MODEL TREE: AN APPLICATION IN REAL ESTATE APPRAISAL

Claudio Acciani, Vincenzo Fucilli, Ruggiero Sardaro

Department of Agricultural Economics and Policy, Evaluation and Rural Planning, via Amendola 165/a, 70126 Bari, Italy

E-mail: claudio.acciani@agr.uniba.it; v.fucilli@agr.uniba.it; ruggiero.sardaro@agr.uniba.it;

Contact: Vincenzo Fucilli. E-mail: v.fucilli@agr.uniba.it

European Association of Agricultural Economists

Università degli Studi della Tuscia

Istituto nazionale di Economia Agraria

Rete di informazione della Commissione Europea

Paper prepared for the 109th EAAE Seminar " THE CAP AFTER THE FISCHLER REFORM: NATIONAL IMPLEMENTATIONS, IMPACT ASSESSMENT AND THE AGENDA FOR FUTURE REFORMS".

Viterbo, Italy, November 20-21st, 2008.

Abstract

In the last twenty years in real estate appraisal there has been a growing interest for new and reliable assessment techniques essentially through the introduction of pluriparametric estimate, in particular of linear regression. However, also these techniques seem having not a great deal of adherence to very complex markets, for which the detection of best suited techniques to investigate market segments is necessary. The aim of the research is to test the applicative possibilities of model tree to land market, in order to highlight possible market segments in the original data set not detectable *a priori*.

Key Words: data mining, model tree, multiple regression analysis, land market appraisal,

JEL Code: **C01**

Introduction

The assessment of the most likely market value is, without doubt, the classic aspect of real estate appraisal, for which, during the last two decades the whole procedural frame has been enriched of more and more reliable methodologies. The more representative example is given by the multiple regression analysis (MRA), statistical-mathematical technique which allows not only to measure the influence of a series of characteristics owned by the good object of appraisal on its trade value but also to understand the logic adopted by dealers in trading. However, MRA can be modestly reliable if the analysis is carried out even on a representative sample but of a very complex market that is characterized by more segments (for example, in urban real estate markets). A solution to this problem can be offered by data mining. In particular, data mining is a technique developed very recently in other research's fields but that may be useful also in real estate appraisal of very complex markets. Therefore, the aim of the research is to test the applicative possibilities of a data mining technique to real estate market. In this paper, after having realized a MRA, a segmentation analysis (through model tree) will be carried out, in order to highlight possible sub-samples in the original dataset, that represent specific market's segments not detectable *a priori*. *Ad hoc* diagnostic indexes will be calculated, to check the validity of the adopted techniques and to compare the obtained models.

Data mining and model tree

Data mining is a very recently developed discipline that combines statistical analysis, computer science, artificial intelligence, database management (and connected areas like machine learning). 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* (Giudici, 2005). In this research, a segmentation analysis for numerical prediction has been carried out through model tree, a non-parametric technique. The segmentation analysis consists of a top-down recursive partitioning procedure. With it a data set of observations is divided step by step into groups on the basis of a criterion function (or dividing criterion). The criterion function detects an independent variable (or attribute) and its threshold value in correspondence of which the partition is carried out. The output of model tree is represented by a tree–structure in which it is possible to distinguish a *root node, parent and child nodes, arches (or branches) and terminal nodes (or leaves)*. Nodes, labeled with the name of the attribute chosen for the partition, are connected among them through arches that are labeled with the threshold value of the attribute in correspondence of which the optimal partition is carried out. Each terminal node, instead, reports a linear regression model calculated on the instances that reach it. If we indicate the non-homogeneity of a node by I, the parent node by M, the i^{th} child node by m_i,

the number of child nodes generated by *s* and the relative frequency of the instances into the i[th] child node by p_i, the criterion function of segmentation is:

$$\phi(M, A) = I(M) - \sum_{i=1}^{s} I(m_i)\, p_i \quad (1)$$

The index *I* (named impurity) is a measure of the variability of observations. There are several impurity indexes implemented into specific partitioning algorithms. The most known algorithms for model tree are M5 (Quinlan, 1992) and M5' (Wang and Witten, 1997) being the last an optimized implementation of the first, inside the open source software WEKA[1] (Witten and Frank, 2005). In them, the impurity index is given by the standard deviation (SD), and therefore, its reduction (SDR) is given by:

