*The Korean Journal of Applied Statistics* (2020) DOI: https://doi.org/10.5351/KJAS.2020.33.5.603
**33**(5), 603–614

# Prediction and factors of Seoul apartment price using convolutional neural networks

Hyunjae Lee[a] · Donghui Son[a] · Sujin Kim[a] · Sein Oh[b] · Jaejik Kim[a,1]

[a]Department of Statistics, Sungkyunkwan University;
[b]Department of Sports Science, Sungkyunkwan University

(Received August 24, 2020; Revised September 7, 2020; Accepted September 8, 2020)

**Abstract**

This study focuses on the prediction and factors of apartment prices in Seoul using a convolutional neural networks (CNN) model that has shown excellent performance as a predictive model of image data. To do this, we consider natural environmental factors, infrastructure factors, and social economic factors of the apartments as input variables of the CNN model. The natural environmental factors include rivers, green areas, and altitudes of apartments. The infrastructure factors have bus stops, subway stations, commercial districts, schools, and the social economic factors are the number of jobs and criminal rates, etc. We predict apartment prices and interpret the factors for the prices by converting the values of these input variables to play the same role as pixel values of image channels for the input layer in the CNN model. In addition, the CNN model used in this study takes into account the spatial characteristics of each apartment by describing the natural environmental and infrastructure factors variables as binary images centered on each apartment in each input layer.

Keywords: convolutional neural networks, image data, spatial data, apartment price

## 1. 서론

오늘날 딥러닝(deep learning) 모형들은 인공지능 분야의 핵심기술로 떠오르고 있고, 이미지(image) 또는 자연어 처리 등에서 독보적인 성능을 보유하며 널리 활용되고있다. 본 연구에서는 딥러닝 모형의 하나로 이미지 데이터에 대한 예측모형으로 널리 활용되고 있는 convolutional neural networks (CNN) 모형을 이용하여 서울 아파트 가격을 예측하고 그들의 가격결정요인들을 분석하고자 한다.

수도권 아파트 가격에 대한 문제는 과거부터 지금까지 지속적으로 이루어져 온 한국 사회의 대표적인 논쟁 중 하나이다. 지금까지 아파트 가격에 대해 경제적 또는 통계적으로 설명하려는 시도가 다수 있어왔다. 그 예로 Yoo 등 (2007)은 통계청의 회사채수익률과 주가변동률을 이용하여 주택가격과 토지가격을 설명하였고, Choi (2010)는 과민변동성검정과 공적분검정을 이용하여 전국, 서울, 강남의 평당 평균 아

The first four authors equally contributed to this paper.
This research was supported by Undergraduate Research Program funded by the Korea Foundation for the Advancement of Science & Creativity (KOFAC).
[1]Corresponding author: Department of Statistics, Sungkyunkwan University, 25-2 Sungkyunkwan-ro, Jongno-gu, Seoul 03063, Korea. E-mail: jaejik@skku.edu

파트 매매가격을 평가하였다. 또한, Lim (2014)은 아파트를 포함한 한국 전체 주택가격에 대한 예측 문제를 다루었다.

아파트의 가격은 그 아파트가 가지는 다양한 요인들에 의해 결정되어지는데 지금까지의 연구는 거시경제학적인 지표에 기반하여 설명하려는 경향이 있었고, 개별 아파트의 가격 결정 요인에 대한 분석이라기보다는 한국, 수도권, 또는 서울 전체 아파트 가격 변동과 예측에 집중하였다. 이에 본 연구에서는 아파트의 위치에 기반한 주변자연환경 및 기반시설 요소와 사회경제적 요소들을 고려하여 개별 아파트들의 가격을 예측하고 각 지역 아파트들의 주요한 가격 결정 요인들을 탐색하고자 한다. 특히 아파트 주변환경과 기반시설 요소들의 공간적 특성을 고려하기 위해 각 아파트를 중심으로 그러한 요소들의 위치를 나타내는 이미지 데이터(image data)를 입력변수로 이용한다.

본 연구에서는 아파트 가격의 예측및 요인을 찾기 위해 이미지 데이터에 대한 예측 및 분류 문제에서 뛰어난 성능을 보여온 CNN 모형을 사용하여 분석을 진행한다. 이러한 이미지 데이터를 이용한 예측 문제에 있어 랜덤포레스트(random forests)와 서포트벡터회귀(support vector regression) 모형 등도 고려될 수 있으나, 이러한 모형들은 CNN 모형의 풀링(pooling) 과정과 같이 입력변수의 차원을 줄이지 않는다면 문제가 있을 수 있다. 트리(tree) 모형을 기반으로 하는 랜덤포레스트는 트리를 만들 때 한 번에 하나의 입력변수를 탐색하기 때문에 이미지의 수 많은 픽셀들이 입력변수가 되는 경우 계산 시간이 오래 걸리는 문제가 있다. 또한 랜덤포레스트와 서포트벡터회귀는 일반적으로 자료의 크기에 비해 차원의 수가 많은 이미지 데이터의 경우 그 예측 성능이 CNN 모형에 비해 떨어진다 (Yoo 등, 2019; Hasan 등, 2019). 따라서, 본 연구에서는 CNN 모형에 집중하여 분석을 진행한다.

CNN 모형을 이용하여 부동산 가격 예측을 시도한 연구가 지금까지 다수 있었는데, Zhang 등 (2017)은 거대도시(mega city)의 관찰을 위해 위성사진을 입력 변수(input variable)로 하는 CNN 모형을 제안하였고, Yu 등 (2018)은 부동산의 공간적 특성에 대한 고려없이 부동산에 대한 정보들을 CNN 모형의 입력층(input layer)에 삽입하여 가격을 예측하였다. 즉, Yu 등 (2018)의 방법은 해당 부동산의 2,000m 안에 존재하는 학교의 수, 버스 정류장의 수, 전철역의 수 등과 같은 13개의 변수값과 이 13개의 변수 중 임의로 선택된 3개의 변수값을 $4 \times 4$의 행렬로 만든 후 이를 CNN 모형의 입력값으로 사용하였다.

