Land Use Policy xxx (xxxx) xxxx

Contents lists available at ScienceDirect

Land Use Policy

journal homepage: www.elsevier.com/locate/landusepol

Understanding house price appreciation using multi-source big geo-data and machine learning

Yuhao Kang[a,b], Fan Zhang[a,*], Wenzhe Peng[c], Song Gao[b], Jinmeng Rao[b], Fabio Duarte[a,d], Carlo Ratti[a]

[a] Senseable City Lab, Department of Urban Studies and Planning, Massachusetts Institute of Technology, Cambridge, MA 02139, United States
[b] Geospatial Data Science Lab, Department of Geography, University of Wisconsin, Madison, WI 53703, United States
[c] Department of Architecture, Massachusetts Institute of Technology, Cambridge, MA 02139, United States
[d] Urban Management Program, PUCPR, Curitiba 80215-910, Brazil

ARTICLE INFO

Keywords:
House price appreciation rate
Street view images
House photos
Human mobility patterns
Geographically weighted regression

ABSTRACT

Understanding house price appreciation benefits place-based decision makings and real estate market analyses. Although large amounts of interests have been paid in the house price modeling, limited work has focused on evaluating the price appreciation rate. In this study, we propose a data-fusion framework to examine how well house price appreciation potentials can be predicted by combining multiple data sources. We used data sets including house structural attributes, house photos, locational amenities, street view images, transportation accessibility, visitor patterns, and socioeconomic attributes of neighborhoods to enrich our understanding of the real estate appreciation and its predictive modeling. As a case study, we investigate more than 20,000 houses in the Greater Boston Area, and discuss the spatial dependency of house price appreciations, influential variables and their relationships. In detail, we extract deep features from street view images and house photos using a deep learning model, merging features from multi-source data and modeling house price appreciation using machine learning models and the geographically weighted regression at two spatial scales: fine-scale point level and aggregated neighborhood level. Results show that the house price appreciation rate can be modeled with high accuracy using the proposed framework (R^2 = 0.74 for gradient boosting machine at neighborhood-scale). We discovered that houses with low house prices and small house areas may have a higher house appreciation potential. Our results provide insights into how multi-source big geo-data can be employed in machine learning frameworks to characterize real estate price trends and help understand human settlements for policy-making.

1. Introduction

As an important aspect of human settlement, house prices are strongly associated with economic activities (Chen et al., 2016). Understanding the trends in house prices can provide suggestions not only for house buyers but also for researchers and decision makers in real estate market, urban planning and development. For decades, researchers from economy, urban planning, geography, politics and computer science have made great efforts in house price-related topics to understand the impacts of property values in different socioeconomic environments (Archer et al., 1996; Cao et al., 2019; Fu et al., 2016; Hu et al., 2019).

Despite large amounts of existing studies, two aspects received insufficient attention. First, most existing literature focuses on the house price modeling but neglects the study of price appreciation rate (Hung and Tu, 2008; Livy, 2017). Compared with absolute values of house prices, which are only snapshots of the property values in a specific time window, house price appreciation rates can reflect the growth or decay of property values from a long-term perspective. In addition, high house price does not equal to a high house price appreciation rate. A same variable may have totally different impacts on house prices and on appreciation rates. Therefore, examining the effect of different variables on house price appreciation is important and promising.

Second, existing models such as the *hedonic pricing model* proposed by Rosen (1974) typically only take *structural attributes* and *locational amenities* into consideration, which may not describe the other aspects of factors influencing the house price appreciation rate comprehensively. In practice, structural attributes contain the tangible assets of the property, including the size of the house, the year built, the number of the bedrooms and bathrooms, etc., which can describe the inner

* Corresponding author.
E-mail address: zhangfan@mit.edu (F. Zhang).

https://doi.org/10.1016/j.landusepol.2020.104919
Received 7 September 2019; Received in revised form 6 July 2020; Accepted 9 July 2020

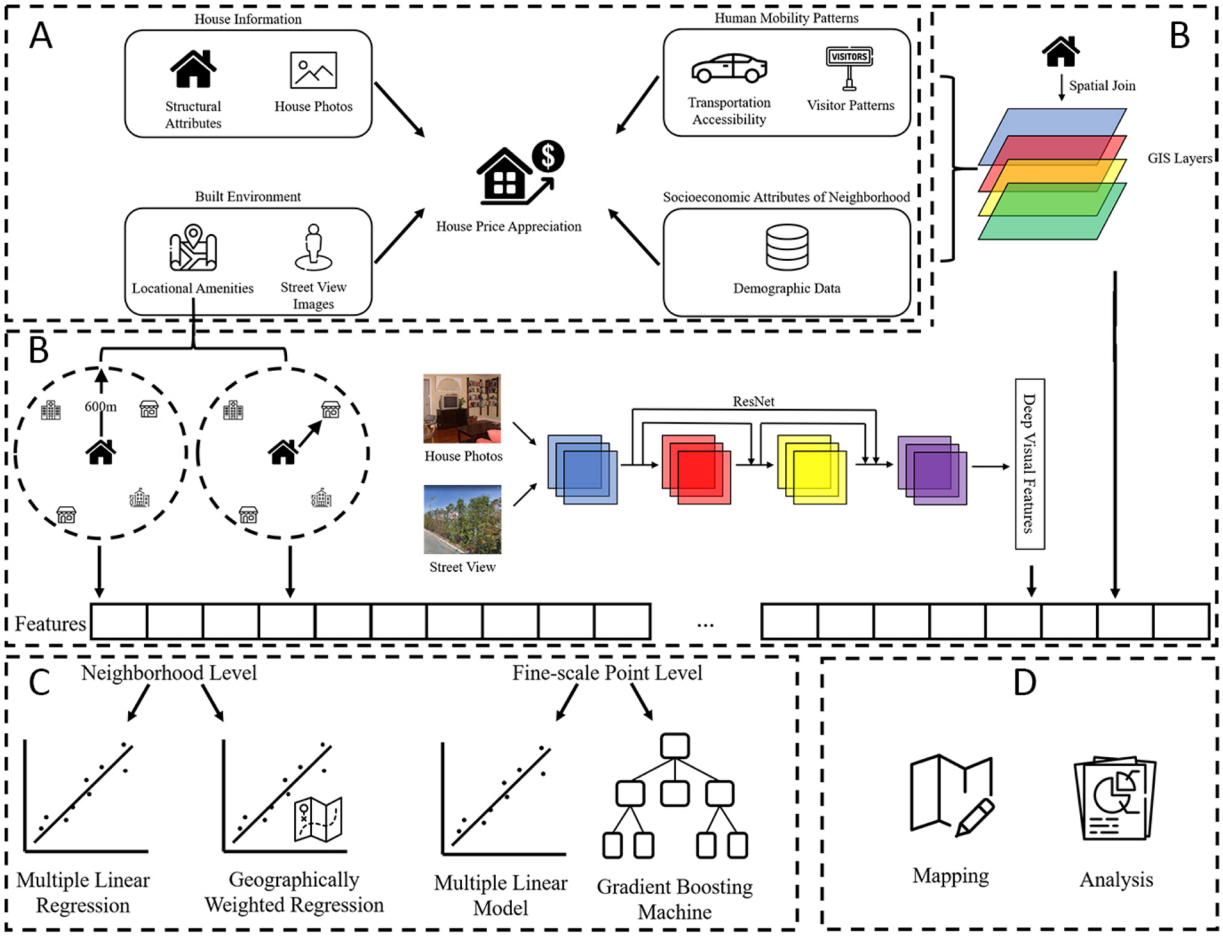

Fig. 1. The workflow of this study: (A) Data collection. (B) Feature construction. (C) Model training. (D) Mapping and analysis.

characteristics of the houses (Can, 1992). Locational amenities refer to geographical-related variables, such as the distance to the nearest facilities, which can reflect the intangible environment nearby (Chau and Chin, 2003). However, the house price appreciation rate might be affected by other variables such as the physical appearance of the house, surrounding physical and social environment settings, and dynamic human mobility patterns (Du et al., 2018). For example, houses with exquisite decoration worth higher values by intuition; houses located in districts and areas with a beautiful visual aesthetic environment, where residents' physical and mental health can be benefited, might have higher appreciation rate; and regions that can attract more visitors may have higher business values. However, due to the lack of quantitative measurements in conventional data collection methods, these key factors were overlooked by most of the previous studies.

