Expert Systems with Applications 39 (2012) 1772–1778

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics

Evgeny A. Antipov *, Elena B. Pokryshevskaya

Higher School of Economics, Faculty of Economics, Sedova St. 55/2, Saint-Petersburg 192171, Russia
The Center for Business Analysis, Prospekt Aviakonstruktorov 20-1-90, Saint-Petersburg 197373, Russia

ARTICLE INFO

Keywords:
Random forest
Mass appraisal
CART
Model diagnostics
Real estate
Automatic valuation model

ABSTRACT

To the best knowledge of authors, the use of Random forest as a potential technique for residential estate mass appraisal has been attempted for the first time. In the empirical study using data on residential apartments the method performed better than such techniques as CHAID, CART, KNN, multiple regression analysis, Artificial Neural Networks (MLP and RBF) and Boosted Trees. An approach for automatic detection of segments where a 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.

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; Kang & Reichert, 1991; Laakso, 1997; Lentz & Wang, 1998; McCluskey & Anand, 1999; Miller, 1982; Theriault, Des Rosers, & Joerin, 2005). In some studies nonparametric regressions have been applied successfully (e.g., Filho & Bin, 2005). Among machine learning methods the most commonly used are neural networks (e.g., Curry, Morgan, & Silver, 2002; Ge, Runeson, & Lam, 2003; Kauko, 2003; Kauko, Hooimeijer, & Hakfoort, 2002; Liu, Zhang, & Wu, 2006; McCluskey & Anand, 1999; Pace, 1995; Selim, 2009; Verikas, Lipnickas, & Malmqvist, 2002; Verkooijen, 1996; Worzala, Lenk, & Silva, 1995). 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 (Kontrimas & Verikas, 2011; Worzala et al., 1995). 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 & Anand, 1999), regression trees (Fan, Ong, & Koh, 2006) and fuzzy logic techniques (Bagnoli & Smith, 1998; Lee, Yeh, & Hsu, 2003; Theriault et al., 2005).

It should be noted that there is a lack of studies, which compare a sufficiently large number of machine learning algorithms for mass appraisal, and, as far as we are concerned, there are no studies where such a modern method as Random forest is used. The existing literature also 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 goals of our study are: (1) to justify the use of Random forest for mass appraisal and empirically compare it with 9 other methods; (2) to develop a segmentational approach for the diagnostics of mass appraisal models quality.

2. Methodology

2.1. Expected benefits of using Random forest for residential estate mass appraisal

Random forest (Breiman, 2001) is a machine learning algorithm used mainly for classification problems, which can be applied to

* Corresponding author at: Higher School of Economics, Faculty of Economics, Sedova St. 55/2, Saint-Petersburg 192171, Russia. Tel.: +7 906 269 26 09.
E-mail addresses: eugene.antipov@gmail.com (E.A. Antipov), e.pokryshevskaya@gmail.com (E.B. Pokryshevskaya).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.08.077

regression tasks as well. A Random forest is in fact a special type of simple regression trees ensemble, which gives a prediction based on the majority voting (the case of classification) or averaging (the case of regression) predictions made by each tree in the ensemble using some input data. All the trees of the ensemble are built independently according to the following algorithm. Let N be the size of the learning sample and M – the total number of predictors. A subset of $m < M$ randomly chosen predictors is used to grow each tree on a bootstrap sample of the training data. For each of the bootstrap samples, an unpruned regression tree is grown, with the following modification: at each node, rather than choosing the best split among all predictors, the best split among a random sample of m variables is made. After a large number of trees are generated, predictions are averaged over the different trees.

Despite the large size of created models, we believe Random forest may become one of the most appropriate techniques for mass appraisal due to the following reasons:

