REMV
REAL ESTATE MANAGEMENT AND VALUATION

Journal homepage:
www.remv-journal.com

HOUSING PRICE PREDICTION - MACHINE LEARNING AND GEOSTATISTICAL METHODS

Radosław Cellmer[1]*, Katarzyna Kobylińska[2]

[1] *Department of Real Estate and Urban Studies, University of Warmia and Mazury in Olsztyn, Prawochenskiego 15, 10-724 Olsztyn, Poland, (RC) e-mail: rcellmer@uwm.edu.pl, ORCID: 0000-0002-1096-8352; (KK) e-mail: katarzyna.kobylinska@uwm.edu.pl, ORCID: 0000-0001-6183-3215*

** Corresponding author*

ARTICLE INFO

Keywords:

machine learning, housing prices, geostatistics

JEL Classification:

C45, C53, R20, R32

ABSTRACT

Machine learning algorithms are increasingly often used to predict real estate prices because they generate more accurate results than conventional statistical or geostatistical methods. This study proposes a methodology for incorporating information about the spatial distribution of residuals, estimated by kriging, into selected machine learning algorithms. The analysis was based on apartment prices quoted in the Polish capital of Warsaw. The study demonstrated that machine learning combined with geostatistical methods significantly improves the accuracy of housing price predictions. Local factors that influence housing prices can be directly incorporated into the model with the use of dedicated maps.

Citation: Cellmer, R., & Kobylińska, K. (2025). Housing price prediction - Machine learning and geostatistical methods. *Real Estate Management and Valuation, 33*(1), 1-10. https://doi.org/10.2478/remav-2025-0001

1. Introduction

Accurate price prediction plays a key role in the decision-making process on the real estate market. In addition to buyers and sellers, the housing market involves other stakeholders, such as the local authorities, real estate agents, financial institutions, and property developers. The real estate market is an important part of the economy because it can drive changes on the stock market, or even catalyze destructive economic events (Park & Bae, 2015). Therefore, accurate methods for predicting market performance are needed, in particular in the face of contemporary global challenges. Construction activity and employment in the real estate sector increase in periods of economic expansion, which increases property prices. These trends are reversed during an economic slowdown. Therefore, accurate price predictions play a very important role in social and economic development, both at the local and the national level (Gu et al., 2011; Plakandaras et al., 2015).

Machine learning algorithms are increasingly often used to predict prices on the real estate market. These methods are being continuously improved, and they generate much more accurate results than conventional econometric models (Çılgın & Gökçen, 2023). Thus, the purpose of the research presented here is not to show that machine learning is more effective than classical regression models, but to demonstrate that the prediction accuracy of machine learning algorithms can be improved in a relatively simple way, by including geostatistical methods in the training process. The spatial distribution of residuals can be incorporated into the model to account for the influence of local factors on real estate prices and to minimize prediction errors. In existing solutions for the application of machine learning to price prediction, location is generally taken into account by a number of additional apriori assumed variables, e.g. distances from characteristic locations or directly in the form of coordinates. The novel approach of the presented method is to assume that all variability (or most of it) caused by location factors is reflected in the random error (residuals) of the model. In this way, all locational factors are included, including those that are difficult to identify directly.

The article is divided into several sections. The applicability of machine learning algorithms for predicting real estate prices, in particular in models that incorporate geospatial data, is briefly discussed in the literature review section. Data sources and the research methodology are described in the third section. The empirical results are presented and

discussed, and potential limitations of the study are identified in the fourth section. The advantages of the proposed approach are summarized, and conclusions are formulated in the last section of the article.

2. Literature review

For decades, hedonic models have been used in empirical research to analyze real estate prices and rents based on property attributes such as utilities and location (Lorenz et al., 2023). However, due to the growing popularity of machine learning algorithms, these methods will probably emerge as one of the main tools for predicting property prices in the near future (Choy & Ho, 2023). The advantages of machine learning algorithms and their applicability for predicting the performance of the real estate market have been extensively investigated in recent years (Hamizah Zulkifley et al., 2020). Machine learning algorithms undoubtedly contribute to the accuracy of housing price prediction models by utilizing advanced computational techniques to analyze large data sets and identify complex patterns in housing market data (Park & Bae, 2015). Selected algorithms were reviewed by Baneriee and Dutta (2017) based on a data set acquired from kaggle.com. They concluded that random forest and support vector machines were particularly useful for predicting property prices. Similar observations were made by Chowhaan et al. (2023) who found that complex algorithms are most suitable for processing large data sets. Machine learning algorithms have been tested on property prices and rents in Fairfax County (Park & Bae, 2015), Washington (Oladunni & Sharma, 2016), Madrid (Baldominos et al., 2018), Thessaloniki (Georgiadis, 2018), Beijing (Truong et al., 2020), New Taipei (Alenany et al., 2021), Frankfurt (Lorenz et al., 2023), Hong Kong (Choy & Ho, 2023), and Ankara (Çılgın & Gökçen, 2023). According to many researchers, machine learning algorithms, in particular random forest and xgboost techniques, outperform hedonic models in terms of accuracy and efficacy (Choy & Ho, 2023). Research has also shown that not only regression algorithms, but also classification algorithms are useful for predicting prices (Durganjali & Pujitha, 2019; Jha et al., 2020; Thamarai & Malarvizhi, 2020).