그러나 본 연구에서는 아파트 주변의 각 환경 변수를 각 입력층에 할당하고 개개의 입력층의 각 픽셀(pixel)에 환경적 요인들의 유무 및 가중값을 입력하는 방식으로 특정 아파트의 자세한 공간적 특성들을 CNN 모형에 반영하는 시도를 하였다. 예를 들어 아파트 주변에 버스 정류장 개수만을 단순히 이용하는 것이 아니라 아파트를 중심으로 버스 정류장이 어느 방향으로 얼마나 떨어져있는지 등에 대한 자세한 정보가 이미지를 통해 입력층에 반영된다. 또한, 이러한 환경적인 변수 뿐만 아니라 건축연도, 학군정보, 치안, 일자리 등과 같은 사회경제학적인 변수들도 하나의 입력층으로 모형에 사용된다. 따라서, 본 연구에서는 환경적/기반시설 요인들의 공간적 정보와 사회경제학적 변수들을 고려하는 CNN 모형에 대한 새로운 시도를 하였고, 이 모형을 이용하여 각 지역의 아파트 가격이 어떠한 요인들에 의해 주로 영향을 받는지 CNN 모형으로 부터 나온 잔차(residual)를 분석하여 조사한다.

본 연구의 구성은 다음과 같다. 2절에서는 기본적인 CNN 모형에 대해 간단히 묘사하고, 3절에서는 서울지역 각 아파트에 대한 정보를 어떻게 CNN 모형의 입력층에 변환하여 고려하였는지 설명한다. 4절에서는 CNN 모형의 결과와 서울 개별 지역 아파트들의 가격결정요인들에 대해 살펴본다.

## 2. CNN 모형

CNN 모형은 이미지에서 얼굴, 사물, 동물과 같은 객체 및 특정 장면을 인식하는 유용한 도구로써 뛰어난 성능을 보여왔으며, 이에 자율주행 자동차 및 안면 인식 분야에서 널리 사용되고 있다. 기본적으로

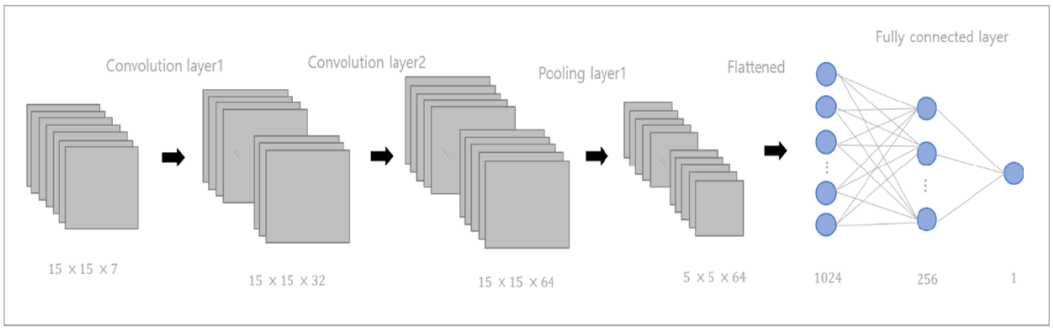



**Figure 2.1.** The structure of convolutional neural networks model used in this study.

CNN 모형은 다른 신경망 모형과 유사하게 입력층(input layer), 은닉층(hidden layer), 출력층(output layer)으로 구성된다. 여기서 구조적으로 일반 신경망 모형과 다른 점은 Figure 2.1에서와 같이 입력 데이터와 일반적인 신경망 모형 사이에 입력된 이미지의 특징(feature)을 추출하는 역할을 하는 컨볼루션층(convolution layer)이 있다는 것이다.

컬러 이미지(color image)가 입력 데이터인 경우 이미지 각 픽셀의 색에 대한 정보를 받기 위해 red (R), green (G), blue (B) 3개의 채널(channel)을 사용한다. 컨볼루션 층은 이미지의 특징을 추출하는 필터(filter)와 이 필터값들을 비선형 값으로 변환해주는 활성화 함수(activation function)로 구성된다. 즉, 이 층에서는 필터를 통해 특징들이 추출되면 활성화 함수를 통해 그 특징의 유무를 나타내는 비선형 값으로 변환이 일어난다. 컨볼루션 층을 통해 추출되고 변환된 특징들은 서브샘플링(subsampling) 과정을 통해 인위적으로 특징의 크기를 줄이는데 그 과정을 풀링이라고 부른다. 이러한 풀링 과정은 특징의 크기를 줄임으로써 계산 시간을 줄일 수 있고 과적합을 방지하는 효과를 갖는다. 이 풀링 과정을 거친 값들은 완전히 연결된(fully connected) 일반적인 신경망 모형의 입력값으로 들어가고 최종적으로 출력층에서 반응변수의 값을 얻게된다. 본 연구에서는 Figure 2.1에서 보이는 바와 같은 2개의 컨볼루션층, 하나의 풀링 층, 완전히 연결된 신경망 모형으로 이루어진 CNN 모형을 분석에 사용한다.

