Applying a CART-based approach for the diagnostics of mass appraisal models

Evgeny Antipov and Elena Pokryshevskaya

Higher School of Economics, Higher School of Economics

1. December 2010

Applying a CART-based approach for the diagnostics of mass appraisal models

Evgeny Antipov, The State University – Higher School of Economics (Russia)

Elena Pokryshevskaya, The State University – Higher School of Economics (Russia)

Abstract

In this paper an approach for automatic detection of segments where a regression model significantly underperforms and for detecting segments with systematically under- or overestimated prediction is introduced. This segmentational approach is applicable to various expert systems including, but not limited to, those used for the mass appraisal. The proposed approach may be useful for various regression analysis applications, especially those with strong heteroscedasticity. It helps to reveal segments for which separate models or appraiser assistance are desirable. The segmentational approach has been applied to a mass appraisal model based on the Random Forest algorithm.

Key words:

CART, model diagnostics, mass appraisal, real estate, Random forest, heteroscedasticity

1. Introduction

According to International Association of Assessing Officers mass appraisal is the process of valuing a group of properties as of a given date using common data, standardized methods, and statistical testing (Eckert, 1990). Expert systems for mass appraisal allow determining the taxable value of a real estate object. The growing number and quality of websites with real estate prices and characteristics help researchers to develop formal models for mass appraisal.

Various methods have been used for real estate mass appraisal, among which parametric regression analysis is the traditional choice (Ball, 1973; Lentz and Wang, 1998; Miller, 1982; Laakso, 1997; Theriault et al., 2005; Kang and Reichert, 1991; McCluskey and Anand, 1999). In some studies nonparametric regressions have been applied successfully (e. g., Filho and Bin, 2005). Among machine learning methods the most commonly used are neural networks (e. g., Verkooijen, 1996; Pace, 1995; McCluskey and Anand, 1999; Verikas et al., 2002; Worzala et al., 1995; Ge et al., 2003; Curry et al., 2002; Kauko, 2003; Kauko et al., 2002; Liu et al., 2006; Selim, 2009). At the beginning of 1990s several authors revealed some problems with neural networks (Worzala *et al.*, 1995). For example, the average absolute error varied significantly depending on the algorithm used in different software packages, i. e. results are often unstable (Worzala *et al.*, 1995; Kontrimas and Verikas, 2010). On the other hand, Nguyen and Cripps (2001) showed that neural networks are effective in the case of large heterogeneous datasets. Other methods, reported to be effective, include, but are not limited to, k nearest neighbors (McCluskey and Anand, 1999), regression trees (Fan et al., 2006) and fuzzy logic techniques (Bagnoli and Smith, 1998; Lee et al., 2003; Theriault et al., 2005).

The existing literature pays little attention to model diagnostics. As a rule, to evaluate model quality aggregated diagnostic indicators are used (coefficient of determination, mean average percentage error etc.), while there are virtually no tools which can be used to reveal problem segments of observations and improve models based on this knowledge. Without such diagnostics, model quality is questionable, since it may give a much higher than average error when objects from particular segments are under consideration. That is why the goal of our study is to suggest a segmentational approach for the diagnostics of mass appraisal models quality.

2. Methodology

2.1. Measures of valuation accuracy

We have chosen the accuracy measures, which allow comparing valuation quality independent of the methodology used and which comply with the existing standards on automated expert systems evaluation.

Average Sales ratio (SR) with a confidence interval

The numerator of the sales ratio for a particular transaction would be the estimated value generated from the model, while the denominator would be the sale price. The 95% confidence interval must overlap 0.9-1.1 range according to international standards (International Association of Assessing Officers, 2003). In our study we use bootstrap confidence intervals because the distribution of SR is not normal.

Mean average percentage error (MAPE)

$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{Y}_i - Y_i}{Y_i}\cdot 100$, where Y_i is the observed and $\widehat{Y}_i$ is the predicted value of object *i*. MAPE is easy to interpret and reflects the accuracy of the model.