In most published studies, the analyses were conducted based on transaction price data, but real estate listings are also a valuable source of information that is easily available and up-to-date (Park & Bae, 2015). Listings do not always accurately reflect transaction prices, but they provide information about trends on the real estate market (Mora-Garcia et al., 2022). Despite the fact that property listings have been successfully used to train machine learning models, the accuracy of the resulting predictions can be undermined by the sellers' excessive optimism or susceptibility to manipulation (Rico-Juan & Taltavull de La Paz, 2021). The main advantage of this data source is that it can be easily acquired through web scraping (Üzümcü & Eliguzel, 2023).

A reliable assessment of location factors plays a very important role in the process of predicting property prices. In general, location denotes the proximity or remoteness of spatial features that could influence the property's price (Choy & Ho, 2023; Çılgın & Gökçen, 2023; Mora-Garcia et al., 2022). Location can also be directly incorporated into machine learning models as geographic coordinates resulting from geocoding, which significantly increases the accuracy of price predictions (Tchuente & Nyawa, 2022). Location factors can also be evaluated with the use of Google Street View images, but this procedure requires additional algorithms (Kang et al., 2021). Geostatistical methods can be widely applied in housing price predictions because location can be regarded as a continuous spatial variable that can be interpolated (Cellmer, 2014). Morillo Balsera et al. (2018) relied on geostatistical methods to predict housing prices in Madrid and concluded that the results of ordinary kriging can improve the accuracy of predictions made by artificial neural networks. Kim et al. (2022) combined geostatistical methods with machine learning algorithms trained on housing transaction prices in Seoul and found that spatial data can be effectively used to train machine learning models. Geostatistical tools were also successfully used to analyze the real estate market in Fukushima, where nine machine learning algorithms and the kriging-regression method were tested for their ability to estimate land prices (Derdouri & Murayama, 2020). However, geostatistical methods are often applied to complement machine learning or, depending on the research concept, machine learning can be used to expand the capabilities of conventional geostatistical methods (Ren et al., 2023).

A review of the literature indicates that geostatistical methods can support machine learning algorithms. Therefore, the present study was undertaken to propose a method for incorporating information about the spatial distribution of residuals estimated by kriging into machine learning algorithms

eISSN: 2300-5289

to increase the accuracy of housing price predictions.

3. Materials and Methods

The analysis was conducted on the real estate market in the Polish capital city of Warsaw, which has a population of more than 1.8 million and spans an area of around 517 km^2. The Warsaw real estate market can be regarded as a highly developed market due to the city's large area and population, and high levels of property development. Warsaw is also characterized by the highest number of property transactions and listings in Poland.

Apartment listings on the gratka.pl website were the source of data for the analysis. A total of 4958 listings were acquired on 15 December 2023 as training data and 2349 listings were acquired on 15 January 2024 as testing data by web scraping. The two data sets are disjoint (have no common elements), which is a necessary prerequisite for eliminating data leaks. The actual number of listings was higher (more than 8000 items in the training set), but incomplete listings (such as offers without a listing price) were eliminated from the data set during a preliminary analysis. The analysis covered a period of one month, which is sufficiently long for new listings to appear on the market, but also sufficiently short to ensure price stability. The distribution of listings in the training set and the testing set is presented in Figure 1.

In both sets, data followed an approximately normal distribution (log-normal distribution). The analyzed housing attributes are presented in Table 1. The selection of variables was determined, on the one hand, by the availability of information contained in the sales offers, while on the other hand, by their substantive importance, as factors which potential buyers of real estate pay particular attention to. Due to the characteristics of data distribution and the possible presence of non-linear relationships, the natural logarithm of the unit price was adopted as the response variable. For the presentation of the results, the logarithms were again converted to 1m^2 prices. Basic descriptive statistics for the training and test set are shown in Table 2

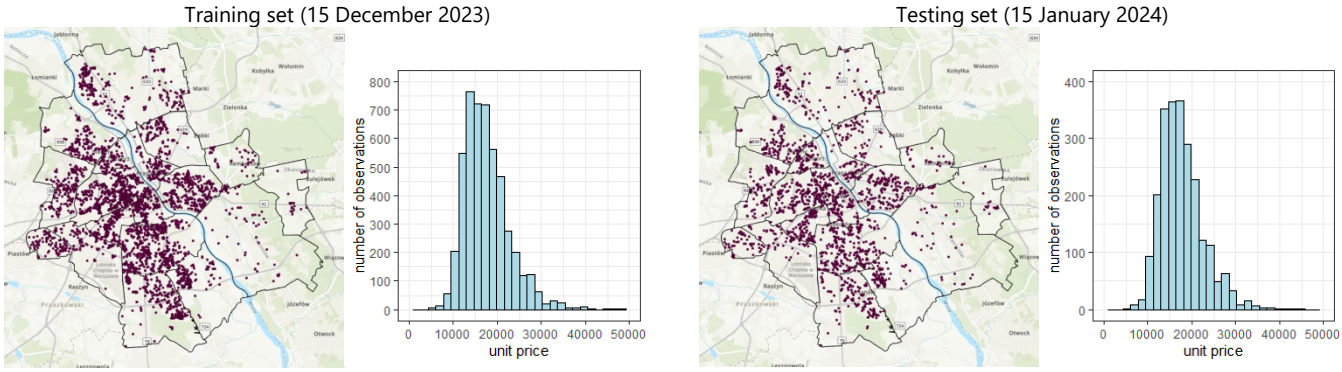

Fig. 1. Distribution of apartment listing prices. *Source*: own elaboration.

