에 관련한 지표이며 부동산 버블 연구에서 주로 분석하는 변수이고 변수의 추이를 통해 버블을 판단할 수 있기 때문에 본 연구의 데이터로 포함시켰다. <표.2>는 건축과 관련한 지표로 주택 공급과 관련성을 가지며 주택 공급과 부동산 버블의 관계성을 파악하기 위하여 데이터에 포함시켰다. <표.3>과 <표.4>는 토지와 관련한 변수이며 기존 연구에서 고려되지 않았던 주택 가격과 토지의 연관성을 파악하기 위하여 데이터에 포함시켰다. <표.5>는 대출과 관련한 변수로, 주로 부동산을 매입할 때 대출을 이용하여 구입하는 대한민국 시장의 특성을 반영하기 위하여 데이터에 포함시켰다. <표.6>은 금융 및 물가 지수로, 부동산 가격에 영향을 미치는 시장 경기를 반영하기 위하여 데이터에 포함시켰다. <표.7>은 인구 관련 변수로 특정 지역의 인구수가 지역 부동산 가격에 영향을 미칠 수 있음을 반영하여 데이터로 포함시켰다. <표.8>은 경제지표 관련 변수로 취업률과 실업률을 통하여 지역의 경제 활성화 정도를 반영하고자 하였다. <표.9>는 주택 거래 관련 변수로 아파트 매매 거래량을 반영하였으며, 주택 매매 거래량은 조사기간에 결측치가 존재하여 반영하지 않았다. <표.10>은 부동산 심리 관련 변수로 거시 경제 변수만이 아니라 부동산 가격에 영향을 미치는 구매자의 심리적 요인을 계량화하고 이를 머신러닝 학습에 반영하기 위하여 관련 지수를 설정하여 데이터로 구성하였다.

Table. 1 Variables related to housing prices

Variable	Source
Apartment Sales Price Index	KOSIS
Apartment Jeonse Price Index	KOSIS
house sale price index	KOSIS
Housing Jeonse Price Index	KOSIS

Table. 2 Variables related to construction

Variable	Source
Construction start area	KOSIS
Building permit area increase/ decrease rate	KOSIS

Table. 3 Variables related to land transactions

Variable	Source
land transaction completion index	Korea Real Estate Agency Real Estate Statistics

Table. 4 Variables related to land supply

Variable	Source
Land price change rate	Korea Real Estate Agency Real Estate Statistics

Table.5 Variables related to loan

Variable	Source
mortgage loan	Bank of Korea economic statistics system
household loan	KOSIS

Table. 6 Variables related to financial and price

Variable	Source
base rate	Bank of Korea economic statistics system
KTB (3 years)	Bank of Korea economic statistics system
Corporate bonds (3 years, AA-)	Bank of Korea economic statistics system
consumer price index	Statistics Korea National Statistics Portal

Table. 7 Variables related to population status

Variable	Source
Resident Registration Population	Statistics Korea National Statistics Portal

Table. 8 Variables related to economic indicators

Variable	Source
employment rate	KOSIS
unemployment rate	KOSIS

Table. 9 Variables related to housing transactions

Variable	Source
Apartment sales volume	Korea Real Estate Agency Real Estate Statistics

Table. 10 Variables related to real estate psychology

Variable	Source
Real estate market consumer sentiment index	Statistics Korea National Statistics Portal
PIR	Housing Finance Statistical System
LIR	Housing Finance Statistical System

