Sagi, Alon; Gal, Avigdor; Czamanski, Daniel Z.; Broitman, Dani

Article

Uncovering the shape of neighborhoods: Harnessing data analytics for a smart governance of urban areas

Journal of Urban Management

Provided in Cooperation with:
Chinese Association of Urban Management (CAUM), Taipei

Suggested Citation: Sagi, Alon; Gal, Avigdor; Czamanski, Daniel Z.; Broitman, Dani (2022) : Uncovering the shape of neighborhoods: Harnessing data analytics for a smart governance of urban areas, Journal of Urban Management, ISSN 2226-5856, Elsevier, Amsterdam, Vol. 11, Iss. 2, pp. 178-187, https://doi.org/10.1016/j.jum.2022.05.005

This Version is available at:
https://hdl.handle.net/10419/271457

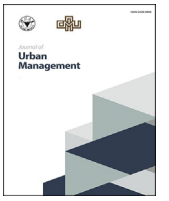
Journal of Urban Management 11 (2022) 178–187

HOSTED BY

Contents lists available at ScienceDirect

Journal of Urban Management

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



Research Article

Uncovering the shape of neighborhoods: Harnessing data analytics for a smart governance of urban areas

Alon Sagi [a], Avigdor Gal [a], Daniel Czamanski [b], Dani Broitman [a,*]

[a] Technion - Israel Institute of Technology, Israel
[b] Rupin Academic Center, Israel

A B S T R A C T

Urban scholars have made great advances to understand the reciprocal relations between households and their immediate environments as a means for the creation of efficient urban administrative systems. However, from an urban management perspective, reliance on geographical areas fixed for long periods of time as basic data collection constitutes a problem. Modern urban areas are in a permanent state of flux because of changing preferences, willingness to pay, location choices, and physical development. In this constantly changing context, what is the most appropriate delimitation of a "neighborhood", defined as a small and relatively homogeneous area in a certain (and temporary) urban configuration? This paper contributes to the growing literature on the use of data analytic tools in urban studies and neighborhood delimitation in housing sub-markets, exploiting big data on real-estate transactions in England and Wales during a long period of time. The results shed light on the importance of organic urban features and the drawbacks of rigid geometric definitions. They also highlight the importance of the usage of deep Machine Learning (ML) tools such as Artificial Neural Network (ANN), alongside with traditional methods. The paper's contribution to urban governance is the suggestion of a smart and dynamic system aimed at defining the most appropriate areas for urban management given a specific period and situation. The suggested framework can be implemented periodically, helping to define homogeneous spatial units (neighborhoods) with large variances among them, allowing for designing urban policies tailored to each one of them.

1. Introduction

Academic researchers, as well as the private and governmental sector practitioners, are trying to analyze and predict constantly housing prices in more accurate and efficient ways. A popular approach is the 'hedonic' method (Rosen, 1974) that estimate the effects of local housing characteristics on housing prices by means of multiple regressions.

Other traditional pricing model is the 'repeated sales index' (Bailey et al., 1963; Case & Shiller, 1987) that link the fluctuations of the housing market over time to macro-economic variables. Unlike the "analyze and interpret" approach of the traditional statistical models, the Machine Learning (ML) approach advocates a "learn and predict" approach: Using statistical algorithms (sometimes the same as in the traditional methods) ML aims to learn patterns and behavior from the existing data samples and then predict the value of a new-coming sample (Bzdok et al., 2018). In our case, learning from large real-estate transactions datasets we intend to predict housing prices. To assess the ability of a ML model to evaluate values, the common practice is to divide the data into a train set, from which the model 'learns', and a validation set from which its predictions can be tested. As will be described in the next section, much of the current ML research regarding housing prices is focused on improving the accuracy of the price estimates. This is done using more advanced ML algorithms, larger or more detailed datasets, and a better delimitation of the data.

At the same time governments collect vast amounts of data organized at various small geographic units. Usually, data are available for administrative geographic units such as "census tracks", "neighborhoods", and "wards". There is lack of agreement regarding the

* Corresponding author. Faculty of Architecture and Town Planning, Technion – Israel Institute of Technology, Haifa, Israel.
E-mail address: danib@technion.ac.il (D. Broitman).

https://doi.org/10.1016/j.jum.2022.05.005
Received 15 December 2021; Received in revised form 5 May 2022; Accepted 5 May 2022
Available online 14 May 2022

geographic boundaries of these fundamental units of study (Shearmur, 2015). What is the best delimitation of a "neighborhood", defined as a small and relatively homogeneous area? What is the smallest geographic unit appropriate for studying urban dynamics? The available data make it difficult to study modern urban areas that are in a permanent state of flux because of changing preferences, willingness to pay, location choices, and physical development (Weaver, 2014).

This paper contributes to the growing literature on housing price estimates using ML methods, data analytic tools and exploiting big data on real-estate transactions in England and Wales during a long period of time. Our dataset contains all the housing transactions that took place in England and Wales between 1995 and 2018. It was collected by the UK government and made available free for public use. However, we take advantage of this large-scale dataset to go beyond finding more accurate housing price prediction algorithms. By using the large-scale geographic coverage, we were able to test different housing price prediction models and to test the homogeneity of various submarkets.

To this end, we compare the results of three statistical and ML algorithms: the "classic" Hedonic regression, random forest, and artificial neural network (ANN). The application of these methods makes it possible to examine several different administrative spatial divisions, from the regional to the neighborhood level. We tested the homogeneity of the housing submarkets in the context of five divisions and a total of 23,200 subdivisions: regions, local authorities, Wards, census statistical areas (MSOA) and a neighborhood scale, density-based random rectangles. Rural, urban, and suburban housing submarkets display behavior consistent with their characteristics. After testing for the most accurate prediction algorithm and the most homogenous spatial division, we examined these urban building blocks for different patterns of results.

The main contribution of this paper to the research concerning urban governance is the proposal of a smart and dynamic system aimed at defining the most appropriate areas for urban management. The aim of our suggested framework is to choose the most appropriate division of neighborhoods for urban governance and management, from a given set of scenarios. The guiding principle is to select the scenario that creates the most homogeneous spatial units (neighborhoods) with the largest variances among the units. This differentiation allows potentially for the designing urban policies tailored to the specific characteristics of each one of the neighborhoods. In addition, this study evidences the advantages of comparing traditional housing market methods of analysis (as linear regressions) with recent and more updated ML tools, particularly Artificial Neural Networks (ANN). In this respect, this paper is part of the recent trend of studies that combine concepts, methods, and data, including ANN, in urban studies (Grekousis, 2019) and other domains (Angrist & Frandsen, 2022).

