Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 207 (2022) 435–445

Procedia Computer Science

www.elsevier.com/locate/procedia

26th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems

Machine learning in house price analysis: regression models versus neural networks

Iwona Foryś*

Institute of Economics and Finance, University of Szczecin, Mickiewicza 64, 71-101 Szczecin, Poland

Abstract

The problem addressed in this paper is automatic house price determination using multiple regression models and machine learning. In the practice of real estate appraisal, discussions about automated valuation (AVM) are increasingly common. Resistance to modern machine learning methods stems from a low level of knowledge about these methods and, as a result, an unawareness to what extent and where they can support the classical process of estimating property value. The problem is much broader, because it affects many aspects of the use of machine learning (including neural networks) in the broader real estate market. The process of real estate management at each stage generates huge information resources, which are used at different levels and to different extents by entities operating in the real estate market. One of such professional groups are real estate appraisers, for whom intelligent systems for monitoring the market and providing necessary data are becoming increasingly common and sought after. In this context, a comparative study of two models: multiple regression and neural networks has been carried out. Both methods were used to determine house prices, based on the same set of input data, and in the next step the effects of using these models for an additional group of objects with known characteristics and transaction prices were compared. The multivariate regression model obtained in the study was of medium quality, but sufficient for the purpose of comparative analysis. In the case of neural networks, the highest quality model was not obtained for the study sample, despite normalizing the variables to the required interval (0;1). In both models, the prices for the control sample were overestimated in most cases. However, this does not deny the relevance of further research and attempts to teach neural networks using larger data sets, especially in the case of properties other than typical residential units or land for residential development. Machine learning methods can be extremely useful especially in the processes of general valuation. In these processes, large sets of properties are estimated in a short period of time and with the same methods.

Peer-review under responsibility of the scientific committee of the 26th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2022)

Keywords: "machine learning, deep learning, neural networks, house value, regression models"

* Corresponding author. Tel.: +48914441964; fax: +48914442130.
E-mail address: iwona.forys@usz.edu.pl

1877-0509
Peer-review under responsibility of the scientific committee of the 26th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2022)
10.1016/j.procs.2022.09.078

1. Introduction

There is an ongoing discussion amongst appraisers and real estate market analysts about the usefulness of statistical models and machine learning in the practice of property valuation. It seems that the real estate market cannot ignore modern IT solutions, not only in the area of valuations performed by professional appraisers. There are in practice numerous applications of machine learning (e.g., Zillow portal), supporting players in the real estate trade, which use artificial intelligence algorithms to estimate property prices. These systems are evolving, constantly calibrating transactions and offers as well as updating input data resources.

The development of modern technologies and computational tool software in confrontation with increasingly easy access to large databases prompts a change in the way of thinking about the appraisal problem, in particular through traditional data selection methods [1]. The availability of a large amount of objective structured data in numerous databases and information systems makes it possible or even necessary to change the approach to the role of the real estate appraiser in the process of data collection and verification of its quality and usefulness in subsequent valuation stages [2], [3]. In the past, the appraisers were challenged by the shortage of good data, now the challenge is the opposite: how to properly sort, examine and formulate conclusions from large sets of available data? The new situation requires a shift in thinking about data collection and its analysis for appraisal purposes, which are fundamental to a properly conducted analytical process. Big data sets require a change in the approach to data analysis. Big Data Analysis is "the extraction of actionable knowledge directly from data through a process of discovery or hypothesis formulation and hypothesis testing" [4].

This paper analyzes two approaches to modeling the price of houses sold in the western part of Poznan. The databases for these transactions were created for the needs of the methodology of determining damages due to the creation of the RUA around the local airport [5],[6]. The first approach assumes classical multivariate regression, commonly used by real estate appraisers [7], [8]. In the second approach, neural networks will be tested as an alternative applied to estimate large sets of properties, as required by mass appraisal [9]. The purpose of the study is to compare the quality of the obtained models and to appraise, on their basis a sample of houses that were subject to sale in the following year. The analysis covers a large group of houses, yet in practice such sets include several hundred or even several thousand objects described by many variables. Hence the attempt to look for more advanced methods than simple multiple regression models.

2. Literature review

Each valuation process is preceded by market analysis, and the degree of sophistication of the methods employed depends on the valuation objective and the quality of data [10]. Correctly conducted analysis helps to determine the specific characteristics of the market [11] and, as a result, to identify its pricing features [12]. However, the development of modern technologies and software computational tools in collision with the increasingly easy access to large databases prompts a change in thinking about property valuation, particularly when it is performed using traditional data selection methods [13]. G. Dell notes that in the past, it was a challenge for valuers to find good data and represent it with purposely selected samples. Today, the challenge is the opposite: how to properly sort, explore, and draw conclusions from large sets of available data, entering the realm of *Big Data Analysis* in the valuation process [14]. The analyst's task is to select the ideal data set using scientific methods combined with the valuers' expertise and to merge the research problem and knowledge of the valuated object (valuation objective) with an appropriate analytical tool [15].

