Land Use Policy 79 (2018) 116–122

Contents lists available at ScienceDirect

Land Use Policy

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

Large-scale simultaneous market segment definition and mass appraisal using decision tree learning for fiscal purposes

Fabián Reyes-Bueno, Juan Manuel García-Samaniego, Aminael Sánchez-Rodríguez*

Departamento de Ciencias Biológicas, Universidad Técnica Particular de Loja.,San Cayetano Alto s/n, PC: 110104, Loja, Ecuador

ABSTRACT

Cadastral assessment aims at guarantying equity in the allocation of property taxes. Therefore, we must be able to massively determine property values through models that reflect, with the minimum error, the behaviour of land market in each region. Despite this imperative need, currently land valuation for cadastral purposes is plagued with subjectivity. A very extended bad practice for instance is to assume that variables of productive performance i.e. land use capacity, are the ones with the highest influence on land value formation in the rural sector. The former assumption largely ignores the plethora of rural land uses that exist nowadays. To open the door to less subjective methodologies of land mass appraisal we borrowed statistical methodologies from the field of data-mining and applied them to a dataset of 410 purchase-sale transactions (2003–2009) of land plots located in the rural sector of the Vilcabamba parish (southern Ecuador). Land market behaviour in Vilcabamba responds to a transition from a pure agricultural territory to a touristic one at which many second-homes are being built for leisure. Our results demonstrate the applicability of methodologies such as model-tress (M5P) and multivariate adaptive regression splines (MARS) to rural land mass appraisal. Both M5P and MARS allow defining market segments while simultaneously establishing the weights of predictor variables for land value formation. We also collected evidence supporting that removing variables of productive performance from land value prediction models do not hamper models predictive power at least in rural areas where gentrification is taking place.

1. Background

Cadastral assessment is a mass appraisal process of property groups used for calculating the real property tax (Baumane, 2010). During cadastral assessment a property value is commonly calculated by value determination models which aim at reducing errors during value estimation. Cadastral assessment is very important to ensure right real property taxation and the principles of equality (Baumane, 2010). In Ecuador, the cadastral value of a given property is the basis for taxation and for establishing rates (*e.g.* the corresponding percentage for fire brigades) and special contributions (*e.g.* to pay for a specific infrastructure work on a given area). Cadastral value of a property in Ecuador is also taking into account expropriation and compensation processes. According to the Ecuadorian legislation (Ecuador, 2010) the cadastral value of a property in a given sector should be established by adding to the land value, the property value itself. Property values are determined by comparing against unit prices of comparable properties from the same sector. The resulting cadastral (market) value is then the most probable price (in terms of money) of a property in a competitive and open market provided the conditions needed to guarantee a fair sale. Both the buyer and seller must act prudently and knowledgeably, and it is commonly assumed that the price is not affected by undue stimulus (Iaao, 2011).

From the methodological point of view, cadastral assessment consists of three stages (Fig. 1):

1.1. *Stage 1: seed points selection*

A prerequisite for cadastral value formation is the identification of comparable properties in the same sector to which the property being assessed belongs. To this end its important to possess enough land market information from the whole study site. Once enough market information is available, one can proceed to define a series of “seed points” (georrefered) around which market values are expected to behave homogeneously.

1.2. *Stage 2: homogeneous zones definition*

From the “seed points” identified during Stage 1, homogeneous zones (HZs) are defined: zones in which land market is expected to behave homogeneously. Market homogeneity in this context means that within a HZ, the coefficient that modify the values taken by variables that have an effect on land value formation remain constant. It could be said then, that HZs definition is no more than finding the geographical areas where coefficients affecting the variables that best explain value formation are truly constant. In this sense, HZs are also known as

[*] Corresponding author.
E-mail addresses: frreyes@utpl.edu.ec (F. Reyes-Bueno), mgarcia@utpl.edu.ec (J.M. García-Samaniego), asanchez2@utpl.edu.ec (A. Sánchez-Rodríguez).

https://doi.org/10.1016/j.landusepol.2018.08.012
Received 4 April 2018; Received in revised form 6 August 2018; Accepted 6 August 2018
0264-8377/ © 2018 Elsevier Ltd. All rights reserved.

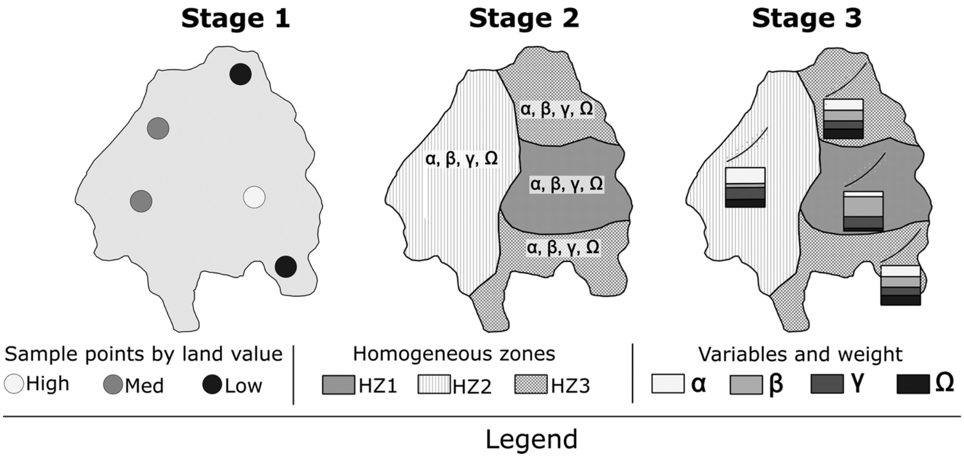

Fig. 1. The process of land value predicting models generation. **Stage 1**: the territory under study is sampled at points of known land value (light grey circles represent high land values, grey circles represent average land values and black circles represent low land values). **Stage 2**: as an example, four homogeneous zones (HZs) are identified (solid black lines) according to the behaviour of spatial variables that most influence land value formation. **Stage 3**: Land formation models are generated and the relative weight of each variable in the resulting equation determined (represented as the height of horizontal bars) within each HZ.

market segments or submarkets. When they are properly defined, HZs can be instrumental for estimating land values and for prioritizing the most important variables for value formation at each HZ (Lozano-Gracia and Anselin, 2012).

