AESTIMUM 58, Giugno 2011: 27-45

Claudio Acciani
Vincenzo Fucilli
Ruggiero Sardaro

Department of Agro-Environmental and Territorial Sciences. University of Bari "Aldo Moro"
e-mail: claudio.acciani@agr.uniba.it
v.fucilli@agr.uniba.it
ruggiero.sardaro@agr.uniba.it

Data mining in real estate appraisal: a model tree and multivariate adaptive regression spline approach[1]

In this paper we adopt two exploratory modelling techniques: Model Trees and Multivariate Adaptive Regression Splines. The objective is the building of two sale price prediction models in order to highlight possible market segments not detectable *a priori*. We show how these novel procedures can help to understand complex patterns and interactions among predictors in real estate appraisal.

Keyword: *real estate appraisal; hedonic price method; data mining; model trees; multivariate adaptive regression splines*

JEL classification: C14, C52, R33.

1. Introduction

The classic aspect of real estate appraisal is the assessment of the most likely market value (Grillenzoni and Grittani 1994). This objective can be attained through Multivariate Regression (MR), which allows measuring the influence of good amenities on its sale value and understanding the logic adopted by dealers in trading.

Obviously, MR is concerned with estimating the *mean* value of the dependent variable on the basis of the known values of the explanatory variables (Gujarati 2009), so that it has a less predictable capacity if the analysis is carried out on heterogeneous markets, i. e. characterized by more segments.

In real estate appraisal practice, the discovery of submarkets is an important requirement for using hedonic price estimation (Acciani and Gramazio 2006; Acciani *et al*. 2008). To that end the implementation of techniques that permit *to mine* a more complex data structure that often hides in high-dimensional datasets is necessary. This goal can be achieved through data mining approaches.

Data mining is a recently developed discipline (Han and Kamber 2006) that combines statistical analysis, computer science, artificial intelligence (and connected areas like machine learning and pattern recognition) and database man-

[1] The authors thank the referees for their helpful comments and suggestions. However, the views expressed in this paper are those of the authors.

agement. It is a selection, exploration and mining process of knowledge from masses of data, through the application of particular techniques, in order to detect possible regularities, trends and associations that are not known *a priori*. In other words, data mining is the extraction of implicit, previously unknown, and potentially useful information from data. The idea is that it is possible *to mine* useful information also from unsuspected zones of the informative space that can be extended (and so generalized) to bigger data sets, in order to get a useful and clear result that allows to take a strategic decision (Witten and Frank 2005).

Undoubtedly, the success of data mining is closely connected to development, over the last few years, both of more powerful and economical hardware and software tools that put together and spread big databanks and of new procedures in informatics and statistics that are necessary to analyse them. Therefore, data gathering has become easier (Web, Data Warehouse) and the implementation of particular algorithms has allowed the use of informative heritage, overcoming the constraints concerning the heterogeneous, redundant and not structured data shape.

Data mining techniques are divided into descriptive and forecasting ones. The first describe the data set in concise and simplified way, presenting interesting general patterns and characteristics. The second ones have the objective to build models, on the basis of data owned, that can be generalized, so to predict the behaviour of new data sets.

The forecasting techniques (e.g. Model Tree, MT) use data in which the value of the dependent variable is known. This approach is named *supervised training*, because the observations are classified on the basis of all the independent variables, supervised by the presence of the dependent variable.

The descriptive techniques (e.g. clustering), instead, are of *unsupervised training*: in this case no dependent variable is specified and the observations are classified on the basis of an indistinct set of available variables (Witten and Frank 2005).

In this research, a segmentation analysis has been carried out through MT and Multivariate Adaptive Regression Splines (MARS), non-parametric techniques that don't need assumptions about the dependent variable distribution. The objective is to test and compare them on real estate market in order to highlight possible sub-samples, representative of specific market segments that are not detectable *a priori*. The segmentation analysis has been conducted on the lively market of Apulian *trulli* (UNESCO world heritage). Furthermore, the prediction accuracy of the models is compared to a standard multivariate linear regression (MLR).

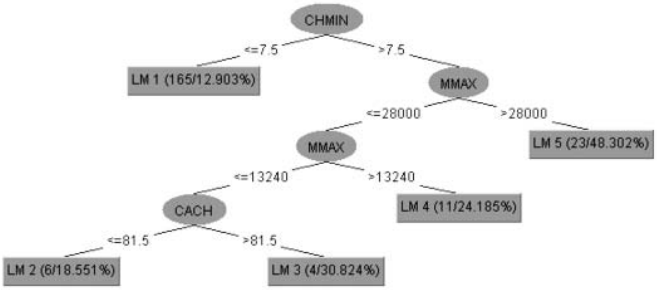
2. Model Tree

Model Tree (MT) is based on a *divide-and-conquer* approach by which is possible *to learn* from a set of instances (Witten and Frank 2005). In particular, a top-down recursive partitioning procedure is carried out, by which a data set of observations is gradually divided into subsets. This segmentation is made on the basis of a *splitting criterion* by which the independent variable (or attribute) and its *threshold value* that maximize the expected error reduction are detected. In particular, all the

possible segmentations in respect to each attribute and relative threshold values are made, choosing, finally, the best partition in terms of error reduction (Quinlan 1986; Quinlan 1992; Quinlan 1993; Witten and Frank 2005).