1. Good results in comparative studies (mainly classification problems), comparable with Support Vector Machines (SVM) and boosting and often better than those of neural networks (e.g., Caruana & Niculescu-Mizil, 2006). It should be noted, however, that little is known about the performance of Random forest in regression problems.
2. Successfully deals with categorical variables with lots of levels. For instance, in the case of multiple regression or neural networks, a large number of qualitative variables leads to an increased number of estimated parameters, which usually results in overfitting. In Random forest a nominal variable with k categories is recoded into $k-1$ dichotomous ones, only a fraction of which is used in building a tree. This helps to avoid most problems, caused by the large number of categories and makes Random forest an especially good technique for tasks with many categorical variables, such as mass appraisal, where there are such non-numeric variables as district, house type, bathroom unit type, etc.
3. Adequately works with missing data. If some data is missing for an observation, the prediction is made based on the part of the tree which had already been built. Therefore, there is no need in excluding any observations or imputing missing values.
4. Thanks to bagging the method is robust to outliers (they seldom appear in bootstrap samples and their influence is reduced).
5. In contrast to single regression trees, the prediction for each object is a unique number, rather than one of discrete values, derived using a set of rules.
6. Since the method is based on regression trees, it allows for nonlinear links and the unsteadiness of variable influence across different segments.
7. The method does not require a detailed model specification, which delivers from accounting for differences in the sets of pricing determinants in different areas.
8. Predictions for new observations are in the same range as already observed ones, which prevents significant overestimation or underestimation of real estate objects.
9. There is a way to measure factor importance by finding the average marginal reduction in the residual sum of squares for each explanatory variable.

Thus Random forest is expected to avoid fallacies of many other methods, commonly used for mass appraisal.

2.2. Formal criteria for model comparison

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

2.2.1. 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.

2.2.2. 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).

2.2.3. Mean average percentage error (MAPE)

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

2.3. 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, $\hat{Y}_i$ – the value predicted using some data analysis method. Then $PE_i = \frac{|\hat{Y}_i - Y_i|}{Y_i}\cdot 100$ 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:

$$inequality2 = \frac{(max(room1.area, room2.area) - min(room1.area, room2.area))}{min(room1.area, room2.area} \cdot 100$$

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 min on foot"; 1 = "6–10 min on foot"; 2 = "11–15 min on foot or 1–5 min by bus"; 3 = "16–20 min on foot or 6–10 min by bus"; 4 = "21–25 min on foot or 11–15 min by bus"; 5 = "16–20 min by bus"; 6 = "more than 20 min 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.

The following features of the data, which will obviously influence the comparative performance of different methods, have been revealed:

1. **Lots of missing values (no data on one or several characteristics for about two thirds of observations)**
   That is why, valuation methods should be able to cope with such a large number of missing values.
2. **Strongly asymmetric distribution of most quantitative variables**
   Due to the heterogeneity of the real estate market in the area, most distributions are positively skewed, so methods should be robust to the violation of normality assumption, which is often used in traditional econometric approaches.
3. **The presence of rare characteristics**
   This problem has been partly solved by merging some categories into one (e.g., we assume that apartments with no bath and those with the shower in the kitchen should cost the same). However, for some other variables a priori assumptions would be too strong. Therefore, the method should be able to merge some categories automatically. We address a related problem further in the text, discussing categorical variables with many levels.
4. **The presence of categorical variables with many levels**
   The variables "district", "house type", "distance from the house to the nearest underground station", "bathroom unit type" and "telephone availability" have 13, 24, 6, 4 and 2 levels correspondingly, which significantly increases the number of estimated parameters, if we use, for example, multiple regression or neural networks.
5. **The heterogeneity of residential apartments market**
   In Saint-Petersburg there are various types of apartment houses built in different decades of the 18th, 19th, 20th and 21st centuries. The marginal effects of characteristics are likely to differ across these segments. That is why standard regression analysis is not a very appropriate modeling tool. However methods based on regression trees can account for such a heterogeneous influence.
6. **Heteroscedasticity**
   Heteroscedasticity essentially results in increasing variance of apartment price when the total area increases. We suppose that a two-step procedure can partly solve this problem: first, the price per meter is predicted, after that the overall price is calculated. We believe that such procedure usually better reflects the logic behind real estate pricing: for instance, the location of an apartment in a prestigious district adds a constant value not to the overall price of the apartment, but more likely to its price per meter.
7. **The occurrence of non-typical transactions**
   Demand and supply are very limited in some segments, while those segments are still an important part of the market. Thus the utilized methods should be robust to non-typical transactions, which appear in the database.

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

Taking into account Random forest features described in the methodological section of the paper, we expect it to be among the most competitive methods for mass appraisal, since it can successfully cope with most of the problems detected in the market data.

