

IMPACT OF THE REGULARIZATION OF REGRESSION MODELS ON THE RESULTS OF THE MASS VALUATION OF REAL ESTATE*

Sebastian Gnat, Ph.D.

University of Szczecin
Institute of Economics and Finance
Mickiewicza 64, 71-101 Szczecin, Poland
e-mail: sebastian.gnat@usz.edu.pl
ORCID: 0000-0003-0310-4254

Received 26 September 2019, Accepted 30 March 2020

Abstract

Research background: Mass appraisal is a process in which multiple properties are appraised simultaneously, with a uniform approach. One of the tools that can be used in this area are multiple regression models. In the valuation of real estate features are often described on an ordinal or nominal scale. Replacing them with dummy variables with an insufficient number of observations leads to multicollinearity. On the other hand, there is a risk of overfitting the model. One of the ways to eliminate or weaken these phenomena is to introduce regularization based on a model's penalization for the high values of its weights.

Purpose: The aim of the study is to verify the hypothesis whether regularized regression reduces the errors of property valuation and which of the analyzed methods is the most effective in this context.

Research methodology: The article will present a study in which two ways of regularization will be applied – ridge and lasso regression, in the context of their impact on the errors of property valuation. The analyzed data set includes over 300 land properties valued by property appraisers. The key aspects of the study are the selection of optimal values of the regularization parameter and its influence on model's errors with a different number of observations in the training sets.

Results: The study showed that regularization improves valuation results and, more specifically, allows for lower average absolute percentage errors. The improvement of model effectiveness was more pronounced in the case of ridge regression. An important result is also that regularization has provided a higher accuracy of valuation compared to multiple regression models for smaller training sets.

Novelty: The article confirms the effectiveness of regularization as a way to eliminate the problem of multicollinearity or overfitting of the model. The results showed that ridge regression can be an effective

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

way of modelling the value of real estate. Especially in the case of a small amount of market data, which is an important conclusion in the context of the real estate market.

Keywords: property valuation, market analysis, regularization

JEL classification: C10, R30

Introduction

In the practice of real property valuation two main trends can be distinguished: individual and mass appraisal. In the process of an individual appraisal the entity valuing real estate focuses on one real property or on a small number of properties. Whereas in the case of mass appraisal the subject of an appraisal involves a large number of real properties of one type (e.g. Hozer, Kokot, Kuźmiński, 2002). When it comes to mass appraisal, the choice of method depends on the objectives and conditions regarding real properties. For example, in Poland, the legislator introduced three main objectives of mass property valuation: general property taxation, updating of perpetual usufruct fees and assessment of the economic effects of adopting and amending zoning plans. Mass valuations can be also useful for e.g. banks, which from time to time update the value of real estate, which are the basis for mortgage collateral. Mass appraisal methods can also support investment decisions. In practice and in the theory of real property mass appraisal, many models and algorithms can be differentiated (Jahanshiri, Buyong, Shariff, 2011).

The most frequently used models of property valuation are multiple regression models. Their popularity in property valuation, but also in other areas, results mainly from their simplicity and ease of interpretation. However, these advantages have a price to pay. In order to build a good model, a number of conditions need to be met (egg. Doszyń, 2012). One of the problems that may occur when estimating multiple regression models is the multicollinearity of variables and model overtraining due to the insufficient number of observations. One way to reduce undesirable effects is to regularize the model, which is typically achieved by constraining its weights (structural parameters). The article will present a study in which two ways of regularization (ridge regression and *LASSO*) will be applied and their influence on the errors of real estate valuation will be presented and discussed.

The aim of the study is to verify the hypothesis whether regularization reduces errors in property valuations and which of the analyzed methods is more effective in this context. The study also included the aspect of the size of the data set (training set), on the basis of which models are

estimated. It will be investigated whether the influence of regularization on valuation errors is stronger when the training set counts less observations.

The subject of the study was 318 properties located in Szczecin – one of the largest Polish cities. These properties were subject to valuation due to the revaluation of perpetual usufruct fees.