$$SDR = sd(M) - \sum_{i=1}^{s} \frac{|m_i|}{|M|} sd(m_i) \quad (2)$$

In this case, the dividing process terminates either when the number of observations into the node is less than a fixed value (generally equal to or less than 4), or when the standard deviation of the instances that reach the leaf is less than a minimum threshold (generally 5%) of the standard deviation of the original instance set. After having obtained the model tree, it is interpreted through the rules "if/then". The rules "if/then" are derived from the tracks, through the branches, that lead from the root to each leaf. Generally, the above described heuristic tend to generate very complex and deep trees, as they divide extremely the instances of data set. The deriving phenomenon is known as *overfitting*: the tree generated fits perfectly to the data set and therefore interprets completely the reality of the sample on which has been developed, but doesn't explain anything about the remaining population from which the sample has been extract. In order to remedy this impasse, it's possible to carry out a *pruning* of the tree (Breiman et al., 1984). This technique consists of letting the tree to grow until his maximum size and then in reducing it, by replacing a sub tree with a terminal node. This reduction is possible if the estimated error of the sub tree is bigger than that of the terminal node.

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

Sample Survey

The aim of the research is to experiment an application of the segmentation technique for the assessment of the most likely land market value, measuring its accuracy through a comparison with the diagnostic indexes obtained by an ordinary MRA. The analysis has been carried out on a sample of 109 transfer of land property instances of vineyards (wine grape) in Barletta, a commune in province of Bari, south Italy. These land properties have been traded from 1998 to 2007. In order to secure statistical reliability, the sample survey has been carried out on the basis of two important prerequisites: accurate selection of trade prices and statistical representativeness (Grillenzoni and Grittani, 1994). With reference to the first prerequisite, real trade prices have been obtained through interviews with real estate agents and, moreover, through triangulation of information obtained with local mediators, a sort of neutral and local brokers of land market exchanges. About the second prerequisite, a statistically representative

number of trade instances has been collected. In particular, the statistical representativeness has been checked correlating the number of instances with that of the independent variables. In fact, MRA puts an important condition about the ratio between the number of instances and the number of predictors to warrant a good applicability of results. Empirical criteria suggest this ratio have to be equal to 4 – 5 (Dilmore, 1981 and Shenkel, 1978) or even to 10 (Dilmore, 1981 and Weaver, 1976). Therefore, having built a sample with a number of independent variables equal to 9, the ratio between the instances and the predictors has been equal to 12, so that the second sample condition has been considered widely satisfied. About the choice of the predictors used into the analysis, from the interviews to the real estate agents and local brokers has emerged that the trade price is influenced essentially by: *surface* (expressed in hectares), *production of grapes* (expressed in 100kg per hectare), *distance* (intended as distance from the nearest town centre and expressed in kilometers), *age of plants* (expressed in years), *cultivar* (expressed with 0,6 if there are “Montepulciano”, “Sangiovese”, “Trebbiano” or “Malvasia Bianca” cultivars; with 1 if there is “Lambrusco” cultivar), *irrigation* (expressed with 0 in the presence of consortium irrigation network, with 0,8 in the presence of irrigation company, with 1 in the presence of private well), *access* (expressed with 0, 0,5 and 1 if farm, respectively, is bounded by country lane, municipal/provincial road and highway), *year of trading* (expressed with values included between 1 and 10, indicating the years when each land property has been traded), *neighbour*, (dummy variable expressed with 1 if the buyer is a neighbour farmer; with 0 if the buyer is not a neighbour farmer).

Regression Analysis

After the sample survey, a stepwise regression analysis has been carried out by the software WEKA. To assess the forecasting accuracy of the analysis on the non-detected instances, the 10-fold cross-validation has been carried out. The model so calculated contains 5 of the 9 independent variables initially selected, because the predictors *surface, distance, cultivar* and *access* are not resulted significant to the level selected (α = 0,05). In particular, the regression equation is given by:

Calculated Price = 13.084 + 72,86 * *Production* + 5.986,90 * *Irrigation* - 479,95 * *Year of Trading* - 595,77 * *Age of Plants* + 2.567,82 * *Neighbour* (3)