The output of a MT is represented by a tree–structure in which it is possible to distinguish a *root node*, *parent* and *child nodes*, *arches* (or branch) and *leaves*. Before the recursive partition, all instances of data set are contained in the root node. During the first step, through the splitting criterion, the whole instances set in the root node are divided into subsets (nodes). In the following step, a new partition is carried out for each node arisen during the first stage in order to create new nodes. In particular, a node is named "parent" with respect to the nodes that it generates, "child" with respect to the node from which it descends. Partition proceeds recursively for each node generated during the various steps until terminal nodes (leaves) are obtained. In particular, each leaf reports a linear regression model (LM) calculated on the instances that it contains. Nodes, labelled with the name of the attribute chosen for the partition, are connected between them through arches that are labelled with the threshold value of the attribute in correspondence to which the partition has been carried out (Witten and Frank 2005). Figure 1 shows an example model tree over 209 different computer configurations, adapted from Witten and Frank (2005).

Figure 1. A model tree for CPU performance data.



In MT the expected error index of splitting criterion is the *standard deviation* and the respective expected error reduction, *SDR*, is given by:

$$SDR = sd(M) - \sum_{i=1}^{s} \frac{|m_i|}{|M|} sd(m_i) \tag{1}$$

SDR is implemented into M5P partitioning algorithm (Wang and Witten 1997), an optimization of M5 algorithm (Quinlan 1992), included in the open source software WEKA[2] (Witten and Frank 2005).

[2] WEKA (Waikato Environment for Knowledge Analysis), is an open source tool developed by the University of Waikato, New Zealand, and written in Java. It is free downloadable from the web (www.cs.waikato.ac.nz/ml/weka).

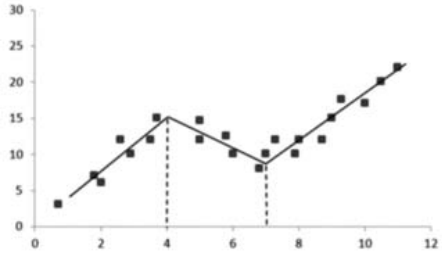
In this case, the dividing process terminates either when the standard deviation of instances that reach the leaf is less than a minimum threshold (generally 5%) of the standard deviation of the original data set or when the number of observations into the node is less than a fixed value.

3. Multivariate Adaptive Regression Splines

Multivariate Adaptive Regression Splines (MARS) is a non-parametric regression technique introduced by Friedman in 1991. It essentially detects relation between a dependent variable and a set of predictors by fitting piecewise linear regressions (Friedman 1991a, Friedman 1991b, Friedman 1991c; Friedman 1993). In particular, MARS builds flexible models by dividing the whole space of each covariate into various subsets and then defining a different regression equation for each area. In this way, separate regression slopes in distinct intervals of the predictors space are individuated (Hastie *et al*. 2009). A key concept is the notion of *knots* that are the points that bound each interval of data in which a distinct regression equation is calculated, i.e. where the behaviour of the modelled function changes.

A simple example of a piecewise linear regression is shown in figure 2, in which the slope of the regression line changes from one interval to the other as the knot points are crossed. In particular, the figure shows two knots (points with abscissa 4 and 7) that delimit three intervals in which different linear relationships are identified.

Figure 2. Example of piecewise linear regression.



In this way, the space of predictors is split into several regions in which truncated *spline functions* or *basis functions* (BFs) are fit. A truncated BF consists of a left-sided (2) and a right-sided (3) segments defined by a knot *t*:

$$b_q^-(x-t)=\left[-(x-t)\right]_+^q=\begin{cases}(t-x)^q, & \text{if } x<t\\ 0 & \text{otherwise}\end{cases} \tag{2}$$

$$b_q^+(x-t)=\left[+(x-t)\right]_+^q=\begin{cases}(x-t)^q, & \text{if } x<t\\ 0 & \text{otherwise}\end{cases} \tag{3}$$

where $b_q^-(x-t)$ and $b_q^-(x-t)$ are the BFs describing the regions to the left and the right of the knot t, q indicates the power (>0) to which the BFs are raised in order to manipulate the degree of smoothness of the resultant regression models (for example, when q=1 only simple linear BFs are considered), the subscript "+" indicates a value of zero for negative argument. The general MARS model equation is given as:

$$\hat{y}=\alpha_0+\sum_{m=1}^{M}\alpha_m B_m(x) \tag{4}$$

where $\hat{y}$ is the dependent variable predicted through the MARS model, M is the number of BFs included into the model, α_0 is the constant term, α_m is the coefficient of the mth truncated BF and $B_m(x)$ is the mth truncated BF that may be a single spline function or a product (interaction) of two or more spline functions (Friedman 1991b; Friedman 1991c).