The rest of this paper consists of 5 sections. Section 2 includes a literature review. In section 3 we present the methods of analysis we utilized. The data are described in section 4. In section 5 we present the results of our analysis. Some conclusions are presented in section 6.

2. Literature review

In recent years there is growing availability of very detailed, fine grained, geographic, and temporal, urban data. These include data concerning real estate transactions, social media participation information, data concerning private and public traffic, and communication data from information and communication (ICT) devices. Efforts to extract useful information about urban dynamics from these big data led to the application of novel ML algorithms, including ANN and other deep learning methods. ML algorithms are used to study gentrification forecasting (Reades et al., 2018), neighborhood location analysis (Kauko et al., 2002), and street facade analysis (Naik et al., 2017). Earlier applications of deep ML algorithms were used to study housing markets. Specht (1991) used ANN to evaluate housing prices. In most cases, the baseline for the evaluation of various of ML housing price prediction algorithms, have been the traditional Hedonic methods, or the 'Case-Shiller index' based on repeated sales data (Caplin et al., 2008; Chau & Chin, 2002; Weigand, 2019).

Many ML methods have were proven suitable for complex pattern recognition and for non-linear systems forecasting (Abidoye & Chan, 2017; Kauko et al., 2002), and are not limited by weaknesses of the hedonic linear regressions (Xiao, 2017, pp. 11–40). Some of these models and algorithms that have shown satisfactory results are random forest or regression trees and gradient boosting methods (Jha et al., 2020; Shahhosseini et al., 2020; Truong et al., 2020; Yan & Zong, 2020); lasso or ridge regressions (Jha et al., 2020; Shahhosseini et al., 2020; Yan & Zong, 2020); support vector machine (Baldominos et al., 2018; Chen et al., 2017; Mu et al., 2014; Shahhosseini et al., 2020); and other algorithms (Barr et al., 2017; Park & Bae, 2015). The most notable prediction algorithm used to replace the hedonic method is ANN. In 2017 (Abidoye & Chan) found that between 1991 and 2015, thirty-one academic papers were published that used ANN to evaluate housing prices. The results of most of these studies showed that ANN is an effective method for housing price estimation. Out of twenty-one studies comparing ANN and the hedonic method, in sixteen the ANN was found to be more accurate, supporting the validity of ANN. But despite a growing list of papers and algorithms, there is no agreement concerning the most accurate method. Comparison of various contributions, among them several ML and deep learning analyses, concluded that no algorithm consistently gives the best results (Baldominos et al., 2018; Santibanez et al., 2015; Shahhosseini et al., 2020).

It is noteworthy that while many ML and particularly ANN, are considered as tools for the analysis of big data, most of applications motivated to improve the accuracy of the traditional methods utilize relatively small housing datasets and geographic coverage. Thus, out of the thirty-one papers which Abidoye and Chan reviewed, most of them (17) used less than 1000 samples, eight used 1000 to 10,000 samples and four used 10,000 to less than 50,000 samples. Few other contributions used much larger datasets for housing price predictions. For example, Truong et al. (2020) used 231,962 samples from Beijing to test several ML methods. Barr et al. (2017) used 1.4 million samples from Los Angeles and the Gradient Boosting algorithm. Caplin et al. (2008) used over 1.5 million samples, also from Los Angeles and combined ANN and regression models. Ng (2015) used 2.4 million samples from London using a Gaussian process method.

In order to improve their results, most of the housing price analysis, research, both traditional and ML, used a property location

variable to define smaller and more homogenous submarkets. Studies used identifiers of census statistical areas (Harrison & Rubinfeld, 1978; Oladunni & Sharma, 2016); zip codes or postal codes (Baldominos et al., 2018; Feng & Humphreys, 2018; Jha et al., 2020; Park & Bae, 2015; Santibanez et al., 2015); and district or county location (Chen et al., 2017; Lam et al., 2008; Nguyen & Cripps, 2001; Truong et al., 2020; Zhou et al., 2019). Some researchers suggest that more accurate predictions might be achieved with different approaches. Thus, Caplin et al. (2008) used a set of 1,550,451 observations and for each prediction their algorithm created 150 sub-sets of the nearest observations in space and transaction date as a basis for multiple regressions. Barr et al. (2017) used over 1,400,000 observations from the Los Angeles metropolitan area and divided the data into 275 subsets by means of zip codes. Then they trained 275 gradient boosted regression tree models, for housing price prediction and tested the mean error for each zip code. In that way the authors could not only reduce the model's error by creating a more homogenous sub market, but to show on a map which of the zip codes had lower or higher prediction accuracy. The authors suggested that areas with lower accuracy were due to the 'heterogeneity of income and variables.

3. Methodology

Our analyses include multiple runs utilizing three regression algorithms to assess housing price: Hedonic regressions, random forest algorithm, and a feed forward ANN.

The hedonic method, based on a multiple linear regression, relies on a 'hedonic price' for each variable. This is the price that the public is willing to pay for one unit of a housing characteristic (Rosen, 1974). For instance, if the average price of new apartment in the training-set is 10,000 £ more than a secondhand apartment, the model sets the hedonic price for new apartments at that price. It means that the model predicts the price of each new apartment 10,000 £ more than a similar second-hand apartment, with this being the only difference.

The basic element of **the random forest algorithm** is a decision tree. A decision tree is a statistic method that attempts to split the observations of the training-set at each step into two homogenous groups by its labels (price, in our case). After some steps the samples 'fall' into very small and homogenous groups (leaves of the tree). This way, it is possible to predict the label of a new sample. Since decision trees tend to overfit (getting a low error rate over the learning-set but high over the validation-set), the random forest algorithm creates multiple decision trees, each one containing only some of the variables, selected randomly. This method results are much more robust (Breiman, 2001).

The structure of all simple feed-forward and fully connected **ANNs** for price prediction (or any other numerical prediction) is similar. The input layer gets 'neurons' as the number of variables of the data samples, the sample vector in the input layer is multiplied by 'weights' in several layers with several neurons, each, and eventually goes out as a price prediction at the 'output layer'. The prediction is compared with the real price and in a process call 'backpropagation', the weights are adjusted accordingly. Repeating this process many times bring the hidden layers' weights to minimize the error between the prediction and the real price (Mora-esperanza, 2004). The ANN sometimes referred as a 'black box' since, unlike in the hedonic method or a decision tree, the correlation between the variables and the output cannot be traced. And this is also one of the reasons why it is hard to determine parameters as the size or number of hidden layers and it is usually set in a process of trial and error (Zhang et al., 1998).