Table 1

Data description

Feature	Description	Type
price	price per square meter	numerical
area	apartment area expressed in square meters	numerical
rooms	number of rooms in the apartment	numerical
floor	location on a specific floor	categorical
high	number of floors in the building	numerical
year	year of construction	numerical
district	location in a specific district	categorical
type	type of residential building	categorical

Source: own study.

Table 2

Basic descriptive statistics of the variables accepted for analysis*

	Training set					Testing set				
	min	median	mean	max	sd	min	median	mean	max	sd
price	4424.05	16784.06	17620.04	67961.17	5392.39	5934.96	17058.82	17738.38	50043.08	4966.80
area	12.50	56.76	65.36	197.45	36.61	17.50	56.49	66.04	150.00	38.77

eISSN: 2300-5289 | Received 2024-04-18 | Revised 2024-07-29 | Accepted 2024-09-29

rooms	1	3	2.73	7	1.01	1	3	2.72	7	1.03
high	1	5	5.85	30	3.85	1	5	5.85	30	3.60
year	1505	2008	1995.89	2023	30.58	1895	2003	1993.78	2024	29.35

**floor, district and type were included as dummy variables*

Source: own study.

Those machine learning algorithms that are most widely used and are effective predictive tools for solving regression problems were selected for the study. The following algorithms were used in the analysis: multiple regression, regression tree, bagging, random forest, extreme gradient boosting (XGBT), and support vector machines (SVM).

Ordinary least squares (OLS) linear multiple regression is the basic econometric model which is applied to find the best equation describing a linear relationship between the response variable and explanatory variables. The algorithm finds the optimal values of regression coefficients which minimize the sum of squared residuals.

The response tree (RT) algorithm builds a decision tree, which is a hierarchical structure composed of nodes and branches. Similarly to conventional decision trees, the RT algorithm divides the data set into subgroups based on the values of the standard deviation reduction (SDR) statistic to express the logical relationship between the elements (Ho et al., 2021). This process is repeated for each subgroup, until maximum tree depth is achieved. In the next step, each point in the data set is assigned a value that is predicted based on the regression equation nested in tree leaves which were obtained by data splitting (Loh, 2011).

Bagging (BAG) is one of the simplest methods for averaging predictions. In decision trees, the bagging algorithm generates multiple bootstrap samples in the training set based on repeated random splits of data. An individual model is created for each subsample, and it is used to make individual predictions. The results are averaged, which decreases variance in the model, improves the model's stability, and minimizes the risk of overfitting (Sutton, 2005).

The random forest (RF) algorithm combines classification and regression trees. This algorithm was originally proposed by Breiman et al. in 1984 (Breiman et al., 2017). To increase prediction accuracy, numerous decision trees are generated by bootstrapping, and they are trained with the use of the original data set. The predictions are averaged to minimize error variance (Breiman et al., 2017). The main difference between random forest and bagging is that the RF algorithm considers a random subset of attributes from the data set; therefore, each tree is trained on a different set of explanatory variables. In ordinary bagging, each tree is trained on a full set of attributes (Ho et al., 2021).

Extreme gradient boost (XGBT) was proposed by Chen and Guestrin in 2016 (Chen & Guestrin, 2016), and it is one of the most popular machine learning algorithms. This algorithm boosts the effectiveness of decision trees by sequentially training a series of them and correcting prediction errors made by the previous trees in each one. As a result, the model's accuracy is increased in each iteration. The main advantage of the XGBT algorithm is that it introduces a regularization technique for minimizing classification errors by filtering decision trees with an excessive number of observations in training nodes (Santhanam et al., 2016).

The support vector machines (SVM) algorithm uses optimization techniques to find a hyperplane in multidimensional space that divides input data into decision classes. Initially, this algorithm was used mainly to solve classification problems (Hearst et al.,, 1998). At present, SVM is also applied to solve regression problems by finding a line that best fits the data. Multiple regression minimizes the mean squared error between the observed and predicted values, whereas the SVM algorithm maximizes the distance between the hyperplane and the nearest data points which are referred to as support vectors (Noble, 2006).

The analyzed data were also spatially interpolated with the use of geostatistical methods. Ordinary kriging was selected for spatial interpolation because other kriging methods require additional assumptions that are difficult to fulfill (for example, in simple kriging, the predicted value of the analyzed process is constant over the entire area). The use of kriging for analyzing the spatial distribution of prices on the property market has been described extensively in the literature (Cellmer, 2014).

The research procedure is presented in Figure 2. Data analyses were conducted separately for every selected algorithm (OLS, RT, RF, BAG, XGBT, SVM).

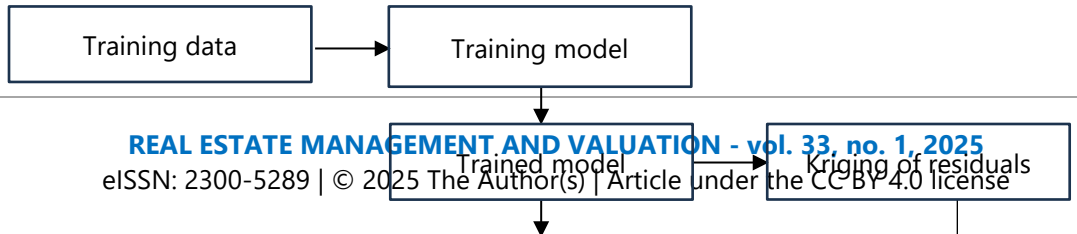

Fig. 2. Research procedure. *Source*: own elaboration.

