INTERNATIONAL REAL ESTATE REVIEW
2007 Vol. 10 No. 2: pp. 42 - 65

Comparing Rough Set Theory with Multiple Regression Analysis as Automated Valuation Methodologies

Maurizio d'Amato
1st Faculty of Engineering, Technical University of Bari, Politecnico di Bari, Italy
E-mail: madamato@interfree.it

This paper focuses on the problem of applying rough set theory to mass appraisal. This methodology was first introduced by a Polish mathematician, and has been applied recently as an automated valuation methodology by the author. The method allows the appraiser to estimate a property without defining econometric modeling, although it does not give any quantitative estimation of marginal prices. In a previous paper by the author, data were organized into classes prior to the valuation process, allowing for the if-then, or right "rule" for each property class to be defined. In that work, the relationship between property and class of valued was said to be dichotomic.

A real estate property may be considered inside or outside a specific class of valued. This paper introduces a valued tolerance relation to allow for more flexible rules, and offers an objective measure of discriminant threshold. In this case, the results have been derived from an explicit and specific relationship. The methodology was tested on 600 transactions in the residential property market of Helsinki, Finland. The sample of property transactions were divided into two parts (wherever possible) to calculate both internal validity on 390 "in-sample" property transactions and valuation accuracy in an "out-of-sample" group of 210 property transactions, thus obtaining interesting results.

Keywords

mass appraisal; property valuation; rough set theory; valued tolerance relation

Introduction

The interest in developing techniques for automated valuation methodologies (AVM) is increasing, because of their wider application in property taxation, insurance real estate investment trusts (REITs) and mortgage management. However, appraising a large sample of properties is limited by temporal and financial constraints (Ward, et al., 1999), since the property market is fixed in terms of (geographical) location, and as illiquid, or highly durable, assets. This paper represents a further step in the application of rough set theory for mass appraisal problems, a valuation methodology applied previously by the author to a small sample of residential property transactions in the real estate market of Bari, Italy (d'Amato, 2004). In the present work, the valuation methodology is applied for mass appraisal problems to a sample of 600 residential property transactions obtained from the Real Estate Market Observatory of the 1st School of Engineering of the University Polytechnic of Bari. These properties are located in the center of Helsinki, Finland. The sample was divided into two parts. Both multiple regression analysis (MRA) and rough set theory (RST) were used for predicting valued in the first half of the sample of 390 property transactions.

Valuation of internal validity and the valuation variation between the two methodologies were calculated for this "in-sample group" of property transactions. Both the econometric model and the if-then rules were tested on the remaining 210 property transactions, in order to have an "out-of sample" valuation and a measure of valuation variability between the two mass appraisal models. The RST model presented here has been enhanced with a particular functional extension defining "Valued Tolerance Relation" (Stefanowski and Tsoukias, 2000) which allows the appraiser to choose the right rough set rule per each object (property), for valuation purposes. Although the MRA remains the most reliable automated valuation method, the approach presented here, applying RST, has yielded interesting results. It must be stressed that this approach may be helpful in underdeveloped real estate markets where an econometric analysis may fail because of the quantity and the quality of data. The application presented is a first version of the integration between RST and the functional extension, Valued Tolerance Relation (VTR). After three articles, a more extensive work on the integration of VTR and RST is included in the Chapter 11 of a forthcoming work from the author (d'Amato and Kauko, 2008). According to the classification offered in the forthcoming book, this application of RST can be defined as "simplified version". The current paper is organized as follows. The first section comprises an overview of the emerging approaches on mass appraisal. The second section offers a brief presentation of RST and the functional extension, VTR. In the third section, a comparison is made

between the RST and a regression analysis model. The paper concludes with final remarks and future directions of research.

An Overview on Emerging Approaches in Automated Valuation Methodologies

Mass appraisal can be defined as the systematic appraisal of groups of properties as given data using standardized procedures and statistical testing. Mass appraisal statistical modeling tries to replicate market behavior through a representative (econometric) model which achieves this aim. Automated valuation methods must explain the behavior of supply and demand patterns for groups of properties. For this reason, these methodologies refer to large groups of properties rather than single properties. The main issue is the same whether the approach is mass, or single, valuation: an accurate assessment of the value of many properties or a single property (McCluskey, et al., 1997).

According to Silverherz (1936) the reappraisal of St. Paul, Minnesota, marked the beginning of scientific mass appraisal. Development accelerated in the 1950s, with the introduction of computers. Several contributions addressed the importance of mass appraisal, especially in the property market, and explored the relationship between property valued, property characteristics, and urban social and economic problems. Market behavior is influenced by property prices, the high durability of this particular asset, and by fixed, geographical location (Robinson, 1979; Harvey, 1996). Hedonic price modeling has been proposed to define an econometric relationship between price and property characteristics. The application of hedonic price theory (Griliches, 1971; Rosen, 1974) is based on demand-side analysis in a static framework.

It is possible to distinguish at least two types of use of multiple regression models in real estate markets. The former group is essentially based on hedonic modeling. In this case, econometric models are aimed at explaining real estate prices and their variations rather than predicting them, and are essentially model-oriented. Recent studies where hedonic modeling has been applied successfully include: constructing constant-quality price indices for apartment buildings and vacant land in Geneva, Switzerland (Hoesli, et al., 1997a); determining rental valueds of apartments in central Bordeaux, France (Hoesli, et al., 1997b); explaining the housing market in Tel Aviv, Israel (Gat, 1996); and confirming the rationality of condominium buyers and markets in Hong Kong (Mok, et al., 1995). In the hedonic modeling literature, locational proxies may be defined in various ways (cf., surveys by Ball, 1973; Miller, 1982; Laasko, 1973; Lentz and Wang, 1998). Kang and Reichert (1991) constructed a locational-quality dummy based on levels of

price per sqm. living-space. Similarly, McCluskey and Anand (1999) used a solution, where the location was captured with a categorical 'ward'-variable comprising seven valueds based on mean transaction prices for a given area. Pace, et al., (1998) argued that empirical real estate practice has employed relatively spaceless tools, despite the recognized importance of location in theory, and frequent observations in the literature regarding the limited optimality of such tools. An interesting direction for research has been the inclusion of the locational variable inside the AVM models (Clapp, 2003; Clapp, et al., 2002). In this case, it is possible to develop a local regression model (LRM) including, in the model, both housing characteristics and a vector of latitude and longitude. The term concerning the locational valued may be estimated in several ways. Unfortunately, the data set obtained by the Real Estate Market Observatory does not give geographical data concerning property transactions. For this reason, the comparison was carried out between the RST and a linear MRA.