Due to the nature of the phenomena observed in the real estate market and due to the distribution of the dependent variable, one of the models most used in this segment is the linear multiple regression model [16]. These models gained popularity in the 1970s, especially in mass valuation of real estate for taxation purposes [17]. It should be noted here that there are also other models derived from multiple regression [18]. The methods of hedonic regression (RHI) involve the use of regression equations, where the dependent variable is the property price, and the explanatory variables are property features measured at different scales. Related to data quality is the issue of time instability of the estimates of the impact of property features on a property price. This problem was recognized long ago and for this reason, the literature recommends hedonic regression models that include a binary time variable [19]. In this regard, there is no clear qualification of methods in the literature. Most often three main approaches to

the construction of real estate price indices are proposed: the hedonic regression method, the repeat sale method and the stratification methods, such as the mix-adjustment method that has gained the greatest popularity with practitioners [20].

A linear hedonic regression model for determining the house price index was described by Fleming and Nells [21]. The quality features in the model take on zero-one values. By estimating the model with the least squares method (LSM), it is possible to identify the effect of the variables Xj on the price in a given period. In the next step, it is necessary to perform standardization by applying a system of weights to the features in the selected period and to determine the indices of change in subsequent periods. The hedonic regression method most often employs linear, semilog and log-linear models. The weakness of the above methods is due to the imperfection of the regression equation itself, such as the absence of coincidence or the presence of spurious regression (significant parameters and high coefficient of determination) or inappropriate choice of the analytical form of the model. Moreover, as noted by Gibbons and Machin, the traditional hedonic approach based on cross-sectional data does not allow solving the potential endogeneity problem [22]. Finally, the model does not capture the change over time of the impact of property features on its price by adopting a weighting system from the base period. The advantage of these methods is that the price indices are more accurate than when determining simple price indices. On the other hand, resale methods (RSM) are also burdened with sampling effects since they take into account only those properties that were resold and the reasons for the decision to sell are unknown. The last group of methods assumes that the best way to determine price changes is to compare these changes in homogeneous groups with respect to a selected criterion. The basic measure is the mean or median price for a given group determined in the first step, while in the next step the group measures are aggregated into one index for the whole market.

The literature emphasizes the occurrence of spatial autocorrelation in the real estate market [23], which offers broad potential to apply spatial regression models [24]. Depending on the type of spatial interactions, two basic spatial regression models are usually used: spatial lag model, and spatial error model [25],[26]. A method that accounts for spatial structure is giving weights to such observations that, due to their location in space, may theoretically have a greater effect on the phenomenon under study than others [27].These are the Geographically Weighted Regression (GWR) models .

In addition to the aforementioned classical models or spatial models, some combinations of models, e.g. non-parametric models or multilevel models, can also be used for particular stages of property market analysis. Due to their functional flexibility, nonparametric models have been proposed to extract the nonlinear structures underlying the hedonic pricing approach [28]. The next step is to use neural networks for property market analysis. Curry et al., in their study based on two-layer networks trained via a polynomial algorithm, found a slight predictive advantage of nonlinear models over linear hedonic models [29]. Linear models or those reduced to a linear form have an advantage due to the computational simplicity and interpretation of the results.

3. Materials and Methods

3.1. Materials

The study drew on house sales transactions concluded in Poznań in 2018 (N=229), while for the control group, the contracts concluded in 2019 (N=98) 1. As a result of interviews with property appraisers and the actual accessibility of data, the set of transactions was described with 10 variables. It was decided to focus on the basic characteristics of land properties developed with single-family houses, even though modern geoinformatics technologies can describe each property in even more detail with regard to the neighborhood or geospatial conditions.
The assumed explanatory variable was C- property price (PLN).
The explanatory variables are:
STU - (technical and functional condition of property) - to be demolished - 1, (minus) average - 2, average - 3, (plus) average - 4, good - 5 (description of the criteria below)
SZ - (development status) - poor - 1, average - 2, good - 3 (description of the criteria below)
DB - (additional buildings) - none 1, present - 2
FZ - (form of development) – terraced - 1, semi-detached, end-terraced- 2, detached - 3
PD - (plot area) - a continuous quantitative value of the registered area

PB - (building footprint area) - continuous quantitative area calculated as the product of PZ and the number of floors, but taking into account the irregularity of the body of the buildings
OC - (distance from the city center) - continuous quantitative value expressed in km
LS - (subjective location) - grade 5.

The variables PD, PB and OC are continuous variables on real scales, having the character of destimulants with respect to unit house prices. The qualitative variables STU, SZ, DB, FZ and LS are described on nominal scales and are stimulants (the higher the numerical value of the primary attribute, the higher the score on the ordinal scale).

The study took all the indicated variables and then built a multivariate regression model and tested it in different variants of MLP network for 2018 data. For the control group of 2019 transactions, house values were determined using the estimated regression model and the best-fit network, comparing the resulting prices with actual prices.