In the first step, prices were predicted with the use of trained models based on the explanatory variables in testing data sets. Testing data sets can be expected to provide a weaker fit between predicted and observed prices than training data sets because data for both sets were acquired on different, but not too distant, dates.

Data were analyzed with the use of geostatistical methods to increase the degree of fit between predicted and observed prices. The analysis was conducted on the assumption that housing prices are affected not only by the adopted explanatory variables (Table 2), but also by local factors. The influence of local factors can be determined by analyzing the spatial distribution of residuals in each model, where the residual for each observation is the difference between the observed and the predicted value. As a result, the observed price is the sum of the predicted price and the residual. In a residual analysis, the value of a residual computed from the training set, rather than the testing set, is added to the predicted value to avoid data leaks. However, the locations of testing and training data generally do not overlap, and the value of a residual at a given point can be difficult to estimate. This problem can be resolved by kriging, which is a spatial interpolation method for estimating the unknown value of a variable at a given point based on points whose value is known. The final price prediction is the sum of preliminary predictions made based on testing data, the trained model, and a hypothetical residual from a residual map developed based on training data samples.

After the final prediction, the results were validated by comparing the predicted prices with the observed prices. The model's fit was assessed with the use of the following metrics: mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) (Table 3).

Table 3

Metrics for evaluating prediction accuracy

Metric	Abbreviation	Formula
Mean squared error	MSE	$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - y_i^p\right)^2$
Root mean squared error	RMSE	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - y_i^p\right)^2}$
Mean absolute error	MAE	$\frac{1}{n}\sum_{i=1}^{n}\left\|y_i - y_i^p\right\|$
Mean absolute percentage error	MAPE	$\frac{100\%}{n}\sum_{i=1}^{n}\frac{\left\|y_i - y^p\right\|}{y_i}$

y_i - *observed values,* y_p - *predicted values, i – index of observations, n – number of observations*

Source: own study.

4. Results and Discussion

Each model was trained with the use of the caret

eISSN: 2300-5289 | Received 2024-04-18 | Revised 2024-07-29 | Accepted 2024-09-29

package in R. Some parameters in the analyzed data set (floor, district, type) were qualitative attributes that were expressed as categories. Categorical variables had to be transformed into dummy (binary) variables. Before training, the training data set was divided into five folds by k-fold cross validation (Wong & Yeh, 2020). In successive steps, each fold contained a validation data set, and the remaining folds were used to train the models and select hyperparameters. The data set was also divided into folds to prevent model overfitting and weak prediction results based on testing data. Hyperparameters were selected iteratively by successive approximations, excluding the OLS model, where the modeled constant was the only hyperparameter. The final model was developed based on all training data with the use of adjusted hyperparameter values. The resulting models' fit to training data is presented in Table 4.

Table 4

Models' goodness-of-fit to training data

	MSE	RMSE	MAE	MAPE
OLS	14374625.47	3791.39	2668.67	14.96
RT	6537684.41	2556.89	1697.32	9.39
RF	1809592.22	1345.21	820.40	4.48
BAG	7935208.70	2816.95	1862.23	10.23
XGBT	4658332.59	2158.32	1483.57	8.26
SVM	14968733.91	3868.94	2643.59	14.59

Source: own study.

The goodness-of-fit analysis was performed only for training data, and is not a reliable indicator of the quality of predictions. The selection of optimal hyperparameter values can be a highly complex task that requires not only extensive knowledge about the studied phenomenon and the applied tool, but also a certain degree of intuition (Yang & Shami, 2020). Therefore, it is possible that a different configuration of variables could yield more accurate results. The goodness-of-fit analysis revealed that decision tree algorithms generate superior results to the multiple regression model. The results of the SVM model were similar to those generated by the OLS model, but the SVM algorithm involved only a linear kernel, which can pose a certain limitation in analyses of non-linear correlations. However, the obtained results confirm the observation that SVM algorithms are highly effective in solving classification problems, but may deliver less satisfactory results in regression problems (Hearst et al., 1998). The value of MAPE was relatively low in the RF algorithm (below 5%), but this result could be attributed to model overfitting.

In the next step, the residuals (differences between the observed values and the predicted values) were assessed in each model to determine their applicability for interpolation and the development of spatial distribution maps. Positive spatial autocorrelations with a p-value of less than 0.05 were noted in all cases; therefore, geostatistical methods (kriging) were applied. The residuals were interpolated by ordinary kriging with semivariograms adjusted to a spherical model. Semivariogram parameters were selected independently for each model. The spatial distribution of the residuals for each model is presented in Figure 3.

The greatest variations in the spatial distribution of residuals were noted in the OLS and SVM models, and similar observations were made based on the values of the calculated metrics. Residual values were highest in the central districts of Warsaw (Śródmieście, western part of Praga-Południe, and north-western part of Mokotów) and lowest in south-eastern districts (Ursynów and Wilanów). The location (district) of the analyzed properties was taken into consideration in each model, but local factors significantly influenced the results, despite the fact that they had a limited reach within every district.

Trained models and residual distribution maps were used to predict listing prices as the sum of the value predicted by the model and the value of the residual read from the map. Based on the adopted research procedure (Fig. 2), training data were used to train models and develop residual distribution maps, whereas testing data were a closed set that had not been used in previous calculations and should be independent from training data. Housing prices were initially predicted based on the attributes in the testing set, and the values of the residuals were read in each location where a new property was placed on the market. The sum of these values was the final price prediction. Therefore, the quality of the prediction was evaluated based on testing data, and the predicted

values were compared with the observed values of new listings. The scatterplots of the predicted and observed prices, including the results of preliminary predictions (without residuals), are presented in Figure 4.