As Barr et al. (2017) did in their research, we also group our samples by geographic subdivisions and for each subdivision run a separate evaluation model. However, unlike them, we repeat this process over various geographic divisions and various algorithms. Specifically, from each sub-division, out of the 1995–2018 transactions, we randomly split the samples into train-set and validation-set by an eighty-to-twenty ratio and utilized three methods. For instance, all the transactions in Ward ID E36000001 between 1995 and 2018 were split to train-set and validation-set, and used as input for the linear regression, random forest and ANN. This was followed by the same process for the transactions in Ward ID E36000002, and so forth. After all the Wards has been 'learned' from the train-set, and the transactions in the validation-set were evaluated by the models, the similarity of the evaluated results to the real transactions was evaluated. The same sequence was done with the MSOA and the other divisions. This stratification strategy was performed to keep the consistency between the different analytical methods at all spatial levels.

We chose each run of the assessment process to return five common statistical measurements to evaluate a model's precision: $\mathbf{R^2}$ indicates the ability of the variables to explain the variance in the transaction's price. **Adjusted $\mathbf{R^2}$** adjusts the R^2 result with the number of explanatory variables. **Less than ten percent error** indicates the proportion of transactions' predictions that were accurate within ten percent of the real price. **The median error** indicates the error in percentage in which half of the transaction was more accurate. And **Root Mean Square Error (RMSE)**- indicates the standard deviation of the error.

For the random forest we chose thirteen for the trees' depth and two for the number of trees. For the neural network model we used a ten-neuron input layer, two hidden layers of ten neurons each and an output layer of one neuron for a continuous numerical output.

4. Data

Naturally, the most important element in ML-based research is the data it uses. The more and detailed data the models are fed, the more accurate results the algorithms will yield. But collecting and producing enough data on real-estate transactions can be expensive and time-consuming. Luckily, in the last decade, governments and companies around the world have started to publish their data freely for public use, providing huge data resources that are accessible and easy to use. Regarding data size, a methodological contribution of the present research is the use of a large and geographically broad dataset, compared with smaller sample sizes used in previous studies

that were mentioned in the literature review.

The main database in the current research is the housing price database, which 'includes information on all property sales in England and Wales that are sold for full market value and are lodged with us for registration'.[1] 23,554,383 transactions are included in the dataset since the beginning of 1995 to July 2018, and each transaction contains the following data features who are relevant to our research:

- Price- Natural logarithm of the sale price as stated on the transfer deed.
- Date of Transfer- Date when the sale was completed.
- Property Type-i.e. Detached, Semi-Detached, Terraced, Flats/Maisonettes or Another property type.
- Old/New-an established residential building or a newly built property.
- Duration- Relates to the tenure: Freehold, Leasehold, etc.

To get robust samples, we eliminated all the transactions labeled 'non-standard price paid category type', resembling price paid entry as the sale of a whole building or of nursing home units, etc. which also eliminated all the samples labeled as house type 'Others'. Eventually our full database contains 23,036,165 valid samples.

The comprehensive record of residential transactions used in this study has also its shortcomings, the most important being the lack of total floor size information. Despite that, it provides the most accurate source about property sales and is used for official statistics calculations (South & Henretty, 2017). Recently there were attempts to add the total floor size by linking the dataset with alternative sources (Chi et al., 2021), but either there were methodological issues (Orford, 2010) or the resulting time span is relatively short (Chi et al., 2021). The second crucial dataset for this research is the geographical divisions used to test the housing price predictions. In total there are 23,300 subdivisions in five division scales. Four are generated by the British Office for National Statistics (ONS) and provided by data.gov.uk. and the fifth was generated by a GIS program:

- Regions: this is the largest sub-division in the U.K. It contains the nine regions of England and Wales as the tenth region.
- Local Authorities Districts: contains the borders of all 352 local government administrations.
- Middle-Layer Super Output Areas (MSOA): the MSOA created for statistical purposes for the censuses of England and Wales. It is a neighborhood scale division it which contains 7201 sub-divisions with an average population of 7200 per area, aims to provide areas consistent with population size and socio-economic characteristics.
- Census Merged Wards (Wards): Wards are the 'sub-division of a local authority drawn up for electoral purposes' usually based on historical boundaries between neighborhoods. Unlike the MSOAs, wards vary in population size. The layer contains 8546 areas completely covering all of England and Wales.[2]

Regions and Local Authority districts are shown in Fig. 1(a). As can be seen from Fig. 1(b) and (c), in urban areas, both the ward and the MSOA present a good division of a neighborhood or sub-neighborhood level. They are small enough to easily cross them by foot, and the borders are based on physical features (streets, parks, river, etc.), that also often coincides with a change in building characteristics or period. Naturally, in small towns or rural areas, where population density is low, the size of the wards and MSOAs are bigger. Towns smaller than a few thousand people are usually included in just one ward and MSOA. Other wards and MSOAs can contain several villages populated with hundreds of people each. While it may be strange to think of a cluster of villages as a 'neighborhood', these villages act as one homogenous sub-market which, for the sake of this research, may be considered a neighborhood.

Past research suggesting that the smaller a housing area, the more homogeneous it is, and the more accurate price prediction models are (Dark & Bram, 2016). Therefore, it might be expected that one of the neighborhood scale division's predictions is the most accurate. However, we believe that not only the small spatial scale determines the homogeneity of a housing area, but also the neighborhoods' borders which been defined by a change in physical housing or in socio-economical characteristics. To test this belief another 'pseudo' neighborhood scale division as a control group:

- Density-based random rectangles: a GIS layer was created via the 'Create Vector Tile' tool in ESRI GIS Pro. This tool tiled the geographic area with rectangles whose size changes according to the density of the wards and MSOA layers. This gives us a layer with roughly the same number of sub-divisions as the neighborhood scale divisions (7,091), but with random borders that do not represent physical attribute such as a street or river (see Fig. 1(b) and (c)).

5. Results

Table 1 presents the average values obtained for all the components of each territorial sub-division using each of the algorithms: linear regression, random forest, and ANN.

The first noticeable information from the table is that different algorithms perform differently at different spatial scales. At the large regional and the local authority scale the random forest gives the most accurate results, but at the neighborhood scale the ANN performs better (results highlighted in bold in Table 1). Compared to both, and at all scales, the linear regression's results are the least accurate. In past research different algorithms gave the most accurate results. Since we tested several scales of geographic divisions, our research

[1] https://www.gov.uk/government/statistical-data-sets/price-paid-data-downloads.

[2] More information about wards and MSOAs can be found on the official Office for National Statistics website: https://www.ons.gov.uk/.

