Hindawi
Complexity
Volume 2020, Article ID 5287263, 12 pages
https://doi.org/10.1155/2020/5287263

*Research Article*

# A Fully Automated Adjustment of Ensemble Methods in Machine Learning for Modeling Complex Real Estate Systems

**José-Luis Alfaro-Navarro**,[1] **Emilio L. Cano**,[2] **Esteban Alfaro-Cortés**,[2] **Noelia García**,[1] **Matías Gámez**,[2] **and Beatriz Larraz**[2]

[1]*Faculty of Economics and Business Administration, University of Castilla-La Mancha, Albacete, Spain*
[2]*Quantitative Methods and Socio-Economic Development Group, Institute for Regional Development (IDR), University of Castilla-La Mancha (UCLM), Albacete, Spain*

Correspondence should be addressed to Beatriz Larraz; beatriz.larraz@uclm.es

Received 12 November 2019; Accepted 19 March 2020; Published 14 April 2020

Guest Editor: Marco Locurcio



The close relationship between collateral value and bank stability has led to a considerable need to a rapid and economical appraisal of real estate. The greater availability of information related to housing stock has prompted to the use of so-called big data and machine learning in the estimation of property prices. Although this methodology has already been applied to the real estate market to identify which variables influence dwelling prices, its use for estimating the price of properties is not so frequent. The application of this methodology has become more sophisticated over time, from applying simple methods to using the so-called ensemble methods and, while the estimation capacity has improved, it has only been applied to specific geographical areas. The main contribution of this article lies in developing an application for the entire Spanish market that fully automatically provides the best model for each municipality. Real estate property prices in 433 municipalities are estimated from a sample of 790,631 dwellings, using different ensemble methods based on decision trees such as bagging, boosting, and random forest. The results for estimating the price of dwellings show a good performance of the techniques developed, in terms of the error measures, with the best results being achieved using the techniques of bagging and random forest.

## 1. Introduction

Since the year 2008, the global economic crisis caused a slowdown in the economy which resulted in a decrease in the price of real estate properties. The appraised valuation of a property is considered key for any transaction related to the property and particularly for its sale or for a mortgage application, so it is essential that the price is a true reflection of its value. Banks also need to periodically review the value of their real estate portfolio by updating their appraised valuation, see the Basel II International Banking Agreement [1, 2]. Normally, appraisals for the purpose of a mortgage are carried out by professional appraisers visiting the property. However, the appraisal procedure developed in this way is expensive, both in terms of time and money, and makes this procedure unsustainable for the valuation of large real estate portfolios. Furthermore, although the physical presence of an appraiser may help to give a more accurate valuation of the property, there is also the possibility of bias from interested parties, such as buyers, sellers, or banks themselves, which may make the valuation more subjective. There is clearly a need for the development of a prediction model which can present unbiased, realistic valuations.

The International Association of Assessing Officers (IAAO) considers mass appraisal as the process of valuing a group of properties using common data, standardized methods, and statistical procedures [3]. These valuation methods have been implemented through models known as Automated Valuation Models (AVM) and have enabled the appraisal of large real estate portfolios without the direct intervention of an appraiser [4, 5]. The development of these estimation procedures has been enhanced by the growth in

the quantity and quality of information related to both real estate prices and property characteristics, accessible to researchers. This level of information allows the application of increasingly sophisticated statistical techniques for the development of estimation procedures of a higher quality and precision. Real estate AVMs allow the valuation of property prices en masse, without the need for the physical presence of an appraiser, by using computer-assisted task appraisal systems [6]. In many cases, the presence of an appraiser is only necessary for those valuations considered to be out of the ordinary [7].

Estimation techniques include parametric regression analysis [8] and nonparametric [9] or machine learning methods such as neural networks [10, 11], decision trees [12, 13], random forests [14, 15], fuzzy logic [16], or ensemble methods [17]. These techniques are used primarily with three objectives in mind: to estimate a real estate property price, to find out the influence of a characteristic of the house on its price, and to create a hedonic price index.

Over the recent decades, the most commonly used procedures has been based on hedonic-based regression [18, 19]. However, these models present certain fundamental problems related to the assumptions of the model: normality of the residuals, homoscedasticity, independence, and the absence of multicollinearity. This situation has led to greater use of pattern recognition techniques, often known as data mining techniques, which include machine learning. These techniques are more flexible about the assumptions related to the distribution of data, they are easier to interpret, and they allow linear and nonlinear relationships to be analysed. In addition, they enable both categorical and continuous variables to be managed [13]. Although these techniques were initially used more as classification methods, in recent years, their application has been used in determining the most influential variables on house pricing and in estimating dwelling prices. Pérez-Rave et al. [20] provide a two-stage methodology for the analysis of big data regression under a machine learning approach for both inferential and predictive purposes.

Accurate and efficient prediction of real estate prices has been and will continue to be an essential but controversial issue, with an impact on the various actors in the economy such as buyers, sellers, commission agents, governments, and banks [21, 22]. Nowadays, the big data paradigm offers exciting possibilities for more accurate predictions and one of the main approaches for dealing with big data is machine learning.

These machine learning methods have been applied to the estimation of real estate properties in very specific locations. Research has been carried out by Jaen [12] in Coral Gables (Florida, USA); Fan et al. [13] in Singapore (Republic of Singapore); Özsoy and Şahin [23] in Istanbul (Turkey); Del Cacho [14] in Madrid (Spain); Pow et al. [17] in Montreal (Canada); Ceh et al. [24] in Ljubljana, (Slovenia); Nguyen [25] in 5 counties of USA; and Dimopoulos et al. [26] in Nicosia (Cyprus). In contrast, Pérez-Rave et al. [20] deal with the estimation of dwelling prices for a whole country, Colombia, using independent variables identifying the city in which each property is located, thus proposing a unique model for the entire country (with a sample of 61,826 properties).

The main new element of this article is that it proposes a new methodology to carry out the automated estimation of real estate prices for an entire country (Spain in this case study), specifying automatically a different model for each municipality, with a sample size of 790,631 real estate properties. The whole country can be considered as a complex real estate system due to the great differences that exist between rural and urban areas and even inside the urban areas. Each municipality is trained with the information available in its training set so that each one will have its own model adapted to its characteristics and needs. This study addresses a program capable of covering a different model for each of the 433 municipalities with more than 100 properties for sale, with a population ranging from 1,559 inhabitants in the smallest municipality to 3,223,334 in Madrid, the biggest one. We focus on the application of this automated valuation system based on machine learning methods in estimating real estate prices and analyse the accuracy of each of them using error measures widely recognised in economic literature. As a rule, to evaluate model quality, aggregated diagnostic indicators are used (coefficient of determination) although there are few contributions in the relevant literature where the quality of the procedure is analysed using a measurement of the estimation error [15]. In this article, we use four measures to analyse the validity of the proposed methods, namely, the mean ratio, the mean absolute percentage error (MAPE), the median absolute percentage error (MdAPE), and the coefficient of dispersion (COD).