3.2. Methods

In the real estate appraisal practice, the multiple regression model has been accepted as a tool adequate for mass appraisal or real estate market analysis, especially since it is well recognized in case of typical properties (apartments, land). On the other hand, the circle of recipients of information on real estate prices is widening, as well as the area of application. Also, increasingly large datasets, as well as wide implementation of modern geoinformatics systems prompt us to look for solutions to problems in the area of machine learning.

Hence, the paper proposes a comparison of the effects of estimating the house price on the model market using the classical multiple regression model and applying one of the machine learning tools, i.e., neural networks.

The most popular, classical multivariate least squares method, its generalizations have been developed, which enables the construction of correct models also when Gauss-Markov assumptions are violated (e.g., homoskedasticity). The linear multivariate regression model has the form:

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \varepsilon_i \qquad (1)$$

where $y_i \in R$ is the response variable, $x_{1i}, \ldots, x_{ki} \in R$ is the set of predictors, $\beta_0,\ \beta_1 \ldots \beta_k \in R$ is the set of parameters, and $\varepsilon_i \sim N(0, \sigma^2)$.

One of the most popular methods for estimating the unknown parameters of equation (1) ordinary least squares (OLS) is based on minimizing the sum of squared residuals (SSR).

$$\hat{\beta} = min \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}))^2 \qquad (2)$$

The advantage of multivariate regression models is that they are easy to apply using available computational packages, as well as to interpret the estimated coefficients with variables, even for market practitioners, provided, of course, that the ceteris paribus principle is adopted [30]. Researcher, and especially market practitioners, follow the various processes and their effects at each computational stage in which they can interfere. For this reason, for real estate appraisers accustomed to "manual" comparisons of property pairs, it is easier to accept than modern machine learning methods where the computational process takes place beyond the user. The model also has many disadvantages, particularly for real estate market data. A good quality model on the one hand requires homogeneous objects (houses), on the other hand high variability of variables adopted to the model. What can be another barrier is the size of the data set of similar properties acquired for the analysis. A good quality model requires linearity of variables to be used [31]. When there are too many predictors in the model, its estimates often have a large variance, which reduces the accuracy of the forecasts. OLS lacks the skill of determining a smaller set of predictors that exhibits the best effects when one aims to find out which predictors most explain the variability of the response variable. Finally, by construction, the OLS method cannot be implemented when the number of parameters exceeds the number of observations. The predictors in the multivariate regression are usually estimated using the OLS method, that is, they are prone to over-fitting because there is no penalty for adding additional predictors. Another procedure for choosing which predictors do and do not enter the model is to compute all possible regression models, examining all possible combinations of predictors.

Neural networks are gaining popularity in practical solutions to complex classification problems (e.g., image analysis), pattern recognition, or prediction of (change) phenomena based on historical phenomena. The machine learning (as well as *data mining*) is classified into methods with a teacher (*supervised learning*) and without a teacher [32]. In supervised learning the goal of the model is to reproduce a distinguished feature (e.g., house price), hence two common problems: a classification problem when the distinguished feature is qualitative, and the regression problem when the distinguished feature is measured on a real scale. In unsupervised learning, there is no distinguished feature, thus there is also no known output (explanatory) variable, and the goal may be, for example, market segmentation.

The learning regression model in neural networks can generally be written:

$$Y = G(X, \varepsilon) \qquad (3)$$

where: Y – response variable, X – explanatory variables, ε – model error, G – - implicit functional dependency realized by neural network

The set of variables consists of a set of neurons [33], while the test subjects are divided into three subsets: learning (L), validation (V), and testing (T). Most often, the learning set is at least twice as large as the validation and testing set combined (e.g. $L=78\%$, $V=15\%$, $T=15\%$). Training of the network consists in searching for weights that would ensure decreasing the total network error (*SSE*), i.e. the difference of the value of the response variable at the input with the variable at the output, after checking all the learning samples, i.e. neurons (the sum of the values of the input signals multiplied by appropriate weighting factors). The signal representing the total excitation of a neuron is in turn transformed by a fixed neuron activation function (which is also sometimes referred to as the neuron transition function). The value calculated by the activation function is the output value of the neuron. Whereby, the total allowable estimation error should not exceed ±25/30% of the actual value of the response variable.

Many network structures are considered in the literature, but due to their practical value, unidirectional networks without feedback are most often used [34]. In such a network, neurons form a fixed structure: input layer, hidden layer and output layer. The number of instances (studied objects) should be ten times the number of connections present in the network. Increased the number of variables causes a non-linear increase in the number of instances. In practice, the number of instances also depends on the functional complexity being modeled. Popular functions like linear, logistic, tanh and exponential are used as activation functions for hidden and output neurons. A small number of instances justifies the use of a linear function.

When designing a neural network, popular network types such as MLP (Multi- Layer-Perceptron) or RBF (Radial Basis Function network) can be used, differing in e.g., the operation of the hidden layer. Networks can be trained with the use of numerous training algorithms [35] such as the back propagation error method, the Levenberg-Marquardt (LM) algorithm, or the coupled gradients (CG) method. In this study, we decided to use the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm programmed in the Statistica package, which has the advantage of fast convergence.