The latter group is data-oriented and involves predicting property price. In this case (see, e.g., reviews by Ball, 1973; and Lentz and Wang, 1998), success has been defined by model fit and model significance, whether each independent variable has the anticipated sign of price association (indicated by the partial correlation coefficients) and whether each independent variable is statistically significant. Sometimes standard model error and various test statistics are also applied as formal criteria of modeling performance. Integration of adjustment grid methods and regression analysis was proposed by Colwell, Cannaday and Wu (1983), in order to integrate the ordinary-least-squares (OLS) estimation of adjustment factors to the standard method.

Here, the multiplicative percentage-grid adjustment method is considered a particularly promising option (Kang and Reichert, 1991). For Mason and Quigley (1996), on the other hand, theory does not provide guidance about choice of functional form, which is considered purely an empirical matter. To this end, non-parametric (and semi-parametric) methods are useful in confirming non-linear relationships. Despite the great number of applications, statistic data analysis has a theoretical weakness (Lentz and Wang, 1998) and may not be efficient in those markets where uncertainty is high because of the unavailability, or the nature, of information and information sources. With alternative approaches such as neural networks (Borst, 1992; McCluskey, et al., 1997; Rossini, 1997; Nguyen, et al., 2001), several criticisms have been raised (Worzala, Lenk, and Silva, 1995). Later, McGreal, et al., (1998) arrived at the same conclusion as Worzala, Lenk and Silva, observing how outcomes vary with different neural network-based models. In another study, Nguyen and Cripps (2001) have argued that artificial neural networks (ANNs) are better than MRA if the data set is large,

and if the right training parameters are found, whereas MRA is, in turn, better if small data sets are run. While results also depend on the functional specification of MRA, MRA still performs better when small samples are used. If sufficient data size and appropriate ANN parameters are found, multilayered perception (MLP) performs better. When MRA was compared with a particular kind of neural network called the self-organizing map (Kauko, 2002), the findings were quite different, however. The ANN-based model performed better than the MRA models, for small data sizes, in recognizing a hypothesized impact between externality and price.

The genetic algorithm (GA) is another new tool from the machine learning paradigm. Recently, GA has been used by Cooley, et al., (1994) to model and add the locational variable to the list of independent variables. In rule-based expert systems, human judgment and decision-making are modeled assuming explicit rules instead of learning automatically from experience. Scott and Gronow (1990) discuss the components of valuation expertise within the mortgage valuation domain and further explore the different levels at which this expertise is exhibited. Another contribution to AVM comes from the application of fuzzy theory. Gonzales, et al., (2002) have built a mass appraisal model, where fuzzy rules are extracted from the ANN. Suitable methods, based on the stated-preferences approach for measurement of utility, include more established tools, such as contingent valuation, conjoint analysis, and perceived diminution (McLean and Mundy, 1998; Ruokolainen, 1999), as well as more experimental ones, such as valued tree models (Kauko, 2002). Some authors freely encourage the use of interview survey methodology for residential valuation. For Lentz and Wang (1998), a questionnaire focusing on individuals' willingness to pay for certain property characteristics, such as aesthetic valued, is an equally valid technique.

The contingent valuation method (CV) is based on a formal questionnaire indicating the respondent's willingness to pay (WTP) or willingness to accept (WTA) a given sum of money. Breffle, et al., (1998) have used contingent valuation to estimate a neighborhood's WTP to preserve a parcel of undeveloped land in Colorado. Another tool for valuation support is the analytic hierarchy process (AHP). This method offers a model based on a valued tree concept, whereby choices are made according to preferences in a multi-attribute problem setting, in contrast to the purely economic CV setting of revealed preferences. The AHP has been applied in several ways within this particular research area. Like Kauko (2002), Fischer (2003) also sees a more qualitative approach as an improvement. The difference between these two specific methods is that Fischer mixes price criteria with the other criteria already present in the model structure, whereas Kauko arrives at a pure quality rank, which can be compared with actual prices at a later stage.

Fischer (2003), like Kauko (2002), concludes that the problem with this approach, if it is being adopted for practical applications, may be the very low time- and cost-efficiency. Ghyoot (2001) uses the AHP for site selection, together with the repertory grid (RG) – a more qualitative method, which is focused on the search for best choice on the basis of in-depth interviews.

Of these developments in multi-criteria methods, RST represents a recent direction for research in mass appraisal (d'Amato, 2002). A methodological improvement to RST has been developed by d'Amato, (2004), to integrate the so-called valued tolerance relation (VTR). This paper represents the first complete application of RST integrated with VTR for a large sample of property transactions.

The Rough Set Theory Method and the Valued Tolerance Relation

Recently, RST has been proposed in the above-mentioned papers as a method for appraising property values by applying the if-then rule without mathematical modeling. As will be shown later, this approach is characterized by methodological steps closer to regression analysis.