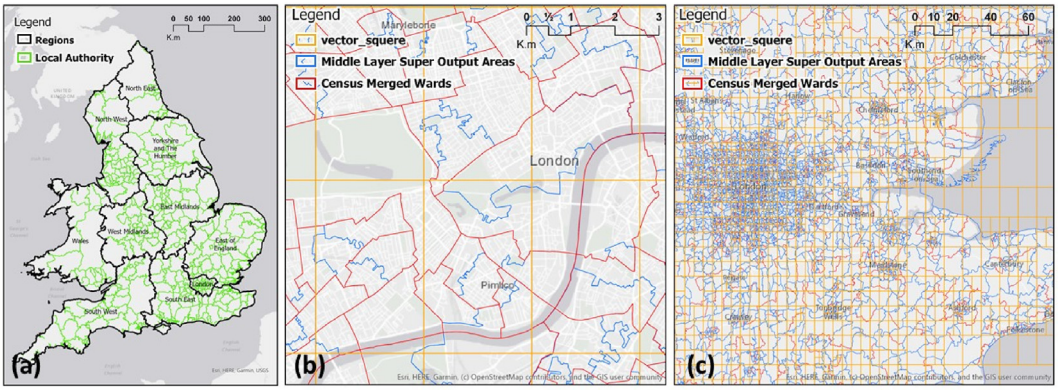


Fig. 1. (a) The 9 main regions of England and Wales and their 352 local authorities. (b) Wards, MSOAs and random rectangles (note that size changes with the density of housing). (c) Zoom-in into wards, MSOAs and random rectangles.

Table 1
The average statistical measurements of the different geographic sub-divisions for the linear regression random forest and ANN. In bold, the most accurate result for the specific division. In red font and underlined: The most accurate result of all the tests.

division	no. of subdivisions	statistical model	R^2	Adjusted R^2	less than 10% error	median error	RMSE
Region	10	Linear regression	0.349	0.349	18.71%	28.69%	124,987
		Random Forest	**0.391**	**0.391**	**21.00%**	**25.88%**	122,010
		ANN	0.249[a]	0.249[a]	18.78%	33.31%	**108,756**[a]
Local Authority	352	Linear regression	0.469	0.448	22.11%	24.38%	105,544
		Random Forest	**0.514**	**0.471**	**26.83%**	**20.93%**	102,099
		ANN	0.46[a]	0.45[a]	26.37%	23.95%	**98,297**[a]
MSOA	7201	Linear regression	0.545	0.539	24.89%	21.94%	82,408
		Random Forest	0.561	0.556	30.53%	19.08%	83,027
		ANN	**0.599**[a]	**0.594**[a]	**31.95%**	**17.87%**	**79,216**
Wards	8546	Linear regression	0.527	0.514	24.28%	22.54%	86,885
		Random Forest	0.523	0.51	29.34%	20.10%	88,603
		ANN	**0.575**[a]	**0.564**[a]	**30.88%**	**18.73%**	**83,438**
Random rectangles	7091	Linear regression	0.45	0.357	22.48%	24.46%	104,116
		Random Forest	0.408	0.301	25.91%	23.00%	108,981
		ANN	**0.502**[a]	**0.455**[a]	**27.46%**	**21.88%**	**93,115**

[a] The upper and lower 0.5% of the R2 and adjusted R2 were removed from the ANN's results due to being outliers.

results give a possible answer to this phenomenon: each algorithm handles the spatial scale of the samples differently. To the best of our knowledge, this is the first time that such an observation can be made in the urban and real-estate field considering a broad and systematic examination of data. The second insight provided by our methodology is that, as the spatial division goes smaller from the regional to the neighborhood scale, the more accurate results the algorithms give. For instance, the average median error of the random forest for the regions is 25.88%, for the local authorities is 20.93% and for the wards and the MSOA it drops to 22.54% and 21.94% respectively. Moreover, for the ANN the improvement in accuracy level is even more impressive as the scale becomes more detailed: From an average median error of 33.31% at the regional scale, to 23.95% at the local authority and then to 18.73% and 17.87% median error at the wards and the MSOA levels respectively.

The results clearly show that not only spatial scale matters in housing price predictions but also borders. For all three algorithms, the predictions' results of the two neighborhood divisions which their borders are based on physical attributes (Wards) and socio-economic attributes (MSOA) are the most accurate with a small advantage for the MSOA. The results of the random rectangles, which also has a neighborhood scale, but with random borders are significantly behind them. This shows that a neighborhood acts as a unit, and it has a crucial role in defining homogenous submarket.

Nevertheless, the average results of all the sub-divisions do not present the full picture. It is much more interesting to examine spatial patterns in housing price estimation over the territory. The map in Fig. 2(a) presents the median error between estimated and real transaction price in each of the 7201 MSOA neighbourhoods. The red areas indicate more accurate estimations and the blue tones, the least accurate. It is clearly visible that the median error does not distribute randomly over the map: some areas of the map tend to be blue while other tend to be pink and red. It appears that around large cities, the model estimate prices better than in the rural areas in the north and the west of the map (with an apparent anomaly in London's central neighborhoods, which present higher errors than the peripheral neighborhoods). Indeed, when comparing the median error map with the Rural/Urban classification map, that has been made by the UK office of national statistics for the 2011 census, it looks like a relationship between the estimated level and urban characteristics of neighbourhoods exists (Fig. 2(b)). Fig. 3 shows that the average estimation error of the urban neighborhoods is

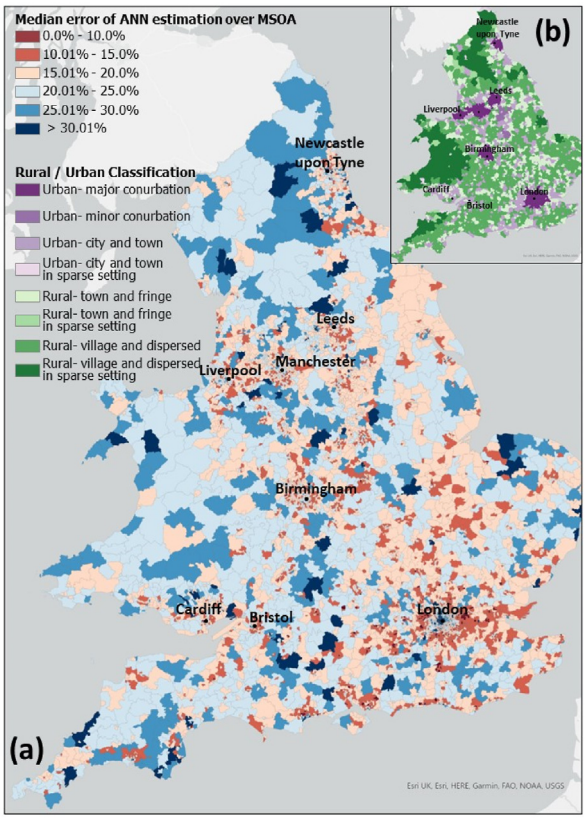


Fig. 2. (a) Median error of estimating housing transactions' price by neighborhood (MSOA) (b) Rural/Urban classification.