In this article, machine learning techniques used in estimating the price of dwellings are based on the decision tree technique. However, in general, it is difficult to build a single tree to make predictions because of incorrect parameter settings, simplicity rules, and tree instability. To overcome these problems and obtain better behaviour when making predictions, techniques of ensemble of decision trees have been developed, such as bagging, boosting, and random methods [27]. In bagging, the models are fitted using random independent bootstrap replicates that are then combined by averaging the output for regression [28]. In boosting, the fitted model is a simple linear combination of many trees that are fitted iteratively and boosted to reweight poorly modelled observations [29]. The random forest model, however, is constructed in a random vector of the data feature space sampled independently [30]. Starting from this base, we automatically design the best ensemble method, including bagging, boosted regression tree, and random forest in each municipality and then make a comparison to analyse their behaviour under different circumstances. In addition, the results obtained with a single decision tree are included in order to analyse and compare the benefit of using an ensemble of decision tree techniques.

A further consideration is the particular emphasis that specialized literature places on the need to include spatial information in hedonic models, given the significant influence that the location of the property can have on its price and therefore on its valuation. Ceh et al. [24] highlight the

growing interest in recent years in applying spatial statistics to hedonic price modeling, besides coupling the geographic information system and machine learning techniques. In this article, we include geographical coordinates in the explanatory variables to draw attention to the importance of the property location when valuing a property.

The article is laid out as follows. After the introduction, Section 2 presents a literature review highlighting the main research to date on the valuation of real estate property prices from the use of regression trees to tree ensemble models. In Section 3, the methodology used is presented with a description of the main techniques as well as the valuation measures of behaviour of the different models. In Section 4, the empirical argument for an application to the whole of Spain is developed and advocates the need for using ensemble methods for Spain. Finally, Section 5 presents the main conclusions and further lines of research.

## 2. Literature Review

The application of machine learning methods in the field of estimating property prices has attracted interest for some years now. However, the application of decision trees is relatively recent, initially being used as a classification technique and for determining which variables had the greatest influence on the price of housing. The application of decision trees was then used as a prediction technique through the so-called regression tree to obtain dwelling price predictions. One of the first proposals for the application of a regression tree was made by Jaen [12] who used information from 15 variables for 1,229 transactions in the city of Coral Gables (Florida) taken from the multiple listing system (MLS). Jaen [12] tests the effectiveness of using stepwise regression, CART decision tree, and neural networks in estimating the price of housing and in determining the most important variables for this prediction. The best results are achieved from CART measuring the estimation capacity with the mean absolute error (MAE), using a smaller number of variables, specifically five versus the nine used in stepwise regression.

Following on from [12], Fan et al. [13] demonstrate the good behaviour of regression trees, using the CART algorithm to identify the main determinants and predict the price of housing. This application is developed for the Singapore resale public housing market. However, although the process used to identify the main variables that affect the price of housing is extensive, its estimation is based solely on the average value in a leaf node of the tree, this value being thought of as a forecasting value or regression value. Özsoy and Şahin [23] develop a CART application in Turkey to determine the most influential characteristics on the price of housing in Istanbul based on a database taken from the Internet in 2007. The results lead them to conclude that the size of the house and the existence of an elevator, security, central heating, and views are the most influential variables on house prices in Istanbul.

Kok et al. [31] show the main advantages of the application of regression trees to predict property prices, taking into account that these models help overcome the problem of regression models in nonlinear relationships. The advantages highlighted by the authors are they are simple to understand and interpret, and their statistical significance is easy to calculate; they can handle categorical variables without creating dummy variables; and they consume little computing time even with large amounts of data. In addition, the authors propose the use of a procedure called stochastic boosting which allows an unlimited number of variables to be handled with good results, including economic and demographic variables and hyperlocal metrics in the prediction model. The limitations of regression trees are they can show unlimited growth vertically until the sample has an observation which may generate models with poor generalization capacity; they are not robust to changes in the training set; and they usually suffer an underfitting effect giving rise to models with little predictive capacity. To solve these limitations, the authors propose the use of tree ensembles such as random forest. Though it is true that these models have been used before in the pioneering works in [14, 15], Breiman [30] produced one of the first papers that highlights the need to improve prediction using ensemble methods.

Following on from these papers, there have been numerous proposals that compare the behaviour of ensemble techniques with classical regression models, concluding that models behave better with machine learning techniques. Likewise, in the work by Pow et al. [17], they use 25,000 web data on Montreal properties with 130 characteristics; 70 related to the housing itself and 60 sociodemographic. These authors use principal component analysis (PCA) to reduce the dimension and four regression techniques to predict property prices: linear regression, support vector machine, K-nearest neighbors (KNN), and random forest regression and an ensemble approach by combining KNN and random forest technique. From the results, the authors highlight the good behaviour of the ensemble approach with a mean absolute percentage difference for the asking price of 9.85. In addition, they show that applying PCA does not improve the prediction error.

Ceh et al. [24] analyse the behaviour of random forest compared to multiple regression to select the most important variables. In the case of multiple regression, an analysis of main components allows it to go from 36 variables to 10 principal components and in the case of random forest, a procedure is carried out to determine the 10 most important variables. Interestingly, for random forest, the date of sale is important but not for ordinary least squares (OLS). Although the behaviour in terms of COD and MAPE of random forest is better than that of OLS, it should be noted that both overestimate the lowest prices and underestimate the highest. Specifically, in the application developed for the price of apartments in Ljubljana (Slovenia) with 7,497 observations for the 6-year period 2008–2013, the results in terms of MAPE for the test set were 7.27% for RF and 17.48% for multiple OLS, while in terms of COD the values obtained were 7.28% and 17.12%, respectively. Although the authors state that their model does not take into account the potential price differences over the 6-year time period under consideration, this price change could

influence their results. In our study, we use a static database of 2018 to avoid this problem.

Nguyen [25] develops an application in five counties in the United States using Zillow group web data by comparing linear regression models, random forest, and support vector machine. The results lead the author to conclude that both random forest and support vector machine behave better than linear regression in terms of the percentage of houses whose estimated prices fall within a 5% range of their actual sold prices. In addition, the conclusion emphasizes that it is not necessary to change the variables used in each county and that the accuracy of the model is practically the same using a series of common attributes for all of them. Dimopoulos et al. [26] develop an application to compare the behaviour of random forest and linear regression in estimating the prices of residential apartments in Nicosia (Cyprus). The results verify that the best behaviour in predictive terms is that of random forest, with average MAPE values of 25.2%. Shinde and Gawande [32] use data based on 3,000 observations with 80 parameters of a database called KaggleInc to compare the behaviour of logistic regression, support vector regression, lasso regression, and decision tree and show that the best behaviour, both in terms of accuracy and of error, is achieved with the decision tree. The variables used to estimate the sale price are area in square metres, overall quality which includes the overall condition and finish of the dwelling, location, the year in which the house was built, number of bedrooms and bathrooms, garage area and number of cars that can fit in the garage, swimming pool area, year in which the house was sold, and price at which the house was sold.

In addition to the comparison in the literature of machine learning techniques with classical regression models, there is a wide range of literature that compares different machine learning methods, concluding that there is no one technique that shows better behaviour than the others but highlights the best behaviour of tree ensemble techniques. For example, Kagie and Wezel [33] use Friedman's LSBoost and LADBoost boosting algorithms designed for regression with three main objectives: to predict dwelling prices in six areas in Netherlands; to determine the most important characteristics; and to build a price index. To do this, they use transaction data from the year 2004 obtained from Nederlandse Vereniging van Makelaars (NVM, Dutch Association of Real Estate Brokers) for the cities of Groningen, Apeldoorn, Eindhoven, Amsterdam, Rotterdam, and Zeeland, with 83 variables and a number of observations ranging from 2,216 for Zeeland to 8,490 for Amsterdam, also including sociodemographic variables. The results show that both boosting models improve the behaviour of linear and nonlinear models in the six areas considered, with improvements in terms of the absolute error of around 25–30% and in relative error of around 33–39%. In addition, they show that the models present a better behaviour in the prediction of errors in terraced houses and apartments and a worse behaviour in predicting errors in detached houses, which is consistent considering that the most influential characteristic on dwelling price is the size of the house.