위 데이터의 열은 총 21개의 변수로 구성되어 있으며, 행은 세종시와 제주도를 제외한 각 시도 2011년 7월에서 2021년 3월의 총 1,755개의 월별 데이터로 구성되어 있다. 본 데이터는 LH의 SEE:REAL 주요 통계표에 포함되지 않은 분류인 '부동산 심리 관련 변수'를 설정하여 만들어졌으며 거시적 경제 변수뿐만이 아니라 주택 가격에 영향을 줄 수 있는 다양한 변수들을 다각도에서 분석하고자 하였고 모든 데이터는 결측치가 없이 정리하여 구성하였다. 재구성 오류 분석을 위한 데이터 학습에 앞서, 정상적인 부동산 가격 데이터와 버블이 발생한 데이터를 분류하는 작업이 필요하다. 버블을 정의하는 방법은 연구마다 차이가 있고 이를 나타내고 버블을 판단하는 방법에도 의견이 갈리지만, 통상적으로 버블이란 소비자의 기대 수준보다 높은 가격에 형성되어 정상적인 시장 양상을 보이지 않고 투기를 유발하는 상태를 의미한다. 따라서 본 연구에서는 정상 데이터와 버블 데이터를 월별로 나누기 위하여 '주택 구입 부담 지수'를 활용하였다. 주택 구입 부담지수는 한국주택금융공사 주택 금융 통계 시스템에서 제공하고 있다. 주택 구입 부담 지수의 사전적 의미는 중간 소득 가구가 표준 대출을 받아 중간 가격 주택을 구입하는 경우의 상황 부담을 나타내는 지표로 대출 상환 가능 소득(원리금 상환액 / DTI)을 중간가구 소득으로 나눈 것이며 지수가 높을수록 주택 구입 부담이 가중됨을 의미한다. 이는 평균적으로 대출을 받고 집을 매입하는 한국 부동산 시장의 특성을 반영하였고 본 연구에서는 이를 절대적인 수치로 판단하기보다는 지수의 변동률을 기준으로 버블 유무를 판단하였으며 이를 바탕으로 버블에 대한 사후적 평가를 내려 학습을 진행하였다.

2011년 7월에서 2021년 3월까지의 전국 주택 구입 부담 지수 평균은 56.6이며, 서울의 경우 평균 111.09, 최댓값 166.2로 타지역에 비해 가장 높은 수치를 보였다. 가장 낮은 수치를 기록한 지역은 전남이며 평균 31.0, 최댓값 31.3를 기록하였다. 또한 지수의 증가율에서도 서울은 가장 높은 수치를 보였는데 20년 1분기 기준, 21년 1분기는 20%가 넘는 상승률을 보이며 현재 부동산 상황이 버블의 양상을 보일 수 있음을 시사하였다. 본 연구에서는 부동산 데이터 학습을 위하여 부동산 부담 지수가 전체 기간 평균을 초과한 60을 넘으며, 2020년 1분기보다 현재 10% 이상 증가한 지역을 제외하고, 분기 간 지수의 평균 변화율을 초과하지 않은 지역을 대상으로 학습 데이터를 구성하였다. 구성된 데이터를 <그림.2>와 같이 그래프로 나타내 보면 서울, 부산, 대구, 대전, 경기도를 제외한 나머지 시도로 정상 부동산 주기 및 데이터를 구성한 것을 확인할 수 있으며, 구체적으로 모형의 성능을 검증하기 위하여 데이터를 학습에 필요한 훈련 데이터와 검증 데이터로 분류하였다. 본 연구에서는 데이터의 30%를 검증 데이터로 설정하였다. 오토인코더에 학습되는 데이터는 전부 정상 데이터로 구성되며, 그렇기 때문에 오토인코더는 이상치 자료를 학습한 경험이 없다. 이렇듯 소수의 이상치를 훈련 데이터로 분배하지 않고 정상 데이터로만 학습 시키는 것을 단일 범주 학습이라 하며, 이런 방식으로 학습된 모형은 이후에 이상치가 입력되었을 경우 복구에 정상 데이터보다 많은 오류를 유발한다.

