Journal of Property Research

ISSN: 0959-9916 (Print) 1466-4453 (Online) Journal homepage: www.tandfonline.com/journals/rjpr20

A machine learning approach to big data regression analysis of real estate prices for inferential and predictive purposes

Jorge Iván Pérez-Rave, Juan Carlos Correa-Morales & Favián González-Echavarría

To cite this article: Jorge Iván Pérez-Rave, Juan Carlos Correa-Morales & Favián González-Echavarría (2019) A machine learning approach to big data regression analysis of real estate prices for inferential and predictive purposes, Journal of Property Research, 36:1, 59-96, DOI: 10.1080/09599916.2019.1587489

To link to this article: https://doi.org/10.1080/09599916.2019.1587489

JOURNAL OF PROPERTY RESEARCH
2019, VOL. 36, NO. 1, 59–96
https://doi.org/10.1080/09599916.2019.1587489

Routledge
Taylor & Francis Group

A machine learning approach to big data regression analysis of real estate prices for inferential and predictive purposes

Jorge Iván Pérez-Rave [a], Juan Carlos Correa-Morales [b] and Favián González-Echavarría [c]

[a]Grupo de investigación IDINNOV, IDINNOV S.A.S, Medellín, Colombia; [b]Escuela de Estadística, Universidad Nacional de Colombia, Medellín, Colombia; [c]Departamento de Ingeniería Industrial, Universidad de Antioquia, Medellín, Colombia

ABSTRACT

The hedonic price regressions have mainly been used for inference. In contrast, machine learning employed on big data has a great potential for prediction. To contribute to the integration of these two strategies, this article proposes a machine learning approach to the regression analysis of big data, viz. real estate prices, for both inferential and predictive purposes. The methodology incorporates a new procedure of selecting variables, called 'incremental sample with resampling' (MINREM). The methodology is tested on two cases. The first is data from web advertisements selling used homes in Colombia (61,826 observations). The second considers the data (58,888 observations) from a sample of the Metropolitan American Housing Survey 2011 obtained and prepared by a reference study. The methodology consists of two stages. The first chooses the important variables under MINREM; the second focuses on the traditional training and validation procedure for machine learning, adding three activities. In both test cases, the methodology shows its value for obtaining highly parsimonious and stable models for different sample sizes, as well as taking advantage of the inferential and predictive use of the obtained regression functions. This paper contributes to an original methodology for big data regression analysis.

ARTICLE HISTORY
Received 21 December 2018
Accepted 22 February 2019

KEYWORDS
Regression analysis; real estate; machine learning; big data; variable selection

1. Introduction

The real estate industry is of interest to academics, practitioners, and governments due to its implications for decision-making in several economic sectors, such as the financial, construction, valuation, and public sectors, among others. Price modeling is one of the most popular topics in the study of real estate because price is the primary proxy for the value of a property (Adetiloye & Eke, 2014).

The study of real estate prices has been guided mainly by the theory of hedonic prices (Rosen, 1974), which has been well accepted and recognized as viable and practicable (Oladunni & Sharma, 2016). In this theory, a property is a composite and heterogeneous object, whose market price is a function of the utility of structural characteristics (e.g. area, bathrooms, rooms), of neighborhoods (i.e., garden, access roads) and environmental

CONTACT Jorge Iván Pérez-Rave investigacion@idinnov.com
Supplemental material for this article can be accessed here.

characteristics (e.g. proximity to parks) (Freeman, Herriges, & Kling, 2014). In this view, the modeling of real estate prices has been developed usually through a multiple regression analysis (Oladunni & Sharma, 2016), with which the relations between the price and several hedonic attributes of a property are studied from a parametric perspective. This perspective demands the fulfillment of a series of statistical assumptions, such as the normality of the residuals, homoscedasticity, independence, and the absence of multicollinearity. From an economic point of view, the use of this type of model has been oriented more towards inference than prediction, guided by questions such as evaluating and estimating the economic value of the hedonic characteristics (Yoo, Im, & Wagner, 2012), identifying which characteristics are most significant for explaining the phenomenon under study (Athey, 2018), determining the causal effect of specific regressors (Athey, 2018) and, in general, how to improve the interpretation of the functions obtained (Mullainathan & Spiess, 2017). Although the hedonic price regression approach has been contributing to the needs of inference for decades, its predictive potential, by itself, is not as mature.

An emerging alternative, increasingly promising for the prediction of real estate prices, is machine learning (Borde, Rane, Shende, & Shetty, 2017; Mullainathan & Spiess, 2017; Trawiński et al., 2017). It is a supervised learning approach (guided by a response variable, in this case, the price of the property), intensive operations and automatic learning technologies, and is useful for prediction using big data. Machine Learning uses at least two subsamples, one for training and another for validation. In this last sample, the predictive capacity of the models under test is evaluated. Despite the benefits of machine learning to meet the needs of prediction, its contribution to the needs of inference is minimal (Athey, 2018; Mullainathan & Spiess, 2017). That is, all the effort is on designing and choosing, among several rival models, the one with the highest predictive capacity in the non-training sample. This tends to neglect inferential analysis, which helps to understand the relevant hedonic attributes, their marginal effects, and other related economic analyses.

The combination of hedonic regression and machine learning can provide valuable opportunities for analyzing real estate prices while facilitating the joint execution of inferential and predictive analysis, based on the obtained models. However, its combination is not elementary, because when using regression from a machine learning perspective to analyze big data, the inferential treatment is still limited, due to reasons such as: (1) the estimates of the coefficients are generally not consistent; in fact, they do not tend to fulfill the essential prerequisites of estimation problems (Mullainathan & Spiess, 2017). (2) Regularization procedures are usually carried out, in order to reduce overfitting in non-training samples; however, this practice can lead to biases due to the omission of variables or, in general, to erroneously specified models (Mullainathan & Spiess, 2017). (3) Because of the high correlations between the variables (typical of big data), models tend to be unstable: that is, the obtained predictors (and their parameters) vary depending on the sample, which means that the resulting models are usually unstable (Cateni & Colla, 2016; Čeh, Kilibarda, Lisec, & Bajat, 2018; Mullainathan & Spiess, 2017). (4) Any minimal difference in the effects of the predictors, even if it is not important from a practical point of view, is statistically significant, due to the high power of the tests.

These shortcomings are to be expected, since the use of hedonic regression from the perspective of machine learning is practically equivalent, in procedural terms, to using any other method of this last approach. For example, the theory of hedonic prices does

not provide any guidance on how to choose the set of significant variables to deduce an eventual hedonic model, nor on its functional form (Anderson, 2000; Freeman, Herriges & Kling, 2014; Chen & McCluskey 2018). Therefore, it is necessary to incorporate other strategies to guide these decisions, while meeting the criteria of importance (only those variables that better summarize the observed phenomenon), efficiency (from the computational point of view and the practical waiting time) and improvement of knowledge (interpretation and understanding of the obtained function) (Cateni & Colla, 2016).

In search of hedonic regression solutions based on machine learning for big data that will facilitate attending to the needs of inference and prediction, a useful tool is the selection of the variables (or features). This consists in choosing a subset of the variables, from the original set, through some selection criteria, in order to obtain the variables that are most important for understanding the phenomenon under study, with the minimum redundancy (Bin et al., 2017; Cai, Luo, Wang, & Yang, 2018; Cateni & Colla, 2016). Thus, the selection of variables plays a preponderant role for the exclusion of those hedonic attributes that are redundant or irrelevant, while facilitating the parsimony and interpretation of the models, computational efficiency, and reducing the risk of overfitting (Bin et al., 2017; Cai et al., 2018; Cateni & Colla, 2016). A fundamental aspect in the selection of variables is stability, understood in terms of the sensitivity of the model to changes in the training sample. This aspect needs to be addressed if we seek to take advantage of the inferential potential of the obtained prediction model (Cateni & Colla, 2016). Hence, Banerjee and Dutta (2017) state that the identification of the optimal number of attributes is the greatest challenge in the study of real estate prices. Additionally, Varian (2014) points out that although machine learning models do not address causal inference, they can help to estimate causal effects when these effects occur; therefore, they emphasize that it is essential to have a mechanism for the selection of variables that helps eliminate variables that generate confusion and hide these effects.

The objective, then, is to propose a methodology for the regression analysis of big data using a machine learning approach, for real estate prices, for both inferential and predictive purposes. This methodology incorporates a new variable selection procedure, called 'incremental sample with resampling' (MINREM).

The validation of the methodology considers two cases, which are summarized below.

The first case uses data collected from web advertisements selling used homes in Colombia. This data was provided by the IDINNOV research group, as a result of the statihouse® project (Pérez-Rave, 2019). This type of ad on the internet has the characteristics of big data; for example, volume (and it is growing daily), variety (structured, semi-structured, and unstructured formats), and value (close relations between the offline and the online prices, as well as a variety of characteristics of the property and of the neighborhood). This online data source has been little used, but today there are some successful experiences, both in real estate and in other sectors (see Beręsewicz, 2015). According to Cavallo (2012, 2017), online price data constitutes a valuable opportunity for statistical analysis, due to the high frequency of its generation, its accessibility, and its availability. Besides, online prices do not tend to present notable differences from offline prices (Beręsewicz, 2015; Cavallo, 2017). Moreover, the data comes from an emerging country, in which the use of massive real estate data and its associated technological tools is in its infancy. In fact, there are still temporary delays in price reports and other characteristics of the sector, as well as high costs in the

administration of face-to-face surveys and newsletters that are limited to descriptive analyses or price evolution. These practical situations generate opportunities for the development of new modeling procedures using machine learning which will contribute both to inference and prediction (Pérez-Rave, 2019). Moreover, this type of solution is in sync with what is being demanded, in Colombia, from the Science and Technology Policy and the Ministry of Technologies, concerning strategies to promote smart cities and the use of mass data.

The second case study is external, using to the dataset obtained and prepared by Mullainathan and Spiess (2017) from a sample of the Metropolitan American Housing Survey 2011. To take advantage of this case, we first analyzed the algorithm shared by those authors (https://www.aeaweb.org/articles?id=10.1257/jep.31.2.87). Then, we identified the code extracts in which these authors loaded the data and stored the results of one of their models (the regression one). Then, using our code, we reproduced the results of that study (only of the OLS regression, they also used other methods). After that, we tested the methodology proposed in this article and made the respective comparisons. Note that this second case uses a different data source than the first case and, also, from a different country, which enriches the evidence. Besides, the study by Mullainathan and Spiess (2017) has become one of the leading international benchmarks in the analysis of real estate from the perspective of machine learning, so it is an important comparative resource for evaluating the performance of the methodology.

In itself, the proposed methodology seeks to contribute to the generation of inputs that can motivate and serve as a guide to meet the needs of: (1) The intersection between economics and machine learning (Athey, 2018; Mullainathan & Spiess, 2017), which merits new strategies to identify the important variables in the context of big data and the development of stable machine learning models, whose variables do not depend substantially on the chosen sample; (2) The analysis of online prices, as a new statistical resource to address big data (Beręsewicz, 2015; Cavallo, 2012); (3) Better use of the functions obtained with machine learning, so as to favor their interpretation and the comprehension of the phenomenon under study (Mullainathan & Spiess, 2017); (4) The use of strategies based on resampling, which, in the face of data-driven methods, helps to overcome problems of high sample dependence, excessive testing power, and standard errors that are unreliable/invalid, or even not estimated (when they use computational methods other than regression) (Athey, 2018). (5) Stimulating and facilitating the scrutiny of the findings in the field of real estate (Krause, 2016). So, this study makes public the general code employed (using *Rmarkdown*) for the treatment and analysis of the data.

The remainder of this paper is organized as follows. Section 1 argues for the need for this study. Section 2 presents the proposed methodology and describes the two test cases. Section 3 presents the descriptive, inferential, predictive and comparative results, including a discussion of these in the light of the evidence. Section 4 provides the main general conclusions derived from the study.

2. Literature review

Nowadays, there is prevalent an overwhelming production of data, with exponential growth in practically all organizational spheres. This phenomenon has generated paradigms such as big data, which demand reproducible strategies and methods to extract,

prepare, process, analyze and visualize information derived from the large volumes of available data (Abdallah & Khashan, 2016; Leung & Jiang, 2014; Varian, 2014). These effects of the information economy also permeate the real estate industry (C.A.R., 2015); according to Puyun and Miao (2016), the exponential growth of the data produced today on real estate is part of the 5V of big data: volume, speed, variety, truth, and value. Besides, the real estate industry is motivating reflections on the importance of dealing with big data, through new strategies for data preparation, analysis and visualization, in order to benefit from it and generate competitive advantages. This is an emerging research field (Holland, 2016; C.A.R., 2015). Among the opportunities in this sector are: mitigating the subjectivity present in traditional methods to estimate the commercial value of a home; predicting and monitoring trends in real time; analyzing the behavior of income and housing demands and their attributes; and generating price maps, hedonic characteristics and development zones, among other needs (Holland, 2016). In other words, there are inferential and predictive challenges.

Regarding prediction, it is worth mentioning that the accurate and efficient prediction of real estate prices has been and will continue to be a fundamental issue, with an impact on the diverse actors (e.g. buyers, sellers, commission agents) and institutions (e.g. government, banks), but it is also controversial (Bin et al., 2017; Dubin, 1998). Now, considering the big data paradigm, the situation is even more challenging and promising, and one of the main approaches to face big data is machine learning. The main benefits of machine learning are the automatic (or semi-automatic) induction of predictive-dynamic models (which learn as the data is updated), as well as the prevention of overfitting, through the division of the sample into at least two parts: the training data (generally 70% of the data) and another dataset for performance validation. In itself, the goal is to obtain models that present the best predictive capacity in non-training samples, which favor the reduction of overfitting and provide empirical guarantees for subsequent uses. Also, machine learning offers a variety of computationally intensive methods to generate competing models for the sake of choosing the best performance for the specific scenario under study.

Among the recent applications of machine learning in real estate, Čeh et al. (2018) compares the predictive performance of random forest (an algorithm based on assembling trees under random subspaces) against multiple regression in the context of apartment prices in Ljubljana, the capital of the Republic of Slovenia. The dataset consists of 7,404 records. The independent variables fall under four typologies: structural/time (e.g. age, areas), accessibility (e.g., proximity to schools, transport routes), neighborhoods (e.g. local unemployment rates) and environmental (e.g., noise, visibility). Čeh et al. (2018) used principal component analysis for the induction of the multiple regression model, while for the model based on random forest they used the ten most essential predictors according to this method (out of a total of 36 original variables). For the performance evaluation, they use MAPE (mean average percentage error) and COD (coefficient of dispersion); the latter is classic in real estate valuation procedures. They conclude that random forest yielded a better performance than multiple regression. However, from a critical point of view, it is worth mentioning that they did not report whether the response variable (price) was transformed before carrying out the regression, as is traditionally suggested to mitigate the high asymmetry in this phenomenon. Also, the methods they analyzed may not have been entirely comparable, since in the regression

they used principal components as predictor variables whereas in random forest they used the original variables. On the other hand, they make a valuable interpretation of the principal components (e.g. quality of the apartment and size, the latter consisting of the variables describing the area and the number of rooms).