The emergence of big data, high-performance computing, and advanced machine learning methods provide unprecedented opportunities to model those intangible assets of houses, which can enhance the estimation of house price appreciation rates. On one hand, in contrast to previous studies which used official statistical data and manual surveys in exploration of house price appreciation rates (Andrew and Meen, 2003; Crone and Voith, 1992; Archer et al., 1996; Quercia et al., 2000), larger volumes, velocities, varieties and veracities of geo-referenced data actively and passively produced by users bring more comprehensive insights into depicting socioeconomic environments in the era of volunteered geographic information (VGI) (Goodchild, 2007) and big geo-data (Gao et al., 2017b). For instance, house photographs that reflect indoor and outdoor scenery of properties, taken from the house owners and seller agents, are uploaded to online websites, which enable people to understand the scenery of houses; and street view images can describe the relationships between urban physical attributes and socioeconomic environments (Gebru et al., 2017; Zhang et al., 2018b; Zhang and Dong, 2018; Liu et al., 2019b). These two data sources make it possible to characterize the living environment from a human's perspective. Furthermore, the wide spread of GPS-embedded devices (e.g., mobile phones and vehicles), makes it possible to track individuals' trajectories to infer people's activities and movements. These dynamic observations of human movements may be taken as supplementary for locational amenities which only characterize the static geospatial aspects of houses. Intuitively, houses located in the areas with high accessibility to other places and higher attractiveness of others, may have higher price appreciation rate because of the travel convenience. A better understanding of the relationship between all these dimensions and house price appreciation rates can provide more comprehensive and valuable information for policy making to improve the overall quality of neighborhoods and stimulate social and economic balances between urban areas.

On the other hand, the development of state-of-the-art computer vision techniques enables us to extract high-level visual features from urban images. Capturing visual features to represent the scenic characteristics of houses as well as their neighborhood settings might help measure real estate appreciation values. In fact, recent works have shown the great potential of visual information in estimating house prices and in exploring culture and socioeconomic characteristics of neighborhoods (Gebru et al., 2017; You et al., 2017; Yao et al., 2018; Law et al., 2018; Fu et al., 2019; Liu et al., 2019a; Chen et al., 2020; Zhang et al., 2020). Accordingly, modeling house price appreciation rate with visual information is promising.

In this work, we propose a comprehensive multi-feature-fusion

framework using machine learning to model the house price appreciation rate. To build the framework, multiple data sources, including house information, built environment, human mobility patterns, and socioeconomic attributes of neighborhoods, are used to understand the value of urban settlements comprehensively. We take the Greater Boston Area as an example to test the feasibility of the proposed framework, and explore factors impacting on house price appreciation rates.

2. Framework

2.1. Overview

The framework is composed of four stages, namely data collection, feature construction, model training, and mapping and analysis (Fig. 1). First, we collect multi-source datasets, including the house information, built environment features, human mobility patterns, and socioeconomic attributes of neighborhoods on a cloud server. Second, by fusing the above datasets, we extract a series of features that are assumed to have an impact on price appreciation rates and use a multidimensional vector for representation. Then, algorithms including machine learning and geographically weighted regression (GWR) are built using the features constructed. The metrics are defined to measure the performance of those algorithms as well. Specifically, two spatial units (points and neighborhood) are tested in this research with different combinations of approaches. Finally, we aim to not only explore better ways for predicting house price appreciation rates, but also interpret the potential variables that are associated with the values of real estate appreciation.

2.2. Data collection

Four different categories of data are used in this study, namely house information, built environment, human mobility patterns, and socioeconomic attributes of the neighborhoods.

2.2.1. House information

House information consists of two subcategories: structural attributes and house photos. Both of them are collected from a popular online real estate website—REDFIN website.[1] House owners and seller agents post the information of their properties to the website for sale with an estimated price of each house provided by the system.

Structural attributes describe the basic characteristics of the house, including the location of the property, the number of bathrooms and bedrooms of the house, the built year, the number of floors and the size of the property, and the house type (single family residential, townhouse, etc.), which have been widely used in traditional hedonic pricing models (Rosen, 1974; Chau and Chin, 2003). Since our main focus is to predict the house price appreciation rates (i.e., price changes), the house prices across a five year period from February 2014 to February 2019, are retrieved. Accordingly, the appreciation rate R of a house with market price P is defined as follows:

$$R = \frac{P_{2019} - P_{2014}}{P_{2014}} \quad (1)$$

House photos are downloaded from the REDFIN website as another important part of the house information (Fig. 2). For each property, sellers upload photos taken by themselves to show the interior and exterior appearance of the house. Because the number of photos shared by sellers varies and not all properties have house photos available, we discarded those houses without photos. After that, the remaining houses with available photos are stored in order to extract meaningful high-level visual features to describe the house scenery.

2.2.2. Built environment

Two datasets are used to depict the built environment of a house: locational amenities and street view images.

Typically, locational amenities refer to the facilities near the house in the *hedonic* pricing model. Here, we use the point of interest (POI) information to show the location characteristics of nearby properties. The SafeGraph POI data[2] is used to provide the location information. Besides the location coordinates, each POI has a specific category code, which follows the standard criteria proposed by the North American Industry Classification System (NAICS).[3] In reference to the existing research (Cao et al., 2019), the following categories of POIs are chosen as illustrated in Table 1.

Street view images are downloaded by utilizing the Google Street View API[4] (Fig. 2). Street view images have been widely used to describe the physical settings of urban environment and neighborhoods, which can infer the relationship between human society activities and physical environment (Gebru et al., 2017; Zhang et al., 2018a; Chen et al., 2020). In order to retrieve the street view data along the roads, road networks are downloaded from the OpenStreetMap.[5] A set of georeferenced sampling points are generated along the roads with a fixed distance interval of 100 m. For each point, eight street view images are analyzed from different angles to show the surrounding urban environment comprehensively. It should be noted that not all street view images collected are used. Only those images within 50 m of each house are retrieved as descriptors to model the visual scenery of the housing built environment.

2.2.3. Human mobility patterns

There are two datasets used in this research to reflect the dynamic human mobility patterns: visitor patterns and transportation accessibility. Both are aggregated at the spatial resolution of Census Block Groups (CBGs).

The visitor patterns of CBGs are retrieved from the SafeGraph mobile phone database which covers about 10% of total population with mobile devices in the United States.[6] SafeGraph aggregates anonymized location data from numerous mobile applications in order to provide insights about physical places. To enhance privacy, SafeGraph excludes CBG information if fewer than five devices visited an establishment in a month from a given census block group. For each CBG, the records of aggregated visitor patterns illustrate how many visitors to the CBG during a specified time window, which could reflect the attractiveness of the CBG. The hourly visit counts are recorded as a 24-dimensional vector to show the dynamic patterns of visitors at CBGs.