## 3. 데이터 전처리

본 연구는 서울 시내의 개별 아파트들이 아닌 각 아파트 단지들을 분석 대상으로 하고 그 아파트들의 가격 예측과 가격 결정 요인 분석을 목적으로 한다. 개별 아파트가 아닌 아파트 단지를 분석 대상으로 정한 이유는 같은 아파트 단지 내의 아파트들은 같은 환경요인 및 기반 시설과 사회경제적 지표들을 가지고 있다고 가정할 수 있기 때문이다. 따라서, 반응변수로는 2018년도 서울 아파트 매매가의 $1m^2$ 당 가격을 통해 얻은 각 단지별 $1m^2$ 당 평균가격을 사용하였고, 이는 국토교통부에서 제공하는 아파트 실거래가 데이터를 통해 계산하였다. 입력변수는 크게 자연환경요소(natural environment factor), 기반시설요소(infrastructure factor), 사회경제적요소(social economic factor)로 구분할 수 있다. 자연환경요소로는 강, 녹지, 고도 데이터를 사용하였고 기반시설요소로서 버스정류장, 지하철역, 학교, 상권(카페, 패스트푸드점, 영화관) 데이터를 활용하였다. 강, 녹지, 고도 데이터는 국토지리정보원에서 제공하는 지리 위치 정보와 속성정보를 통해 이용 가능하다. 사회 경제적 요소로는 아파트의 건축연도, 해당 아파트 단지가 속한 지역의 서울대학교 합격자 수(교육 요인), 범죄율, 그 지역의 일자리 수를 사용하였고, 이러한 데이터는 '서울시 열린 데이터 광장'을 통해 얻을 수 있다.

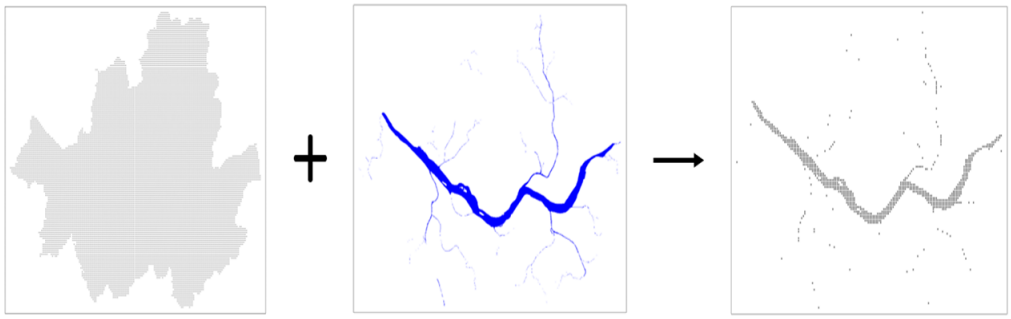


**Figure 3.1.** Transformation of polygon data into point data.

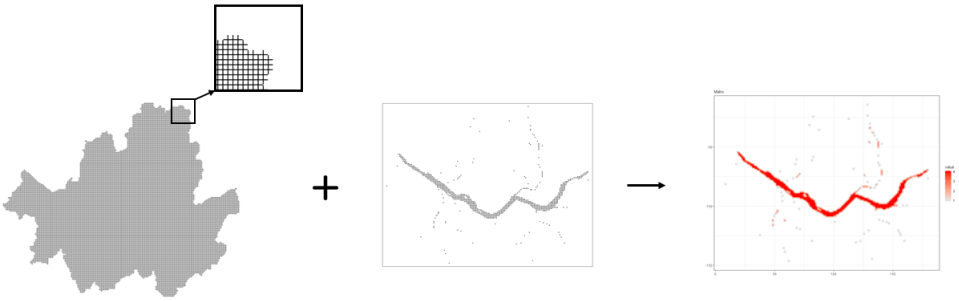


**Figure 3.2.** Transformation of polygon data into matrix format.

본 연구에 사용된 데이터들 중 등록(registration)이 필요한 데이터는 점데이터(point data)와 폴리곤 데이터(polygon data) 이다. 점데이터는 지점별로 데이터가 주어지며 버스정류장, 지하철역, 학교 등이 점데이터에 해당한다. 폴리곤 데이터는 일정한 면적에 대해 값이 주어지며 강과 녹지 등이 이에 해당한다.

점데이터와 폴리곤 데이터를 통합하여 분석에 사용하기 위해 두 데이터를 행렬로 변환하는 과정이 필요한데, 본 연구에서는 폴리곤 데이터와 점데이터를 통합하기 위해 폴리곤 데이터를 점데이터의 형태로 변환하였다. 이를 위해 서울시내에 일정한 간격으로 포인트를 부여하고 폴리곤 데이터를 이용해 폴리곤 내에 있는 점들만 남기는 방식을 사용하였다. 이를 이용해 남겨진 점들은 해당 데이터가 존재하는 지역이라는 정보를 담고 있다. 즉, 폴리곤 데이터 내에 있는 점에 대해서는 가중값 1을 부여하였고, 점이 없는 지역은 0의 값을 할당하였다. Figure 3.1은 폴리곤 데이터로 나타난 강을 포인트 데이터로 변환하는 과정을 보여준다. Figure 3.1에서 왼쪽의 첫 번째 그림은 서울시내가 일정한 크기의 점들로 빈틈없이 배치된 것을 보여주며, 두 번째 그림은 서울시내의 강의 위치를 나타내는 폴리곤 데이터를 보여준다. 이 두 그림을 겹쳤을 때 첫 번째 그림의 점이 두 번째 그림의 폴리곤 데이터의 강 면적에 위치하면 그 점에 1의 값을 부여하고 그 외의 점에는 0의 값을 부여하였다. 결과적으로 0과 1의 이진값으로 강을 표현한 것이 세 번째 그림이다.