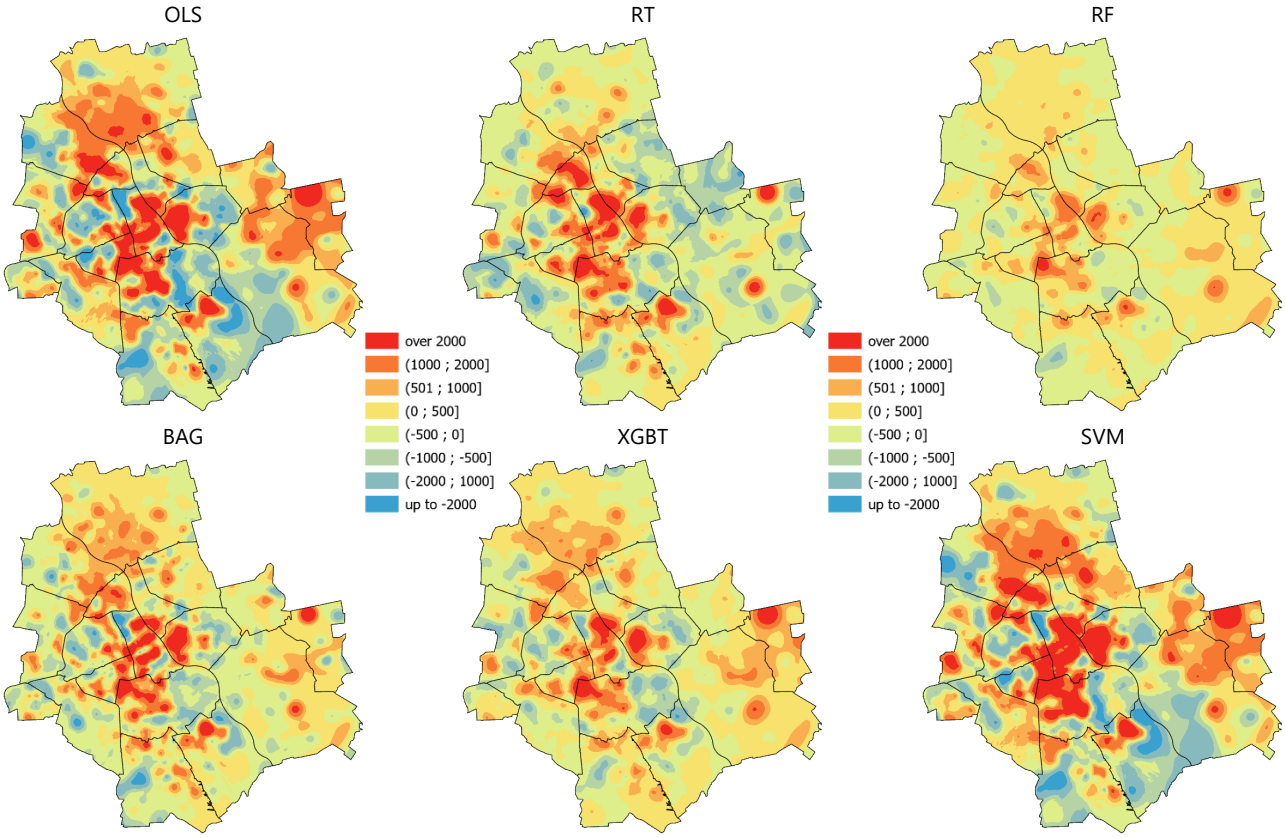

Fig. 3. Spatial distribution of residuals (PLN/m^2) calculated based on training data. *Source*: own elaboration.

Linear fit was improved in all analyzed models, and the greatest improvement was noted in OLS and SVM models, which yielded highly similar results. The prediction accuracy of each model was quantified with the use of selected metrics in Table 5.

The presented results indicate that prediction accuracy was highest in the RF model, regardless of the adopted metric. The MAPE was less than 10% in the RF model, but it exceeded 13% in the OLS model. The RMSE ranged from 2749 to 3333 PLN/m^2, whereas MAE values were much lower in the range of 1719-2379 PLN/m^2 because this metric is less sensitive to outliers, and the testing data set comprised prices that significantly diverged from the average price (Fig. 4).

The analysis revealed that residuals from model fitting improve the accuracy of the prediction, regardless of the adopted evaluation method. The improvement in prediction accuracy is particularly noticeable when RMSE values are taken into consideration. In the SVM model, RMSE decreased by nearly 500 PLN/m^2, which implies that the model's fit decreased by around 12.5% in absolute terms. The smallest change in the values of the adopted criteria was observed in the RF model, which was also characterized by the lowest metric values, including in the variant without residual kriging.

Prediction accuracy was also influenced by observations that differed considerably from average prices and can be regarded as outliers (difference of more than three standard deviations). The empirical distribution of apartment listing prices (Fig. 1) indicates that outliers occurred mainly on the right side of the histogram and led to price underestimation in each case. According to the author, any attempts to remove outliers indiscriminately can distort the testing data set and lead to biased results.