1. Literature review

The general review of quantitative methods used in mass appraisal could be found in (Pagourtzi, Assimakopoulos, Hatzichristos, French, 2003). In the article the methods are divided into traditional (multiple regression, comparable, cost, income, profit, contractors methods) and advanced, such as *ANN* (Artificial Neural Networks), hedonic pricing methods, spatial analysis, fuzzy logic, and *ARIMA* models.

An interesting comparison of modern approaches in mass appraisals is presented in (McCluskey, McCord, Davis, Haran, McIlhatton, 2013). In the survey such modelling approaches as multiple regression (*MLR*), spatial autoregression (*SAR*), geographically weighted regression (*GWR*), and *ANN* are compared. The neural network is widely used in the real estate market. The most common cases regard valuation, but market rents are also modelled (Muczyński, Walacik, 2017).

T. Kauko and M. d'Amato (2008) classify appraisal methods into four groups: model driven methods, data driven methods, methods based on machine learning and expert methods.

The literature concerning the possibility of applying econometric methods in appraisal is fairly extensive e.g. (Benjamin, Randall, Guttery, Sirmans, 2004; Isakson, 1998; Dell, 2017). Econometric methods are sometimes also used not directly in appraisal but, for example, to identify outlier transactions (e.g. Doszyń, Gnat, 2017).

Multiple linear regression models have known issues and there have been many disputes regarding the legitimacy of use of these models for real estate valuation, (Pawlukowicz, 2007; Barańska, 2010; Ligas, 2010; Zurada, Levitan, Guan, 2011; Bieda, 2018). There are some methods that can tackle some of the multiple regression problems, such as multicollinearity, little variance of explanatory variables and overfitting for small training data sets. One of those methods is regularization also known as shrinkage methods (Hastie, Tibshirani, Friedman 2008, p. 61). There is a plethora of studies regarding the regularization of regression models. Some of them concern research related to improvements in classical forms of regularization (Lipovetsky, 2010; Toker, Kaçiranlar, 2013; Hurvich, Simonov, Tsai, 1998; Durage 2014, Assaf, Tsionas,

Tasiopoulos, 2019). Studies related to regularization also concern the methods of determining the level of regularization hyperparameter (Golub, Heath, Wahba, 1979; Khalaf, Shukur, 2005; Ohishi, Yanagihara, Fujikoshi, 2020). There are also studies on the comparison of regularized models (Melkumova, Shatskikh, 2017; Fraley, Hesterberg, 2009; Rakesh, Suganthan, 2017; Pereira, Basto, Silva, 2016).

From the perspective of this study, the most important publications are those on the use of regularized models in the real estate market. It should be stated here that the literature in this area is not broad. The paper (Kubus, 2016) presents the possibility of using the local regularization of regression models. The proposed procedure has proved to be effective. However, this study concerns modelling on the basis of a large dataset. In the real estate market, it is not uncommon that the number of available transactions is very limited. Therefore, the presented application of regularization in the case of a small set of data on real estate is a novelty in the scope of the mass valuation of real estate.

2. The Szczecin Algorithm of Real Estate Mass Appraisal

As was previously mentioned, there are a number of methods of mass appraisal. One example of such a method is the Szczecin Algorithm of Real Estate Mass Appraisal (*SAREMA*). In the survey, an econometric form of this algorithm constitutes a point of reference to regularized models:

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

where:

- w_{ji} – market value of 1 square meter of i -th real estate in j -th location attractiveness zone,
- N – number of real estates ($i = 1, 2, \dots, N$),
- J – number of location attractiveness zones ($j = 2, 3, \dots, J$),
- $surf_i$ – surface of i -th real estate (in m^2),
- α_0 – constant term,
- K – number of real estate attributes,
- k_p – number of states of k -th attribute,
- α_{kp} – impact of p -th state of attribute k ,
- x_{kpi} – zero–one variable for p -th state of attribute k ,
- α_j – market value coefficient for j -th location attractiveness zone,

laz_{ji} – dummy variable equal one for j -th location attractiveness zone,
 u_i – error term.