Currently in Ecuador, HZs are being defined in a complete subjective fashion. In the rural sector for instance, HZs definition is based on variables related to land use capacity such as slope, soil texture, effective depth, stoniness and drainage. Other variables commonly used for HZs definition in rural Ecuador are climatic ones e.g. precipitation, hydric deficit and temperature (Dinac et al., 1989; Magap, 2008) which are most of the time employed with redundancy. To date, there is no analysis that effectively demonstrates a decrease in heterogeneity during land value estimation thanks to the use of land use capacity or climatic variables. In the present work, we will use a plethora of mathematical approaches for HZs definition in an unsupervised fashion i.e. data-driven detection. The use of unsupervised methods is done to avoid any source of subjectivity during HZs definition i.e. prioritizing certain variables over the other (see further).

1.3. Stage 3: land value predicting models generation

It is important to note that at each HZ, land has a “base value” which is determined by the magnitude of several coefficients, each affecting a variable that resulted important for value formation (the set of prioritized variables at each HZ during Stage 2). In a perfectly modelled land market, the land base value should change from one HZ to the other as the coefficients (and the variables) change reflecting a locally adapted model. However, in the Ecuadorian context, the determination of the base value associated with a given HZ is so unreal that the weight assigned to the value-forming variables is the same among all the zones identified in a territory (Magap, 2008).

There is no universally accepted method or technique for the identification of HZs (their surface, limits, etc.) and to model the process of land value formation at each of the HZs (Kennedy et al., 1997). Among the techniques that have been used to this end, are: the geographic weighted regression (Hayles, 2006; Manganelli et al., 2014), the cluster analysis (Hayles, 2006; Kennedy et al., 1997), the main component analysis (Kennedy et al., 1997), and CART decision tree (Valenti et al., 2015).

By the end of Stage 3, we should obtain a model for the prediction of the land value. Such model is mainly based on the weights assigned to the coefficients as to modify the value-forming variables in a way that best reproduce the value changes that occur among all HZs in a given territory. An ideal model should on the one hand comprise only the variables that best explain the variance in the input data. On the other hand, it should give information on the weight each of these variables has on land value formation, which could in fact be different across HZ. One of the most widely used techniques for obtaining land value models is the multiple regression analysis (MRA) (Buurman, 2003; Elad et al., 1994; Hayles, 2006), although its application at large scales e.g. a canton, is not appropriate (Kauko and d’Amato, 2009; Mora-Esperanza, 2004). The former is because MRA is not able to properly capture the spatial (non-linear) dependence that the land value has on the land market dynamics that occur in a large territory.

As we have seen so far, the application of methodologies such as MRA for the generation of land value predicting models are based on the pre-definition of the HZs from which a final model is constructed. However, there are alternative techniques that could simultaneously segment the market and model the relationships of the variables impacting value formation (combining stage 2 and 3 into a single process). The main advantage of simultaneously defining HZs and variable weights is that the models (one per each HZ) are generated from the input data as a whole which maximizes the amount of variation finally explained by the models. That is, HZs definition becomes a data-driven process within a single dataset. Resulting models could then be used to identify the influence exerted by the explanatory variables on land value in each homogeneous zone (Clifton and Spurlock, 1983).

However, the techniques that would allow the simultaneous execution of Stages 2 and 3 have been poorly used in the field of cadastral assessment in a systematic way. Among such techniques, decision trees (DT), which are very advantageous for land markets, allow the model to adapt to local characteristics that condition it. There are several DT algorithms, including Model Tree (MT) and Multivariate Adaptive Regression Splines (MARS), which generate subsets of values showing small variations among them and generates regression functions for each subset (Wang and Witten, 1996). These techniques can face problems of classification and regression, are easy to interpret, and are of great help to analyze linear and nonlinear relationships between the dependent variable and the independent ones (Fan et al., 2006). The Model Tree technique was applied by (Acciani et al., 2008) to model the price of 109 vineyard properties in Southern Italy, obtaining more satisfactory results than with MRA on the same dataset.

Taking into account how little explored the DT method has been for cadastral valuation, in the present study we seek to answer the following questions: is the DT method suitable for the rural cadastral valuation? Can the variable land capacity improve the predictive power of land value predicting models? Have the variables affecting value formation the same weight across HZs? To answer such questions, the present study was carried out in the parish of Vilcabamba (Loja province, Ecuador). Due to the accelerated process of land transfer that has experienced Vilcabamba (Reyes-Bueno et al., 2016), we were able to have enough samples of rural properties to model land market in this territory. Four techniques for the generation of land value predicting models were compared: Linear regression, M5P model tree, M5P model tree with Bagging, and Multivariate Adaptive Regression Splines - MARS. The results show that model trees outperform all other methods and allow generating market segments that in turn decrease the prediction error of the land value.