OLS OLS with kriging RT RT with kriging

eISSN: 2300-5289 | Received 2024-04-18 | Revised 2024-07-29 | Accepted 2024-09-29

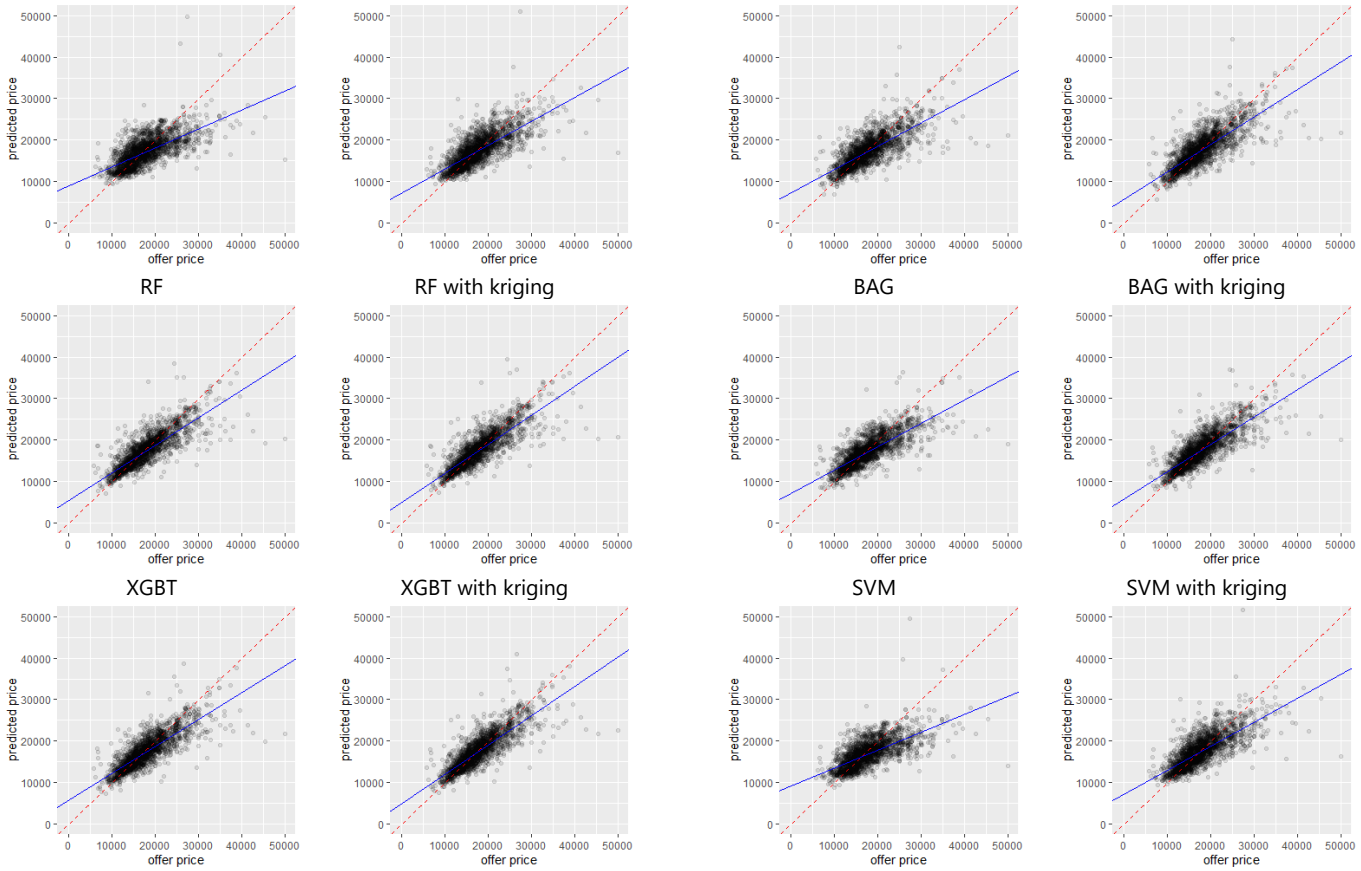

Fig. 4. Scatterplots of predicted and observed property prices based on testing data. *Source*: own elaboration.

Table 5

Prediction accuracy of the analyzed models without and with residual kriging

	MSE		RMSE		MAE		MAPE	
OLS	14079954.75	11109902.85	3752.33	3333.15	2672.39	2378.69	14.93	13.52
RT	11205829.68	9763325.26	3347.51	3124.63	2276.02	2119.60	12.68	11.99
RF	8088911.28	7540685.38	2844.10	2746.03	1781.14	1718.90	9.90	9.64
BAG	10280203.13	8570367.32	3206.27	2927.52	2197.72	2006.97	12.21	11.38
XGBT	8503755.91	7556888.17	2916.12	2748.98	1958.28	1825.63	10.91	10.30
SVM	14550227.42	11095818.22	3814.48	3331.04	2683.61	2368.73	14.74	13.43

Source: own study.