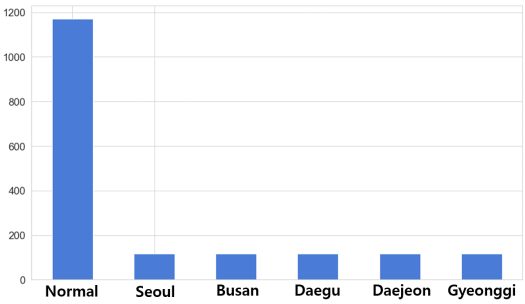

Fig. 2 The number of normal data and bubble data

3.2. 오토인코더 모델

오토인코더 학습을 위한 데이터 분류 후 정상 가격 데이터는 0의 Class를, 나머지는 지역별로 1~5의 Class를 부여하였다. 학습은 Python3으로 <그림.3> 의 코드를

Algorithm 1: Data setting

```
X_train, X_test = train_test_split(data, test_size=0.3,
random_state=RANDOM_SEED)
X_train = X_train[X_train.Class == 0]
X_train = X_train.drop(['Class'], axis=1)

y_test = X_test['Class']
X_test = X_test.drop(['Class'], axis=1)

X_train = X_train.values
X_test = X_test.values

X_train.shape
```

Fig. 3 Data setting

활용하여 진행하였다. Train data의 shape는 (804, 21)로 구성되어 있고, 행은 월별 분류, 열은 변수별 분류이다. 원자료 데이터를 그대로 학습에 사용하기에는 데이터 간 단위가 달라, 모든 변수별 데이터는 MinMax 정규화를 통해 0에서 1사이의 값을 가지며, 오토인코더 encoder와 decoder의 activation function은 <그림.4>와 같이 'relu'와 'tanh'를 사용하였다. 오토인코더는 차원 축소와 복원 과정에서 모델의 파라미터들을 오류가 최소화되도록 학습시키는데 이때 오류를 판단하고 이를 계량화하는 방법으로는 MSE(Mean squred error)을 사용하였다. 옵티마이저로는 'adam' 사용하였으며 정규화 방법으로는 'L1 정규화'를 사용하였다. <그림.5>의

Algorithm 2: Autoencoder

```
input_dim = X_train.shape[1]
encoding_dim = 14

input_layer = Input(shape=(input_dim, ))

encoder =
Dense(encoding_dim, activation="tanh",
      activity_regularizer=regularizers.l1(10e-5))(input_layer)

encoder =
Dense(int(encoding_dim / 2),activation="relu")(encoder)

decoder =
Dense(int(encoding_dim / 2),activation='tanh')(encoder)

decoder = Dense(input_dim, activation='relu')(decoder)

autoencoder = Model(inputs=input_layer, outputs=decoder)
```

Fig. 4 Autoencoder

'ModelCheckpoint'를 이용하여 가장 우수한 성능을 가진 오토인코더 모델을 파일로 저장할 수 있다. 분석을 위해 구상한 오토인코더는 4개의 fully connected layer로 구성되어 있으며, 최초 2개의 layer는 encoder로 나머지 2개의 layer는 decoder로 구성되어 있다.

Algorithm 3: Autoencoder Training

```
nb_epoch = 100
batch_size = 32
autoencoder.compile(optimizer='adam',
                    loss='mean_squared_error',
                    metrics=['accuracy'])
checkpointer = ModelCheckpoint(filepath="model.h5",
                               verbose=0,
                               save_best_only=True)
tensorboard = TensorBoard(log_dir='./logs',
                          histogram_freq=0,
                          write_graph=True,
                          write_images=True)
history = autoencoder.fit(X_train, X_train,
                    epochs=nb_epoch,
                    batch_size=batch_size,
                    shuffle=True,
                    validation_data=(X_test, X_test),
                    verbose=1,
                    callbacks=[checkpointer, tensorboard]).history
```

