*European Research Studies Journal*
*Volume XXIII, Issue 4, 2020*

*pp. 1230-1245*

# Parametric and Non-parametric Methods in Mass Appraisal on Poorly Developed Real Estate Markets*

*Submitted 30/08/20, 1st revision 24/09/20, 2nd revision 08/10/20, accepted, 11/11/20*

Sebastian Gnat[1], Mariusz Doszyń[2]

***Abstract:***

***Purpose:*** *The objective of the article is to identify machine learning methods that provide the best real estate appraisals for small-sized samples, particularly on poorly developed markets. A hypothesis is verified according to which machine learning methods result in more accurate appraisals than multiple regression models do, taking into account sample sizes.*
***Design/Methodology/Approach:*** *Four types of regression were employed in the study: a multiple regression model, a ridge regression model, random forest regression and k nearest neighbours regression. A sampling scheme was proposed which enables defining the impact of a sample size in training datasets on the accuracy of appraisals in test datasets.*
***Findings:*** *The research enabled drawing several conclusions. First of all, the greater the training set was, the more precise the appraisals in a test set were. The conclusion drawn is that a reduction of a training set causes the deterioration of modelling results, but such deterioration is not substantial. Secondly, ridge regression model appeared to be the best model, and thereby the one most resistant to a low number of data. This model, apart from demonstrating the greatest resistance, additionally has the advantage of being a parametric, hence allowing inference.*
***Practical Implications:*** *Presented considerations are important, for instance in the case of valuations conducted for fiscal purposes, when it becomes necessary to determine the value of every type of real properties, even the ones featuring sporadically occurring states of properties.*
***Originality/Value:*** *The study contains modelling of the values defined by property appraisers, and not prices, as in the majority of studies. This decision enabled increasing the diversity of states of real estate properties, thereby including in the modelling process not just those real properties which are most typically traded.*

***Keywords:*** *Property mass appraisal, multiple regression, machine learning algorithms, appraisal accuracy measures, real estate market.*

***JEL classification:*** *C01, C14, C21, R31.*

***Paper Type:*** *Research study.*

---

[1]*University of Szczecin, Poland, Institute of Economics and Finance, Department of Econometrics and Statistics, ORCID ID: 0000-0003-0310-4254, sebastian.gnat@usz.edu.pl*
[2]*University of Szczecin, Poland, Institute of Economics and Finance, Department of Econometrics and Statistics, ORCID ID: 0000-0002-3710-1177, mariusz.doszyn@usz.edu.pl*
**This article has been selected as one of the best presented in ICABE 2020 to be published in this issue.*

***Acknowledgement:*** *The research was conducted within the project financed by the National Science Centre, Project No 2017/25/B/HS4/01813.*

## 1. Introduction

In a dynamically developing legal, economic and technical environment more and more often it becomes necessary to apply property mass appraisal methods. Mass appraisal becomes a necessity when the knowledge of the values of multiple real properties is required, determined at the same moment with a uniform methodological approach.

Problems related to real estate mass appraisal frequently result from the lack of complete and reliable databases regarding transactions on local markets. On poorly developed real estate markets the number of transactions is typically limited. It especially applies to the market segments other than the residential real estate. A low number of transactions concerns e.g. the market of undeveloped land real property in urban areas. This market segment is the subject of analysis undertaken in this article. A limited number of transactions may prevent mass appraisal, even when simple methods are applied. It is for that reason, inter alia, that in this article real estate values, estimated by a team of certified property appraisers, are taken into account instead of transaction prices. Those values were determined in individual (not mass) appraisals. The selection of real properties to the data base enables ensuring its greater representativeness. In turn, greater cost ineffectiveness, resulting from the need to remunerate property appraisers for the preparation of individual valuations, may constitute a drawback of adopting values instead of prices. It is for this reason, among other things, that the objective of this article involves demonstrating such valuation methods that compare best for small-sized (training) datasets.

Recapitulating, a limited number of transactions on a local real estate market ceases to have a decisive significance if real estate values, and not transaction prices, are adopted for a database. It is especially important on poorly developed real estate market. This type of a market – a market of undeveloped land real estate in an urban area – is the subject of analyses in this article.

In the article the authors verify a hypothesis according to which machine learning methods result in better valuations than those yielded by multiple regression models, especially in the cases of small-sized samples. The above hypothesis will be verified on the grounds of a database comprising the values of undeveloped land real estate in urban areas. If a database can contain information on a lower number of real properties (in a training dataset), the cost of constructing it is lower. In that case the cost of mass appraisal becomes lower as well. The creation of databases containing real estate values determined by property appraisers is cost ineffective. That is why the properties of individual mass appraisal method will be examined for samples of decreasing sizes. The purpose of the study is to identify the methods, chiefly machine learning methods, which provide the best appraisals, taking into

consideration sample sizes.

## 2. Literature Review

A review of property mass valuation models could be found e.g. in (Jahanshiri *et al.*, 2011). In the paper, mass valuation methods are divided into non-spatial and spatial models. An outline of the AVM (Automated Valuation Models), which also might be useful in mass appraisals, is presented in (d'Amato, 2017). In this article different methods (multiple regression models, spatial models) and their evolution in the last decades are also described. The general review of quantitative methods used in mass appraisal could be likewise found in (Pagourtzi *et al.*, 2003). In the article methods are divided into traditional ones (multiple regression, comparable, cost, income, profit, contractors methods) and advanced ones, such as ANN (Artificial Neural Networks), hedonic pricing methods, spatial analysis, fuzzy logic, ARIMA models.