Additionally, one of the most recognized studies on the subject is that of Mullainathan and Spiess (2017), which discusses the importance and implications of machine learning for econometrics. These authors provide a framework for the development of applications in this regard. Also, they provide an application case that uses data from the American Housing Survey of 2011, using more than 50,000 records and more than 150 covariates. Regarding this application, they report that multiple regression (under OLS) explained 41.7% (R^2) of the variability in the non-training sample, being higher than regression trees (34.5%), but lower than LASSO (43.3%), random forest (45.5%) and ensemble (45.9%). They emphasize the value of these two last methods for their predictive capacity. However, it is worth mentioning that random forest and ensemble presented notable over-fitting, considering their performance in the training sample (85.1% and 80.4%, respectively). Additionally, they provide the code used for the treatment, analysis, and visualization of the data, which facilitates the reproducibility of the results and the development of new proposals on the subject. Likewise, they report several challenges for the use of machine learning with economics (e.g., inference, instability of models), some of them already described in the Introduction to the present article.

Despite the benefits of machine learning for prediction, its contribution to the processes of inference is virtually nil. According to Mullainathan and Spiess (2017), machine learning addresses the needs of prediction, but most applications in economics revolve around the estimation and interpretation of the parameters of the model. Athey (2018) is consistent with this, pointing out that causal inference is typical in the economics literature, but that machine learning does not address the needs of estimation. Overcoming this limitation is crucial in the context of real estate, since the interpretation of the obtained function, the identification of the most important variables, and the estimation of the marginal effects, among other things, are fundamental topics for the generation of knowledge and economic decision making. In fact, in the field of real estate the theory of hedonic prices is worth employing (Fletcher, Gallimore & Mangan, 2000; Rosen, 1974): it has been contributing significantly to inferential purposes, since it allows understanding the price of a property as the sum of the contributions of internal and external characteristics attributable to the property. Hence, econometric models are the main practical drivers of this theory (Čeh et al., 2018; Yoo et al., 2012). However, when analyzing big data from such an econometric point of view, several limitations are presented, among them: (1) the impossibility of manual – dynamic treatment, due to the enormous volume of data that is produced daily, as well as its characteristics (Varian, 2014); (2) the instability of the models, due to the high correlation between the variables (Cateni & Colla, 2016; Čeh et al., 2018; Mullainathan & Spiess, 2017); (3) the high risk of irrelevant variables, as a consequence of the high power of the parametric tests.

So, either from machine learning or from the econometric approach, the analysis of big data in the context of real estate is limited in its ability to simultaneously satisfy the needs of inference and prediction. For the same reason, studies like Varian (2014), Mullainathan and Spiess (2017), and Athey (2018) draw attention to the importance of integrating these two

fields, in order to provide a more comprehensive use for the functions created from the data. In this sense, the use of variable selection methods is important (Bin et al., 2017; Cai et al., 2018; Cateni & Colla, 2016; Varian, 2014), to seek to reduce the set of variables to those that are more important, while favoring a reduction of the computational burden (efficiency), and help the interpretation, prediction and theorization about the phenomenon of study. For example, Banerjee and Dutta (2017) used data from a machine learning competition from the kaggle.com platform and were interested in identifying the number of attributes that contribute most to the prediction of the direction of movement of the price of the property (binary: 1 is a price increase; 0 is a decrease). To do this, they used: VIF (inflation value of variance) <2, principal component analysis (PCA), and a non-parametric criterion of the value of the information. The PCA is a useful method to mitigate multicollinearity problems and redundancy in the set of variables, before carrying out the model training. Among the publications that have resorted to this method, in the context of real estate, there are: Oladunni and Sharma (2016) (properties in Washington DC metropolitan area, comparison between principal component regression, support vector regression and *k*-nearest neighbor; principal component regression had a slight edge over the others; 14 observable variables, 5 components explained at least 85% of the variance); Wang and Zhang (2013) (influencing factors in the development of China's real estate market using PCA, 12 observable variables, 2 components explained 96.377% of the variance); and Shi (2009) (Chinese real estate market, PCA and artificial neural networks; 17 observable variables, 7 components explained 89.7% of the variance).

There are three types of procedures for selecting the variables. The first are the filter methods (Cateni & Colla, 2016), which do not depend on the type of model. Instead, they depend on a metric that relates the dependent variable to the candidate predictors, either from a bivariate perspective or multivariate. Among the metrics used are: Pearson correlation, distance measurements, or coefficients based on information theory (information value, Banerjee & Dutta, 2017). This typology is a pre-processed strategy, before carrying out the training of the model. Another method within this category is PCA, through which we seek to obtain a lower number of theoretically independent and latent variables (linear combinations), which summarize the original variables as best as possible. Some applications that use this technique are Shi (2009) and Oladunni and Sharma (2016). The second type of variable selection method is that of wrappers (Cateni & Colla, 2016), which are based on some type of sequence to exploit the performance of the learning models. Among the most recognized, from the econometric perspective, are forward selection and backward selection, which, iteratively, define the final set of variables, based on a stopping criterion for the model (AIC, R^2, ...). The other type of method is embedded (Cateni & Colla, 2016), which is executed as part of the training of the model, considering all the variables first and then determining the importance of each one. Some of these methods are regression (or classification) trees, random forest and those based on regularization. Their main advantage is the closeness to the nature of the training algorithm. However, it should be remembered that when training (including the selection of variables) is based merely on machine learning methods, they, despite their predictive potential, they act as a black box (James et al., 2015), in the sense that they limit the understanding of and interpreting the effects of the predictors on the dependent variable. In addition, the resulting models do not

necessarily possess inferential stability (but predictive) properties, as evidenced by Mullainathan and Spiess (2017).

Therefore, in this study, it is important to take advantage of the potential simplicity and practicality of the hedonic regression for the measurement and interpretation of the hedonic effects, as well as take advantage of the potential of machine learning for the prediction and treatment of big data. Additionally, the literature has warned one that the integration of these two perspectives requires a strategy for the selection of the variables, one that will help reduce the space of attributes, excluding irrelevant or redundant variables. In this sense, a new method of selecting variables will be proposed, based on resampling and an analysis of the stability of the variables (as the sample size increases). With all this, machine learning, hedonic regression, and selection of variables are intended to enable the training of more stable and interpretable models, which facilitate and make viable a better understanding and use of model induction driven by big data.

3. Methodology

The proposed methodology is detailed in Figure 1. It has of two stages. The first consists in the selection of the relevant variables, under a proposed strategy called incremental sample with resampling (MINREM). The second stage is the classical training and validation of models from the machine learning approach, with three additional activities: (1) a comparison of the debugged model under MINREM with the model with all variables; (2) the inclusion of double interactions for the refined model; and (3) the estimation of metrics not only on a transformed scale (classically the natural logarithm of the property price) but also on the original monetary scale.

Emphasizing, initially, in MINREM, in general, this strategy consists in establishing an incremental set of m sample sizes ($N_1, N_2, \ldots N_m$); for each sample size N_i, k samples (repetitions) with replacement are obtained. Then, from each sample, a model is trained, and the regression coefficients are estimated (for all available variables). Then, considering these coefficients for the k samples, we calculate the proportion of cases in which each variable was significant (in the regression, p-values less than 0.05, in other methods, such as regression trees, the inclusion of the variable in the final tree). A variable is important if it was significant in at least 95% of the cases, otherwise, it will be considered irrelevant. Next, a refined model is trained using the subset consisting of only the variables that, under MINREM, are stable-important considering *Ndepura* (maximum sample size that will be tolerated for a variable to meet the criterion of being important). That is, it seeks to refine the whole set of independent variables so that it emphasizes the 'few' that are 'vital,' in analogy with the Pareto principle.

The MINREM strategy, in the framework of the proposed methodology, was tested on two cases (sets of real estate data). The first comes from web advertisements about used homes for sale in Colombia; for this dataset (61,826 observations), the model of interest (hedonic regression) will be tested, comparing the results of with those from the traditional training vs. MINREM. Additionally, we consider the results of regression trees, since they are a non-parametric method of machine learning based on information theory, which, in some cases, is also used as an embedded variable selection procedure (Cateni & Colla, 2016). It is also worth saying that, in this first study case,

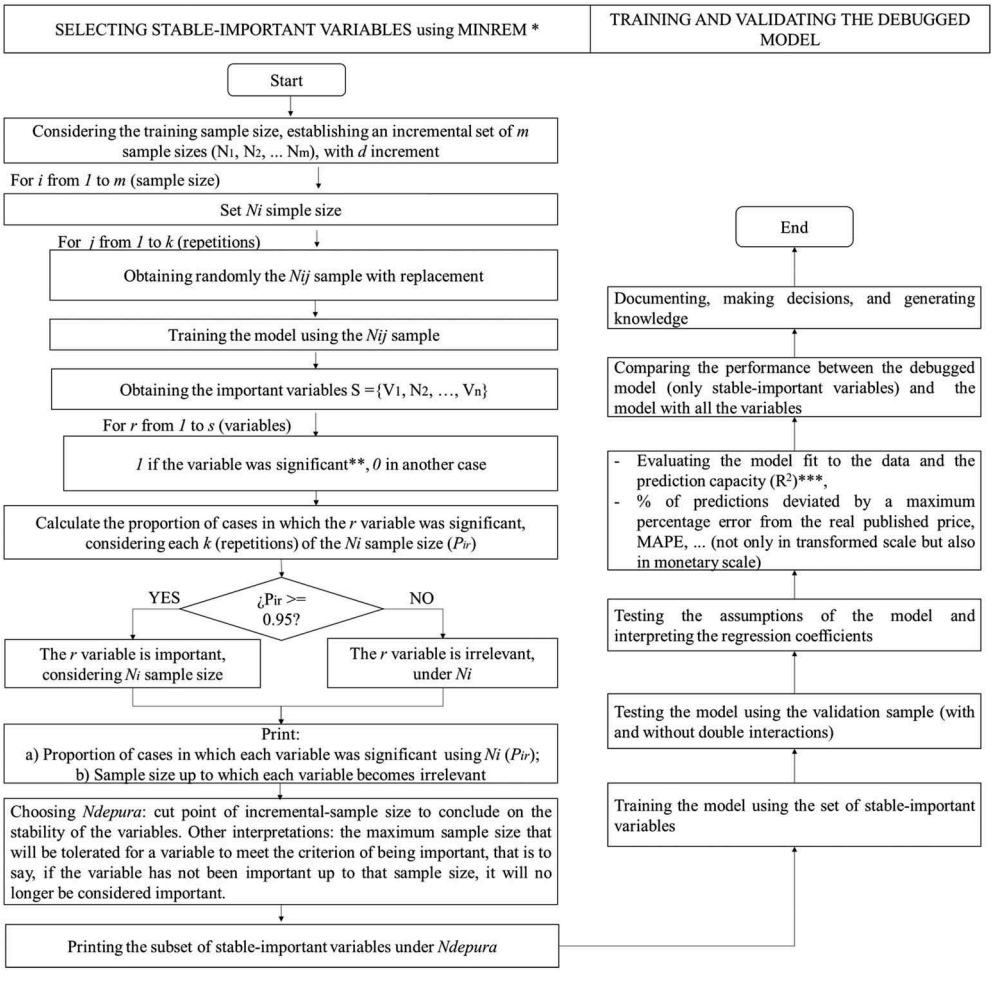

Figure 1. Stages and steps of the methodology used and details of the MINREM debugging strategy.

* Incremental sample with resampling; ** In the regression, p-values less than 0.05, in other methods, such as regression trees, the inclusion of the variable in the final tree; *** Square of the correlation between predicted and real values.

the regression model will not only be meant for predictive purposes, but also inferential, in order to take advantage of the interpretive richness of the resulting function. The second case study uses the dataset that Mullainathan and Spiess (2017) cleaned, processed and analyzed to train and validate several methods from a machine learning approach (among them, regression). We will compare the OLS regression trained by them with a regression model trained using MINREM.

A function custom-designed by us, called 'min_rem.R,' will facilitate the use of the MINREM training strategy. The following lines explain this function:

- *base.dat*: the database with all records and all variables (must be indicated).
- *prop.entrena*: the proportion of records that will constitute the training sample (0.7 by default).

- *nrep*: number of repetitions of each incremental sample, under a resampling approach with replacement (100 by default).
- *secu.ini*: start of the MINREM incremental sample size sequence; that is, the minimum sample size in test (300 observations by default).
- *p.fin*: the proportion of the number of observations of the total training sample, which is the maximum sample under test (0.15 by default).
- *delta*: increment for the sequence of incremental sample sizes (300 by default).
- *type.mod*: type of model to be tested ('ml': linear regression, 'tree': regression tree).
- *refer*: if MINREM analyzes a set of data specified by the user (refer = 0, in this case, that of houses-used in Colombia) or for the reference case (refer = 1, that is, in this context, the dataset prepared by Mullainathan & Spiess, 2017).
- *Ndepura*: cut point of incremental-sample size to conclude on the stability of the variables. That is, it is equivalent to the maximum sample size that will be tolerated for a variable to meet the criterion of being important (if it was significant in at least 95% of the cases), that is to say, if the variable has not been important up to that sample size, it will no longer be considered important (1,000 observations by default).

4. Results and discussion

4.1. Case 1: big data real estate in Colombia

4.1.1. Description of the dataset

The sample consists of 61,826 observations and 18 variables, from web advertisements selling used houses in Colombia, offered in the period January 2016 – August 2018, which were provided by the IDINNOV research group, as a result of the Statihouse® project (Pérez-Rave, 2019). The variable to be explained was the total sale price of the property, published by the seller (total.price). As explanatory variables, there were 17 variables, concerning physical aspects, amenities, and neighborhoods. Table 1 summarizes statistically the variables included in the sample.

Most of the independent variables described in Table 1 can be easily understood through their name, but this does not happen with 'strat.rec' and with 'pre.m2.mean.zone.' The variable 'strat.rec' is a recoding of the socioeconomic stratum corresponding to each property. In Colombia, the stratum is an ordinal variable of categories 1 (less favored socioeconomic conditions) to 6 (most favored socioeconomic conditions). Through it, the government geographically classifies sectors within neighborhoods and, therefore, also classifies real estate depending on its location. In Colombia, such a classification is used to set rates in public services and health, among others. In this regard, instead of six categories of the stratum, five were used, after joining strata 1 and 2 (hence the name 'strat.rec'), since they do not represent significant differences in living conditions and, in addition, the percentage of observations in the stratum '1' is minimal (1.97%).

The variable 'pre.m2.mean.zone' denotes the price of the average square meter of all the used houses, available in the sample, located in the same city to which a certain property of interest belongs. It is a proxy variable to summarize possible attributes of the city where the property is located, which may influence the price. It is worth noting that for the calculation of the 'pre.m2.mean.zone' assigned to a certain property, the data of the property were first excluded, in order to eliminate eventual endogeneity. The

Table 1. Statistical summary of variables under study for the total sample (*N*: 61,826 observations).

