Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 192 (2021) 3542–3550

Procedia Computer Science

www.elsevier.com/locate/procedia

25th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems

Impact of Categorical Variables Encoding on Property Mass Valuation

Sebastian Gnat*

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

Abstract

The main aim of the article was to present impact of categorical variables encoding on property mass valuation. Categorical variables are often used for describing important properties' characteristics. In some countries, i.e., Poland, description of properties is mainly conducted with categorical variables, both nominal and ordinal. When property mass valuation is carried out it is important to introduce this kind of variables in best way to achieve most accurate results. There are many techniques of categorical variables encoding. In this study some of them were used in data pre-processing to determine whether the choice of encoding technique affects valuation results obtained with several regression algorithms. Three types of regression models were used in the research: a ridge regression model, *k* nearest neighbours regression and random forest regression algorithm. Each algorithm used explanatory variables coded using five techniques: one hot encoding, catboost encoding, Helmert encoding, target encoding and ordinal encoding. The results show that mass valuation results vary depending on how the encoding of categorical variables occurs. The regression algorithms used in the study respond differentially to the variable encoding techniques. Nevertheless, the one-hot encoding technique proved to be the best choice. The practical implications of the study are related to the reform of property taxation in Poland. Under this reform, values would become the basis for property taxation. This will be a complex undertaking, requiring the testing of various types of computational techniques to accurately determine the value of an enormous number of properties. The machine learning techniques presented in the study could be a part of a decision support system for introducing a new way of property taxation.

Keywords: categorical varibles encoding; property mass valuation; valuation accuracy

* Corresponding author. Tel.: +48 444 19 94
E-mail address: sebastian.gnat@usz.edu.pl

10.1016/j.procs.2021.09.127

1. Introduction

Various machine learning algorithms are used in almost every sphere of human activity. One such sphere is undoubtedly the real estate market. The theoretical and practical tasks encountered concern both classification and regression. Mass real estate valuation belongs to the regression tasks. Using available information about properties and their surroundings, an attempt is made to predict prices or values as accurately as possible. There are many situations in which mass real estate valuation can find application:

- monitoring the value of real estate portfolios that serve as collateral for credit exposures held by the bank [1], [2],
- property valuation for the purpose of estimating the economic effects of adopting or amending local zoning plans,
- real estate taxation [3],
- other, in which it is necessary to appraise the value of multiple real properties at the same time.

No matter what algorithm is used for mass real estate valuation one has to face the problem of proper data preparation. This is a particularly important step in the process of applying machine learning, whether it is real estate valuation or any other research area. Data used in real estate value modeling can be both quantitative and qualitative in nature. Qualitative variables often represent important factors that strongly determine property value. They must therefore be included in statistical models or machine learning algorithms.

This paper will present the results of a study on the impact of the applied qualitative variables encoding techniques on the accuracy of property valuations. Five such methods were selected for the study and used at the data pre-processing stage. The prepared data sets were used in several machine learning algorithms to determine the repeatability of the obtained results. The purpose of the study is to examine whether the method of categorical variables encoding significantly affects the accuracy of property valuations.

2. Literature review

Accurately determining the value of many properties, i.e., mass valuation, is a major research challenge. A number of studies have emerged related to the application of various types of machine learning algorithms to determine property values. Constructing a valuation model can be aimed at determining the impact of individual property attributes on their value or at creating a tool that accurately predicts the price or value of real estate. 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 [4]. Wang and Li [5] conducted a review of over 100 articles related to mass valuation methods from nearly twenty years. They point 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. Examples of the application of the following machine learning algorithms in property valuation can be found in scientific papers: [6] – multiple regression, [7] – ridge regression, [8] – regression trees, [9] – random forests, [10] – support vector machines, [11] – artificial neural networks, [12] – *XGBoost*. There is an ongoing discussion regarding comparison of multiple regression models and machine learning algorithms. Their superiority over multiple regression models was demonstrated on the case of New York [13]. On the other hand, there are studies proving no significant differences between e.g. neural networks and multiple regression, or even studies in which neural networks occurred to be an inferior solution [14].