The explained variable is a natural logarithm of a real estate unit value. Real estate values are determined by certified appraisers in the individual appraisals. Real estate attributes are qualitative characteristics measured on an ordinal scale, so they are introduced into the model (1) through dummy variables for each state of an attribute.

In the model (1) there is a constant term. In order to avoid the strict multicollinearity of the explanatory variables, each dummy variable for the worst attribute states are skipped. Hence the summation of $p = 2, \dots, k_p$ in the formula (1). In the interpretation, the ignored state of an attribute serves as a point of reference for the remaining states.

There are also market value coefficients (α_j) in the model (1). They could be treated as a proxy for location. They are estimated by introducing dummy variables for each location attractiveness zone. Location attractiveness zones are constructed by experts. They are defined as areas with a similar impact of location. Therefore, location attractiveness zones are constructed in such a way that the impact of location in the given area is homogenous.

Because of the strict collinearity of explanatory variables the worst (cheapest) location attractiveness zone is skipped. The omitted location attractiveness zone creates a point of reference.

3. Linear models regularization

Regularization is achieved by setting constrains for the weights of the model. Different kinds of algorithms implement those constraints in different ways. Two types of regularization will be used in this article. The first one is ridge regression and the second one is lasso regression. In multiple regression models, model weights are determined by minimizing the sum of squares of the residuals of the model ($RSS \rightarrow \min$). When it comes to ridge regression a regularization term equal to $\beta \sum_{i=1}^n \alpha_i^2$ is added to RSS cost function (Lesmeister, 2019, p. 107) of equation (1).

The hyperparameter β controls how much you want to regularize the model. If $\beta = 0$ then ridge regression is just a 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 data's mean (Geron, 2017, p. 201). Therefore setting β is a crucial stage of creating a model in order to achieve high quality results. In the case of lasso regression, the model weights are regularized by entering in the cost function

an expression $\beta \sum_{i=1}^n |\alpha_i|$. An important feature of this type of regularisation is the elimination of the least significant variables from the model. What causes lasso regression to be used for the selection of explanatory variables in addition to the regularization of weights of the regression equation? For both methods of regularization, the key stage is the selection of its strength – β . Figure 1 presents the results of preliminary ridge regression models for 500 draws, in which 250 properties were a training set and 68 a test set (more about the set of modelled properties in the next section). Nine different values of the regularization strength have been adopted. For each model, a mean absolute percentage error (*MAPE*) has been calculated. As it turned out, the level of the coefficient β strongly influenced the average *MAPE* for 500 models. This confirms that determining its optimum level of β is an important stage in modelling with the use of regularization.

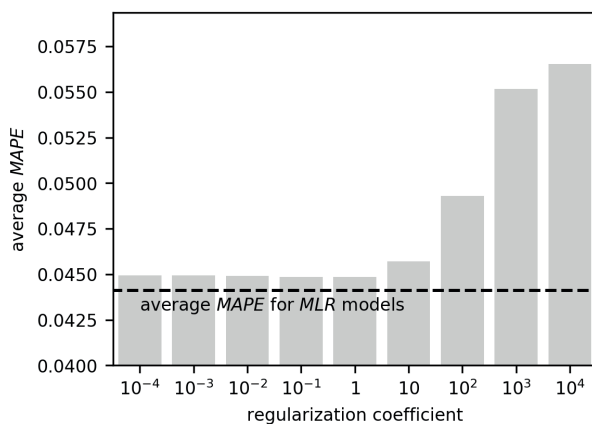


Figure 1. Comparison of average MAPEs for multiple regression models with ridge regression models taking into account different levels of regularization strength

Source: own work.

The selection of the value of this parameter can be carried out in a number of ways. From a completely random, arbitrary value, to the use of different machine learning techniques to optimize this parameter. In this study, the aim of adopting the appropriate strength of regularization is to obtain the best possible valuation results. The accuracy of the valuations obtained with estimated models will be determined by comparing them with the valuations carried out by certified property valuers.

