land

MDPI

Article

Property Mass Valuation on Small Markets

Sebastian Gnat

Department of Econometrics and Statistics, Institute of Economics and Finance, University of Szczecin, Mickiewicza 64, 71-101 Szczecin, Poland; sebastian.gnat@usz.edu.pl

Abstract: The main bases for land taxation are its area or value. In many countries, especially in Eastern Europe, reforms of property taxation, including land taxation, are being carried out or planned, introducing property value as a tax base. Practice and research in this area indicate that such a change in the tax system leads to large changes in land use and reallocation. The taxation of land value requires construction of mass valuation system. Different methodological solutions can serve this purpose. However, mass land valuation requires a large amount of information on property transactions. Such data are not available in every case. The main objective of the paper is to evaluate the possibility of applying selected algorithms of machine learning and a multiple regression model in property mass valuation on small, underdeveloped markets, where a scarce number of transactions takes place or those transactions demonstrate little volatility in terms of real property attributes. A hypothesis is verified according to which machine learning methods result in more accurate appraisals than multiple regression models do, considering the size of training datasets. Three types of models were employed in the study: a multiple regression model, *k* nearest neighbor regression algorithm and XGBoost regression algorithm. Training sets were drawn from a larger dataset 1000 times in order to draw conclusions for averaged results. Thanks to the application of KNN and XGBoost algorithms, it was possible to obtain models much more resistant to a low number of observations, a substantial number of explanatory variables in relation to the number of observations, a low property attributes variability in the training datasets as well as collinearity of explanatory variables. This study showed that algorithms designed for large datasets can provide accurate results in the presence of a limited amount of data. This is a significant observation given that small or underdeveloped real estate markets are not uncommon.

Keywords: land valuation; land reallocation; automated valuation models

Citation: Gnat, S. Property Mass Valuation on Small Markets. *Land* **2021**, *10*, 388. https://doi.org/10.3390/land10040388

Academic Editors: Mirosław Bełej, Małgorzata Krajewska and Izabela Rącka

Received: 21 March 2021
Accepted: 7 April 2021
Published: 8 April 2021

1. Introduction

The system of land taxation has a great influence on the use of land and its reallocation. In countries with established market economies, land and other property is usually taxed on the basis of its value. This system has been in place for hundreds of years. In post-communist countries, land taxation was usually based on its area, which distorted the nature of relations between market participants. Some post-communist countries have already reformed land taxation, while some of them are just considering it; Poland belongs to the latter group. The introduction of property value taxation will significantly change the way real estate market players behave. In order to carry out a reform of property taxation, it is necessary to carry out a mass valuation of property, which is a general term signifying a set of methods used for valuation of a large number of real properties in a uniform manner, determined at the same moment and carried out in a short time period. It could be said that property mass appraisal constitutes a certain general concept. The concept of a uniform approach to the valuation of multiple properties of one type in a short time period does not constitute a tool in itself. Such a tool must only be used within the scope of property mass appraisal. Grover [1] pointed to a series of conditions that need to be met in order to be able to effectively carry out a process of property mass appraisal. According to Grover, the use of instruments of property mass valuation depends on the

degree of property market development and transparency as well as on the institutional structure capable of gathering and keeping up-to-date data on property appraisals and attributes. He also stated that countries introducing mass valuation of real estate may be forced to work on improving institutional bases in that regard, which is a pre-requisite for a successful implementation of mass appraisal. The author rightly believes that in the process of mass valuation, the focus ought not to be just on the improvement of statistical models, but also on the issue of the availability and quality of data used in valuation. Econometric models are one of multiple instruments available in that respect. However, employing them in property mass valuation is not an easy feat or a solution that works in every situation. These models require a series of theoretical and practical conditions to be fulfilled. Failure to satisfy the conditions may lead to appraisal of values significantly different from the actual state of a given real estate market. One of the problems that may be encountered when attempting to use econometric modeling in property mass valuation is the issue of insufficient data required for obtaining a good model. Thus, a model that would ensure achieving adequate appraisal accuracy is needed. A fragment of the market that, on account of various conditions, does not permit obtaining sufficient amount of information is called, for the purposes of this study, an undeveloped market. However, such a market condition does not eliminate the circumstances in which valuation of a large number of real properties located on it is required. Several market and administrative situations can be indicated in which property mass valuation may be useful:

- Monitoring the value of the portfolios of real properties, constituting a security of credit exposures held by a bank [2,3],
- Property valuation for the purpose of estimating the economic effects of adopting or amending local zoning plans,
- General real estate taxation [4],
- Situations in which it is necessary to appraise the value of multiple real properties at the same time.

Transaction prices, which typically constitute the source of data for property mass appraisal in a given period, may refer only to real properties of a specific type (of similar location, similar attributes, etc.). In this study, instead of using information on transaction prices, which may occur in an insufficient number and which may demonstrate little variation both in terms of attributes and location, individual valuations of a drawn real property sample are used, which are here called "representatives." Applying real properties' values instead of transaction prices enables building databases that satisfy the requirements of statistical modeling, and thereby, of property mass appraisal. Thanks to the use of representative properties valuations, it is possible to obtain information on the properties in the entire area covered by property mass valuation. The variability of real property attributes may also be taken into consideration. The process of mass appraisal based on the values determined by property appraisers enables achieving greater variability of real properties in the database used for mass appraisal.