An interesting comparison of modern approaches in mass appraisals is presented in (McCluskey *et al.*, 2013). Also many useful proposals of prediction accuracy measures are given. Prediction accuracy measures are also discussed in (Doszyń, 2019). A classification of quantitative methods useful in a mass appraisal process is presented in (Kauko and d'Amato, 2008). Generally, appraisal methods are classified into four groups: model-driven methods, data-driven methods, methods based on machine learning and expert methods. This classification could be treated as a benchmark, because it includes most groups of mass appraisal methods.

Model-driven methods include mostly econometric models, hedonic regression models and spatial econometric models. The literature concerning the possibility of applying econometric methods in appraisal is fairly extensive e.g. (Benjamin *et al.*, 2004; Isakson, 1998; Dell, 2017). The basics of appraisals performed by means of multiple regression are presented in (Benjamin *et al.*, 2004). In his study Dell (2017) emphasized that regression and other analytical tools could be very useful in appraisal, but in real applications they are often misused. Econometric methods are sometimes also used not directly in appraisal but, for example, to identify specific transactions (Doszyń and Gnat, 2017), (Doszyń *et al.*, 2017).

Nowadays, often spatial econometric models are applied in mass valuations. Many authors assume that spatial effects could be treated as a proxy for location. A comparison of different methods in modelling and predicting house prices is described in (Bourassa *et al.*, 2009). An interesting proposal of modelling spatial variation in housing prices is presented in (Fik *et al.*, 2003). Data-driven methods include non-parametric models, such as GWR. Methods based on machine learning are nowadays applied equally frequently. This class of tools accounts for ANN, rough set theory, fuzzy logic, genetic algorithms, etc. An interesting application of machine learning methods could be found e.g. in (Zurada *et al.*, 2011). If the quality of databases is low, expertise methods, such as AHP (Analytic Hierarchy Process),

Conjoint Analysis, CV (Contingent Valuation) could efficiently support mass appraisal process. The application of an expert system in mass appraisal is presented in e.g. (Kilpatrick, 2011).

Despite extensively developed research and the employment of various types of econometric models in real estate valuation, the search for increasingly better solutions is unrelenting. As was already mentioned, machine learning also constitutes a methodological area used in determining real estate value. Literature is very extensive in this regard and it can be divided into two trends. The first one comprises studies within the scope of which authors use and try to improve the existing solutions within the framework of multiple regression (Zaddach and Alkhatib, 2014), regression trees (McCluskey *et al.*, 2014), random forests (Antipov and Pokryshevskaya, 2012), support vector machines (Wang *et al.*, 2014), market segmentation (Ciuna *et al.*, 2017), artificial neural networks (Ćietković *et al.*, 2018; Liu *et al.*, 2011). What is more, attempts at using approaches never before applied in valuation can be found. Such studies include the employment of purpose programming (Morano *et al.*, 2018) or innovative methods of real estate clustering aimed at improving prediction accuracy (Shi *et al.*, 2015). Machine learning is also used for the specification of variables in hedonic models (Yoo *et al.*, 2012). A comparison of various ensemble techniques to increase prediction accuracy (Graczyk *et al.*, 2010) led to a conclusion that, in general, property valuation results obtained with stacking of utilized models were characterized by the lowest prediction error but the outcome tended to vary. Ensemble techniques were also used on real estate market for projects classification (Paireekreng and Choensawat, 2015).

The second trend focuses on comparing several algorithms in order to determine which one yields the best results. An example of such work is the paper of (Park and Bae, 2015), in which the effectiveness of real estate price predictions in Fairfax County, Virginia, was analysed. English housing rental market was subjected to mass appraisal using generalized linear regression, machine learning and a pseudo practitioner based approach (Clark and Lomax, 2018). Apart from the conclusions regarding the fact that machine learning models proved to be superior to multiple regression, the authors argue that the use of machine learning is computationally demanding, which was also confirmed in this study. Whereas a comparison of the random forest and multiple regression in the Cyprus market (Dimopoulos *et al.*, 2018) demonstrated that the random forest outperformed the linear models.

Comparative studies frequently use artificial neural networks as representatives of machine learning. Their superiority over multiple regression models was proven on the example of New York (Khamis and Kamarudin, 2014). Machine learning models are also compared with expertise approach (Trawiński *et al.*, 2017). In this study machine learning algorithms occurred to be better as well. In the article the authors use an expert method of dividing the appraised area into smaller, more homogenous areas, which was also employed in this study.

Despite the examples demonstrating the advantage of machine learning methods, it is possible to find studies in which there are no significant differences between e.g. neural networks and multiple regression, or even such studies in which neural networks appeared to be a worse solution. For instance, in the work of (Del Giudice *et al.*, 2017) the conducted experiment proved the superiority of models based on Markov chains over neural networks.