The optimal MARS model is built by a two-stage process: a forward selection procedure followed by a backward-pruning procedure. The forward procedure starts with just the constant term in the model and then, by an iterative way, selects the best pairs of BFs that improve the global model (each pair is constituted by one left-sided and one right-sided truncated spline function). In particular, the algorithm evaluates all possible predictors, as well as all possible knot locations for each predictor, selecting the pairs that minimize a "lack of fit" criterion. Additionally, at the end of each iteration, the algorithm checks whether the introduction of an interaction improves the model. In this case, the maximum number of BFs that interact indicates the *order* of a MARS model, so that for order equal to 1 the model is additive, whereas for order equal to 2 the interactions between a maximum of 2 predictors are verified as well. The brute search continues until a researcher-defined maximum number of BFs is included into the global model. This value should be at least two to four times the size of the "truth". Thus, if previous experience suggests that a robust model has approximately 10 predictors, the maximum number of BFs should be set to at least 20 and more likely 40.

This forward stepwise selection of BFs leads to a very complex and over fitted model that has poor predictive abilities for new data. So, in the backward stage, the "lack of fit" criterion is used to evaluate the contribution of each BF to the descriptive abilities of the model and the BFs with the lowest contribution are removed one at a time.

The "lack of fit" criterion used by MARS is the generalized cross-validation (GCV) criterion, that is, the mean square error divided by a penalty dependent on the model complexity. It is given by:

$$GCV(M) = \frac{1}{n} \frac{\sum_{i=1}^{n} \left(y_i - \hat{y}\right)^2}{\left[1 - \frac{C\left(M\right)}{n}\right]^2} \quad (5)$$

where n is the number of observations in the data set, M is the number of non-constant BFs, and $C(M)$ is the cost-complexity measure of the model containing M BFs. $C(M)$ increases with number of BFs and has the purpose to penalize model complexity in order to avoid over fitting. It is defined as:

$$C\left(M\right) = M + dM \quad (6)$$

where d is a cost penalty factor for adding a BF. The higher value of d the fewer BFs included in the final model. Eventually, the selection of the optimal model is performed in a third step. The selection is based on evaluation of the predictive properties of the different models, which often are determined using cross-validation or a new independent test set. Further details on MARS modelling are given in Friedman (1991c).

4. MT or MARS?

In order to evaluate the predictive capacity of a model, its assessment on unseen data is necessary since any performance estimate based on the original data set (*training set*) is optimistic (*over fitting*). As a rule, in presence of large data sets, to obtain a realistic estimate of predictive power of a model the original sample is partitioned into two different subsets: the *training set* to fit the model and the *test set* to validate it. However, when the amount of data is limited, *k-fold cross-validation* can be used to obtain nearly unbiased estimators of prediction error. For a small data set with n observations, k-fold cross-validation randomly divides the data set into k approximately equal subsamples. Hence each subsample in turn is used for testing while the remainder k-1 subsamples are used for training. The cross-validation process is then repeated k times, with each of the k subsamples used exactly once for testing. The k results from the folds then are averaged to produce a single estimation. The advantage of this method is that all observations are used for both training and testing so that each observation is used for testing exactly once (Witten and Frank 2005). In this study a 10-fold cross-validation has been used.

4.1 Prediction performance of fitted models

In order to compare the prediction performance of the different models, the statistical indexes summarized in Table 1 have been calculated, where a_i is the ith

actual value of the dependent variable while p_i is the corresponding predicted value from the model. In particular, the *correlation coefficient* measures the statistical correlation between the estimated values p and the actual ones a of the target variable. In its formula, *cov(p,a)* is the covariance between the estimated values and the actual ones, while σ_p e σ_a are the respective standard deviations. This coefficient ranges from +1 (ideal situation of perfect direct correlation) to -1 (perfect inverse correlation), with coefficient equal to 0 in absence of correlation. Of course, negative values should not occur for reasonable prediction methods. The *root mean-squared error* is measured in the same unit of dependent variable and ranges from 0 (ideal situation) to infinity. The *mean absolute error* is equal to the average of errors without their sign, while the *root relative squared error* calculates the relative error in respect to the average of the actual values of the target variable. Finally, *relative absolute error* is similar to the root relative squared error and ranges from 0 to 1 (Witten and Frank 2005).

Table 1. Performance measures.