Del Cacho [14] compares different ensemble methods for housing valuation in Madrid, based on a sample of 25,415 observations taken from an online real estate portal. The results show a better behaviour of ensemble of M5 model trees with a better behaviour of bagging unpruned decision trees, with a mean relative error of 15.25%. Similar results with a median percentage error of 15.11% and 13.18% are obtained for the English private rental market using gradient boost [34] and Cubist [35], respectively, by Clark and Lomax [36]. Graczyk et al. [37] use six machine learning algorithms: multilayer perceptron (MLP); radial basis function neural network for regression problems (RBF); pruned model tree (M5P); M5Rules (M5R); linear regression model (LRM); and NU-support vector machine (SVM) for the three ensemble methods of additive regression (an implementation of boosting in WEKA), bagging, and stacking, in Waikato Environment for Knowledge Analysis (WEKA). The results show that there are differences between the simple and ensemble methods used although all of them with good behaviour in terms of MAPE had values ranging from 19.02% to 15.89%. Bagging results are the most stable, with better results using SVM. However, the best results are obtained using stacking and SVM. The general conclusion of the study is that there is no single algorithm that produces the best results and, therefore, it is necessary to investigate the behaviour of different alternatives.

Antipov and Pokryshevskaya [15] show the best behaviour of random forest when estimating prices per square metre rather than for the total price due to heteroscedasticity and other real estate data problems. They propose comparing the behaviour of 10 algorithms: multiple regression; CHAID; exhaustive CHAID; CART; k-nearest neighbors (2 modifications); multilayer perceptron neural network (MLP); radial basis function neural network (RBF); boosted trees; and random forest. In the evaluation of each method, habitual metrics are used in the validation of the predictive capacity of automated valuation models such as the average ratio sale (SR), the coefficient of dispersion (COD), and the mean average percentage error (MAPE). All the analysed techniques showed acceptable values for all the metrics, both in the training set and in the test set and with better results for random forest with a MAPE of 17.25 and a COD of 16.97 while, using a two-step procedure, these are 14.86 and 14.77, respectively. In addition, this study proposes a classification of variables according to their relevance, highlighting the importance of the type of house and the district in which it is located. It also recommends a segmentation-based diagnostic method that determines segments based on the total area and the district in which the house is located, with any overestimated or underestimated value highlighting the need for the intervention of an appraiser. However, the main drawback of this study is that the data are too limited, focusing on 2-bedroom apartments with an area of up to 160 $m^2$ and a price below 30 million rubles. Such a limited profile is an unrealistic reflection of most cities.

Lasota et al. [38] propose that instead of using a single expert machine learning system, a combination of these should be used. They argue that in this way, the risk of selecting a poor model would be reduced in some of the cases

and large volumes of data could be analysed efficiently by applying the procedure to small partitions of the data and combining the results. This proposal is compared with the individual methods with two ensemble machine learning methods: mixture of experts (MoE) and Adaboost.R2 (AR2), concluding that Adaboost and this mixture of machine learning procedures show better behaviour, with no significant differences between the methods. In the case of MoE, the algorithms, multilayer perceptron, general linear model, and support vector regression, are used, while for AR2, multilayer perceptron, general linear model, and a regression tree are used. It is the mixture of machine learning procedures with multilayer perceptron and general linear model that show a better behaviour, without significant differences between MoE and AR2. However, in the study by Lasota et al. [38], they used information for the period 1998–2011 with the problem, as highlighted by the authors, of comparability in the data. They also use only four characteristics as explanatory variables which can give rise to an oversimplified model and could be the reason for a good behaviour of ensemble techniques with simple basic techniques such as the general linear model.

Another comparison, in this case of random forest with other machine learning methods was developed by Yoo et al. [39]. Machine learning is used to determine the variables which have the greatest impact on the price of housing in Onondaga (New York) and to establish a way of estimating dwelling prices. Specifically, OLS regression methods are compared with Cubist and random forest. In terms of determining the most important variables, though OLS uses a stepwise selection based on the level of significance, RF or Cubist uses boosting or bagging techniques that permit the handling of nonlinear models as they are nonparametric procedures. For predictability, the behaviour of the two machine learning techniques is better, highlighting RF in terms of root mean squared error (RMSE) with values, in relative terms with respect to their average, of 25.04 considering a neighbourhood within a radius of 100 metres and 22.47 within a radius of 1 km, for the test set. In addition, the model also incorporates environmental variables which have not previously been included in these types of models. The authors highlight that the application of machine learning methods in the selection of variables allows key variables to be selected without being based on a level of significance. These methods also allow a sufficiently parsimonious set of important variables to be found for good prediction, which means it is not so important that the model contains all relevant variables, as long as the prediction works well. Park and Bae [40] compare C4.5, RIPPER, Naive Bayesian, and Adaboost in the residential market of the county of Fairfax, Virginia, concluding that the best behaviour is achieved with RIPPER. In addition, their study uses these techniques as classification techniques, not regression, when classifying properties based on the presence of a positive or negative value in the difference between what they call closing (sold) prices and listing (for sale) prices.

Shahhosseini et al. [41] compare the behaviour of several ensemble models for the prediction of dwelling prices using two databases, widely cited in the relevant literature, the Boston metropolitan area dataset [42] and the sales database of residential homes in Ames (Iowa) presented in [43]. To demonstrate the validity of the ensemble models, they use the following algorithms: multiple learners including lasso regression, random forest, deep neural networks, extreme gradient boosting (XGBoost), and support vector machines with three kernels (polynomial, RBF, and sigmoid). Based on the results of the median price prediction error, for Boston, the best performance in terms of MAPE appears for XGBoost and random forest with MAPE values of 16.44% and 16.35%, respectively. In the case of Ames housing, lasso and random forest are the models with the best MAPE with values of 0.66% and 0.77%, respectively. Incredibly low errors are attributable to the quantity and quality of information related to the 80 available variables as well as the huge sample size (2,930) in relation to the population size of Ames (Iowa, USA) of 50,781 inhabitants. Therefore, these results lead to us conclude that there is no one model which performs better than the others.

Finally, Neloy et al. [44] develop a model for predicting the rental price of houses in Bangladesh through a website database of 3,505 homes with information on 19 characteristics. To develop the model, the following simple algorithms are selected for prediction: advance linear regression, neural network, random forest, support vector machine (SVM), and decision tree regressor. In addition, the ensemble learning is stacked with the following algorithms: ensemble AdaBoosting regressor, ensemble gradient boosting regressor, and ensemble XGBoost. Also, ridge regression, lasso regression, and elastic net regression are used to combine the advanced regression techniques. The best results, in terms of accuracy, are obtained by the ensemble gradient boosting with 88.75% and the worst by the ensemble AdaBoosting with 82.26%. In terms of root mean square error (RMSE), the behaviour is similar, with values of 0.1864 and 0.2340, respectively.