Regardless of which algorithm is used to estimate property values, a very important aspect of computational experimentation is the issue of the quality of the data at one's disposal. Grover [15] points out that the issue of the availability and quality of data used in valuation should not be neglected. The stage of specifying variables, which have a significant impact on a dependent variable is very important. Metzner and Kindt [16] tried to itemize the variables determining real property values used by researchers in various studies. The authors, having reviewed the literature, specified over 400 real estate attributes used in mass appraisal models. They postulate the need for determining a certain fixed set of attributes, which would allow creating more stable and comparable valuation models. The importance of data preparation in real estate is not often emphasized. Krause and Lipscomb [17] stress that little discussion in the real estate literature is given to acquiring, managing, cleaning, and preparing datasets. They provide evidence of general state of real estate data. They also define some characteristics that make properties' data hard to work with. Those datasets suffer from inconsistent levels of observation, non-matching unique identifiers, temporal inconsistencies, field standardization, conflicting observations form different sources.

Machine learning algorithms require that the data used be numerical in nature. Qualitative variables, on the other hand, are a quite common way to describe the reality around us. Some data pre-processing steps ought to be made in order to prepare data for machine learning algorithms. One of the elements of data pre-processing is categorical variables encoding. There are many techniques designed for this purpose [18], [19]. The authors emphasize that the variety of computational procedures available makes it necessary to test them and evaluate. Broad survey on categorical data encoding for neural network presented by Hancock and Khoshgoftaar [20] investigates current techniques for representing qualitative data for use as input to neural networks. They state that new techniques of encoding are still emerging and not all of them can be used for any given data set. There are supervised and unsupervised categorical encoding techniques. When encoding relies only on categorical columns there is an unsupervised encoding. On the other hand, when encoding is based on other, numerical, column it is a supervised one. Since there is such a wide range of techniques to choose from, it is only when the best technique is identified that the next steps in the investigation can be carried out. Literature on categorical variable encoding for real estate valuation is not broad. Parygin et al. [21] present research on categorical data processing for real estate. They focus on one of the most popular encoding technique – one hot encoding. Small number of publications regarding categorical variables encoding makes a research gap. This study presents results of properties mass valuation with different encoding methods selected on pre-processing stage of machine learning experiments. The main goal is to determine whether the choice of encoding technique impacts the valuation accuracy. Presented study is a part of broader research regarding property mass valuation. Scientists develop and test various numerical procedures to build as good as possible valuation algorithm [22], [23], [24].

3. Material and Methods

Three types of regression models were used in the research: a ridge regression model, *k* nearest neighbours regression and random forest regression algorithm. The first one is a parametric model, whereas the remaining two models are non-parametric algorithms.

Nomenclature

w_j	unit market value of *j–th* real estate
α_0	constant term
α_i	structural coefficients
β	regularization coefficient
u_i	random component
$\widehat{w}_j$	theoretical value of *j–th* real estate
$\overline{w}_j$	mean of actual values of real estates
$RMSE$	root mean square error
$rRMSE$	relational root mean square error

The ridge regression model is as follows:

$$ln(w_j) = \alpha_0 + \sum_{i=1}^{k} \alpha_i x_i + \beta \sum_{i=1}^{k} \alpha_i^2 + u_i \quad (1)$$

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 need to be encoded in order to be used in machine learning algorithms.

In multiple regression models, model weights are determined by minimizing the sum of squares of the residuals of the model ($RSS \rightarrow \min$). When it comes to ridge regression, a regularization term equal to $\beta \sum_{i=1}^{k} \alpha_i^2$ is added to *RSS* cost function [25] of equation (1). The hyperparameter β controls how much one wants to regularize the model. If $\beta = 0$, then ridge regression is just pure multiple regression. If β is very large, then all weights end up very close to zero and the result is a flat line going through the dependent variable mean value [26]. Therefore, setting β is the crucial stage of creating a model to achieve high quality results.