Performance measure	Formula
Correlation coefficient	$r = \mathrm{cov}(p,a) / \sigma_p \sigma_a$
Root mean-squared error	$RMSE = \sqrt{\sum_{i=1}^{n} (p_i - a_i)^2 \Big/ n}$
Mean absolute error	$MAE = \sum_{i=1}^{n} \lvert p_i - a_i \rvert \Big/ n$
Root relative squared error	$RRSE = \sqrt{\sum_{i=1}^{n} (p_i - a_i)^2 \Big/ \sum_{i=1}^{n} \left(a_i - \hat{a} \right)^2}$
Relative absolute error	$RAE = \sum_{i=1}^{n} \lvert p_i - a_i \rvert \Big/ \sum_{i=1}^{n} \lvert a_i - \hat{a} \rvert$

4.2 Significant test

To evaluate the presence of a significant difference among the fitted models, the *absolute residual error* (ARE) and the *magnitude of relative error* (MRE) have been calculated. They are defined as:

$$ARE_i = \lvert p_i - a_i \rvert \tag{7}$$

$$MRE_i = \frac{\lvert p_i - a_i \rvert}{p_i} \tag{8}$$

Hence, the Wilcoxon signed-rank test for correlated samples has been used, a non-parametric test which checks the difference between means when the population cannot be assumed to be normally distributed (Wilcoxon 1945). Also, it deals with the signs and ranks of the values and not with their magnitude, so that it is not influenced by outliers. In particular, the test calculates the differences between the paired observations, ranks them from the smallest to largest by absolute value, gives the sign of each difference to the corresponding rank, sum separately ranks having the plus sign and ranks having the minus sign, obtaining the two values W_+ and W_-, respectively[3]. For small datasets ($n \leq 20$) the calculated value of the test statistic W (W_+ or W_-) must be compared to the tabulated critical value of an exact sampling distribution (defined by n and the level of significance α). So, for a two-tail test, if the observed value of W equals or is greater than the upper critical value or is equal to or less than the lower critical value, the null hypothesis is rejected. On the contrary, for $n > 20$, the test statistic W (W_+ or W_-) is approximately normally distributed with mean μ_W and standard deviation σ_W equal to:

$$\mu_W = \frac{n(n+1)}{4} \tag{9}$$

$$\sigma W = \sqrt{\frac{n(n+1)(2n+1)}{24}} \tag{10}$$

respectively[4]. Thus, the Z-ratio is defined as:

$$Z = \frac{|W - \mu_W| - 0.5}{\sigma_W} \tag{11}$$

where 0.5 is the correction for continuity. For a two-tail test and for a particular level of significance α, if the computed $|Z\text{-ratio}|$ is greater than or equal to the critical value, then the null hypothesis H_0 ($\mu_1 = \mu_2$) is rejected, where μ_1 and μ_2 are two population means of matched pairs (Conover 1999). In this study, the null hypotheses using ARE and MRE are:

H_0: $ARE_{MARS} = ARE_X$
H_0: $MRE_{MARS} = MRE_X$

where X denotes MT or MLR. The level of significance $\alpha$for the null hypothesis rejection has been set equal to 0.05.

[3] For tied observations, respective ranks are added together and divided by the number of ties. Also, the cases with the zero difference are removed.

[4] The cutoff varies among authors, so that some put it lower (10 or 15) or higher (25).

5. The data set

The aim of the research is to experiment the application of MT and MARS for the detection of real estate submarkets and the assessment of most likely market value. The transactions of farms in which are present *trulli*[5] (Figure 3) have been considered. In particular, these typical constructions are mostly concentrated in the *Valle d'Itria* area, which spreads over the Provinces of Bari, Brindisi and Taranto and coincides with the south-east area of the *Murgia* plateau.

Constructed without cement or mortar as dwellings or storehouses by local farmers, after a dereliction period, nowadays trulli are popular among Italian and northern Europe tourists (above all from England and Germany) and entrepreneurs which buy and restore them as holiday home, bed and breakfast or holiday farmhouse. Since 1996, trulli are in the United Nations Educational, Scientific and Cultural Organization (UNESCO) world heritage list.

The choice of this property typology is connected to its peculiarity. Because of the high touristic attraction of the area, over the last years the demand of trullo-inclusive farms from Italian and foreign buyers has been increasing, causing a strong rising trend of their market value. As a consequence, a different assessment has been arisen for trullo-inclusive lands, causing the formation of a new property typology that has progressively assumed intermediate characteristics between urban real estate and ordinary farms in terms of vivacity and market segmentation for the huge variability of their characteristics.

In this survey, the analyses have been carried out on a dataset of 169 trading instances of trullo-inclusive farms in the Ceglie Messapica, Cisternino, Fasano and Ostuni countryside (Figure 3). About the sample numerousness, it may seem scanty in the international framework, but is sizeable if related to the Italian real estate situation (scarce transparency), the peculiarity of the investigated market and the economic trend over the last years.

The transactions occurred over the period October 2008 - June 2010 and the data have been gathered by the estate agencies of the area. About the choice of predictors, through interviews to local opinion leaders has emerged that the sale price is influenced essentially by 13 amenities. Examples are: distance from the nearest town, size of farm, size of trullo and annexe, renovation, etc. (Table 2).

In order to warrant a high statistical reliability, the ratio between the number of instances and the number of the covariates has been tested. In particular, empirical criteria suggest this ratio should be equal to 4 – 5 (Dilmore 1981; Shenkel 1978) or even to 10 (Dilmore 1981; Weaver 1976). In this study the ratio has been equal to 13, so the condition has been satisfied.