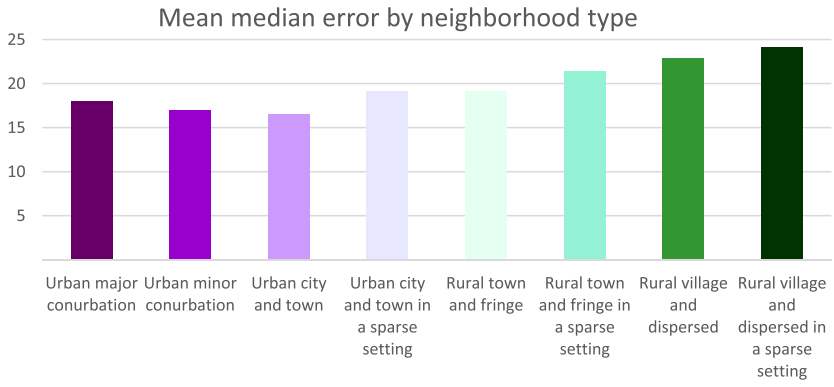


Fig. 3. The mean median error of evaluating housing transactions in neighborhoods by their rural/urban classification.

relatively low and goes from 16.49% in the 'Urban city and town' neighborhoods to 19.18% in the 'Urban city and town in sparse setting' neighborhoods while the rural neighborhoods' estimation error is higher and goes from 19.11% in the 'Rural town and fringe' to 24.14% in the 'Rural village and dispersed in sparse areas'.

However, variables such as neighborhood area, number of transactions and average price may affect the accuracy of estimation and may also be correlated with the rural/urban classification. To measure this correlation, we apply a simple linear regression in order to reveal the effect of the different urban classes on the ability to evaluate housing price in addition to other variables. Seven of the eight urban/rural classes were entered as dummy variables, with the 'urban major conurbation' class as the base variable. The average neighborhood housing price, the area of the neighborhood and the total number of transactions were entered as continuous variables while the median error of the neighborhoods is the depended variable.

Table 2 shows that for the MSOA, the urban neighborhoods have lower standardized Beta coefficient values than the rural ones. It means at a significant level, that if the neighborhood lies in an urban area, the ANN estimated its housing price more accurately, even

Table 2
Results of a linear regression measuring the effect of different MSOA neighborhoods' characteristics with the ability of the ANN model to estimate neighborhoods' transaction prices. (*** indicates significant at a 1% level, ** significant at a 5% level and * significant at a 10% level).

	Coeff.	Std. Error	Beta	t
(Constant)	15.209	.194		78.301***
Urban minor conurbation	.354	.321	.012	1.101
Urban city and town	-.856	.136	-.080	-6.305***
Urban city and town in a sparse setting	1.894	1.046	.019	1.811*
Rural town and fringe	.928	.224	.050	4.147***
Rural town and fringe in a sparse setting	2.721	.907	.032	2.999***
Rural village and dispersed	2.916	.294	.147	9.922***
Rural village and dispersed in a sparse setting	2.918	.739	.059	3.951***
Total transactions	.000	.000	-.018	-1.579
Price	.000	.000	.291	25.958***
Area	.000	.000	.123	6.981***