4. Empirical study

The described *SAREMA* procedure will be used for the appraisal of 318 land plots located in the northern part of Szczecin, which is the capital of the West Pomeranian voivodeship, one of 16 Polish voivodeships. The real properties constitute a set for which an update of annual perpetual usufruct charges was conducted. The real properties were located in three clusters (referred as *LAZs*) of various numbers of real properties. The area within which the appraised real properties lie is shown in Figure 2.

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



Figure 2. Location of the valued properties

Source: own work.

Table 1. Real estate attributes and their states

Yeah.	Attribute	Attribute category/symbol
1	Utilities	None Incomplete Complete
2	Neighbourhood	Onerous Unfavourable Average Favourable
3	Transport availability	Unfavourable Average Favourable
4	Physical plot properties	Unfavourable Average Favourable
5	Plot area	Large (>1,200 m ²) Average (500–1,200 m ²) Small (<500 m ²)

Source: own work.

The accuracy of the valuations will be assessed on the basis of the absolute percentage error (*APE*):

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

where:

w_i – the actual unit value of the property determined by the property valuer,

\hat{w}_i – theoretical, unit value of the property determined from the model.

The empirical study was carried out according to the following scheme. The collection of 318 properties was divided 500 times into a test set of 68 properties and a training set of 250 properties. For each of the 500 training sets, the *SAREMA* model and its variants with ridge and *LASSO* regularization were estimated. In order to select the optimal strength of regularization, the procedure of a 3-fold cross-validation was carried out. 70 different values of the β coefficient from 0.0001 to 1,000 were tested. The models with best β were used to estimate the value of properties in the test sets.

The same procedure for estimating and testing the models was repeated with a reduced number of properties in the training sets. This time they consisted of 50 properties, with an unchanged number of test sets, i.e. 68 properties. This scheme has resulted in the estimation of more than 400,000 models (including those estimated for cross validation). The final result of the estimation was 3,000 models.

The target variable in the models was the value of 1 square meter of properties. The models reflect valuers estimates rather than the market as such. The results obtained allow us to determine how well the econometric models imitate the results of real estate appraisal conducted by valuers.

Figures 3–5 show the collective results of the comparison of non-regularized models with models for which regularization has been applied. Figure 3 shows the values of the coefficients of the determination of the estimated models. There are two main elements here. Firstly, because of the regularization, the *SAREMA* model's plain form R^2 is on average higher than that of the regularized models. This is expected since regularization tends to flatten the results. Secondly, determination coefficients for smaller training sets indicate another feature associated with multiple regression models (*MLR*), namely their tendency to overfit. With a smaller training set, the difference between R^2 coefficients for classic and regularized models is greater. In general, the values of these coefficients for models based on 50 properties are on average higher than for models based on 250 properties. Whether or not this is actually proof for overfitting will be revealed after the analysis of valuation errors in the test sets.

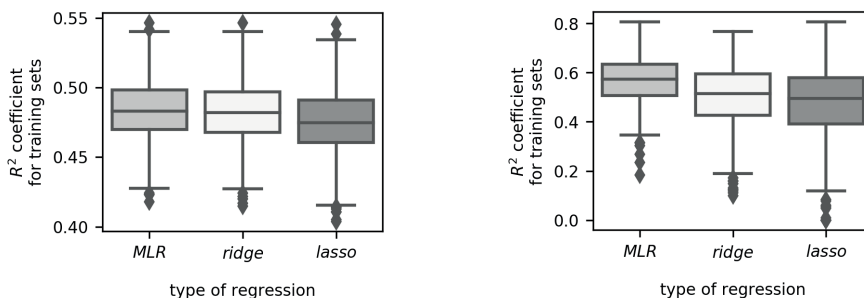


Figure 3. Distribution of determination coefficients for *SAREMA* models in the form of plain multiple, ridge and lasso regression. Training sets of 250 items (left side), training sets of 50 items (right side)

Source: own work.