이를 최종적인 행렬의 헝태로 만들기 위해 본 연구에서 사용한 방법은 다음과 같다. 우선 Figure 3.2의 첫 번째 그림에서와 같이 서울시내를 100m 간격의 격자로 만든 후 격자 내에 위치한 점들의 값을 모두 합한 값으로 서울시 전체에 대한 행렬을 구성한다. Figure 3.2의 세 번째 그림은 강에 대해 각 격자 내의 점들의 가중값을 모두 합한 값을 나타낸다. 최종적으로 이를 아파트 단지별 행렬로 만들기 위해 서울

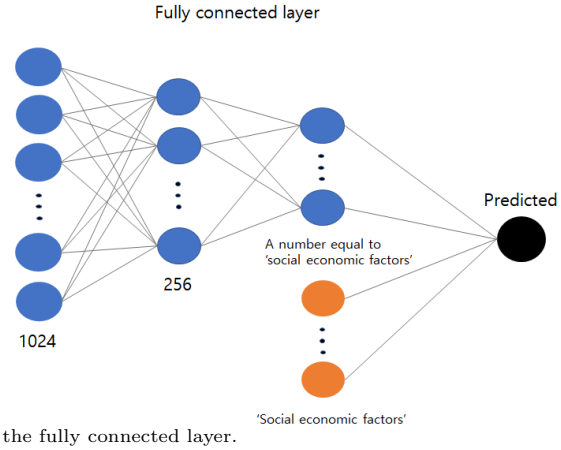



**Figure 3.3.** Structure of the fully connected layer.

시의 격자 한 칸은 각 아파트 단지 행렬의 각 원소에 대응하도록 한다.

본 연구에서는 자연환경요소, 기반시설요소, 사회경제적요소를 입력변수로 하고, 2018년도 아파트 $1m^2$ 당 평균매매가를 출력변수로 하는 CNN 모형을 고려한다. 이처럼 다양한 형태의 데이터를 적절한 방식으로 CNN 모형에 입력하기 위해 이미지 특성 저장소인 채널과 이미지가 아닌 입력값을 받는 단일층 두 가지 방식이 사용된다. 2절에서 설명한 바와 같이 일반적으로 컬러 이미지 분석에서 채널은 이미지의 red (R), green (G), blue (B) 색상 값을 저장하는 곳이다. 단일층은 색상이 아닌 수치적 변수들을 추가할 수 있는 기본적인 신경망 모형의 구조이다. 본 연구에서는 이미지 데이터의 픽셀이 행렬의 형태로 적용되는 CNN 모형의 일반적인 입력방식에서 기반하여 각 채널마다 색상이 아닌 픽셀단위의 공간정보 행렬(자연환경요소 및 기반시설 요소)을 만들어 모형이 특정 공간에 대한 정보를 학습하게 한다.

채널에 입력되는 변수로는 앞서 언급한 자연환경요소(녹지, 강, 고도)와 기반시설요소(버스정류장, 지하철역, 학교, 상권)의 행렬 데이터를 사용하였다. 이는 대상이 되는 아파트 단지를 중심으로 하는 행렬의 형태로 가공되어 입력변수의 값으로 사용된다. 각 행렬 원소 하나는 100m × 100m 내의 정보를 의미하며 실제 모형에는 입력값으로 $15 \times 15$ 행렬의 채널을 사용하였다. 이를 통해 해당 아파트 단지 주변 750m 반경의 환경/기반시설요소들이 고려될 수 있다. 본 연구에서는 아파트를 중심으로 750m를 아파트 거주자들이 도보로 접근할 수 있는 최대 거리로 간주하였고, 이 거리 내의 시설 및 환경들은 거주자가 상대적으로 쉽게 이용 가능하기 때문에 아파트 가격을 결정하는데 있어 중요한 정보들이 그 거리 내에 대부분 존재한다고 가정한다. 본 연구에서는 서울시내 전체를 100m 간격의 격자로 나누었기 때문에 반경 750m의 정보는 $15 \times 15$ 행렬로 표현되었다. 물론, 서울시내 전체를 10m 또는 50m의 격자로 더 세밀하게 나눌 수 있을 것이다. 이 경우 세밀한 공간 정보를 고려할 수 있는데 반해 행렬의 요소 수가 증가하면서 입력변수 역시 증가하여 과적합이 발생할 가능성이 존재한다.

CNN 모형의 출력변수는 행렬의 가운데 픽셀에 위치한 아파트 단지의 $1m^2$ 당 평균가격으로 설정하였다. 이 과정에서 서울 변두리에 위치해 서울이 아닌 지역을 포함하게 되는 아파트 단지들은 관측값으로 고려하지 않았다. 단일층 입력 변수로써 사회경제적요소인 아파트 단지의 건축연도, 교육(서울대학교 합격자 수), 치안(범죄율), 일자리 데이터를 고려한다. 각 아파트 단지의 교육, 치안, 일자리 변수들의 정보는 해당 아파트 단지가 위치하는 행정동의 정보를 이용하였고, 이 변수들은 Figure 3.3에서 보여지는 것과 같이 풀링과정을 통해 나온 자연환경요소, 기반시설요소의 변환된 값들과 더불어 완전히 연결된 층에 하나의 단일층으로 입력된다.

**Table 4.1.** Convolutional neural networks models considered in this study

| Model | Input variable |
|---|---|
| Model0 (M0) | No factor (null model) |
| Model1 (M1) | Natural environmental factors |
| Model2 (M2) | Natural environmental factors, infrastructure factors |
| Model3 (M3) | Natural environmental factors, infrastructure factors, education |
| Model4 (M4) | Natural environmental factors, infrastructure factors, criminal rate |
| Model5 (M5) | Natural environmental factors, infrastructure factors, the number of jobs |
| Model6 (M6) | Natural environmental factors, infrastructure factors, building year |
| Model7 (M7) | Natural environmental factors, infrastructure factors, social economic factors |

모형의 예측력을 올리기 위해 학습 데이터가 약 4,900개로 많지 않기 때문에 데이터의 양을 늘리는 방법을 시도하였다. 이를 위해 CNN 모형에서 사용하는 기본적인 방식인 회전(rotation) 기법을 사용하였다. 본 연구에서는 이미지가 아닌 아파트 단지에 대한 행렬을 90°, 180°, 270°로 회전시킴으로써 입력 데이터를 4배로 증가시킬 수 있었다.