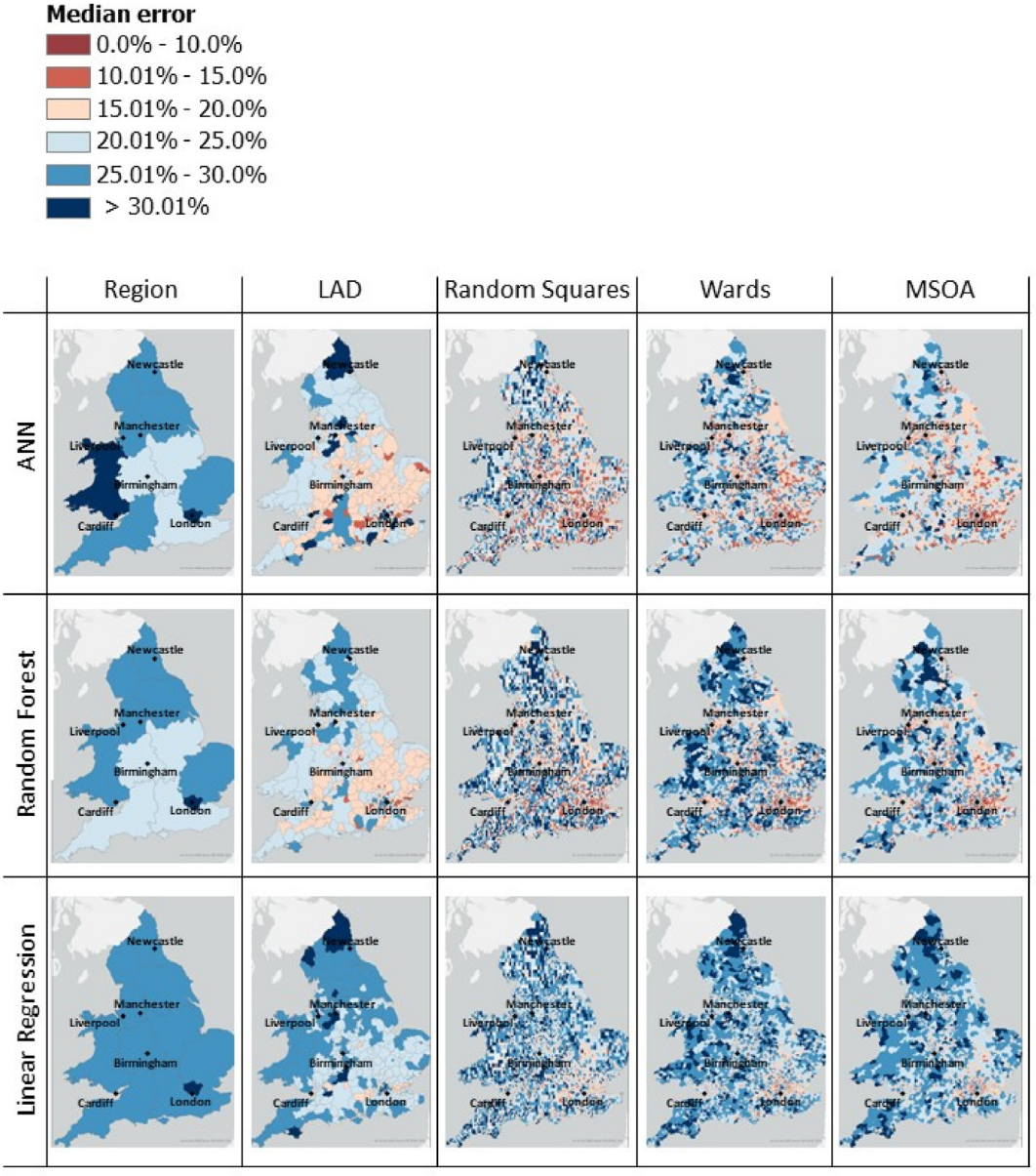


Fig. 4. Median error of estimating housing transactions' price in all five divisions and all three ML models.

when controlling for other neighborhood variables. This result explains, at least partially the pattern observed on the map shown in Fig. 2(b).

Fig. 4 shows the median error estimated for the five division and over the three statistical methods. We can clearly see that the same rural/urban pattern appears at all three neighborhood scales, and all three methods. At the local authority scale the higher errors anomaly in London's central neighbourhoods, seems to spread also to the Greater Manchester LADs, pointing on an in-LAD inequality in those areas. In the region scale, the Greater London region shows the most inaccurate results, highlighting that the London metropolitan area as the most heterogeneous region.

6. Conclusions

ML algorithms seems to gradually replace the traditional methods in the urban and real estate studies. However, at the information era of big data, not only the algorithms supposed to be changed, but the whole thinking about the possibilities, the limits, and the scale of research. At the current research, we used the full housing transaction dataset available for England and Wales to predict housing price. Some past research tested several price prediction algorithms to decide which one of them is the most accurate. We added 69,900 tests to this body of knowledge, 23,300 runs for each of the hedonic method, random forest, and ANN. As most of past research showed, the ANN and the random forest perform much better than the hedonic method. However, since we tested the algorithms over five different spatial divisions, we have found out that the random forest is more accurate at the regional and the local authority scale and the ANN is more accurate at the neighborhood scale.

Using open and free governmental databases, also has some downsides. The main one in our case is the “shallowness” of the data features, and the lack of some important transaction characteristics. While most of the reviewed papers used more than 10 data features, and all of them conclude data about the size of the soled Property (floor area or number of bedrooms), the UK governmental dataset includes only five data features which none of them is the housing size. And indeed, the average prediction accuracy we present here is lower than the results of most similar research used more detailed data. Nevertheless, the use of the full dataset and the methodology that we presented, gave us the opportunity to test for the whole land several times and find the spatial scale that minimizes the prediction error and achieve reasonable results with data and tools which are accessible to all.

One thing any ML model cannot deal well with is a ‘noisy’ data. It means that if several samples in the same area are identical in their characteristics but has very different prices, the estimation results for this area tends to be low. Therefore, according to the results presented here, the housing market in neighborhoods which highly accurate evaluated are more homogeneous in terms of housing prices.

The fact that the minimal error was achieved on the two neighborhood scale divisions is not surprising as much as the fact that a similar scale division with random borders achieved considerably higher error. It means that physical urban borders as streets and parks cluster homogenous submarket, which housing characteristic and the public preferences inside it are more similar to each other than to those in neighborhoods outside it, even to adjacent ones. This result highlights the importance of organic urban features (as the shape of continuous built areas, and topographical characteristics) and the drawbacks of rigid geometric definitions. What we call “rigid geometric definitions” are generally grids, that may be defined at different scales. In our case we use random rectangles as a control group, that is compared with other spatial configurations. As commented previously, the main objective of our paper is to find the most appropriated spatial configuration both for data collection and for urban management. But beyond this and keeping in mind that the MSOA's borders were designed for the purpose of the UK census, it is surprising that this territorial delimitation provides the most homogenous housing submarket and not the historic Wards delimitation. From a census-oriented point of view, MSOAs are spatial artifacts designed to provide consistency between workplace zones and residential areas, while keeping the number of included households between certain thresholds.[3] The results of our analysis suggest that while building this artifact, the English Office for National Statistics achieved more than intended initially. Perhaps without intention, an innovative tool for sustainable urban governance arise together with the MSOA definition. These areas are, on one hand, functional regarding quantitative and data aspects of the daily life of the city. On the other, they manage to capture the socio-economic realities expressed by mortar-and-bricks physical structures, and the willingness to pay for living in a certain place. The relative homogeneity of the property market within these areas demonstrates the instrumental value of MSOAs both as a monitoring spatial unit, as well as an urban decision-making area.

Perhaps the most surprising finding of this research is the spatial distribution of the levels of accuracy. The clear connection between the housing price prediction accuracy and the urban/rural classification of a neighborhood or an area, as shown in our results, is not trivial. Basically, it shows that different types of neighborhoods have different type of housing homogeneity, but in a deeper level, it implies that inner metropolitan, central urban, suburban, and rural areas act as different type of submarkets which may need to have different attention when dealing with their planning issues or examining their real estate situation. A concrete illustration of this type of specific territorial dynamics may be the migration of wealthy population to from the inner cities and the suburbs to the rural areas. These population movements change the characteristics of some rural zones, slowly converting them in peri-urban neighborhood, a process usually called “counter-urbanization” or “rural gentrification” (Phillips, 2010). Counter-urbanization is led by former urban populations that move to rural zones, while maintaining economic ties (mainly related to working activities) with the metropolitan area (Herslund, 2012; Karsten, 2020). However, this process is not homogeneous since not all rural areas experience the same rate of rural gentrification (Kalantaridis, 2010). It is possible that these types of geographic, demographic, and economic dynamics are behind the observed heterogeneity of rural areas.

[3] As described in “Census geography” (https://www.ons.gov.uk/methodology/geography/ukgeographies/censusgeography).

A similar conclusion can be drawn from the observed apparent anomaly in London's central neighborhoods, which present higher errors than the peripheral neighborhoods, evident in Figs. 2(a) and figure 4. The gentrification of central London is one of the most common and well-known urban research topics in urban studies (Atkinson, 2000; Butler, 2016; Moran, 2007). Moreover, in London, a 'super gentrification' phenomena are occurring, where only the top of the upper class is moving to some neighborhoods (Butler & Lees, 2006). The spatial-temporal dynamics of inner London reflects a spatial separation between wealthy neighborhoods, generally in the West, and their less affluent neighbors located in the East (Atkinson, 2000; Butler, 2016; Hamnett, 2003). Also in this case, ongoing processes of inner migration seems to create a heterogeneous and transitional housing markets which impairs the housing price estimating.