The k nearest neighbours algorithm is a non-parametric algorithm. Though mainly applied in classification problems, the *KNN* algorithm can also be used in regression problems [27]. The operation of the algorithm comes down to two steps. In the first step for a given point x_o, one finds k training points $x(r)$, $r = 1, \ldots, k$ located closest to x_o. 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.

Introduced for the first time by Breiman [28] random forest is ensemble machine learning technique. It utilises set number of regression or classification trees to achieve prediction based on majority voting in case of classification problems or averaging in case of regression of predictions made by each tree in the ensemble. All trees that make up the random forest are constructed independently of each other. For any given tree a subset of predictors is chosen, and bootstrap sample of training data is used to build a tree. After generating pre-set number of trees predictions are made.

Of the many techniques available for encoding qualitative variables, the study used 5: one hot encoding [18], catboost encoding [29], Helmert encoding [30], target encoding and ordinal encoding [31].

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*):

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

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

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 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.

The study was conducted using dataset of 318 land plots located in one of the largest cities of Poland – Szczecin. The location of the city of Szczecin in Poland and boundaries of valuated region are presented in Figure 1. Whereas Figure 2 presents valuated land plots within that region.

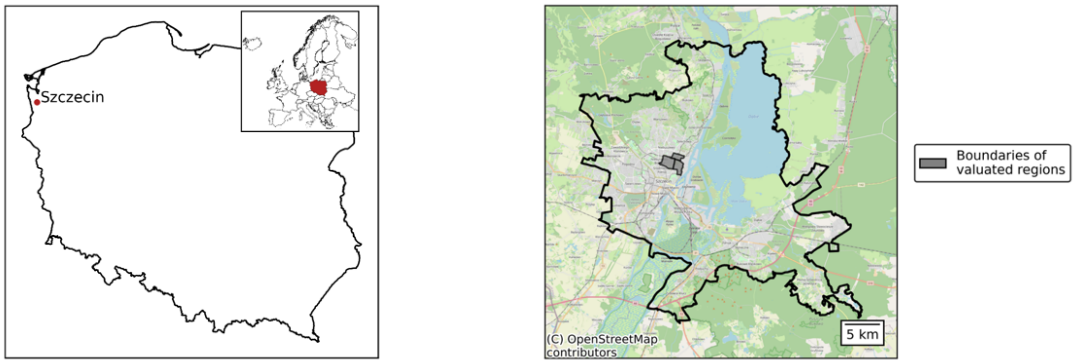

Fig. 1. Location of the city of Szczecin within Poland and location of valuated area.

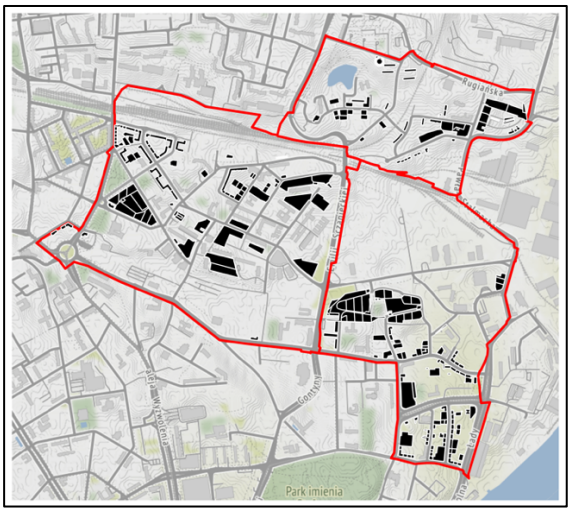

Fig. 2. Location of the valued real estates.