Other uses of the decision tree include the application of the CART algorithm to segment the observations and to improve the ability to estimate the model by applying different models by segments or even with the assistance of an appraiser if necessary [45]. To do this, a CART algorithm is applied, using the percentage error (estimated value less real value in absolute value divided by real value) as a dependent variable and the sales ratio (estimated value divided by real value) to determine segments of observations that allow them to go from a general MAPE in the simple training of 12,688 to a value in the best segment of 9,783 and in the simple test of 14,859 to 12,364. Pérez-Rave et al. [20] propose a methodology that incorporates a variable selection procedure called simple incremental with resampling (MINREM). This procedure is used in combination with a principal component analysis in two cases; 61,826 homes sold in Colombia, and the data used in [46] from the 2011 Metropolitan American Housing Survey with 58,888 observations. The results show a MAPE value of 27% without using interactions and 20.9% using the procedure proposed, in the case of housing in Colombia.

From all these studies, it follows that the analysis of the behaviour of different machine learning techniques to

analyse the price of housing has been widely covered in the literature. While the majority of the applications stress the importance of determining the most influential variables on the price of housing, there are few applications which focus on prediction and, above all, there are few studies that use measures such as MAPE or COD to help evaluate the predictive capacity of the models; the majority are based on measuring the predictive capacity using the coefficient of determination. In addition, the applications developed focus on specific areas or cities without trying to cover a wide geographical area (except in the case of Colombia where the study use the same model for the whole country). In this study, we cover a wider geographical area by developing, through an automated procedure for estimating models, a model to be applied to each of the Spanish municipalities where information is available. This gives us a total of 433 municipalities.

## 3. Methodology

As it has been stated, the aim of this article is to develop an automatic application that contains, for each municipality, a model capable of accurately estimating the price of housing. Several models are fitted in each municipality, among a range of competing machine learning techniques. Then, they will be analysed in order to check if there exists one best method that achieves optimal results in terms of the error measures explained at the end of this section.

The selected models are bagging, boosting, and random forest. All of them are ensemble algorithms, and we use regression trees as base learners. For this reason, the results of the single decision tree model will also be displayed as a reference alongside the results of the more complex models. The ensemble methods usually provide good prediction results although it is true that they sacrifice in some way the possibility of interpretation of the relationships between the predictor variables and the target. In our context, given the large number of models that will be estimated to completely cover the Spanish territory, accurate predictions are more important that easily interpret models.

The following briefly shows what each of these ensemble methods consists of. To begin with, bagging is an ensemble method proposed in [47] from the basis of bootstrapping and aggregating methods. The main advantage of this methods is the reduction of noise presence in the observations in the random samples obtained with replacement from the original set. Once the trees are fitted over the bootstrap samples, the outputs are averaged. The noise reduction coupled with the instability often shown by individual predictors lead bagging to improvements, especially for unstable procedures.

For its part, boosting [48] is an ensemble method capable of converting a weak learner into one with much higher accuracy. Boosting, similar to bagging, applies an iterative learning process. The differential characteristic of this method is that each iteration is not independent of the previous ones but uses a reweighting system to focus the attention of the learning process on the observations that in former steps have been estimated with higher errors. The chosen algorithm to implement boosting in this article is gradient boosting [34] that consists in adding weak learning models, such as decision trees, by using a gradient descent procedure to minimize a loss function.

Random forest was also proposed in [30], and it could be seen as a variation of the bagging method, with a higher dose of randomness. This added randomness is given because when constructing the successive trees, the optimal division is not sought among all the available predictive variables, but only among a subset randomly chosen in each node. The main advantage of this method is that it incurs a lower risk of overfitting and, therefore, usually provides more accurate estimates. It should be noted that bagging is a special case of random forest when the subset of variable candidates contains the total number of predictors available.

All the models have been applied using the statistical environment R [49]. Specifically, the R packages rpart [50], gbm [51], and randomForest [52] have been used for fitting individual trees, boosting and bagging, and random forest, respectively.

Due to the large number of models to be fitted in this complex problem, the parameter tuning in random forest has been optimized for each model through the caret R library [53]. There are three main parameters to be set in random forest. The first two are the number and size of trees to be grown. The number should not be too small to ensure that every input row participates in the learning process at least a few times. The size of trees depends on the minimum size of terminal nodes. Setting this number larger leads to smaller trees and quicker learning procedure. Another important parameter in random forest is the number of predictors randomly sampled as candidates at each split. With regard to bagging, it has been treated as a particular case of random forest.

Regarding boosting, there are four main parameters to be set in the gbm model. The first one is the learning rate (shrinkage), with values 0.001, 0.01, and 0.1, which controls how large the changes are from one iteration to the next one, similarly to the learning rate in neural networks. Secondly, the complexity of the tree is controlled by two parameters, interaction depth (tested among 1, 3, 5, and 10) and the minimum number of observations per node, similar to random forest (taking 1, 5, 10, and 20). Finally, but also very important, the number of trees (iterations), an ensemble of 1,000 trees is generated and then pruned according to the minimum cross-validation error.

The loss function chosen for the optimization of each supervised method is the mean square error (MSE). In order to guarantee a good generalization ability avoiding overfitted models, 2/3 of the observations in the sample has been randomly assigned to the train set and the other 1/3 to the validation set. Once the best model of each technique (regression tree, bagging, boosting, and random forest) has been chosen in each municipality, the comparison of the final behaviour of four models has been analysed by the following error measures. They will be able to analyse the goodness of fit and validate the predictive capacity of the models.

(a) Mean ratio (average sales ratio) is the average of the $SR_i$, $SR$ being the sales ratio defined as

$$SR_i = \frac{\widehat{y}_i}{y_i}, \tag{1}$$

where $y_i$ is the property value $i$ and $\widehat{y}_i$ is the estimated value.

(b) Mean absolute percentage error (MAPE) or relative mean error:

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \widehat{y}_i}{y_i}\right| \cdot 100. \tag{2}$$

The measure is in percentage terms, so it is comparable among different models.

(c) Median absolute percentage error (MdAPE):

$$\text{MdAPE} = \text{Me}\left(\left|\frac{y_i - \widehat{y}_i}{y_i}\right|\right) 100, \tag{3}$$

where Me() is the median, i.e., the value separating the higher half from the lower half of the absolute percentage errors.

(d) Coefficient of dispersion (COD):

$$\text{COD} = \sum_{i=1}^{n}\frac{(1/n)\sum_{i=1}^{n}\left|\text{SR}_i - \text{Me}(\text{SR})\right|}{\text{Me}(\text{SR})}100, \tag{4}$$

where Me(SR) is the median of the $\text{SR}_i$. Its interpretation does not depend on the assumption of normality.

In line with the study by Pérez-Rave et al. [20], we carry out the estimation and measurement of errors using monetary values since any transformation of the variable to be estimated (price), such as the logarithmic transformation, can lead to an improvement in results from a fictitious statistical point of view. In addition, the estimated value of the price is made in monetary terms and does not require any transformation for its interpretation and comparison between locations.