2. Materials and methods

2.1. Land market transactions dataset

Vilcabamba parish is located in the Loja province, southern Ecuador and has surface of approximately 156 km^2. Rural Vilcabamba has undergone a strong process of land fragmentation and a concomitant very active land market. During the period 2003–2009, the 78% of all rural land transactions involved properties of less than one hectare (Reyes-Bueno et al., 2016). Land market behaviour in Vilcabamba responds to a transition from a pure agricultural territory to a touristic one at which many second-homes are being built for leisure. We considered 2003 as the starting point of our data universe mostly because in this year land purchase-sale prices stabilized after of the macro-inflation that hit the Ecuadorian economy during early 2000s (Jiménez, 2005). The dataset comprises 410 purchase-sale transactions (2003–2009) of land plots that are located in the rural sector of the Vilcabamba parish.

2.2. Pre-selection of rural land value explanatory variables

To model the process of rural land value formation, an initial set of 29 variables with potential predictor value were selected on the basis of a bibliographic search (Buurman, 2003; Elad et al., 1994; Hayles, 2006; Kostov et al., 2008; Perry and Robison, 2001; Tsoodle et al., 2006; Vandeveer et al., 2001). Once defined, a survey was applied to settlers of rural Vilcabamba to validate the applicability of the selected 29 variables as well as to identify new potential predictor variables. Out of the survey, three new predictor variables were identified and that resulted were very helpful to eliminated outliers:

- It was evident from the results of the survey that, when the buyer has a different nationality to the Ecuadorian one, the final negotiation price is higher that if an Ecuadorian buyer would buy the same land plot. As consequence, we deleted all records that involved foreign buyers.
- It was also evident that, when there was a friendship or consanguinity link with the buyer, the final negotiation price was lower in a comparison to the normal price.
- Finally, the formalization of the property was identified as a variable that affects the value of the land. Land without deeds or property rights were sold at less than half their normal price.

After the debugging process, a total of 132 transactions (out of the 410 original ones), were kept in the final dataset.

Prior to variable selection, an adjustment to the final transaction value of the 132 records in our dataset was made using the Consumer Price Index (CPI) (available at http://www.ecuadorencifras.gob.ec/indice-de-precios-al-consumidor/). The CPI index allowed us to calculate the nominal value of transactions by considering the inflation in Ecuador as of July 2009. The list of selected variables is shown in Table 1. In the case of the variables measuring a sort of distance *e.g.* from the plot in question to populated centers or roads, a unit was added to the actual variable value to calculate their logarithm since there were values of zero in some cases. Formula 1 was applied to normalize continuous variables between 0 and 1.

$$a_i = [v_i - min(v_i \ldots n)]/[max(v_i \ldots n) - min(v_i \ldots n)] \tag{1}$$

where a_i is the i-th normalized value; v_i is i-th original value prior normalization; $max(vi \ldots n)$ and $min(vi \ldots n)$ represent the minimum and maximum value respectively that the variable in question takes in the entire dataset.

Finally, a Spearman correlation analysis was performed to detect highly correlated variables from the entire dataset. As result, a total of 11 with a correlation higher than 0.6 were discarded resulting in a definitive set of 18 variables (Table 1).

Table 1
Final set of potential predictor variables of land value.

Variable	Type	Description
RS	Discreet	Road surface (asphalt, concrete, composite pavement, gravel)
I_d_ce	Continuous	Distance index to Vilcabamba parish economic center (distance by road / Euclidian distance)
LUC	Discreet	Land use capacity (crop, permanent crop, pasture, forest farming, non-farming forest)
Crops	Continuous	Proportion of annual or semi-perennial crop
CP	Continuous	Percentage of permanent crop
Grass	Continuous	Grass proportion
Forest	Continuous	Forest ratio
Mato	Continuous	Scrub proportion
Habitat	Continuous	Land to building ratio
Slope	Discreet	Slope of the terrain (weak, soft, moderate, strong, very strong, abrupt)
Irrigation	Continuous	Proportion of land with irrigation
Water	Binary	Access to clean water
Electricity	Binary	Access to electricity
Ln_surface	Continuous	Natural logarithm of the surface (m2)
Ln_d_roads + 1	Continuous	Natural logarithm of distance (m) to roads
Ln_t_ce	Continuous	Natural logarithm of the time (in minutes) needed to access the nearest economic center
Ln_c_pob + 1	Continuous	Natural logarithm of the road distance (in meters) separating the plot in question from the nearest town
Ln_V_Actual	Continuous	Natural logarithm of the market value ($/m2) of the plot in question

A second dataset (LUC- dataset) was generated from the original one (LUC + dataset) by just removing the variable land use capacity (LUC, Table 1) from it. This was done to answer the question of whether LUC can improve the predictive power of land value predicting models.

2.3. Mass appraisal models

A total of four statistical techniques were evaluated for their applicability to mass appraisal each of which are described below:

Linear regression is a statistical method for studying the linear relationship between a dependent variable and a single or multiple independent variables. It is the most common and orthodox technique used for mass appraisal. In the present work the following multiple regression equation was used:

$$Y = a_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n + \varepsilon \tag{2}$$

where a_0 is the regression constant; $b_1 \ldots b_n$ are the regression coefficients and ε is the error term

M5P is a model trees algorithm developed by Quinlan to predict continuous variables for regression (Quinlan, 1992). There are three major steps when applying the M5P methodology: (1) tree construction; (2) tree pruning; and (3) tree smoothing. The formula that the M5P algorithm uses for tree generation:

$$SDR = sd(T) - (SUM\left(\frac{|T_i|}{|T|}\right) * sd(T_i)) \tag{3}$$

where *SDR* is the reduction of the expected error; *sd* is a vector in which each element corresponds to the standard deviation of all values in the dataset for a given variable; *T* represents the entire dataset; T_i are the resulting subsets of both values and variables.