In publications on real estate valuation, the subject of the impact of a data size on the quality of models is rarely undertaken. The issue of small-sized training datasets is analysed in the research on artificial neural networks (Shaikina *et al.*, 2015; Barz and Denzler, 2019). In those studies it has been demonstrated that it is possible to obtain high quality supervised training results despite small-sized datasets. In the research related to space, in the interpolation of hydrocarbon deposits it was also demonstrated that even a small number samples enables obtaining valuable results (Malvić *et al.*, 2019).

## 3. Research Methodology

Four types of regression models were used in the research. A multiple regression model (MR), ridge regression model (RR), random forest regression (RF) as well as *k* nearest neighbours regression (*k*NN). The first two types are parametric models, whereas the remaining two models are non-parametric algorithms.
In the survey a non-linear econometric model constitutes a point of reference:

$$ln(w_{ji}) = \alpha_0 + \sum_{k=1}^{K} \sum_{p=2}^{k_p} \alpha_{kp} x_{kpi} + \sum_{j=2}^{J} \alpha_j laz_{ji} + u_i \quad (1)$$

where:
$w_{ji}$ – unit market value of *i*–th real estate in *j*–th location attractiveness zone,
$N$ – number of real estate $(i = 1, 2, \dots, N)$,
$J$ – number of location attractiveness zones $(j = 2, 3, \dots, J)$,
$\alpha_0$ – constant term,
$K$ – number of real estate attributes,
$k_p$ – number of states of *k*–th attribute,
$\alpha_{kp}$ – impact of *p*–th state of attribute *k*,
$x_{kpi}$ – zero–one variable for *p*–th state of attribute *k*,
$\alpha_j$ – market value coefficient for *j*–th location attractiveness zone,
$laz_{ji}$ – dummy variable equal one for *j*–th location attractiveness zone,
$u_i$ – random component.

The explained variable is a natural logarithm of a real estate unit value. Real estate values are determined by certified appraisers in the individual appraisals. Real estate

attributes are qualitative characteristics measured on an ordinal scale, so they are introduced into the model (1) through dummy variables for each state of an attribute. In model (1) there is a constant term. In order to avoid strict collinearity of the explanatory variables, each dummy variable for the worst attribute states is skipped. Hence, we arrive at the summation of $p = 2, \ldots, k_p$ in the formula (1). In the interpretation, the ignored state of an attribute serves as a point of reference for the remaining states.

There are also market value coefficients $(\alpha_j)$ in model (1). They could be treated as a proxy for a location. They are estimated by introducing dummy variables for each location attractiveness zone. Location attractiveness zones are constructed by experts. They are defined as areas with similar impact of location. Therefore, location attractiveness zones are constructed in such a way that the impact of a location in a given area is homogenous. Because of the strict collinearity of explanatory variables, the worst (cheapest) location attractiveness zone is skipped. The omitted location attractiveness zone creates a point of reference.

Model (1) was a starting point for the application of the remaining machine learning methods (ridge regression, random forest regression, *k*NN regression).

In multiple regression models, model weights are determined by minimizing the sum of squares of the residuals of the model ($RSS \rightarrow \min$). When it comes to ridge regression, a regularization term equal to $\beta \sum_{i=1}^{n} \alpha_i^2$ is added to *RSS* cost function (Lesmeister, 2019) of equation (1). The hyperparameter $\beta$ controls how much one wants to regularize the model. If $\beta$ = 0, then ridge regression is just pure multiple regression. If $\beta$ is very large, then all weights end up very close to zero and the result is a flat line going through the data's mean (Geron, 2017). Therefore, setting $\beta$ is the crucial stage of creating a model in order to achieve high quality results.

Random forest (Breiman, 2001) constitutes a machine learning algorithm mostly used in classification problems. Yet it is also possible to use it as a regression algorithm. A random forest regressor is a type of a simple regression trees ensemble, which gives a prediction based on averaging predictions made by each tree in the ensemble. A clear summary of RF algorithm was presented e.g. by Antipov and Pokryshevskaya (2012).

The *k* nearest neighbours algorithm, similarly to a random forest algorithm, is a non-parametric algorithm. Though mainly applied in classification problems, the *k*NN algorithm can also be used in regression problems (Pace, 1996). The operation of the algorithm comes down to two steps. In the first step for a given point $x_0$ we find *k* training points $x(r)$, $r$ = 1, …, *k*, located closest to $x_0$. In the second step a prediction is made based on averaging of target variable value of every training point. The machine learning part of the algorithm regards choosing an optimal *k* for the highest accuracy of prediction in testing sets.

Data accuracy will be evaluated on the basis of the following errors:

– Percentage Error (*PE*):

–

$$PE_i = \frac{w_i - \hat{w}_i}{w_i} \cdot 100\% \quad (2)$$

where:

$w_i$ – real unitary real estate value determined by a property appraiser,

$\hat{w}_i$ – theoretical unitary real estate value determined in a model,

– Absolute Percentage Error (APE):

–

$$APE_i = \frac{|w_i - \hat{w}_i|}{w_i} \cdot 100\% \quad (3)$$

## 4. Description of the Database

The data base used in the study contains information not on transaction prices, but on real estate values, which were determined by property appraisers in individual valuations. In a short period transactions may refer to the real properties having attributes that differ very little. Low variability of attributes (explanatory variables) translates into e.g. low effectiveness of econometric model estimators. When commissioning the appraisal of real properties of various attribute states this problem can be avoided, since the variance of explanatory variables (attributes) is greater.