Attributes and their states are presented in Table 1. All the attributes were treated as qualitative variables. It is typical situation in Poland, because market participants treat real estate features in such a way. All features are usually described on an ordinal or nominal scale. This was also endorsed by appraisers. 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. Some research has provided evidence that segmenting property market often improves mass valuation [32]. A procedure of determining submarkets has been introduced in the study as well. In original dataset there are variables that could be treated as a proxy for a location. Location attractiveness zones were in these cases constructed by experts. They are constructed in such a way that the impact of a location in each area is homogenous. 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. Recalculation of perpetual usufruct annual fees is conducted, according to Polish regulations, for the land only. Therefore, only land was the object of evaluation, without taking development into account.

Table 1. Real estate attributes and their levels.

Attribute	Attribute Category
Utilities	None Incomplete Complete
Neighborhood	Onerous Unfavourable Average Favourable
Transport availability	Unfavourable Average Favourable
Physical plot properties	Unfavourable Average Favourable
Plot area	Large (>1200 m2) Average (500–1200 m2) Small (<500 m2)
Location attractiveness zone	Zone 1 Zone 2 Zone 3

4. Empirical results

The survey was conducted as follows. Three types of models were employed in the study: a multiple regression model, *k* nearest neighbour regression model and random forest regression model. Training sets were sampled 500 times in order to obtain averaged results. Of the 318 properties, 250 were sampled for the training set, while the remaining 68 were the test sets in each of the 500 reiterations. The accuracy of appraisal was analyzed on the basis of relational root mean square error (*rRMSE*). After each draw, the explanatory variables were coded using five different techniques. In this way, 7500 thousand models were built (three types of regression algorithms, five techniques for categorical encoding, 500 repetitions). To build regression models, in all cases a 10-fold grid search with cross validation was used in order to tune hyperparameters. Calculations were performed using two Python programming language libraries – Scikit-Learn [33], Category encoders [34]. Figures from 3 to 5 present kernel density estimations of *rRMSE* distributions obtained with chosen encoding techniques. The main objective of the study is to evaluate how different encoding techniques affect valuation errors (models' residuals). The results obtained are inconclusive. The *rRMSE* distributions obtained for the different models and coding techniques differ from each other. In the case of ridge regression and the *KNN* algorithm, the individual distributions differ from each other. However, it can be observed that two techniques, one-hot encoding and target encoding allow the lowest valuation errors. On the other hand, the catboost encoding technique proved to be the worst for both mentioned regression algorithms. A different situation occurred in the third algorithm used - random forest. Helmert encoding, target encoding and one hot encoding gave similar *rRMSE* distributions. Again, catboost encoding proved to be the least favorable choice. Table 2 contains selected descriptive statistics of the distributions discussed.

Table 2. Descriptive statistics of relational root mean square errors.

encoding technique	regression algorithm					
	ridge regression		*KNN* regression		random forest regression	
	mean	SD	mean	SD	mean	SD
one hot encoding	0.0449	0.00596	0.0407	0.00484	0.0389	0.00504
catboost encoding	0.0523	0.00437	0.0471	0.00476	0.0474	0.00458
Helmert encoding	0.0455	0.00374	0.0423	0.00536	0.0391	0.00477
target encoding	**0.0448**	0.00335	**0.0396**	0.00500	**0.0387**	0.00444
ordinal encoding	0.0596	0.00418	0.0433	0.00547	0.0397	0.00454

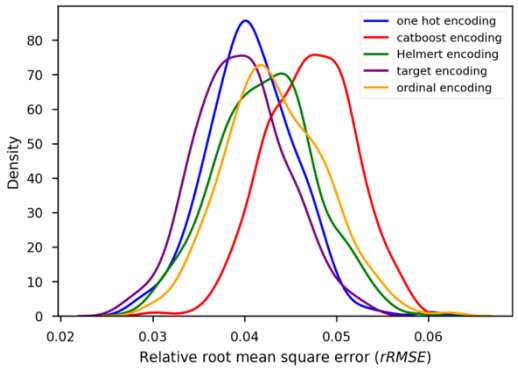

Fig. 3. Kernel density estimation of relational root mean square errors (ridge regression models).