Prediction accuracy is undoubtedly affected by the specificity of the adopted models and the choice of hyperparameters in the training process. A number of other models, e.g., quantile regression or spatial econometric models, can also be considered during the analysis, although prediction based on spatial models may pose additional problems related to the construction of spatial weighting matrices. However, the aim of the present study was not to select the best model, but to demonstrate that spatial information about the distribution of residuals can increase prediction accuracy. The results could suggest that the RF model is most effective, but in reality, prediction accuracy is affected by numerous factors, including the set of explanatory variables, choice of hyperparameters, or splitting of the training data set into folds. Every additional variable (quality standards, neighborhood, utility costs, availability of parking) would undoubtedly improve the results, but not all listings contain a full list of property attributes. During the study, possible misspecifications due to improper selection of variables were not tested (this was not the subject of the study), but for practical applications of the presented method, this would be a necessary step. The assumption of independence is also an important consideration in training and testing data sets, in particular when some properties from the training set are also present in the testing set because they have been listed by different agencies and have somewhat different descriptions.

eISSN: 2300-5289

Accurate information about a listed property's location also plays a very important role in analyses of geospatial data. In online listings, a property's location is described with a certain margin of error, which can also affect the accuracy of prediction, in particular when geostatistical methods are used. However, despite these limitations, the results generated by the analyzed models can be regarded as promising.

5. Conclusions

Apartment prices quoted on the real estate market in Warsaw were predicted with the use of selected machine learning algorithms combined with geostatistical methods (residual kriging). The study demonstrated that information about the spatial distribution of errors generated by the analyzed models can substantially increase the accuracy of predictions. Local factors that influence housing prices can be directly incorporated into the model with the use of dedicated maps. Thus, a new approach to the problem presented in the paper was to use additional information about local location-induced price volatility in price prediction. The greatest improvement in prediction accuracy was noted in models that generated the largest errors during preliminary price prediction.

The proposed approach can be applied to diagnose real estate market trends, and spatial data available in the public domain can be used to improve the quality of modeling and analyze the performance of the real estate market in an era characterized by the rapid development of spatial data infrastructure.

Advanced computational techniques have considerable potential for improving the accuracy of predictions on the real estate market, but mathematical models alone are unable to completely eliminate the impact of behavioral factors on housing prices. The present findings confirm that complex models (random forest, extreme gradient boost) predict housing prices with much greater accuracy than classical statistical models (multiple regression). However, it should be noted that the choice of the optimal model requires a compromise between prediction accuracy and the ease of implementing a given tool.

Accurate prediction models undoubtedly deliver many advantages for financial institutions, mortgage lenders, and property developers. These models can promote well-informed decisions and minimize risk factors on the housing market. Further research directions should focus on the possibility of integrating machine learning algorithms and geostatistical methods to maximize the potential of data of a spatial nature, and this certainly includes real estate transaction data.

References

Alenany, E., Lekham, L. A., & Lu, S. (2021). Integrated clustering regression for real estate valuation. *Real Estate Finance*, Available at SSRN: https://ssrn.com/abstract=3835967

Baldominos, A., Blanco, I., Moreno, A. J., Iturrarte, R., Bernárdez, Ó., & Afonso, C. (2018). Identifying real estate opportunities using machine learning. *Applied Sciences (Basel, Switzerland), 8*(11), 2321. Advance online publication. https://doi.org/10.3390/app8112321 Preprint at https://doi.org/10.20944/preprints201810.0297.v1

Banerjee, D., & Dutta, S. (2017). Predicting the housing price direction using machine learning techniques. *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*, 2998–3000. https://doi.org/10.1109/ICPCSI.2017.8392275

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). Classification and regression trees. *Classification and Regression Trees*, 1–358. https://doi.org/10.1201/9781315139470

Cellmer, R. (2014). The possibilities and limitations of geostatistical methods in real estate market analyses. *Real Estate Management and Valuation, 22*(3), 54–62. https://doi.org/10.2478/remav-2014-0027

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13-17-August-2016*, 785–794. https://doi.org/10.1145/2939672.2939785

Chowhaan, M. J., Nitish, D., Akash, G., Sreevidya, N., & Shaik, S. (2023). Machine learning approach for house price prediction. *Asian Journal of Research in Computer Science, 16*(2), 54–61. https://doi.org/10.9734/ajrcos/2023/v16i2339

Choy, L. H. T., & Ho, W. K. O. (2023). The Use of Machine Learning in Real Estate Research. *Land (Basel), 12*(4), 740. Advance online publication. https://doi.org/10.3390/land12040740

Çılgın, C., & Gökçen, H. (2023). Machine learning methods for prediction real estate sales prices in Turkey. *Revista de la Construcción, 22*(1), 163–177. https://doi.org/10.7764/RDLC.22.1.163

Derdouri, A., & Murayama, Y. (2020). A comparative study of land price estimation and mapping using regression kriging and machine learning algorithms across Fukushima prefecture, Japan. *Journal of Geographical Sciences, 30*(5), 794–822. https://doi.org/10.1007/s11442-020-1756-1

Durganjali, P., & Pujitha, M. V. (2019). House resale price prediction using classification algorithms. *6th IEEE International Conference on Smart Structures and Systems, ICSSS 2019*. https://doi.org/10.1109/ICSSS.2019.8882842

Georgiadis, A. (2018). Real estate valuation using regression models and artificial neural networks: An applied study in Thessaloniki. *RELAND: International Journal of Real Estate & Land Planning, 1*(0), 292–303. https://doi.org/10.26262/RELAND.V1I0.6485

Gu, J., Zhu, M., & Jiang, L. (2011). Housing price forecasting based on genetic algorithm and support vector machine. *Expert Systems with Applications, 38*(4), 3383–3386. https://doi.org/10.1016/j.eswa.2010.08.123

Hamizah Zulkifley, N., Abdul Rahman, S., Ubaidullah, N. H., & Ibrahim, I. . (2020). House price prediction using a machine learning model: A survey of literature. *International Journal of*

Modern Education and Computer Science, 12(6), 46–54. https://doi.org/10.5815/ijmecs.2020.06.04