Quantitative	Description	Min	Max	Mean	SD	Median	Kurto	Symmet
Total.price	Offered for sale (millions of pesos)	17	2,135	433.7	360.8	319.9	3.57	1.82
Area.build.m2	Nat.log of built area (m^2)	2.89	6.90	5.06	0.59	5.05	−0.43	0.08
Bedrooms	Nat.log of number of bedrooms	0.00	3.50	1.38	0.38	1.39	1.30	0.72
Bathrooms	Nat.log of number of bathrooms	0.00	2.89	1.05	0.48	1.10	0.13	−0.47
pre.m2.mean.zone	Average price of m2 in the zone (millions of pesos)*	0.80	4.58	2.33	0.52	2.40	−0.86	0.11

Binary	Description	Freq 1.Yes	% 1.Yes
Gas	Gas service	32,131	51.97
Patio	Patio	27,948	45.20
Floor.tile.mar	Floor in tile or marble	20,631	33.37
Integral.kitch	Integral kitchen	28,804	46.59
Admon	Administration services	19,182	31.03
Garaje	Garage	24,991	40.42
School	Schools near	28,770	46.53
Garden	Garden	15,015	24.29
Commercial	Commercial area nearby	9,883	15.99
Park	Parks nearby.	27,518	44.51
Transport.rout	Transport routes	34,479	55.77
Bogo.atri	In capital city (Bogotá).	21,582	34.91

Categorical (ordinal)
Socioeconomic stratum reported for the property, according to Colombian classification (recoded)

Levels	Freq	%
1 o 2	12,618	20.4
3	21,908	35.4
4	13,608	22.0
5	8,579	13.9
6	5,113	8.3

*The average price of the m^2 of the houses located in the same city, without taking into account the property under study. Cont.: Continuous; Disc.: Discrete; Bin: Binary; Nat.log: natural logarithm; SD: standard deviation Kurto: kurtosis; Symmet: Symmetry; Freq: Frequency.

use of 'strato.rec' and 'pre.m2.mean.zone' from a 'proxy' perspective is theoretically due to beliefs that cultural, geographical and socioeconomic aspects of particular areas influence the housing prices (Banerjee & Dutta, 2017)

The variables described in Table 1 allowed forming four sets of variables, one of them the original (ori) and the others reduced through scores of linear combinations, using PCA. In Table 2 these sets of variables are described.

Table 2 shows four sets of variables that will be used to train/validate the models. The 'ori' label set consists of the 17 original variables previously shown in Table 1, while the other datasets are reduced, through the inclusion of at least one latent variable, which summarizes information from some of the original variables. For example, the set 'cp. global.size' agglomerates a variable called 'global.size,' which is a linear combination of physical variables of the property (all in natural logarithm scale): constructed area, number of bathrooms and rooms, and combinations of these last two (area built per bathroom, as well as per room). Likewise, in the case of the 'cp.amenities' group, PCA is used to summarize the characteristics of the property's amenities (gas network, patio, garage, …). The use of PCA not only contributes to overcoming possible problems of collinearity between some variables highly associated from the theory or logic of the phenomenon, but also facilitates the parsimony of the models and computational efficiency. These last are

Table 2. Description of the sets of independent variables to study.

Sets of independent variables	Description	Number of independent variables for study		
		Originals	Latents*	Total
ori	Only original variables (see Table 1).	17	0	17
cp.global.size	Global size (scores of the 1st principal component using area. build.m2, bedrooms, bathrooms, area.build.m2 per bedrooms, and area.build.m2 per bathrooms.) and original variables not used in global size.	14	1	15
cp.amenities	Amenities (scores of the 1st principal component of binary variables, except for 'bogo.atri') and original variables not used in amenities.	6	1	7
cp.glob.size.amenit	Global size, amenities, and original variables not incorporated neither in global size nor in amenities.	3	2	5

*Obtained using Principal Components Analysis and represent the scores of the first component resulting in each case

Table 3. Summary of the PCA for global size and amenities.

Global size % Variability PC_1: 52.7%		Amenities % Variability PC_1: 65.9%			
Variables*	Loads	Variables**	Loads	Variables	Loads
Area.build.m2	0.615	Gas	0.376	Garage	0.289
Bedrooms	0.281	Patio	0.332	Scholl	0.352
Bathrooms	0.387	Floor.tile.mar	0.254	Garden	0.190
Area.m2.bed	0.493	Integral.kitch	0.337	Commercial	0.118
Area.m2.bath	0.386	Admon	0.189	Park	0.340
				Transport.rout	0.402

	Statistical summary for scores of the PC_1								
Variables	Min	Max	Mean	SD	Median	Q1	Q3	Kurto	Symmet
Global.size	−5.97	5.07	0	1.62	−0.03	−1.21	1.17	−0.42	0.09
Amenities	0	3.18	1.37	1	1.47	0.19	2.24	−1.35	−0.06

*Natural log scale; **Binary (0, 1); PC_1: Principal Component 1.

primarily useful when working with big data. Table 3 describes the composition of the latent variables originated from global size and amenities, the result of the PCA. Note that PCA is used only to obtain three more test data sets (cp.global.size, cp.amenities, cp.glob. size.amenit), but it is not a requirement for MINREM nor is it part of it.

In Table 3 it can be seen that the latent variables 'global.size' and 'amenities' consisting of the scores of the first principal component in each situation, account for more than 50% of the variability in the data ('global.size': 52.7% and 'Amenities': 65.9%). The scores of the variable 'global.size' can be interpreted as a measure of the overall size of the property, where a higher score refers to the property as being more spaciousness (more area, bathrooms, ...), and vice versa. Similarly, as the score for the variable 'comfort' increases, we are facing property with greater amenities (e.g., gas network, patio).

4.1.2. Traditional model training

In this first case, for each available data set (ori, cp.global.size, cp.amenities, and cp.glob. size.amenit) a training sample is taken, and an OLS regression model is trained, using the available variables (between 5 and 17 variables, see Table 2). Said sample consists of 43,278 randomly chosen observations, which are equivalent to 70% of the total sample. To facilitate the automation of operations, we resort to the *lm()* function in R. The natural logarithm of the price *ln(total.price)* is the dependent variable, which favors normality and homoscedasticity, as well as a better interpretation of the parameters. Additionally, regression tree models are trained (using function *rpart* in R), from the same variables (dependent and independents) and observations used for the OLS regressions. The results are presented separately for each model type (regression or tree).

4.1.2.1. Regression models. Table 4 shows the regression models trained in a traditional way, from the four datasets previously described (in Table 2). Table 4 shows four regression models for *ln(total price)*, one for each dataset, from the training sample (*N*: 43,259 observations). Thus, the 'regre.ori' model represents the regression analysis on the original dataset (ori: 17 variables), 'regre.cp.global.size' on the dataset that incorporates the latent variable 'global.size,' 'regre.cp.amenities' executes the regression on the data that summarize the amenities of the property in the latent variable 'amenities,' and 'regre.cp.glob.size.amenit' represents the use of both latent variables ('global.size' and 'amenities'). In each case, these variables are used in conjunction with the other originals: 'strat.rec,' 'pre.m2.mean.zone,' and 'bogo.atri.'

Table 4 shows the high R^2 obtained for the four trained models, oscillating between 80% ('regre.cp.glob.size.amenit') and 82.2% ('regre.ori'). These results show its usefulness in summarizing the 17 original variables, employing linear combinations of some of them, which can favor not only parsimony, but also avoid collinearity problems for an inferential phase. In turn, it is highlighted in Table 4 that practically all the independent variables tested (except 2/17 in the 'regre.ori' model), were statistically significant (*p*-values less than 0.05). However, this result should be treated with caution, since the large sample size can highly influence this significance. In other words, the high number of observations tends to increase the power of the tests and, therefore, the results can be seen as statistically significant, although they may not be important from a practical point of view. We will investigate this in a later section.

Table 4. Regression models traditionally trained (OLS with the available variables in the sample, using *lm* in R) for ln*(total.price)*.

Variables/Models	regre.ori	regre.cp.global.size	regre.cp.amenities	regre.cp.glob.size.amenit
area.build.m2	0.569*** (0.004)		0.571*** (0.004)	
bedrooms	−0.035*** (0.006)		−0.051*** (0.006)	
bathrooms	0.281*** (0.005)		0.294*** (0.005)	
gas	−0.020*** (0.004)	−0.015*** (0.004)		
patio	**−0.005 (0.004)**	−0.011*** (0.004)		
floor.tile.mar	−0.025*** (0.004)	−0.024*** (0.004)		
integral.kitch	0.063*** (0.004)	0.072*** (0.004)		
admon	**0.0002 (0.004)**	0.018*** (0.004)		
garage	0.030*** (0.004)	0.036*** (0.004)		
school	−0.024*** (0.005)	−0.017*** (0.005)		
garden	0.044*** (0.004)	0.051*** (0.005)		
commercial	0.052*** (0.005)	0.055*** (0.005)		
park	−0.027*** (0.005)	−0.029*** (0.005)		
transport.rout	−0.010** (0.005)	−0.014*** (0.005)		
bogo.atri	0.180*** (0.004)	0.196*** (0.004)	0.175*** (0.004)	0.187*** (0.004)
strat.rec3	0.336*** (0.004)	0.362*** (0.005)	0.349*** (0.004)	0.380*** (0.005)
strat.rec4	0.605*** (0.006)	0.659*** (0.006)	0.632*** (0.005)	0.699*** (0.005)
strat.rec5	0.749*** (0.007)	0.810*** (0.007)	0.779*** (0.006)	0.861*** (0.006)
strat.rec6	1.017*** (0.008)	1.079*** (0.008)	1.055*** (0.008)	1.143*** (0.008)
pre.m2.mean.zone	0.325*** (0.004)	0.335*** (0.004)	0.334*** (0.004)	0.351*** (0.004)
global.size		0.248*** (0.001)		0.247*** (0.001)
amenities			0.001 (0.002)	0.010*** (0.002)
Constant	1.381*** (0.018)	4.429*** (0.009)	1.352*** (0.017)	4.380*** (0.009)
Observations	43,278	43,278	43,278	43,278
R^2	82.2%	80.4%	81.9%	80.0%
Adjusted R^2	82.2%	80.4%	81.9%	80.0%
Residual Std. Error	0.324 (df = 43,257)	0.340 (df = 43,259)	0.327 (df = 43,267)	0.343 (df = 43,269)
F Statistic	9,996.819*** (df = 20; 43,257)	9,874.327*** (df = 18; 43,259)	19,588.250*** (df = 10; 43,267)	21,671.250*** (df = 8; 43,269)

*p < 0.1; **p < 0.05; ***p < 0.01; standard error in parenthesis next to the regression coefficients

4.1.2.2. Tree models. Figures 2–3 provide the regression trees resulting from traditional training, prior to the use of the *rpart* function in R. Figure 2 provides the regression tree obtained with the set of original variables (ori: 17 variables). This tree coincides with the one obtained when using the dataset that includes the latent variable 'amenities,' which summarizes the variables of amenities. Note the remarkable parsimony of this tree, in comparison with the previously obtained regression models (Table 4), since the tree only uses four independent variables: constructed area (in natural logarithm scale), number of baths (in natural logarithm), recoded stratum, and price of one square meter in the city. It will be necessary to see if such a parsimony does not sacrifice prediction quality, which we will evaluate in a later section.

Figure 3 provides the regression tree trained with the dataset that includes the variable 'global.size,' which summarizes the amplitude characteristics of the property. In turn, this tree coincides with the one obtained for the dataset that incorporates the two latent variables (global size and comfort). The present regression tree includes the following variables: recoded stratum, global size, and average price of one square meter of the city (zone).

In the trees presented in Figures 2 and 3, it is worth noting that none of the amenity variables was chosen, neither in their original form (e.g., gas network, transport) nor in summarized form ('amenities'). Besides, the root of each tree is a function of an amplitude variable, either as an original variable (constructed area: Figure 2) or as a latent variable (global size: Figure 3).

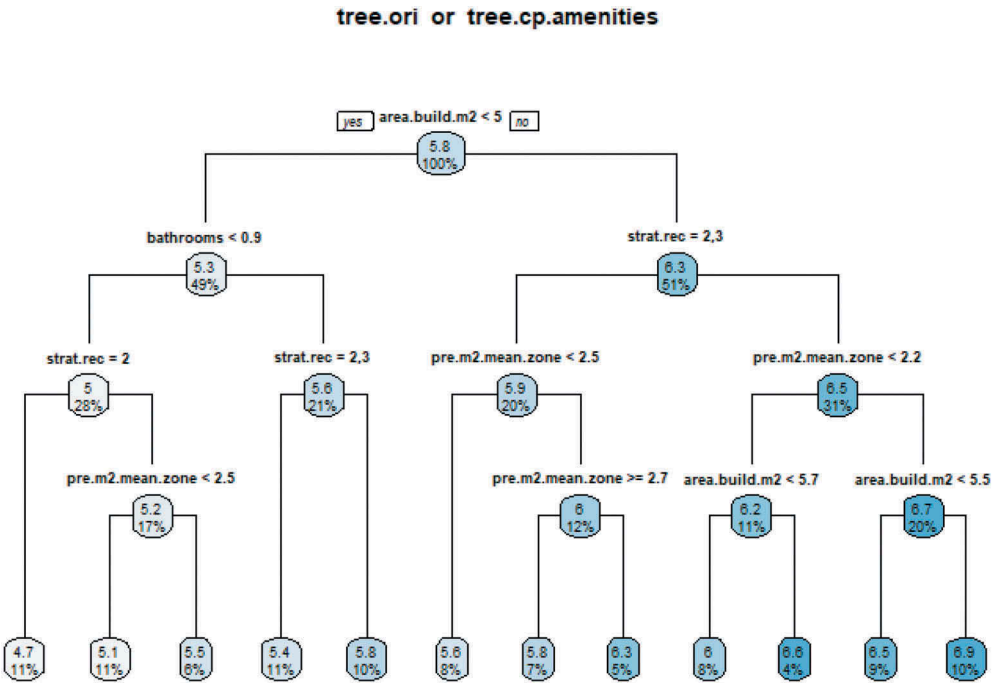

Figure 2. Regression tree for ln(total.price) trained by only the original variables (ori), as well as being trained using the variables of amenities and the remaining originals (cp.amenities). Note: *strat.rec = 2* corresponds to levels 1 and 2 of the stratum.

From a general view to the traditional procedure of model training, there is an outstanding controversy about which are the most important attributes for explaining the *ln(total price)*, since in the regression models practically all the variables (original or latent) were significant; however, in the tree-based models, only a few variables were included. Possibly this parsimony occurs since regression tree is a nonparametric approach that uses information theory and is not affected by the sample size in terms of the significance of the variables, as it does in classical regression (high test power). However, it can be affected by multicollinearity or redundancy, which could yield useful variables for prediction but erroneous for inference. Also, it will be necessary to analyze if the parsimony observed in the trained trees does not sacrifice prediction quality. Additionally, tree methods do not present the benefits of classical regression in terms of the interpretative possibilities of the study phenomenon.

4.1.3. Model training under incremental sample with resampling (MINREM)

The models employing the MINREM strategy used the default conditions in the function already described and programmed in *R*:

These function parameters correspond to the following initial conditions:

- 70% of the observations for the training sample (43,278 obs) and the remaining 30% for the validation (prop.entrena).

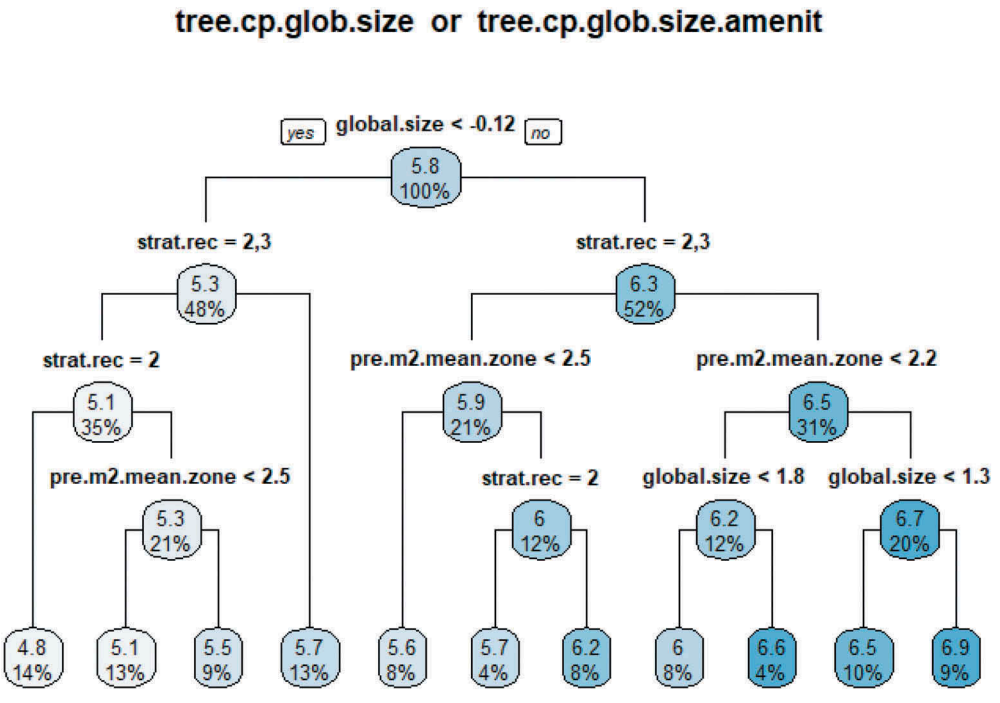

Figure 3. Regression trees for *ln(total price)*, trained from variables global size and remaining originals (cp.global.size), as well as from variables global size, amenities and remaining originals (cp.glob.size.amenit).