To generate a more robust model *i.e.* with reduced variance and to avoid the presence of noise in the data during model generation (Witten and Frank, 2005), the M5P technology was applied together with the Bagging or Bootstrap aggregating technique. This technique is a machine learning ensemble meta algorithm to improve the accuracy and stability of the resultant model. With the Bagging technique, training subsets are generated by randomly selecting and replacing (there may be repeated examples) a sample of *m* training examples from the original training of *n* examples (Hernández et al., 2007). When MP5 run with the Bagging algorithm, *m* models (one per each subset of training examples) are built which are later combined to generate an averaged model.

MARS, or multivariate adaptive regression splines is a decision tree algorithm introduced by Friedman, 1991 that models the relationship between a set of input (predictive) variables and dependent variables. Training data is modelled by separate piecewise linear segments (splines) of differing slopes known as basic functions. MARS generates basis functions by searching in a stepwise manner where an adaptive regression algorithm is used for selecting the knot (endpoints of the segment) locations. MARS models are constructed in a two-phase procedure. The forward phase adds functions and finds potential knots to improve the performance, resulting in an overfitted model. The backward phase involves pruning the least effective terms (Zhang et al., 2015). The main difference between M5P and MARS is that MARS is continuous at the borders of the partitioned regions, while M5P is discrete.

2.4. Land transaction data analysis

Data analysis applying i) linear regression; ii) M5P and iii) MP5 together with bagging was performed in Weka (Hall et al., 2009), while iv) MARS was executed in R using the 'earth' package. A detailed explanation of the experimental design can be found in Figure S1. Due to the limited number of samples, the cross-validation technique with 10 iterations was used in all cases to evaluate the predictive quality of the model. With this technique the original dataset was randomly divided into 10 subsets, nine of which were used for model training and the remaining one for validation purposes. Fig. 2 shows mock-examples of the application of the four statistical techniques when plot surface (ln_area) and plot accessibility (ln_t_ce) are use a predictive variables of the land value (ln_actual_value).

The multiple regression analysis was performed with the stepwise technique that incorporates only the most relevant variables into the final model ($p\text{-value} < 0.05$ associated to the F statistics). For MP5 model generation we specified that each branch should have at least 20 cases. To prevent local overfitting during the generation of the tree model, a tree pruning step was included (a detailed description of the algorithm behind the pruning step can be found in (Frank et al., 1998; Witten and Frank, 2005). In the bagging method, the number of iterations was selected by comparing the responses of the evaluators. To do this, an cross-validation experiment was first performed with 10 subsets, executed five times for each iteration (50 runs in total). Figure S2 shows that the results of the evaluators vary a lot at the beginning, although between iterations 19 to 25 it gets different results with lower prediction error. Note that the prediction error begins to stabilize from iteration 17, which was finally used to generate the models. The MARS method was applied with a backward pruning technique. The maximum number of terms in the pruned model (*nprune*) was set at 20.

2.5. Performance evaluation

Assessing the accuracy of land value estimation achieved by the four statistical techniques applied in the present study was at the center of our attention. In this direction, 'goodness-of-fit' tests (Witten and Frank, 2005) were performed: correlation coefficient, mean absolute error, relative error of the absolute values. We also calculated the magnitude of the relative error obtained by each methodology to apply the Wilcoxon test which allowed us to identify significant differences between models and databases (LUC + and LUC- datasets). We finally applied the Moran test to calculate if there was spatial autocorrelation among the predictor variables that end up in each model. All formulas used in each case can be found in the supplementary material (Section 1.1).

3. Results

3.1. Models performance

A detailed description of the results obtained by each of the four techniques when applied to both datasets (LUC + and LUC-) can be found in Supplementary Material (Section 2.1). According to the three performance metrics calculated to compared across models, those generated with MARS showed a better performance compared to the rest (Table 2).

That MARS outperformed Linear Regression and M5P (both with and without bagging) is corroborated when calculating estimation errors (Table S4). MARS and M5P predicted land values with an estimation error of less than the 30% for a large fraction of test data points (67% of the test data points in MARS 44% and 44% in M5P). In contrast, the corresponding number for Linear Regression was only 29% of the test data points. In addition, 51% of the estimations made by Linear Regression have an estimation error higher than 50% (this was only the 27% for the estimations made by MARS and 39% of those obtained by M5P). Once the Wilcoxon rank test was applied to the generated residues (Table 3), it makes evident that the predictive capacity of the linear regression model is significantly lower than in the other models analyzed. Also, the models generated by MARS show fewer errors in the estimates, being a significant difference with respect to all the other models.

Table S4 shows that when comparing the two databases (LUC + vs LUC-), there is a slight increase in the percentage of test data points with an estimation error higher than 50% (except in M5P) in the LUC- dataset. There is also a decrease in the percentage of test data points with an estimation error lower than 30% for the models obtained by Linear Regression and M5P with Bagging. However, according to the Wilcoxon rank test applied to the residuals generated by each model on both datasets, these differences are not statistically significant (Table 4).