Rough set theory

First developed by Zdzislaw Pawlak a polish mathematician in two well-known works (Pawlak, 1982; Pawlak, 1991), RST is a rule-based approach to handle uncertain information, which considers a real estate transaction as an element or object to be related to a specified piece of information. Such information as property price, and technical characteristics of real estate or tenancy, are considered "attributes" of the "real estate transactions." If the real estate price is considered as an object, the only available information is the specific characteristics or attributes relating to property. These characteristics can be "owned" in different ways by an object set. The relationship between an "object" and its "attributes" can be described by three regions of knowledge: "certainly," "possibly," and "certainly not". The relationship between the object and its attribute can be defined as "certainly not" for a property (object) inside a group of property transactions (universe) without a garden (attribute). Among the property transactions (universe), properties with the same attributes can be considered indiscernible at a certain level of information. An indiscernible element is defined as an "elementary set," which cannot be confused with any other element. Two properties (objects) with similar technical features (attributes), and found in the same area, at the same price, can be considered indiscernible.

It has been pointed out that: "… (the) indiscernibility relationship…is the mathematic foundation of rough sets theory, it is the brick on which is built the building of knowledge of reality"(Matarazzo, 1997).[1] As this method does not rely on any assumptions, the first stage of analysis is concerned with the relationship between objects (property transactions) and their attributes (property characteristics). This analysis starts with the so-called "informative table," where rows represent universe units or objects, and columns represent different attributes belonging to objects of a universe. Considering the RST application to real estate valuation, all the attributes (panoramic quality, maintenance need, area, etc.) are listed in columns, with each measured in a different domain. Rows contain single attributes of the universe, or the real estate under consideration. This stage is very important because valuation depends on data quality and homogeneity. Each cell contains the quantitative or qualitative description of the relationship between an object and its attribute. The presence, or the absence, of a parking space (attribute) within a property (object) is marked with a dummy variable, while the area dimension (attribute) of a property (object) is expressed in square meters. It is easy to see a great similarity with the first step of regression analysis. The informative table S is expressed in formal terms, as in equation (1):

$$S = \langle U, Q, V_q, f \rangle \quad . \tag{1}$$

In this equation, U is the universe, or a finite element set (or property transactions). Q is the finite set of attributes or features (property features sold). V_q is the attribute with a q domain, and f is the information function, which describes the relationship between the object and the attribute belonging to the Q set, which varies inside the V_q domain. The information function can be defined formally as follows:

$$f : U \times Q \to V \wedge f(x,q) \in V_q \forall q \in Q \wedge \forall x \in U \quad . \tag{2}$$

This equation shows that the information function works in a universe U of data in which information can be classified through a set of attributes Q. Each object (x) is linked to its attribute (q) through a function. This happens for each attribute inside the Q-set, and for each object (x) inside the universe U. The property transaction $x \in U$ is described by a line which can also be seen as a vector. Each element of this vector represents the value given to the relative attribute of the x object, and which can be defined as Des$Q$ (x). The following relationships among objects can be highlighted. There is an indiscernibleness or equivalency relationship between two objects that belong to the same universe U, when respective attributes are identical. For example, two 100 sqm area properties will be indiscernible with regard to

this attribute. The indiscernibility relations can be expressed formally, considering a non-empty subset *N* of the *Q*-attribute set for $N \subseteq Q$:

$$IN = \left\{(x, y) \in U \times U : f_q(x) = f_q(y), q \in N\right\}. \quad (3)$$

In other words, two objects (properties) can be defined as indiscernible if they have identical characteristics. The pair of coordinates (*x*, *IN*) defines the so-called approximative space. If $(x, y) \in IN$, then it is possible to say that *x* and *y* are *N*-indiscernible. In this case, the indiscernibility relation may only have two valueds: 1 or 0. Moreover if $N = Q$, where the *Q* elementary sets are known as atoms, all the elements are indiscernible. If all the *X* set units of the *U* universe are analyzed according to the *N* attribute set, and if they are similar (for example, where all properties are 100 sq meters and are near the center), then they are indiscernible. Two real estate properties may be characterized by a single difference – but a relevant difference in price – or by two or more differences, but the same price. For this reason two important concepts must be added. Assuming *U* as the universe, *X* as a universe object set (real estate properties with known prices), *Q* as the attribute set (that belongs to *U* universe), and *N* as an attribute subset, the lower approximation can be defined as follows:

$$N_{\_}(X) = \left\{x \in U : N(x) \subseteq X\right\}. \quad (4)$$

If a real property unambiguously has an attribute included in this attribute subset, then it can be defined as part of its positive or lower region, and can be defined in the following terms:

$$N^{-}(X) = \left\{x \in U : N(x) \cap X = 0\right\}. \quad (5)$$

The upper approximation is defined by the set which shows a non-empty intersection with *X*. If there are some elements in set *N* that belong to *X* and others that do not, then the attribute can be described by the upper approximation.

RST evaluates each uncertain phenomenon through these approximations. The difference between the upper or lower regions is represented by a "boundary region" of rough sets. Comparing data in the rough sets to the consumption of an orange, the edible part of the orange is defined through the difference between an inner region that is edible in all its points - its contents - and an outer region where the fruit is not edible at all. The yellow-colored intermediate content is a boundary region that is partially edible (assuming a common taste). The boundary region is expressed formally as:

$$BN_N(X) = N^{-}(X) - N_{\_}(X). \quad (6)$$

The three regions described are useful to define "granular" information. Both qualitative and quantitative attributes are useful to describe an object. If the boundary region is not empty, the rough set is defined through upper approximation and lower approximation union. The granular nature of information is influenced by different aspects such as attribute characteristics, attribute numbers and each attribute domain *Q*. As with MRA, this procedure is heavily dependent on quality of information, ability to classify information and adequately describe single attributes, as well as levels of confidence in knowledge and problems with knowledge itself.