Most often, a single hidden layer is used in practice, while the researcher independently determines the number of neurons in the hidden layer. In other studies, authors suggest a number equal to half the sum of input and output neurons, or the root of the product of these quantities [36]. Networks having too many hidden layers or neurons in hidden layers may lead to the network overfitting.

In regression networks, the quality of the network is expressed by the Pearson correlation coefficient of the output response value and the network predictions. The most important feature indicating the quality of the model is the correlation coefficient for the validation set. An equally important parameter is the evaluation of the network error, especially the error functions (sum of squares or SOS) for each output neuron as the sum of squares of the differences between the set values and the values at the outputs of each, resulting in the SSE (Sum Square Errors) index written:

$$SSE = \sum_{j=1}^{M} \sum_{i=1}^{N} \left(d_{ji} - y_{ji}\right)^2 \qquad (4)$$

where: d_{ji} - the model response for learning incident j at the output of network i, y_{ji} - the actual value that appears at the output.

In practice, artificial neural network committees are also used. Their application is an alternative to training several or more networks and choosing one of them. The ensemble can consist of different types of networks with

different architectures, trained using different algorithms. The use of committees of networks (or committee machines) can lead to satisfactory results especially in the case of small or noisy learning datasets [37]. The output signal of a committee of networks is a combination of output signals of individual networks (experts), calculated as an arithmetic mean, The committees of artificial neural networks are also used when the number of input data is small.

Neural networks have the advantage of creating nonlinear models without having to formulate the multivariate model themselves. However, their disadvantage is the uncertainty of whether the error obtained during the network training is minimal. Additionally, the obtained model is not subject to further training and error correction. The networks create these models as a result of learning from user-specified examples, and the network learning algorithm itself consists in creating a data structure in its memory, based on the input information.

4. Research Results

In the first step, the correlation of the explanatory variables (i=9) with the response variable C was determined. The analysis was conducted for a set of N=229 transactions concluded in 2018. The marked correlation coefficients are statistically significant at p<.0500 (see Tab.1). Due to the character of the variables, the Spearman rank coefficient was used (non-parametric). At the assumed level, the correlation between a house price and its neighborhood (OT) and the type of development (FZ) was statistically insignificant. In case of the plot or building size, the correlation was statistically significant. There was also a strong relationship (value of correlation coefficient =.8089) between the price and technical and functional property condition (STU) and development status (SZ). A strong and statistically significant relationship was seen between some explanatory variables like a plot size (PD) and form of development (FZ). The last relationship shows a strong correlation (high value of correlation coefficient =.5059 between PD and FZ). The indicated relationships are understandable in view of the potential to locate houses on appropriately sized plots as well as in terms of the assessment of their technical condition related to the plot development. However, due to the comparison with the results of training neural networks in further calculations, all variables were left in place, which will consequently affect the quality of the multiple regression model.

Table 1. Pearson correlation coefficients (p <.05000).

Variable	*C*	*STU*	*SZ*	*DB*	*OT*	*FZ*	*PD*	*PB*	*OC*	*LS*
C	1									
STU	0.459048	1								
SZ	0.354036	0.808879	1							
DB	0.083409	0.066525	0.007659	1						
OT	0.111706	0.13846	0.152182	-0.03174	1					
FZ	0.184705	-0.04202	-0.03383	0.327559	-0.14408	1				
PD	0.435946	0.059223	0.035641	0.116205	-0.10051	0.505941	1			
PB	0.327667	0.160552	0.107448	0.123402	-0.09548	0.141545	0.180788	1		
OC	-0.29895	0.042622	0.104505	-0.04879	-0.01305	0.203776	0.221123	-0.36333	1	
LS	0.104434	0.049718	0.020237	0.041797	-0.22954	-0.00385	-0.06622	-0.11254	-0.20813	1

Based on the correlations obtained, the decision was made to remove the variables STU and DB.

In the next step, multivariate regression models were estimated using classical regression (see Tab.2). First, a classical multivariate regression model was estimated. using the OLS. All the estimated parameters were statistically significant at the assumed level (p <.05000), except for the parameters for the variables: building floor area (PB) and type of development (FZ). The models obtained the fits of R2 close to 0.57, which is satisfactory but not the best result in analyses of real economic data.