The objective of the paper is to define the effectiveness of applying several types of models used for property mass appraisal in a situation when they are employed on an underdeveloped market, i.e., a market where a low number of transactions takes place or such transactions demonstrate little variety. The models applied include a multiple regression model (in the form proposed by Doszyń [5]) and *k* nearest neighbor regression model as well as XGBoost regression model. With these computation procedures, the value of real properties with two datasets of varying number of observations will be calculated. The size of both datasets is small and it is intended to simulate the so-called underdeveloped market. The accuracy of the resultant valuation will be subject to assessment. As previously mentioned, these types of markets may also require property mass valuation. It is worth investigating whether it is possible to obtain valuations that are close to the ones conducted by licensed property appraisers, while having only a limited number of observations.

Models of property mass valuation are understood as various types of econometric and statistical models, both of parametric nature, in which the value of a property is modeled

on the grounds of an equation comprised of the assessment of structural parameters describing the relations between explanatory variables and a property value or price along with a random component, as well as models of non-parametric nature, where a property value is estimated without a model form through the employment of various methods dividing the applied real property data. Irrespectively of the approaches to property value modeling, it is postulated in the literature that a dataset on a modeled property market ought to be extensive and varied, ensuring suitable data variability and providing an opportunity of determining the relations between property attributes and their prices or values. The last decade in particular has been a period of development of various model applications in property appraisal, and much more broadly, in data modeling in general. Such development is conditional upon two main factors. Firstly, the computing power of contemporary personal computers allows for the use of complex calculations within an acceptable timeframe. Secondly, the 21st century has been the century of data. It is said that data are the oil of the present century, while access to various data is currently easier than ever before. Contemporary scientific literature provides numerous examples of the use of parametric and non-parametric data modeling in the sphere of property appraisal. There is a view presented in the literature that parametric models are mainly applied to examining relations between property attributes and prices, whereas non-parametric models provide a stronger predictive power [6].

A plethora of various scientific works feature a review and classification of property mass appraisal models [7]. In the article, mass valuation methods were divided into non-spatial and spatial models. An interesting review of Automated Valuation Models (AVM) was presented by d'Amato [8]. Various methods (multiple regression models and spatial models) were described in the paper along with the evolution they underwent over the last decades. A general review of quantitative methods applied in mass valuation can also be found in [9]. In the article, the methods were divided into traditional ones (multiple regression as well as comparative, cost-based and income-based valuation methods) and advanced ones, such as artificial neural networks (ANN), spatial analysis, fuzzy logic and ARIMA models. Another comparison of modern approaches in mass appraisal was presented in [10]. In the paper, a comparison was made between modeling approaches such as multiple regression, spatial autoregression (SAR), geographically weighted regression (GWR) and ANNs. Yet another classification of quantitative models used in property mass valuation was undertaken by d'Amato and Kauko [11], who divided valuation methods into four groups: model-based methods, data-based methods, methods based on machine learning as well as expert methods. Wang and Li [12] conducted a review of over 100 articles concerning models and mass valuation methods from the years of 2000–2018. They pointed out that property mass valuation models can be classed into three basic groups: machine learning models (artificial intelligence models), models based on spatial information systems and mixed models. Moreover, they define the so-called mass valuation 2.0, i.e., a procedure of model building, analysis and examination of a property dataset at a given moment, combined with artificial intelligence, geo-information systems and mixed methods, in order to better model property values with reference to both non-spatial and spatial data. Therefore, they see the future of mass valuation in combining the possessed data resources with GIS software and machine learning. It seems that such a vision has a high chance of being fulfilled. An interesting example of using a GIS-based information tool for the evaluation of properties is presented on the example of the Italian corporate real estate market [13]. The main goal of that study was to propose and evaluate model in order to support various institutions involved in the corporate properties market segment. GIS in this scenario allowed to develop a platform for presenting and interpretating obtained results to all, even non-expert users. This is an important feature in the circumstances when mathematically advance models are used and the quantity of data is significant.

The literature concerning the use of machine learning models for property valuation is very extensive and it can be divided into two trends. The first trend encompasses studies in which authors apply and try to improve existing solutions within the framework of multi-

ple regression [14], regression trees [15], random forests [16], support vector machines [17] or artificial neural networks [18–20]. The second trend focuses on the comparison of several algorithms in order to determine which one of them yields better results. An example of such work is the article [21], in which the effectiveness of property prices forecasting was analyzed in Fairfax county, Virginia. In another study, the English housing rental market was subjected to mass appraisal with the use of generalized linear regression, machine learning and expert approach [22]. Two procedures of mass appraisal in the Italian residential property market are presented by Morano, Tajani and Locurcio [23]. The authors tested the utility additive method, which interprets the process of the property price formation as a multi-criteria selection of multi-objective typology, where the selection criteria are the property characteristics that are decisive in the real estate market. This approach is compared to hybrid data-driven technique, called evolutionary polynomial regression, which uses multi-objective genetic algorithms to search those models' expressions that simultaneously maximize accuracy of data and parsimony of mathematical functions. One of the conclusions indicates the possibility of joining presented techniques to obtain more accurate results.