From the "informative table," it is possible to develop a "decisional table," by dividing the attributes in conditional set *C* and decisional set *D*. This distinction between conditional and decisional attributes allows us to establish a causal relation between attributes. Defining price as a decisional variable and attributes as a conditional variable, the RST allows us to see how conditional attributes (property characteristics) influence the decisional attribute (price). Through this procedure, an object can be analyzed and evaluated to determine lower and upper approximation based on the relationships between the set of elements containing the price (decisional attribute) and the set containing other attributes (conditional attributes) which influence price behavior. As with regression analysis, the appraiser selects the conditional attribute in the same way he defines the independent variables that affect the valued in regression analysis. In the final step, the appraiser analyzes the relationships between conditional and decisional attributes. The relation is analyzed by taking into account the lower and the upper approximations between the decisional set *D* "of the price attribute" and set *C* of the attributes that have been selected as conditional. There are two general types of decisional rules. The former is the "exact decisional or deterministic rule," where the decisional set (price) contains the conditional attributes (area or other features). The latter is the "approximative decisional rule," in which only some conditional attributes (in our case sq. meters, date, date of construction, n. of rooms) are included in the decisional set (price). According to previous research, deterministic rules seem to be the most suitable for real estate valuation purposes (d'Amato, 2002; d'Amato, 2004).

In this case, causal relationships between property features and price can be evaluated without the problem of uncertainty. The "granularity" of the system or its uncertainty can be increased if information is based on various observations. Property valued is determined by comparing property characteristics with the rules defined for comparative properties. This phase is very important. In previous papers, classes of valued were obtained, and if-then rules applied to these classes, which then gave a specific class of valued to properties with particular characteristics.

This paper represents a further step in the application of RST to property valuation for mass appraisal purposes. Our previous work was based on a crisp indiscernibility relation (complete, reflexive, symmetric and transitive relation valued in the following domain {0, 1}). In this paper, a valued tolerance relation (Tsoukias and Vincke, 2000) as opposed to a crisp tolerance relation used in the traditional version of RST, will be used to give an objective measure for k-threshold. Following our previous experience with a small data set for the property market of Bari (d'Amato, 2004), a comparison between regression analysis and RST has been proposed. In previous applications of RST (d'Amato, 2002), rules were developed to consider classes of property value, while in the second application (d'Amato, 2004) a valued tolerance relation (VTR) (Tsoukias and Vincke, 2000) was applied for the first time, even though the measure of k-threshold was subjective. A larger sample has now been considered.

Valued tolerance relation (VTR)

VTR can be considered a more flexible way to deal with the indiscernibility relation.

Classical rough set theory relies on the crucial concept of indiscernibility relation as a crisp equivalence relation. Two properties may be indiscernible only if they have the same attributes. In property markets where this is a very powerful assumption, the VTR is a functional extension of RST and allows the appraiser to develop upper or lower approximation with different "degrees" of indiscernibility relation. In our forthcoming works, the concepts of lower and upper approximation will be replaced by lower and upper approximability. The formal relation is indicated below (Tsoukias and Vincke, 2000).

$$R_j(x, y) = \frac{\max(0, \min(c_j(x), c_j(y)) + k - \max(c_j(x), c_j(y)))}{k} \quad . \qquad (7)$$

The relation R_j may assume continuous values included in the interval [0, 1]. It is a variation ratio based on sets where membership function may have values included in the interval [0, 1]. As a consequence, the VTR brings flexibility to traditional RST. In this context, the choice of the minimum in the membership function represents the intersection between two sets, while the maximum in membership function results in the union between the two sets. Two objects *x* and *y* may have different levels of indiscernibility, according to a discriminant threshold *k,* which measures the characteristic c_j. The *k*-threshold can be applied to different measures of these characteristics for all objects. For example, the indiscernibility relation between two objects

(properties A and B) whose sqm area are 120 and 150 for a *k*-threshold of 10 sqm can be calculated as follows:

$$R(a,b) = \frac{\max(0;120+10-150)}{10} = \frac{\max(0;-20)}{10} = \frac{0}{10} = 0 \quad . \tag{8}$$

The two objects cannot be considered similar. The result of the application of a VTR with the same *k* to two objects (say, property transactions) whose sq.m. areas are 120 and 125 is:

$$R(a,b) = \frac{\max(0;120+10-125)}{10} = \frac{\max(0;5)}{10} = \frac{5}{10} = 0,5 \quad . \tag{9}$$

As one can see, the measure of indiscernibility relation is not crisp, but may have different degrees. If the valued of R_j equals 1, the *k*-threshold of the two objects is similar; if, as in equation (8), R_j is equal to 0, the *k*-threshold is completely different. This mathematical formula can also be used for the relationship between the object of a universe (properties) and a R_j set of rules developed for valuation purposes, where the characteristics of the object (property transaction) are compared with the conditional part of the rule considered, indicated in the following equation as $c_j(\rho)$. In this case, it is modified as follows (Stefanowski and Tsoukias, 2000):

$$R(x,\rho) = \frac{\max(0;\min(c_j(x),c_j(\rho)) + k - \max(c_j(x),c_j(\rho)))}{k} \quad . \tag{10}$$

In the formula output, there is a level of indiscernibility relation between the object and the rule, assuming a k level of threshold for the measure of the attribute. In the previous paper (d'Amato, 2004), the measure of *k*-threshold was found to be subjective, due to the preferences and characteristics of the specific property market. In this work, an objective measure of *k*-threshold is proposed as the standard deviation of each attribute. This is an important step in defining a specific application of RST to property valuation.

With the so-called discriminant thresholds used to define the indiscernibility relation between two objects, analysis of both rules and objects must be stressed. If rules concern properties with similar characteristics, then the threshold (standard deviation) is low. The threshold is high, on the other hand, when the rules refer to a sample of elements containing properties with different features. Table 1 indicates *k*-thresholds obtained from the in-sample group.