Table 2. Estimation of multivariate regression model parameters (Price PLN)

	Multivariate Regression Model I				Multivariate Regression Model II			
N=229	b	Standard Error	T Stat	p	b	Standard Error	T Stat	p
Constant	0.038345	0.037175	1.031468	0.3034	0.068732	0.028569	2.40582	0.016952
STU	0.229168	0.025075	9.139303	4.16E-17	0.236026	0.024201	9.75286	0.000000
OT	0.061935	0.024027	2.5777	0.0106	0.052479	0.022929	2.28876	0.023030
FZ	0.021773	0.021328	1.020833	0.3084	0.024728	0.021088	1.17262	0.242200
PD	0.808315	0.088349	9.149072	3.90E-17	0.811973	0.086741	9.36086	0.000000
PB	0.028129	0.042579	0.660646	**0.5095**				
OC	-0.30555	0.040067	-7.62595	7.11E-13	-0.326922	0.034157	-9.57130	0.000000
LS	0.028442	0.022077	1.288313	**0.19898419**				
R^2=0.57 p <.05000						R^2=0.57 p <.05000		

The signs at the estimated statistically significant parameters of the variables in the models were also consistent with expectations. The transaction price of properties decreased with distance from downtown (*OC*). In further analyses, model II and five explanatory variables *STU, OT, FZ, PD* and *OC* as well as the response variable Price *C* will be used.

In the next step, neural network training was performed for the same data (N=229). The data were randomly divided into three subsets: learning (L=70%), validation (V=15%) and testing (T=15%). In numerous attempts to test the network, this structure was also changed to L=60% and L=80% and accordingly changed equal subsets of validation and testing samples. However, the results obtained for the assumed network variants turned out to be the best for the first subset structure.

In the input layer of the network there were 6 neurons, which was due to the number of variables describing each object (house sale), while the output layer (transaction price) contained one neuron.

When designing a neural network, the networks like MLP and RBF can be used, but the results of learning in RBF network were of very low quality, hence the study focused on MLP type networks. In the hidden layer, networks with values of 2, 3 were trained, which stemmed from the number of variables (neurons in the input) and the rule of the average of input and output neurons or the root of their sum. Simulations were performed in Statistica package using available activation functions: linear, logistic, tanh and exponential. In the study, a regression one-way multilayer network (supervised learning method) was repeatedly trained. As a result, the best outcomes were obtained for MLP network (5-3-1) with 5 neurons in the input, 3 neurons in the hidden layer and one neuron in the output layer (response variable of house price). The table below summarizes selected parameters of the tested network (Tab. 3).

Table 3. Selected parameters of tested MLP (5-3-1) network

Network name	Quality (learning)	Quality (testing)	Quality (validity)	Quality (learning)	Error (testing)	Error (validity)	Learning algorithm	Error function	Activation (hidden)	Activation (output)
MLP 5-3-1	0.8695	0.6877	0.8172	0.0032	0.0016	0.0026	BFGS	SOS	Logistic	Linear
MLP 5-3-1	0.8607	0.7131	0.8386	0.0034	0.0012	0.0022	BFGS	SOS	Tanh	Logistic
MLP 5-3-1	0.8940	0.7374	0.8471	0.0027	0.0016	0.0019	BFGS	SOS	Logistic	Linear
MLP 5-3-1	**0.8720**	**0.7145**	**0.8682**	**0.0032**	**0.0015**	**0.0017**	**BFGS**	**SOS**	**Tanh**	**Logistic**
MLP 5-3-1	0.8642	0.7163	0.8600	0.0033	0.0012	0.0018	BFGS	SOS	Tanh	Linear
MLP 5-3-1	0.8662	0.7094	0.8553	0.0033	0.0014	0.0020	BFGS	SOS	Logistic	Logistic
MLP 5-3-1	0.8726	0.7294	0.8495	0.0032	0.0014	0.0020	BFGS	SOS	Logistic	Linear
MLP 5-3-1	0.8880	0.7161	0.8330	0.0028	0.0014	0.0020	BFGS	SOS	Logistic	Logistic
MLP 5-3-1	0.8937	0.7222	0.8393	0.0027	0.0014	0.0020	BFGS	SOS	Tanh	Logistic
MLP 5-3-1	0.8907	0.7278	0.8090	0.0027	0.0015	0.0024	BFGS	SOS	Exponential	Linear
MLP 5-3-1	0.9012	0.7172	0.8319	0.0025	0.0015	0.0021	BFGS	SOS	Exponential	Linear

The results obtained (especially for learning quality) are within the assumed accuracy range (25-30%) and seem satisfactory for the sample. In most cases the input function is a logistic or tanh function and the output one is a linear or logistic function. The best quality assessment values were obtained for learning (0.8720), testing (0.7145) and validation (0.8682). On the other hand, to assess the validity of the variables for the network, the value of global sensitivity coefficients was checked. Values greater than one indicate the importance of the explanatory variable and should be included in the model (Table 4). From the results in the table, all the proposed variables can be accepted because the parameter values are greater than one (except for the environment variable and form of development variable that were close to one). The distance from the center variable (*OC*) was the most significant in the model under study.