## 4. Empirical Application

To develop the empirical application, a database is constructed based on the information obtained from freely available real estate websites. The data from advertisements on the Internet allow the development of the application of big data techniques for the analysis of dwelling prices with greater precision because the volume of accessible data is large and enriched daily, both ideal characteristics to be able to apply these techniques. In addition, the data are quite varied and the sale value on the web and the offline sale value are seen to be of similar magnitude. The Internet source also offers information on a variety of property and neighbourhood characteristics that are difficult to find from other sources [20]. However, this source of information has been little used despite the existence of applications developed with great success in both the real estate sector as well as other sectors [54]. These same aspects are highlighted in [55, 56] in which the authors point out that web prices offer a valuable opportunity for statistical analysis due to the constant generation of information, their accessibility, and availability as well as there being little notable differences compared to offline prices. Within real estate, applications developed using web data are used by Özsoy and Şahin [23] in Istanbul; Del Cacho [14] in Madrid; Larraz and Larraz and Población [57, 58] in Spain; Pow et al. [17] in Montreal; Larraz and Población [59] in Czech Republic; Nguyen [25] in the United States; Clark and Lomax [36] in England; Pérez-Rave et al. [20] in Colombia; or Neloy et al. [44] in Bangladesh.

In our study, the database contains information related to the price of the property (flat nonsingle-family home) and its reliable geolocation as well as information that refers specifically to the characteristics of each property. We have access to information on properties for sale in all Spanish municipalities during 2018. The information includes the price of the property and the following 33 variables which represent the characteristics of each property: a text variable that shows the postal code in which the property is located; three numerical variables that include the constructed surface area, the number of bedrooms, and the number of bathrooms; and 29 attributes that have been categorized into different levels. Among these, the variables considered the most influential on dwelling prices by the implemented methods are location (longitude and latitude coordinates), constructed area, number of bedrooms, number of bathrooms, floor (basement, normal, or attic), the state of conservation (new, with important improvements, adequate for the age, or need for major improvements), and the presence of air conditioning, heating, lift, garage, terrace, green areas, swimming pool, and storage room. Therefore, these variables are used to estimate the price of each property.

In the first phase, data of adverts with possible errors are removed, for example, properties with zero unit price. Subsequently, a descriptive analysis is performed to decide what type of variables to work with. Finally, a multivariate analysis of outliers with the available variables is carried out, based on Mahalanobis distance. After this preliminary analysis, we obtain a substantial database whose elements are uniformly distributed throughout the territory. From this database, we work with those municipalities where there are at least 100 sample observations, 100 homes for sale that allow the procedure to choose the best model in each case. Therefore, our study presents an empirical application developed in 433 municipalities (out of the 8,125 in Spain) which have more than 100 dwellings on sale during the period of research. To be precise, this amounts to information on 790,631 real estate properties distributed in 48 Spanish provinces (out of the 52 in Spain) made up of 433 municipalities.

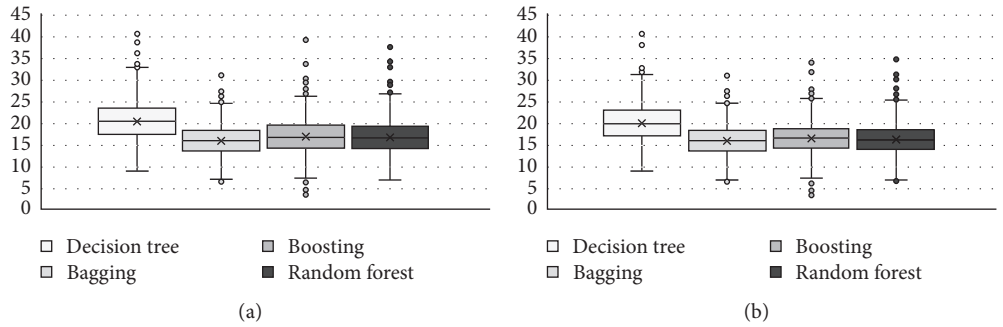


Figure 1: MAPE (a) and COD (b) main error measures of the four techniques (decision tree, bagging, boosting, and random forest) corresponding to the 433 municipalities in Spain.

As it has been said in Section 3, for the application of the different regression techniques, the data set is divided into two; a training set and a test set in a proportion of two-thirds and one-third, respectively. The data in the training set allow to fit the best combination in an automated way for each case, while assessments made on the properties of the test set are used to compare the suitability of the different techniques.

The errors obtained in the valuation of property prices in the 433 municipalities with information available show an average MAPE value of 20.49 with the decision tree technique, 16.54 with bagging, 16.98 with boosting, and 16.69 with random forest, while the average value of the COD was 20.03, 16.03, 16.53, and 16.23, respectively. Statistical dispersion is depicted through the box and whiskers plots in Figure 1. These results show a good performance of bagging, boosting, and random forest techniques given the heterogeneity of the sample and the wide geographic range analysed. However, the decision tree technique, overall, does not give satisfactory results.

A further line of research could be to what extent the error measures depend on the population size and even the sample size, or the number of properties available to be used as examples in the estimate. It is worth finding out whether there is a better or worse behaviour in the methods analysed in small, medium, or large cities, or if the maxim of "larger sample size, better estimates" is met. In fact, Table 1 shows how the results of the 4 techniques show a practically zero linear correlation between the different population and sample sizes with the different error measures. This may be due to both the quality of the starting information and the great arbitrariness present in the prices of housing in Spain. Since the quality of the starting information was controlled in the early stages of the analysis, the second option is considered more plausible.

Nor, are there any nonlinear correlation between error measures and population or sample sizes. Just as an example, Figure 2 shows the scatter plot for two main error measures, MAPE and COD versus population size for bagging results. Four techniques present very similar results. No regression structure can be deduced from the graphs. As observed in Figures 2(a) and 2(c), the biggest cities in Spain, Madrid, and Barcelona could be hiding the real correlation. But after having eliminated both cities (see Figures 2(b) and 2(d)), the graphs do not show any relation between the errors and the population size. Coefficient of determination is stated in Table 2, having computed linear, exponential, potential, and logarithmic coefficients. Note that all of them are almost zero.

Table 1: Linear correlation coefficients between the MAPE and COD values obtained from the 433 municipalities where the valuations and population sizes (inhab.) and sample sizes (N) of said municipalities have been calculated.

| | Technique | | | |
|---|---|---|---|---|
| Correlation coefficient | Decision tree | Bagging | Boosting | Random forest |
| MAPE vs inhab. | 0.10 | −0.07 | −0.04 | −0.07 |
| MAPE vs N | 0.26 | −0.03 | 0.02 | −0.04 |
| COD vs inhab. | 0.09 | −0.07 | −0.03 | −0.06 |
| COD vs N | 0.24 | −0.02 | 0.03 | −0.03 |

Note. Own elaboration.

Because most of the assessments of real estate portfolios will be carried out in the largest cities, we decide to analyse the average results of the municipalities with a population greater than 100,000 inhabitants (see Table 3), 63 in all, and show the results of each of these in more detail. The results obtained for each of the 63 selected municipalities are presented in four tables in the appendix of this study (see Tables S1–S4 in the Supplementary Material for a more detailed analysis), one for each of the techniques used.