To look for signs of spatial autocorrelation on the estimated land values, the model generated by MARS on the LUC- dataset was taken as a basis. Results obtained by calculating the Moran index, allow us to say that the pattern that estimated land value show is not significantly different than randomness (Fig. 3). These results are supported by the calculated Z-value of 0.17 (which is within the range of -1.96 to 1.96) and the associated p-value of 0.88 for the acceptance of the null hypothesis.

4. Discussion

Our results show that locally adaptive nonparametric techniques (M5P, M5P with Bagging and MARS) generate significantly better models than those obtained by global methods such as Linear Regression when modeling the value of rural land. It has been shown that techniques able to generate models that adjust to local variations (e.g. Mobile Window Regression or Geographically Weighted Regression) have a better performance compared to global techniques susceptible to local variations when modeling the value of the land (Buyong et al., 2008; Miller and Jiawei, 2009). Among the techniques used in our analysis, MARS generated the model with the least relative error in the predictions. MARS proposes the generation of an equation in which the weight of the variables of importance for the formation of the value varies according to ranges or sections of the domain of those variables. In this way, a continuous space is generated at the borders of the partitioned dataset (Fig. 2), trying to adapt to the existing interaction between the value-forming variables (Thomaes et al., 2008).

In the case of Vilcabamba, the model generated by MARS comprised some accessibility variables (index of distance to the economic center, time of access to the economic center and distance to population centers), as well as variables of land use (pasture), area, slope and access to irrigation. The MARS model depicted at least two ranges of the domain

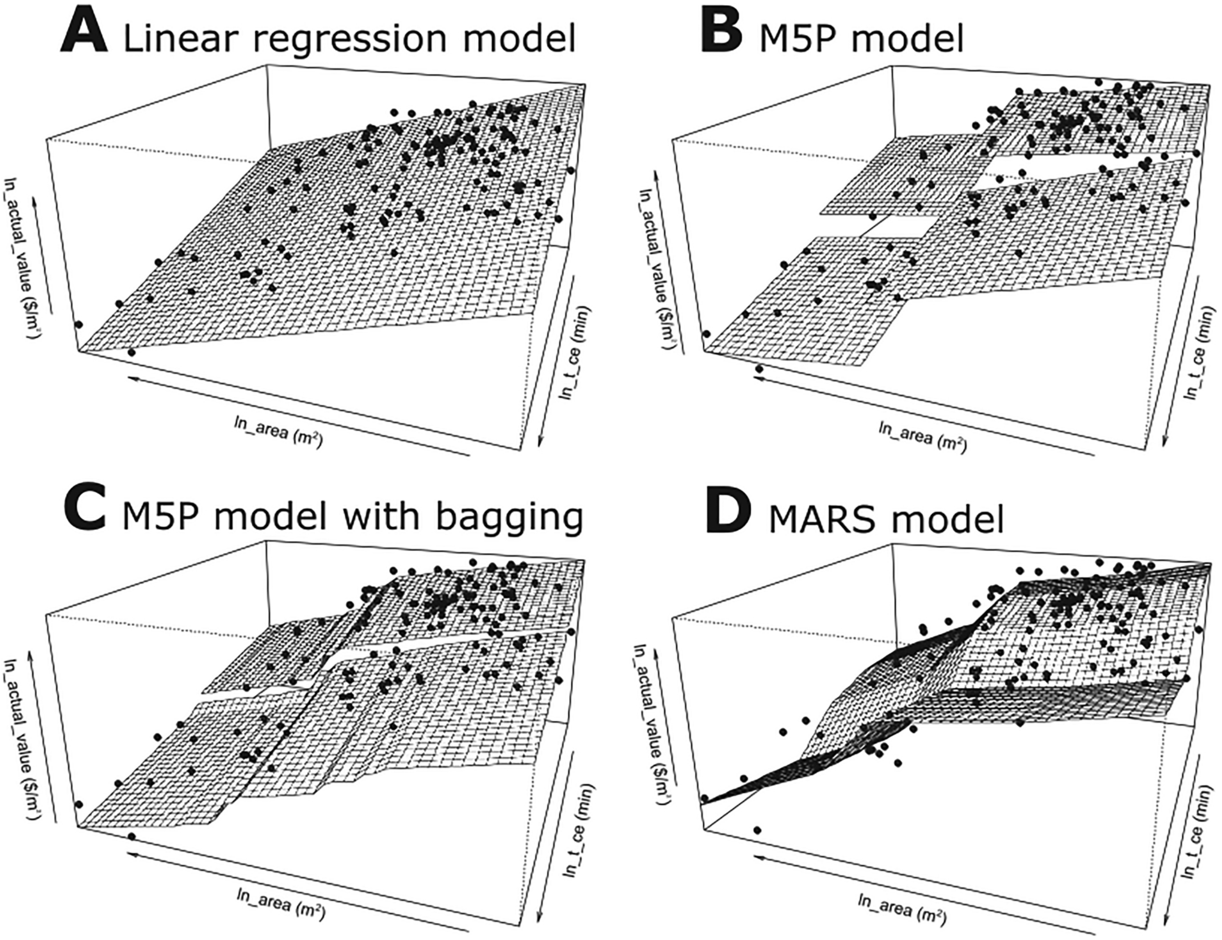

Fig. 2. Graphical representation of land value predicting models (wireframe meshes) obtained by four statistical techniques (panels A–C). Plot surface (ln_area) and plot accessibility (ln_t_ce) were used in each case as predictive variables of the land value (ln_actual_value). Black circles represent actual data points from the land transactions dataset. MP5 (panel B) and MP5 with bagging (panel C) produce more than one wireframe mesh (predictive models) each modeling a different portion of the data. Linear regression (panel A) and MARS (panel D) produce a single wireframe mesh (model) that describes the entire dataset as a whole.