Attributes and their states are presented in Table 1. It can be noted that all attributes, except surface area, are qualitative variables. They are introduced into econometric model (1) as a dummy variable for each state of an attribute (with the exclusion of the first, worst state). Surface area is a quantitative variable, but it is treated as a qualitative one. This is because market participants often treat this variable in this way. This conclusion was presented by appraisers. With respect to real estate unit value, it is assumed that a small surface is better than average one, and average is better than large.

***Table 1.*** *Real estate attributes and their states*

| No. | Attribute | Attribute category |
|---|---|---|
| 1 | Utilities | None<br>Incomplete<br>Complete |
| 2 | Neighbourhood | Onerous<br>Unfavourable<br>Average<br>Favourable |
| 3 | Transport availability | Unfavourable<br>Average |

| | | |
|---|---|---|
| | | Favourable |
| 4 | Physical plot properties | Unfavourable<br>Average<br>Favourable |
| 5 | Plot area | Large (>1200 $m^2$)<br>Average (500 – 1200 $m^2$)<br>Small (<500 $m^2$) |

***Source:*** *Own work.*

The study encompassed 318 land plots located in one of the largest cities of Poland – Szczecin. The basic positional measurements calculated for the employed set of 318 real properties are presented in Table 2. Unitary values of real properties were within the range of 502.11 PLN/1$m^2$ – 701.43 PLN/1$m^2$, with a median equal to 592.28 PLN/1$m^2$. In the case of all attributes, except for the neighbourhood, the median was equal to a maximum value of an attribute. Variability measured with quartile deviation and positional coefficient of variation was rather small.

***Table 2.*** *Descriptive statistics if unitary values (in PLN – Polish zlotys) of real properties and their attributes defined for a set of 318 real properties*[3]

| Statistics | Values of 1$m^2$ | Plot area | Utilities | Transport availability | Neighbourhood | Physical properties |
|---|---|---|---|---|---|---|
| *Min* | 502.11 | 1 | 3 | 1 | 1 | 1 |
| $Q_{1.4}$ | 569.26 | 2 | 3 | 2 | 3 | 2 |
| *M* | 592.28 | 3 | 3 | 3 | 3 | 3 |
| $Q_{3.4}$ | 623.52 | 3 | 3 | 3 | 3 | 3 |
| *Max* | 701.43 | 3 | 3 | 3 | 4 | 3 |
| *Q* | 27.13 | 0.5 | 0 | 0.5 | 0 | 0.5 |
| $V_Q$ (%) | 4.58 | 16.667 | 0 | 16.667 | 0 | 16.667 |

***Source:*** *Authors' calculations.*

## 5. Description of the Database

In the research a set of 318 real properties was divided into training and test sets according to the following scheme. Out of the entire set, 68 properties were drawn 1000 times (test sets). The remaining 250 properties were used to create training sets. The initial training sets included 250 properties each. These sets were randomly reduced by 25 properties in a few steps. The smallest training sets consisted of 50 properties. In this way, 9000 pairs of training and test sets were created. For each of these pairs, four different regression models were built, in order to compare which of them is more resistant to the reduction of the training set.

[3]*Real estate attributes are encoded in such a manner that a worse variant equals 1, a subsequent variant is 2, etc. Min is a minimum value,* $Q_{1.4}$ *– first quartile, M – median,* $Q_{3.4}$ *– third quartile, max – is maximum value, Q – quartile deviation* $V_Q$ *– positional coefficient of variation.*

On account of the fact that four regression types were used in the research, and altogether 36000 models were evaluated. Hyperparameters optimization with grid search with cross validation was carried out for ridge regression models, *k* nearest neighbours and random forest regression. For the purpose of accelerating optimization calculations, whose total duration was estimated at several weeks of computer operation, out of every 1000 training sets 50 sets were drawn. Those sets of hyperparameters, which occurred most frequently as optimal ones (which minimum mean absolute error), were chosen for creating models. Calculations were conducted with scikit-learn package in Python programming language (Pedregosa *et al. 2011).*

Table 3 presents average $R^2$ coefficients of determination obtained for individual models in training datasets. The lowest average fit was achieved for a random forest regression model, whereas the highest one – for the regression based on *k*NN algorithm. All the models showed a tendency for overfitting demonstrated by the fact that along with a decreased training set size, average model fit was increasing. It is worth supplementing the comparison of the level of $R^2$ coefficients of determination by verifying which of the models is the most resistant to overfitting, taking into account a reduction of a sample size.

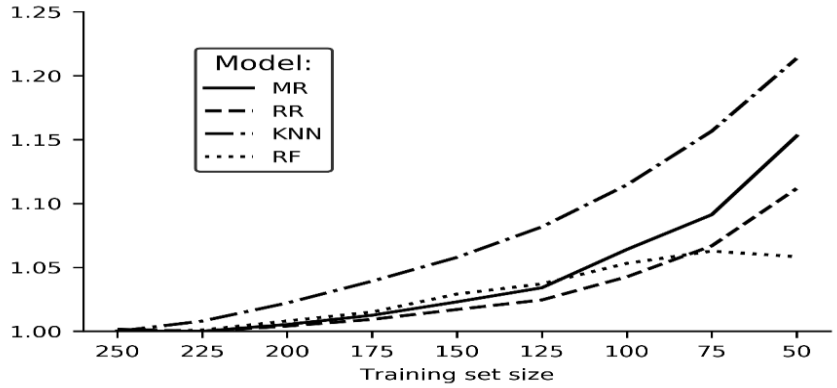
***Table 3.*** *Average determination coefficients*