Table 1: Definition of k-threshold

Standard deviation			
Sqm	**n.of rooms**	**date**	**Years**
25.02501563	0.984987169	3.691223846	7.28743735
k=25	*k*=1	*k*=4	*k*=7

The relationship among all the attributes of an object and the conditional part of the "rules" is calculated assuming the "intersection" of all sets (Stefanowski and Tsoukias, 2000). The intersection is obtained by comparing object with rule. As a consequence, it is possible to obtain several R_j s according to the *n* characteristics of the property and the rules. The select R_j is, as indicated in equation (11), as the minimum R_j among *n* comparisons between rule and object:

$$R_j(x,\rho) = \min_{j=1}^{n}(R_j(x,\rho)) \quad . \qquad (11)$$

Where R_j is the VTR , *x* is a characteristic of the property considered, ρ is the same feature belonging to the conditional part of the rule and *n* is the number of characteristics of a property and the conditional part of the rule. The $R(x,\rho)$ gives a flexible (not crisp) measure of this relationship. Since an object may have more than one attribute, the appraiser has to take into account the minimum R_j among all attributes, as indicated in equation (11). For example, the property indicated in table 2 is an object in the universe of property transactions *U* (in a given sample group of property), which has the following features or attributes:

Table 2: The attributes of object n.2

sqm	**n.of rooms**	**Date**	**Year**
139	5	35	28

In order to select the appropriate rule for property n.2, equation (10) is used to calculate the $R(x,\rho)$, comparing each attribute of the property (object) and each attribute of the conditional part of all rules indicated in the conditional part of 70 property transactions selected. This comparison gives several $R(x,\rho)$ with a valued per each attribute referring to the *k*-threshold indicated in table 1. The minimum of these values, as indicated in equation (11), is the output of the comparison between the single rule and property. Comparison is repeated for the 70 rules until the rule with the highest $R_j(x,\rho)$ valued can be chosen. The selected rule has to satisfy the following first criteria, indicated in equation (12):

$$R_j(x,\rho)_{1^{st}\ criteria} = \max_{j=1}^{m}(R_j(x,\rho)) \quad . \tag{12}$$

The higher the R_j, the greater the similarity among single objects and rules. Applying the above to the property markets of Bari and Amsterdam, it was found that more than one rule had the same minimum R_j. In this case, the appraiser considers as second criteria the rule with the highest sum of R_j calculated in comparison between property and the single rule (absolute maximum).

$$R_j(x,\rho)_{2^{nd}\ criteria} = \max_{j=1}^{m}(\sum_{j=1}^{n} R_j(x,\rho)) \quad . \tag{13}$$

In fact, a property with a greater sum presents a higher R_j than other objects. It may happen that the highest sum of R_j, indicated as second criteria, does not match the first criteria. In this case, it is possible to select the rule using a third criteria. The correct rule is given as the highest sum among those rules satisfying the first criteria (relative maximum). These criteria must be considered fundamental for choosing the right rule for mass appraisal purposes. By applying these rules, a comparison between MRA and RST on a large sample has been possible.

A Comparison between Regression Analysis and Rough Set Theory

The sample of property transactions was divided into two parts. Appendix 2 highlights the characteristics of this sample. The first part, which is composed of 390 observations, allowed us to calculate an in-sample internal validity of two automated valuation methods, namely MRA and RST. The remaining set of 210 observations allowed the calculation of out-of-sample valuation accuracy and the variability of results between MRA and RST. Both regression and RST models were tested on 210 out-of-sample properties. In order to analyze valuation accuracy (out-of-sample) and internal validity (in-sample), the market price of each property was compared with the valued predicted from RST rules. Error measurement was calculated with the forecasting error, in the following equation:

$$FOR.ERROR = \left| \frac{PS_i - AS_i}{AS_i} \right| \times 100 \quad , \tag{14}$$

where PS_i and AS_i are predicted selling price and actual selling price of property i in the set of m properties. Another measure was calculated using the mean absolute percentage error (MAPE) indicated in the equation below:

$$MAPE = \frac{\left(\sum_{i=1}^{n} \left| \frac{PS_i - AS_i}{AS_i} \right| \times 100 \right)}{m} . \quad (15)$$

The measure of forecasting error was divided into four different categories: those where the forecasting error falls within an interval between 0-10%; those with a forecasting error between 10%-20%; those with a forecasting error between 20%-30%, and, finally, those with a forecasting error more than 30%.

Appendix 1 shows the output of regression analysis. The date referred to is January 2005. The model is significant, as indicated by the F- test, the R^2 is 0.927, and the Adj. R^2 is 0.926. The *t* test shows interesting results for the significance of parameters. The in-sample of 390 observations is shown in Appendix 2. The MRA model is indicated in equation (16):

$$\begin{aligned} \text{PRICE} &= 39.975,805 + 1.290,916\text{SQM} + 6507,988\text{ROOMS} \\ &-397,107\text{DATE} - 165,287\text{YEAR OF CONSTR} \end{aligned} . \quad (16)$$

Price in euros is dependent on location, number of square meters (SQM), number of rooms (ROOMS), date of transaction (DATE), and, finally, on years of construction (YEAR OF CONSTR). The in-sample application of regression analysis showed a 0.06 MAPE. The internal validity of the MRA is highlighted in table 3:

Table 3: A comparison between market prices and MRA estimated prices to calculate the internal validity (in-sample) of 390 property transactions

Proportion of Errors	0-10%	10-20%	20-30%	More than 30%	Total	MAPE
N.of Observations	323	63	3	1	390	
Percentage FOR.ERROR	82,82	16,15	0,77	0,26	100	0,06

The proportion of error within the interval of 0-10% shows interesting results for MRA. RST was applied to the same in-sample group of 390 observations. The rules for the application of RST were based on 70 real transactions inside this sample. In a similar way, the application of RST refers to the following four conditional attributes for RST (independent variables for MRA): SQM - square meters; ROOMS – number of rooms; DATE – number of months; YEARS – number of years of construction. For example, the first rule listed can be read as follows:

IF SQM = 36 ∧ ROOMS = 1 ∧ DATE = 41 ∧ YEARS = 17 → PRICE = 70.632, 00 € (17)

Rules were generated comparing the object features to 70 property transactions. In this case, rules coincide with decision tables. This happens because no "class" of value has been considered, as there was in the first work on RST (d'Amato, 2002). In order to analyze the "quality" of the rule, there are two important indexes: "coverage" of rule and "accuracy" of rule. The former index is a ratio between the number of properties which satisfy both the conditional and the decisional part of a rule, and the number of properties, which satisfies only the decisional part (Pawlak, 1997). The latter index measures the probability that the decisional part is exact. In other words, it is the ratio between the number of properties that satisfy both conditional and decisional parts of a rule and the number of properties that satisfy only the conditional part. All rules given in table 3 have the highest level of accuracy and coverage (equal to 1).