본 연구에서는 공간 데이터의 전처리와 시각화를 통계 소프트웨어 `R`과 패키지(package) `sp` (Pebesma와 Bivand, 2005), `tidyverse` (Wickham, 2017), `automap`을 이용하였고, CNN 모형 적합을 위해 프로그래밍 언어 `Python`의 오픈 소스 라이브러리(open source library)인 `tensoflow` (Abadi 등, 2015)를 사용하였다.

## 4. 분석 모형과 결과

본 연구에서는 서울 아파트의 $1m^2$ 당 매매가를 예측하기 위해 2개의 컨볼루션 층, 1개의 풀링층, 2개의 완전히 연결된 은닉층을 갖는 CNN 모형을 사용하였고, 활성함수로는 rectified linear unit (ReLu)를 사용하였다. 입력변수로는 녹지, 강, 고도를 고려한 자연환경요소, 학교, 버스역, 지하철역, 상권을 나타내는 기반시설요소, 교육, 치안, 일자리, 건축연도를 포함하는 사회경제적요소가 예측을 위해 사용된다. 3에서 언급했듯이 자연환경요소의 녹지, 강, 고도와 기반시설요소인 학교, 버스역, 지하철역, 상권은 각 아파트 단지를 중심으로 7개의 $15 \times 15$ 행렬들로 입력된다. 4개의 사회경제적요소는 단일층으로 완전히 연결된 층에 바로 입력되는 형태를 갖는다. 입력값은 미니배치(mini batch) 형태로 넣었으며 손실함수(loss function)는 root mean squared error (RMSE)를 사용하였다. 또한, 모형의 과적합을 피하기 위해 탈락(dropout)기법을 적용하였다. 각 모형에 대한 최적의 조정모수(tuning parameter)들은 원자료를 트레이닝(training)과 테스트셋(test set)으로 무작위로 나누는 것을 반복하여 테스트 오차(test error)를 측정하여 최소가 되도록 설정하였다.

Table 4.1은 분석에서 고려된 8개의 모형(M0–M7)을 보여준다. M0는 어떠한 설명변수없이 서울 전체 아파트 단지들의 $1m^2$ 당 매매가들의 평균으로만 예측한 모형이고, M1은 자연환경요소만, M2는 자연환경요소와 기반시설요소만을 입력변수로 고려한 모형이다. M2–M6는 자연환경요소와 기반소설요소에 부가하여 교육, 치안, 일자리, 건축연도와 같은 요소들을 각각 고려한 모형이다. 최종적으로 M7은 자연환경, 기반시설, 사회경제요소 모두를 고려한 완전 모형(full model)이다.

Figure 4.1은 모형 M7에 대한 CNN 모형의 모수들이 추정되어지는 과정에서 트레이닝과 테스트 RMSE의 변화를 보여준다. 모수의 추정이 진행됨에 따라 트레이닝 RMSE는 계속 감소하고 테스트 RMSE는 작아지다가 다시 약간 커지는 추세를 보여준다. 본 연구에서는 과적합을 방지하기 위해 테스트 RMSE가 가장 작아지는 지점에서 최종적인 CNN 모형을 구하였다.

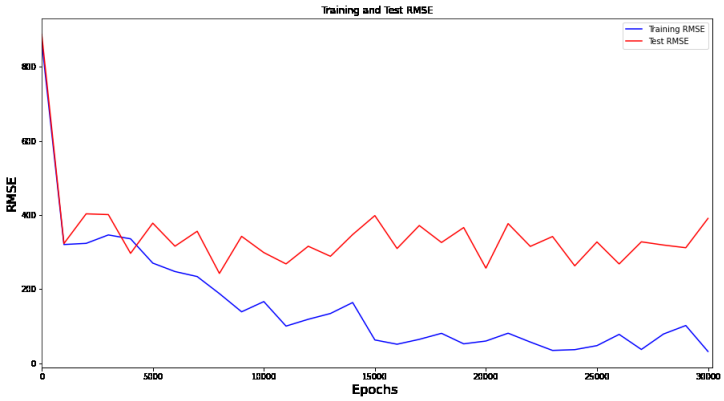
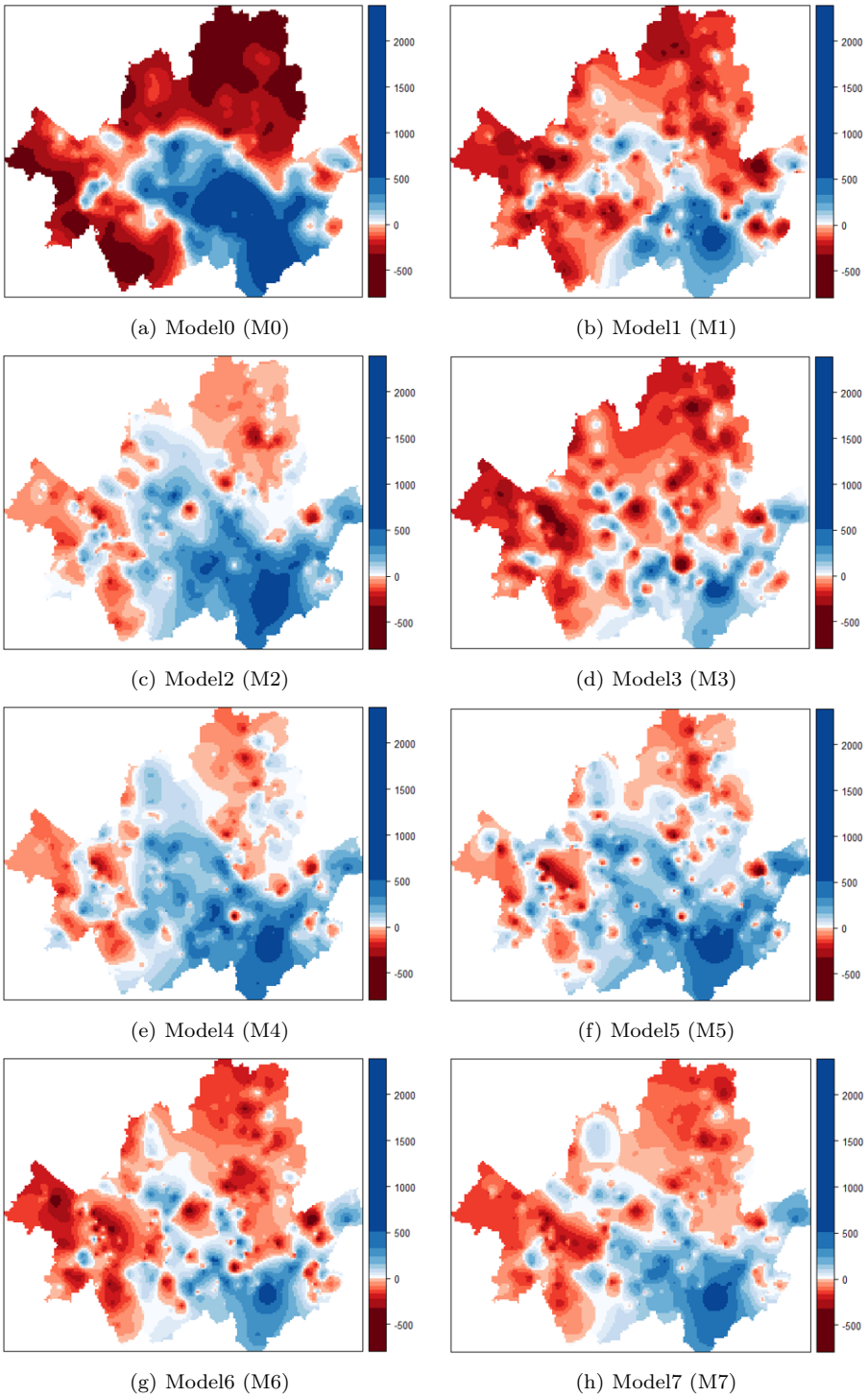



