Elżbieta Jasińska *, Edward Preweda *, Jan Ruchel *

Modeling Transaction Prices of Properties Based on Qualitative and Quantitative Features **

1. Introduction

A property is a special type of commodity, which has features which can be divided into measurable – such as area and immeasurable – such as street or district. Such a division requires dual approach to model the transaction price, depending on the type of accepted attributes. The existing methods of testing the real estate market are divided into qualitative and quantitative methods. The first approach is based on sociological techniques, for example the relevance of various features of the property for potential buyers, and based on experience, knowledge and intuition, which allow to express opinion on the development of market phenomena, if they are determined by immeasurable factors. Among the quantitative methods, dominate statistical analysis and econometric models, which depend on modeling transaction prices and real estate values [2]. The methods used so far in Poland, for the real estate market analysis, are mostly based on assigning numerical values to qualitative characteristics. There is also used an approach, which separates the quality from quantitative traits. After the initial analysis is done, e.g. based on a ranking, the property value is corrected by quantitative methods.

The aim of this study is to extend the recently used methods on the real estate market by multi-dimensional analysis, allowing comparison of the impact of single variables, as well as the evaluation of these attributes, without assigning them arbitrary numerical values. The study was carried out by the C&RT (Classification and Regression Trees), which does not require scaling the attributes, which can describe by the quality scale. This proposal extends the existing research by taking into account the quantitative characteristics and at the same time qualitative ones (no need to assign numerical values to them.) This allows the introduction of location of the premises by the chosen street, which has been ignored so far.

* Faculty of Mining Surveying and Environmental Engineering, AGH University of Science and Technology, Krakow

** Article created as part of the statutory research in the Chair of the Geomatics No 11.11.150.006

The research was done on the basis of about 109 housing properties located in two selected districts of the City of Krakow. Those dwellings were sold between November 2008 and March 2009.

2. Data Analysis Method by C&RT

Studies conducted by classification trees allow to incorporate the attributes that are difficult to convert to a quantitative scale (the name of a district, precinct number, street name). It is an unquestionable advantage, because it is expected that the final decision about buying a property or the transaction price will depend on the characteristics of this type. It is possible to omit using numerical values for these characteristics without sacrificing the ones maintained in the analysis, by using the proposed model. The definition "classification tree" is used as a general term, which depending on how the variable is measured, enables to build discriminatory or regressive models. The discriminatory model assumes a qualitative discriminatory nature of the dependent variable, indicating its affinity to a particular class in the final node. The regression model (regression tree) comes up as a product of segmentation of the sample based on a dependent variable of a quantitative nature, in the final node, setting the average value and variance of this variable [1].

The process of identifying the rules characterizing the market price was carried out by setting the following rules:

- trimming variance – if any of the nodes in the descendants of the variance does not decrease, then this division is eliminated in the process of trimming the tree;
- maximum number of nodes: 1000;
- the minimum number of nodes: 5;
- the proper tree size was based on the cost of a 10-fold cross-validation (understood as a variance based on prediction of the continuous dependent variable); the tree with the best predictive ability was considered as a basis, while selecting attributes of price-setting behavior and researching the relationships between them and the price of the transaction.

In the graph created this way, the ranking of variable was established. The validity of the predictor, presented on it is the reciprocal of the total re-substitution cost in all nodes of the created tree. It is expressed in a scale of 0 to 1 (scaled so that its maximum is 1), which may be analogous to coefficients correlation [3], but the validity of the rankings cannot be determined whether the predictor influences the value of the dependent variable positively or negatively. It may also occur that the predictor, which has not been a criterion for final distribution of the tree, will get a high position in the ranking. It is possible, when such an attribute in most divisions was second, in the possibility of reducing the variance in the lymph-descendants. Despite that, ultimately it has not been placed on a graph, its predictive ability is higher than such an attribute, which "used the power of predictive" in the first divisions of the tree, and then was not essential for further segmentation.

3. Presentation of the Results

Use of the C&RT algorithm does not require conversion scale features of the property to the quantitative scale, which allowed the inclusion of the characteristics "Street" to the test, which was usually ignored. An example of such an analysis is presented in the selected districts and they are Lagiewniki-Borek Falecki and Debniki.