3.2. Algorithms comparison

We have conducted a comparative analysis of performance achieved by 10 algorithms (Multiple regression, CHAID, Exhaustive CHAID, CART, k-Nearest Neighbors (2 modifications), Multilayer Perceptron neural network (MLP) and Radial Basis Function neural network (RBF)), Boosted Trees and Random forest). As far as we know, the results of applying Boosted Trees and Random forest to the mass appraisal of residential apartments have not been reported by any researchers yet.

We used algorithms available in StatSoft STATISTICA 8 and SPSS 18. Below we mention some key settings we used. For all methods that require setting the significance level the standard probability of 0.05 was used. For stepwise regression we used the entry probability equal to 0.05 and the removal probability equal to 0.1. For CHAID, Exhaustive CHAID and CART the minimum size of nodes was set to 50 to avoid overfitting.

When applying k-Nearest Neighbors method (Euclidian distance) we used normalized values and used automatic feature selection; optimal number of neighbors between 1 and 200 was chosen automatically by minimizing the error in the learning sample; weights, reflecting the importance of characteristics were used. In the case of neural networks, Random forest and Boosted Trees the training sample was divided into the actual learning sample and the validation one in 4:1 proportion. Then the automatic search was conducted to minimize the validation sample error. We used hyperbolic tangent as an activation function in MLP, softmax – in RBF, identity – as an output layer function in RBF.

As for Random forest we have determined by trial and error that good results are achieved when 10 out of 13 predictors are used for building each tree on the basis of a 70% bootstrap sample. Such algorithm settings are also in compliance with our intuitive beliefs about the ensembling in the case of the residential apartments mass appraisal. It is worth mentioning that small deviations from the above mentioned specification did not influence the prediction error significantly, which positively characterizes the robustness of Random forest to small changes in the parameters of the algorithm. In the case of boosting, each tree was the simplest tree with 3 nodes; 50% of the training sample was used to build each tree.

The following criteria, described in the methodological section of the paper, were used in order to assess the prediction accuracy achieved with each method: sales ratio, MAPE and COD. These three indicators were calculated for both learning and testing samples, which allowed making conclusions about the degree of overfitting for each method. The comparison is presented in Tables 2 and 3.

The following conclusions can be made from Tables 2 and 3:

1. For all algorithms and for any of the procedures (either one step or two-step) sales ratio is in the acceptable range of 0.9–1.1.
2. For the majority of the algorithms MAPE and COD decreased on both training and test samples after the two step procedure had been used. This leads to a conclusion that it is reasonable to use this procedure instead of assessing the overall price of a real estate object. In our study we will use the two-step procedure as the main one.
3. Despite the presence of a validation sample to avoid overfitting, neural networks (MLP and RBF), often considered as the best class of methods for real estate appraisal, are not among the best performing techniques in our study. Neural networks could probably deliver better results after some fine-tuning, but we think mass appraisal algorithms should be as independent of an analyst and as universal as possible. The main reason of neural networks poor performance is the small number of observations without any missing values and the large number of explanatory variables: in such a case it is very difficult to prevent overfitting.
4. Independent of the estimation method (either one step or two-step) Random forest has shown the best results. Whether Random forest will be the best choice on other datasets can be revealed only after a large number of comparative studies. However, we already now can suppose that this algorithm has advantage over neural networks in the cases when there are lots of explanatory variables and many missing values. Besides, in our case the prediction error was almost invariably falling with the growth of the number of trees in the ensemble, which indicates that Random forest is not prone to overfitting. This makes it a less risky choice for the use in automated expert systems compared to neural networks.

We have segmented the methods by two criteria: MAPE and COD (Fig. 1). It is easy to see that there are three groups of methods:

1. Providing high accuracy and low sales ratio variability (Random forest, Boosting, KNN (mean)).
2. Providing average model quality (CHAID, Exhaustive CHAID, CART, RBF).
3. Providing relatively low accuracy and high sales ratio variability (Regression, KNN (median), MLP).

The stability of the revealed segments should be verified on other datasets having similar structure.

3.3. The Random forest model diagnostics

3.3.1. Assessing feature importance

The importance of every variable is proportional to the average decrease in the residual sum of squares after splitting by this variable is done. The most important variable gets the score of 1; scores for other variables are derived by standardizing their aver-

Table 2
Prediction accuracy in the case of direct valuation.