The other dataset is released by the Uber Movement project.[7] This publicly available open data platform provides the observed travel times between two CBGs based on the movements of Uber vehicles. We calculate the mean travel time of each CBG to all other CBGs in the whole year of 2018. Note that the mean travel time may vary among CBGs, so the standard deviation of the travel time is also computed.

2.2.4. Socioeconomic attributes of neighborhoods

We used the CBG data released from the American Community Survey (ACS), which contains all kinds of demographic data. It is widely used in socioeconomic studies to estimate the neighborhood social identities at the CBG level. Specifically, we retrieved the population, ethnicity, income, and unemployment rate of the CBGs. Population of seven ethnicity groups as well as their ratio of each ethnicity are

[1] https://www.redfin.com/.

[2] https://www.safegraph.com/.

[3] https://www.naics.com.

[4] https://developers.google.com/maps/documentation/streetview/intro.

[5] https://www.openstreetmap.org/.

[6] https://www.safegraph.com/blog/what-about-bias-in-the-safegraph-dataset.

[7] https://movement.uber.com/?lang=en-US.

Fig. 2. Left: Study area. Red dots indicate the location of houses and blue polygons represent the boundary of census block groups. Middle: Examples of street view images. © 2019 Google. Right: Examples of house photos. © 2019 Redfin.

Table 1
POI categories with NAICS code.

NAICS code	Categories
445110	Grocery Store
452319	Stores
611110	School
611310	Universities
622110	Hospital
712190	Nature Parks
713110	Amusement Parks
722511	Full-Service Restaurants
722513	Limited-Service Restaurants
722515	Snack and Nonalcoholic Beverage Bars

recorded in each CBG. For each CBG, both the population and the ratio of each ethnicity are computed. In addition, the average income and average unemployment rate, which could reflect the identity and class of the neighborhood are also retrieved for further analysis.

2.3. Feature construction

Assume that each house appreciation rate r_i is influenced by a set of features from four perspectives: $r_i = F(h_i, b_i, m_i, s_i)$, in which h_i refers to house information, b_i refers to built environment, m_i refers to human mobility patterns, and s_i refers to socioeconomic attributes. In order to integrate these four types of factors to predict the house price appreciation rate, a set of features are extracted and constructed following the steps below.

2.3.1. General features

Structural attributes are constructed as features from the data source and attached to each house directly. It is worth noting that we use the natural logarithm of house price, which is normally distributed rather than the original values as they distributed skewed. Features of human mobility patterns and socioeconomic attributes of the neighborhood are attached to the houses after the spatial join operation between house locations and CBG polygons.

2.3.2. Locational features

For locational amenities, to better characterize the living convenience of neighborhoods thoroughly, we construct features from two aspects: the distance from the house to the nearest facilities, and the number of nearby amenities. The assumption is that different facilities may have different urban functions, which result in different mobility patterns of people and neighborhood vibrancy (Liu et al., 2012; Gao et al., 2017a; Yue et al., 2017). For example, people usually prefer to go to the nearest transportation hubs including metro and bus stops instead of farther away options. Thus, the distance to the nearest facility matters while the number of these transportation stations nearby may have limited impacts on people's travel mode. However, for amenities such as shops and restaurants, their quantity and variety of offerings influence the convenience to people living in a certain area. Therefore, we calculate the total number of these types of POIs. As suggested by studies from urban planning and geography (Neilson and Fowler, 1972; Murray and Wu, 2003), we select 600 m as the distance threshold for our distance analysis, which is suitable to represent the preferred coverage of human physical activity by walking, and is used to evaluate the walkability of the neighborhood convenience (Ellis et al., 2016). In other words, POIs in each category within 600m of house properties would be counted as the descriptors of a house.

2.3.3. Visual features

Furthermore, we extract deep features of street view imagery and house photos using a deep convolutional neural network (DCNN). The model is adapted from ResNet18, a commonly used architecture that has been proved efficiency in various computer vision tasks (He et al., 2016). It can extract high-dimensional visual features which can reveal hidden scenery information captured in photos. In order to learn efficient visual features from the images, we train the model with a house price prediction task. Accordingly, we take the images as model inputs and the house price value as the output. To deal with the skewed distribution (power law) of the house prices and accelerate the training process, we discretize the house price values into 10 levels and formulate the training as a 10-category classification task. A similar strategy was adopted in Zhang et al. (2019). The pre-trained model is then used to extract 512-dimensional features from each image, which

is considered as an efficient visual representation of the indoor/outdoor scene depicted in the image. Here, we only take the Greater Boston Area as a case study. The framework is also expected to be employed in other cities. With such a high-dimensional feature representation, the scenery of all collected photos can be represented comprehensively. We conduct the training process for house photos and street view images separately. Given the training process of the high-dimensional features (especially for the images) is time-consuming. Therefore, we adopt the principle component analysis (PCA) to reduce the feature dimension while preserving major feature characteristics. For each image, the first twenty components with about 60% of the total explained variance are maintained as the image feature. Please note that the major feature characteristics remained may vary across cities due to different urban environments and spatial dependency place by place. The first 20 components selected here can represent the visual scenery of built environment specifically in the Greater Boston Area only. Finally, we average the image features from multiple images associated with the same house (for both house photos and street view images).

2.4. Modeling algorithms

Two spatial analysis units are used in the experiment: fine-scale point level and aggregated neighborhood level. We assume that there are two kinds of target purposes using the proposed framework given the demand difference from two different groups. For house buyers and real estate industry, good machine learning models for individual house prices is more informative because of the high accuracy for appreciation estimation. The higher the model accuracy, the higher users' satisfaction is. Therefore, a fine-scale prediction of house value appreciation with advanced machine learning models is essential (Law et al., 2018; Hu et al., 2019). In comparison, economists, geographers, and policy makers are more interested in analyzing the macroscopical trend of house prices, and discuss the hidden economic and geographic factors influencing house price appreciation. The accuracy of results is not the only metric to consider when choosing the best model, while a macroscopic perspective of the real estate appreciation rate may be more helpful. The efficacy of geographically weighted regression (GWR) that could explain the spatial heterogeneity of variables in regression has been demonstrated in house price modeling (Cao et al., 2019; Wu et al., 2019; Liu et al., 2020). Therefore, spatially explicit models such as the GWR at the neighborhood-scale are favored.

2.4.1. Fine-scale level

At the fine-scale point level, all properties are treated equally with the entire set of the abovementioned features. We compare the multiple linear regression (MLR) approach with one machine learning approach—gradient boosting machine (GBM) with decision trees (Friedman, 2001)—to test the efficiency of the proposed framework. Although there are various machine learning methods, we only use the GBM as a representative machine learning model to make comparison with the linear regression model according to the following reasons: The accuracy and efficiency of GBM have been proved in various prediction tasks (Natekin and Knoll, 2013); And the main focus in this paper is to explore whether those extended data features can provide useful information for house price appreciation rate prediction, while not focusing on which machine learning algorithm performs the best. We conduct the k-fold cross-validations which split data into two parts: one is the training dataset and the other is the testing dataset, to mitigate overfitting problem in model training and prediction. The importance of each variable for GBM is also recorded to provide helpful suggestions for decision makings.