It turned out that the smallest valuation errors were obtained in each regression model for the target encoding technique. This technique belongs to supervised methods and is associated with data leakage risk. It is worth noting, therefore, that also in each case the second position on the list of the smallest valuation errors fell to the immensely popular one hot encoding technique. While applying it one should remember about omitting in the matrix of explanatory variables one of the levels of each attribute describing real estate.

Analyzing, in turn, the regression algorithm itself, the smallest errors were obtained for random forests. They were on average 20% smaller than for ridge regression and about 5% smaller than for *KNN* regression algorithm.

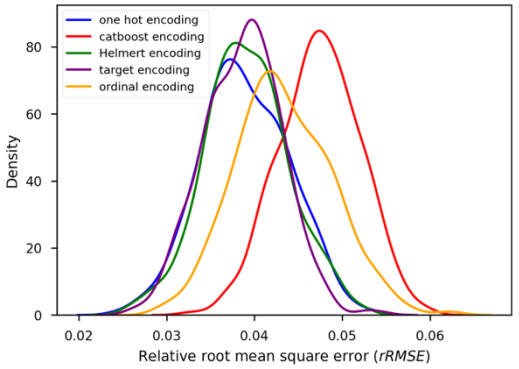

Fig. 4. Kernel density estimation of relational root mean square errors (*k* nearest neighbours algorithm).

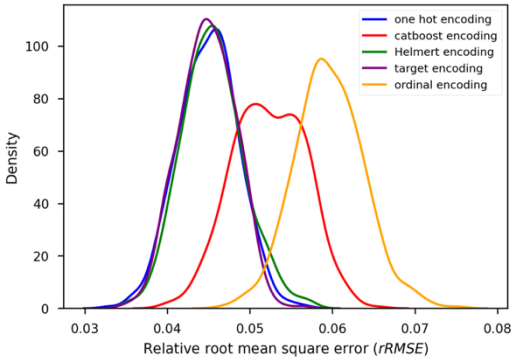

Fig. 5. Kernel density estimation of relational root mean square errors (random forest algorithm).

5. Conclusions

The study evaluated the effect of categorical variables encoding on the accuracy of mass real estate valuation. Five methods of converting qualitative variables into quantitative variables were used for three different regression algorithms at the data pre-processing stage. The selected methods belonged to both supervised and unsupervised groups of methods. The modeling was repeated for 500 splits of the analyzed data set into training and test sets. For the results obtained with the training data, the *rRMSE* for the data in the test sets was determined. The results that were obtained demonstrate that the categorical encoding technique has an impact on valuation errors. Furthermore, whether a particular technique produces better (lower errors) valuations also depended on the regression algorithm used. In the case of ridge regression, the error distributions obtained from data encoded by the different techniques differed from each other. The best techniques in this case turned out to be target encoding and one hot encoding. A similar situation was observed for the *KNN* algorithm. In both cases, the largest valuation errors were obtained for data encoded with the catboost encoding technique. The situation was different for the random forest algorithm. It proved to be more robust to the encoding technique for qualitative data. Three of the techniques analyzed gave similar *rRMSE* distributions. The catboost and ordinal encoding techniques gave higher errors. This means that while the

random forests proved to be less sensitive to the encoding method, it is still possible to make a poor choice and degrade the modeling results if one of the techniques to which the algorithm is more robust is not used.

The study presented here is one component of a larger study related to evaluating the applicability of machine learning methods in small real estate markets. The need for mass property valuation may also occur in markets where the number of transactions is small and the properties that are traded on the market are similar to each other. Such situations make it difficult to apply different types of algorithms. Thus, there is a need to study what computational solutions should be used to meet the need for valuing many properties in the conditions existing on small markets.