Fig.5 Autoencoder Training

Ⅳ. 실험 및 결과

4.1. 모델 학습 데이터

신경망의 학습에서는 적절한 epoch(학습 횟수)와 batch(학습 단위)를 구성하는 것이 중요하다. epoch란 전체 훈련 세트가 신경망을 통과한 수를 의미하고 오토인코더에서는 encoder와 decoder의 hidden layer의 수를 의미한다. batch는 위의 학습에서 모델의 가중치를 업데이트할 경우 사용되는 묶음에 포함된 요소의 개수를 의미한다. epoch은 overfitting(과다 적합)을 유발할 수 있으며, 이 경우 학습 완료 후 새로 입력한 신규 데이터에는 그 성능을 발휘하지 못하는 문제가 있다. 또한 너무 큰 batch는 메모리 부족 문제를 야기할 수 있고, 반대로 너무 작은 batch는 너무 적은 샘플을 고려하여 가중치를 업데이트하기 때문에 업데이트가 자주 발생하고 훈련이 불안정하게 흔들리는 양상을 보일 수 있다. 따라서 본 연구에서는 이를 변화시키며 <그림.6> 과 같이 MSE가 수렴하는 추세를 확인하였으며, 가장 안정적으로 MSE가 수렴하는 지점을 epoch 100으로 확인하였다.

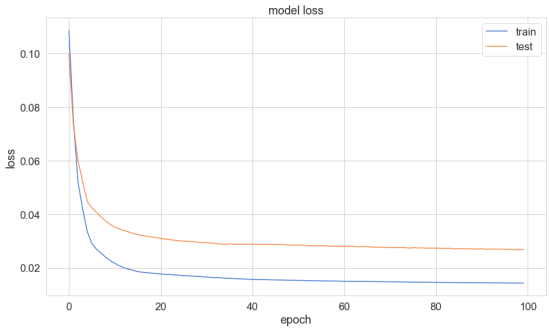

Fig. 6 Reconfiguration error per epoch

Train data로 오토인코더를 수행하고, 이후 Test 데이터로 재구성 오류를 분석해본 결과, <그림.7> 와 같이 정상 부동산 가격 데이터는 아래와 같이 최대 0.04의 재구성 오류를 기록하고 0.005에서 0.030 사이에 대부분 분포하는 것을 확인할 수 있다.

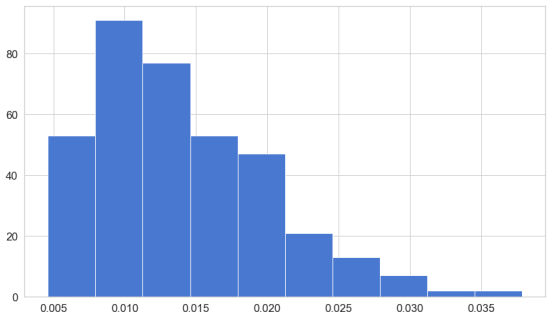

Fig. 7 Distribution of reconstruction error of normal price data.

이와는 달리 버블이 의심되는 서울 지역의 데이터의 경우 <그림.8>에서 확인할 수 있듯이 MSE가 0.015에서 0.035 사이에 분포가 되어 있는 것을 볼 수 있으며, 이는 기존 정상 가격 데이터와 버블 의심 지역의 데이터가 부동산 변수의 크기와 관계에 있어 차이를 보인다는 것을 의미한다.

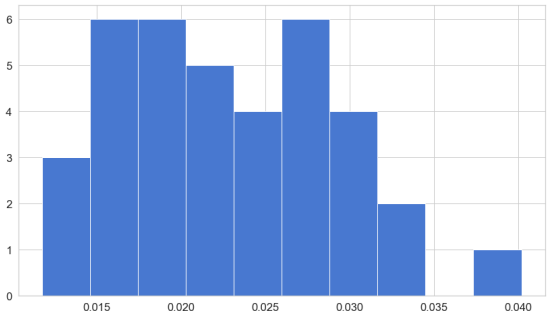

Fig. 8 Distribution of reconstruction error of bubble data.