Table 3 reports the improvement in all cases when using tree ensemble methods (bagging, boosting, and random forest) compared to individual decision trees, with improvements in the estimation capacity measured through the MAPE of around six percentage points. The average MAPE values for the 63 municipalities show the best performance for bagging and random forest with an average value for the set of 63 municipalities with the largest population of 15.73 and 15.93, respectively. Both techniques achieve the same minimum MAPE value (8.58). However, in terms of a maximum value, although the value reached by random forest is a point higher than that reached by bagging, both achieve very acceptable values. In terms of COD, a similar situation is observed, with the results of bagging and random forest being the best, at around 15%, followed by those of the boosting technique that has an average COD of 16.3%. The

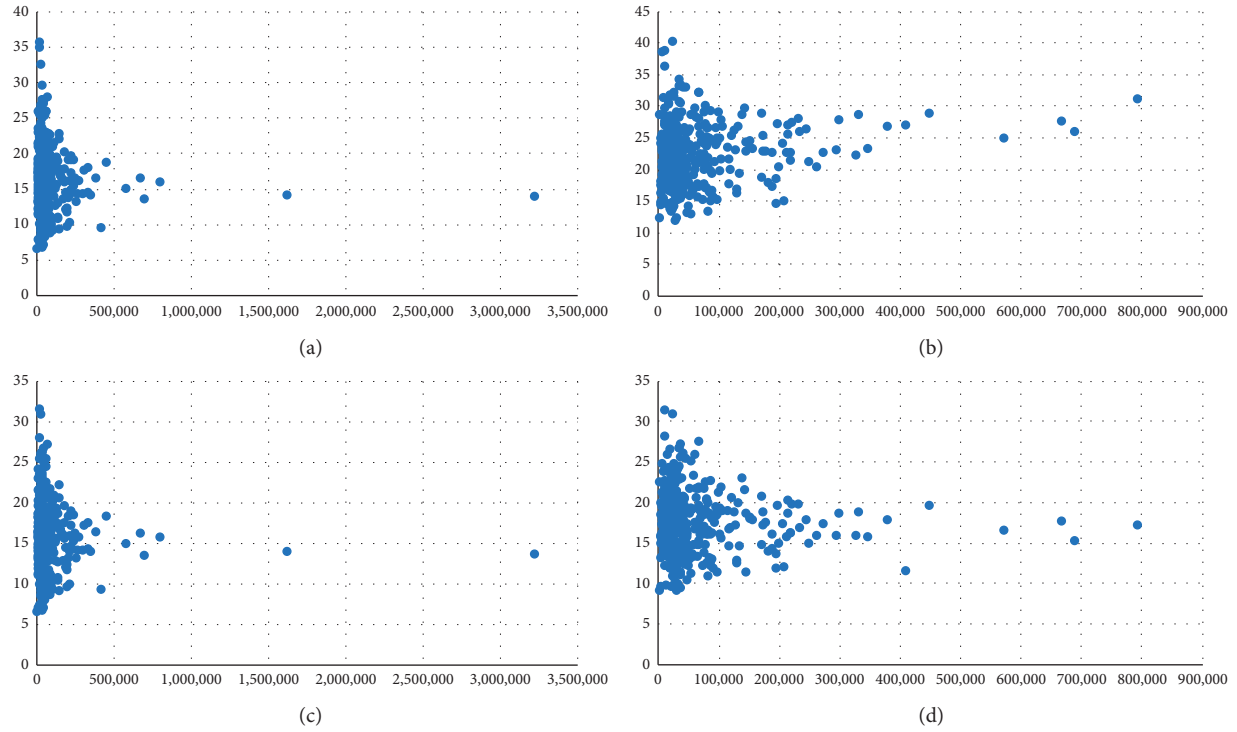


Figure 2: Bagging's MAPE and COD versus municipality population considering all the locations and without Madrid and Barcelona. (a) MAPE vs inhabitants, bagging. (b) MAPE vs inhabitants, bagging without Madrid and Barcelona. (c) COD vs inhabitants, bagging. (d) COD vs inhabitants, bagging without Madrid and Barcelona. Note. Own elaboration.

Table 2: Coefficients of determination of the regression between Bagging's MAPE and COD and municipality population considering all the locations and without Madrid and Barcelona.

| | Technique: bagging | | | |
|---|---|---|---|---|
| Coefficient of determination | Linear | Exponential | Logarithmic | Potential |
| *Considering all locations* | | | | |
| MAPE vs inhab. | 0.0053 | 0.0034 | 0.0075 | 0.0040 |
| COD vs inhab. | 0.0044 | 0.0026 | 0.0049 | 0.0023 |
| *Without Madrid and Barcelona* | | | | |
| MAPE vs inhab. | 0.0252 | 0.0246 | 0.0129 | 0.0148 |
| COD vs inhab. | 0.0053 | 0.0028 | 0.0038 | 0.0017 |

Note. Own elaboration.

Table 3: Average, minimum, and maximum results for MAPE and COD values obtained for the 63 municipalities with more than 100,000 inhabitants.

| | 63 municipalities (>100,000 inhab.) | | | | | |
|---|---|---|---|---|---|---|
| Technique | MAPE average | MAPE minimum | MAPE maximum | COD average | COD minimum | COD maximum |
| Decision tree | 21.92 | 11.68 | 30.40 | 21.22 | 11.65 | 27.95 |
| Bagging | 15.73 | 8.58 | 22.93 | 15.41 | 8.48 | 22.19 |
| Boosting | 16.62 | 10.15 | 23.45 | 16.36 | 10.06 | 23.28 |
| Random forest | 15.93 | 8.58 | 23.92 | 15.56 | 8.48 | 23.08 |

COD of the decision tree technique was alone in achieving a value above the recommended 20%. Figure 3 graphically shows these results along with the dispersion of MAPE and COD measures. Note that all the outliers, municipalities with MAPE, and COD abnormally high or low have disappeared. They corresponded to municipalities with less than 100,000 inhabitants.

From the analysis of errors made in the valuation of properties of the test set for each of the 63 municipalities with more than 100,000 inhabitants of Spain, it is worth highlighting the good behaviour of the mean ratio that, in almost all cases, shows values between 0.98 and 1.1 (see Tables S1–S4 in the Supplementary Material). From the value of MdAPE, in the case of bagging, the smallest value

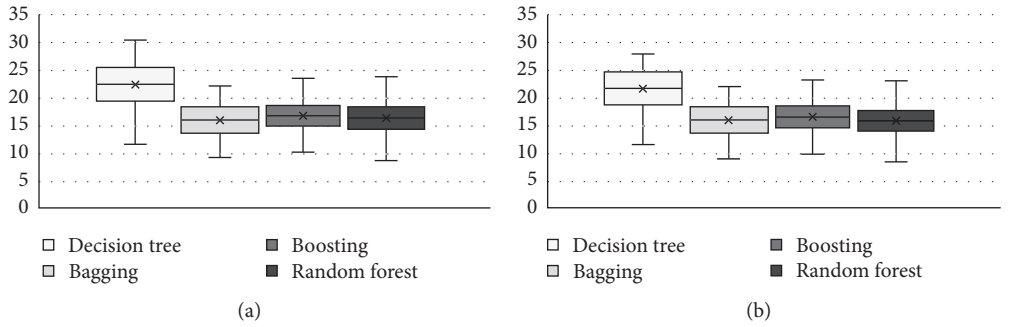


Figure 3: MAPE (a) and COD (b) main error measures of the four techniques (decision tree, bagging, boosting, and random forest) corresponding to the largest 63 municipalities in Spain.

for this measure is 6.59 and the maximum 18.91, which indicates that in the best-case municipalities, fifty percent of the valuations with the lowest error have an error of less than 6.59 and, in the worst-case municipalities, this error is no greater than 18.91. For random forest, these values are 6.15 and 19.64, respectively, reaching a higher value in terms of maximum and a lower value in terms of minimum compared to bagging. In addition, given that choosing between bagging and boosting is made more difficult by the similarity in the results obtained, it should be noted that bagging presents a better behaviour in terms of MdAPE since only 4 of the 63 municipalities analysed present values superior to 16% and 10 superior to 15%, while with random forest these values are 6 and 14, respectively.