Furthermore, XGBoost algorithm, which is highly recognized both in the sphere of science as well as practice, owing it to its high effectiveness, is employed in property mass appraisal [24]. The algorithm effectiveness was additionally confirmed in article [25] when it was applied on the South Korean property market. Apart from the conclusions regarding the fact that machine learning models proved to be better than multiple regression, the authors state that the application of machine learning is computationally demanding, which has been confirmed in this study as well. In comparative research, artificial neural networks are frequently used as representatives of machine learning. Their superiority over multiple regression models was demonstrated on the case of New York [26]. Furthermore, machine learning models are compared to expert approach [27]. In the study, machine learning algorithms also appeared to be better. Zurada et al. [28] presented comparative research in which several regression methods and artificial intelligence were used to appraise property. The results indicate that non-traditional methods based on regression are slightly superior. Moreover, it was emphasized that the results obtained in the study to a large degree depend on the specificity of a property market, the real property type or the size of an analyzed dataset. Despite the fact that the examples demonstrate an advantage of employing machine learning methods, certain studies can be found which showed no significant differences between, e.g., neural networks and multiple regression, or even studies in which neural networks occurred to be an inferior solution [29]. Such ambiguity of research results indicates the need for conducting further studies in the field of comparing multiple regression with broadly understood machine learning models, particularly in the context of a view claiming that data science and big data constitute the future of real property valuation [30].

The development of modern valuation methods reaches even further. Studies are conducted that test an option of valuing property on the basis of available photographic documentation [31]. In their work, the authors indicate that at present, real estate agents provide their customers with easy online access to detailed information on real properties. Researchers undertook an attempt of valuing a real property price on the grounds of such large amounts of easily available data.

As can be concluded from the presented course of research, the question of employing quantitative methods to real property valuation is extremely broad, starting with multiple regression, through spatial models, to deep neural networks. The models presented here most certainly do not exhaust the subject matter. New proposals are and will be made, the purpose of which is to create quick and reliable mass valuation models. A particular task that stands before researchers is achieving the highest possible accuracy of valuations from a model [32].

When modeling real property values, the stage of particular importance is specifying the variables which have a significant impact on a dependent variable. In their work,

Metzner and Kindt [33] tried to itemize the variables determining the real property values used by researchers all over the world. The results of their work are not hard to guess. Real estate markets demonstrate local characteristics and significant variability. The authors, having reviewed the literature, itemized more than 400 real estate attributes used in mass appraisal models. They postulate the need for determining a certain core in that set of attributes, which would allow creating more stable and comparable valuation models.

In the context of defining property attributes, attention needs to be paid to the second dimension of data used in mass valuation, i.e., the number of observations. Various studies concerning the application of models and computation algorithms frequently fail to undertake the subject of data scarcity. In publications concerning real property valuation, the issue of the impact that data size exerts on model quality is rarely mentioned. The question of small training sets is examined in studies on artificial neural networks [34,35]. It was demonstrated in those studies that despite sparse datasets, it is possible to achieve high-quality results. In the examples of mass property valuation typically presented in literature, the problem of data availability is not raised. Nevertheless, it needs to be remembered that not every local real estate market provides the opportunity of gathering information on a large number of transactions.

Studies related to mass valuation of real estate, including land, in connection with determination of its cadastral value are conducted in different contexts [36,37]. Kilić Pamuković et al. proposed a model to assess the bonitet of private cadastral parcels based on the Expert System (ES) of fuzzy logic within the knowledge component, which would reduce uncertainty and increase the objectivity of the evaluation. Gnat argues that the replacement of tax based on the area of real estate with tax calculated on its value causes significant shifts in the tax burden of individual landowners. He states that the percentage of land plots, the financial burden of which after the introduction of cadastral tax will be close to the current burden of property tax, is small. This indicates that the reform of property taxation will not be a simple replacement of one tax by another but may have a significant impact on the land market. The implementation of land tax reform in Poland will rationalize land use policy. It will prevent peculiar situations in which, despite large demand for land in cities, vacant land will not be developed and will be maintained only for speculative purposes. The increase in value will lead to an increase in tax burdens and will motivate owners to conduct actions generating more income from real estate or to it will force them to sell the land.

The problems of property valuation for tax purposes and the convergence of valuations with market prices are related to the important concept of vertical inequity [38,39]. The authors define progressive and regressive inequity. Vertical inequity occurs when assessed value-to-sales price ratios are not uniform across property value categories. Studies indicate that expensive homes are underassessed more often. The studies regarding inequity present and evaluate different models measuring this phenomenon. They indicate that in addition to linear, linear transformable or simple quadratic relation types, more complex forms of inequity may also exist. They require models suitable to this kind of situation. Benson and Schwartz [40] gave the example of improving the accuracy of valuations for tax purposes in differing property appreciation periods. The use of an appropriate model is, therefore, not only related to the valuation process, but also to the modeling of phenomena that affect the assessment of the tax system by property owners.

2. Materials and Methods

Three types of regression models were used in the research: a multiple regression model (MR), *k* nearest neighbors regression (KNN) and XGBoost. The first one is a parametric model, whereas the remaining two models are non-parametric algorithms.