| Training set size | Model type | | | |
|---|---|---|---|---|
| | Multiple regression | Ridge regression | *k* nearest neighbours regression | Random forest regression |
| 250 | 0.485 | 0.479 | 0.588 | 0.473 |
| 225 | 0.485 | 0.478 | 0.593 | 0.474 |
| 200 | 0.487 | 0.48 | 0.601 | 0.477 |
| 175 | 0.491 | 0.483 | 0.611 | 0.48 |
| 150 | 0.496 | 0.487 | 0.622 | 0.487 |
| 125 | 0.501 | 0.49 | 0.636 | 0.491 |
| 100 | 0.516 | 0.499 | 0.656 | 0.498 |
| 75 | 0.529 | 0.51 | 0.681 | 0.503 |
| 50 | 0.559 | 0.532 | 0.714 | 0.501 |

***Source:*** *Authors' calculations.*

Relative changes of average fit indices in training datasets are presented in Figure 1. Average model fit for training datasets of the biggest size, i.e. the ones that number 250 real properties, were adopted as the basis. *k*NN models demonstrated the greatest susceptibility to overfitting. In turn, the most stable average fit occurred in the case of regression based on random forests. These are contradictory conclusions in relation to the analysis of the values of the determination coefficients.

***Figure 1.*** *Relative changes of average $R^2$ in training sets*




***Source:*** *Authors' calculations.*

In the next stage of the study unitary values of real properties from test datasets were determined with the use of estimated models for various training set sizes. As a reminder, the test set sizes were constant and equal to 68, whereas training set sizes were decreasing from 250 to 50, by 25.

The quality of appraisals was evaluated with the use of mean percentage errors (*MPE*) and mean absolute percentage errors (*MAPE*). As previously mentioned, the valuations obtained through models were compared to the valuations performed by certified property appraisers. Mean percentage errors are presented in Table 4. In every case they were values close to 0, which proves that the results obtained with the use of employed models, irrespectively of a training set size, did not demonstrate any bias.

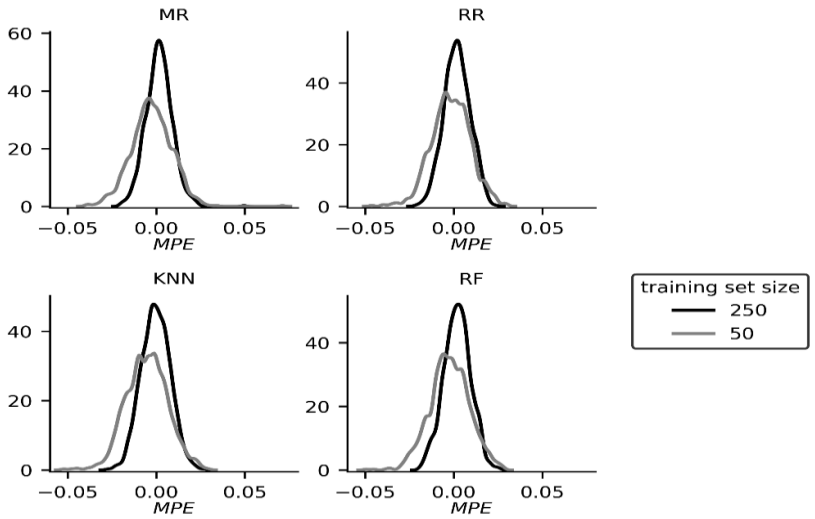
***Table 4.*** *Average mean percentage errors (MPE)*

| Training set size | Model type | | | |
|---|---|---|---|---|
| | Multiple regression | Ridge regression | *k* nearest neighbours regression | Random forest regression |
| 250 | 0.001559 | 0.001845 | -0.000335 | 0.001934 |
| 225 | 0.000896 | 0.001110 | -0.001118 | 0.001164 |
| 200 | 0.000461 | 0.000605 | -0.001705 | 0.000642 |
| 175 | -0.000058 | 0.000247 | -0.002011 | 0.000330 |
| 150 | -0.000655 | -0.000068 | -0.002560 | -0.000169 |
| 125 | -0.000811 | -0.000433 | -0.003295 | -0.000543 |
| 100 | -0.001169 | -0.000735 | -0.003687 | -0.000979 |
| 75 | -0.002148 | -0.001143 | -0.004654 | -0.001497 |
| 50 | -0.002622 | -0.001589 | -0.005331 | -0.002294 |

***Source:*** *Authors' calculations.*

More precise analyses of the distribution of percentage errors, which were visualized (for biggest smallest training sets) with kernel density estimation in Figure 2, demonstrate a displacement of the distributions to the left in the case of smaller training sets. It means that there is a greater probability of overestimating a real estate value for smaller training sets.

***Figure 2.*** *KDE distribution of mean percentage errors for largest and smallest train set size*




***Source:*** *Authors' calculations.*

Table 5 presents mean absolute percentage errors of appraisals in test sets, also for various training set sizes. The results obtained with the analysed models occurred to be similar to one another. A ridge regression model turned out to be the best one on account of the *MAPE* level and *k*NN regression models proved to perform the most poorly. The *MAPE* increase accompanied by a decrease in training set sizes was observed for all the models.