The highest level of both indicators is very important for mass appraisal problems. In this case, there are no "better" or "worse" rules, but, rather, all the rules give an equal contribution for valuation purposes. It must be stressed that these thresholds allow the appraiser to have a flexible measure of indiscernibility relation. Accuracy ranges are given in order to highlight the internal validity of RST application:

Table 4: Comparison between market prices and RST estimated prices to estimate the internal validity (in sample) of 390 property transactions

Proportion of Errors	0-10%	10-20%	20-30%	More than 30%	Total	MAPE
N.of Observations	237	110	29	14	390	0,0884
Percentage FOR.ERROR	60,77	28,21	7,44	3,59	100,00	

The difference between the internal validity of MRA and RST is quite small, even if the advantage of MRA on RST is evident. MRA performs better in all the intervals, based on proportion of errors. For example, the proportion of errors included in the interval 0-10% is 60.77% for RST, while 82.82% for MRA. The interval between 10-20% shows a 28.21% proportion of error for RST, compared to 16.15% for regression analysis. The proportion of error between 20-30% shows a 7.44% proportion of error for RST, compared to 0.77% for MRA. For proportion of error more than 30%, it is possible to observe a 3.59% for RST, compared to 0.26% for MRA.

Table 5: Comparison between market prices and MRA estimated prices. Valuation accuracy (out-of-sample) of 210 property transactions.

Proportion of Error	0-10%	10-20%	20-30%	More than 30%	Total	MAPE

N.of Observations	71	138	0	1	210	0,11
Percentage FOR.ERROR	33,81	65,71	0,00	0,48	100	

The results concerning the out-of-sample variation or valuation accuracy (Brown, 1985; Brown, et al., 1998) show the difference between predicted prices with MRA computed. Actual prices are indicated in table 5.

In table 6, the valuation accuracy of the RST method is calculated with the difference between actual and predicted prices computed with RST.

Table 6: Comparison between market prices and RST estimated prices. Valuation accuracy (out-of-sample) of 210 property transactions.

Proportions of Errors	0-10%	10-20%	20-30%	more than 30%	Total	MAPE
N. of Observations	143	27	19	21	210	0,0997
Percentage FOR.ERROR	68,42	12,86	9,05	10,00	100	

As one can see, the differences between predicted valued and actual valued of property as computed with RST and MRA are within an interval of 10% and 20%, where valuation accuracy has also been included. For valuation accuracy, RST performs better than MRA. In the interval 0-10%, there is 68.42% proportion of errors for RST and 33.81% for MRA. Error percentage, in the interval of 10-20%, is 12.86% for RST and 65.71% for MRA. In the interval between 20-30%, it is possible to observe 9.05% for RST and 0% for MRA. It should be noted that the proportion of error superior to 30% is higher in RST (10%) than MRA (0.48%). The variability in results between the two methodologies, RST and MRA, have been calculated for the 390 “in-sample data”. It can be defined as “in-sample valuation variability” and shows the percentage difference between MRA estimated price, and the RST estimated price per number of observations and proportion of the sample. The final results are reported in table 7.

Table 7: Variability of predicted price between MRA and RST for the 390 in-sample property transactions.

	0-10%	10-20%	20-30%	More than 30%	Total	IN-SAMPLE VALUATION VARIABILITY
N. of Observations	255	109	16	10	390	0,0925049
Percentage	65,38	27,95	4,10	2,56	100	

As one can see, the results between MRA and RST are closer in the 65.38% sample.

Similarly, in table 8, it is possible to observe the out-of -sample valuation variability between the estimated prices of the two different mass appraisal methodologies.

Table 8: Variability of predicted price between MRA and RST for the 210 out-of-sample property transactions.

	0-10%	10-20%	20-30%	more than 30%	Total	OUT-OF-SAMPLE VALUATION VARIABILITY
N.of Observations	70	95	27	18	210	0,152569
Percentage	33,33	45,24	12,86	8,57	100	

Table 8 shows a higher distance between MRA estimated prices and RST estimated prices.

Some differences between RST and MRA must, however, be highlighted. Multiple regression analysis allows the appraiser to define the price of each property characteristic considered in the model, while rough set theory does not give information about hedonic – marginal prices. The RST valuation procedure does not give information for the location variable or the relationship between marginal prices.

For this reason, the quality of outputs from statistical mass appraisal methodologies still remains superior to those obtainable from the RST approach. MRA relies on econometric modeling which reproduces market behavior based on a probability framework. RST is not based on a mathematical model. In fact, the results of this mass appraisal valuation technique are dependent on simple observation of market reality. In a prior application of RST (d'Amato, 2002), valuation results were given as classes of valued instead of crisp valued; only through the valued tolerance relation could single values be estimated. Regression theory has, on the other hand, statistical control indexes built into model assumptions. In RST no such assumption is made; control indexes are restricted to the two main indexes, "accuracy" and "coverage" of rules.

It is important to bear in mind that, although there are data limits with RST, a high number of observations allow the appraiser to develop rules at the highest level of coverage and accuracy. The two valuation procedures are similar in other respects. As one can see, both the application of RST and MRA are based on cross-sectional analyses. The valuation process starts with the definition of "attributes" in RST and independent variables in MRA. In fact, a cause/effect relationship is assumed in both MRA and in RST. With MRA, output is a mathematical model, while in RST the output is a

boolean sum, or an if-then rule. Both valuation procedures give the same results, starting from the same sample and the same group of attributes.
There is no risk of different results coming from different "algorithms", as, for example, with neural networks. In neural networks, the final result is dependent on the choice of learning algorithms and network features. As a conesquence, choosing different information tools and starting from the same set of attributes different results may be obtained (Worzala, Lenk, and Silva, 1995; McGreal, et al., 1998). Application of RST is recommended for mass appraisal in those markets where the property market is not transparent, such as in Eastern European Countries.