Method	Test sample			Training sample		
	Mean SR	MAPE (%)	COD (%)	Mean SR	MAPE (%)	COD (%)
Regression	1.05	20.02	19.83	1.04	16.79	16.37
CHAID	1.02	17.37	17.44	1.05	15.06	15.15
Exhaustive CHAID	1.02	17.54	17.49	1.05	14.91	15.00
CART	1.04	19.89	19.82	1.05	15.73	16.38
KNN (mean)	1.06	21.72	21.14	1.05	18.46	17.68
KNN (median)	1.01	20.87	21.22	1.01	16.34	16.31
Boosted Trees	1.04	18.33	18.14	1.05	14.66	15.11
Random forest	1.03	17.25	16.97	1.04	12.57	12.82
MLP	1.04	19.79	19.49	1.05	16.38	15.50
RBF	1.05	20.53	19.37	1.06	19.25	18.47

Table 3
Prediction accuracy in the case of the two-step procedure.

Method	Test sample			Training sample		
	Mean SR	MAPE (%)	COD (%)	Mean SR	MAPE (%)	COD (%)
Regression	1.04	18.33	18.10	1.04	13.97	14.31
CHAID	1.04	16.92	16.93	1.04	13.94	13.54
Exhaustive CHAID	1.04	17.02	16.92	1.04	14.34	13.53
CART	1.04	17.36	17.16	1.04	14.62	13.72
KNN (mean)	1.02	15.63	15.41	1.03	14.93	14.44
KNN (median)	1.02	18.53	18.70	1.02	13.62	14.75
Boosted Trees	1.04	15.71	15.52	1.04	12.69	13.22
Random forest	1.03	14.86	14.77	1.04	11.70	12.25
MLP	1.09	20.53	19.75	1.02	16.89	11.54
RBF	1.03	16.90	16.78	1.05	14.94	16.17

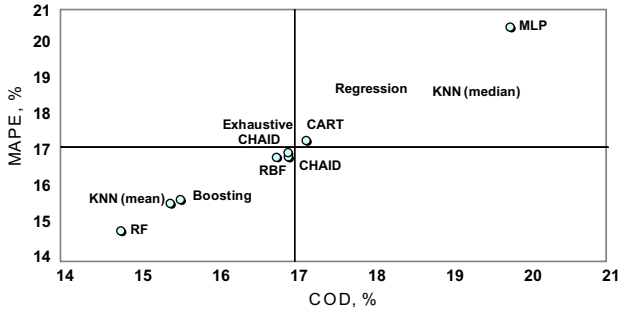

Fig. 1. Model quality indicators for methods which are being compared.

age relative to the largest one. The importance of each predictor is presented in Table 4 (the dependent variable is the price of an apartment per square meter, since in the case of the total price the area is obviously the most important factor).

The district appeared to be the most important factor, which is why its importance score is 1. We broke down all the predictors into 4 groups:

1. Very important factors:
- District
- Time to the city center by underground
- House type

2. Important factors:
- Total area of the apartment in square meters
- Bathroom unit type

Table 4
The importance of price per meter predictors.

Variable	Importance score (max = 1)
district	1.000
time_to_downtown	0.923
house_type	0.903
total_area	0.608
bathroom_unit	0.591
kitchen_area	0.451
living_area	0.419
distance_from_underground	0.411
number_of_floors	0.383
inequality2	0.291
inequality1	0.290
floor	0.220
telephone	0.085

Table 5
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 $\leqslant$ 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

3. Factors of average importance:
- Kitchen area in square meters
- Living area in square meters
- Distance from the house to the nearest underground station
- Number of floors in the house

4. Not so important factors:
- Absolute percentage difference between room areas
- Herfindahl index for room areas
- The floor, on which the apartment is situated
- Telephone availability

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

The calculation of COD and MAPE allowed us to carry out a preliminary comparison of different methods. Unfortunately, using these indicators 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 Section 2.3. Using this approach we will make the diagnostics of Random forest predictions that were considered the best in our comparative study.

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 5) 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 m^2, while MAPE is 19.4% for apartments with greater area,

Table 6
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

Table 7
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

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.

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 6). 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).

We calculated bootstrap confidence intervals for the average SR in each segment (see Table 7).