***Table 5.*** *Average mean absolute percentage errors (MAPE)*

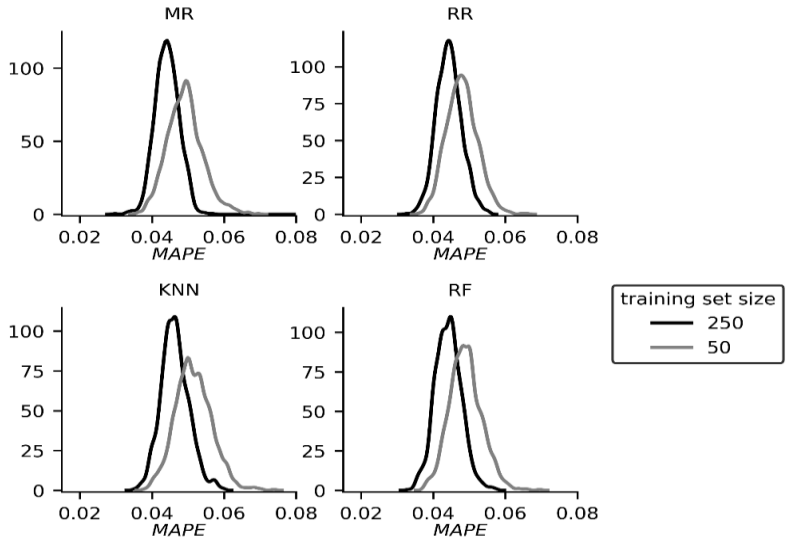
| Training set size | Model type | | | |
|---|---|---|---|---|
| | Multiple regression | Ridge regression | *k* nearest neighbours regression | Random forest regression |
| 250 | 0.0442 | 0.0444 | 0.0464 | 0.0442 |
| 225 | 0.0445 | 0.0446 | 0.0468 | 0.0445 |
| 200 | 0.0447 | 0.0447 | 0.0471 | 0.0448 |
| 175 | 0.0449 | 0.0449 | 0.0476 | 0.0452 |
| 150 | 0.0452 | 0.0452 | 0.0480 | 0.0456 |
| 125 | 0.0456 | 0.0455 | 0.0486 | 0.0461 |

| Training set size | Model type | | | |
|---|---|---|---|---|
| | Multiple regression | Ridge regression | *k* nearest neighbours regression | Random forest regression |
| 100 | 0.0463 | 0.0460 | 0.0492 | 0.0469 |
| 75 | 0.0473 | 0.0467 | 0.0504 | 0.0477 |
| 50 | 0.0492 | 0.0480 | 0.0515 | 0.0490 |

***Source:*** *Authors' calculations.*

Kernel density estimation of *MAPE* for extreme training set sizes is presented in Figure 3. As mentioned earlier, valuation errors estimated with the use of models based on smaller training sets are higher. It is demonstrated in *MAPE* distributions shifted to the right. An analysis of these distributions once again shows that a ridge regression model is the best model type. In the case of smaller training sets errors for this model are higher, similarly as for the remaining models, but the distribution dispersion is the smallest in this case.

***Figure 3.*** *KDE distribution of mean absolute percentage errors for largest and smallest train set size*

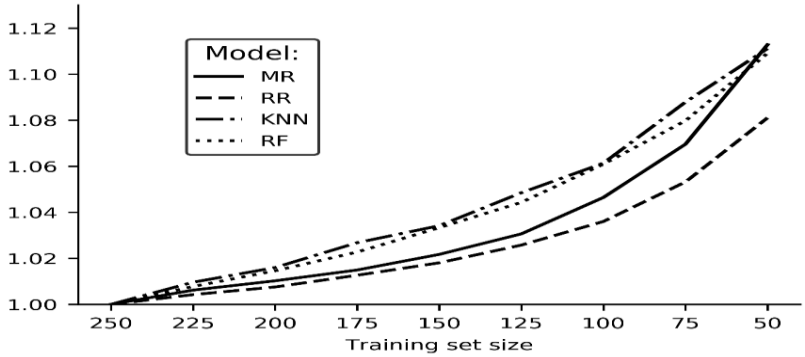



***Source:*** *Authors' calculations.*

The main objective of the study is demonstrating which model type is the best in the case of small-sized training sets. In order to find the answer to the question thus formulated, *MAPE* errors were presented in relative terms, where average values of those measures for the largest training sets constituted the point of reference. Figure 4 shows how *MAPE* rises for individual models along with a decrease of training set sizes. As can be observed, the smallest rise in appraisal errors occurred in the case of ridge regression. It means that this model type ought to be recommended in the

event of reduced training set sizes. In the case of the remaining models the deterioration of *MAPE* averages was similar as in the case of the smallest trainings sets.

***Figure 4.*** *Relative changes of average MAPE*




***Source:*** *Authors' calculations.*

## 6. Conclusions

The article presents the application of several regression models on a poorly developed real estate market. The main objective of the paper was verifying the possibility of using those models in the case of limited number of observations. At the age of big data one must not forget that a lot of easily available information can be found in every area. Local real estate markets, where the number of concluded transactions is low, often constitute such an area.