Final Remarks and Future Directions of Research

In this paper rough set theory has been applied to a large sample of property transactions for mass appraisal problems. Valuation procedures have been applied together with a valued tolerance relation to obtain a crisp valued instead of a class of property values. The results of RST are encouraging, even if quality of output for MRA is higher than RST. While no information is available for location variable or marginal prices, RST gives similar results to MRA and a superior performance in out–of-sample valuation accuracy. In particular, RST may be a useful tool in those markets where econometric modeling cannot be applied because of lack of quality and quantity of property market data sources.

Because of lack of information for geographic location, a comparison between RST and regression methods, which includes the location variable, unfortunately, was not possible.

An interesting direction for future research could involve a comparison between MRA and RST in other urban contexts, to see whether empirical results obtained here for the residential property market in Helsinki are confirmed by the wider picture. In particular, MRA models that take into account the locational variable can be compared to RST performance.

Another idea could be to work on an application of RST that tries to include the location variable. This may be possible by developing a variable that divides the dataset in "class of distance" from the central business district (CBD), including the class of distance as a conditional variable. Furthermore, a house price index could be built on this mass appraisal procedure. Finally, one should stress the valued of applying RST to underdeveloped markets, where it is difficult to apply econometric modeling.

Acknowledgments

The author is grateful for the useful comments and help of Prof. Tom Kauko and an anonymous referee of this paper. It is possible to download previous articles on rough set theory applied for mass appraisal from the research website of the author: www.noaves.com.

References

Ball, Michael J. (1973). Recent empirical work on the determinants of relative house prices, *Urban Studies*, 10, 2, 213-233.

Borst, R.A. (1992). Artificial neural networks: the next modelling/ calibration technology for the assessment community, *Artificial Neural Networks*, 64-94.

Breffle, William S., Morey, Edward R. and Lodder, Tymon S. (1998). Using contingent valuation to estimate a neighborhood's willingness to pay to preserve undeveloped urban land, *Urban Studies*, **35**, 4, 715-727.

Brown, G. (1985). Property investment and performance measurement: a reply, *Journal of Valuation*, **4**, 33-44.

Brown, G.R., Matysiak, G.A., and Shepherd, M. (1998). Valuation uncertainty and the Mallinson Report, *Journal of Property Research*, **15**, 1, 1-13.

Clapp, J.M., Kim, H-J., and Gelfand, A.E. (2002). Predicting spatial patterns of house prices using LPR and Bayesian Smoothing, *Real Estate Economics*, **30**, Winter, 502-532.

Clapp, J. M. (2003). A semiparametric method for valuing residential locations: application to automated valuation, *Journal of Real Estate Finance and Economics*, **27**, 3, 303-320.

Colwell, Peter F.; Cannaday, Roger E. and Wu, Chunchi. (1983). The analytical foundations of adjustment grid methods, *Real Estate Economics*, **11**, 1, 11-29.

Cooley, R.E., Pack, A.D., Hobbs, M. and Clewer, A.D.E. (1994) A genetic algorithm for modelling locational effects on residential property prices, *The Cutting Edge 1994 Conference Proceedings*, 179-193.

d'Amato, M. (2002). Appraising properties with rough set theory, *Journal of Property Investment and Finance*, **20**, 4, 406-418.

d'Amato, M. (2004). A comparison between RST and MRA for mass

appraisal purposes. A case in Bari, *International Journal of Strategic Property Management*, **8**, 205-217.

d'Amato, M. (2008). Rough set theory as property valuation methodology: the whole story, in d'Amato, M. and T. Kauko, eds., Advances in Mass Appraisal.An International Perspective, RICS Real Estate Series, Blackwell Publisher

Fischer, D. (2003). Multi-criteria analysis of ranking preferences on residential traits, *10th ERES conference*, Helsinki, Finland, 10-13 June.

Gat, D. (1996). A compact hedonic model of the greater Tel Aviv Housing market, *Journal of Real Estate Literature*, **4**, 163-172.

Ghyoot, V.K. (2001). Using management science in site selection: a case study in office site selection for a regional authority in South Africa, *3rd AfRES / TIVEA / RICS foundation conference*, Arusha, Tanzania, October.

Gonzalez, M.A.S., Soiberman, L. and Formoso, C.T. (2002) Explaining Results in a Neural-Mass Appraisal Model. *9th European Real Estate Society Conference (ERES).*

Griliches, Z. (1971), Price Indexes and Quality Change, Cambridge Harvard University Press.

Harvey, J., (1996), Urban Land Economics, 4th ed.,MacMillan, London.

Hoesli, M., Giacotto, C. and Favarger, P. (1997a). Three new real estate price indices for Geneva, Switzerland, *Journal of Real Estate Finance and Economics*, **15**, 1, 93-109.

Hoesli M., Thion, B. and Watkins, C. (1997b). A hedonic investigation of the rental valued of apartments in central Bordeaux, *Journal of Property Research*, **14**, 15-26.

Kang, Han-Bin and Reichert, Alan K. (1991). An empirical analysis of hedonic regression and grid-adjustment techniques in real estate appraisal, *Real Estate Economics*, 19, 1, 70-91.

Kauko, J.T. (2002), Modelling the Locational Determinants of House Prices: Neural Network and Valued Tree Approaches, Labor Graphimedia, Utrecht,The Netherland.

Laasko, S. (1997). Urban housing prices and the demand for housing characteristics, The Research Institute of the Finnish Economy (ETLA) A 27, Helsinki.

Lentz, G.H. and Wang, K.(1998). Residential appraisal and the lending process: a survey of issues, *Journal of Real Estate Research*, **15**, 11-39.

Mason, Carl and Quigley, John M. (1996). Non-parametric hedonic housing

prices, *Housing Studies*, 11, 3, 373-385.