Table 2
Performance of models generated by four statistical techniques on two datasets (LUC + and LUC-). LR: multiple linear regression. MP5: M5P model tree algorithm. MP5-Bagging: M5P model tree algorithm together with the bagging technique. MARS: multivariate adaptive regression splines. CC: correlation coefficient. MAE: mean absolute error. RAE: relative absolute error. Details on how each metric was calculated can be found in supplementary material.

Performance metric	Statistical technique (on LUC+)				Statistical technique (on LUC-)			
	LR	M5P	M5P-Bagging	MARS	LR	M5P	M5P-Bagging	MARS
CC	0.88	0.92	0.93	0.96	0.88	0.92	0.93	0.96
MAE	0.62	0.49	0.48	0.38	0.62	0.48	0.49	0.38
RAE	48.93	38.56	37.95	30.11	48.93	38.18	38.46	30.10

Table 3
Results of the Wilcoxon signed rank test applied to the residues associated to the land value predictions generated by each model on the LUC + dataset. Numbers within each cell represent the p-value associated with the null hypothesis being true: no difference between the residuals generated by a given pair of models. LR: linear regression; M5P: M5P model tree; M5P-B: M5P model tree with bagging; MARS: multivariate adaptive regression splines.

	LR	M5P	M5P-B	MARS
LR		0.00	0.00	0.00
M5P			0.57	0.06
M5P-B				0.00
MARS				

Table 4
Results of the Wilcoxon signed rank test applied to the residues generated by each model during land value estimation on the LUC + and LUC- datasets. Numbers within each square represent the p-value associated with the null hypothesis being true: no difference between the residuals generated by the same model on both datasets. LR: linear regression; M5P: M5P model tree; M5P-B: M5P model tree with bagging; MARS: multivariate adaptive regression splines.

		LUC+			
		LR	M5P	M5P-B	MARS
LUC-	LR	0.62			
	M5P		0.89		
	M5P-B			0.08	
	MARS				0.95

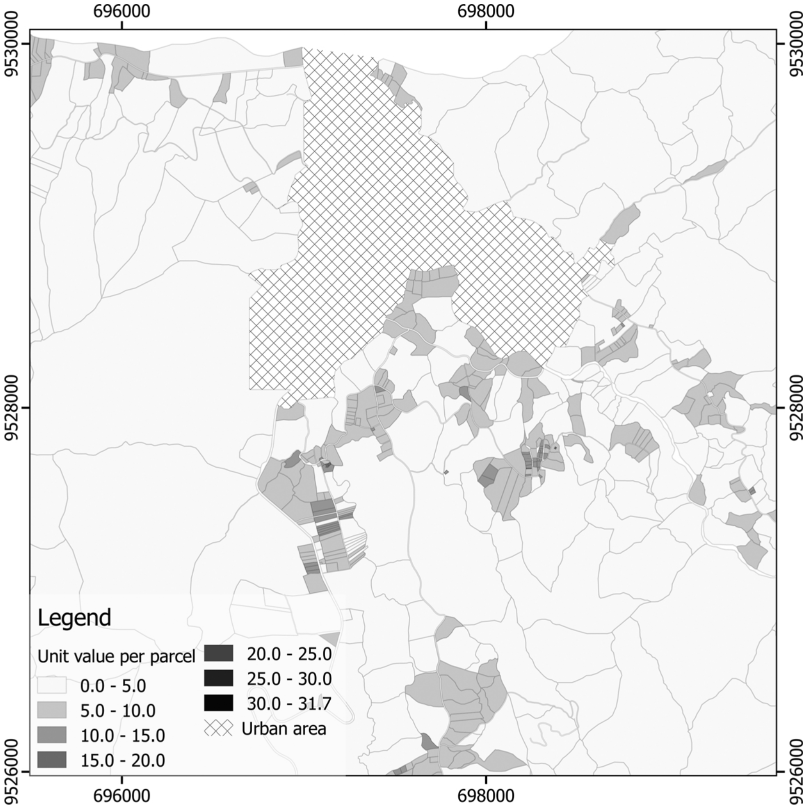

Fig. 3. Classes of land values spatialized over rural Vilcabamba. Estimates were obtained by applying MARS to the LUC- dataset. Geographic coordinates are also shown.

of those variables each with different variable coefficients. The integration of the different ranges of the used variables could then be used to generate market segments or HZs in rural Vilcabamba. The main advantage of MARS lies in its capacity to produce simple, easier-to-interpret models that result from modeling an intricate multi-dimensional web of variable relationships (Zhang et al., 2015). MARS main difference with respect to M5P is that the latter generates tangible submarkets, that is, it clearly delimits discrete areas at which a best fitting linear equation can be drawn (Fig. 2). The multi discrete models generated by M5P have therefore an overall lower predictive capacity than the single continuous model generated by MARS. Its elevated performance together with its modest computational demands position MARS as a very attractive alternative for mass appraisal tasks.

Land use capacity, which classifies the soils according to their productive capacity, was not of significant importance when evaluating the predictive strength of rural land valuation models. The Vilcabamba case shows that “land use capacity as a decisive variable for land value formation” is not a rule of thumb. Although agrological variables have traditionally been used to determine the value of rural land, due to its direct relationship with the income that the land could generate, the contribution of land to the production process has decreased. More importantly is to note that land uses in the rural sector has diversified towards several purposes (Caballer, 2002; Moya and García-Rodrigo, 2001). In Vilcabamba, a large number of land plots are being sold for recreation purposes in a scenario that some authors say have reached rural gentrification (Reyes-Bueno et al., 2016). Rural gentrification with no doubt modifies the weight of variables important to HZs detection and land value formation compared to the classical perspective of rural areas.