**Figure 4.1.** Trends of training and test root mean squared error by epochs.

Figure 4.2은 각 모형에서 구한 각 아파트 단지의 예측값과 실제 매매가의 차이를 나타내는 잔차의 크기를 크리깅(kriging) 기법을 사용하여 서울 전체에 대해 시각화한 것이다. 빨간색은 실제 매매가가 모형의 예측 매매가보다 작은 음의 잔차를 나타내고 색깔이 진해질수록 잔차의 크기가 커짐을 의미한다. 반대로 파란색은 실제 매매가가 예측 매매가 보다 큰 양의 잔차를 나타내고 이 역시 색깔이 짙어질수록 실제 매매가와 예측 매매가의 차이가 크다는 것을 의미한다. 색깔이 흰색에 가까울수록 실제 매매가와 예측 매매가의 차이가 작아진다. Figure 4.2(a)은 매매가 전체 평균만으로 예측한 모형(M0)으로 서울 전역에서 매우 진한 빨간색과 파란색을 보임을 알 수 있다. 이 색깔을 기준으로 보면 강남구, 서초구, 송파구 등의 강남 3구와 용산구, 목동 지역 등이 매우 진한 파란색으로 서울 전체 평균보다 매매가가 높음을 알 수 있다. Figure 4.2(b)는 아파트 단지 주변의 녹지, 강, 고도의 자연환경요소만을 고려한 모형(M1)의 결과로 Figure 4.2(a)와 비교하여 전반적으로 색깔이 많이 옅어졌음을 알 수 있고, 이는 자연환경요소가 아파트 매매가를 예측하는데 있어 중요한 요인 중의 하나임을 알 수 있게 해준다. 그러나 여전히 서울의 여러지역에서 진한 빨간색 또는 파란색들이 존재하고 있으므로 자연환경요소 이외의 매매가를 예측하는데 중요한 변수들을 추가할 필요가 있어 보인다.

Figure 4.2(c)는 자연환경요소와 기반시설요소를 고려한 모형에 대한 잔차를 나타낸 그림이다. 이를 Figure 4.2(b)와 비교하면 기반시설요소를 자연환경요소에 추가함으로써 전반적으로 색깔이 옅어졌음을 알 수 있고, 이는 실제 매매가와 모형의 적합값의 차이가 줄었음을 의미한다. 특히 모형의 적합값보다 실제매매가가 낮았던 많은 지역(빨간색 지역)들이 옅은 파란색으로 변했음을 보여준다. 이는 이 지역들이 기반시설요소에 의해 예측력이 향상되었음을 나타낸다. Figure 4.2(d)부터 Figure 4.2(g)까지의 결과는 자연환경요소와 기반시설요소에 교육, 치안요소, 일자리수, 아파트의 건축연도를 각각 추가한 모형들의 잔차를 나타낸 그림이다. 교육요소를 추가한 Figure 4.2(d)의 결과는 진한 파란색의 강남지역이 많이 옅어졌음을 보여주고 있다. Figure 4.2(c)에 비교하여 Figure 4.2(d)와 Figure 4.2(f)의 치안요소, 일자리수를 고려한 모형들은 잔차의 그림은 전반적으로 아주 뚜렷한 차이를 보여주지 못했다. 건축연도를 고려한 Figure 4.2(g)는 Figure 4.2(c)에 비해 진한 빨간 구역들이 강서구, 금천구와 강북지역을 중심으로 증가하였다. Figure 4.2(h)는 자연환경요소, 기반시설요소, 교육, 치안, 일자리, 건축연도의 사회경제적요소 모두를 고려한 모형이다. 이를 Figure 4.2(b)와 비교해보면 흰색에 가까운 부분들이 증가하여 실제 매매가와 모형 적합값이 근접한 지역들이 많아졌음을 알 수 있다. 또한, 색깔은 많이 옅어졌으나, 강남 3구와 용산, 강동구 등이 여전히 모형의 값보다 실제 매매가가 높은 지역들이고, 강북과 강서를 중심으로 한 지역들은 여전히 모형의 값보다 실제 매매가가 낮음을 나타낸다.