Note: *strat.rec = 2* corresponds to levels 1 or 2 of the stratum.

- 100 repetitions with replacement (nrep) for each incremental sample size to be tested.
- A sequence of samples with an increment of 300 (delta), ranging from 300 to 6,300 observations (43,278 for training × 0.15, the latter is p.fin).
- It will be executed for the case indicated by the user (refer = 0) and not for the reference case (refer = 1, which will be the second one to be dealt with in this article).
- Apply to regression models (type.mod = 'ml') and extract the important variables before a maximum sample size of 1,000 (Ndepura).

4.1.3.1. Importance and stability of the independent variables. Figure 4 provides three types of visualizations of the evolution of the importance of the variables in the case of regression with original variables (regre.ori), according to sample size. On this occasion, the sample size ranged between 300 and 6,300 observations, which is a sufficient range to understand the evolution.

Figure 4 shows what is needed for an independent variable to qualify as important or not under a regression approach driven by big data. Note that as the sample size increases, certain variables tend to be important (at least 95% of the repetitions) and other variables not. In this figure, three subsets of variables stand out. One of these agglomerates variables that remained important throughout the observation range ('area.build.m2,' 'bathrooms,' 'bogo.atri,' 'strat.rec,' and 'pre.m2.mean. zone'); another

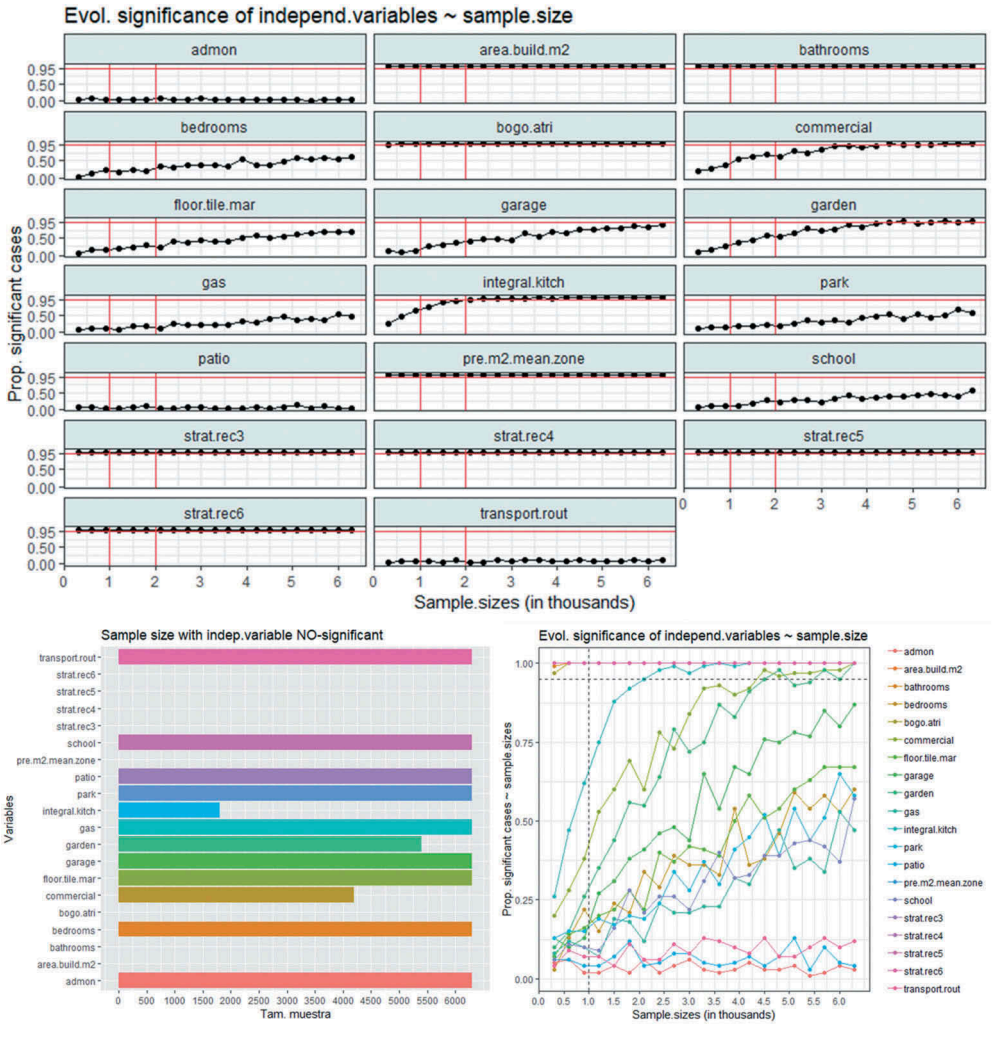

Figure 4. Plots for the importance (a variable is important if it was significant in at least 95% of the cases, otherwise, it will be considered irrelevant) of the independent variables according to sample size, using the original dataset (ori).

includes variables with a growing tendency to be important from a certain sample size included in the observation region ('integral kitchen': approx. 1,800 obs.; 'commercial': approx. 4,200 obs.; 'garden': approx. 5,300 obs.); and another where until the end of the observation range it remained irrelevant (eg. 'transport.rout', 'gas'). Under the traditional training of the regressions, using the 43,278 observations, practically all the variables tested were statistically significant (Table 4). However, under the MINREM strategy, only five of the variables were stable regarding their importance, considering incremental sample sizes from 300 to 6,000 observations.

Thus, the regression model obtained under MINREM (for the variable selection, with Ndepura = 1,000), which initiated the training with the 17 original variables, finally incorporates only 5 of them ('area.build.m2,' 'bathrooms,' 'bogo.atri,' 'strat.rec,' and 'pre.m2.mean.zone'), and will be referred to as 'regre.ori.minrem'.

Table 5. Regression models trained for ln(total.price), with previous use of MINREM* for ln(total.price), considering the training sample (43,278 observations).

Variables	regre.ori.minrem ó regre.cp.amenities.minrem	regre.cp.global.size.minrem ó regre.cp.glob.size.amenit.minrem
area.build.m2**	0.558*** (0.004)	
bathrooms**	0.276*** (0.004)	
bogo.atri	0.168*** (0.004)	0.188*** (0.004)
strat.rec3	0.352*** (0.004)	0.382*** (0.005)
strat.rec4	0.644*** (0.005)	0.705*** (0.005)
strat.rec5	0.798*** (0.006)	0.867*** (0.006)
strat.rec6	1.079*** (0.007)	1.151*** (0.008)
pre.m2.mean.zone	0.339*** (0.004)	0.351*** (0.004)
global.size		0.247*** (0.001)
Constant	1.351*** (0.017)	4.387*** (0.009)
Observations	43,278	43,278
R^2	81.9%	80.0%
Adjusted R^2	81.9%	80.0%
Residual Std. Error	0.327 (df = 43,269)	0.343 (df = 43,270)
F Statistic	24,429.530*** (df = 8; 43,269)	24,745.110*** (df = 7; 43,270)

*Variables that were significant in 95% of cases using Ndepura of 1,000 observations. ** Variables in natural logarithm scale; *** $p < 0.01$; standard error in parentheses, next to the regression coefficients.

4.1.3.2. Regression models. Table 5 shows the regression models obtained under MINREM (Ndepura = 1,000), using the four datasets of traditional training (described in Table 2).

Here, two models (Table 5) are obtained instead of four, since those models trained from the set of original variables ('regre.ori.minrem') and from the set that summarizes the amenities ('regre.cp.amenities.minrem') turned out to coincide. Similarly, the induced model of the dataset that incorporated the 'global.size variable' ('regre.cp.global.size.minrem') coincided with the model that included that latent variable and 'amenities' (regre.cp.glob.size.amenit.minrem).

In itself, the first model (left side of Table 5: 'regre.ori.minrem or regre.cp.amenities.minrem') is a function of only the original variables; while the second (right side: 'regre.cp.global.size.minrem or regre.cp.glob.size.amenit.minrem') excludes the area and the number of bathrooms and adds the latent metric 'global.size'. Additionally, note that none of the amenity variables were statistically significant (neither in their original nor latent form). The two resulting models are highly parsimonious and have an R^2 of about 80%, which is very similar to that achieved with almost all the variables under traditional training (R^2 between 80% and 82.2%, see Table 4). These results reinforce the need to be cautious when judging a certain variable as important or not, under a data-driven approach. Nevertheless, the prediction capacity of the two obtained models has yet to be evaluated, using the non-training sample, which we will analyze in a later section.

Note that MINREM does not use standard procedures for selecting variables in the regression, such as forward, backward, stepwise (combination of the first two) or best subset (all possible combinations). In fact, resorting to these procedures would be impractical when dealing with big data, for three main reasons: 1) if, within the classic procedures, some parametric criterion of significance is used (e.g., p-value), the high power of the test usually leads to unrealistic and erroneous models, since any derisory difference would be considered statistically significant. 2) High computational inefficiency, not only because of the high number of observations but of variables. 3) Solution instability, as a result of the

high multicollinearity and redundancy that characterize big data. In fact, George et al. (2016, p. 1501) specify: '… the number of combinations of the explanatory variables explodes with the number of variables, rendering this approach unfeasible in practice'. Frické (2015, p. 657) states: 'exhaustive analysis of these is not a practical possibility here' [in big data]. Also, again, George et al. (2016, p. 1501), relying on Hastie, Tibshirani, & Friedman (2009), express: 'Inclusion of all variables in a model, such as a regression model, is typically impeded by high multicollinearity'. On the other hand, MINREM, thanks to its focus on the selection of the most stable-important variables along different incremental sample sizes, also helps to address common problems in machine learning approaches, such as the instability warned by Mullainathan and Spiess (2017). Additionally, one of the standard alternatives in machine learning is the regularization of the models (penalty function). However, even such an alternative as LASSO regression, continues to be affected by multicollinearity and redundancy (Hastie et al., 2009, p. 614) and, therefore, although it tends to generate fewer complexes (and is useful for predictive purposes), these are often wrong (Mullainathan & Spiess, 2017, p. 98).

Instead, MINREM: 1) choose a training sample. 2) Using resampling for various incremental sample sizes, train the model that contains all available variables. 3) In turn, for each sample size, calculate the proportion of cases (repetitions) in which each variable (parameter) was significant (in regression: p-values<0.05; in other such methods as regression trees, consider the inclusion of the variable in the resultant model). 4) In each sample size, MINREM considers that a variable is important if it was significant at least 95% of the repetitions, otherwise, it will be considered irrelevant. 5) According to an established Ndepura (see Section 3), the stable-important variables are chosen (next used to train the debugged model).

Note, moreover, that the execution of MINREM is flexible since the user can modify the default parameters (e.g, % of observations for training, number of repetitions in each sample size, the increment for the sequence of incremental sample sizes; see Section 3). Likewise, MINREM does not need constant human intervention, as it is done automatically, through the 'min_rem.R' function.

4.1.3.3. Tree models. Figure 5 shows the tree model resulting from using the set of original variables (ori), which was the same obtained from the set that summarized the amenities ('comfort').

The trained tree with previous use of MINREM (Figure 5) includes one fewer less variables (excludes the number of baths) than the traditionally trained tree. Additionally, it is worth mentioning that the tree supported by MINREM using the dataset that incorporates 'global.size' is the same as the tree trained from the data set that includes that variable and 'amenities,', and is also equivalent to the tree induced traditionally, previously presented in Figure 3, which included 'global.size,' 'strat.rec,' and 'pre.m2.mean.zone.'

4.1.4. Validation of the model

Table 6 compares the fit of the models obtained under the traditional training and MINREM strategies, using both the training sample and the validation sample, the latter consisting of 18,548 properties (30% of the total sample).

In order to illustrate the interpretation of Table 6, see the rows of the trained regression model with the original variables ('regre.ori'). The row of the traditional

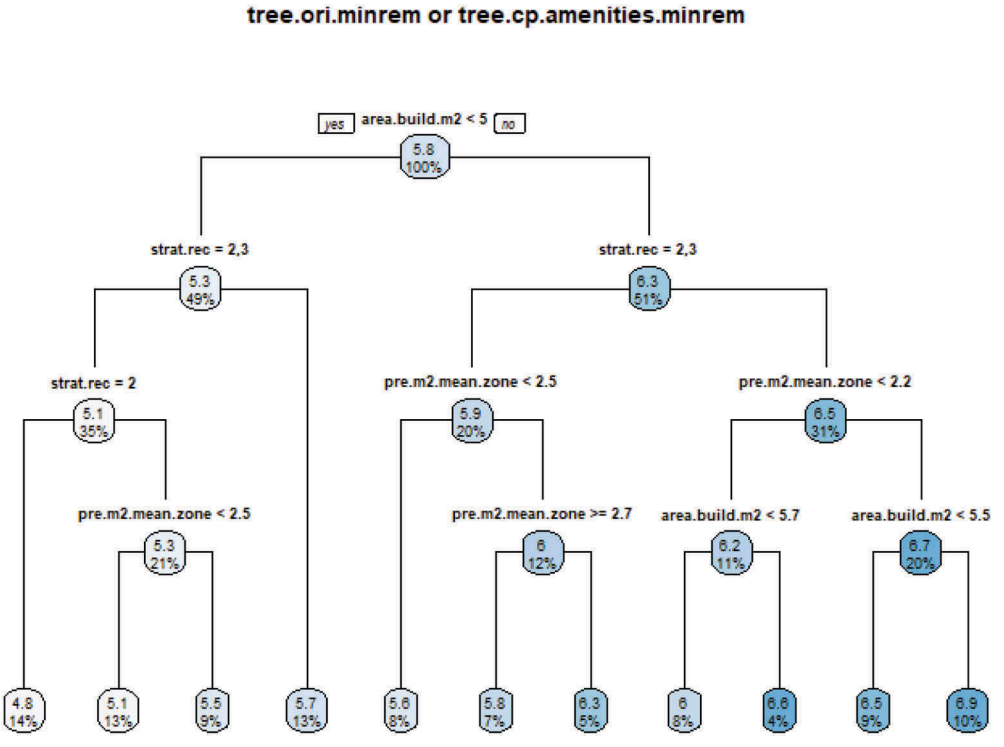

Figure 5. Regression trees trained for ln(total.price) incorporating MINREM (Ndepura = 1,000) previously on the training sample of 43,278 observations.

Table 6. Fit of models in training (43,278 observations) and validation samples (18,548 observations) according to traditional training strategies and MINREM.