The main conclusion of the study is that the most common and used technique, i.e. one hot encoding in each regression algorithm was one of the best choices and for the data analyzed in the study it can be recommended for use. Further research should look at extending both the set of encoding techniques and regression algorithms. Research should also go into the analysis of other real estate datasets. The practical implications of the study are related to the reform of property taxation in Poland. Under this reform, property values would become the basis for taxation. Enacting such a reform will require many legal, organizational, technical decisions, and decisions related to the mass appraisal process itself. There is no doubt that creating a decision support system for this endeavor will have a positive impact on the reform. Such a system should include, among others, various types of data pre-processing and machine learning solutions to support property mass valuation.

Acknowledgements

Project financed under the program of the Minister of Science and Higher Education under the name "Regional Initiative of Excellence" in the years 2019–2022, project no. 001/RID/2018/19, financing amount 10,684,000.00 PLN. The research was co-financed by the National Science Centre, Project No 2017/25/B/HS4/01813.

References

[1] Korteweg, A. G., and M. Sorensen. (2016) "Estimating loan-to-value distributions." *Real Estate Economics* **44 (1)**: 41–86. doi:10.1111/1540-6229.12086

[2] Tzioumis, K. (2017) "Appraisesrs and valuation bias: an empirical analysis." *Real Estate Economics* **45 (3)**: 679–712. doi:10.1111/1540-6229.12133

[3] Breiman, L. (2001) "Random Forests." *Machine Learning* **45 (1)**: 5–32. doi:10.1023/A:1010933404324

[4] Pérez-Rave, J. I., J. C. Correa-Morales, and F. A. González-Echavarría. (2019) "Machine learning approach to big data regression analysis of real estate prices for inferential and predictive purposes." *Journal of Property Research* **36 (1)**: 59-96. doi:10.1080/09599916.2019.158748

[5] Wang, D., and V. J. Li. (2019) "Mass appraisal models of real estate in the 21st century: A systematic literature review." *Sustainability* **11 (24)**: 7006. doi:10.3390/su11247006

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

[7] Gnat, S. (2020) "Impact of the regularization of regression models on the results of the mass valuation of real estate." *Folia Oeconomica Stetinensia* **20 (1)**: 163-176. doi:10.2478/foli-2020-0009

[8] McCluskey, W. J., D. Daud, and N. R. Kamarudin. (2014) "Boosted regression trees: An application for the mass appraisal of residential property in Malaysia." *Journal of Financial Management of Property and Construction* **19 (2)**: 152-167. doi:10.1080/09599916.2013.781204

[9] Antipov, E. A., and E. B. Pokryshevskaya. (2012) "Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics." *Expert Systems with Applications* **39 (2)**: 1772–1778. doi:10.1080/10835547.2004.12091602

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

[11] Zhou, G., Y. Ji, X. Chen, and F. Zhang. (2018) "Artificial neural networks and the mass appraisal of real estate." *International Journal of Online Engineering* **14 (3)**: 180–187. doi:10.3991/ijoe.v14i03.8420

[12] Kim, Y., S. Choi, and M. Y. Yi. (2020) "Applying comparable sales method to the automated estimation of real estate prices." *Sustainability* **12 (14)**: 5679. doi:10.3390/su12145679

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

[14] Del Giudice, V., P. De Paola, F. Forte, and B. Manganelli. (2017) "Real estate appraisals with bayesian approach and Markov chain hybrid Monte Carlo method: an application to a central urban area of Naples." *Sustainability* **9 (11)**: 1-17. doi:10.3390/su9112138

[15] Grover, R. (2016) "Mass valuations." *Journal of Property Investment & Finance* **34 (2)***:* 191-204. doi:10.1108/JPIF-01-2016-0001

[16] Metzner, S., and A. Kindt. (2018) "Determination of the parameters of automated valuation models for the hedonic property valuation of residential properties: A literature-based approach." *International Journal of Housing Markets and Analysis* **11 (1)**: 73-100. doi:10.1108/IJHMA-02-2017-0018

[17] Krause, A., and C. A. Lipscomb. (2016) "The Data Preparation Process in Real Estate: Guidance and Review." *Journal of Real Estate Practice and Education* **19 (1)**:15-42. doi:10.1080/10835547.2016.12091756