The process of creating a final model, on the basis of which proceeded the modeling of the transaction price of a real estate, started from the creation of the most sophisticated model. The next step was trimming the tree, which was based on the value of replacement cost and value of cross-validation. These values increase as the trim is being continued (for the next trees, the number of terminal nodes decreases, further tree presents a more general criteria for the division). The value of re-substitution costs provide the dependent variable value (close or equal to the average cost) in the chosen leaf. It is interesting that the cost of cross-validation for the tree number 8 (in the sample test) reaches a similar value to the first model, maintaining the general accuracy of increasing its value while the number of nodes is decreasing. This behavior confirms the aptness of this scheme, as the best final cut. Other trees provide less prognostic relevance in attempts (other sub samples *v*, drawn from the data), as shows figure 1.

The best division of a given node is the one which gives the greatest decrease in the cost of replacement. C&RT algorithm aims to allow the separation of high and low values of the dependent variable, which means that in a properly built model, to one node go the higher values and to the second node go values lower then the ones in the parent node [4].

On the base on the characteristics of a selected regression tree, features were ranked relatively to the ability of distributing real estates, as illustrated in figure 2. Basing on this schema it is possible to distinguish four characteristics, with coefficients higher than 0.8, "Rooms layout", "Street" at which the premises is located, "Standard" and "Neighborhood." These are mainly the characteristics of a qualitative nature, a special case of which is the "Street", it is the most difficult to express by numerical value. On the next position, the authors lined up: "Expenditure", "Surface", "Floor", "Year of Building", belonging of a "Basement" to the premises, and the later we have – "Communication Access".

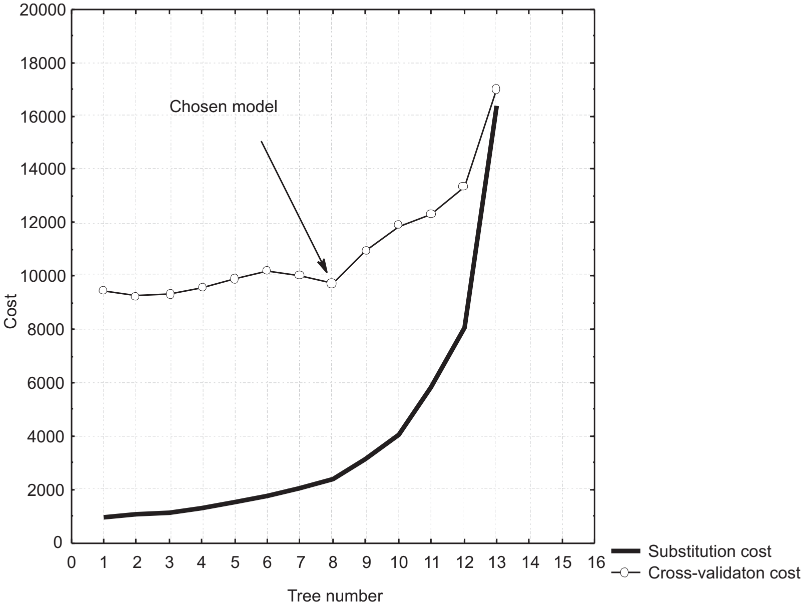

Fig. 1. Sequence chart of substitution costs and costs of cross-validation for the district Borek Falecki-Lagiewniki

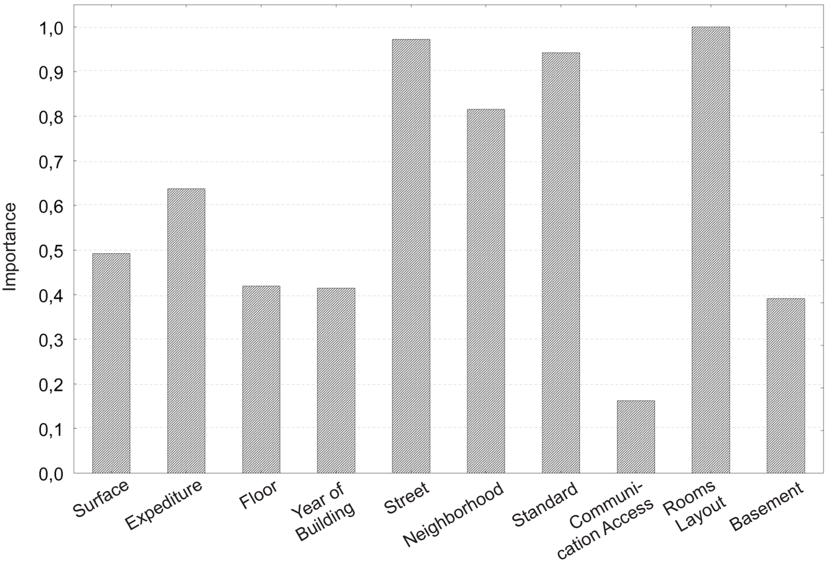

Fig. 2. The validity of the attributes for the district Borek Falecki-Lagiewniki

The analyzed area is characterized by a highly developed network of communication, both public (two tram depots) and roads (vehicle traffic). Features directly related to the location of apartments in the building and age of the building are not as important as the standard of the residence or functionality, as most buildings have been built in a similar time.