		Independent variables				R^{2}*			
Models	Stra	Pro	Sig	%Sig	Change	Train	Change	Valid	Change
regre.ori	T	17	15	88.2%		82.2%		82.4%	
	M	17	5	29.4%	**−66.7%**	81.9%	**−0.36%**	82.1%	**−0.36%**
regre.cp.global.size	T	15	15	100.0%		80.4%		80.6%	
	M	**15**	**4**	**26.7%**	**−73.3%**	**80.0%**	**−0.50%**	**80.2%**	**−0.50%**
regre.cp.amenities	T	7	6	85.7%		81.9%		82.1%	
	M	7	5	71.4%	**−16.7%**	81.9%	**0.00%**	82.1%	**0.00%**
regre.cp.glob.size.amenit	T	5	5	100.0%		80.0%		80.2%	
	M	5	4	80.0%	**−20.0%**	80.0%	**0.00%**	80.2%	**0.00%**
tree.ori	T	17	4	23.5%		71.1%		71.3%	
	M	17	3	17.6%	**−25.0%**	70.4%	**−0.98%**	70.2%	**−1.54%**
tree.cp.amenities	T	7	4	57.1%		71.1%		71.3%	
	M	7	3	42.9%	**−25.0%**	70.4%	**−0.98%**	70.2%	**−1.54%**
tree.cp.global.size	T	15	3	20.0%		70.3%		70.1%	
	M	15	3	20.0%	0.0%	70.3%	**0.00%**	70.1%	**0.00%**
tree.cp.glob.size.amenit	T	5	3	60.0%		70.3%		70.1%	
	M	5	3	60.0%	0.0%	70.3%	**0.00%**	70.1%	**0.00%**

Stra: Strategy (T: Traditional, M: MINREM). Pro: probed variables; % Sig: percentage of significant variables concerning those tested. Change: % change (M-T)/T. * Square of the correlation between ln(total.price) real vs predicted.

strategy (T) shows that, of the 17 variables initially tested, 15 were significant (p-values less than 0.05), which is equivalent to 88.2% of the variables under test (% Sig). In turn,

this model showed an R^2 of about 82% in both the training sample and the validation sample. In contrast, the model trained under the MINREM (M) strategy only included 29.4% of the variables initially tested, and the R^2, both in training (81.9%) and validation (82.1%) samples, was practically equivalent to the traditional strategy. In itself, with the MINREM strategy using the original baseline data (17 variables), MINREM led to 66.7% fewer independent variables in contrast to traditional training and distanced itself from the latter by only – 0.36% in R^2. A similar behavior occurs in the other regression models. For example, in 'regre.cp.global.size' (using the data that includes the variable global size), with MINREM, there were 73.3% fewer variables than in traditional training (from 15 variables to only 4), and the difference in the fit in the training and validation samples is negligible (−0.5%). As was inferred from previous sections, the impact of the MINREM is more noticeable in the hedonic regression than in the tree, although in the latter there is also a model with one less variable ('tree.ori or tree. cp.amenities') and there is only a minimal impact on the reduction of R^2 (−0.98% in training and −1.54% in validation).

On the other hand, when comparing the models trained only with original variables and with at least one latent variable ('global.size' or 'amenities'), the reduction of original variables by a linear combination of some of them favors the parsimony of the models and did not significantly affect the resulting fit, neither with the traditional strategy nor under MINREM. For example, in 'regre.ori' traditionally trained: original (R^2: 82.2%) and with PCA (R^2: 80.4%); in the validation sample: original (R^2: 82.4%), and with PCA (R^2: 80.6%). Another aspect to be highlighted is the superiority of the regression approach to that of the regression tree, since the former exceeded the latter in terms of R^2, by approximately ten percentage points.

Considering the parsimony of the models under test and their values of R^2, both in training and validation samples, the model trained with support under MINREM was chosen, which includes the variable 'global.size' to summarize some of the original variables of amplitude. In this model, of the 17 original variables, only 4 variables were chosen, and, even so, the fit was almost the same as using all the variables, both in the training sample (43,278 observations) and in the validation sample (18,548 observations), with R^2 close to 80%.

4.1.5. Using the chosen model for inference and prediction

A supervised learning model should support two practical needs. One of them is the inference about the phenomenon of study, so that it allows a better knowledge about it, regarding, at least, what are the effects of the important independent variables on the dependent one. The other is the prediction of future events. In the context of machine learning and big data, the use of supervised models has been primarily prediction. In some cases, for the interest in obtaining an answer to solve practical problems, without demands to nourish the understanding of the phenomenon under observation and the data generation process. In other cases, even if one is interested in inference, there are barriers to the use of big data, which turn out to be observational data that are mostly approached from the perspective driven by the data and not by the theory or logic of the phenomenon. Likewise, collinearity tends to be present, affecting the attitudes about the significance of the parameters, as also happens with the large sample sizes, among

other aspects. These situations lead to unstable models or models that contain redundant or irrelevant variables.

Here, the model chosen will be tested for inferential purposes, which is facilitated by the fact that this model is quite parsimonious (which facilitates its interpretation) and includes variables whose importance is stable under different sample sizes under MINREM. Also, from the logic of the phenomenon (used houses) and the theory, there are elements which make reasonable the inclusion of these variables to explain the *ln(total price)*. For example, aspects of the size of the home (areas, number of bathrooms, …), which have been considered in this study through the main component 'global size,' have been reported as determinants of the price of a property in studies such as Dubin (1998), Limsombunchai (2004), Pardoe (2008), and Lowrance (2015). On the other hand, Bonetti et al. (2016) (Milan, Italy, the effects of water sources, 10,530 observations from a web scrape), Dorr (2016) (New York, the effects of available community services 122 observations) and Seo et al. (2017) (California, pavement condition, 19,608 observations), among others, have provided evidence for the importance of socioeconomic and environmental conditions of areas near the property, and the relations of these with the price of housing. Although the present paper, being limited to observational-web data, does not have the specific variables studied by these authors, it has represented general socioeconomic and environmental conditions, alluding to micro and macro neighborhoods of the properties (with different delimitations), through the proxy variables 'stra.rec' (sub-neighborhood), 'pre.m2.mean.zone' (local neighborhoods), and 'bogo.atri' (global neighborhood, capital city or not).

Considering then the stability of the variables of the chosen model and the theoretical elements, we proceed with the interpretation of the parameters and the provision of new findings on the response variable, in the context of used houses for sale in Colombia through the Internet.

4.1.5.1. Use of the model for inferential purposes. Table 7 provides estimates of the coefficients of the chosen model ('regre.cp.global.size') along with their 95% confidence intervals. These estimates are presented for both the training sample, 43,278 observations, and for a subsample of this, with only 1,000 observations. It is worth remembering that even with 1,000 observations, all the variables present in the chosen model were already important and thus they remained during the test, as evidenced by MINREM (see Figure 4).

Table 7. Coefficients and 95% confidence intervals for the model Y = ln(total.price) chosen (regre.cp.global.size), using the training sample and a subsample.

	Coef		CI.95%		% Change on Y: $100 \times (e^{\beta} - 1)$	
Variables	N: 43,278	N: 1000	N: 43,278	N:1000	N: 43,278	N:1000
bogo.atri	0.19	0.19	(0.18; 0.20)	(0.13; 0.24)	21	21
strat.rec3	0.38	0.41	(0.37; 0.39)	(0.35; 0.47)	46	51
strat.rec4	0.70	0.73	(0.69; 0.72)	(0.66; 0.80)	101	108
strat.rec5	0.87	0.91	(0.85; 0.88)	(0.83; 0.98)	139	148
strat.rec6	1.15	1.16	(1.14; 1.17)	(1.06; 1.26)	216	219
pre.m2.mean.zone	0.35	0.37	(0.34; 0.36)	(0.31; 0.42)	42	45
global.size	0.25	0.26	(0.24; 0.25)	(0.25; 0.28)	28	30

*Coef: regression coefficients; CI: confidence intervals at 95 %

Note, in Table 7, that the estimates of the model parameters are stable, independent of the sample sizes tested. For example, in the case of 'bogo.atri' there was no variation in the parameter estimates (0.19). In 'global.size' and in 'stra.rec' the estimates only varied by one-tenth. The level with more variation was 'strata.rec5,' which ranged from 0.87 (*N*: 43,278) to 0.91 (*N*: 1,000), but remains minimal. Also, in all cases, the confidence intervals overlap considerably, with it being expected that with 1,000 records these intervals are more extensive than with 43,278 observations.

Table 7 also provides, in the last column, '% Change on Y'; this refers to a more precise interpretation of the marginal effect of each independent variable on the dependent variable, keeping the others constant. In this regard, it is worth presenting the following descriptions of the average sale price of used homes, offered by web ads in Colombia, considering the estimates of the training sample (43,278 observations):

- The variable 'bogo.atri' is a proxy, at an aggregate level, which seeks to summarize factors of global neighborhoods (capital or not of the country), which are seen neither locally (specific city) nor at the micro level (property or nearby sub-neighborhoods). In fact, Clavijo, Janna, and Muñoz (2005) use aggregate data from Bogotá (new housing price index) as a proxy to infer situations at the national level. Now, the fact that the property is located in the capital city (Bogotá, in this case), represents an increase of about 21% in the average price published on the Internet for the property. In this regard, it is worth mentioning that, in Colombia, when considering the last five years of reports on 'The price index of used housing (IPVU),' available on 11/17/2018 (2013–2017; city, http://www.banrep.gov.co), in all these years 'Bogotá' (capital city) presented the highest (real) values, in comparison with the rest of the cities. Additionally, when considering other contexts, not necessarily comparable, such as China, there are also exploratory inputs that reinforce the importance of the variable 'capital city': '… clearly, land is becoming more expensive relative to structure in the capital city' (Wu, Gyourko & Deng, 2012, p. 537).
- The higher the socioeconomic stratum registered for the property, the higher its published price. For example, compared to stratum 1–2, if the property belongs to stratum 3, its price tends to increase by 46%; but if it is from stratum 4, the expected increase is 101%, or 139% for stratum 5, and 216% for stratum 6. In other words, the change between strata seems exponential. For example, when moving from stratum 1–2–3 (46%), from stratum 3 to 4 (55%), from stratum 4 to 5 (84%) and from stratum 5–6 (132%). These results are a reflection of the gap between 'rich' and 'poor' in countries like Colombia. In addition, the variable 'stra.rec' seeks to summarize socio-economic conditions of sub-neighborhoods (neighborhood, blocks, streets, …), but also some amenities of the house, as a result of greater purchasing power, better public services, among others. In other words, the stratum is also a latent variable – proxy. For example, Florez and Árias (2010, p. 335), alluding to Colombia, states that 'the socioeconomic stratum is a measure of the resources and facilities of a place.' In the Argentine scenario, Tuñón & Halperin (2010, p. 8) point out: '[…] attribute of the home that can be extended to all its members […] considers the main assets of the household; […]; how is access to goods and services; and those that refer to the economic head of the household,

such as the highest level of education attained and, the occupational situation.' In Paraguay, Castillo (2015, p. 34) states: 'variable constructed from information on access to basic services in housing, such as water, electricity, sanitary service, possession of household appliances, means of transportation and average of people housing and bedroom.' In fact, in the site statihouse.com (on the statistical exploration of houses in Colombia), in the section 'Segment,' it can be seen that when the stratum increases, the proportion of houses with amenities such as a gas network also tends to increase, same with kitchen, payment of administration, garage, garden and patio (Pérez-Rave, 2019). It is worth remembering that, in the present study, 11 amenities were summarized in the latent variable 'amenities' (first main component, 65.9% of the variability, Table 3), which competed together with 'stra.rec' and other variables under the MINREM strategy, and in no case 'amenities' was important (neither in multiple regression nor in regression trees). On the other hand, 'strata.rec' was important in both types of models, which throws elements in favor of the integrality and relevance of the stratum. Additionally, the socioeconomic stratum is an important predictor of variables such as: consumption (Herrán-Falla, Prada-Gómez, & Patiño-Benavidez, 2003, shows in Colombia); academic performance (Tuñon & Poy, 2016; sample in Argentina) and infection risk (Fonseca et al., 2005, sample in Colombia); to name a few.

- Considering the variable 'pre.m2.mean.zone,' also of a proxy nature, it stands out that for each unit increase in the average price of one square meter of the city in which the used house is located (without including the price of the property), an increase of 42% in the average price is expected (published online). The variable in question ('pre.m2.mean.zone') represents an aggregate level of data, so it tries to summarize attributes at the local level (specific city) that influence the offer price; these attributes are not specific to the micro level (property and its sub-neighborhoods) or the global level (e.g. capital of the country, department, country). Among these attributes can be: housing demand, construction costs (Clavijo et al., 2005); population density (Figueroa, 1992), as well as other factors; but all of them added to the level of the city in which the housing is located.
- 'Global.size' is a latent metric, with zero mean, created from a linear combination of physical variables of size, using PCA (built area, number of bathrooms and bedrooms, built area/number of bathrooms, and constructed area/No. of rooms; all them in Natural log scale). To the extent that the score of 'global.size' of a property increase by one unit, an increase of 28% in the average price (published on the Internet) is expected. This result is consistent with findings from studies such as Dubin (1998), Limsombunchai (2004), Pardoe (2008), and Lowrance (2015), which have reported the positive influence of size variables, such as area and number of rooms on housing prices.

In order to supplement the above with an additional empirical exploration, it is worth observing whether the estimated effects for the regression model are reflected not only in the training sample but also in the validation sample. In this sense, for each of the important attributes, Figure 6 presents graphs that relate the median of the total price of the property (original scale in millions of Colombian pesos) with the type of samples

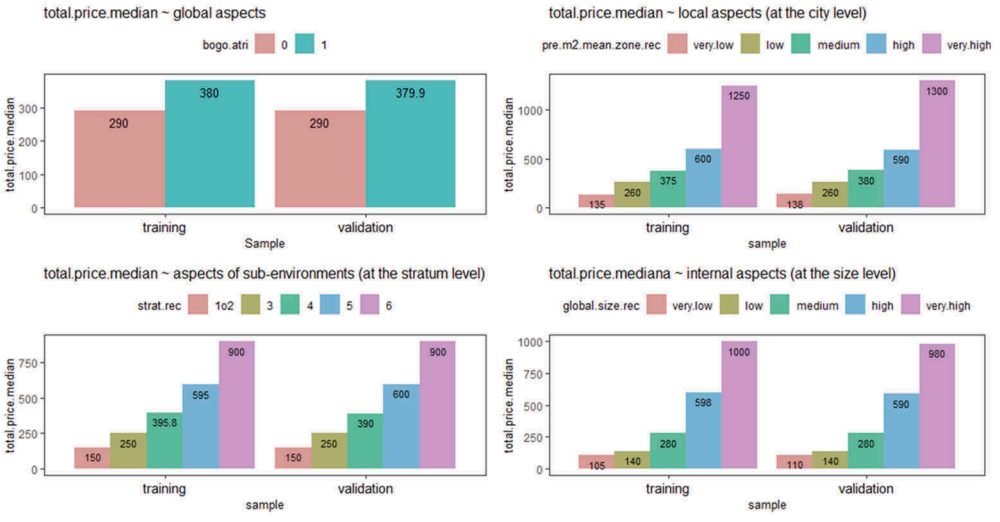

Figure 6. Median of the total price according to sample type for each attribute of the chosen model.

used (training, validation). For ease of interpretation, the numerical attributes were recoded into five levels, ranging from very low to very high.