There are negative signs of the coefficients concerning some variables. Specifically for *year of trading*, because of the average downward trend of the prices during the period examined, and for *age of plant* for obvious reasons connected to the ageing of the vineyards, therefore coherently with what was expected. To evaluate the performance of the method, the following diagnostic indexes have been used:

correlation coefficient (r): measures the statistical correlation between the estimated values *p* and the actual ones *a* of the target variable. It is computed as:

$$r = \frac{\mathrm{cov}(p,a)}{\sigma_p \sigma_a} \quad (4)$$

where *cov(p,a)* is the covariance between the estimated values and the actual ones, while σ_p and σ_a are the respective standard deviations. This coefficient ranges from +1 (ideal situation of perfect correlation) to -1 (uncorrelated results), with coefficient equal to 0 in absence of correlation.

root mean-squared error (RMSE): measured in the same unit of measure of the dependent variable, it ranges from 0 (ideal situation) to infinity. It is computed as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (p_i - a_i)^2}{n}} \quad (5)$$

mean absolute error (MAE): it is equal to the average of errors without their sign, and is given by:

$$MAE = \frac{\sum_{i=1}^{n} | p_i - a_i |}{n} \quad (6)$$

root relative squared error (RRSE): it calculates the relative error with respect to the average of the actual values of the target variable. It is given by:

$$RRSE = \sqrt{\frac{\sum_{i=1}^{n}(p_i - a_i)^2}{\sum_{i=1}^{n}(a_i - \bar{a})^2}} \quad (7)$$

relative absolute error (RAE): similar to the RRSE, it ranges from 0 to 1 and is calculated by the following formula:

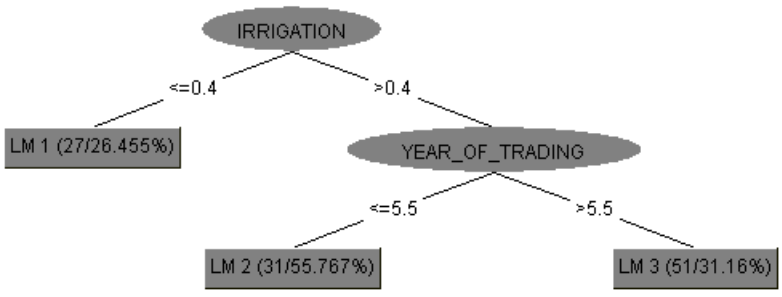
$$RAE = \frac{\sum_{i=1}^{n}| p_i - a_i |}{\sum_{i=1}^{n}| a_i - \bar{a} |} \quad (8)$$

These indexes highlight a fairly good performance of the regression model calculated: r equal to 0,77, RMSE equal to 4.338 €/ha (about 14% of the average of the prices), MAE equal to 3.296 €/ha, RRSE equal to 62,7% and RAE equal to 59,8% (Table 2).

Segmentation Analysis

The segmentation has been carried out through the algorithm M5'. Also in this case, the 10-fold cross-validation has been used. The parameters of the classifier have been set on those of default, with the exception of the pruning level initially set at 0 (maximum size of tree) and after at 2.0 (pruned tree), in order to obtain the easiest interpretation model. Initially an excessive complex model tree formed by 48 terminal nodes has been obtained. After pruning, instead, a new tree with 3 leaves has been obtained (Figure 1).

Figure 1 - Pruned tree



About the linear models calculated, the signs of the coefficients are in keeping with those expected (Table 1). Finally, diagnostic indexes have highlighted a good performance of the classifier: r equal to 0,86, RMSE equal to 3.514 €/ha (about 11% of the average of the prices), MAE equal to 2.568 €/ha, RRSE equal to 50,8% and RAE equal to 46,6% (Table 2).