We agree with the statement that the rapid urbanization of our planet calls for smart and evidence-based responses, both from the policy side and from the urban research community. The need for a framework that allows dynamic territorial urban governance, is evidenced by a wide range of management problems characterized by a strong spatial dimension. For example, from natural hazards (Cutter & Finch, 2008), epidemiology (Grubesic & Matisziw, 2006) and deprivation (Durán & Condorí, 2019), to urban renewal (Liu et al., 2019) and the delineation of election districts (King et al., 2018). However, available practices make it difficult to study modern urban areas that are in a permanent state of flux because of changing preferences, willingness to pay, location choices, and physical development. In this changing context, traditional methods of urban management, relying on geographically fixed areas for basic data collection and governance units may be outdated. This paper contributes to the growing literature on the use of data analytic tools in urban studies and neighborhood delimitation in housing sub-markets, exploiting big data on real-estate transactions in England and Wales during a long period of time.

This paper offers an approximation to the crucial question of what the most appropriated delimitation of a "neighborhood" is, defined as a small and relatively homogeneous area in a certain (and momentary) urban configuration. On the practical level, the homogeneity of the most appropriated delimitations allows us to discuss their relevance as a basic building block of future smart and dynamic system of urban governance: A method able to define a system of neighborhoods that are as homogeneous as possible internally, and as varied as possible among them. Once such configuration is achieved, different urban policies could be focused at solving the specific problems that stem from each neighborhood's characteristics. The framework suggested in this paper can be enriched by adopting emergent methodological analyses together with the best available tools, whether established or new. Above all, it can be implemented periodically, allowing for a flexible and updated management, adapted to the ever-changing conditions of urban areas.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abidoye, R. B., & Chan, A. P. C. (2017). Artificial neural network in property valuation: Application framework and research trend. *Property Management, 35*(5), 554–571. https://doi.org/10.1108/PM-06-2016-0027

Angrist, J. D., & Frandsen, B. (2022). Machine labor. *Journal of Labor Economics, 40*(S1), S97–S140.

Atkinson, R. (2000). The hidden costs of gentrification: Displacement in central London. *Journal of Housing and the Built Environment, 15*(4), 307–326. https://doi.org/10.1023/A:1010128901782

Bailey, M. J., Muth, R. F., & Nourse, H. O. (1963). A regression method for real estate price index construction. *Journal of the American Statistical Association, 58*(304), 933–942. https://doi.org/10.1080/01621459.1963.10480679

2018, undefined Baldominos, A., Blanco, I., Moreno, A., & sciences, R. I.-A. (2018). Identifying real estate opportunities using machine learning. *Applied Sciences, 8*(11). Retrieved from https://www.mdpi.com/2076-3417/8/11/2321.

Barr, J. R., Ellis, E. A., Kassab, A., Redfearn, C. L., Srinivasan, N. N., & Voris, K. B. (2017). Home price index: A machine learning methodology. *International Journal of Semantic Computing, 11*(1), 111–133. https://doi.org/10.1142/S1793351X17500015

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Butler, T. (2016). *Living in the Bubble: Gentrification and its "Others" in North London: Urban studies* (Vol. 40, pp. 2469–2486). https://doi.org/10.1080/0042098032000136165, 12.

Butler, T., & Lees, L. (2006). Super-gentrification in Barnsbury, London: Globalization and gentrifying global elites at the neighbourhood level. *Transactions of the Institute of British Geographers, 31*(4), 467–487. https://doi.org/10.1111/J.1475-5661.2006.00220.X

Bzdok, D., Altman, N., & Krzywinski, M. (2018). Points of significance: Statistics versus machine learning. *Nature Methods, 15*(4), 233–234. https://doi.org/10.1038/NMETH.4642

Caplin, A., Chopra, S., Leahy, J. V., LeCun, Y., & Thampy, T. (2008). Machine learning and the spatial structure of house prices and housing returns. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1316046

Case, K., & Shiller, R. (1987). Prices of single family homes since 1970: New indexes for four cities. In *National Bureau of economic research working paper series*. https://doi.org/10.3386/w2393

Chau, K. W., & Chin, T. L. (2002, June 12). *A critical review of literature on the hedonic price model*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2073594.

Chen, J. H., Ong, C. F., Zheng, L., & Hsu, S. C. (2017). Forecasting spatial dynamics of the housing market using Support Vector Machine. *International Journal of Strategic Property Management, 21*(3), 273–283. https://doi.org/10.3846/1648715X.2016.1259190

Chi, B., Dennett, A., Oléron-Evans, T., & Morphet, R. (2021). *A new attribute-linked residential property price dataset for England and Wales, 2011 to 2019*. UCL Open: Environment Preprint.

Cutter, S. L., & Finch, C. (2008). Temporal and spatial changes in social vulnerability to natural hazards. *Proceedings of the National Academy of Sciences, 105*(7), 2301–2306.

Dark, S. J., & Bram, D. (2016). *The modifiable areal unit problem (MAUP) in physical geography: Progress in Physical Geography: Earth and Environment* (Vol. 31, pp. 471–479). https://doi.org/10.1177/0309133307083294, 5.

Durán, R. J., & Condorí, M.Á. (2019). Deprivation index for small areas based on census data in Argentina. *Social Indicators Research, 141*(1), 331–363.