Coefficient of dispersion (COD)

COD measures the average percentage deviation of SR from its median value. It is often considered to be the most useful measure of sales ratio's variability, because its interpretation is not dependent on the normality assumption. In accordance with international standards COD of 5-20% is acceptable (International Association of Assessing Officers, 2003).

2.2. A segmentational approach for model accuracy diagnostics

Besides average indicators of prediction accuracy, the homogeneity of valuation quality across different segments is important, especially in the context of mass appraisal. If there are segments in which the predicted values are systematically over- or underestimated, the model cannot be considered satisfactory. This is also true in the case of the segments, where prediction errors are significantly higher than average, which also puts tax payers in unequal position. For problem segments it is reasonable to apply appraiser assisted AVMs, which still simplify experts' job, but are controlled by them.

Despite active development of statistical methods, there are hardly any universal and easy-to-use approaches to diagnose and correct the heterogeneity of valuation quality. We propose an approach to revealing segments with high and low prediction error in the context of mass appraisal problem.

1. Let Y_i be the observed market value for object *i*, $\widehat{Y}_i$ – the value predicted using some data analysis method. Then $PE_i = \frac{\left|\widehat{Y}_i - Y_i\right|}{Y_i}\cdot 100$ is the percentage error of prediction for observation *i*.
2. On the training sample build the decision tree, using the CART algorithm with PE_i as a dependent variable and with all the predictors used for valuation purposes as the explanatory variables. The tree splits the sample into segments, differing by MAPE. We suggest setting a reasonably large minimum number of cases per node (at least several hundred).
3. If the regression tree does not reveal significantly different segments, then either the accuracy of the model may be considered homogeneous or another regression tree algorithm can be tried instead of CART. We do not recommend increasing the significance level (I type error), since in order to transfer our conclusions to the testing sample, we should be confident enough in the regularity of the revealed differences.
4. If the regression tree reveals significantly different segments, then acceptability of MAPE in each segment should be considered. In the case of high MAPE in some segments, appraiser assistance may be required for objects belonging to those segments. Building separate models for different segments may also lead to an increased overall accuracy.

Revealing segments with systematically under- and overestimated sales prices requires repeating steps 1 – 4 of the previous procedure using SR_i instead of PE_i.

It should be noted, that the proposed tree-based approach can be used for diagnostics and correction of the prediction quality in various regression problems in the presence of a reasonably large training sample. Instead of a percentage error, an absolute error or squared residuals may be used depending on a researcher's purpose. The latter case, for instance, gives a tool for heteroscedasticity diagnostics, capable not only of detecting heteroscedasticity of any type, but also of describing the detected segments, which gives our approach a competitive advantage compared to standard econometric tests.

3. Empirical analysis

3.1. Data

The dataset is based on the largest in Saint-Petersburg (Russia) real estate catalog "Real estate bulletin" (www.bn.ru). The content of the bulletin is moderated by its publisher, which increases the data quality.

Our initial sample consisted of 2848 two-room apartments, sold in the spring of 2010 in Saint-Petersburg. In order to record prices closest to the actual sales prices, we collected the last values, which appeared in the bulletin for each object. We have noticed, however, that these values are usually equal to the initial prices. A scatter diagram ("total area - apartment price") helped us to exclude three likely outliers. Thus the empirical analysis is based on the objects with the area of up to 160 m^2 and the price of up to 30 million rubles. Such a range is still very wide due to the heterogeneity of apartments in the city, which makes the valuation difficult. The final version of the dataset was split into the training sample (2695 observations) and the testing sample (150 observations).

Each object is characterized by the following variables:

1. Apartment price in thousand rubles (price)
2. Price per square meter in thousand rubles (price_per_meter)
3. Total area of the apartment in square meters (total_area)
4. Living area in square meters (living_area)
5. The area of the first room in square meters (room1_area)
6. The area of the second room in square meters (room2_area)
7. Herfindahl index for room areas:

$$inequality1 = \left(\left(\frac{room1\_area}{living\_area}\right)\cdot 100\right)^2 + \left(\left(\frac{room2\_area}{living\_area}\right)\cdot 100\right)^2$$