While analyzing the regression tree, a segmentation of 45 house properties in the district Lagiewniki-Borek Falecki was carried out. Based on figure 3 conditional sentences, assigning accommodations to one of the selected sub-groups, were created:

- If the property is located by one of the streets: Borsucza, Chmielna, Slupska, Sowia, and "Rooms layout" is "Very favorable" or " Favorable" the market value of that property is 6700 zł/m^2 ± 47 zł. The transaction price of the premises located by one of the mentioned streets, witch "Average" or "Disadvantage" "Rooms layout" "is 6550 zł/m^2 ± 237 zł.
- If the property is located by one of the streets: Brozka or Cegielniana, and its "Standard" is "High" the transaction price is 7000 zł/m^2 ± 82 zł. The analyzed database has three such properties.
- If the apartment is located by the street: Borkowskie Błonia, Brozka or Zdunow, and its "Standard" is not "High" its transaction price is 6728 zł/m^2 ± 45 zł.

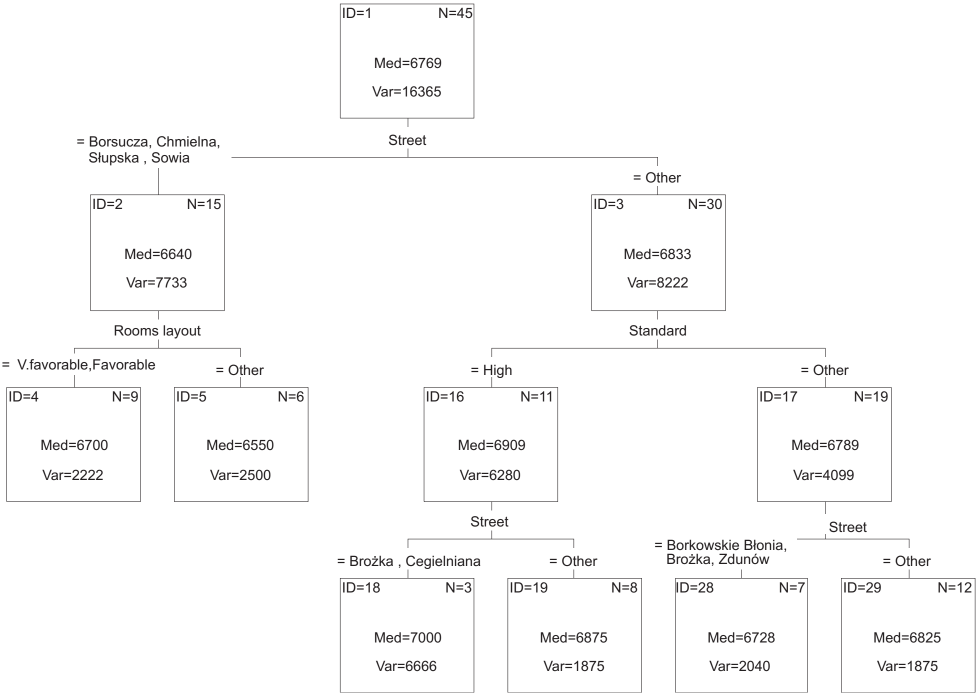

Fig. 3. Regression tree system for real estate housing for the district Lagiewniki-Borek Falecki