(a) Model0 (M0) (b) Model1 (M1)

(c) Model2 (M2) (d) Model3 (M3)

(e) Model4 (M4) (f) Model5 (M5)

(g) Model6 (M6) (h) Model7 (M7)

**Figure 4.2.** Change of residuals for apartment price.

**Table 4.2.** Test root mean squared error (RMSE) of each model

| Model | M0 | M1 | M2 | M3 | M4 | M5 | M6 | M7 |
|---|---|---|---|---|---|---|---|---|
| Test RMSE | 440.0 | 332.7 | 317.6 | 293.3 | 300.3 | 302.2 | 305.8 | 290.1 |

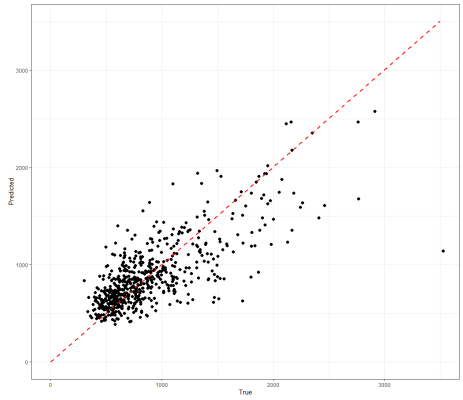



**Figure 4.3.** Scatter plot of real apartment prices and predicted values from M7.

**Table 4.3.** Test root mean squared error (RMSE) values of models with two factors

| Model | Input variable | Test RMSE |
|---|---|---|
| M2 | Natural environmental factor, infrastructure factor | 317.6 |
| M8 | Natural environmental factor, social economic factor | 299.3 |
| M9 | Infrastructure factor, social economic factor | 323.5 |
| M7 | All factors | 290.1 |

Table 4.2는 각 모형들의 테스트 RMSE 값들을 보여준다. Table 4.2로 부터 모든 변수들을 입력변수로 하는 완전 모형 M7이 가장 낮은 테스트 RMSE를 갖는다는 것을 알 수 있다. Figure 4.3는 모든 요소들을 고려한 M7 모형의 출력변수값과 실제 매매가의 산점도를 보여준다. 전반적으로 45°선을 중심으로 점들이 분포되어있음을 보여주고 있고 이는 모형의 예측에 대한 적합성이 나쁘지 않음을 의미한다. M7이 자료의 전체 변동을 설명하는 비율은 62.5%이다 (Test RMSE = 290.1).

CNN 모형은 모형 내부에서 모든 변수들의 선형결합의 비선형 함수로써 연결되므로 M7이 가장 낮은 테스트 오차를 갖는다고 해서 M7에 포함된 모든 변수들이 아파트 가격 예측에 중요하다는 의미는 아니다. 다만, 모든 변수들이 모형에 입력변수로 사용되었을 때 그 변수들의 어떤 조합이 가격을 예측하는데 중요한 역할을 한다고 볼 수 있다. 이러한 관점에서 자연환경요소, 기반시설요소, 사회경제요소 중 서울 아파트 가격 예측에 필수적인 요소가 무엇인지 알아보기 위해 M7의 테스트 RMSE를 기준으로 각 요소를 제거한 세 개의 모형들의 테스트 RMSE를 살펴보았다.

Table 4.3에서 보듯이 세 개의 모형 (M2, M8, M9) 중에서 M7을 기준으로 자연환경요소를 제거한 M9의 테스트 RMSE가 가장 크게 증가했음을 알 수 있다. 그 다음으로 크게 테스트 RMSE가 증가한 모형은 사회경제요소를 제거한 M2였다. 이를 통해 서울 아파트 요인을 결정하는데 있어 가장 필수적인 요소는 자연환경요소임을 알 수 있고, 그 다음으로는 교육과 일자리수를 포함하는 사회경제요소가 중요함을 알 수 있다. 기반시설요소가 서울 지역 아파트 가격 예측에서 상대적으로 중요도가 떨어지는 이유는 전체적으로 서울지역 아파트의 경우 많고 적음의 차이는 있겠지만 주변에 교통이나 상점들이 존재하기 때문에 가격결정에 있어 아주 큰 영향은 없는 것으로 추측된다.

## 5. 결론 및 토의

본 연구는 이미지 데이터에 대한 예측 모형으로 뛰어난 성능을 보여온 CNN 모형을 이용하여 서울 아파트 가격의 예측과 서울 지역 아파트들의 가격결정 요인들을 전차의 변화를 통해 조사하는데 집중하였다. 이를 위해 자연환경요소, 기반시설요소, 사회경제요소들을 입력변수로 고려하였고, CNN 모형이 이미지 데이터에 좋은 성능을 보여온 것에 착안하여 이 입력변수들의 값들을 CNN 모형 입력층으로써 이미지 채널의 픽셀값과 같은 역할을 하도록 변환하여 예측모형을 만드는 시도를 하였다.