Table 1 - Linear models of the pruned tree

Parameters	LM1	LM2	LM3
Intercept	3.112,32	-6.093,45	2.135,65
Surface	1.814,23	-	-
Production of grapes	74,24	128,07	76,78
Distance	-	-502,90	-
Age of plants	-227,60	-296,41	-539,16
Cultivar	-	-	-
Irrigation	2.416,24	4.004,73	9.591,76
Access	-	-	1.995,16
Year of trading	-32,58	1.110,69	153,39
Neighbour	-	-	-

Table 2 - Diagnostic indexes

Algorithm/index	*r*	*RMSE*	*MAE*	*RRSE*	*RAE*
Linear regression	0,77	4.338	3.296	62,7%	59,8%
M5'	0,86	3.514	2.568	50,8%	46,6%

Interpretation of Results

Comparison among the diagnostic indexes has highlighted a better forecasting capacity of the model tree with respect to the linear regression because segmentation has a better capacity of fitting to the sample. In fact, the gradual partition of the original data set has allowed to carry out a straightforward analysis "modeled" on the possible market realities which, otherwise, would have been hardly detectable by the researcher. Moreover, it is appropriate to highlight also the considerable interpretative capacities of segmentation, at least with respect to land market analysis. In fact, during the first stage, the classifier has divided the original data set with respect to the attribute *irrigation*. This segmentation has generated a first leaf (LM1) containing all the instances with the value of the above-named

variable ≤ 0,4 and corresponding to the vineyards which are included in consortium irrigation network. For this farm typology, the coefficient of the variable *year of trading* is negative (-32,58 €/ha), indicating a downward trend, even if light, of the prices during the time-frame considered. *Vice versa*, the subset concerning vineyards included in irrigation company or provided with private well (threshold value > 0,4) has been further divided with respect to the variable *year of trading* (threshold value equal to 5,5). Therefore two more leaves have been generated, LM2 and LM3, in which the attribute *irrigation* influences positively the target variable, even if with different intensity, though the global downward trend of prices. It means that the presence of irrigation company or private well has a great positive influence on prices, being these irrigation typologies more reliable than consortium irrigation network. This tendency has been confirmed by the real estate agents and the local brokers interviewed and is due essentially to the discontent generated by the non optimal service offered by the consortium managing authority (malfunctioning of the network, etc.). Finally, it is important to note the constant presence of the attributes *production*, *age of plants* and *irrigation* inside the three models calculated. This is in line with what has been affirmed by the subjects interviewed about the local intensive grape farming, aimed at producing high output as a strategy to contrast the strong price decrease in the last five years.

Conclusions

Aim of the research has been to check the forecasting and interpreting capacities that a segmentation analysis can offer in real estate appraisal, both to assess the most likely market value and to study the land market dynamics. After the survey about the vineyards exchanges (from 1998 to 2007), regression and segmentation analyses have been carried out and compared. With reference to the diagnostics indexes calculated (r, RMSE, MAE, RRSE and RAE), model tree has been characterized by a better performance and reliability of the results obtained. In particular, segmentation has proved to offer some advantages, allowing to identify aspects of the local market that are not detectable through an ordinary MRA. The research, therefore, has allowed to highlight the validity of segmentation analysis in real estate appraisal, overall as alternative analysis tool of heterogeneous and complex samples.

References

Acciani C., Gramazio G., L'Albero di Decisione Quale Nuovo Possibile Percorso Valutativo, AESTIMUM n. 48, 2006.

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

Dilmore G., Quantitative Techniques in Real Estate Counseling, Lexington Books, Massachussetts, 1981.

Giudici P., Data Mining, Metodi informatici, statistici e applicazioni, The McGraw-Hill, 2005.

Grillenzoni M., Grittani G., ESTIMO: Teoria, Procedure di Valutazione, Casi Applicativi, Calderini, Bologna, 1994.

Quinlan J. R., Induction of Decision Trees, Machine Learning n. 1, Kluwer Academic Publishers, Boston, 1986.

Quinlan J. R., Learning with Continuous Classes, in Proceedings of the 5th Australian Joint Conference on Artificial Intelligence, World Scientific, Singapore, 1992.

Quinlan J. R., C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, San Mateo, CA, 1993.

Shenkel W.M., Modern Real Estate Appraisal, McGraw Hill, New York, 1978.

Wang Y., Witten I. H., Inducing Model Trees For Continuous Classes, Proceedings of the 9th European Conference on Machine Learning, , University of Economics, Faculty of Informatics and Statistics, Prague, 1997.

Weaver W.C., «To Regress or Not to Regress: That is the Question», The Real Estate Appraiser and Analyst n. 6, 1976.

Witten I. H., Frank E., Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann Publishers, San Francisco, CA, 2005.