2.4.2. Neighborhood level

As for the neighborhood level (Fig. 3), the average values of all features for properties in one specific CBG are calculated as the feature set for the CBG. Ordinary least squares regression (OLS) and geographically weighted regression (GWR) are used to estimate the variables that influence house price appreciation rate. Compared with global regression model which ignores spatial non-stationary and only illustrates global impacts of variables, the GWR model constructs spatial relationships between independent and dependent variables with the following equation (Fotheringham et al., 2003):

$$R_i = \alpha_{0(u_i,v_i)} + \sum_{k=1}^{m} a_{k(u_i,v_i)} X_{k(u_i,v_i)} + \varepsilon_i \tag{2}$$

where R_i refers to the house price appreciation rate at location i, and the coordinate is (u_i, v_i); $\alpha_{0(u_i,v_i)}$ refers to the intercept parameter at location i; $a_{k(u_i,v_i)}$ refers to the local regression coefficient for the kth independent variable at location i; $X_{k(u_i,v_i)}$ refers to the kth attribute of location i; and ε_i indicates the random error. By using this model, the derived coefficients may vary across the research area and show the spatial heterogeneity of impact factors.

The main assumption in our study is that by embedding new data sources, including house photos, street view images, human mobility data and socioeconomic data, a model with better performance could be built. We expect such a model can achieve higher accuracy and provide better explanations of the reasons for house price appreciation values. Therefore, the traditional *hedonic* model (Rosen, 1974) fed with structural attributes and locational amenities only is considered as the baseline. New models fed with extended data sources are added respectively. Finally, a hybrid model using all data sources is also tested.

2.5. Evaluation

Two metrics are used for the evaluation of model performance, namely, the root mean square error (RMSE) and the coefficient of determination R^2. The RMSE is calculated as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(r_{0i} - r_i)^2} \tag{3}$$

where r_{0i} is the actual house appreciation rate and r_i is the predicted price appreciation rate of a house i. And the R^2 is calculated as follows:

$$R^2 = \frac{1}{m} * \sum_{i=1}^{m} \frac{(r_i - \bar{r})*(r_{0i} - \bar{r_0})}{\rho_r * \rho_{r0}} \tag{4}$$

where $\bar{r}$ and $\bar{r_0}$ refer to the average values of the predicted and the observed house price appreciation rates, and ρ_r and ρ_{r0} are the standard deviations of the predicted and the observed house price appreciation rate respectively.

3. Experiment and results

We take the Greater Boston Area as the study area. As shown in Fig. 2, the red dots represent houses (fine-scale) and the blue polygons are the CBGs (neighborhood-level). In this study, there are 21,928 houses with 125,000 house photos and about 470,000 street view images in total. All the house-related datasets are spatially aggregated into the 867 CBGs based on their point-in-polygon relationship (Fig. 3).

We train the machine learning models with multi-sources of data. The RMSE and R^2 with k-fold cross-validations are calculated to evaluate the model performance between the predicted and the actual value of house price appreciation rate. We conduct the experiments at the fine-scale and at the neighborhood-scale respectively.

3.1. Fine-scale house price appreciation estimation

At the fine-scale, we take each house as the basic unit, and conduct five experiments with different combinations of explanatory variables. The baseline experiment only takes house attributes and locational amenities as the explanatory variables. Then, four additional

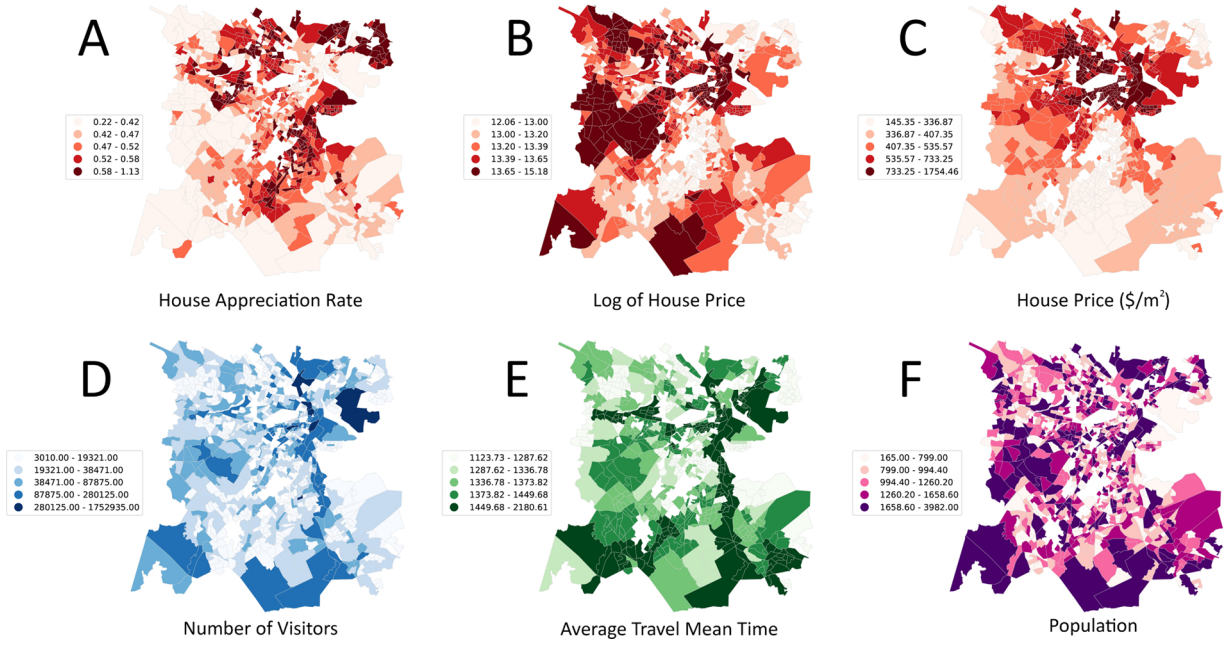

Fig. 3. Data distributions at census block group (CBG) level: (A) average house appreciation rates. (B) The natural logarithm of house prices. (C) Average house price per square meter. (D) Number of visitors to each CBG. (E) Averaged travel mean time to other CBGs. (F) Population.

experiments are conducted with house photos, street view images, human mobility patterns, and socioeconomic factors by feeding these features into each model step-by-step. Finally, we train the model using all variables.

Fig. 4 shows the scatter plots between the observed and the predicted house price appreciation rate. Table 2 illustrates the RMSE for all models. In general, the machine learning model using GBM (R^2 = 0.74; RMSE = 0.077) outperforms the MLR (R^2 = 0.48; RMSE = 0.103). This is expected, as the relationships between house price appreciation rate and features are not linear and the decision tree-based machine learning approach can better model non-linear relationships among variables. Results also show that combining multiple data sources indeed improves the performance of the models. In particular, the models that incorporate street view images got the lowest RMSE and improved the R^2 to a large extent. Most importantly, the model incorporating all the variables achieved the best performance. It proves that the four groups of variables characterize the appreciation value of a house from different perspectives and contribute differently to the variation of the house price appreciation rates.

Table 2
Model performance with *RMSE* in different combinations of data aspects using multiple linear regression (MLR) and gradient boosting machine (GBM) at fine-scale point level.