To compare *MLR*, ridge and lasso models, the relative differences between *MAPE* for *MLR* and ridge models and *MLR* and lasso for both sizes of training sets were determined. Negative differences indicate that a lower *MAPE* occurred in a given test set for the classic *SAREMA* form; positive differences meant lower errors of the regularized models. Figures 4 and 5 show the distribution of these differences. In each comparison the number of positive differences was higher. This means that in most cases, the models with regularization resulted on average in lower valuation errors than the models without regularization. For models

estimated on the basis of larger training sets, ridge regression gave lower valuation errors in 299 cases out of 500. For lasso regression it was 269 cases. With small training sets, the advantage of regularized models was more frequent: 340 and 311 times, respectively. This confirms the hypothesis that the *MLR* models tend to overfit more often than regularized ones, especially in the case of a smaller number of properties in the training set. Interestingly, although lower errors were more often obtained for regularized models, in extreme cases that non-regularized models had a greater advantage over regularized ones. This was visible in the longer left tails of the distributions (especially for smaller training sets).

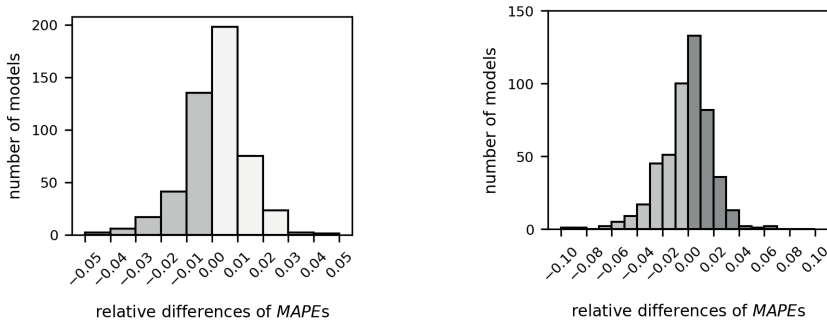


Figure 4. Distribution of relative *MAPE* differences between *SAREMA* models in the form of plain multiple and ridge regression (left side) and plain multiple and lasso regression (right side). Training collections of 250 properties

Source: own work.

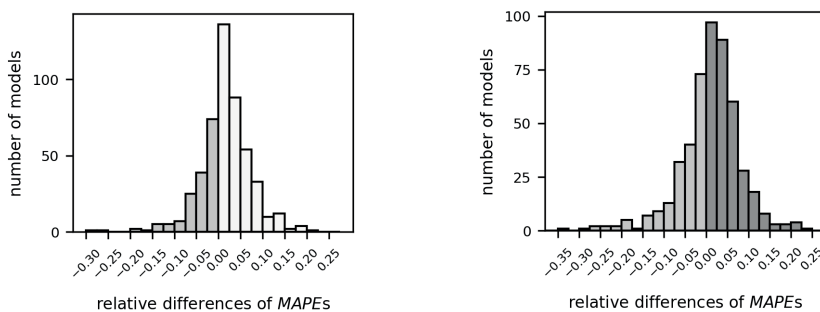


Figure 5. Distribution of relative *MAPE* differences between *SAREMA* models in the form of multiple and ridge regression (left side) and multiple and lasso regression (right side). Training collections of 50 properties

Source: own work.

The shares of regularized models with lower and higher valuation errors than plain multiple regression are presented in (Table 2). The use of regularization was particularly beneficial when the training sets were smaller. It allows avoiding overfitting and eliminates problems resulting from the multicollinearity of variables.

Table 2. Font sizes of headings (%)

	Ridge regression		Lasso regression	
	<i>N</i> = 250	<i>N</i> = 50	<i>N</i> = 250	<i>N</i> = 50
Number of <i>MAPE</i> models in a test set lower than for multiple regression	59.8	68.0	53.8	62.2
Number of <i>MAPE</i> models in the test set higher than for multiple regression	40.2	32.0	46.2	37.8

Source: own work.