Feng, X., & Humphreys, B. (2018). Assessing the economic impact of sports facilities on residential property values. *Journal of Sports Economics, 19*(2), 188–210. https://doi.org/10.1177/1527002515622318
Grekousis, G. (2019). Artificial neural networks and deep learning in urban geography: A systematic review and meta-analysis. *Computers, Environment and Urban Systems, 74*, 244–256.
Grubesic, T. H., & Matisziw, T. C. (2006). On the use of ZIP codes and ZIP code tabulation areas (ZCTAs) for the spatial analysis of epidemiological data. *International Journal of Health Geographics, 5*(1), 1–15.
Hamnett, C. (2003). Gentrification and the middle-class remaking of inner London, 1961-2001. *Urban Studies, 40*(12), 2401–2426. https://doi.org/10.1080/0042098032000136138
Harrison, D., & Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management, 5*(1), 81–102. https://doi.org/10.1016/0095-0696(78)90006-2
Herslund, L. (2012). The rural creative class: Counterurbanisation and entrepreneurship in the Danish Countryside. *Sociologia Ruralis, 52*(2), 235–255. https://doi.org/10.1111/J.1467-9523.2011.00560.X
Jha, S. B., Babiceanu, R. F., Pandey, V., & Jha, R. K. (2020). *Housing market prediction problem using different machine learning algorithms: A case study*. ArXiv. Retrieved from http://arxiv.org/abs/2006.10092.
Kalantaridis, C. (2010). In-migration, entrepreneurship and rural–urban interdependencies: The case of East Cleveland, North East England. *Journal of Rural Studies, 26*(4), 418–427. https://doi.org/10.1016/J.JRURSTUD.2010.03.001
Karsten, L. (2020). Counterurbanisation: Why settled families move out of the city again. *Journal of Housing and the Built Environment, 35*(2), 429–442. https://doi.org/10.1007/S10901-020-09739-3, 2020 35:2.
Kauko, T., Hooimiejer, P., & Jacco, H. (2002). Capturing housing market segmentation: An alternative approach based on neural network modelling. *Housing Studies, 17*(6), 875–894. https://doi.org/10.1080/02673030215999
King, D. M., Jacobson, S. H., & Sewell, E. C. (2018). The geo-graph in practice: Creating United States Congressional districts from census blocks. *Computational Optimization and Applications, 69*(1), 25–49.
Lam, K. C., Yu, C. Y., & Lam, K. Y. (2008). An artificial neural network and entropy model for residential property price forecasting in Hong Kong. *Journal of Property Research, 25*(4), 321–342. https://doi.org/10.1080/09599910902837051
Liu, Z., Wang, S., & Wang, F. (2019). Isolated or integrated? Planning and management of urban renewal for historic areas in Old Beijing city, based on the association network system. *Habitat International, 93*, Article 102049.
Mora-esperanza, J. (2004). *ARTIFICIAL INTELLIGENCE APPLIED TO REAL ESTATE VALUATION an example for the appraisal of Madrid*. Citeseer. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.564.2570.
Moran, J. (2007). Early cultures of gentrification in London. *1955–1980: Journal of Urban History, 34*(1), 101–121. https://doi.org/10.1177/0096144207306611
Mu, J., Wu, F., & Zhang, A. (2014). Housing value forecasting based on machine learning methods. *Abstract and Applied Analysis*. https://doi.org/10.1155/2014/648047, 2014.
Naik, N., Kominers, S. D., Raskar, R., Glaeser, E. L., & Hidalgo, C. A. (2017). Computer vision uncovers predictors of physical urban change. In *Proceedings of the National Academy of Sciences of the United States of America* (Vol. 114, pp. 7571–7576). https://doi.org/10.1073/pnas.1619003114, 29.
Ng, A. (2015). Machine learning for a London housing price prediction mobile application. In *doc.ic.ac.UK*. Retrieved from http://www.doc.ic.ac.uk/~mpd37/theses/2015_beng_aaron-ng.pdf.
Nguyen, N., & Cripps, A. (2001). Predicting housing value: A comparison of multiple regression analysis and artificial neural networks. *Journal of Real Estate Research, 22*(3), 313–336. Retrieved from https://ideas-repec-org.ezlibrary.technion.ac.il/a/jre/issued/v22n32001p313-336.html.
Oladunni, T., & Sharma, S. (2016). Hedonic housing theory — a machine learning investigation. In *2016 15th IEEE International Conference on machine learning and applications (ICMLA)* (pp. 522–527). https://doi.org/10.1109/ICMLA.2016.0092
Orford, S. (2010). Towards a data-rich infrastructure for housing-market research: Deriving floor-area estimates for individual properties from secondary data sources. *Environment and Planning B: Planning and Design, 37*(2), 248–264.
Park, B., & Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert Systems with Applications, 42*(6), 2928–2934. https://doi.org/10.1016/j.eswa.2014.11.040
Phillips, M. (2010). Counterurbanisation and rural gentrification: An exploration of the terms. *Population, Space and Place, 16*(6), 539–558. https://doi.org/10.1002/PSP.570
Reades, J., De Souza, J., & Hubbard, P. (2018). Understanding urban gentrification through machine learning. *Urban Studies*. https://doi.org/10.1177/0042098018789054, 004209801878905.
Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy, 82*(1), 34–55. https://doi.org/10.1086/260169
Santibanez, S. F., Kloft, M., & Lakes, T. (2015, May). Performance analysis of machine learning algorithms for regression of spatial variables. A case study in the real estate industry. In *13th International Conference of GeoComputation, Dallas, USA* (pp. 20–23).
Shahhosseini, M., Hu, G., & Pham, H. (2020). *Optimizing ensemble weights for machine learning models: A case study for housing price prediction*. https://doi.org/10.1007/978-3-030-30967-1_9
Shearmur, R. (2015). Dazzled by data: Big Data, the census and urban geography. *Urban Geography, 36*(7), 965–968.
South, B., & Henretty, N. (2017). House price statistics for small areas: Using administrative data to give new insights. *Statistical Journal of the IAOS, 33*(3), 609–614.
Specht, D. F. (1991). A general regression neural network. *IEEE Transactions on Neural Networks*. https://doi.org/10.1109/72.97934
Truong, Q., Nguyen, M., Dang, H., & Mei, B. (2020). Housing price prediction via improved machine learning techniques. *Procedia Computer Science, 174*, 433–442. https://doi.org/10.1016/j.procs.2020.06.111
Weaver, R. C. (2014). *Urban geography evolving: Toward an evolutionary urban geography*.
Weigand, A. (2019). Machine learning in empirical asset pricing. *Financial Markets and Portfolio Management*, 1–12. https://doi.org/10.1007/s11408-019-00326-3
Xiao, Y. (2017). *Hedonic housing price theory review* (pp. 11–40). Springer Geography. https://doi.org/10.1007/978-981-10-2762-8_2
Yan, Z., & Zong, L. (2020). Spatial prediction of housing prices in Beijing using machine learning algorithms. *ACM International Conference Proceeding Series*, 64–71. https://doi.org/10.1145/3409501.3409543
Zhang, G., Eddy Patuwo, B., & Hu, Y. M. (1998). Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting*. https://doi.org/10.1016/S0169-2070(97)00044-7
Zhou, X., Tong, W., & Li, D. (2019). Modeling housing rent in the atlanta metropolitan area using textual information and deep learning. *ISPRS International Journal of Geo-Information, 8*(8), 349. https://doi.org/10.3390/ijgi8080349