RMSE	MLR	GBM
Baseline	0.111	0.082
Baseline + house photos	0.110	0.081
Baseline + street view	0.106	0.079
Baseline + mobility data	0.107	0.080
Baseline + socioeconomic	0.109	0.080
All data sources	0.103	0.077

Moreover, we ask which variables contribute most to modeling the house price appreciation rate. The variable importance is calculated by the GBM. Fig. 5 ranks the top 20 variables of the model. In addition, we calculated the correlation coefficients between these variables and the house price appreciation rate to explore how these factors influencing house price appreciation rate. Fig. 6 shows the Pearson correlation

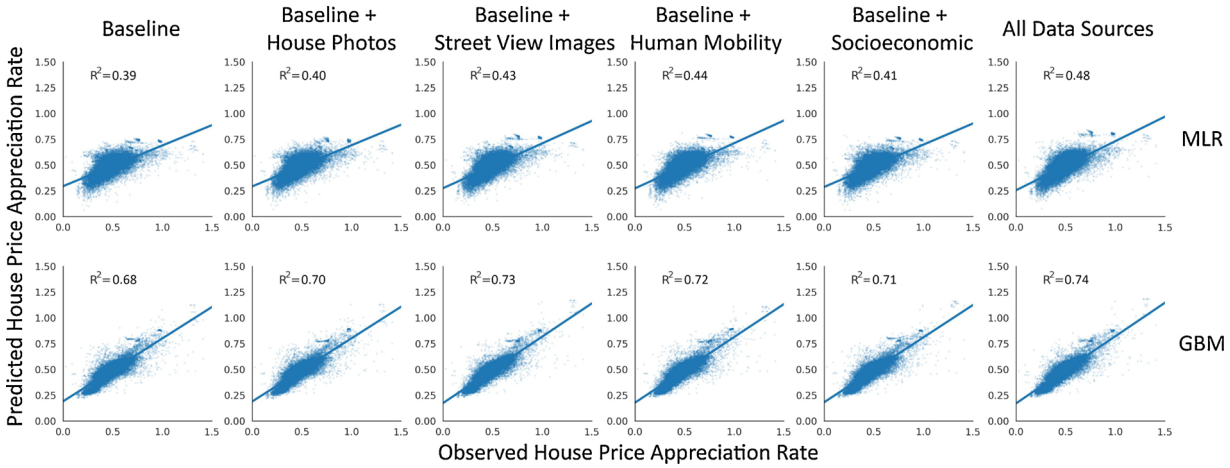

Fig. 4. Model performance with R^2 in different combinations of data sources using multiple linear model (MLR) and gradient boosting machine (GBM) at fine-scale point level.

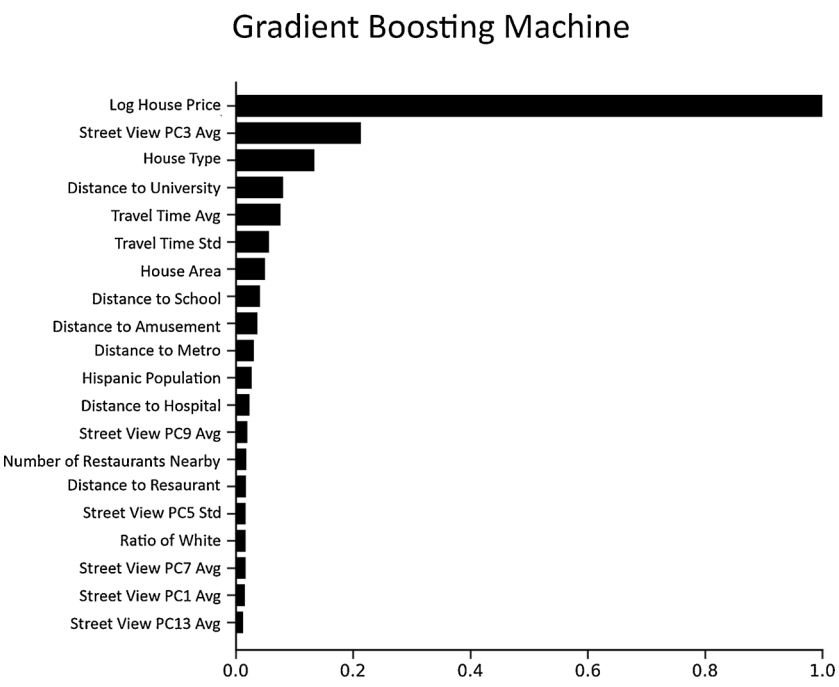

Fig. 5. Importance of top 20 variables using GBM with all data sources.

coefficients of several selected variables with *p*-values less than 0.01, which means that they are statistically significant.

The results show that the logarithm house price is the most important variable among all models with a correlation coefficient of −0.55, indicating that within the last five years, low-cost houses had a higher price appreciation in the Great Boston Area. The type of houses such as townhouse, single family house, etc., and the house area (with absolute correlation coefficient 0.49), also have great impacts on house price appreciation. It illustrates that structural attributes can influence not only house prices, as illustrated in the traditional hedonic pricing model, but also the house price appreciation rate. Moreover, we noticed that the street view image feature is one of the most important variables for all the models. Among them, the average values of the third component of visual features (represented as StreetView PC3 AVG), which has great contributions to the scenery captured by street view images, has moderate influence on house price appreciation rate with negative correlation at -0.30. In addition, the impacts of the ninth, fifth, seventh, first and thirteenth components of visual features also ranked in the top 20 among all variables. Though it is hard to explain the specific meaning of these visual features, it indeed indicates that high-level visual features could capture parts of important perspectives that are related to real estate appreciation rate. The results support our hypothesis that the detailed visual information of the house surrounding environment plays an important role in real estate appreciation evaluation as the street view images contain the overall environment of a neighborhood (Li et al., 2015; Gebru et al., 2017). Besides, for locational amenities, a house that is closer to a school (−0.13), amusement park (−0.21), metro station (−0.19), hospital (−0.09), or surrounded by more restaurants (0.01), may have a higher price appreciation rate. Similarly, the mean travel time that reflects the transportation convenience of a neighborhood is negatively correlated with house price appreciation rate, which means the less the mean travel time to other regions, the higher the appreciation rate. Interestingly, the study corroborates what is widely discussed in real estate studies: proximity to amenities matters, proximity to transportation hubs matters, shorter travel time matters, and physical quality of the surroundings matters.

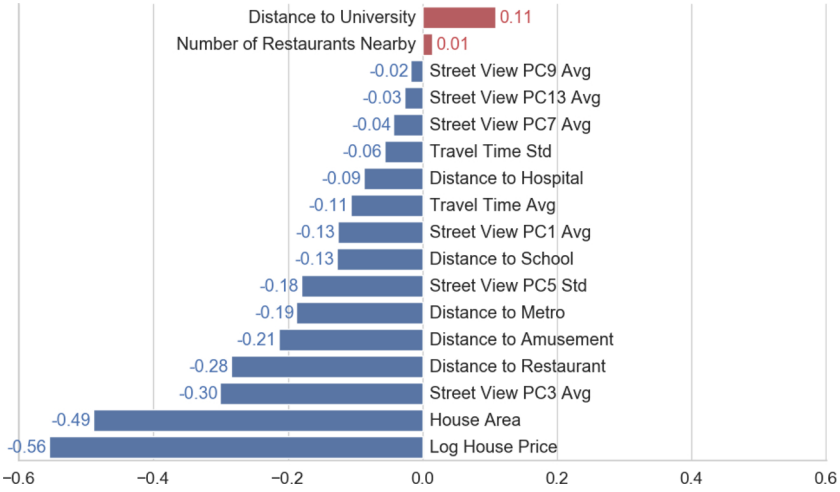

Fig. 6. Correlation coefficients of variables selected at fine-scale point level.