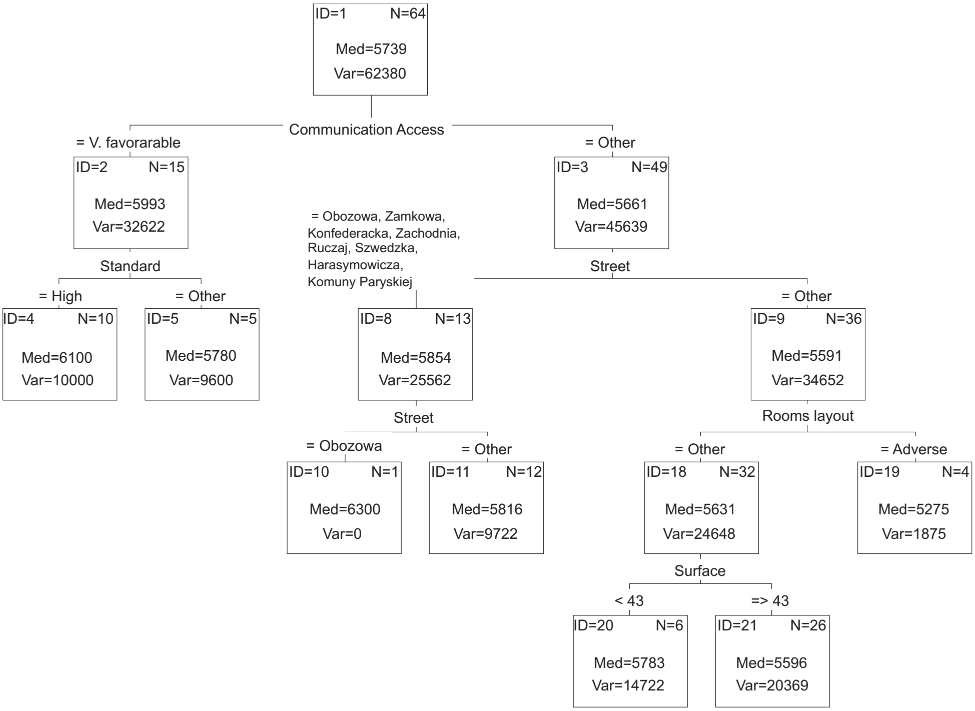

Fig. 4. Regression tree for the Debniki district

Based on the diagram shown in figure 3, it can be seen that node 3 is split into nodes 16 and 17, and these nodes into 18 and 19, and later into 28 and 29. This numeration is a result of the most extensive model. The presented tree consists of 5 nodes and it is split into 6 end nodes, and is the smallest node fulfilling the rule of a single standard deviation. While dividing the set, the feature "Street" was included three times, which in predictors is ranked on the second place, and the feature "Rooms layout" only once, which is the most important. On this basis, it can be concluded that the next considered for the distribution of nodes feature was just room apartment layout. As it was previously mentioned, the ranking of predictors does not necessarily have to coincide with the graph tree, because it is a summary of all such lists, created during each division.

The graph presented in figure 4 can be presented in the form of a few conditional sentences such as:

- If the housing property is characterized by "Very favorable" "Communication Access", and its "Standard" is "High", the transaction price of the property is 6100 zł/m^2 ± 100 zł – in the database ten such properties exists. The transaction price of the same property, with the same "Communication Access" but with a lower "Standard" stands at 5780 zł/m^2 ± 98 zł. Database accumulates five such properties.

- If “Communication Access” to this apartment is “Favorable”, “Average” or “Adverse”, and the residence is located by one of the streets: Zamkowa, Konfederacka, Zachodnia, Szwedzka, Harasymowicza, Komuny Paryskiej or Ruczaj, its transaction unit price is 5816 zł/m^2 ± 99 zł. Moreover, one can distinguish a property situated by Obozowa street, which cost 6300 zł/m^2.
- A dwelling located by other than the streets mentioned above, and its “Room layout” is “Adverse” the unit price is on the level 5275 zł/m^2 ± 43 zł. If the “Rooms layout” is “Average”, “Favorable” or “Very Favorable”, and “Surface” is smaller than 43 m^2, its transaction unit price is 5783zł/m^2 ± 121 zł. If its “Surface” comes at least to 43 m^2, the price of 1 m^2 can be estimate as 5596 zł ± 143 zł.

Analyzing the ranking of predictors for each stage of division, it is possible to rank the attributes of the property considering the ability to create homogeneous groups in terms of the transaction price. The graph shown in figure 5 systematizes the attributes forming the transaction price of the real estate. Again the qualities outweigh the measurable characteristics of the premises. The main criteria while selecting a dwelling are: “Street”, “Room layout”, “Neighborhood” and “Standard”. Further items are: “Expenditure”, “Surface”, “Floor”, “Year of Buiding”, “Basement” was ranged at the last level.