8. Absolute percentage difference between room areas:

$$inequality2 = \frac{\left(max\left(room1\_area, room2\_area\right) - min\left(room1\_area, room2\_area\right)\right)}{min\left(room1\_area, room2\_area\right)}\cdot 100$$

9. Kitchen area in square meters (kitchen_area)
10. Bathroom unit type (bathroom_unit): 1="no bath/shower in the kitchen/bath in the kitchen/shower only"; 2="the bathroom unit including the toilet"; 3="the toilet separate from the bathroom"; 4="2 or more bathroom units"

11. Telephone availability (telephone): 0="not available"; 1="available"
12. The floor, on which the apartment is situated (floor)
13. Number of floors in the house (number_of_floors)
14. House type (house_type): 24 categories
15. Distance from the house to the nearest underground station (distance_from_underground): 0="1-5 minutes on foot"; 1="6-10 minutes on foot"; 2="11-15 minutes on foot or 1-5 minutes by bus"; 3="16-20 minutes on foot or 6-10 minutes by bus"; 4="21-25 minutes on foot or 11-15 minutes by bus"; 5="16-20 minutes by bus"; 6="more than 20 minutes by bus"
16. Time to the city center by underground (time_to_downtown)
17. District (district): 13 categories

Descriptive statistics for quantitative variables are given in Table 1.

Table 1

Descriptive statistics for quantitative variables

	Number of valid cases	Min	Max	Mean	Std. Deviation	Coefficient of variation, %
price	2695	1500.0	26500.0	4826.4	2456.1	50.9
price_per_meter	2695	29.4	375.0	82.1	26.7	32.5
total_area	2695	22.0	156.0	57.7	13.9	24.0
living_area	1697	15.0	75.0	33.0	6.3	19.1
room1_area	2020	7.0	75.0	19.1	6.1	32.1
room2_area	1905	6.0	48.0	14.8	4.1	27.9
kitchen_area	1623	4.0	50.0	10.6	5.0	47.7
floor	2652	1.0	25.0	5.1	3.9	75.4
number_of_floors	2688	2.0	27.0	9.5	5.3	56.4
time_to_downtown	2695	0.0	6.0	1.6	1.3	82.6

3.2. The diagnostics of the Random forest model accuracy using a segmentational approach

Using the indicators COD and MAPE, it is difficult to give recommendations on how to increase accuracy homogeneity across different segments and decrease prediction error. That is why we use the approach for homogeneity of model accuracy diagnostics introduced in Subsection 2.2. Using this approach we will make the diagnostics of Random forest predictions (we use Random Forest predictions because they appeared to be the best in our comparison study, the results of which are not going to be covered in this paper).

To begin with, we build a regression tree that will allow revealing apartment segments which differ the most in the average MAPE. As we want to pick out the most stable segments, we set the minimum number of observations in a node equal to 300.

The diagnostics (see Table 2) showed that the pooled model based on all observations of the training sample gives an average error of less than 9.8% for apartments with area of below 61.5 sq. meters, while MAPE is 19.4% for apartments

with greater area, among which MAPE for districts 4, 5, 6, 9, 11, 12 is 12.9% and for other districts – 23.6%. Hence we can recommend the correction of valuations in the third segment with the help of experts or by developing another model for this segment. Our experience showed that the separate model building for this segment did not decrease the error. This can be partly explained by the fact that transactions of relatively big apartments in these districts have many features that are hard to take into account in mass appraisal models: therefore, the error can hardly be significantly reduced by applying some other method without adding other variables to the dataset. The segment that requires special attention accounts to approximately 18% of the market. It is easy to ascertain that the revealed regularity is stable and the differences among the obtained segments appear on the test sample, as well as on the training sample.