## 5. Conclusions

The need for a rapid and economical appraisal of real estate and the greater availability of up-to-date information accessible through the Internet have led to the application of big data techniques and machine learning to carry out real estate valuation.

At the forefront of these machine learning techniques are tree ensemble methods, in particular, bagging, boosting, and random forest. So far, these techniques have been applied in many cases for purposes other than the estimation of property prices, and when they have been applied to real estate valuations, they have been done in a limited way to very specific geographical areas. In order to advance understanding of the value of the techniques of tree ensemble on an automated and massive scale, this study shows the results obtained from the application of different techniques for the whole of Spain with a total of 433 municipalities spread across 48 provinces. The article presents an automated algorithm which selects the best model for each technique in each municipality. Their behaviour in terms of estimation capacity is measured through error measures widely cited in the literature.

The results show that the behaviour of the tree ensemble clearly outperforms individual trees although of the three methods analysed (bagging, boosting, and random forest), none has a clear advantage over the others. Even so, looking more closely at the behaviour of the bagging and random forest methods, it seems that the slightly better results of bagging in terms of MAPE and COD together with the results in terms of MdAPE would make us opt for the use of bagging in the case of Spain.

Reviewing the literature available so far, it can be concluded that the results obtained in terms of MAPE are better than those obtained in [26] with a value of 25.2% in Nicosia or in [46] for the US with 20.9%. The results are similar to those obtained in [14] with a value that ranges from 19.02% to 15.89% in Madrid and worse than those obtained for Ljubljana in [24] with an average MAPE of 7.28. However, it should be borne in mind that these applications focused on specific geographical areas while the application developed in this study covers the entire Spanish territory. The error measures provided are means of the MAPE and COD of each municipality, with municipalities of very different population, sample sizes, and socioeconomic characteristics.

From the global analysis of the 433 municipalities as a whole, it can be concluded that the error measures do not depend on the population size or the size of the sample set. This fact suggests the presence of a certain random component in the determination of sales prices since the greater the available sample information, the better the results should be.

Finally, it should be noted that this study has other active lines of research that are already being developed, such as the inclusion of a dynamic database that allows the handling of information with different temporal references or the inclusion of ensemble methods that allow machine learning techniques, not just simple trees, to be combined.

## Data Availability

The data base used to support the findings of this study were supplied by COHISPANIA, Consultoría y Valoración, under license and so cannot be made freely available. Requests for access to these data should be made to http://www.cohispania.com/contacto.

## Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

## Acknowledgments

The authors want to thank the collaboration of Compañía Hispania de Tasaciones y Valoraciones, S. A. Emilio L. Cano

was partially funded by the Spanish "Agencia Estatal de Investigación" via the MTM2017-86875-C3-1-R AEI/FEDER, UE project. The present work was financed through the R&D contract between the University of Castilla-La Mancha and Cohispania with ref: UCTR180093.

## Supplementary Materials

Table S1: main results for decision trees. Table S2: main results for bagging. Table S3: main results for boosting. Table S4: main results for random forest. (*Supplementary Materials*)

## References

[1] Bank for International Settlements, *International Convergence of Capital Measurement and Capital Standards*, Basel Committee on Banking Supervision, Basel, Switzerland, 2006.

[2] European Council, "Directive 2006/48/EC of the European Parliament and of the Council of 14 June 2006 relating to the taking up and pursuit of the business of credit institutions," *Official Journal of the European Union*, vol. L177, pp. 1–200, 2006.

[3] J. K. Eckert, *Property Appraisal and Assessment Administration*, International Association of Assessing Officers, Chicago, IL, USA, 1990.

[4] V. Kontrimas and A. Verikas, "The mass appraisal of the real estate by computational intelligence," *Applied Soft Computing*, vol. 11, no. 1, pp. 443–448, 2011.

[5] R. Schulz, M. Wersing, and A. Werwatz, "Automated valuation modelling: a specification exercise," *Journal of Property Research*, vol. 31, no. 2, pp. 131–153, 2014.

[6] O. Kettani and M. Oral, "Designing and implementing a real estate appraisal system: the case of Québec Province, Canada," *Socio-Economic Planning Sciences*, vol. 49, pp. 1–9, 2015.

[7] M. Mooya, "Of mice and men," *Urban Studies*, vol. 48, no. 11, pp. 2265–2281, 2011.

[8] W. McCluskey and S. Anand, "The application of intelligent hybrid techniques for the mass appraisal of residential properties," *Journal of Property Investment & Finance*, vol. 17, no. 3, pp. 218–239, 1999.

[9] C. M. Filho and O. Bin, "Estimation of hedonic price functions via additive nonparametric regression," *Empirical Economics*, vol. 30, no. 1, pp. 93–114, 2005.

[10] H. Selim, "Determinants of house prices in Turkey: hedonic regression versus artificial neural network," *Expert Systems with Applications*, vol. 36, no. 2, pp. 2843–2852, 2009.

[11] N. García, M. Gámez, and E. Alfaro, "ANN+GIS: an automated system for property valuation," *Neurocomputing*, vol. 71, no. 4–6, pp. 733–742, 2008.

[12] R. D. Jaen, "Data mining: an empirical application in real estate valuation," in *FLAIRS Conference*, S. M. Haller and G. Simmons, Eds., pp. 314–317, AAAI Press, Palo Alto, CA, USA, 2002.

[13] G.-Z. Fan, S. E. Ong, and H. C. Koh, "Determinants of house price: a decision tree approach," *Urban Studies*, vol. 43, no. 12, pp. 2301–2315, 2006.

[14] C. Del Cacho, *A Comparison of Data Mining Methods for Mass Real Estate Appraisal*, University Library of Munich, Munich, Germany, 2010, https://mpra.ub.uni-muenchen.de/id/eprint/27378MPRA Paper No. 27378,.

[15] E. A. Antipov and E. B. Pokryshevskaya, "Mass appraisal of residential apartments: an application of Random forest for valuation and a CART-based approach for model diagnostics," *Expert Systems with Applications*, vol. 39, no. 2, pp. 1772–1778, 2012.

[16] M. Thériault, F. Des Rosiers, and F. Joerin, "Modelling accessibility to urban services using fuzzy logic," *Journal of Property Investment & Finance*, vol. 23, no. 1, pp. 22–54, 2005.

[17] N. Pow, E. Janulewicz, and L. Liu, "Applied machine learning project 4 prediction of real estate property prices in montreal," 2014, http://rl.cs.mcgill.ca/comp598/fall2014/comp598_submission_99.pdf.

[18] O. Bin, "A prediction comparison of housing sales prices by parametric versus semi-parametric regressions," *Journal of Housing Economics*, vol. 13, no. 1, pp. 68–84, 2004.

[19] Shabana, G. Ali, M. K. Bashir, and H. Ali, "Housing valuation of different towns using the hedonic model: a case of Faisalabad city, Pakistan," *Habitat International*, vol. 50, pp. 240–249, 2015.

[20] J. I. Pérez-Rave, J. C. Correa-Morales, and F. González-Echavarría, "A machine learning approach to big data regression analysis of real estate prices for inferential and predictive purposes," *Journal of Property Research*, vol. 36, no. 1, pp. 59–96, 2019.