Table 4. Sensitivity analysis of MLP network (5-3-1)

Network name	*OC*	*PD*	*STU*	*FZ*	*OT*
MLP 5-3-1	2.602791	1.644690	1.520844	1.000447	0.995077
MLP 5-3-1	3.323387	1.866232	1.849285	1.176747	0.995521
MLP 5-3-1	2.718666	1.825995	1.646689	1.011411	0.993233
MLP 5-3-1	2.573238	1.812128	1.586713	1.035844	0.998712
MLP 5-3-1	4.830331	1.724698	1.562634	1.043636	1.023943
MLP 5-3-1	2.760233	1.652887	1.626895	0.977823	0.976945
MLP 5-3-1	4.061864	1.838565	1.612630	0.998164	1.002897
MLP 5-3-1	2.824136	1.941243	1.817333	0.989010	1.012871
MLP 5-3-1	2.878003	2.196496	1.877713	1.102479	1.018555
MLP 5-3-1	3.893548	1.936713	2.160602	1.112041	1.066213
MLP 5-3-1	4.433288	1.864087	2.031290	1.077835	1.012418

Table 3 shows 15 cycles of network learning, because subsequent cycles no longer improved the quality of learning, which can be seen in the learning error plot of the learning sample (Fig. 1). The shape of the error graph for the learning sample shows a clear decrease at the beginning to gently approach zero after several attempts. This means that after a dozen or so learning attempts, subsequent cycles do not significantly change the learning error of the network.

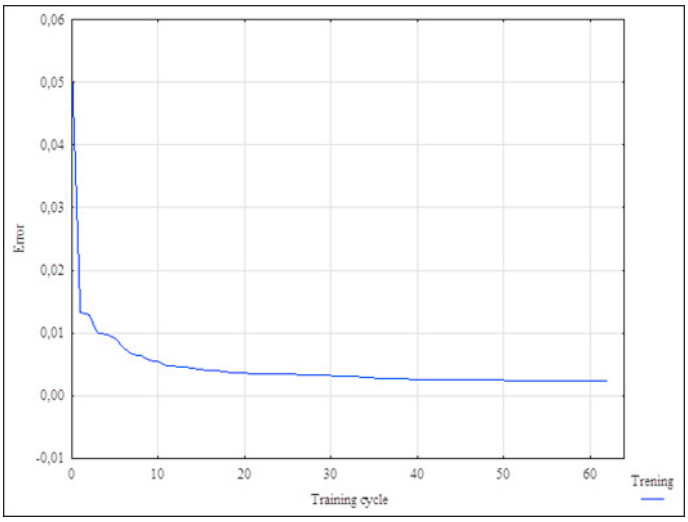
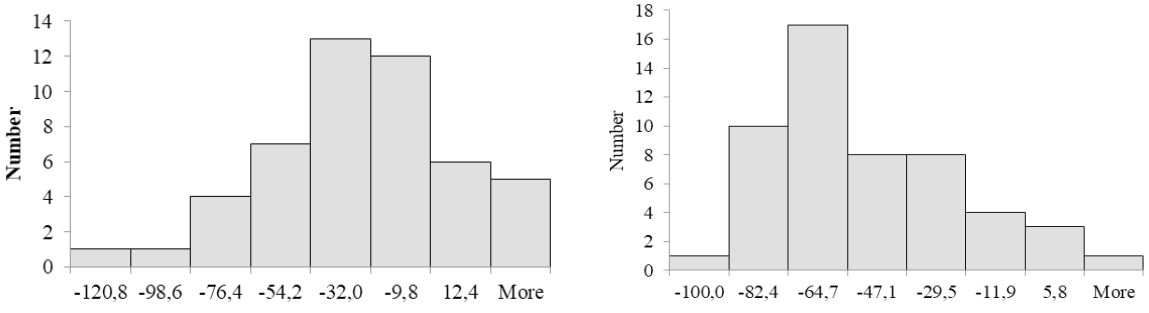


Fig. 1. Learning error plot of the learning sample.

While using the estimated regression model (Table 2) and the neural network (Table 3), it is possible to predict house prices in the next period. Data from the first quarter of 2019 (*STU, OT, FZ, PD, OC* variables) were used for this purpose, so as not to distort the results with price changes over time (N=52). The predicted prices (from both models) were compared with the actual transaction prices, determining the rate of underestimation by model and the network (for new data) of the actual house price according to the relationship:

$$[(actual\ price - estimated\ price\ )/(estimated\ price)] \cdot 100\% \qquad (5)$$

The distributions shown below were constructed for the obtained indicators and both sequences of values (Fig. 2).



Multivariate Regression Model II

Neural Networks MLP (5-3-1)

Fig. 2. Distributions of estimation errors in the multiple regression model and when using neural networks.

The obtained distributions of house price underestimation indices are slightly asymmetric, with the distribution of the index in the regression model being left-skewed and the index determined based on prices estimated by the neural network being right-skewed. This means that one model overestimates prices more and the other less. In the interesting range of overestimation not exceeding 25% there are prices estimated by the regression model. Such a high inaccuracy may be due to poor predictive quality of the model (average value of the coefficient of determination). As regards neural networks, the quality of the validation sample at 0.8 may be insufficient. Negative values of indices mean that real prices are lower than those estimated by the model. In the case of both the regression and the neural network, this means overestimation of house prices in relation to actual prices paid on the market.