Matarazzo, B. (1997), L'Approccio dei Rough Set all'Analisi delle Decisioni, Proceedings XXI Annual Meeting A.M.A.S.E..S., Rome, 10-13 September.

McCluskey, W. and Anand, S. (1999). The application of intelligent hybrid techniques for the mass appraisal of residential properties, *Journal of Property Investment & Finance*, 17, 3, 218-239.

McCluskey, W., Deddis, W., McBurney, R.D., Mannis, A., and Borst, R. (1997). Interactive application of computer assisted mass appraisal and geographic information systems, *Journal of Property Valuation and Investment*, **15**, 448-465.

McGreal, S., Adair, A. McBurney, D. and Patterson, D.(1998). Neural networks: the prediction of residential valueds, *Journal of Property Valuation and Investment*, **16**, 16, 57-70.

McLean, D.G. and Mundy, B. (1998). The addition of contingent valuation and conjoint analysis to the required body of knowledge for the estimation of environmental damages, *Journal of Real Estate Practice and Education*, 1, 1, 1-19.

Miller, Norman G. (1982). Residential property hedonic pricing models: a review, *Research in Real Estate*, 2, 31-56.

Mok, Henry M. K., Chan, Patrick P. K., and Cho, Yiu-Sun. (1995). A hedonic price model for private properties in Hong Kong, *The Journal of Real Estate Finance and Economics*, 10, 1, 37-48.

Nguyen, N. and Cripps, Al. (2001). Predicting housing valued: a comparison of multiple regression analysis and artificial neural network, *Journal of Real Estate Research*, **22**, 3, 313-336.

Pace, R. Kelley, Barry, Ronald, Clapp, John M. and Rodriquez, Mauricio. (1998). Spatiotemporal autoregressive models of neighborhood effects, *The Journal of Real Estate Finance and Economics*, 17, 1, 15-33.

Pawlak, Z. (1982). Rough sets, *International Journal of Information and Computer Sciences*, **11**, 341-356.

Pawlak, Z. (1991). Rough Sets. Theoretical Aspects of Reasoning about Data, Kluwer Academic Publisher, Dordecht.

Pawlak Z. (1997). Rough set approach to knowledge-based decision support, *European Journal of Operational Research*, **99**, 48-57.

Robinson, R. (1979). Housing Economics and Public Policy, MacMillan, London.

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

Rossini, P.A. (1997). Application of artificial neural network to the valuation of residential property, 3rd Annual Pacific RIM Real Estate Society Conference, Palmerston North New Zealand, January.

Ruokolainen, A. (1999). A decision support system for investing in owner-occupied dwellings. Tampere University of Technology, Publications 256. Tampere.

Scott I. and Gronow S. (1990). Valuation expertise: its nature and application, *Journal of Valuation*, **8**, 4, Summer, 362-375.

Silverherz, J.D. (1936). The Assessment of Real Property in the United States, SP10 N.Y. State Tax Commission, (Albany: J.B.Lyon Company).

Stefanowski, J. and Tsoukias, A. (2000). Valuedd Tolerance and Decision Rules, in W.Ziarko and Y. Yao (eds), Proceedings of the RSCTC 2000 Conference, Banff, 180-187.

Tsoukiàs, A. and Vincke, Ph. (2000). A characterization of PQI interval orders, *to appear in Discrete Applied Mathematics*.

Ward, R., Weaver, J.R. and German, J.C. (1999). Improving CAMA models using geographic information systems/response surface analysis location factors, *Assessment Journal*, **6**, 30-38, January.

Worzala, E., Lenk, M. and Silva, A., (1995). An exploration of neural networks and its application to real estate valuation, *The Journal of Real Estate Research*, **10**, 2, 185-201.

Appendix 1

Descriptive Statistics

	Mean	Std. Deviation	N
PRICE	151727,7897	38676,2515	390
SQM	84,8026	25,0250	390
ROOMS	3,2846	,9850	390
DATE	39,4077	3,6912	390
YEAR	20,8615	7,2874	390

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,963	,927	,926	10487,7256

Analysis of Variance

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	539539527090,563	4	134884881772,641	1226,311	,000
	Residual	42347069454,197	385	109992388,193		
	Total	581886596544,759	389			

Coefficients

		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
Model		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	39975,805	6196,352		6,452	,000					
	SQM	1290,916	34,221	,835	37,723	,000	,956	,887	,519	,386	2,594
	ROOMS	6507,988	855,447	,166	7,608	,000	,808	,362	,105	,398	2,511
	DATE	-397,107	144,217	-,038	-2,754	,006	-,070	-,139	-,038	,998	1,002
	YEAR	-165,287	76,293	-,031	-2,166	,031	,251	-,110	-,030	,915	1,093

Collinearity Diagnostics

		Eigenvalued	Condition Index	Variance Proportions				
Model	Dimension			(Constant)	SQM	ROOMS	DATE	YEAR
1	1	4,818	1,000	,00	,00	,00	,00	,00
	2	8,829E-02	7,387	,00	,06	,09	,00	,62
	3	7,170E-02	8,197	,03	,04	,04	,04	,34
	4	1,801E-02	16,357	,00	,89	,87	,00	,01
	5	4,104E-03	34,264	,97	,00	,01	,96	,02

Appendix 2

With kind permission of Statistics Finland, a sample of 600 observations of single-family house transactions in the center of Helsinki, Finland, for the year 2001, has been obtained by Real Estate Market Observatory, 1st Faculty of Engineering, Technical University Politecnico di Bari.

Though, unfortunately, longitude and latitude, or other geographical information, has not been compiled, property characteristics include date of sale in months, number of rooms, year of construction, and price.

The sample has been divided into an in-sample of 390 property transactions for the MRA model and an out-of-sample of 210 property transactions, to test MRA and RST efficiency.

To apply RST to the first group of property transactions, a further sample of 70 property transactions have been selected from the in-sample group of 390 properties.