Table 2

Segments with different MAPE revealed by CART algorithm

Segment number	Segment description	MAPE (training sample)	MAPE (test sample)	% of the market (training sample)	% of the market (test sample)
1	Total area≤61.5	9.783	12.364	69.8	69.3
2	Total area>61.5	19.401	20.498	30.2	30.7
3	Total area>61.5 and districts 4, 5, 6, 9, 11, 12	12.852	14.423	11.9	10.0
4	Total area>61.5 and districts 1, 2, 3, 7, 8, 10, 13	23.643	23.438	18.3	20.7
Total sample		12.688	14.859	100	100

In order to verify if there are segments with systematically under- and overvalued objects, we build a similar tree with SR as a dependent variable (see Table 3). As a result of our analysis, 2 segments were revealed that are likely to systematically overestimate the predicted price compared to real sales prices (SR for one of the segments is 1.018, for the other – 1.073).

Table 3

Segments with different SR revealed by CART algorithm (training sample)

Segment number	Segment description	MAPE	% of the market
1	Districts 3, 4, 5, 6, 8, 9, 11, 12	1.018	68.5
2	Districts 1, 2, 7, 10, 13	1.073	31.5

Total sample	1.035	100

We calculated bootstrap confidence intervals for the average SR in each segment (see Table 4).

Table 4

Confidence intervals for the average SR

Segment	Point estimate of the average SR	Lower bound of the average SR confidence interval	Upper bound of the average SR confidence interval
1	1.018	1.011	1.025
2	1.073	1.055	1.088

We use the lower and the upper bound of the confidence interval as well as the point estimate of the average SR as correction coefficients. If values predicted by Random forest are divided by the lower bound of the confidence interval in the corresponding segment, MAPE was 14.06% in the test sample (reduced by 0.80 percentage points); in the case of using the point estimate of the average SR as the correction coefficient, MAPE decreased to 13.98% (reduced by 0.88 percentage points); finally, when the upper bound of the confidence interval was used, MAPE decreased to 13.95% (reduced by 0.91 percentage points). Taking into account already relatively low error provided by the Random forest algorithm, the obtained improvements should be considered quite substantial. Meanwhile, we suppose that using lower bound of 95%-confidence interval is the most conservative and safe variant.

While the effectiveness of the proposed correction method requires further inquiry, the segmentational approach itself, which allows revealing problem segments, undoubtedly helps carry out substantially deeper diagnostics of automated appraisal systems in comparison with calculating just a few integral accuracy indicators for the whole sample of objects.

4. Conclusion and future research

In our study we have proposed and applied the segmentational approach to the model accuracy diagnostics that, in contrast to a number of widely used integral indicators, allows not only to evaluate the overall quality of a model, but to pick out the market segments which differ the most in the average MAPE and to detect segments with systematically under- and overvalued predictions. The proposed approaches may be useful for various regression analysis applications, especially those with strong heteroscedasticity.

A deeper diagnostics using the proposed segmentational diagnostic approach has been conducted for the Random forest model built using Saint-Petersburg residential apartments dataset. The diagnostics showed that the pooled model based on all observations of the training sample gives an average error of less than 9.8% for apartments with area under 61.5 sq. meters, while MAPE is 19.4% for apartments with greater area, among which MAPE for districts 4, 5, 6, 9, 11, 12 is 12.9% and for other districts – 23.6%. Hence we can recommend the correction of valuations in the problem segment with the help of experts or by developing another model for

this segment. The diagnostics of systematically under- and overestimated values and calculating bootstrap confidence intervals for the average SR in the segments revealed by the procedure allowed to implement the correction coefficients and reduce MAPE in the test sample by 0.80-0.91 percentage points depending on the choice of correction coefficient.

The use of correction coefficients for segments with systematically under- or overestimated predicted values of the dependent variable seems to be very promising, however it requires a deeper theoretical and empirical study of the entailed consequences. We also plan to study how building separate models for underperforming segments may help reduce prediction error.