In the survey, a non-linear multiple regression 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 estates $(i = 1, 2, \ldots, N)$,
J—number of location attractiveness zones $(j = 2, 3, \ldots, J)$,
α_0—constant term,
K—number of real estate attributes,
k_p—number of states of k-th attribute,
α_{kp}—impact of p-th state of attribute k,
x_{kpi}—dummy variable for p-th state of attribute k,
α_j—market value coefficient for j-th location attractiveness zone,
laz_{ji}—dummy variable equal one for j-th location attractiveness zone,
Second bullet;
u_i–random component.

The dependent variable is a natural logarithm of a real estate unit value. Real estate values are determined by certified appraisers in 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 state is skipped. Hence, we arrive at the summation of $p = 2, \ldots, k_p$ in Formula (1). In the interpretation, the ignored state of an attribute serves as a point of reference for the remaining states.

Some research has provided evidence that segmenting property market often improves mass valuation [41]. A procedure of determining submarkets has been introduced in model (1) as well. There are coefficients (α_j) in model (1) that could be treated as a proxy for a location. They are estimated by introducing dummy variables for defined, so-called location attractiveness zones. Location attractiveness zones were in these cases constructed by experts. They 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 (KNN regression and XGBoost).

The k nearest neighbors algorithm is a non-parametric algorithm. Though mainly applied in classification problems, the KNN algorithm can also be used in regression problems [42]. 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 $x(r)$, $r = 1, \ldots, k$ located closest to x_0. In the second step, a prediction is made based on averaging of a 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.

The XGBoost [43] is an open-source library providing the implementation of a gradient boosted decision trees algorithm. XGBoost is an ensemble learning method, involving a combination of the predictive power of multiple models (decision trees in this case). The effect of ensemble learning is an aggregated result from a specific number of models. The models that create an ensemble are defined as base ones and they may be models of the same or of different type. Bagging and boosting are two widely applied approaches in ensemble learning. The most frequently used base models include decision trees. Some of the most important features that cause XGBoost to be so extensively applied include regularization, which helps prevent overfitting, handling sparse data, block structure for effective usage of computer cores and out of core computing, which is helpful when dealing with datasets that do not fit into memory. The algorithm was devised in such a way so that it can operate effectively even in the case of billions of observations. Without a doubt, its testing at the other end of the spectrum of the number of observations is valuable from a scientific perspective.

The most important part of the study involves comparing valuation errors obtained with the use of model (1) and other models. In each case, once model valuations (obtained with the application of a model) have been computed, their error was determined by comparing property appraisers' valuations with the results achieved with regression models. The error is a relative root mean square error (*rRMSE*):

$$RMSE = \sqrt{\frac{\sum_{j=1}^{n}\left(\hat{w}_j - w_j\right)^2}{n}} \tag{2}$$

$$rRMSE = \frac{RMSE}{\overline{w}_j} \tag{3}$$

where:

w_j—actual property value defined by a property surveyor,
$\hat{w}_j$—theoretical property value,
$\overline{w}_j$—mean actual property value,
n—number of real properties,
RMSE—root mean square error,
rRMSE—relative root mean square error.

The error in percentage terms indicates by how much valuations obtained from a model differ on average from the valuations carried out by property appraisers.

The dataset used in the study contains information not on transaction prices, but on real estate values, which were determined by property appraisers in individual valuations. All individual appraisals have been conducted by the group of four certified valuers. In Poland, as well as in other countries, there are several types of real estate value. In this research, the market value of land plots was estimated by appraisers. In a short period, transactions may refer to the real properties having attributes that differ very little. A 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 the attributes were treated as 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). Land plot area is a quantitative variable, but it is treated as a qualitative one. This is because market participants often treat this variable in such a way. This conclusion was also presented by appraisers. With respect to the real estate unit value, it is assumed that a small surface is better than an average one, and the average surface is better than a large one. The use of only qualitative variables in the model is related to the specifics of the real estate valuation methodology used in Poland. It is based on describing the property using several most important characteristics of the property, which determine the value. All these features are described on an ordinal or nominal scale. Mass valuation in this study was intended to mimic the commonly used approach in terms of explanatory variables. It is also worth noting that there were three location attractiveness zones established. Attributes used in the study origin from the dataset obtained from appraisers who conducted evaluation of these properties in the process of recalculation of perpetual usufruct annual fees.

The study encompassed 318 land plots located in one of the largest cities of Poland—Szczecin. The location of the city of Szczecin in Poland is presented in Figure 1. Land plots were developed with residential houses. Recalculation of perpetual usufruct annual fees is conducted, according to Polish regulations, for the land only. Thus, developed plots were treated as undeveloped. Only land was the object of evaluation. The properties' value levels reflected the market prices as of the second half of 2018.

Table 1. Real estate attributes and their states.

No.	Attribute	Attribute Category (State)
1	Utilities	None Incomplete Complete
2	Neighborhood	Onerous Unfavorable Average Favorable
3	Transport availability	Unfavorable Average Favorable
4	Physical plot properties	Unfavorable Average Favorable
5	Plot area	Large (>1200 m^2) Average (500–1200 m^2) Small (<500 m^2)

Figure 1. Location of the city of Szczecin within Poland.

The location of the three designated location attractiveness is presented in Figure 2 and the location of the valuated properties within those zones is presented in Figure 3.