이를 산점도로 구성해보면 <그림.9>과 같이 나타낼 수 있는데 세로축은 재구성 오류, 가로축은 월별 데이터로 기록한 것이다. 가로축 1번부터 117번까지는 서울시의 데이터를 2011년 7월에서 2021년 3월까지 월별로 순차적으로 정리한 것이며, 그 후 117개씩 순서대로 부산광역시, 대구광역시, 인천광역시, 광주광역시, 대전광역시, 울산광역시, 경기도, 강원도, 충청북도, 충청남도, 전라북도, 전라남도, 경상북도, 경상남도의 데이터이다. 재구성 오류의 분포로 서울과 경기도가 다른 지역들에 비해 확연하게 큰 편차를 보이고 있으며, 이는 서울과 경기도의 부동산 자산이 다른 지역에 비해 버블 가능성이 높게 평가되고 있음을 의미한다.

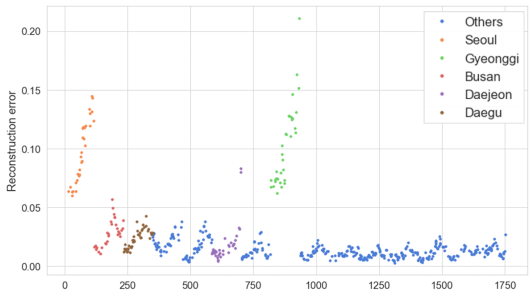

Fig. 9 MSE distribution by region

또한 두 지역 모두 시간이 흐를수록 재구성 오류가 상승하는 것을 확인할 수 있는데 이는 지난 기간 동안 두 지역의 버블 가능성이 지속적으로 증가하는 경향성을 보이고 있음을 시사한다. 또한 MSE가 0.060를 넘는 지역으로 서울, 경기도를 제외하고도 대전 지역을 꼽을 수 있으며, 대전의 경우 2010년대 초반에는 다른 정상 부동산 가격 지역과 유사한 MSE를 보이다가, 지속적으로 버블 가능성이 증가하여 2021년도에 이르러서는 버블로 판단할 수 있을 만큼 높은 증가율과 MSE를 보였다. 다음으로 높은 MSE를 기록한 지역으로는 부산, 대구, 광주, 인천이 있으며 모든 지역에서 시계열적으로 MSE가 증가하는 양상을 보이고 있다.

특히 광주와 인천의 경우 데이터 학습 과정에서 정상 부동산 가격 데이터로 분류하였지만 재구성 과정에서 높은 MSE를 보였다. 인천의 경우 평균 주택 구입 부담 지수가 58.2로 60에 미치지 못하였지만, 2020년 1분기와 2021년 1분기의 지수를 비교하였을 때 14.1%p 가 증

가하며 다른 정상 부동산 가격 지역에 비하여 높은 증가율을 보였다. 따라서 절대적인 부동산 가격의 부담은 적더라도 가격의 상승률이 높다면 부동산 버블 위험성이 높음을 판단할 수 있다. 또한 조사 첫 시점부터 높은 MSE를 보였던 서울과 경기와 달리 부산, 대구, 인천, 광주, 대전, 울산은 시계열적으로 지속적인 증가를 보였으며 후에는 타지역에 비해 높은 MSE를 기록하였다.

현재 국내에서는 정부 기관에서 부동산 버블 관련 지수를 발표하고 있지 않다. 그렇기 때문에 본 연구의 결과가 신뢰성을 가지는지를 판단하기 위하여 국토연구원에서 2021년 발표한 '국토이슈리포트'에 수록되어 있는 'UBS 글로벌 버블 지수를 활용한 부동산 버블 분석' 리포트를 참고하여 검증하고자 하였다. 국토연구원에서는 스위스 글로벌 금융기업(UBS)에서 발표하는 UBS 글로벌 버블 지수를 국내의 도시의 변수에 대입하여 각 시도별 버블 위험 정도를 '버블 위험', '고평가', '적정 수준', '저평가', '침체'의 5단계로 구분하였는데 리포트의 결과와 본 연구의 결과를 비교하여 신뢰성을 평가해 보았다. 그 결과 리포트에서 2020년 1분기에서 2021년 1분기에 버블 위험이라 판단한 적 있는 서울, 경기, 대전을 본 연구에서도 MSE 상위권의 버블 위험 지역으로 구분하였고, 리포트의 고평가 진단을 받은 지역이었던 인천, 광주, 대구, 부산을 본 연구에서도 버블 가능성이 있는 지역으로 분류하였다.