About the dependent variable, the sale price is taken as the price per hectare, while the size of trullo and its annexe (if existing) has been suitably elaborated in

[5] Traditional limestone dwellings with characteristic conical roof, widespread in the Apulia region (southern Italy).

order to obtain the incidence of these buildings per hectare, homogenizing the data. All other covariates have been introduced as such (DIST, LDSIZE) or in dichotomous format (BZONE, RENOV, ELECT, etc.). Tables 3 and 4 report common descriptive statistics for continuous and categorical predictors.

Figure 3. Research area and traditional limestone dwellings (trulli) in Valle d'Itria, Apulia region.

Table 2. Variables included in the analysis.

Variable	Definition
SP	Sale price, in Euros (dependent variable)
DIST	Distance from the nearest town, in kilometres
LDSIZE	Size of farm + size of trullo, in square metres
TRINDEX	Size of trullo/size of farm * 10000
ANINDEX	Size of annexe/size of farm * 10000
BZONE	Trullo is nearby nature reserves, tourist centres, protected areas, archaeological sites, panoramic areas, etc. = 1, 0 otherwise
RENOV	Renovation of trullo in the last 5 years = 1, 0 otherwise
ACC	Farm overlooks highway = 1, 0 otherwise
POOL	Presence of pool = 1, 0 otherwise
ELECT	Presence of electricity network = 1, 0 otherwise
WELL	Presence of well = 1, 0 otherwise
WATER	Presence of water system = 1, 0 otherwise
COND	Presence of conditioner = 1, 0 otherwise
PHONE	Presence of telephone system = 1, 0 otherwise

Table 3. Descriptive statistics of the continue predictors.

Variable	Max.	75%	Median	25%	Min.	Mean	Stand. Dev.	Skewness	Kurtosis
SP	701896	164528	78062	56344	9949	132818.04	127299.20	1.99	4.04
DIST	15	9	7	5	2	7,42	3.02	0.27	-0.54
LDSIZE	30381	15512	12184	9384	2417	12663	5011.60	0.71	1.32
TRINDEX	264	119	75	58	29	90.98	46.00	1.20	1.35
ANINDEX	180	60	0	0	0	29.11	39.30	1.07	0.20

Table 4. Descriptive statistics of the categorical predictors.

Variable	Data	Frequency	%
RENOV	0	90	53
	1	79	47
BZONE	0	150	89
	1	19	11
ACC	0	131	78
	1	38	22
POOL	0	160	95
	1	9	5
ELECT	0	116	69
	1	53	31
WELL	0	130	77
	1	39	23
WATER	0	141	83
	1	28	17
COND	0	154	91
	1	15	9
PHONE	0	149	88
	1	20	12

6. Empirical results

6.1 MLR analysis

In the MLR analysis, the stepwise selection technique has been used (Table 5). The entry and excluding criterion employed is the p-value of F-statistic, set equal to 0.05. So the final model includes 6 of the 13 predictors initially considered. In

particular, LDSIZE, TRINDEX, ANINDEX, RENOV, ELECT and WATER have been selected, giving the following equation:

$$SP = 24397 - 4.2 * LDSIZE + 955 * TRINDEX + 858 * ANINDEX + \\ + 75203 * RENOV + 30462 * ELECT + 33098 * WATER \quad (12)$$

Statistical tests and indexes indicate a good model and the parameters have the expected sign, except for LDSIZE. In this case, a sale price directly proportional to farm size was expected, but results indicate the contrary, likely because of the low interest of buyers with respect to land. In particular, purchasers are tourists exclusively interested in trullo in order to restore and use it during holidays while land is considered rather a constraint than an amenity. In addition, in Table 6, the partial influence of the selected covariates is reported. In this case it is possible to note that roughly the 60% of the variation of the dependent variable is mostly explained by TRINDEX and RENOV, as expected.

Table 5. Estimate by stepwise selection.

Parameter	Estimate	S.E.	t-Ratio	p-Value	VIF
Constant	24396.7	23847.2	1.02	0.308	0
LDSIZE	-4.2	1.2	-3.42	0.000	1.226
TRINDEX	954.8	162.7	5.87	0.000	1.786
ANINDEX	858.3	172.6	4.97	0.000	1.467
RENOV	75203.0	14026.0	5.36	0.000	1.571
ELECT	30462.5	12744.5	2.39	0.018	1.121
WATER	33098.2	15076.7	2.20	0.030	1.008
R^2	0.686			F-statistic	59.12
R^2 adj	0.675			p-Value	0.000
S.E.of regression	72588				

Table 6. Stepwise selection summary.