5. Discussion and Conclusion

As pointed out by researchers, neural networks have a valuable advantage over other classical methods - because they deal well with data burdened with subjective evaluation of the researcher, they are resistant to bad hierarchization on weak scales or errors in measurement. Some authors show that the machine learning models significantly improve the accuracy compared not only to linear multiple regression, but also to spatial econometric models, and the performance of the stacking model is better than that of standalone machine learning models [38]. Others explained that machine learning is a powerful weapon to deal with the big data challenge [39]. An additional advantage is the fact that neural networks cope well not only with large amounts of data, but also with variables that are most often nonlinear. However, the condition of linearity is often an important assumption for many traditional econometric models.

The presented neural modeling and comparative analysis with classical regression model showed that the application of both neural networks and the regression model resulted in overestimation of house prices in the

control sample. In this study, the best results of network training were obtained when applying the BFGS algorithm and MLP network with the 5-3-1 structure, with the quality level of the learning sample at 0.87. It was not possible to obtain a network with better quality parameters of the learning sample, which undoubtedly affected the outcome of its comparison with the real price in the control sample.

Further work is planned to extend the research with the application of neural network committees, which can improve the quality of property price estimations in comparison to single networks and traditional regression models, also in specific submarkets. Additionally, the direction of future research is associated with the size and structure of the learning, validation and testing samples of individual datasets.

Although the user of neural networks needs empirical knowledge on how to select and prepare input data, how to choose the type of neural network and interpret the results, the level of theoretical knowledge to successfully build a model is lower than the expertise needed when using traditional modeling methods [40]. Hence, the use of neural networks in real estate practice is becoming increasingly common. Not without significance are also the expectations of market participants who expect to obtain quick answers to the questions posed, the supply of decision-making tools supported by modern computing techniques that cope well with large data sets, including the historical ones. Neural networks fit perfectly into these expectations. The integration of machine learning into property appraisal practice will move the whole industry from the era of manual work to the era of electronic work [41].

Acknowledgements

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.

In addition, project SOWA 2020, realized at the Cracow University of Economics.

References

[1] Dell, George (2017). “Regression, Critical Thinking, and the Valuation Problem Today.” *The Appraisal Journal*, Summer 2017: 217-230. www.appraisalinstitute.org

[2] Benjamin, D John, Randall, S, Guttery, Sirmans, R Stacy. (2004) “Mass appraisal: An introduction to Multiple Regression Analysis for real estate valuation.” *Journal of Real Estate Practice and Education* **7(1)**: 65–77.

[3] Wu, Hao, Jiao, Hongzan, Yu, Yang, Li, Zhigang, Peng, Zhenghong, Liu, Lingbo, and Zeng, Zheng. (2018) "Influence Factors and Regression Model of Urban Housing Prices Based on Internet Open Access Data." *Sustainability* **10(5)**: 1-17. https://doi.org/10.3390/su10051676.

[4] “*Big Data Interoperability Framework, Volume 1, Definitions*.” (2015) National Institute of Standards and Technology, NIST US Department of Commerce, Washington, DC.

[5] Batog, Jacek, Foryś, Iwona, Gaca, Radosław, Głuszak, Michał, and Konowalczuk, Jan. (2019) “Investigating the Impact of Airport Noise and Land Use Restrictions on House Prices: Evidence from Selected Regional Airports in Poland.” *Sustainability* **11(2)**: 1-18. https://doi.org/10.3390/su11020412.

[6] Drobiec, Łukasz, Foryś, Iwona, Habdas, Magda, and Konowalczuk, Jan. (2021) "*Value of real estate in the neighborhood of airports - methodology of estimating compensation and losses*.” Publishing C.H.Beck, Warszawa.

[7] Isakson, R. Hans. (1998) "The Review of Real Estate Appraisals Using Multiple Regression Analysis." *Journal of Real Estate Research* **15(2)**: 177-190.

[8] Rencher, C. Alvin. (2002) “*Methods of Multivariate Analysis*.” A John Wiley & Sons, Inc. Publication, New Jersey.

[9] Seckin, Yilmazer, Sultan, Kocaman.(2020) A mass appraisal assessment study using machine learning based on multiple regression and random forest”. *Land Use Policy* **99(8)**:104889.

[10] Fanning, F. Stephen. (2014) “*Market Analysis for Real Estate. Concepts and Applications in Valuation and Highest and Best Use*.” Appraisal Institute, Chicago, Il.

[11] Braun, A. David. (2012) “Market Delineation.” *The Appraisal Journal* **80(2)**:122-129.

[12] Emerson, M. Don. (2008) “Subdivision Market Analysis and Absorption Forecasting.” *The Appraisal Journal* **76(4)**: 377-390.

[13] Dell, George. (2017) “Regression, Critical Thinking, and the Valuation Problem Today.” *The Appraisal Journal* **85(3)**: 217-230.

[14]“*Big Data Interoperability Framework*” (2015) National Institute of Standards and Technology, NIST (Washington, DC: US Department of Commerce, September 16, 2015): 8.

[15] Wolverton, L. Marvin. (2009) "*Introduction to Statistics for Appraisers*." Appraisal Institute, Chicago.