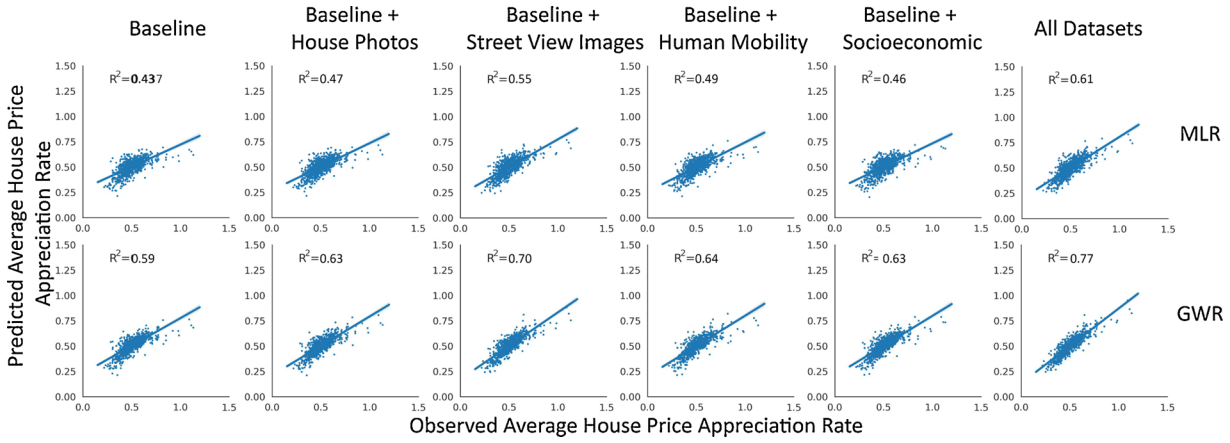

Fig. 7. Model performance with R^2 in different combinations of data sources using multiple linear regression (MLR) and geographically weighted regression (GWR) at aggregated-neighborhood level.

3.2. Neighborhood-scale house price appreciation estimation

The relationship between the independent variables and the house price appreciation rate may vary over space due to the spatially non-stationarity (Fotheringham et al., 2003). To investigate how spatial relationships change across the research area, we employed the geographically weighted regression (GWR) at the neighborhood-scale and compared the results with global multiple linear regression (MLR).

Fig. 7 shows the performance of the two models with data from various sources. Similar to the results at the fine-scale, the GWR achieves a better performance (R^2 = 0.774) than the MLR (R^2 = 0.608), which confirms the spatial heterogeneity of the study phenomenon over the research area.

Fig. 8 (A) shows that the coefficient of determination (R^2) of the GWR model is generally consistent across the study area. However, in the Boston downtown area, the R^2 is a little bit lower than the surrounding area (about 0.70 vs. 0.78). This indicates downtown area to be a more complex region which requires more latent factors that determine house appreciation rates. Fig. 7(B)–(F) depict several selected correlation coefficient distributions over the study area. Results show that the logarithm house price (changes from −0.30 to about −0.55) and the distance to metro (changes from −0.05 to about −0.25) have weak to moderate negative effects on the appreciation rate of house prices and such a relationship change spatially. In contrast, the effect of mean travel time (vary from about −0.05 to 0.125), distance to hospital (vary from about −0.12 to 0.04), and distance to university varied (vary from about −0.05 to 0.125) from negatively to positively across the study area. For instance, in the southeast region, the closer to a hospital, the lower the house appreciation rate (positive correlation of about 0.04). Whereas for other regions, a house price appreciation rate increase is associated with a decrease in the distance to hospital (negative coefficient of about −0.12). Results of the GWR model indicate that house price appreciation rates in Boston have spatial heterogeneous patterns. In other words, the coefficients of each variable and their impacts on house price appreciation rates vary across space and should be modeled place by place. Hence, it is necessary to explicitly embed spatial relationships for the predictive modeling of the house price appreciation rate.

3.3. Model and determinants analysis

Results of this study show promising findings in estimation of house price appreciation rate. We compare a series of models at two spatial scales, interpreting the results of these models, and explaining the

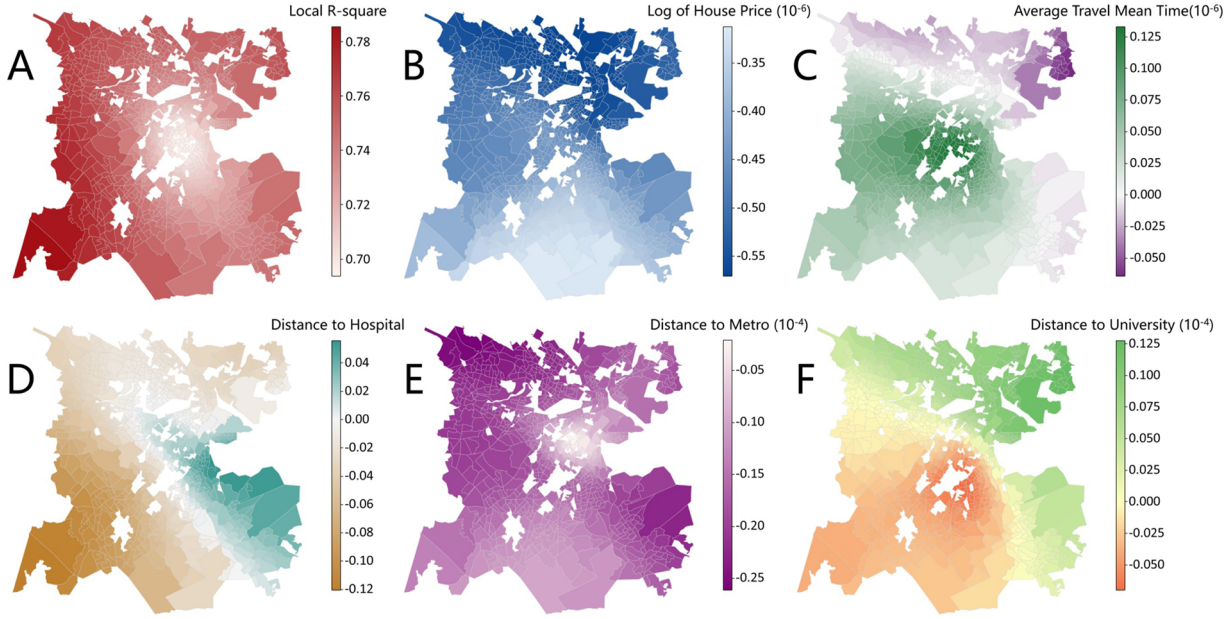

Fig. 8. Spatial distribution of GWR coefficients at the neighborhood scale.

spatial patterns of the house price appreciation rates. Different from the MLR which mostly models linear relationships, the decision tree-based machine learning method can build non-linear relationships between features and house price appreciation rate, and the GWR can model spatial non-stationarity between the variables, which indeed provide better prediction and a more holistic explanation.

We also examined the importance and the impacts of the factors related to house price appreciation at fine-scale and neighborhood-scale. The emerging sources of house photos, street view images, human mobility patterns and socioeconomic attributes, enable us to examine house price appreciation rate comprehensively from various aspects. Results show that by combining visual scenery of a house, built environment, dynamic human mobility patterns, and socioeconomic attributes of neighborhoods with machine learning approaches, the estimation accuracy of the price appreciation rate can be improved by a large margin. Among them, high-dimensional visual features extracted from street view images can provide important information related to house price appreciation rate. Such visual features capture intangible information which was not explored and discussed before. With better quantifying high-level semantic information from visual features, the procedure of policy making might be improved from these new insights.

In addition, several interesting findings are discovered in this study. For instance, at fine-scale, we found that houses with lower prices and small house area may have higher house appreciation potential. We dig into this discovery further and attempt to quantify and explain such a relationship.

As shown in Fig. 9, the logarithm of house prices in 2014 and the actual house price appreciation rates follow an exponential decay with $\lambda = -2.758$. It means that with the logarithm of house prices increases, the lower the rate of appreciation of house prices, and the slower the decay slope. The reasons might be traced from two aspects. On the one hand, a greater percentage of increment is not equal to a greater actual increment of prices. Therefore, houses may have larger price increment while less increased percentage. On the other hand, there are fewer houses with high prices compared with medium and low price houses. For those houses with high prices, since their prices have already been at a high level, there is limited room for house appreciations and thereby are more stable.