References

1. Bagnoli, C., & Smith, H. (1998). The theory of fuzzy logic and its application to real estate valuation. *Journal of Real Estate Research, 16*, 169–199.
2. Ball, M. J. (1973). Recent empirical work on the determinants of relative house prices. *Urban Studies, 10*, 213–233.
3. Curry, B., Morgan, P., & Silver, M. (2002). Neural networks and nonlinear statistical methods: An application to the modelling of price quality relationships. *Computers & Operations Research, 29*, 951–969.
4. Eckert, J. K. (1990). *Property Appraisal and Assessment Administration*. International Association of Assessing Officers, Chicago, IL.
5. Fan, G., Ong, Z. S. E., & Koh, H. C. (2006). Determinants of house price: A decision tree approach. *Urban Studies, 43(12)*, 2301–2315.
6. Filho, C. M., & Bin, O. (2005). Estimation of hedonic price functions via additive nonparametric regression. *Empirical Economics, 30(1)*, 93–114.
7. Ge, J. X., Runeson, G., & Lam, K. C. (2003). Forecasting Hong Kong housing prices: An artificial neural network approach. *International Conference on Methodologies in Housing Research*, Stockholm, Sweden.
8. International Association of Assessing Officers (2003). Standard on automated valuation models (AVMs). www.iaao.org. Approved September, 2003.
9. Kang, H.-B., & Reichert, A. K. (1991). An empirical analysis of hedonic regression and grid-adjustment techniques in real estate appraisal. *AREUEA Journal, 19(1)*, 70–91.
10. Kauko, T. (2003). On current neural network applications involving spatial modelling of property prices. *Journal of Housing and the Built Environment, 18(2)*, 159–181.
11. Kauko, T., Hooimeijer, P., & Hakfoort, J. (2002). Capturing housing market segmentation: An alternative approach based on neural network modeling. *Housing Studies, 17(6)*, 875–894.
12. Kontrimas, V., & Verikas, A. (2010). The mass appraisal of the real estate by computational intelligence. *Applied Soft Computing*, In Press.
13. Laakso, S. (1997). Urban housing prices and the demand for housing characteristics. The Research Institute of the Finnish Economy (ETLA) A 27, Helsinki.
14. Lee, Y.-L., Yeh, K.-Y., & Hsu, K.-C. (2003). Fair evaluation of real estate value in urban area via fuzzy theory. *10th ERES Conference*, Helsinki, Finland, 10–13 June.

15. Lentz, G. H., & Wang, K. (1998). Residential appraisal and the lending process: A survey of issues. *Journal of Real Estate Research, 15(1/2)*, 11–39.
16. Liu, J., Zhang, G. X. L., & Wu, W. P. (2006). Application of fuzzy neural network for real estate prediction. *LNCS, 3973*, 1187–1191.
17. McCluskey, W.J., & Anand, S. (1999). The application of intelligent hybrid techniques for the mass appraisal of residential properties. *Journal of Property Investment and Finance, 17(3)*, 218–238.
18. Miller, N. G. (1982). Residential property hedonic pricing models: A review. In: Sirmans C.F. (ed.), Urban Housing Markets and Property Valuation. Research in Real Estate, Vol. 2, Jai Press Inc., Greenwich, CT, pp. 31–56.
19. Nguyen, N., & Cripps, A. (2001). Predicting housing value: A comparison of multiple regression analysis and artificial neural networks. *Journal of Real Estate Research, 22(3)*, 313–336.
20. Pace, R. K. (1995). Parametric, semiparametric, and nonparametric estimation of characteristic values within mass assessment and hedonic pricing models. *Journal of Real Estate Finance and Economics, 11*, 195–217.
21. Selim, H. (2009). Determinants of house prices in Turkey: Hedonic regression versus artificial neural network. *Expert Systems with Applications, 36*, 2843–2852.
22. Theriault, M., Des Rosers, F., & Joerin, F. (2005). Modelling accessibility to urban services using fuzzy logic. A comparative analysis of two methods. *Journal of Property Investment & Finance, 23(1)*, 22–54.
23. Verikas, A., Lipnickas, A., & Malmqvist, K. (2002). Selecting neural networks for a committee decision. *International Journal of Neural Systems, 12(5)*, 351–362.
24. Verkooijen, W. J. H. (1996). Neural networks in economic modelling. Doctoral dissertation, Tilburg University, Center for Economic Research, 205 pp.
25. Worzala, E., Lenk, M., & Silva, A. (1995). An exploration of neural networks and its application to real estate valuation. *Journal of Real Estate Research, 10(2)*, 185–201.