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, that 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 validated the application of the Random forest method to the mass appraisal that is characterized by the stability to outliers, the ability to work properly with missing values and categorical variables with many levels. We have also proposed and validated 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.

The effectiveness of Random forest has been supported by the empirical research based on Saint-Petersburg residential apartments dataset. A comparative study has shown that all algorithms perform better if the price per meter is predicted, followed by calculating the total price. We believe that using this two-step procedure instead of valuating the overall price of a real estate object is likely to increase model performance in most mass appraisal expert systems due to the heteroscedasticity and some other problems inherent in real estate data. Feature importance diagnostics has revealed that the district, time to the city center by underground, house type, total area of the apartment and bathroom unit type comprise the two most important groups of price per square meter predictors. The factors of low importance include indicators of inequality between room areas, the floor, on which the apartment is situated, and telephone availability. A deeper diagnostics using the proposed segmentational diagnostic approach has been conducted for the best model (Random forest). 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 m^2, 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.

Although our study provides not only empirical, but also some theoretical grounds for the wide use of the Random forest algorithm, we plan to conduct a comparative analysis on other datasets that, as we expect, will prove the superiority of Random forest over at least most methods used for the mass appraisal nowadays.

It is worth noting that Random forest requires choosing the optimal settings (the number of variables selected for building each tree and the bootstrap pseudosample size). Our preliminary observations have shown that the optimal parameters for the mass appraisal seemingly differ from the recommended by the algorithm authors and used as the default settings in some statistical software. However, the choice of these settings in our research is not rigorously grounded yet. Taking into account that Prinzie and Van den Poel (2008) showed Random forest prediction accuracy (for classification problems) to be rather sensitive to the parameters set by an analyst, we plan to compare the accuracy of several Random forest specifications in the future.

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 study of the entailed consequences.

References

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.

Ball, M. J. (1973). Recent empirical work on the determinants of relative house prices. *Urban Studies, 10*, 213–233.
Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.
Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on machine learning, Pittsburgh, PA* (pp. 161–168).
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.
Eckert, J. K. (1990). *Property appraisal and assessment administration*. International Association of Assessing Officers, Chicago, IL.
Fan, G., Ong, Z. S. E., & Koh, H. C. (2006). Determinants of house price: A decision tree approach. *Urban Studies, 43*(12), 2301–2315.
Filho, C. M., & Bin, O. (2005). Estimation of hedonic price functions via additive nonparametric regression. *Empirical Economics, 30*(1), 93–114.
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*.
International Association of Assessing Officers (2003). Standard on automated valuation models (AVMs). www.iaao.org. Approved September, 2003.
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.
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.
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.
Kontrimas, V., & Verikas, A. (2011). The mass appraisal of the real estate by computational intelligence. *Applied Soft Computing, 11*(1), 443–448.
Laakso, S. (1997). Urban housing prices and the demand for housing characteristics. The Research Institute of the Finnish Economy (ETLA) A 27, Helsinki.
Lee, Y.-L., Yeh, K.-Y., & Hsu, K.-C. (2003). Fair evaluation of real estate value in urban area via fuzzy theory. In *10th ERES Conference, Helsinki, Finland* (Vol. 5)
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.
Liu, J., Zhang, G. X. L., & Wu, W. P. (2006). Application of fuzzy neural network for real estate prediction. *LNCS, 3973*, 1187–1191.
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.
Miller, N. G. (1982). Residential property hedonic pricing models: A review. In C. F. Sirmans (Ed.), *Urban housing markets and property valuation. Research in real estate* (Vol. 2, pp. 31–56). Greenwich, CT: Jai Press Inc.
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.
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.
Prinzie, A., & Van den Poel, D. (2008). Random forests for multiclass classification: Random multinomial logit. *Expert Systems with Applications, 34*, 1721–1732.
Selim, H. (2009). Determinants of house prices in Turkey: Hedonic regression versus artificial neural network. *Expert Systems with Applications, 36*, 2843–2852.
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.
Verikas, A., Lipnickas, A., & Malmqvist, K. (2002). Selecting neural networks for a committee decision. *International Journal of Neural Systems, 12*(5), 351–362.
Verkooijen, W. J. H. (1996). Neural networks in economic modelling. Doctoral dissertation. Tilburg University, Center for Economic Research, 205p.
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.