Basic positional measurements calculated for the employed set of 318 real properties are presented in Table 2. Real estate attributes are encoded in such a manner that the worst variant equals 1, a subsequent variant is 2, etc. Min is the minimum value, $Q_{1.4}$ is the first quartile, M is the median, $Q_{3.4}$ is the third quartile, max is the maximum value, Q is the quartile deviation and V_Q is the positional coefficient of variation. 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 neighborhood, the median was equal to the maximum value of an attribute. The variability measured with quartile deviation and positional coefficient of variation was rather small.

Figure 2. Designated location of the attractiveness zones.

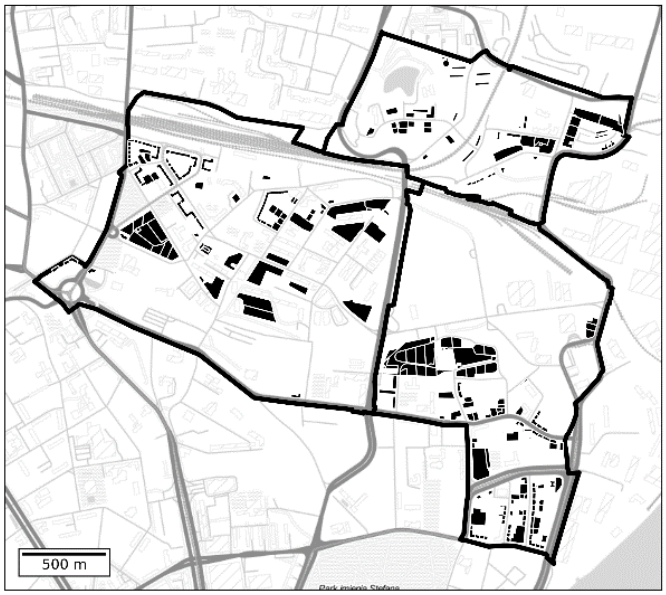

Figure 3. Location of the valued properties.

Table 2. Descriptive statistics in unitary values (in PLN—Polish zlotys) of real properties and their attributes defined for a set of 318 real properties.

Statistics	Values of 1 m^2	Utilities	Neighborhood	Transport Availability	Physical Properties	Plot Area
Min	502.11	3	1	1	1	1
$Q_{1.4}$	569.26	3	3	2	2	2
M	592.28	3	3	3	3	3
$Q_{3.4}$	623.52	3	3	3	3	3
Max	701.43	3	4	3	3	3
Q	27.13	0	0	0.5	0.5	0.5
V_Q (%)	4.58	0	0	16.667	16.667	16.667

3. Results

As previously mentioned, the study encompassed 318 real properties. The value of all the properties was determined by property appraisers and all of the properties

will be subjects of modeling in the study, which was devised in the following manner. By simulating a limitation in the availability of data on underdeveloped markets, two training datasets were drawn 1000 times from the original dataset. The first of them held 118 observations, while the other one held 68. Repeated sampling of training sets was meant to enable the averaging of results and eliminating the risk, of which the results will be characteristic for a single dataset; more general conclusions cannot be constructed on the basis of the results. The multiple regression model (1) was built on the grounds of each of the sets drawn, and with it, the value of all 318 properties was specified. In that manner, the theoretical values were obtained, which were compared to the values determined by the property appraisers. Following that, on the grounds of the same training sets, the values of the properties were determined by employing the KNN and XGBoost algorithms. Then, the valuation errors arising in property value modeling were compared. KNN and XGBoost algorithms are machine learning methods, and one of their characteristics is that they provide a possibility, or even a need, to optimize their input parameters (hyper-parameters), the right selection of which enables achieving better results that feature smaller errors. In both models, the value of the selected hyper-parameters underwent optimization. For the KNN algorithm, the combinations of (k) neighbors and weights used for determining property values on the basis of value in k closest points were tested using a grid search with a cross-validation procedure. The number of neighbors was selected from a range between 3 and 20. In turn, two variants were designated for weights: weights were either based on a property's distance from a neighbor in the space of explanatory variables, or no weights were used for neighbors. For the XGBoost algorithm, which possesses multiple hyper-parameters, the testing involved a maximal depth of a single decision tree and a percentage of explanatory variables accounted for in a single decision tree. Owing to the fact that hyper-parameter optimization was conducted 1000 times, the optimization of a greater number of hyper-parameters of the algorithm was not conducted. The time needed to obtain the results of such an experiment would exceed the acceptable limits.

Kernel density estimations of *rRMSE* distributions for models based on 118 and 68 observations training datasets are presented, respectively, in Figures 4 and 5. Selected measures of the distribution of *rRMSE* errors obtained in individual draws are presented in Tables 3 and 4. From the gathered results, it arises that the multiple regression model generated greater appraisal errors. In a certain portion of draws, the training sets featured high collinearity of explanatory variables and low variability. This resulted in valuations demonstrating high errors. Such unfavorable results to a far greater extend occurred in the case when training sets in models had 68 observations with a total of 13 explanatory variables. Non-parametric models worked better both in the case of 118- and 68-element training sets. Slightly lower mean valuation errors were observed for the XGBoost algorithm. Errors in non-parametric models demonstrated significantly lower variability. This proves that they were more resistant to real properties drawn into the training sets. This is a valid observation, since the collinearity of explanatory variables may frequently occur on underdeveloped markets. In the results obtained on the basis of 118-element training sets, a mean valuation error was approximately 40% greater for multiple regression models than mean errors for KNN and XGBoost models. In the case of smaller sets, the difference was even greater, i.e., approximately 75%, owing to very substantial errors resulting from the model (1) in single trainings sets created unfavorably in some training samples. This is evidenced by maximum recorded valuation errors, which in the case of multiple regression models and XGBoost amounted to, respectively, 14.17% and 5.68% for 118-element training sets and 66.68% and 6.06% for 68-element training sets. Another important observation is that although valuation errors rise as a result of a decrease in training set sizes, in the case of KNN and XGBoost algorithms, those errors grow significantly less than in the case of multiple regression models.