[21] J. Bin, S. Tang, Y. Liu et al., "Regression model for appraisal of real estate using recurrent neural network and boosting tree," in *Proceedings of the 2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA)*, IEEE, Beijing, China, pp. 209–213, September 2017.

[22] R. A. Dubin, "Predicting house prices using multiple listings data," *The Journal of Real Estate Finance and Economics*, vol. 17, no. 1, pp. 35–59, 1998.

[23] O. Özsoy and H. Şahin, "Housing price determinants in Istanbul, Turkey," *International Journal of Housing Markets and Analysis*, vol. 2, no. 2, pp. 167–178, 2009.

[24] M. Ceh, M. Kilibarda, A. Lisec, and B. Bajat, "Estimating the performance of random forest versus multiple regression for predicting prices of apartments," *International Journal of Geo-Information*, vol. 1, p. 168, 2018.

[25] A. Nguyen, "Housing price prediction," 2018, https://pdfs.semanticscholar.org/782d/3fdf15f5ff99d5fb6acafb61ed8e1c60fab8.pdf.

[26] T. Dimopoulos, H. Tyralis, N. P. Bakas, and D. Hadjimitsis, "Accuracy measurement of random forests and linear regression for mass appraisal models that estimate the prices of residential apartments in Nicosia, Cyprus," *Advances in Geosciences*, vol. 45, pp. 377–382, 2018.

[27] M. Skurichina and R. P. W. Duin, "Bagging, boosting and the random subspace method for linear classifiers," *Pattern Analysis & Applications*, vol. 5, no. 2, pp. 121–135, 2002.

[28] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*, Chapman & Hall, New York, NY, USA, 1993.

[29] J. Elith, J. R. Leathwick, and T. Hastie, "A working guide to boosted regression trees," *Journal of Animal Ecology*, vol. 77, no. 4, pp. 802–813, 2008.

[30] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[31] N. Kok, E.-L. Koponen, and C. A. Martínez-Barbosa, "Big data in real estate? From manual appraisal to automated valuation," *The Journal of Portfolio Management*, vol. 43, no. 6, pp. 202–211, 2017.

[32] N. Shinde and K. Gawande, "Valuation of house price using predictive techniques," *International Journal of Advances in Electronics and Computer Science*, vol. 5, no. 6, pp. 34–40, 2018.

[33] M. Kagie and M. V. Wezel, "Hedonic price models and indices based on boosting applied to the Dutch housing market,"

*Intelligent Systems in Accounting, Finance & Management*, vol. 15, no. 3-4, pp. 85–106, 2007.

[34] J. H. Friedman, "Machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[35] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, Wadsworth, Belmont, CA, USA, 1984.

[36] S. D. Clark and N. Lomax, "A mass-market appraisal of the English housing rental market using a diverse range of modelling techniques," *Journal of Big Data*, vol. 5, no. 1, p. 43, 2018.

[37] M. Graczyk, T. Lasota, B. Trawiński, and K. Trawiński, "Comparison of bagging, boosting and stacking ensembles applied to real estate appraisal," in *Intelligent Information and Database Systems. ACIIDS 2010. Lecture Notes in Computer Science*, R. Goebel, J. Siekmann, and W. Wahlster, Eds., vol. 5991, pp. 340–350, Springer, Berlin, Germany, 2010.

[38] T. Lasota, B. Londzin, Z. Telec, and B. Trawiński, "Comparison of ensemble approaches: mixture of experts and AdaBoost for a regression problem," in *Intelligent Information and Database Systems. ACIIDS 2014, Lecture Notes in Computer Science*, N. T. Nguyen, B. Attachoo, B. Trawiński, and K. Somboonviwat, Eds., vol. 8398, Springer, Cham, Switzerland, 2014.

[39] S. Yoo, J. Im, and J. E. Wagner, "Variable selection for hedonic model using machine learning approaches: a case study in Onondaga County, NY," *Landscape and Urban Planning*, vol. 107, no. 3, pp. 293–306, 2012.

[40] B. Park and J. K. Bae, "Using machine learning algorithms for housing price prediction: the case of Fairfax County, Virginia housing data," *Expert Systems with Applications*, vol. 42, no. 6, pp. 2928–2934, 2015.

[41] M. Shahhosseini, G. Hu, and H. Pham, "Optimizing ensemble weights for machine learning models: a case study for housing price prediction," *Smart Service Systems, Operations Management, and Analytics*, Springer, Berlin, Germany, 2019, https://lib.dr.iastate.edu/imse_conf/185/.

[42] D. Harrison and D. L. Rubinfeld, "Hedonic housing prices and the demand for clean air," *Journal of Environmental Economics and Management*, vol. 5, no. 1, pp. 81–102, 1978.

[43] D. De Cock, "Ames, Iowa: alternative to the Boston housing data as an end of semester regression project," *Journal of Statistics Education*, vol. 19, no. 3, p. 115, 2011.

[44] A. Neloy, M. Sadman Haque, and M. Mahmud Ul Islam, "Ensemble learning based rental apartment price prediction model by categorical features factoring," in *Proceedings of the 2019 11th International Conference on Machine Learning and Computing*, pp. 350–356, Zhuhai, China, 2019.

[45] E. B. Pokryshevskaya and E. A. Antipov, "Applying a CART-based approach for the diagnostics of mass appraisal models," *Economics Bulletin*, vol. 31, no. 3, pp. 2521–2528, 2011.

[46] S. Mullainathan and J. Spiess, "Machine learning: an applied econometric approach," *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 87–106, 2017.

[47] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[48] R. E. Schapire, "The strength of weak learnability," *Machine Learning*, vol. 5, pp. 197–227, 1990.

[49] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2019, https://www.R-project.org/.

[50] T. Therneau and B. Atkinson, *Rpart: Recursive Partitioning and Regression Trees. R Package Version 4.1-15*, 2019, https://CRAN.R-project.org/package=rpart.

[51] B. Greenwell, B. Boehmke, and J. Cunningham, *Gbm: Generalized Boosted Regression Models. R package version 2.1.5*, 2019, https://CRAN.R-project.org/package=gbm.

[52] A. Liaw and M. Wiener, "Classification and regression by random forest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[53] M. Kuhn, J. Wing, and S. Weston, *Caret: Classification and Regression Training. R Package Version 6.0-84*, 2019, https://CRAN.R-project.org/package=caret.

[54] M. E. Beręsewicz, "On representativeness of Internet data sources for real estate market in Poland," *Austrian Journal of Statistics*, vol. 44, no. 2, pp. 45–57, 2015.

[55] A. Cavallo, *Scraped Data and Sticky Prices*, Social Science Research Network, Rochester, NY, USA, SSRN Scholarly Paper ID 1711999, 2012.

[56] A. Cavallo, "Are online and offline prices similar? Evidence from large multi-channel retailers," *American Economic Review*, vol. 107, no. 1, pp. 283–303, 2017.

[57] B. Larraz, "An expert system for online residential properties valuation," *Review of Economics & Finance*, vol. 2, pp. 69–82, 2011.

[58] B. Larraz and J. Población, "An online real estate valuation model for control risk taking: a spatial approach," *Investment Analysts Journal*, vol. 42, no. 78, pp. 83–96, 2013.

[59] E. Hromada, "Mapping of real estate prices using data mining techniques," *Procedia Engineering*, vol. 123, pp. 233–240, 2015.