그림. 10은 산점도를 2011년 7월부터 2021년 3월까지 기간별로 나열한 경우를 보여준다. 버블 위험 지역을 중심으로 시간이 지날수록 전체적으로 우상향하는 양상을 보이는 것을 알 수 있다. 이는 버블이 시계열적으로 악화되고 있음을 시사하며, 비교적 증가율이 낮은 지방에 비하여 수도권은 비약적으로 버블 가능성이 상승하니 이를 분석하고 대책을 강구할 때 지역별 구분을 통한 접근이 필요함을 알 수 있다.

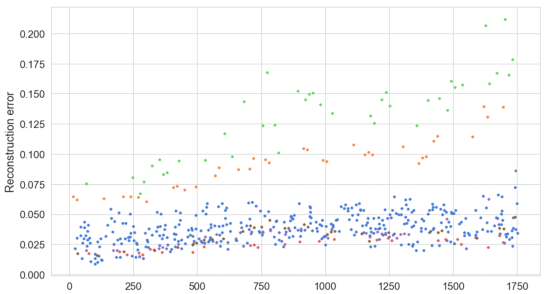

Fig. 10 MSE distribution chart by period

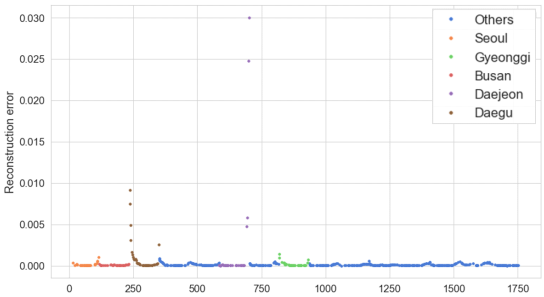

Fig. 11 MSE analysis using housing-related variables

또한 <그림.11>에서 확인할 수 있듯이, 기존 연구에서 버블 관련 분석의 변수로 사용하였던 주택 관련 지수(아파트 매매 가격 지수, 아파트 전세 가격 지수, 주택 매매 가격 지수, 주택 전세 가격)지수만을 이용하여 같은 기간 오토인코더로 분석을 할 경우 MSE의 분포가 모든 지역에서 거의 차이가 나지 않는다는 것을 알 수 있다. 이는 부동산 버블 분석에 단순히 주택 매매가와 전세가의 비율로는 설명하기 어려우며, 부동산 버블에 있어 다각도의 변수 분석이 필요함을 알 수 있다.

Ⅴ. 결 론

본 연구는 기존에 부동산 연구에 사용되지 않았던 부동산 심리 지수 등의 변수를 오토인코더를 이용하여 분석에 포함시켰고 이를 통하여 버블 위험 지역을 분리해 내었다. 부동산 버블 예측에 사용한 21가지의 변수는 총 10가지의 분류하였고, 직접적인 부동산 가격을 나타내는 변수를 제외하고도 부동산 심리 지수와 등 부동산 버블에 있어서 그동안 포함되기 어려웠던 심리적 요인을 정규화하여 분석하였다는 것에 의의가 있다.