Step	Introduced Variable	Removed Variable	Partial R2	R2 model	F-statistic	p-Value
1	TRINDEX	-	0.506	0.506	170.95	0.000
2	RENOV	-	0.081	0.587	32.55	0.000
3	ANINDEX	-	0.045	0.632	19.97	0.000
4	LDSIZE	-	0.035	0.667	17.45	0.000
5	ELECT	-	0.010	0.677	5.16	0.024
6	WATER	-	0.009	0.686	4.82	0.030

6.2 MT analysis

The MT analysis has been carried out through the M5P algorithm, a classifier able to generate model trees in which a linear equation is calculated for each leaf on the basis of the instances that reach it. The minimum number of cases for each leaf has been set equal to 20, in order to obtain a statistically significant ratio between selected covariates and number of instances included in each leaf. Furthermore, 10-fold cross-validation has been used and, in order to obtain a thrifty model, a pruned tree has been calculated. Hence, the final tree has 4 leaves (Figure 4) whose linear models (LM) are reported in Table 7.

Figure 4. Pruned tree.

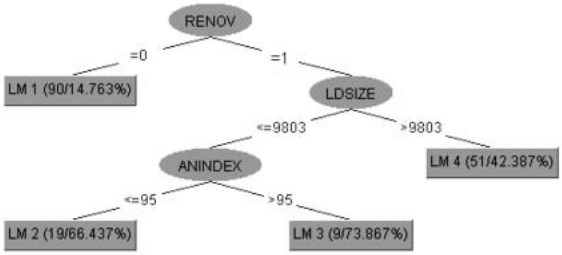


Table 7. Linear models in the pruned tree.

Parameter	LM1	LM2	LM3	LM4
Constant	51539.1	211371.9	252548.7	130562.9
DIST	-2869.8	-	-	-
RENOV	10997.3	12284.2	12284.2	12284.2
LDSIZE	0.4	-4.3	-4.3	-6.1
TRINDEX	283.1	472.9	472.9	881.0
ANINDEX	406.2	780.9	959.1	276.8

The classifier has carried out a first partition of the data set with respect to RENOV. Thus, instances with predictor value equal to 1 have been further divided on the basis of LDSIZE (threshold value equal to 9803 m^2). Finally, instances with farm size lower than 9803 m^2 have been split on the basis of ANINDEX, with threshold value equal to 95 m^2.

Signs in each linear model are consistent with those expected, save LDSIZE, whose coefficients in LM2, LM3, and LM4 are negative as in MLR, with a peak of -6.1 €/m^2 for properties with renovated trullo and land size higher than 9803 m^2. As stated before, this peculiarity may be connected with the low interest of buyers to size land, not being farmers or however interested to the agricultural field in general.

In LM1, instead, the LDSIZE coefficient is positive (0.4 €/m^2), meaning the absence of renovation makes trullo-inclusive lands slightly similar to ordinary farms, even though coefficient magnitude is definitely lower than known values found by real estate practice in the research area.

DIST, left out by MLR, has been selected through MT only for the first linear model. Also in this case, as for LDSIZE, its negative sign suggests this property typology verges on ordinary farm in the absence of renovated trullo, picking out a different market segment. Therefore, analysis of figure 4 permits to understand the logic underlying the mechanism of the value formation adopted by purchasers.

6.3 MARS analysis[6]

The maximum number of BFs has been set at 15 and second-order MARS was applied, so that the basis functions of the models consist of linear and second-order splines. During the pruning step to obtain sequentially smaller models the generalized cross-validation criterion was alternated with 10-fold cross-validation. In Table 8 are detailed the basis functions used as decision points to determine which value will be used in the MARS model at a given knot. In particular, for BF_1, if TRINDEX – 75 > 0, then BF_1 = TRINDEX – 75. Otherwise, if TRINDEX – 75 < 0, then BF_1 = 0. Table 9 provides a ranking of the independent variables, showing their relative importance in terms of percentage of the highest –gcv (the loss in GCV), that is the highest reduction of goodness of fit among all variables. Variables having no impact at all are not shown. Noteworthy is the similarity of this ranking with the MLR one in Table 6 and, in general, with the predictors selected by the MT analysis. Table 10 shows the optimal MARS model calculated for the trulli data set, whose R^2adj is 0.88, much larger than the MLR one (0.67). In it, several interaction terms are integrated in the model, so that it contains a constant, 1 basis function as single spline (BF_6), i. e. defined by only one real estate attribute, and 5 basis functions as second-order interactions of two real estate characteristics. If we consider a simple basis function in the model, for example BF_6, it is interpreted as follows:

$$\left(ANINDEX - 20\right)_+ = \begin{cases} ANINDEX - 20 & \text{if} \quad ANINDEX > 20 \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

This means that when ANINDEX > 20, the fourth element of the second term of the equation (4) is 911.6 * (ANINDEX - 20), otherwise it is 0. In the presence of an interaction between two amenities, as in BF_3, we obtain:

---

[6] MARS v3.0 has been used (www.salford-systems.com).

$$(LDSIZE - 9384)_+ (TRINDEX - 75)_+ = \begin{cases} (LDSIZE - 9384)(TRINDEX - 75) & \text{if } LDSIZE > 9384 \text{ and } TRINDEX > 75 \\ 0 & \text{otherwise} \end{cases} \quad (14)$$

This means that, when LDSIZE > 9384 and TRINDEX > 75, the second element of the second term of the equation (4) is -0.3 (LDSIZE – 9384) (TRINDEX - 75), otherwise it is 0. So, MARS has carried out a deeper analysis, being less restrictive in terms of curvature properties that it can capture. Noteworthy is the predictors interaction. In particular, as said above, if LDSIZE > 9384 m^2 and TRINDEX > 75, property value decreases, adding information about the sale logic adopted through buyers.