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems & their Applications, 13*(4), 18–28. https://doi.org/10.1109/5254.708428

Ho, W. K. O., Tang, B. S., & Wong, S. W. (2021). Predicting property prices with machine learning algorithms. *Journal of Property Research, 38*(1), 48–70. https://doi.org/10.1080/09599916.2020.1832558

Jha, S. B., Pandey, V., Jha, R., & Babiceanu, R. (2020). Machine learning approaches to real estate market prediction problem: A case study. ArXiv:2008.09922. https://doi.org/10.48550/arXiv.2008.09922

Kang, Y., Zhang, F., Peng, W., Gao, S., Rao, J., Duarte, F., & Ratti, C. (2021). Understanding house price appreciation using multi-source big geo-data and machine learning. *Land Use Policy, 111*, 104919. Advance online publication. https://doi.org/10.1016/j.landusepol.2020.104919

Kim, J., Lee, Y., Lee, M.-H., & Hong, S.-Y. (2022). A comparative study of machine learning and spatial interpolation methods for predicting house prices. *Sustainability 2022, Vol. 14, Page 9056, 14*(15), 9056. https://doi.org/10.3390/su14159056

Loh, W. Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery, 1*(1), 14–23. https://doi.org/10.1002/widm.8

Lorenz, F., Willwersch, J., Cajias, M., & Fuerst, F. (2023). Interpretable machine learning for real estate market analysis. *Real Estate Economics, 51*(5), 1178–1208. https://doi.org/10.1111/1540-6229.12397

Mora-Garcia, R. T., Cespedes-Lopez, M. F., & Perez-Sanchez, V. R. (2022). Housing price prediction using machine learning algorithms in COVID-19 times. *Land, 11*(11), 2100. https://doi.org/10.3390/land11112100

Morillo Balsera, M. C., Martínez-Cuevas, S., Molina Sánchez, I., García-Aranda, C., & Martinez Izquierdo, M. E. (2018). Artificial neural networks and geostatistical models for housing valuations in urban residential areas. *Geografisk Tidskrift, 118*(2), 184–193. https://doi.org/10.1080/00167223.2018.1498364

Noble, W. S. (2006). What is a support vector machine? *Nature Biotechnology 2006 24:12, 24*(12), 1565–1567. https://doi.org/10.1038/nbt1206-1565

Oladunni, T., & Sharma, S. (2016). *Hedonic housing theory – A machine learning investigation*. 522–527. https://doi.org/10.1109/ICMLA.2016.0092

Park, B., & Bae, K. J. (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. https://doi.org/10.1016/j.eswa.2014.11.040

Plakandaras, V., Gupta, R., Gogas, P., & Papadimitriou, T. (2015). Forecasting the U.S. real house price index. *Economic Modelling, 45*, 259–267. https://doi.org/10.1016/j.econmod.2014.10.050

Ren, X., Mi, Z., & Georgopoulos, P. G. (2023). Socioexposomics of COVID-19 across New Jersey: A comparison of geostatistical and machine learning approaches. *Journal of Exposure Science & Environmental Epidemiology, 34*, 197–207. https://doi.org/10.1038/s41370-023-00518-0 PMID:36725924

Rico-Juan, J. R., & Taltavull de La Paz, P. (2021). Machine learning with explainability or spatial hedonics tools? An analysis of the asking prices in the housing market in Alicante, Spain. *Expert Systems with Applications, 171*, 114590. https://doi.org/10.1016/j.eswa.2021.114590

Santhanam, R., Uzir, N., Raman, S., Banerjee, S., & Nishant Uzir Sunil, R. R. S. (2016). Experimenting XGBoost Algorithm for Prediction and Classification of Different Datasets Experimenting XGBoost Algorithm for Prediction and Classifi cation of Different Datasets. *International Journal of Control Theory and Applications, 9*. https://www.researchgate.net/publication/318132203

Sutton, C. D. (2005). Classification and regression trees, bagging, and boosting. *Handbook of Statistics, 24*, 303–329. https://doi.org/10.1016/S0169-7161(04)24011-1

Tchuente, D., & Nyawa, S. (2022). Real estate price estimation in French cities using geocoding and machine learning. *Annals of Operations Research, 308*(1–2), 571–608. https://doi.org/10.1007/s10479-021-03932-5

Thamarai, M., & Malarvizhi, S. P. (2020). House price prediction modeling using machine learning. *International Journal of Information Engineering and Electronic Business, 12*(2), 15–20. https://doi.org/10.5815/ijieeb.2020.02.03

Truong, Q., Nguyen, M., Dang, H., & Mei, B. (2020). Housing price prediction via improved machine learning techniques. *Procedia Computer Science, 174*, 433–442. https://doi.org/10.1016/j.procs.2020.06.111

Üzümcü, A. C., & Eliguzel, N. (2023). Predictive Analysis Using Web Scraping for the Real Estate Market in Gaziantep. *Bitlis Eren Üniversitesi Fen Bilimleri Dergisi, 12*(1), 17–24. https://doi.org/10.17798/bitlisfen.1155725

Wong, T. T., & Yeh, P. Y. (2020). Reliable accuracy estimates from K-fold cross validation. *IEEE Transactions on Knowledge and Data Engineering, 32*(8), 1586–1594. https://doi.org/10.1109/TKDE.2019.2912815

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing, 415*, 295–316. https://doi.org/10.1016/j.neucom.2020.07.061