[16] Isakson, R. Hans. (1998) "The review of real estate appraisals using multiple regression analysis." *Journal of Real Estate Research* 15(2): 177-190.

[17] Mark, Jonathan, Goldberg, Michael. (1988). "Multiple regression analysis and mass assessment: a review of the issues." *The Appraisal Journal* **56(1):**89–109.

[18] Radermacher, Walter. (2013) "Handbook on Residential Property Prices Indices (RPPIs)." Statistical Office of the European Union (Eurostat), Belgium.

[19] Shiller, J. Robert. (1991) "Arithmetic Repeat Sales Price Estimators." *Journal of Housing Economics* **1(1):**110–126.

[20] Foryś, Iwona. (2012) "Mix-adjustment method of determining residential real estate price indices on the example of cooperative premises." *Studies and Materials of the Scientific Society for Real Estate* **20(1)**: 41–52.

[21] Fleming, C. Michael, Nellis, G. Joseph. (1994) "The Measurement of UK House Prices: a review and Aprisal of the Proncipal Sources." *Journal of Housing Finance* **24**:6-16.

[22] Gibbons, Stephen, Machin, Stephen. (2005) "Valuing Rail Access Using Transport Innovations." *Journal of Urban Economics* **57(1)**: 148–69.

[23] Wilhelmsson, Mats. (2022) "Spatial Model in Real Estate Economics." *Housing Theory and Society* **9(2)**:92–101.

[24] Anselin, Luc. (1988) "*Spatial Econometrics: Methods and Models*." Kluwer Academic Publishers.

[25] Páez, Antonio, Scott, M. Darren. (2004) "Spatial statistics for urban analysis: A review of techniques with examples." *GeoJournal* **61**:53-67.

[26] Bourassa, Steven, Cantoni, Eva, Hoesli, Martin. (2010) "Predicting house prices with spatial dependence: a comparison of alternative methods." *Journal of Real Estate Research* **32(2)**: 139-160.

[27] Charlton, Martin, Fotheringham, Stewart, Brunsdon, Chris. (2009) "Geographically weighted regression." *White paper. National Centre for Geocomputation. National University of Ireland Maynooth* 2.

[28] Peterson, Steven, Flanagan, Albert. (2009) "Neural network hedonic pricing models in mass real estate appraisal." *Journal of real estate research* **31(2)**: 147-164.

[29] Curry, Bruce, Morgan, Peter, Silver, Mick. (2022) "Neural networks and non-linear statistical methods: an application to the modelling of price–quality relationships." *Computers & Operations Research* **29(8)**: 951-969.

[30] Isakson, Hans, R. (2001) "Using Multiple Regression Analysis in real estate appraisal." *Appraisal Journal* **69(4)**: 424-430.

[31] Mark, Jonathan, Goldberg, A. Michael. (2001) „Multiple regression analysis and mass assessment: A review of the issues." *Appraisal Journal* **56 (1)**:89-109.

[32] Haykin, Simon. (1994). "*Neural Networks: A Comprehensive Foundation*". MacMillan College Publishing Co. New York.

[33] Lula, Paweł. (1999). "*Jednokierunkowe sieci neuronowe w modelowaniu zjawisk ekonomicznych*." Zeszyty Naukowe. Akademia Ekonomiczna w Krakowie. Seria Specjalna, Monografie 140.

[34] Tadeusiewicz, Ryszard, (1993) „Sieci neuronowe". Akademicka Oficyna Wydawnicza RM, Warszawa.

[35] Bishop, Christopher M. (1995) "Neural Networks for Pattern Recognition". Oxford University Press, Oxford.

[36] Masters, Timothy. (1993) Practical Neutral Network Recipes in C++. Academic Press. San Diego, CA United States.

[37] Boussabaine, Halim A. (1996) "The use of artificial neural networks in construction management: a review." *Construction Management and Economics* 1**4(5)**: 427-436.

[38] Lulin, Xu, Li, Zhongwu. (2021) "A new appraisal model of second-hand housing prices in China's firs-ttier cities based on machine learning algorithms." *Computational Economics* **57(2)**: 617-637.

[39] Taigel, Fabian, Tueno, Anselme, K, Pibernik, Richard. (2018) "Privacy-preserving condition-based forecasting using machine learning." *Journal of Business Economics* **88(5)**: 563–592.

[40] Wu, Hao, Hongzan Jiao, Yang Yu, Zhigang Li, Zhenghong Peng, Lingbo Liu, Zheng Zeng. (2018). "Influence Factors and Regression Model of Urban Housing Prices Based on Internet Open Access Data." *Sustainability* **10(5)**: 1676.

[41] Worzala, Elaine, Souza, Lawrence A, Koroleva, Olga, China, Martin, Becker, Alicia, Derrick, Nathaniel. (2021). The technological impact on real estate investing: Robots vs humans: New applications for organizational and portfolio strategies." *Journal of Property Investment & Finance* **39(2)**: 170-177.