The research enabled drawing several conclusions. First of all, the greater the training set was, the more precise the appraisals in a test set were. This is fairly obvious. However, the scale of results deterioration as a consequence of diminished training set size is more interesting. Therefore, the conclusion drawn is that a reduction of a training set causes the deterioration of modelling results, but such deterioration is not substantial.

Secondly, ridge regression model appeared to be the best model, and thereby the one most resistant to a low number of data. This model, apart from demonstrating the greatest resistance, additionally has the advantage of being a parametric, enabling the evaluation of the impact of individual real estate parameters on the real estate value. It is a highly important feature for some experts determining a real estate value. It is worth pointing out that although multiple regression models performed slightly worse that ridge regression models, but still far better than random forest and *k*NN regression models did, both from the standpoint of models fit in training sets, as well as valuation accuracy in test sets.

In the subsequent research on the possibility of applying regression models on poorly developed real estate markets the most important conclusions will be verified on other markets and for other real estate types.

**References:**

Antipov, E.A., Pokryshevskaya, E.B. 2012. Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics. Expert Systems with Applications, 39(2), 1772-1778, doi: 10.1016/j.eswa.2011.08.077.

Barz, B., Denzler, J. 2019. Deep Learning on Small Datasets without Pre-Training using Cosine Loss. CoRR, abs/1901.09054. http://arxiv.org/abs/1901.09054, doi: 10.1109/WACV45572.2020.9093286.

Benjamin, J.D., Randall, S., Guttery, R.S., Sirmans, C.F. 2004. Mass Appraisal: An Introduction to Multiple Regression Analysis for Real Estate Valuation. Journal of Real Estate Practice and Education, 7(1), 65-77, doi: 10.1080/10835547.2004.12091602.

Bourassa, S.C., Cantoni, E. & Hoesli, M. 2009. Predicting House Prices with Spatial Dependence: A Comparison of Alternative Methods. Working Paper, 15th Conference of the Pacific Rim Real Estate Society (PRRES) in Sydney.

Breiman, L. 2001. Random forests. Machine Learning, 45(1), 5-32, doi: 10.1023/A:1010933404324.

Ćetković, J., Slobodan, L., Lazarevska, L., Žarković, M., Vujošević, S., Cvijović, J., Gogić, M. 2018. Assessment of the Real Estate Market Value in the European Market by Artificial Neural Networks Application. Complexity 2018, Article ID 1472957, doi: 10.1155/2018/1472957.

Ciuna, M., Milazzo, L., Salvo, F. 2017. A Mass Appraisal Model Based on Market Segment Parameters. Buildings, 7(2), 34, doi: 10.3390/buildings7020034.

Clark, S.D., Lomax, N. 2018. A mass-market appraisal of the English housing rental market using a diverse range of modelling techniques. Journal of Big Data, 5, 43, doi: 10.1186/s40537-018-0154-3.

d'Amato, M. 2017. A Brief Outline of AVM Models and Standards Evolutions. In M. d'Amato, T. Kauko (Ed.), Advances in Automated Valuation Modelling, 3-21. Springer International Publishing.

Del Giudice, V., De Paola, P., Forte, F., Manganelli, B. 2017. Real Estate Appraisals with Bayesian Approach and Markov Chain Hybrid Monte Carlo Method: An Application to a Central Urban Area of Naples. Sustainability, 9(11), 1-17, doi: 10.3390/su9112138.

Dell, G. 2017. Regression, Critical Thinking, and the Valuation Problem Today. Appraisal Journal, 85(3), 217-230.

Dimopoulos, T., Tyralis, H., Bakas, N.P., Hadjimitsis, D. 2018. Accuracy measurement of Random Forests and Linear Regression for mass appraisal models that estimate the prices of residential apartments in Nicosia, Cyprus. Advances in Geosciences, 45, 377-382, doi: 10.5194/adgeo-45-377-2018.

Doszyń, M. 2019. Intermittent demand forecasting in the Enterprise: Empirical verification. Journal of Forecasting, 38(5), 459-469, doi: 10.1002/for.2575.

Doszyń, M., Gnat, S. 2017. Econometric Identification of the Impact of Real Estate Characteristics Based on Predictive and Studentized Residuals. Real Estate Management and Valuation, 25(1), 84-93, doi: 10.1515/remav-2017-0005.

Doszyń, M., Gnat, S., Bas, M. 2017. The Econometric Procedures of Specific Transactions Identification. Folia Oeconomica Stetinensia, 17(1), 20-30, doi: 10.1515/foli-2017-0002.

Fik, T.J., Ling, D.C., Mulligan, G.F. 2003. Modelling Spatial Variation in Housing Prices: A Variable Interaction Approach. Real Estate Economics, 31(4), 623-646, doi: 10.1046/j.1080-8620.2003.00079.x.

Geron, A. 2017. Hands-On Machine Learning with Scikit-Learn and TensorFlow. O'Reilly.

Graczyk, M., Lasota, T., Trawiński, B., Trawiński, T. 2010. Comparison of Bagging, Boosting and Stacking Ensembles Applied to Real Estate Appraisal. In Nguyen N.T., Le M.T., Świątek J. (Eds), Intelligent Information and Database Systems. ACIIDS 2010. Lecture Notes in Computer Science, vol 5991. Berlin, Heidelberg: Springer, doi: 10.1007/978-3-642-12101-2_35.