현재 서울과 경기권 위주로 부동산 버블 가능성이 타 지역에 비해 급속도로 상승하고 있고, 그 뒤로 대전과 부산 등의 지역이 이를 따라가고 있다. 부동산 버블이란 명확한 기준을 잡기 어렵고 하나의 지표로 버블을 판단하기는 불가능하기 때문에 다양한 변수를 포함하여 실증적으로 이를 나타낸 점이 본 연구가 기존 연구에 비해 가지는 차별점이다. 부동산 버블은 주로 사후적으로 측정되고 현재 버블이라고 생각했던 가격이 차후에 평가가 될 땐 정상 가격이라 생각될 수도 있다. 그러므로 버블에 관한 정책과 대책을 세울 때는 부동산 가격과 매매

가 증가율 만을 평가하기보다는 부동산 가격에 영향을 미치는 각 변수 간의 상관관계를 분석하여 그것이 버블의 형성에 어떠한 영향을 미치는지 판단하고 가격 규제 말고도 경제 전반에 걸친 복합적인 정책 제안이 필요하다고 사유한다.

본 연구에서는 과거와 현재의 데이터와 지역별 데이터의 분석을 통해 그 차이를 그래프로 나타내었지만, 연구에서 사용한 MSE는 버블의 발생 가능성을 나타낼 뿐 결국 이를 버블로 판단하는 것은 판단자의 몫이라는 한계점이 있다. 또한 연구에 사용된 데이터는 결측치의 문제로 지역의 세부적인 분류를 하지 못하였고 경기도 등의 지역은 지역별 상승 폭이 크게 다름에도 지수가 평균적으로 기록되어 일반적인 상식과는 다른 결과를 도출하기도 하였다. 앞으로 사전적으로 버블을 측정하고자 할 때 본 연구의 결과를 활용하여 현재 나타나있는 지수들의 조합으로 과거의 데이터와 얼마나 차이를 보이는지 계측할 수 있을 것이라 전망한다.

ACKNOWLEDGEMENT

Following are results of a study on the "Convergence and Open Sharing System" Project, supported by the Ministry of Education and National Research Foundation of Korea.

REFERENCES

[1] S. W. Bae and J. S. Yu, “Predicting the Real Estate Price Index Using Machine Learning Methods and Time Series Analysis Model,” *Housing Studies Review*, vol. 26, no. 1, pp. 107-133, Feb. 2018.

[2] C. H. Jung, “A Study on the Estimation of the Housing Market Bubble by Region,” *Asia-pacific Journal of Multimedia Services Convergent with Art, Humanities, and Sociology*, vol. 9, no. 10, pp. 891-900, Oct. 2019.

[3] W. H. Kim and W. C. Kang, “Study on the Real Estate Bubble Measurement -Focusing on Apartments-,” *Journal of the KRSA*, vol. 28, no. 2, pp. 129-142, Jun. 2012.

[4] K. H. Kim, D. Y. Yang, and E. J. Kang, “Global Real Estate Price analysis Using Big Data,” *World Economy Today*, vol. 19, no. 10, May. 2019.

[5] B. H. Kim, “A Further Investigation of House Price Bubbles in Korea: Kalman Filter Approach,” *Social Studies*, vol. 6, no. 1, pp. 147-180, 2005.

[6] H. J. Chun, “An Empirical Study on the Estimate of Rational Real Estate Bubble in Korea,” *Journal of the Economic Geographical Society of Korea*, vol. 17, no. 1, pp. 147-159, 2014.

[7] Y. T. Hwang, “A Study on the Estimation of Apartment Price Index: Focused on the Machine Learning Algorithm,” *Journal of money & finance*, vol. 33, no. 3, pp. 51-83, Sep. 2019.

[8] H. J. Chun, “Analysis of Factors Influencing the Retail Property Auction Price Ratio Using the Bayesian Network Approach,” *Journal of the Korea Real Estate Management Review*, vol. 21, pp. 259-277, Jun. 2020.

김윤서(Yoonseo Kim)

성균관대학교 경영학과

※관심분야 : 머신러닝, 딥러닝, 부동산

박종찬(Jongchan Park)

성균관대학교 경제학과

※관심분야 : 머신러닝, 딥러닝, 추천시스템

오하영(Hayoung Oh)

Sungkyunkwan University Professor (2019~)
Ajou University Professor (2016~2019)
Soongsil University Professor (2013~2016)
Ph.D. in computer engineering at Seoul National University (2013)