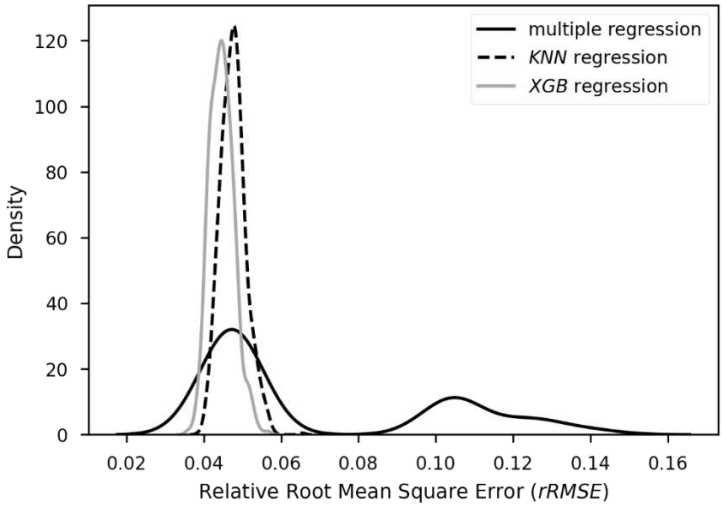

Figure 4. Kernel density estimations of *rRMSE* distributions for models based on training datasets with 118 observations.

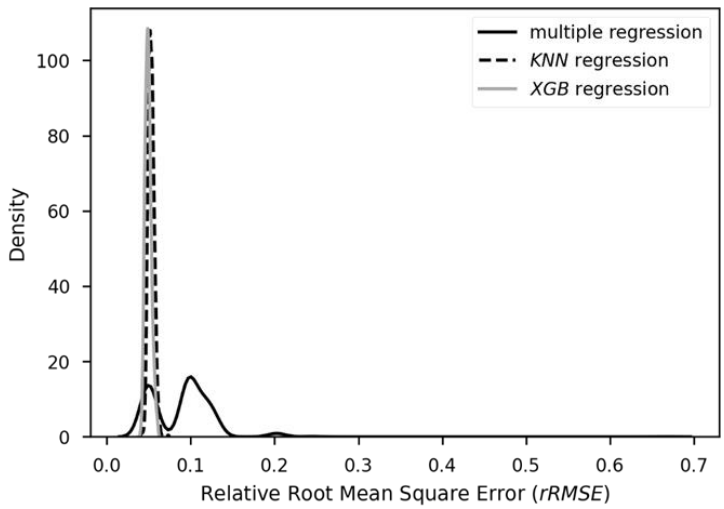

Figure 5. Kernel density estimations of *rRMSE* distributions for models based on training datasets with 68 observations.

Table 3. Selected measures of a central tendency of relative root mean square errors for the analyzed models in 1000 draws (a training set of 118 observations).

Model	Mean	Standard Deviation	Minimum	First Quartile	Median	Third Quartile	Maximum
KNN	4.76%	0.33%	3.88%	4.53%	4.75%	4.95%	6.54%
XGB	4.46%	0.32%	3.58%	4.23%	4.45%	4.67%	5.68%
MR	6.95%	3.18%	4.16%	4.65%	4.86%	10.41%	14.17%

Table 4. Selected measures of a central tendency of relative root mean square errors for the analyzed models in 1000 draws (a training set of 68 observations).

Model	Mean	Standard Deviation	Minimum	First Quartile	Median	Third Quartile	Maximum
KNN	5.29%	0.36%	4.13%	5.04%	5.28%	5.51%	7.31%
XGB	4.93%	0.37%	4.03%	4.67%	4.90%	5.15%	6.06%
MR	9.01%	3.91%	4.42%	5.11%	9.68%	11.37%	66.68%