[18] Potdar, K., T. S. Pardawala, and C. D. Pai. (2017) "A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers." *International Journal of Computer Applications* **175 (4)**: 7-9. doi:10.5120/ijca2017915495
[19] Hien, D. T. T., C. T. T. Thuy, K. Anh, D. T. Son, and C. N. Giap. (2020). "Optimize the Combination of Categorical Variable Encoding and Deep Learning Technique for the Problem of Prediction of Vietnamese Student Academic Performance." *International Journal of Advanced Computer Science and Applications* **11 (11)**: 274-280. doi:10.14569/IJACSA.2020.0111135
[20] Hancock, J. T., and T. M. Khoshgoftaar. (2020) "Survey on categorical data for neural networks." *Journal of Big Data* **7 (28)**: 1-41. doi:10.1186/s40537-020-00305-w
[21] Parygin D. S., Malikov V. P., Golubev A. V., Sadovnikova N. P., Petrova T. M., and Finogeev A. G. (2018) "Categorical data processing for real estate objects valuation using statistical analysis." *Journal of Physics: Conference Series* **1015 (3)**: 032102. doi:10.1088/1742-6596/1015/3/032102
[22] Doszyń, M. (2020) "Algorithm of real estate mass appraisal with inequality restricted least squares (IRLS) estimation." *Journal of European Real Estate Research* **13 (2)**: 161-179. doi:10.1108/JERER-11-2019-0040
[23] Gnat, S, and M. Doszyń. (2020) "Parametric and Non-parametric Methods in Mass Appraisal on Poorly Developed Real Estate Markets.", *European Research Studies Journal*, **23 (4)**: 1230-1245. doi:10.35808/ersj/1740
[24] Dmytrów, K., and S. Gnat. (2019) "Application of AHP Method in Assessment of the Influence of Attributes on Value in the Process of Real Estate Valuation." *Real Estate Management and Valuation* **27 (4)**: 15-26. doi:10.2478/remav-2019-0032
[25] Lesmeister, C. (2019) *Mastering Machine Learning with R*, Pact Publising.
[26] Geron, A. (2017). *Hands-On Machine Learning with Scikit-Learn and TensorFlow*. O'Reilly.
[27] Pace, R. K. (1996) "Relative performance of the grid, nearest neighbor, and OLS estimators." *Journal of Real Estate Finance and Economics* **13**: 203-218. doi:10.1007/BF00217391
[28] Bradbury, K. L., C. J. Mayer, and K. E. Case. (2001) "Property tax limits, local fiscal behavior, and property values: Evidence from Massachusetts under Proposition 2,5." *Journal of Public Economics* **80 (2)**: 287–311. doi: 10.1016/S0047-2727(00)00081-5
[29] Prokhorenkova, L., G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin. (2018) "CatBoost: unbiased boosting with categorical features." *Advances in Neural Information Processing Systems* **31**.
[30] Hutcheson, G. (2011), "Categorical explanatory variables." *Journal of Modelling in Management* **6 (2)**. doi:10.1108/jm2.2011.29706baa.002
[31] He, D., and L. Parida. (2016) "Does encoding matter? A novel view on quantitative genetic trait prediction problem." *BMC Bioinformatics* **17 (Suppl 9)**: 272. doi:10.1109/BIBM.2015.7359667
[32] Usman, H., M. Lizam, and M. U. Adekunle. (2020) "Property price modelling, market segmentation and submarket classifications: a review." *Real Estate Management and Valuation* **28 (3)**: 24-35. doi:10.1515/remav-2020-0021
[33] Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. (2011) "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research* **12**: 2825–2830. arXiv:1201.0490v4
[34] McGinnis, W. D., Ch. Siu, A. S, and H. Huang. (2018) "Category Encoders: a scikit-learn-contrib package of transformers for encoding categorical data." *Journal of Open Source Software* **3(21)**: 501. doi:10.21105/joss.00501