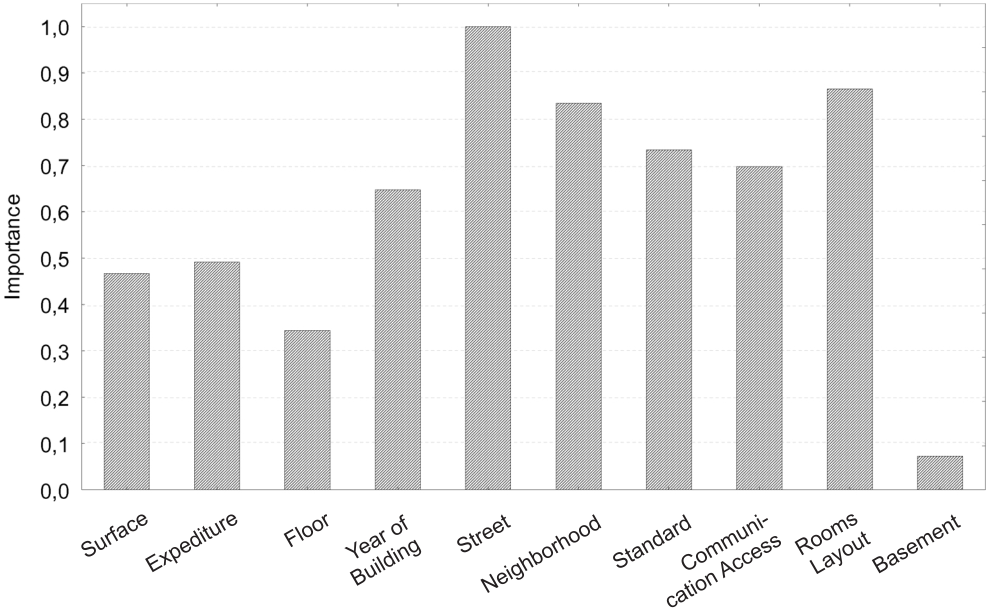

Fig. 5. The validity of attributes for the district Debniki

The analyzed district occupies an extensive area and combines the areas of pre-war buildings and a newly built ones. Dwellings also differ substantially. It is therefore difficult to speak about the same technologies or the same creation period. Because of diversification of the market, it is difficult to establish a clear criteria division, simultaneously it is extremely important for investors because it allows them to look for valuable areas for further investments. What is interesting, the territorial division (because of its location at a particular street) does not create a closed cluster, but it is a group of selected streets, from Szwedzka Street to Obozowa Street.

Combining predictor ranking, with the tree graph, can give us a searching analysis of the controlling processes at the local market. In this study, the features of a qualitative nature ("Street", "Rooms layout", or "Neighborhood") are the most important criteria for the evolution of property prices. Apart from determining which attributes are especially valuable to buyers, sets of homogeneous properties in terms of unit price can be allocated.

4. Conclusion

Simultaneous consideration of the qualitative and quantitative features in research of the real estate market is a challenge for analysts, that is why we should pay attention to considerating concerning regression trees.

The introduction of the feature "Street" – as an attribute specifying the location of the property has made it possible to deepen the analysis. This feature is in the forefront of established predictor rankings, however, because of its nominal nature, it cannot be used in classical studies. Because such analysis are limited to an area of the one district, where the address of the real estate is not distinguished, the only criterion for estimating the property is the neighborhood expressed in a numerical scale.

In addition, it is worth noting that the qualitative features are on the first places in the ranking features that affect the market price of the property. Finding a place with a similar surfaces in the same neighborhood is not a problem, it is much more difficult narrow down the search to the same street and a similar standard of finish. Therefore, C&RT models can be successfully used as a tool helping at work real estate experts, as they enable the selection of similar properties, even in very complex sets.

Also other sectors related with the real estate market, such as banks, developers and real estate agencies could define the attractiveness of new investments on the basis of C&RT trees, or estimate how to modernize the existing properties.

References

[1] Breiman L. et al.: *Classification and Regression Trees*. Chapman & Hill/CRC, New York 1998.

[2] Cellmer R.: *Zasady i metody analizy elementów składowych rynku nieruchomości.* Olsztyn 1999.

[3] Czaja J., Preweda E.: *Analiza statystyczna zmiennej losowej wielowymiarowej w aspekcie korelacji i predykcji*. Geodezja, T. 6, z. 2, Kraków, 2000, pp. 129–145.

[4] Fayyad U.M, Piatetsky-Shapiro G., Smyth P.: *Advances in Knowledge Discovery and Data Mining*. AAAI Press, 1996.

[5] Kucharska-Stasiak E. (red.): *Międzynarodowe Standardy Wyceny* (wyd. polskie). Polska Federacja Stowarzyszeń Rzeczoznawców Majątkowych, Warszawa 2005.

[6] Kafkowski L.: *Rynek nieruchomości w Polsce.* Tweeger, Warszawa 2003.

[7] *Statistica (data analysis software system), version 8.0*. StatSoft, 2007 [on-line:] www.statsoft.com.