4. Discussion

The results of applying two non-parametric regression algorithms in property mass valuation on an underdeveloped market were presented in this paper and they were compared to a multiple regression model. The results obtained are highly promising. Thanks to the application of KNN and XGBoost algorithms, it was possible to achieve models that are more resistant to a low number of observations, a substantial number of explanatory variables in relation to the number of observations, low property attributes variability in the drawn datasets as well as collinearity of explanatory variables. Not only were extremely high valuations errors avoided, but mean errors and median errors were lower than in the case of classic multiple regressions models. Such results, obtained in 1000 random samples, allow the author to believe that also in the case of other sets, not only the ones based on individual valuations, but also on transaction prices, the presented non-parametric algorithms will improve the quality of mass valuation. The cost of non-parametric modeling involves losing the possibility to interpret estimation of models' structural parameters. Whether such a cost is acceptable depends on the purpose of constructing a mass appraisal model. If the objective is to precisely define the relations of individual attributes describing properties in relation to their prices, then non-parametric modeling is not the right step. If, however, the objective is to obtain property value as close to reality as possible, then the application of XGBoost algorithm, in particular, seems to be the proper approach. In subsequent research, the intention is to repeat the experiment on other real estate markets with different datasets, in order to determine how repeatable the results produced in the presented study are. Nevertheless, it can already be concluded that even in the event of working on small datasets, one may expect accurate valuation results, which is a highly significant conclusion in the case of mass property valuation on an undeveloped market. The main practical implication of the study is that it demonstrates the ability to effectively use algorithms designed for large datasets in small, underdeveloped real estate markets. A potential property tax reform in Poland that introduces value as a tax base will be carried out nationwide. In some areas, there will not be enough data to effectively apply a variety of property value modeling methods. A study demonstrating the ability to effectively deal with the small amount of data and the resulting consequences may reduce potential concerns about the feasibility of mass valuation in Poland and beyond.

Funding: This research was conducted within the project financed by the National Science Centre, Project No 2017/25/B/HS4/01813. The project is financed within the framework of the program of the Minister of Science and Higher Education under the name "Regional Excellence Initiative" in the years 2019 – 2022; project number 001/RID/2018/19; the amount of financing PLN 10,684,000.00.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy restrictions.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Grover, R. Mass valuations. *J. Prop. Investig. Financ.* **2016**, *34*, 191–204. [CrossRef]
2. Korteweg, A.; Sorensen, M. Estimating loan-to-value distributions. *Real Estate Econ.* **2016**, *44*, 41–86. [CrossRef]
3. Tzioumis, K. Appraisesrs and valuation bias: An empirical analysis. *Real Estate Econ.* **2017**, *45*, 679–712. [CrossRef]
4. Bradbury, K.L.; Mayer, C.J.; Case, K.E. Property tax limits, local fiscal behavior, and property values: Evidence from Massachusetts under Proposition 2,5. *J. Public Econ.* **2001**, *80*, 287–311. [CrossRef]
5. Doszyń, M. Econometric support of a mass valuation process. *Folia Oeconomica Stetin.* **2020**, *20*, 81–94. [CrossRef]
6. Pérez-Rave, J.; Correa-Morales, J.; González-Echavarría, F. A machine learning approach to big data regression analysis of real estate prices for inferential and predictive purposes. *J. Prop. Res.* **2019**, *36*, 59–96. [CrossRef]
7. Jahanshiri, E.; Buyong, T.; Shariff, A.R.M. A review of property mass valuation models. *Pertanika J. Sci. Technol.* **2011**, *19*, 23–30.
8. d'Amato, M. *Advances in Automated Valuation Modeling*; d'Amato, M., Kauko, T., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2017.