In several countries, a characteristic of mass appraisal for cadastral purposes is the application of adjustment factors (variable weights) to adjust the land value of a given lot. The former is done by increasing or decreasing the land value depending on the characteristics such as slope, land use capacity, etc. Although, the adjustment factors vary according to the characteristics of the land of a given plot, the same adjustments (weights) are applied across all HZs defined in a certain territory (Boo and González, 2009). This is exactly the case of Ecuador. The results of the present work show that weights assigned to predictor variables for land value formation varies across HZs. This highlights the importance to reduce subjectivity during HZs definition since this will have a significant impact on the estimation of the weights assigned to predictor variables within each HZs and finally on obtaining a model that effectively reflects the value formation dynamics in the area under scrutiny.

Cadastral appraisal is the basis for the determination of various taxes (land tax, sale tax, contribution for improvements, capital gain, etc.) but also for the definition of the value of land plots affected by expropriation. For these reasons, cadastral appraisal must follow the principles of equality and proportionality (COOTAD 2010). Therefore, is fundamental to have valuation models that are sufficiently strong for the determination of cadastral values as equidistant and as close as possible to those practiced in the market (Peña-Medina, 2016). In order to fulfill these principles, one must start from the definition of HZs (market segments) that reflect a similar market behavior, and therefore eliminating subjectivity in its definition. Decision trees as M5P and MARS prove to be applicable techniques to cadastral appraisal since they allow defining HZs while simultaneously establishing the weights of predictor variables for land value formation.

Declarations of interest

None

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.landusepol.2018.08.012.

References

Acciani, C., Fucilli, V., Sardaro, R., 2008. Model tree: an application in real estate appraisal. Proceedings of the 19th Seminar of European Association of Agricultural Economists Retrieved from. http://ideas.repec.org/p/ags/eaa109/44853.html.

Baumane, V., 2010. Cadastral valuation models. Economic Science for Rural Development Conference Proceedings 22, 68–75. Retrieved from. http://search.ebscohost.com/login.aspx?direct = true&db = a9h&AN = 77483928&lang = es&site = eds-live.

Boo, I.D., González, F.B., 2009. Modelos De Valoración Inmobiliaria En Iberoamérica. Retrieved from. https://dialnet.unirioja.es/servlet/libro?codigo = 400613.

Buurman, J., 2003. Rural Land Markets a Spatial Explanatory Model. Retrieved from. Vrije Universiteit, Amsterdam. http://dare.ubvu.vu.nl/bitstream/1871/10537/1/6001.pdf.

Buyong, T., Ismail, S., Sipan, I., Hashim, M.-G., Azhar, M.-F., 2008. Moving window regression (MWR) in mass appraisal for property rating. Symposium (IRERS): Benchmarking, Innovating and Sustaining Real Estate Market Dynamics 9 Retrieved from. http://eprints.utm.my/5684/.

Caballer, V., 2002. Nuevas tendencias en la valoración territorial. CT/Catastro 45, 11. Retrieved from. http://www.catastro.meh.es/documentos/publicaciones/ct/ct45/11.pdf.

Clifton, I.D., Spurlock, S.R., 1983. Analysis of variations in farm real estate prices over homogeneous market areas in the southeast. South. J. Agric. Econ. 15 (01) Retrieved from. http://econpapers.repec.org/RePEc:ags:sojoae:30227.

Dinac, D.N., de, Ay.C., del, E., 1989. Reglamento De Avalúos De Predios Rurales. Registro Oficial No 913.

Ecuador, 2010. October). Código Orgánico de Organización Territorial, Autonomía y Descentralización. COOTAD. Retrieved from. http://www.mcpolitica.gov.ec/mp3/COOTAD.pdf.

Elad, R.L., Clifton, D., Epperson, J.E., 1994. Hedonic estimation applied to the farmland market in Georgia. Journal of Applied Agricultural Economics 26 (02), 351–366. Retrieved from. http://econpapers.repec.org/RePEc:ags:joaaec:15179.

Fan, G.-Z., Ong, S.E., Koh, H.C., 2006. Determinants of house price: a decision tree approach. Urban Stud. 43 (12), 2301–2315. https://doi.org/10.1080/00420980600990928.

Frank, E., Wang, Y., Inglis, S., Holmes, G., Witten, I., 1998. Using model trees for classification. Mach. Learn. 32 (01), 63–76. https://doi.org/10.1023/A:1007421302149.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., 2009. The WEKA data mining software: an update. SIGKDD Explorations 11 (1), 9. Retrieved from. http://www.kdd.org/explorations/issues/11-1-2009-07/p2V11n1.pdf.

Hayles, K., 2006. A Property Valuation Model for Rural Victoria. School of Mathematical and Geospatial Science. Retrieved from. RMIT University. http://researchbank.rmit.edu.au/eserv/rmit:6265/Hayles.pdf.

Hernández, J., Ramírez, M., Ferri, C., 2007. Introducción a la Minería de datos. PEARSON EDUCACIÓN, S.A., Madrid.

Iaao, I. A. of A. O, 2011. Standard on Mass Appraisal of Real Property. Retrieved from. pp. 21. http://www.iaao.org/uploads/StandardOnMassAppraisal.pdf.