Figure 6 shows that the conclusions about the global aspects (belonging to the capital city or not), local (at the city level, average m^2 price of the properties for sale in the city, excluding the property of interest), sub-environments (represented in the socioeconomic stratum) and internal (global size of the property), are consistent, when considering the sample of training and validation. For example, the price of the property (median) is shown to be higher in the capital city than in the others. Also, considering the socioeconomic stratum and local aspects (pre.m2.mean.zone.rec), it is seen that at a higher level, the price (median) tends to increase. The same happens if the size of the property is observed (global size). The results shown in Figure 6 reinforce statements of Mok, Chan & Cho (1995, p. 46), that 'the valuation of a property is sensitive to changes in the locational, structural, and neighborhood traits.'

4.1.5.2. Testing the assumptions of the regression model. Next, it will be verified that there are no extreme violations of the traditional assumptions of the regression analysis for the chosen model ('regre.cp.global.size'). However, evidence has already been provided in favor of the consistency of the estimates of the parameters of the model, considering samples of 1,000 and 43,278 observations, and also for the importance of the included variables. However, it is worth exploring whether there are any extreme deviations from normality and homoscedasticity, as well as any problems of multicollinearity. In this sense, Figure 7 shows two groups of traditional graphs, that of the upper part for the totality of the training sample, and that of the lower part for the subset of 1,000 observations.

The graphs shown in Figure 7 do not reflect patterns in the residuals. On the other hand, the following stand out: symmetry of the residuals, good fit between the real and predicted values, and randomness when plotting residuals vs. different levels of the variables. It is worth noting that, for strata 1–3, a higher dispersion is observed in the residuals than for the other strata, when considering the total training sample. However,

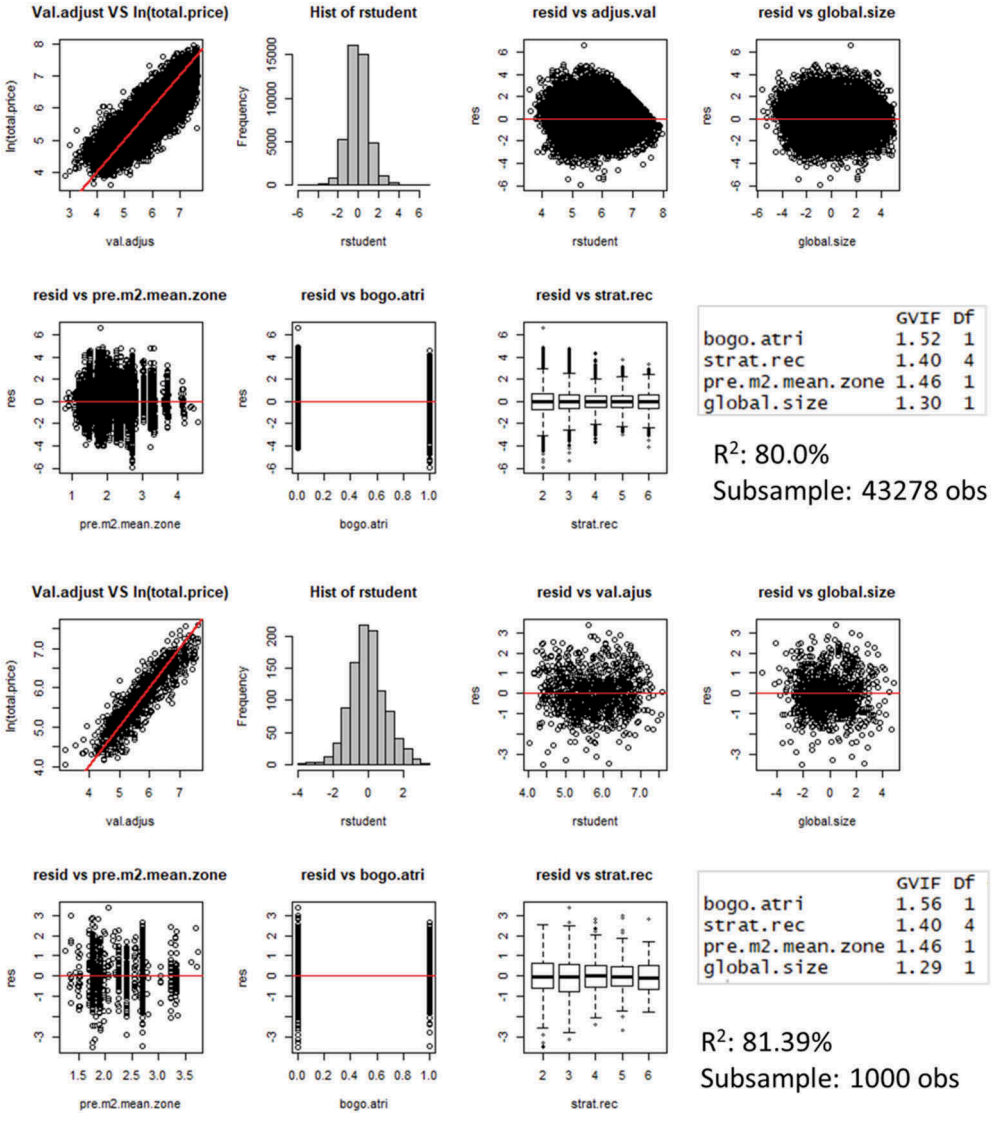

Figure 7. Graphs of real values vs predicted, and of residuals for two groups of samples: training (43,278 observations, top) and subsample (1,000 observations). VIF is also included.

when viewing the subsample (1,000 observations), the pattern is not so clear. Furthermore, if there were problems there, their influence on the estimates of the parameters would not be critical, due to the amount of data, the method of choosing the variables (MINREM), and the stability of the estimates.

Another aspect to emphasize is that there are no of collinearity problems. All VIFs were less than 5, and no square root of these was greater than 2.

4.1.5.3. Use of the model for predictive purposes. In this section, four metrics for the quality of predictions made by the chosen model are used ('regre.cp.global.size'), but,

this time, not using the scale *ln(total price)*, but the original scale, in millions of pesos, prior to the conversion of the predictions with the function 'exp.' Such a conversion is relevant in this context of use since the user will not use the predictions in logarithmic terms but in monetary terms. Despite this, it is common to find that the models under machine learning report the quality of prediction in transformed scales and not in the scale that will be used by the user to make the decisions.

Additionally, given the fact that the chosen model is quite parsimonious (only four variables) and looking to use it in this section only for predictive purposes (not inferential), it will also be tested with the use of double interactions. This alternative is practically nullified for big data prediction under machine learning, since, in addition to a large number of observations, there may also be a large number of independent variables (as will be seen in the following case study), so, according to Mullainathan and Spiess (2017, p. 90), the use of interactions would be infeasible.

Table 8 shows the comparative results of the prediction quality of the chosen model ('regre.cp.global.size').

Regarding the chosen model, it is worth mentioning that, on average, the predictions deviated by 27% from the real published price to offer the property online when using the model without interactions, and by 20.9% when considering double interactions. It is worth noting that using double interactions, in 63% of the cases the prediction error did not exceed 20% of the real price; without interactions, this percentage was 48.2%. It is important to publish this type of results in the original scale when working with transformed variables, for better transparency for the user. For example, in the logarithmic scale, a MAPE of 4.7% was obtained for the model without interactions and 3.6% with double interactions. In addition, 99.3% of the predictions by the model without interactions deviated by no more than 20% of the real value, and 99.5% with double interactions.

4.2. Case 2: big data real estate prepared in Mullainathan and Spiess (2017)

4.2.1. Description of the dataset

In addition to our case, it is also interesting to explore the performance of MINREM on another dataset, which has a larger number of variables. For this, Mullainathan and Spiess (2017) has been taken as a reference. It uses the logarithm of the value of the property as a dependent variable. Its data came from the 2011 American Housing Survey, which included 150 covariates. However, after cleaning procedures and data preparation, executed by the authors mentioned above, we find that they incorporated additional variables. Therefore, the starting dataset for the training of several

Table 8. MAPE and percentage of predictions that did not exceed a specific percentage error in the validation sample, using the model 'regre.cp.global.size'.

	In original scale (Millions of Colombian pesos)**			
Versions of regre.cp.global.size	MAPE*	Err.max10%	Err.max15%	Err.max20%
Without interactions	27.0%	25.7%	37.2%	48.2%
With doubles interactions	20.9%	41.2%	53.8%	63.0%

Err.max: Percentage of predictions that did not exceed a specific error%.
*MAPE: *Mean Absolute Percent Error*
**Prior conversion of the predicted values using: $e^{(prediction)}$

prediction models, including one of OLS regression, consisted of 162 regressors and 10,000 observations. In turn, Mullainathan and Spiess (2017) performed a validation of the models on the 48,888 remaining observations, using an eight-fold strategy and averaging the results. The data, code, article and supplementary material of Mullainathan and Spiess (2017) is available at https://www.aeaweb.org/articles?id=10.1257/jep.31.2.87

For the presentation of the following results, initially, we will provide evidence of the reproducibility of the regression model trained in the reference study, followed by applying MINREM to the same dataset prepared by Mullainathan and Spiess (2017), and finally, we will present a comparison.

4.2.2. Reproducing the reference regression model

Mullainathan and Spiess (2017) do not explicitly present their regression model. They consolidated their findings and presented the R^2 both for the training sample and for the validation. However, for the reproduction of their trained models, it is enough to study the code they shared and execute the following lines in *R* (after execution of the preparation steps, indicated by those authors): 'linmod <- lm(modelformula,df[from,])', which is in the file 'algo_linear.R', folder 'MullainathanSpiess/predictiontools.' Next, we reproduce the content of 'modelformula' which stores the expression of the regression model.

"LOGVALUE ~ PHONE + KITCHEN + MOBILTYP + WINTEROVEN + WINTERKESP +\n WINTERELSP + WINTERWOOD + WINTERNONE + NEWC + DISH + WASH +\n DRY + NUNIT2 + BURNER + COOK + OVEN + REFR + BATHS + BEDRMS +\n DENS + DINING + FAMRM + HALFB + KITCH + LIVING + OTHFN +\n RECRM + BUILT + LOT + UNITSF + CLIMB + ELEV + DIRAC + PORCH +\n AIRSYS + WELL + WELDUS + STEAM + OARSYS + FRPL + FRPLI +\n FLOT + FPLWK + FPINS + DISPL + TRASH + TYPE + EOTEAPP + ENOEAPP +\n ECNTAIR + EAIRC + EHEATUT + EFRIDGE + EDRYER + EWASHR + EDISHWR +\n ETRSHCP + AIR + NUMAIR + WATER + WATERD + WELLDIS + WELLDIS2 +\n SEWDIS + HOTPIP + PUBSEW + SEWDISTP + SEWDUS + KEXCLU + SINK +\n GARAGE + INCP + BUSPER + EXCLUS + LAUNDY + OTHRUN + DRSHOP +\n FLOORS + CONDO + CELLAR + WHNGET + FRSTOC + PREOCC + EBAR + \n SHARAT + SHARFR + OTBUP + NUNITS + PLUGS + OWNLOT + ROOMS +\n PLUMB + ZADEQ + LEAK + IFTLT + ILEAK + WHYCD1 + WHYCD2 +\n WHYCD3 + WHYCD4 + WHYCD5 + RLEAK + BLEAK + WLEAK + OTLEAK +\n PLEAK + PILEAK + WTRHRL + NLEAK1 + NLEAK2 + RATS + MICE +\n NOTSUR + EGOOD + HOWH + WATERS + BSINK + TOILET + ELEVWK +\n EROACH + EVROD + M12ROACH + M12ROD + RATFREQ + ROACHFRQ +\n CRACKS + EBOARD + EBROKE + ECRUMB + EHOLER + EMISSR + EMISSW +\n ESAGR + ESLOPW + HOLES + IFBLOW + NUMBLOW + FREEZE + IFCOLD +\n IFDRY + IFSEW + NUMCOLD + NUMDRY + NUMSEW + NUMTLT + OTHCLD +\n NOWIRE + REGION + METRO + METRO3 + LOTMISS + UNITSFMISS + \n CLIMBMISS + DIRACMISS + NUMAIRMISS + BUSPERMISS + EXCLUSMISS +\n HOWHMISS + NUMCOLDMISS + NUMDRYMISS + NUMSEWMISS + NUMTLTMISS".

Table 9 provides a brief extract of the summary of the regression, since its complete presentation would be too taxing, having 163 variables (including the dependent one).

The prediction model trained in Mullainathan and Spiess (2017) yielded an R^2 of 47.26%, 1,926 degrees of freedom, and incorporated 162 independent variables, 41 of which presented *p*-value less than 0.05. However, as they warn, the models resulting from analyzing big data and using machine learning approaches should be applied mostly for predictive purposes, instead of inferential ones, for reasons such as

Table 9. Summary of the regression model trained and evaluated in Mullainathan and Spiess (2017).

Variables	Coef. (Std. Error)	Variables that were statistically significant
REGION2	−0.1816*** (0.0309)	
REGION3	−0.1258*** (0.0358)	Using the training simple with seed: 'set.seed (20170313)':
REGION4	0.4616*** (0.0352)	
METRO2	0.0550 (0.0600)	'REGION'; 'PHONE';"WINTEROVEN"; 'DISH'; 'WASH';
METRO3	0.1438 (0.1169)	'BATHS';"BEDRMS";"DENS"; 'DINING'; 'FAMRM'; 'HALFB';
METRO4	0.2583 (0.1596)	'OTHFN'; 'RECRM'; 'BUILT'; 'UNITSF'; 'FPLWK'; 'FPINS'; 'TRASH';
METRO7	0.1274* (0.0693)	'TYPE'; 'EWASHR'; 'AIR'; 'NUMAIR'; 'WATERD'; 'SEWDIS';
METRO32	−0.1252* (0.0673)	'SEWDUS'; 'FLOORS'; 'CELLAR'; 'FRSTOC'; 'OTBUP'; 'PLUMB';
METRO39	NA	'RATS'; 'HOWH'; 'WATERS'; 'EVROD'; 'RATFREQ'; 'EBROKE';
PHONE1	−0.0654 (0.0433)	'HOLES'; 'IFDRY'; 'IFSEW'; 'NUMSEW'; 'HOWHMISS'
PHONE2	−0.2638*** (0.0798)	
…	…	
…	…	Tested variables 162
…	…	Degrees of freedom **9,726**
HOWHMISS-7	0.2026 (0.1840)	Significant variables [a] **41**
HOWHMISS0	0.2182*** (0.0764)	Observations **10,000**
NUMCOLDMISS0	0.3105 (0.4322)	R^2 [b] **0.4726**
NUMDRYMISS0	−0.2961 (0.5561)	Adjusted R^2 0.4578
NUMSEWMISS0	NA	Residual Std. Error 0.7782
NUMTLTMISS0	NA	F Statistic 31.9229*** (df = 273; 9,726)
Constant	20.8715*** (2.4278)	*p < 0.1; **p < 0.05; ***p < 0.01

[a]Only considering *p-value* less than 0.05 (for exploration and without testing the assumptions of the regression model)
[b]The reported values apply to include all 162 independent variables in the model, as was done in Mullainathan and Spiess (2017)

collinearity between several of the regressors, redundant or irrelevant variables, and a data-driven analyses. In fact, when we are trying to explore the VIF in *R*, this reports the following message: 'there are aliased coefficients in the model'.

4.2.3. Using MINREM on the reference training sample

For the use of MINREM on the dataset prepared by Mullainathan and Spiess (2017), the sample size was varied from 1,000 observations to 10,000 observations (the total training sample), with increments of 1,000. See, in Figure 8, the behavior of the importance and stability of the variables under MINREM, according to the sample size.