In addition, we conducted the correlation analysis between house area and house prices, and the result shows these two variables are significantly highly correlated with a coefficient of 0.62. Since houses with low prices typically have a small area, the correlation coefficient between house price appreciation rates and house area is thereby negative (correlation coefficient −0.49) as well. Therefore, compared with houses with high prices, those houses with low prices and small areas may have greater house price appreciation rates.

Besides, the more convenience the house with nearby facilities and higher transportation accessibility, the higher the house price appreciation rate is. At the neighborhood-scale, it shows that the spatial heterogeneity of variables exist and their influences to the distribution of the house price appreciation rate are different. Coefficients of variables such as distance to hospital, average travel mean time, even diverge from negative to positive. Therefore, it is necessary to model the spatial relationships between house price appreciation rate and these variables to better interpreting the underlying factors.

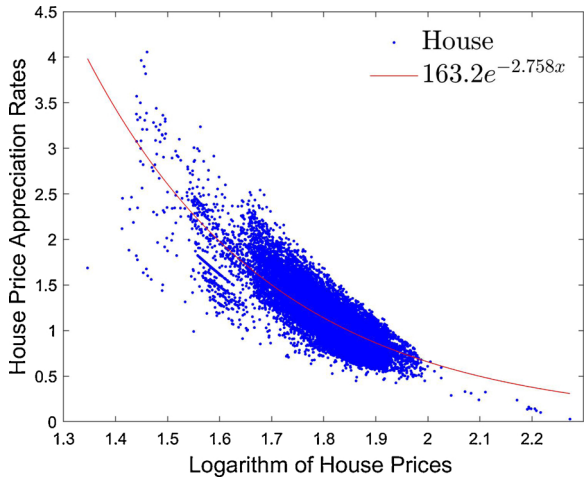

Fig. 9. Fit curve between logarithm of house prices (2014) and house price appreciation rates.

4. Discussion

4.1. Implications of policies

Understanding the variability and dynamic changes of house price appreciation rates are crucial for the government policy decision making. On the one hand, house price appreciation rates are closely related to various groups of people in cities, such as newly married couples, workers as labors, and youths who need school district housing. Hence, house price appreciation rate-related information can provide tutorials for their daily lives and house buying. It also helps the policy makers in planning housing development to fit with the job opportunities distribution and various educational facilities as well. On the other hand, the paper addresses the relationships between several factors and house prices, which might be helpful for sustainable city planning and urban infrastructure construction. The results and conclusions may help the government a more coordinated manner with empirical data support in urban development. In addition, the data-driven paradigm and advanced machine learning methods show potentials in providing insights for decision makings. For instance, street view images can be employed as a useful tool for urban environment observation and monitoring. The urban environment and neighborhood scenery can be captured comprehensively and processed efficiently with deep learning algorithms, which indeed will benefit people in cities. The data fusion of different aspects of big data also illustrates the data-driven paradigm in discovering and addressing the development of cities. Therefore, it is helpful for policy makers to understand the dynamics of housing appreciation rates when formulating housing policies.

4.2. Limitations and future directions

Here, we also discuss several limitations of this work that should be paid more attentions in future studies. Firstly, we only take one region (the Boston area) as a case study area. Since the built environment, human mobility patterns, and social class conditions of neighborhoods may vary in different cities, more regional factors can be taken into consideration in the future. For example, large vs. small, eastern vs. western, and coastal vs. inland cities can be considered to improve the generalization ability and replicability of the proposed framework for estimating house price appreciation rates.

Secondly, as we collected the house photos from VGI sources, the uncertainty of the data is a common concern for quality assurance. Models using house photos do not perform as good as the other three data sources, which might result from the low quality of sample data in capturing various scenery of a house. Instead, a better DCNN or other models that are not biased to samples and can differentiate complex visual features of different houses should be trained in order to improve the accuracy of the framework.

Thirdly, the urban renewal may occur in the past 5 years in Boston. Though the house price data are collected between 2014 and 2019, other attributes are limited to a specific time period. For example, POI data provided by SafeGraph was collected in the year 2018, only human mobility data in the year 2018 was collected from Uber Movement project, and the most recent street view images (may range from 2012

to 2018 for a specific place) were harvested, etc. However, the development of urban construction indeed has impacts on house price appreciation rates as well. These issues have not been addressed in this paper due to the restrictions of data sources. In the future, we expect to involve dynamics of urban land use changes into the framework with richer datasets.

Lastly, although we attempt to understand the patterns of house price appreciation rate, deeper exploration and more explanations could be added in future works. For example, the differences of appreciation rates between houses with different price ranges and in different geographic regions can be compared. Also, our framework focuses more on evaluating the value of several emerging data sources in house price appreciation and discovering the spatial distribution of house price appreciation rates with their spatial dependencies. While the causality relationships between variables and house price appreciation rates are also necessary for policy decision making. In the future, we will try to involve more time-series data and approaches from economy to build such relationships and improve the interpretability of deep learning models.

5. Conclusion

In summary, we present a multi-source-data-fusion framework to estimate the house price appreciation rates from various perspectives by the utilization of several big geo-data sources and the state-of-the-art machine learning approaches. Particularly, we extract high-level visual features from street view images and house photos to depict inner and outer appearance of houses using deep learning methods, which have certain impacts on house price appreciation rates.

This study offers insights into the potential of machine learning and spatial statistical approaches in modeling complex urban environments using multi-source geospatial big data. The contribution of the study is threefold: First, we propose to predict the house price appreciation rate, which differs notably from existing research for the absolute price estimation. Second, we build a big-data-driven multi-feature-fusion framework which utilizes various data sources from different aspects, especially with visual features extracted from house photos and street views, in order to enrich the knowledge of the house price appreciation modeling with state-of-the-art machine learning approaches at two spatial units. Third, we focus not only on improving the accuracy of a model, but also seeking explanations for what factors would influence the property value to provide suggestions for housing policies. Our research integrates computer science and social science research by utilizing advanced techniques with emerging data sources, and could provide new insights for researchers from economy, geography and urban planning towards future land-use studies.

Conflict of interest

None declared.

Acknowledgement

The funding support for this research is provided by the National Natural Science Foundation of China under Grant 41901321 and 41671378, the Office of Vice Chancellor for Research and Graduate Education at the University of Wisconsin-Madison with funding from the Wisconsin Alumni Research Foundation, and the Trewartha Research Award, Department of the Geography, University of Wisconsin-Madison. The authors would like to thank Timothy Prestby, UW-Madison for his generous help of proofreading. We thank Safegraph for providing anonymous mobile location data and POI visit patterns. The authors would also like to gratefully thank the members of the MIT Senseable City Lab Consortium: RATP, Dover Corporation, Teck Resources, Lab Campus, Anas S.p.A., Ford, SNCF Gares & Connexions, Brose, Allianz, ENEL Foundation, Laval, Curitiba, Stockholm, Amsterdam, Victoria State Government, KTH Royal Institute of Technology, UTEC—Universidad de Ingenierıa y Tecnologıa, Politecnico di Torino, Austrian Institute of Technology, Fraunhofer Institute, Kuwait-MIT Center for Natural Resources, SMART—Singapore-MIT Alliance for Research and Technology, and AMS Institute for supporting this research.