Isakson, H.R. 1998. The Review of Real Estate Appraisals Using Multiple Regression Analysis. Journal of Real Estate Research, 15(2), 177–190, doi: 10.1080/10835547.1998.12090922.

Jahanshiri, E., Buyong, T., Shariff, A.R.M. 2011. A Review of Property Mass Valuation Models. Pertanika Journal of Science & Technology, 19, 23-30.

Kauko, T., d'Amato, M. 2008. Mass Appraisal Methods: An International Perspective for Property Valuers. Blackwell Publishing Ltd, doi: 10.1002/9781444301021.

Khamis, A.B., Kamarudin, N. 2014. Comparative study on estimate house price using statistical and neural network model. International Journal of Scientific & Technology Research, 3(12), 126-131.

Kilpatrick, J. 2011. Expert Systems and Mass Appraisal. Journal of Property Investment & Finance, 29(4/5), 529-550, doi: 10.1108/14635781111150385.

Lesmeister, C. 2019. Mastering Machine Learning with R. Pact Publising.

Liu, X., Deng, Z., Wang, T. 2011. Real estate appraisal system based on GIS and BP neural network. Transactions of Nonferrous Metals Society of China, 21, 626-630, doi: 10.1016/S1003-6326(12)61652-5.

Malvić, T., Ivšinović, J., Velić, J., Rajić, R. 2019. Interpolation of Small Datasets in the Sandstone Hydrocarbon Reservoirs, Case Study of the Sava Depression, Croatia. Geosciences, 9(5), 201, doi: 10.3390/geosciences9050201.

McCluskey, W.J., Daud, D. Z., Kamarudin, N. 2014. Boosted regression trees: An application for the mass appraisal of residential property in Malaysia, Journal of Financial Management of Property and Construction, 19(2), 152-167. doi: 10.1108/JFMPC-06-2013-0022.

McCluskey, W.J., McCord, M., Davis, P.T., Haran, M., McIllhatton, D. 2013. Prediction Accuracy in Mass Appraisal: a Comparison of Modern Approaches. Journal of Property Research, 30(4), 239-265, doi: 10.1080/09599916.2013.781204.

Morano, P., Tajani, F., Locurico, M. 2018. Multicriteria analysis and genetic algorithms for mass appraisals in the Italian property market. International Journal of Housing Markets and Analysis, 11(2), 229-262, doi: 10.1108/IJHMA-04-2017-0034.

Pace, R.K. 1996. Relative performance of the grid, nearest neighbor, and OLS estimators. Journal of Real Estate Finance and Economics, 13(3), 203-218, doi: 10.1007/BF00217391.

Pagourtzi, E., Assimakopoulos, V., Hatzichristos, T., French, N. 2003. Real Estate Appraisal: a Review of Valuation Methods. Journal of Property Investment & Finance, 21(4), 383-401, doi: 10.1108/14635780310483656.

Paireekreng, W., Choensawat, W. 2015. An Ensemble Learning Based Model for Real Estate Project Classification. Procedia Manufacturing, 3, 3852-3859, doi: 10.1016/j.promfg.2015.07.892.

Park, B., Bae, J.K. 2015. Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. Expert Systems with Applications, 42(6), 2928-2934, doi: 10.1016/j.eswa.2014.11.040.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., et al. 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

Shaikhina, T., Lowe, D., Daga, S., Briggs, D., Higgins, R., Khovanova, N. 2015. Machine Learning for Predictive Modelling based on Small Data in Biomedical Engineering. IFAC-PapersOnLine, 48(20), 469-474, doi: 10.1016/j.ifacol.2015.10.185.

Shi, D., Guan, J., Zurada, J., Levitan, A.S. 2015. An Innovative Clustering Approach to Market Segmentation for Improved Price Prediction. Journal of International Technology & Information Management, 24(1), 15-32.

Trawinski, B., Telec, Z., Krasnoborski, J., Piwowarczyk, M., Talaga, M., Lasota, T., Sawilow, E. 2017. Comparison of expert algorithms with machine learning models for real estate appraisal. IEEE International Conference on INnovations in Intelligent SysTems and Applications (INISTA), doi: 10.1109/INISTA.2017.8001131.

Wang, X., Wen, J., Zhang, Y., Wang, Y. 2014. Real estate price forecasting based on SVM optimized by PSO. Optik, 125(3), 1439-1443, doi: 10.1016/j.ijleo.2013.09.017.

Yoo, S., Im, J., Wagner, J.E. 2012. Variable selection for hedonic model using machine learning approaches: A case study in Onondaga County, NY. Landscape and Urban Planning, 107(3), 293-306, doi: 10.1016/j.landurbplan.2012.06.009.

Zaddach, S., Alkhatib, H. 2014. Least squares collocation as an enhancement to multiple regression analysis in mass appraisal applications. Journal of Property Tax Assessment & Administration, 11(1), 47-66.

Zurada, J., Levitan, A.S., Guan, J. 2011. A Comparison of Regression and Artificial Intelligence Methods in a Mass Appraisal Context. Journal of Real Estate Research, 3, 349-387, doi: 10.1080/10835547.2011.12091311.