The purpose of Figure 8 is not to show the exact label of each variable (in fact, this was excluded for better visualization), but to show the implicit complexity, the behavior of the importance of the variables as the sample size increases, and the sets of variables that stand out in the observation region. Although in the regression model trained by Mullainathan and Spiess (2017), 162 independent variables were used (for predictive purposes), in this case only a minority of them are robust (not very sensitive) to changes in the sample size. Note that some variables are important (significant at least 95% of the repetitions) already beginning with the first levels of the sample, whereas others remain stably irrelevant, throughout the study region. Other variables have an increasing behavior concerning their importance as the sample size increases, which means that these are the variables that present the analyst the most uncertainty. These results lead us to conclude that, faced with the big data approach from the regression analysis, three subsets of independent variables can stand out. One that is important even from low sample sizes, another that is irrelevant throughout the region of observation, and another, requiring greater care and affording reasonable doubt, whose conclusions can change markedly, depending on the observations that are used.

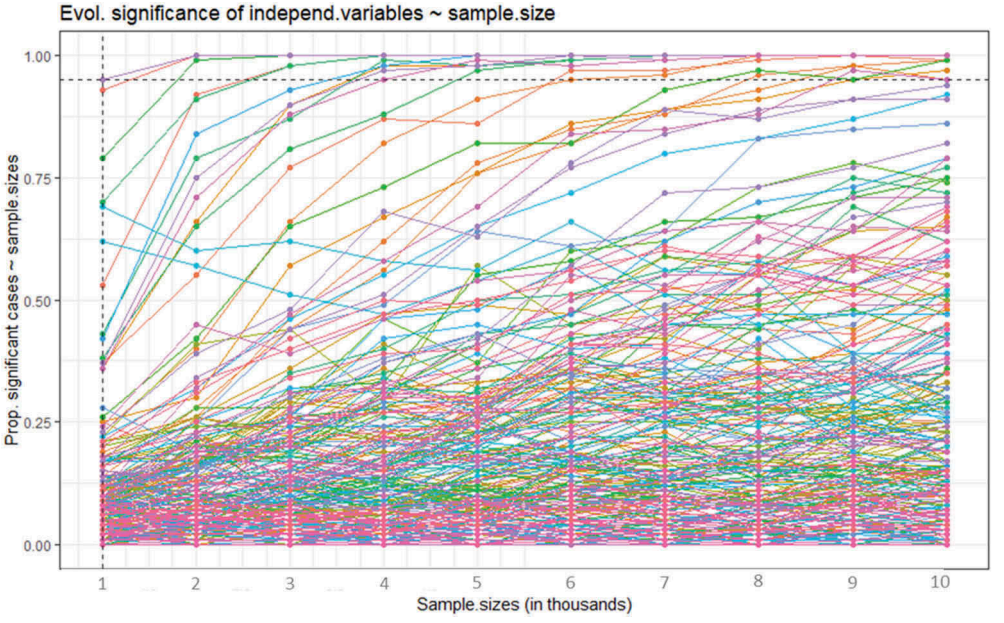

Figure 8. Evolution of the proportion of cases in which the independent variables were significant, according to sample size, using the dataset prepared by Mullainathan and Spiess (2017).

For greater details of the levels of the variables that under MINREM were more important (significant in 95% of the cases), as well as the most irrelevant, see Table 10.

In Table 10 it can be seen that, in total, the 162 independent variables yielded 273 levels (increase as a result of the categorical variables); hence, the degrees of freedom of the reference model, which included the 162 independent variables, was 9,726 (the size of the training sample minus 273 levels, minus 1). Table 10 details 30 of the total levels of the variables (first and last 15 levels). The first 13 levels (ending in 'CELLAR2') are generated by 12 variables since 'REGION2' and 'REGION4' belong to the same categorical variable 'REGION.' Note that, at most, a sample of 60% of the total training observations was necessary, so that these 13 levels went from irrelevant to important. It is noteworthy that the REGION4 level was important (i.e., at least 95% of the cases yielded p-values less than 0.05) even with 1,000 observations (10% of the training sample).

For this exploration of the utility of MINREM, after considering Figure 8 and Table 11, the model to train used as criterion Ndepura of 6,000 observations. It is worth mentioning that when going from 1,000 to 6,000 observations, with an increase of 1,000, in each case a regressor variable entered the area of importance (95%). On the other hand, when going from 6,000 to 7,000 observations, there was no change in this set of variables; one had to wait until the next change (8,000 observations) for another regressor to be included. It should also be noted that when going from 5,000 to 6,000 observations, the R^2 of the resulting model increased by 10.3%, but when going from 6,000 to 8,000 observations, the increase in R^2 was only 0.35% (the variable 'FAMRM' was added). Table 11 provides evidence for the the trained regression model from the 12 important variables, obtained with Ndepura of 6,000 observations.

Table 10. The proportion of cases in which, under MINREM, the level of the independent variable begins to be statistically significant (p-val < 0.05), according to sample size (subject to the predefined increment of 1,000 observations).

		Incremental sample										
No.	Levels of variables	1000	2000	3000	4000	5000	6000	7000	8000	9000	10,000	Relevant from:
1	REGION4	0.95	1	1	1	1	**1**	1	1	1	1	10%
2	BATHS	0.93	1	1	1	1	**1**	1	1	1	1	20%
3	FPLWK2	0.79	0.99	1	1	1	**1**	1	1	1	1	20%
4	BEDRMS	0.53	0.92	0.98	1	1	**1**	1	1	1	1	30%
5	HALFB	0.70	0.91	0.98	1	1	**1**	1	1	1	1	30%
6	REGION2	0.37	0.75	0.90	0.97	0.98	**1**	0.99	1	1	1	40%
7	OTBUP2	0.42	0.84	0.93	0.98	1	**1**	1	1	1	1	40%
8	DENS	0.36	0.66	0.90	0.98	0.98	**0.99**	1	1	1	1	40%
9	HOWH	0.43	0.79	0.87	0.99	0.98	**0.99**	1	1	1	1	40%
10	UNITSF	0.36	0.71	0.88	0.95	0.99	**0.98**	0.99	1	1	1	40%
11	FRSTOC1	0.38	0.65	0.81	0.88	0.97	**0.99**	1	1	1	1	50%
12	BUILT	0.37	0.55	0.77	0.87	0.86	**0.97**	0.97	0.99	1	0.99	60%
13	CELLAR2	0.23	0.4	0.66	0.82	0.91	**0.95**	0.96	1	1	1	**60%**
14	CELLAR1	0.19	0.26	0.46	0.62	0.78	0.85	0.88	0.96	0.98	0.99	80%
15	TRASH2	0.11	0.32	0.47	0.58	0.69	0.84	0.85	0.88	0.97	0.95	90%
	…	…	…	…	…	…	…	…	…	…	…	0%
259	RATFREQ2	0.03	0.02	0.03	0	0.04	0	0.04	0.02	0.04	0	
260	ROACHFRQ2	0.03	0.01	0	0.03	0.03	0	0.04	0.03	0.04	0.02	0%
261	IFBLOW2	0.01	0.02	0.02	0.03	0.01	0	0.02	0.04	0.02	0.05	0%
262	NUMBLOW1	0.03	0.01	0	0.01	0.01	0	0.02	0	0	0	0%
263	NUMBLOW2	0.03	0.01	0	0	0.01	0	0.02	0	0	0.01	0%
264	NUMBLOW4	0.03	0.01	0	0	0.01	0	0.01	0	0	0.01	0%
265	NUMBLOW6	0.02	0	0	0.01	0.01	0	0.01	0.01	0.01	0.02	0%
266	NUMBLOW7	0	0	0	0	0	0	0	0	0	0.01	0%
267	NUMBLOW8	0	0	0	0	0.01	0	0	0	0.01	0.01	0%
268	FREEZE1	0.09	0.06	0.02	0.04	0.01	0	0.01	0.02	0.02	0.01	0%
269	IFCOLD1	0.09	0.04	0.02	0	0	0	0	0	0.01	0	0%
270	IFCOLD2	0	0	0	0	0	0	0	0	0	0	0%
271	OTHCLD1	0	0	0	0	0	0	0	0	0	0	0%
272	NOWIRE2	0.03	0.05	0.04	0.01	0.05	0	0.02	0.02	0.04	0.01	0%
273	NUMDRYMISS0	0	0	0	0.02	0.01	0	0	0.03	0	0	0%

*The proportions within each sample size were calculated using resampling of 100 repetitions.

Table 11 shows that with only 12 of the 162 return variables tested, the trained model achieves an R^2 of 36.39% and 9,979 degrees of freedom.

4.2.4. Comparative results

Table 12 presents the comparative results of case 2, derived from training the regression models under the traditional strategy (refer) vs. the MINREM strategy (refer.minrem and refer.minrem2, the latter with double interactions), not only considering the sample of training, but also the totality of the validation sample, 41,808 observations.

From Table 12, it is worth noting the following findings:

- Regarding the degrees of freedom, the models obtained using MINREM (for the variable selection) exceeded the reference model (trained traditionally) by 2.6% without considering interactions (refer.minrem) and by 0.8% including double interactions (refe.minrem2).

Table 11. Training of the regression model under MINREM, using the same reference case dataset.

Variables	Coef. (Std. Error)	Variables that were important under MINREM [a]	
REGION2	−0.2464*** (0.0309)		
REGION3	−0.2545*** (0.0351)	"REGION" (Census: 1. Northeast, 2. Midwest, 3. South, 4. West)	
REGION4	0.4846*** (0.0347)	"BATHS" (Number of full bathrooms in unit)	
BATHS	0.2616*** (0.0154)	"BEDRMS" (Number of bedrooms in unit)	
BEDRMS	0.0736*** (0.0124)	"DENS" (Number of dens/libraries/tv rooms in unit)	
DENS	0.1118*** (0.0222)	"HALFB" (Number of half bathrooms in unit)	
HALFB	0.2080*** (0.0163)	"BUILT" (Year unit was built) [b]	
BUILT	−0.0029*** (0.0005)	"UNITSF" (Square Footage of Unit)	
UNITSF	0.00003*** (0.00001)	"FPLWK" (Unit has useable fireplace; 1: Yes; 2: No)	
FPLWK2	−0.3357*** (0.0195)	"CELLAR" (Unit has a basement; from 1 to 5; −6: Not applicable)	
CELLAR1	0.8207*** (0.0381)	"FRSTOC" (Current occupants are first occupants; 1: First occupants; 2: Previously occupied; −1: missing values)	
CELLAR2	0.7739*** (0.0441)	"OTBUP" (Other building on property used as living quarters; 1: Yes; 2: No; −1 missing values)	
CELLAR3	0.9897*** (0.0377)	"HOWH" (Rating of unit as a place to live; 1: worst… to 10:best)	
CELLAR4	0.8055*** (0.0353)	Tested variables	162
CELLAR5	1.0662*** (0.0759)	Degrees of freedom	**9,979**
FRSTOC1	0.0932*** (0.0310)	Stable-important variables	**12**
FRSTOC2	0.0493* (0.0298)	Observations	**10,000**
OTBUP1	−0.0723 (0.0676)	R^2 [a]	**0.3639**
OTBUP2	−0.0333* (0.0189)	Adjusted R^2	0.3627
HOWH	0.0647*** (0.0061)	Residual Std. Error	0.8437 (df = 9,979)
Constant	15.5939*** (1.0103)	F Statistic	285.4821*** (df = 20; 9,979)
		*p < 0.1; **p < 0.05; ***p < 0.01	

[a]Detail of names or levels of the variables from 'Codebook for the American Housing Survey, 1997–2011, March 2013, Version 2.1', (Econometrica, 2013).

[b]In order to maintain the same training conditions of the reference case, BUILT was assumed as a quantitative variable; however, it might be more useful to treat it as a categorical variable.

- About the number of variables finally used, under MINREM the model went from 162 independent variables to only 12, which is a reduction by nearly 93%, compared to the traditional strategy.
- In the training sample, MINREM led to models with an R^2 lower by 23% without interactions and by 8.5% with double interactions, compared to traditional training. In contrast, in the validation sample (41,808 observations) the difference was much smaller (15% and 2.5%, respectively).
- Considering the original variable in dollars, and compared to the reference model, 'refer.minrem' (without interaction) produced between 6.2% and 6.5% fewer properties with maximum prediction errors of 10%, 15%, and 20%. However, when including double interactions (refer.minrem2), there was an increase of between 1.3% and 2.6% in the number of properties within these margins of error, in contrast to the reference model.

In summary, the training strategy under MINREM, compared to traditional training, has considerably favored the parsimony and efficiency of the resulting models, while at the same time it has achieved an efficacy close to that achieved with the whole set of available variables (especially in the validation sample and original monetary scale). The reduction of variables has even made it feasible to include double interactions in the model and, in some cases, we have obtained a better predictive capacity. As mentioned in the Colombian case, this addition of double interactions is a useful alternative merely for predictive purposes. However, although previously it was assumed to be infeasible

Table 12. Comparing the traditional training vs. MINREM in the reference case. Training sample: 10,000 observations; Validation: 41,808 observations.

Original:

		refer	refer.minrem	refer.minrem2
Degrees of freedom	df.training Change	9,726	9,979	9,804
			2.6%	**0.8%**
Independent variables	Used	**162** (41)[a]	**12**	**12**
	Change		**−92.6%**(-70.7%) [a]	
R^2*	Training	47.3%**	36.4%	43.3%
Data in natural log	Change		**−23.0%**	**−8.5%**
	Validation	41.9%***	35.6%	40.7%
	Change		**−15.0%**	**−2.9%**
Percentage of properties with a maximum prediction error of *r*%. Data in original scale (dollars), using *exp*	Err.máx 10%	15.4%	14.4%	15.8%
	Change		**−6.5%**	**2.6%**
	Err.máx 15%	23.1%	21.6%	23.4%
	Change		**−6.5%**	**1.3%**
	Err.máx 20%	30.6%	28.7%	31.1%
	Change		**−6.2%**	**1.6%**

Refer: Reference model (Mullainathan & Spiess, 2017); refer.minrem: the model trained under MINREM strategy using Ndepura = 6,000; refer.minrem2: model refer.minrem with double interactions. df.training: Degrees of freedom when training the model with the 10,000 observations of the reference case. Err.max: Percentage of predictions that did not exceed a specific % error. Change: % change concerning the results of the reference model (refer).

[a]41 variables of the 162 were significant, considering p-values less than 0.05.

*Square of the correlation between real LOGVALUE vs. predicted.

**It coincided with that reported by Mullainathan and Spiess (2017, p. 90) and was reproduced in Table 9, previous use of the data and function of Mullainathan and Spiess (2017). All the values exposed for the reference model (refer), in Table 12, were obtained using all the variables tested (162) and not only the 41 with p-values less than 0.05, which was thus done in Mullainathan and Spiess (2017, p. 90).

***Mullainathan and Spiess (2017) used 8-fold validation and averaged the results, obtaining an R^2 was 41.7%, and bootstrap intervals 95% of [39.7%, 43.7%].

due to the large number of variables that may be present in big data (Mullainathan & Spiess, 2017), given the parsimony achieved under MINREM it is much more feasible to use it to favor the prediction capacity of the models, under a machine learning approach. Additionally, thanks to the proposed variable selection procedure, the inclusion of irrelevant or redundant variables can be mitigated, so that the inference of possible causal effects if these effects existed, would be more viable.

5. Conclusions

This work provides a two-stage methodology for the analysis of big data regression under a machine learning approach, which facilitates inference and prediction regarding real estate data.

The first stage exploits the selection of the important variables, under a strategy called incremental sample with resampling (MINREM). The second stage is oriented to the training and validation of models from the classic approach of machine learning, but adding three activities: 1) comparison of the debugged model under MINREM against the model trained traditionally, 2) inclusion of double interactions for the refined model, and 3) estimation and interpretation of metrics not only on a transformed scale (classically the natural logarithm of the property price) but, above all, in the original monetary scale.