References

Andrew, M., Meen, G., 2003. House price appreciation, transactions and structural change in the British housing market: a macroeconomic perspective. Real Estate Econ. 31 (1), 99–116.
Archer, W.R., Gatzlaff, D.H., Ling, D.C., 1996. Measuring the importance of location in house price appreciation. J. Urban Econ. 40 (3), 334–353.
Can, A., 1992. Specification and estimation of hedonic housing price models. Reg. Sci. Urban Econ. 22 (3), 453–474.
Cao, K., Diao, M., Wu, B., 2019. A big data-based geographically weighted regression model for public housing prices: a case study in Singapore. Ann. Am. Assoc. Geograph. 109 (1), 173–186.
Chau, K.W., Chin, T., 2003. A critical review of literature on the hedonic price model. Int. J. Housing Sci. Appl. 27 (2), 145–165.
Chen, L., Yao, X., Liu, Y., Zhu, Y., Chen, W., Zhao, X., Chi, T., 2020. Measuring impacts of urban environmental elements on housing prices based on multisource data – a case study of Shanghai, China. ISPRS Int. J. Geo-Inform. 9 (2), 106.
Chen, M., Liu, W., Lu, D., 2016. Challenges and the way forward in China's new-type urbanization. Land Use Policy 55, 334–339.
Crone, T.M., Voith, R.P., 1992. Estimating house price appreciation: a comparison of methods. J. Housing Econ. 2 (4), 324–338.
Du, Q., Wu, C., Ye, X., Ren, F., Lin, Y., 2018. Evaluating the effects of landscape on housing prices in urban China. Tijdsch. Econ. Soc. Geogr. 109 (4), 525–541.
Ellis, G., Hunter, R., Tully, M.A., Donnelly, M., Kelleher, L., Kee, F., 2016. Connectivity and physical activity: using footpath networks to measure the walkability of built environments. Environ. Plann. B: Plann. Des. 43 (1), 130–151.
Fotheringham, A.S., Brunsdon, C., Charlton, M., 2003. Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. John Wiley & Sons.
Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 1189–1232.
Fu, X., Jia, T., Zhang, X., Li, S., Zhang, Y., 2019. Do street-level scene perceptions affect housing prices in Chinese megacities? an analysis using open access datasets and deep learning. PLOS ONE 14 (5), e0217505.
Fu, Y., Xiong, H., Ge, Y., Zheng, Y., Yao, Z., Zhou, Z.H., 2016. Modeling of geographic dependencies for real estate ranking. ACM Trans. Knowl. Discov. Data 11 (1), 11.
Gao, S., Janowicz, K., Couclelis, H., 2017a. Extracting urban functional regions from points of interest and human activities on location-based social networks. Trans. GIS 21 (3), 446–467.
Gao, S., Li, L., Li, W., Janowicz, K., Zhang, Y., 2017b. Constructing gazetteers from volunteered big geo-data based on hadoop. Comput. Environ. Urban Syst. 61, 172–186.
Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Aiden, E.L., Fei-Fei, L., 2017. Using deep learning and Google street view to estimate the demographic makeup of neighborhoods across the united states. Proc. Natl. Acad. Sci. U.S.A. 114 (50), 13108–13113.
Goodchild, M.F., 2007. Citizens as sensors: the world of volunteered geography. GeoJournal 69 (4), 211–221.
He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 770–778.
Hu, L., He, S., Han, Z., Xiao, H., Su, S., Weng, M., Cai, Z., 2019. Monitoring housing rental prices based on social media: an integrated approach of machine-learning algorithms and hedonic modeling to inform equitable housing policies. Land Use Policy 82, 657–673.
Hung, S.Y., Tu, C., 2008. An examination of housing price appreciation in California and the impact of alternative mortgage instruments. J. Housing Res. 17 (1), 33–47.
Law, S., Paige, B., Russell, C., 2018. Take a Look Around: Using Street View and Satellite Images to Estimate House Prices. arXiv:180707155.
Li, X., Zhang, C., Li, W., Kuzovkina, Y.A., Weiner, D., 2015. Who lives in greener neighborhoods? the distribution of street greenery and its association with residents' socioeconomic conditions in Hartford, Connecticut, USA. Urban Forest. Urban Green. 14 (4), 751–759.
Liu, F., Min, M., Zhao, K., Hu, W., 2020. Spatial-temporal variation in the impacts of urban infrastructure on housing prices in Wuhan, China. Sustainability 12 (3), 1281.
Liu, X., Andris, C., Huang, Z., Rahimi, S., 2019a. Inside 50,000 living rooms: an assessment of global residential ornamentation using transfer learning. EPJ Data Sci. 8 (1), 4.
Liu, Y., Wang, F., Xiao, Y., Gao, S., 2012. Urban land uses and traffic 'source-sink areas': evidence from GPS-enabled taxi data in Shanghai. Landsc. Urban Plann. 106 (1), 73–87.
Liu, Z., Yang, A., Gao, M., Jiang, H., Kang, Y., Zhang, F., Fei, T., 2019b. Towards feasibility of photovoltaic road for urban traffic-solar energy estimation using street view image. J. Clean. Prod. 228, 303–318.
Livy, M.R., 2017. The effect of local amenities on house price appreciation amid market shocks: the case of school quality. J. Housing Econ. 36, 62–72.
Murray, A.T., Wu, X., 2003. Accessibility tradeoffs in public transit planning. J. Geograph. Syst. 5 (1), 93–107.

Natekin, A., Knoll, A., 2013. Gradient boosting machines, a tutorial. Front. Neurorobot. 7, 21.
Neilson, G.K., Fowler, W.K., 1972. Relation between transit ridership and walking distances in a low-density Florida retirement area. Highway Res. Rec.(403).
Quercia, R., McCarthy, G., Ryznar, R., Can Talen, A., 2000. Spatio-temporal measurement of house price appreciation in underserved areas. J. Housing Res. 11 (1), 1–28.
Rosen, S., 1974. Hedonic prices and implicit markets: product differentiation in pure competition. J. Pol. Econ. 82 (1), 34–55.
Wu, C., Ren, F., Hu, W., Du, Q., 2019. Multiscale geographically and temporally weighted regression: exploring the spatiotemporal determinants of housing prices. Int. J. Geograph. Inform. Sci. 33 (3), 489–511.
Yao, Y., Zhang, J., Hong, Y., Liang, H., He, J., 2018. Mapping fine-scale urban housing prices by fusing remotely sensed imagery and social media data. Trans. GIS 22 (2), 561–581.
You, Q., Pang, R., Cao, L., Luo, J., 2017. Image-based appraisal of real estate properties. IEEE Trans. Multimedia 19 (12), 2751–2759.
Yue, Y., Zhuang, Y., Yeh, A.G., Xie, J.Y., Ma, C.L., Li, Q.Q., 2017. Measurements of poi-based mixed use and their relationships with neighbourhood vibrancy. Int. J. Geograph. Inform. Sci. 31 (4), 658–675.
Zhang, F., Wu, L., Zhu, D., Liu, Y., 2019. Social sensing from street-level imagery: a case study in learning spatio-temporal urban mobility patterns. ISPRS J. Photogram. Rem. Sens. 153, 48–58.
Zhang, F., Zhang, D., Liu, Y., Lin, H., 2018a. Representing place locales using scene elements. Comput. Environ. Urban Syst. 71, 153–164.
Zhang, F., Zhou, B., Liu, L., Liu, Y., Fung, H.H., Lin, H., Ratti, C., 2018b. Measuring human perceptions of a large-scale urban region using machine learning. Landsc. Urban Plann. 180, 148–160.
Zhang, Y., Dong, R., 2018. Impacts of street-visible greenery on housing prices: evidence from a hedonic price model and a massive street view image dataset in Beijing. ISPRS Int. J. Geo-Inform. 7 (3), 104.
Zhang, F., Zu, J., Hu, M., Zhu, D., Kang, Y., Gao, S., Zhang, Y., Huang, Z., 2020. Uncovering inconspicuous places using social media check-ins and street view images. Comput. Environ. Urban Syst. 81 p.101478.