Jiménez, R., 2005. Compendio de la Dolarización en Ecuador. Facultad de Contaduría, Administración E Informática de La UNAM. pp. 80. Retrieved from. http://yumka.com/docs/dolarizacion-ecuador.pdf.

Kauko, T., d'Amato, M., 2009. Automated valuation methods, empirical modelling of value. and Systems for Market Analysis 305–319. https://doi.org/10.1002/9781444301021. ch14.

Kennedy, G., Henning, S., Vandeveer, L., Dai, M., 1997. Multivariate procedures for identifying rural land submarkets. J. Appl. Agric. Econ. 29 (02) Retrieved from. http://econpapers.repec.org/RePEc:ags:joaaec:15066.

Kostov, P., Patton, M., McErlean, S., 2008. Nonparametric analysis of the influence of buyers' characteristics and personal relationships on agricultural land prices. Agribusiness 24 (2), 161–176. https://doi.org/10.1002/agr.20152.

Lozano-Gracia, N., Anselin, L., 2012. Is the price right?: assessing estimates of cadastral values for Bogotá, Colombia*. Reg. Sci. Policy Pract. 4 (4), 495–508. https://doi.org/10.1111/j.1757-7802.2012.01062.x/full. Retrieved from.

Magap, 2008. Metodología de Valoración de Tierras Rurales 2008. PROPUESTA (p. 368). Retrieved from http://app.sni.gob.ec/sni-link/sni/Portal SNI 2014/GEOGRAFICA/Conage/Documentos/Metodologias/Metodologia_valoracion_tierras_rp.pdf. .

Manganelli, B., Pontrandolfi, P., Azzato, A., Murgante, B., 2014. Using geographically weighted regression for housing market segmentation. Int. J. Bus. Intell. Data Min. 9 (2), 161–177. https://doi.org/10.1504/IJBIDM.2014.065100.

Miller, H., Jiawei, H., 2009. Geographic Data Mining and Knowledge Discovery, 2 ed. C. R. C. Press, Minneapolis, pp. 486.

Mora-Esperanza, J., 2004. La inteligencia artificial aplicada a la valoración de inmuebles, vol. 50. Un ejemplo para valorar Madrid, CT Catastro, pp. 51–67. Retrieved from. http://www.catastro.meh.es/esp/publicaciones/ct/ct50/-2E.pdf.

Moya, M., García-Rodrigo, Á., 2001. Catastro, Valoración y Tributación Inmobiliaria Rústica.

Peña-Medina, S., 2016. El impuesto predial en Ciudad Juárez desde una perspectiva de equidad/ Property tax in Ciudad Juarez from an equity perspective. Economía, Sociedad y Territorio 16 (51), 519–542. Retrieved from. http://search.proquest.com/openview/2bc9c69df0a1b6b9535e887e9ca2163a/1?pq-origsite = gscholar&cbl = 2026683.

Perry, G.M., Robison, L., 2001. Evaluating the influence of personal relationships on land sale prices: a case study in Oregon. Land Econ. 77 (3), 385–398. Retrieved from. http://econpapers.repec.org/.

Quinlan, J.R., 1992. Learning With Continuous Classes. World Scientific, pp. 343–348.

Reyes-Bueno, F., Tubío Sánchez, J., Gracía Samaniego, J., Miranda Barrós, D., Crecente Maseda, R., Sánchez-Rodríguez, A., 2016. Factors influencing land fractioning in the context of land market deregulation in Ecuador. Land use policy 52, 144–150. https://doi.org/10.1016/j.landusepol.2015.12.021.

Thomaes, A., Kervyn, T., Maes, D., 2008. Applying species distribution modelling for the conservation of the threatened saproxylic stag beetle (Lucanus cervus). Biol. Conserv. 141 (5), 1400–1410. https://doi.org/10.1016/j.biocon.2008.03.018.

Tsoodle, L.J., Golden, B.B., Featherstone, A.M., 2006. Factors influencing kansas agricultural farm land values. Land Econ. 82 (1), 124–139. https://doi.org/10.3368/le.82.1.124.

Valenti, A., Giuffrida, S., Linguanti, F., et al., 2015. Decision trees analysis in a Low Tension Real Estate Market: the case of Troina (Italy). In: In: Gervasi, O., Murgante, B., Misra, S., Gavrilova, M.L., Rocha, A.M.A.C., Torre, C. (Eds.), Computational Science and Its Applications – ICCSA 2015 SE - 17, vol. 9157. Springer International Publishing, pp. 237–252. https://doi.org/10.1007/978-3-319-21470-2_17.

Vandeveer, L., Henning, S., Niu, H., Kennedy, G., 2001. Rural Land Values at the Urban Fringe, (Spring). Retrieved from. http://www.lsuagcenter.com/en/communications/.

Wang, Y., Witten, I., 1996. Induction of Model Trees for Predicting Continuous Classes. Working Paper 96/23, 13. Retrieved from. http://www.cs.waikato.ac.nz/pubs/wp/1996/uow-cs-wp-1996-23.pdf.

Witten, I., Frank, E., 2005. Data mining: practical machine learning tools and techniques. Morgan Kaufmann Series in Data Management Systems, second edition. Morgan Kaufmann.

Zhang, W., Goh, A.T.C., Zhang, Y., 2015. Multivariate adaptive regression splines application for multivariate geotechnical problems with big data. Geotech. Geol. Eng. 34 (1), 193–204. https://doi.org/10.1007/s10706-015-9938-9.