The methodology showed its practical value, being successfully tested both on an original Colombian case and an international reference. The first case also served to

demonstrate the potential of web advertisements for the sale of properties, as one of the representations of massive online data, which generate extensive opportunities for statistical learning due to its high frequency, accessibility and availability (Cavallo, 2012, 2017). The second case allowed deepening the study of Mullainathan and Spiess (2017) and taking another look hedonic regression under machine learning, finding it much more promising. Regarding prediction, some findings obtained with our model were close to the prediction capabilities of the reference case, but with much more parsimony (and the benefits that this implies). In turn, other findings obtained with our model were superior to those of the reference model, while also maintaining high parsimony, especially when double interactions were used. These results are essential, since, from the perspective of machine learning, it is common to find statements about the disadvantages of hedonic regression for big data, in comparison with intensive computer methods. In this regard, it would be worth trying, in specific contexts, to see whether applying MINREM previously and then using double interactions (viable due to parsimony), the performance of intensive computer methods, such as random forest, could be superior or not to the hedonic regression models. Regarding inference, in the first case, the interpretative use of the created function from the big data was seen to lead to a better understanding of the logic of the phenomenon and the effects of four fundamental attributes, notably stable throughout the study region and also backed by previous studies.

This methodology is justified by the need to combine the strengths of hedonic regression (it potential for inference) and those of machine learning (its predictive potential for big data). However, this integration is not elementary, but requires the incorporation of a variable selection procedure (Bin et al., 2017; Cai et al., 2018; Cateni & Colla, 2016; Varian, 2014). In this way, MINREM established criteria to obtain a subset of stable-important variables based on a semi-parametric approach, which uses the significance of the variables in the regression and resampling with incremental sample sizes. The need for this type of strategy is supported by the high instability reported for machine learning models, in inferential terms (but not predictive). Additionally, the hedonic regression, which facilitates understanding the effects of the regression coefficients and, in general, interpreting the resulting functions, is limited when working with big data, due to the high correlations between variables, the presence of useless variables, and the high power of the tests, as well as the disadvantages for prediction of using big data. In turn, the hedonic price theory does not provide inputs that guide the selection of the important variables or the establishment of the functional form of the model (Andersson, 2000; Freeman, 2003). Although the proposed methodology cannot guarantee the possibility of making a causal inference, it does help to estimate causal effects when these effects occur, which is in sync with approaches such as those of Varian (2014).

In general, this study contributes to complement the traditional rollout of OLS regression with machine learning approach strengths, among them: validation in non-training samples to mitigate overfitting; variable selection strategies using resampling (but this time under a new strategy: MINREM); creation/use of algorithmic functions to manipulate big data; and evaluation of the predictive capacity of the models (not only on a transformed scale, but also on a monetary scale). Likewise, the machine learning approach used in this study has been nourished with interpretive strengths, thanks to the following: analysis of the importance and stability of the

variables, incremental samples with automated resampling; validation of compliance with OLS regression fundamental assumptions; and theoretical considerations or findings from other studies to reinforce the pertinence of the interpretations generated. All this was tested in two sets of data (case studies) that differ in terms of geographical context, source, number of observations, and type and number of variables.

The type of regression discussed and addressed under machine learning in this study was OLS, which is the most common type in the analysis of real estate prices. Future works could consider other econometric resources, such as spline estimates and other non-parametric techniques, in order to continue contributing to the approach to big data management for inferential and predictive purposes. In this sense, when focusing on big data, these studies must face, at least, the same problems mentioned here (instability of the variables, …). Therefore, both MINREM and the proposed methodology (Figure 1) for big data regression analysis constitute useful tools for future studies. Additionally, highlighting MINREM in particular, other studies, this time focused on machine learning prediction, could be used to identify more stable-important variables, before starting with the final training of the models; it would help the stability of the solutions and computational efficiency.

In order to facilitate the kind of scrutiny which is highly demanded in the field of real estate (Krause, 2016), this article includes supplementary material, which contains the general code (using *Rmarkdown*), employed for the treatment and analysis of the data (see footnote, first page).

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Jorge Iván Pérez-Rave Industrial Engineer (Universidad de Antioquia, Colombia). Specializations: (1) Statistics, and (2) Systems Engineering (Universidad Nacional de Colombia). Masters: (1) Systems Engineering (Universidad Nacional de Colombia), and (2) Visual Analytics and Big Data (UNIR España). Director of the IDINNOV research group. Ph.D candidate in (Systems Engineering, Universidad Nacional de Colombia), and Ph.D candidate in (Business Management; Universidad de Valencia, España).

Juan Carlos Correa-Morales Statistician (University of Medellín), Master of Statistics (University of Kentucky), PhD. in Statistics (University of Kentucky). Professor (Universidad Nacional de Colombia)

Favián González-Echavarría Industrial Engineer (Universidad de Antioquia, Colombia), Master's Degree in Economics (Universidad de Antioquia). Professor (Universidad de Antioquia), Ph.D candidate in Business Management (Universidad de Valencia, España).

ORCID

Jorge Iván Pérez-Rave http://orcid.org/0000-0003-1166-5545
Juan Carlos Correa-Morales http://orcid.org/0000-0002-9368-4725
Favián González-Echavarría http://orcid.org/0000-0002-1540-9859

References

Abdallah, S., & Khashan, D. (2016). *Using text mining to analyze real estate classifieds*. 2nd International Conference on Advanced Intelligent Systems and Informatics, *533*(2017), 193–202). Cairo, Egypt: Springer International Publishing. https://link.springer.com/chapter/10.1007/978-3-319-48308-5_19.

Adetiloye, K., & Eke, O. (2014). A review of real estate valuation and optimal pricing techniques. *Asian Economic and Financial Review*, *4*(12), 1878–1893.

Andersson, D. E. (2000). Hypothesis testing in hedonic price estimation–On the selection of independent variables. *The Annals of Regional Science*, *34*(2), 293–304.

Athey, S. (2018). The impact of machine learning on economics. In Ajay K. Agrawal, Joshua Gans, and Avi Goldfarb, editors, *The economics of artificial intelligence: An Agenda*. University of Chicago Press.

Banerjee, D., & Dutta, S. (2017, September). Predicting the housing price direction using machine learning techniques. In 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI) (pp. 2998–3000). India: IEEE.

Beręsewicz, M. E. (2015). On representativeness of Internet data sources for real estate market in Poland. *Austrian Journal of Statistics*, *44*(2), 45–57.

Bin, J., Tang, S., Liu, Y., Wang, G., Gardiner, B., Liu, Z., & Li, E. (2017, September). *Regression model for appraisal of real estate using recurrent neural network and boosting tree*. Computational Intelligence and Applications (ICCIA), 2017 2nd IEEE International Conference on (pp. 209–213). Beijing, China: IEEE.

Bonetti, F., Corsi, S., Orsi, L., & De Noni, I. (2016). Canals vs. streams: to what extent do water quality and proximity affect real estate values? A hedonic approach analysis. *Water*, *8*(12), 577. doi:https://doi.org/10.3390/w8120577

Borde, S., Rane, A., Shende, G., & Shetty, S. (2017). Real estate investment advising using machine learning. *International Research Journal of Engineering and Technology (IRJET)*, *4*(3), 1821–1825.

Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing*, *300*, 70–79.

California Association of Realtors® (C.A.R.) (2015). Big data and real estate: adapting to the new information economy. White paper. Author/Researcher: Brown, J. Editor: Framroze. Center for California Real Estate. Roundtable series, 1–20, http://centerforcaliforniarealestate.org/publications/car_big_data_final_102715_pgs_web.pdf

Castillo, M. (2015). Violencia de pareja en el Paraguay según la Encuesta Nacional de Demografía y Salud Sexual y Reproductiva 2008. *Revista Latinoamericana de Población*, *5*(9), 27–48.

Cateni, S., & Colla, V. (2016). Variable selection for efficient design of machine learning-based models. In: C. Jayne & L. Iliadis (Eds), Engineering applications of neural networks: 17th international conference, EANN 2016 (pp 352–366). Aberdeen, UK, September 2–5, 2016, proceedings, Retrieved from https://link.springer.com/chapter/10.1007/978-3-319-44188-7_27

Cavallo, A. (2012). Scraped Data and Sticky Prices. SSRN Scholarly Paper ID 1711999. Social Science Research Network, Rochester, NY.

Cavallo, A. (2017). Are online and offline prices similar? Evidence from large multi-channel retailers. *American Economic Review*, *107*, 283–303.

Čeh, M., Kilibarda, M., Lisec, A., & Bajat, B. (2018). Estimating the performance of random forest versus multiple regression for predicting prices of the apartments. *ISPRS International Journal of Geo-Information*, *7*(5), 168.

Chen, K., & McCluskey, J. (2018). Impacts of expert information on prices for an experience good across product segments: tasting notes andwine prices. *Journal Of Agricultural and Resource Economics*, *43*(3), 388–402.

Clavijo, S., Janna, M., & Muñoz, S. (2005). La vivienda en Colombia: Sus determinantes socioeconómicos y financieros. *Revista Desarrollo y Sociedad*, (55), 101–165.

Dorr, T. (2016). Real Estate Home Pricing: Hedonic model variables and Community Amenities Roles. *Journal of Management and Innovation*, *2*(2), 2016. Recovered 08/ 15/2017, from http://jmi.mercy.edu/index.php/JMI/article/view/27

Dubin, R. (1998). Predicting house prices using multiple listings data. *Journal of Real Estate Finance and Economics*, *17*(1), 35–39.

Econometrica (2013). Codebook for the American Housing Survey, Public Use File: 1997-2011, March 2013, Version 2.1. Prepared for U.S. Department of Housing & Urban Development Office of Policy Development & Research. Recovered 12/12/2018, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.369.5653&rep=rep1&type=pdf

Figueroa, E. (1992). Determinantes del precio de la vivienda en Santiago: Una estimación hedónica. *Estudios de Economía*, *19*(1), 67–84.

Fletcher, M., Gallimore, P., & Mangan, J. (2000). Heteroscedasticity in hedonic house price models. *Journal Of Property Research*, *17*(2), 93–108.

Flórez, R., & Arias, N. (2010). Evaluación de conocimientos previos del aprendizaje inicial de lectura. *Revista internacional de investigación en educación*, *2*(4), 329–344.

Fonseca, D., Gutiérrez, A., Mateus, H., Silva, C., Contreras, N., & Giraldo, A. (2005). Análisis de muestras de orina para la detección molecular de enfermedades infecciosas. Aplicación en la identificación de citomegalovirus humano. *Revista Ciencias de la Salud*, *3*(2), 136–147.

Freeman, A. M., III, Herriges, J. A., & Kling, C. L. (2014). *The measurement of environmental and resource values: Theory and methods*. New York, NY: Routledge.

Frické, M. (2015). Big data and its epistemology. *Journal of the Association for Information Science and Technology*, *66*(4), 651–661.

George, G., Osinga, E. C., Lavie, D., & Scott, B. A. (2016). Big data and data science methods for management research. *Academy of Management Journal, from the Editors*, *59*(5), 1493–1507.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference and prediction*, (2nd ed.). Berlin, Germany: Springer.

Herrán-Falla, O. F., Prada-Gómez, G. E., & Patiño-Benavidez, G. A. (2003). Canasta básica alimentaria e índice de precios en Santander, Colombia, 1999-2000. *Salud Pública Mex*, *45*, 35–42.

Holland, J. (2016). Integrating data science and commercial real estate. NAI partners, 2016, Retrieved on February 05, 2019, http://www.naipartners.com/Portals/248/DATA_SCIENCE_BRO.pdf

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2015). *An introduction to statistical learning: with applications in* R(6th ed.). New York: Springer.

Krause, A. (2016). Reproducible research in real estate: a review and an example. *Journal Of Real Estate Practice and Education*, *19*(1), 69–85.

Leung, C., & Jiang, F. (2014, December). A data science solution for mining interesting patterns from uncertain big data. In 2014 IEEE Fourth International Conference on Big Data and Cloud Computing (pp. 235–242). Sydney; Australia: IEEE.

Limsombunchai, V. (2004, June). *House price prediction: Hedonic price model vs. artificial neural network*. New Zealand Agricultural and Resource Economics Society Conference (pp. 25–26). Blenheim, New Zealand.

Lowrance, R. (2015). *Predicting the market value of single-family residential real estate* (Doctoral dissertation). New York University.

Mok, H., Chan, P., & Cho, Y. (1995). A hedonic price model for private properties in Hong Kong. *The Journal of Real Estate Finance and Economics*, *10*(1), 37–48.

Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87–106.

Oladunni, T., & Sharma, S. (2016). *Hedonic housing theory—A machine learning investigation*. Machine Learning and Applications (ICMLA), 2016 15th IEEE International Conference on (pp. 522–527). Anaheim, United States: IEEE.

Pardoe, I. (2008). Modeling home prices using realtor data. *Journal of Statistics Education*, *16*(2), read on. doi:10.1080/10691898.2008.11889569

Pérez-Rave. (2019). Statihouse®: Desarrollo tecnológico basado en Ciencia de Datos para explorar estadísticamente el sector inmobiliario. *Ingeniare: Revista Chilena De Ingeniería*, *27*(1), ISSN: 0718-3305 versión en línea, enero – marzo, in press.

Puyun, B., & Miao, L. (2016, March). *Research on analysis system of city price based on big data*. 2016 IEEE International Conference on Big Data Analysis (ICBDA) (pp. 1–4). Hangzhou, China: IEEE. retailers. American Economic Review, 107(1):283–303.

Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy*, *82*(1), 34–55.

Seo, K., Salon, D., Shilling, F., & Kuby, M. (2017). Pavement Condition and Residential Property Values: A Spatial Hedonic Price Model for Solano County, CA (No. 17-05145). Recovered 08/15/2017, from http://docs.trb.org/prp/17-05145.pdf

Shi, H. (2009). *Determination of real estate price based on principal component analysis and artificial neural networks*. Intelligent Computation Technology and Automation, ICICTA'09. Hunan, China: Second International Conference on IEEE 314–317. 1, 2009.

Trawiński, B., Telec, Z., Krasnoborski, J., Piwowarczyk, M., Talaga, M., Lasota, T., & Sawiłow, E. (2017, July). *Comparison of expert algorithms with machine learning models for real estate appraisal*. INnovations in Intelligent SysTems and Applications (INISTA), 2017 IEEE International Conference on (pp. 51–54). Gdynia, Poland: IEEE.

Tuñón, I., & Halperin, V. (2010). Desigualdad social y percepción de la calidad en la oferta educativa en la Argentina urbana. *Revista electrónica de investigación educativa*, *12*(2), 1–23.

Tuñón, I., & Poy, S. (2016). Factores asociados a las calificaciones escolares como proxy del rendimiento educativo. *Revista electrónica de investigación educativa*, *18*(1), 98–111.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, *28*(2), 3–28.

Wang, X., & Zhang, J. (2013). *Principal component analysis of influencing factors of the development of China's real estate market*. ICCREM 2013: Construction and Operation in the Context of Sustainability (pp. 1027–1035), Karlsruhe, Germany.

Wu, J., Gyourko, J., & Deng, Y. (2012). Evaluating conditions in major chinese housing markets. *Regional Science and Urban Economics*, *42*(3), 531–543.

Yoo, S., Im, J., & Wagner, J. E. (2012). Variable selection for hedonic model using machine learning approaches: A case study in Onondaga County, NY. *Landscape and Urban Planning*, *107*(3), 293–306.