아파트 매매가는 입지를 나타내는 자연환경요소, 기반시설요소, 학군, 일자리 등의 요인들 이외에도 설명하기 힘든 인간의 심리적 요인들에 의해 결정되기도 하고 같은 1년의 기간 안에서도 지역마다 변동의 차이가 크기 때문에 모형의 예측 정확성을 담보하는 것은 쉽지 않은 문제이다. 이러한 이유로 본 연구에서 사용한 이미지 데이터와 전통적인 입력변수 모두를 고려한 CNN 모형은 전체 자료 변동의 62.5%만을 설명하였다. 또한, 기본적으로 입력변수들의 출력변수에 대한 해석 및 예측에서의 중요도를 측정하기 힘든 신경망 모형을 사용하였기 때문에 각 지역들의 가격결정요인들을 명확하게 밝히는 것은 쉬운 일이 아니었다. 그럼에도 본 연구에서는 서울 각 지역들에 대해 아파트 매매가를 결정하는 요인들을 알아보기 위해 여러 입력변수들의 조합을 고려한 모형들의 잔차들의 변화를 살펴보았고, 이를 통해 가격결정요인들에 대한 해석을 시도하였다. 그러나 신경망 모형 그 자체는 모든 입력변수들의 비선형 변환들이 여러 개의 층을 통해 연결되어 있기 때문에 모형들의 잔차의 변화가 단지 특정 변수의 추가 또는 삭제만으로 명확하게 설명되기 어려운 한계가 있다.

본 연구에서는 1년 동안 발생된 매매가에 대해 분석을 진행하였다. 특정 아파트 단지 내에서도 선호하는 동과 층수에 따라 가격이 천차만별일 수 있고, 단지내 아파트 면적별 구성에 의해서도 가격이 달라질 수 있기 때문에 이러한 요소들도 고려하고 더 긴 기간에 대해 자료를 수집한다면 본 연구에서 제안한 모형이 좀 더 향상될 수 있을 것으로 기대한다.

또한 향후에 적절한 차원 축소 또는 변수선택 방법을 결합한 랜덤포레스트를 이용하여 서울 아파트 가격 예측 및 예측에 중요한 역할을 하는 변수를 선택하여 CNN 모형의 결과와 비교해보는 것도 좋은 시도가 될 것으로 보인다.

## References

Abadi, M., Agarwal, A., and Barham, P., *et al.* (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Available from: https://tensorflow.org

Choi, C. S. (2010). A study on the existence of price bubbles in Korean housing market, *Journal of the Korea Real Estate Society*, **32**, 177–199.

Hasan, M., Ullah, S., Khan, M. J., and Khurshid, K. (2019). Comparative analysis of SVM, ANN and CNN for classifying vegetation species using hyperspectral thermal infrared data, *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2/W13.

Lim, S. S. (2014). A study on the forecasting models using housing price index, *Journal of the Korean Data & Information Science Society*, **25**, 65–76.

Pebesma, E. J. and Bivand, R. S. (2005). Classes and methods for spatial data in R, *R News*, **5**. Available from: https://cran.r-project.org/doc/Rnews/.

Wickham, H. (2017). tidyverse: Easily install and load the 'Tidyverse'. R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse.

Yoo, C., Han, D., Im, J., and Bechtel, B. (2019). Comparison between convolutional neural networks and random forest for local climate zone classification in mega urban areas using Landsat images, *ISPRS Journal of Photogrammetry and Remote Sensing*, **157**, 155–170.

Yoo, J.-S., Lim, K.-C., and Kie, S.-D. (2007). The empirical analysis of the bubble in housing price and

land price, *Journal of Industrial Economics and Business*, **20**, 2245–2264.

Yu, L., Jiao, C., Xin, H., Wang, Y., and Wang, K. (2018). Prediction on housing price based on deep learning, *International Journal of Computer and Information Engineering*, **12**, 90–99.

Zhang, F., Du, B., and Zhang, L. (2017). A multi-task convolutional neural network for mega-city analysis using very high resolution satellite imagery and geospatial data. arXiv preprint arXiv:1702.07985.

# CNN 모형을 이용한 서울 아파트 가격 예측과 그 요인

이현재[a] · 손동희[a] · 김수진[a] · 오세인[b] · 김재직[a,1]

[a]성균관대학교 통계학과, [b]성균관대학교 스포츠과학과

(2020년 8월 24일 접수, 2020년 9월 7일 수정, 2020년 9월 8일 채택)

**요 약**

본 연구는 이미지 데이터에 대한 예측 모형으로 뛰어난 성능을 보여온 convolutional neural networks (CNN) 모형을 이용하여 서울 아파트 가격의 예측과 서울 각 지역 아파트들의 가격결정요인들을 연구한다. 이를 위해 강, 녹지, 고도와 같은 자연환경요인, 버스정류장, 지하철역, 상권, 학교 등과 같은 기반시설요소, 일자리수, 범죄율 등의 사회경제요소들을 설명변수로 고려하고, CNN 모형이 이미지 데이터에 좋은 성능을 보여온 것을 기반으로 이 설명변수들의 값들을 CNN 모형 입력층으로써 이미지 채널의 픽셀값과 같은 역할을 하도록 변환하여 아파트 가격의 예측과 가격결정요인에 대한 해석을 시도한다. 덧붙여 본 연구에서 사용된 CNN 모형은 자연환경요인과 기반시설요인 변수들을 각 아파트를 중심으로 하는 각 입력층의 채널에 이진의 이미지로 표현함으로써 각 아파트의 공간적인 특성을 고려할 수 있다.

주요용어: CNN모형, 이미지데이터, 공간데이터, 아파트가격

첫 네 명의 저자는 본 논문의 공동제1저자임.

이 논문은 2019년도 한국과학창의재단의 학부생 연구 프로그램의 지원을 받아 수행된 연구임.

[1]교신저자: (03063) 서울시 종로구 성균관로 25-2, 성균관대 통계학과. E-mail: jaejik@skku.edu