9. Pagourtzi, E.; Assimakopoulos, V.; Hatzichristos, T.; French, N. Real estate appraisal: A review of valuation methods. *J. Prop. Investig. Financ.* **2003**, *21*, 383–401. [CrossRef]
10. McCluskey, W.J.; McCord, M.; Davis, P.T.; Haran, M.; McIllhatton, D. Prediction accuracy in mass appraisal: A comparison of modern approaches. *J. Prop. Res.* **2013**, *30*, 239–265. [CrossRef]
11. Kauko, T.; d'Amato, M. (Eds.) *Mass Appraisal Methods: An International Perspective for Property Valuers*; Blackwell Publishing Ltd.: Hoboken, NJ, USA, 2008.
12. Wang, D.; Li, V. Mass appraisal models of real estate in the 21st century: A systematic literature review. *Sustainability* **2019**, *11*, 7006. [CrossRef]
13. Locurcio, M.; Morano, P.; Tajani, F.; Di Liddo, F. An Innovative GIS-Based Territorial Information Tool for the Evaluation of Corporate Properties: An Application to the Italian Context. *Sustainability* **2020**, *12*, 5836. [CrossRef]
14. Zaddach, S.; Alkhatib, H. Least squares collocation as an enhancement to multiple regression analysis in mass appraisal applications. *J. Prop. Tax Assess. Adm.* **2014**, *11*, 47.
15. McCluskey, W.J.; Daud, D.Z.; Kamarudin, N. Boosted regression trees. *J. Financ. Manag. Prop. Constr.* **2014**, *19*, 152–167. [CrossRef]
16. Antipov, E.A.; Pokryshevskaya, E.B. Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics. *Expert Syst. Appl.* **2012**, *39*, 1772–1778. [CrossRef]
17. Wang, X.; Wen, J.; Zhang, Y.; Wang, Y. Real estate price forecasting based on SVM optimized by PSO. *Optik* **2014**, *125*, 1439–1443. [CrossRef]
18. Ćetković, J.; Slobodan, L.; Lazarevska, L.; Žarković, M.; Vujošević, S.; Cvijović, J.; Gogić, M. Assessment of the real estate market value in the european market by artificial neural networks application. *Complexity* **2018**, *2018*, 1472957. [CrossRef]
19. Demetriou, D. A spatially based artificial neural network mass valuation model for land consolidation. *Environ. Plan. B Urban Anal. City Sci.* **2017**, *44*, 864–883. [CrossRef]
20. Zhou, G.; Ji, Y.; Chen, X.; Zhang, F. Artificial neural networks and the mass appraisal of real estate. *Int. J. Online Eng.* **2018**, *14*, 180–187. [CrossRef]
21. Park, B.; Bae, J.K. Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert Syst. Appl.* **2015**, *42*, 2928–2934. [CrossRef]
22. Clark, S.D.; Lomax, N. A mass-market appraisal of the English housing rental market using a diverse range of modelling techniques. *J. Big Data* **2018**, *5*, 43. [CrossRef] [PubMed]
23. Morano, P.; Tajani, F.; Locurcio, M. Multicriteria analysis and genetic algorithms for mass appraisals in the Italian property market. *Int. J. Hous. Mark. Anal.* **2018**, *11*, 229–262. [CrossRef]
24. Zhao, Y.; Chetty, G.; Tran, D. Deep learning with XGBoost for real estate appraisal. In Proceedings of the IEEE Symposium Series on Computational Intelligence (SSCI), Xiamen, China, 6–9 December 2019; pp. 1396–1401. [CrossRef]
25. Kim, Y.; Choi, S.; Yi, M.Y. Applying comparable sales method to the automated estimation of real estate prices. *Sustainability* **2020**, *12*, 5679. [CrossRef]
26. Khamis, A.; Kamarudin, N.K. Comparative study on estimate house price using statistical and neural network model. *Int. J. Sci. Technol. Res.* **2014**, *3*, 126–131.
27. Trawiński, B.; Telec, Z.; Krasnoborski, J.; Piwowarczyk, M.; Talaga, M.; Lasota, T.; Sawilow, E. Comparison of expert algorithms with machine learning models for real estate appraisal. In Proceedings of the IEEE International Conference on INnovations in Intelligent SysTems and Applications (INISTA), Gdynia, Poland, 3–5 July 2017. [CrossRef]
28. Zurada, J.; Levitan, A.S.; Guan, J. A comparison of regression and artificial intelligence methods in a mass appraisal context. *J. Real Estate Res.* **2011**, *33*, 349–388. [CrossRef]
29. Del Giudice, V.; De Paola, P.; Forte, F.; Manganelli, B. Real estate appraisals with bayesian approach and Markov chain hybrid Monte Carlo method: An application to a central urban area of Naples. *Sustainability* **2017**, *9*, 2138. [CrossRef]
30. Dell, G. Regression, critical thinking and the valuation problem today. *Apprais. J.* **2017**, *85*, 217–230.
31. You, Q.; Pang, R.; Cao, L.; Luo, J. Image-based appraisal of real estate properties. *IEEE Trans. Multimed.* **2017**, *19*, 2751–2759. [CrossRef]
32. Bogin, A.N.; Shui, J. Appraisal Accuracy and Automated Valuation Models in Rural Areas. *J. Real Estate Financ. Econ.* **2020**, *60*, 40–52. [CrossRef]
33. Metzner, S.; Kindt, A. Determination of the parameters of automated valuation models for the hedonic property valuation of residential properties: A literature-based approach. *Int. J. Hous. Mark. Anal.* **2018**, *11*, 73–100. [CrossRef]
34. Shaikhina, T.; Lowe, D.; Daga, S.; Briggs, D.; Higgins, R.; Khovanova, N. Machine learning for predictive modelling based on small data in biomedical engineering. *IFAC-PapersOnLine* **2015**, *48*, 469–474. [CrossRef]
35. Barz, B.; Denzler, J. Deep learning on small datasets without pre-training using cosine loss. *arXiv* **2019**, arXiv:1901.09054. [CrossRef]
36. Kilić Pamuković, J.; Rogulj, K.; Jajac, N. Assessing the Bonitet of Cadastral Parcels for Land Reallocation in Urban Consolidation. *Land* **2021**, *10*, 9. [CrossRef]
37. Gnat, S. Analysis of communes' potential fall in revenue following introduction of ad valorem property tax. *Real Estate Manag. Valuat.* **2018**, *26*, 1. [CrossRef]
38. Sirmans, G.S.; Diskin, B.A.; Friday, H.S. Vertical Inequity in the Taxation of Real Property. *Natl. Tax J.* **1995**, *48*, 71–84. [CrossRef]

39. Sunderman, M.A.; Birch, J.W.; Cannaday, R.A.; Hamilton, T.W. Testing for Vertical Inequity in Property Tax Systems. *J. Real Estate Res.* **1990**, *5*, 319–334. [CrossRef]
40. Benson, E.D.; Schwartz, A.L., Jr. An Examination of Vertical Equity Over Two Reassessment Cycles. *J. Real Estate Res.* **2000**, *19*, 255–274. [CrossRef]
41. Usman, H.; Lizam, M.; Adekunle, M.U. Property price modelling, market segmentation and submarket classifications: A review. *Real Estate Manag. Valuat.* **2020**, *28*, 24–35. [CrossRef]
42. Pace, R.K. Relative performance of the grid, nearest neighbor, and OLS estimators. *J. Real Estate Financ. Econ.* **1996**, *13*, 203–218. [CrossRef]
43. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [CrossRef]