Obviously, there are some differences about the covariates included in the model as for ELECT, selected by MLR but not through MT. Other differences concern the cutoff values on which discriminant functions are identified, for algorithm differences.

Table 8. MARS Basis functions (BFi).

BFi	Definition
BF_1	(TRINDEX - 75)+
BF_3	(LDSIZE – 9384)+ * BF_1
BF_5	(RENOV = 0 or RENOV = 1) * BF_1
BF_6	(ANINDEX - 20)+
BF_7	(20 - ANINDEX)+
BF_9	(5 - DISTANCE)+ * BF_1
BF_{10}	(ELECT = 0 or ELECT = 1) * BF_6
BF_{11}	(RENOV = 0 or RENOV = 1) * BF_7
MARS prediction function: $Y = 54597.1 - 0.3 * BF_3 + 2264.9 * BF_5 + 911.6 * BF_6 + 1281.2 * BF_9 + 1586.0 * BF_{10} + 2731.1 * BF_{11}$	

Table 9. Variable importance.

Variable	Importance	-gcv
RENOV	100.00	.985756E+10
TRINDEX	86.21	.809028E+10
ANINDEX	51.00	.476544E+10
LDSIZE	48.24	.457704E+10
ELECT	35.52	.384326E+10
DISTANCE	31.62	.366331E+10

Table 10. The MARS model.

Parameter	Estimate	S.E.	t-Ratio	p-Value
Constant	54597.1	4811.81	11.35	0.000
BF_3	-0.3	0.03	-10.43	0.000
BF_5	2264.9	128.81	17.58	0.000
BF_6	911.6	174.50	5.22	0.000
BF_9	1281.2	173.23	7.40	0.000
BF_{10}	1586.0	196.21	8.08	0.000
BF_{11}	2731.1	520.21	5.25	0.000
R^2	0.884		F-statistic	205.53
R^2 adj	0.880		p-Value	0.000
S.E.of regression	44.173			

6.4 Performance of the fitted models

The indexes described in section 4.1 have been used to assess the prediction performances (Table 11). So, it's possible to note that MARS has the better performance for all 5 indexes: r equal to 0.94, RMSE equal to 43249 €/hectare, MAE equal to 31621 €/hectare, RRSE and RAE equal to 34%. The greatest difference is between MARS and MLR. In particular, the indexes RMSE, MAE, RRSE and RAE for MARS are about 40% smaller than MLR while the difference for r is roughly 12%. Smaller is the difference between MT and MARS (6% for r, 27% for RMSE, 14% for MAE, 28% for RRSE and 15% for RAE).

Table 11. Diagnostic indexes.

Algorithm	r	RMSE	MAE	RRSE	RAE
MLR	0.83	71068	51608	56%	56%
MT	0.88	59559	36873	47%	40%
MARS	0.94	43249	31621	34%	34%

6.5 Significant difference among the fitted models

In Table 12 the Z ratios of the two-tailed Wilcoxon signed-rank are reported. As can be seen for both ARE and MRE, there is a significant difference between MT and MLR, but also between MARS and MLR. On the contrary, MARS is not significantly different from MT, highlighting a similar performance between the non-parametric models. In addition, calculating the more likely market value of a

hypothetical trullo-inclusive farm[7] (Table 13) has emerged that MLR tends to overestimate, while MARS supply the lowest value, with a difference of roughly 50000 € (27%).

Table 12. Wilcoxon signed-rank test.

Algorithm	MLR		MT		MARS	
	ARE	MRE	ARE	MRE	ARE	MRE
MLR						
MT	-5.829 [a]	-5.726 [a]				
	(0.0000)	(0.0000)				
MARS	-5.968 [a]	-5.534 [a]	-0.004 [b]	0.087 [b]		
	(0.0000)	(0.0000)	(0.9969)	(0.9306)		

[a] : W+ < W-

[b] : W+ > W-

Table 13. Hypothetical market values calculated by the examined models.

Algorithm	Market value (€)
MLR	187143
MT (LM4)	151141
MARS	135710

7. Discussions and conclusions

Aim of the research has been to check the forecasting and interpreting capacities of two data mining techniques in real estate appraisal in the presence of segmented markets. In particular, the heterogeneity of real estate markets causes the formation of segments referred to various characteristics, as location, use, structural characteristics, neighbourhood, etc. In real estate practice the individuation and analysis of submarkets is an important requirement in order to avoid questionable assessments (Acciani and Gramazio 2006). To this end, MT and MARS, two techniques that permit *to mine* market segments, have been implemented and compared to canonical MLR.

Starting from the sample survey about sales of trullo-inclusive farms located in the Apulia region, MT and MARS analyses have pointed out market characteris-

[7] Calculated on the basis of a standard property, i. e. DISTANCE=3, LDSIZE=13000, TRINDEX=90, ANINDEX=30, RENOV=1, BZONE=0, ACCESS=1, POOL=0, ELECT=1, WELL=1, WATER=0, COND=0, PHONE=0.

tics not detectable *a priori*, unlike MLR approach. In particular, MT results have revealed the presence of 4 submarkets with quite different linear models, outcomes remarked by MARS. With reference to the diagnostics indexes (r, RMSE, MAE, RRSE and RAE), MT and MARS have a higher performance, warranting a better reliability of results. In fact, as said above, while standard regression analyses are carried out on whole samples obtaining medium implicit marginal prices and highlighting general phenomena, MT and MARS, through progressive sample partition, allow identifying possible patterns representing market segments. So, both data mining models outputs are constituted of several linear models with higher statistical significance than MLR. In addition, the Wilcoxon signed-rank test rejects the hypothesis of same predictive accuracy between MLR and non-parametric models, underlining the absence of sample influence. Hence the superior estimating performance of MT and MARS is confirmed empirically.

This study represents an example of data mining approach on hedonic price models and adds essential information on real estate. MARS and MT perform well and are computationally feasible even with small datasets (typical constraint in the Italian real estate framework). Hence, we hold that implementation of data mining modelling techniques deserves further theoretical and empirical research to obtain in-depth knowledge and understanding of real estate markets.

References

Acciani C., Fucilli V., Sardaro R. (2008). Model Tree: an application in real estate appraisal. *The CAP after the Fischler reform: national implementation, impact assessment and the agenda for future reforms*. (109° European Association of Agricultural Economists). Viterbo, November 20-21.

Acciani C. and Gramazio G. (2006). L'Albero di decisione quale nuovo possibile percorso valutativo. *Aestimum* 48: 19-38.

Breiman L., Friedman J.H., Olshen R.A., Stone C.J. (1984). *Classification and regression trees*. Belmont, CA, Wadsworth.

Christensen L.R., Jorgenson D.W., Lau L.J. (1973). Transcendental logarithmic production frontiers. *Review of Economics and Statistics* 55, 1: 28-45.

Conover W.J. (1999). *Practical nonparametric statistics*. New York, John Wiley & Sons Inc.

Dilmore G. (1972). Notes and comments: multiple regression analysis as an approach to value. *The Appraisal Journal* July: 459-461.

Dilmore G. (1981). *Quantitative techniques in real estate counselling*. Massachussetts, Lexington Books.

Follain J.R., Malpezzi S. (1980). *Dissecting housing value and rent*. Washington D.C., The Urban Institute.

Friedman J.H. (1991a). Adaptive spline networks. *Department of Statistics*, *Stanford University*, Technical Report LCS 107.

Friedman J.H. (1991b). Estimating functions of mixed ordinal and categorical variables using adaptive splines., *Department of Statistics*, *Stanford University*, Technical Report LCS 108.

Friedman J.H. (1991c). Multivariate adaptive regression splines. *Annals of Statistics* 19, 1: 1-67.

Friedman J.H. (1993). Fast MARS. *Department of Statistics*, *Stanford University*, Technical Report LCS 110.

Greene W.H. (2007). *Econometric analysis*. Prentice Hall.

Grillenzoni M., Grittani G. (1994). *Estimo - teoria, procedure di valutazione, casi applicativi*. Bologna, Calderini.

Gujarati D. (2009). *Basic Econometrics*. McGraw Hill.
Halvorsen R., Pollakowski H.O. (1981). Choice of functional form for hedonic price equations. *Journal of Urban Economics* 10, 1: 37-49.
Han J., Kamber M. (2006). *Data Mining: concepts and techniques*. Morgan Kaufmann Publishers.
Hastie T., Tibshirani R., Friedman J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer.
Quinlan J.R. (1986). Induction of decision trees. *Machine Learning* 1: 81-106.
Quinlan J.R. (1992). *Learning with continuous classes*. (Proceedings of the 5th Australian joint conference on artificial intelligence). Singapore, World scientific.
Quinlan J.R. (1993). *C4.5: programs for machine learning*. San Mateo, CA, Morgan Kaufmann Publishers.
Rosen S. (1974). Hedonic prices and implicit market: product differentiation in pure competition. *Journal of political Economy*, 82: 34-55.
Shenkel W.M. (1978). *Modern real estate appraisal*. New York, McGraw Hill.
Wang Y., Witten I.H. (1997). *Inducing model trees for continuous classes*. (Proceedings of the 9th European conference on machine learning). University of Economics, Faculty of Informatics and Statistics, Prague.
Weaver W.C. (1976). To regress or not to regress: that is the question. *The Real Estate Appraiser and Analyst* 6: 31-38.
Wilcoxon F. (1945). Individual comparisons by ranking methods. *Biometrics* 1, 6: 80-83.
Witten I.H., Frank E. (2005). *Data Mining: practical machine learning tools and techniques*. San Francisco, CA, Morgan Kaufmann Publishers.