Conclusions

As indicated in the literature, the linear model can be improved, by replacing plain least squares fitting with some alternative fitting procedures (James, Witten, Hastie, Tibshirani, 2017, p. 203). Such techniques allow to maintain the advantages of linear regression models, while improving the accuracy of prediction, eliminating the problem of the collinearity of variables or their low volatility (which may occur in the real estate market in particular) and increasing the interpretability of models. The article presents an example of using the regularization of multiple regression models in order to improve the results of the mass valuation of real estate. The results of the valuation of 68 properties estimated using the econometric form of the Szczecin Algorithm of Real Estate Mass Appraisal (*SAREMA*) with models supplemented with a component responsible for regularization in 500 repetitions were compared. Results obtained from 3,000 models show that regularization has in most cases reduced valuation errors in the test sets. The improvement in performance was more pronounced in the case of small training sets. Better results for both 250- and 50- element sets were obtained using ridge regression than lasso regression, so in this particular set of properties this first type of regularization proved to be a more effective way to minimize valuation errors. For small training sets the differences in *MAPE* were much higher than in larger sets (both in cases indicating an advantage of regularization and indicating an advantage of plain *MLR*). This means that with a larger training set, the impact of regularization improves (or worsens) the results of valuations to

a lesser extent. It can be concluded that the less data you have on the real estate market, the more worthwhile it is to apply regularization.

Further planned research will be aimed at verifying the results obtained in other markets as well as for other types of real estate.

References

- Assaf, A.G., Tsionas, M., Tasiopoulos, A. (2019). Diagnosing and correcting the effects of multicollinearity: Bayesian implications of ridge regression. *Tourism Management*, 71, 1–8. DOI: 10.1016/j.tourman.2018.09.008.
- Barańska, A. (2010). Modele multiplikatywne w procesie wyceny nieruchomości. *Studia i Materiały Towarzystwa Naukowego Nieruchomości*, 18 (1), 65–82.
- Benjamin, J.D., Randall, S., Guttery, R.S., Sirmans, C.F. (2004). Mass Appraisal: An Introduction to Multiple Regression Analysis for Real Estate Valuation. *Journal of Real Estate Practice and Education*, 7 (1), 65–77.
- Bieda, A. (2018). Conditional Model of Real Estate Valuation for Land Located in Different Land Use Zones. *Real Estate Management and Valuation*, 26 (1), 122–130.
- Dell, G. (2017). Regression, Critical Thinking, and the Valuation Problem Today. *Appraisal Journal*, 85 (3), 217–230.
- Dorugade, A.V. (2014). New ridge parameters for ridge regression. *Journal of the Association of Arab Universities for Basic and Applied Sciences*, 15, 94–99. DOI: 10.1016/j.jaubas.2013.03.005.
- Doszyń, M. (2012). Ekonometryczna wycena nieruchomości. *Metody Ilościowe w Ekonomii, Studia i Prace Wydziału Nauk Ekonomicznych i Zarządzania*, 26, 41–52.
- Doszyń, M., Gnat, S. (2017). Econometric Identification of the Impact of Real Estate Characteristics Based on Predictive and Studentized Residuals. *Real Estate Management and Valuation*, 25 (1), 84–93.
- Frabley, C., Hesterberg, T. (2009). Least angle regression and LASSO for large datasets. *Statistical Analysis and Data Mining*, 1 (4), 251–259. DOI: 10.1002/sam.10021.
- Golub, G.H., Heath, M., Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21 (2), 215–223. DOI: 10.1080/00401706.1979.10489751.
- Hastie, T., Tibshirani, R., Friedman, J. (2008). *The Elements of Statistical Learning*. New York, NY, USA: Springer New York Inc.

- Hozer, J., Kokot, S., Kuźmiński, W. (2002), *Metody analizy statystycznej rynku w wycenie nieruchomości*. Warszawa: Polska Federacja Stowarzyszeń Rzeczoznawców Majątkowych.
- Hurvich, C.M., Simonoff, J.S., Tsai, C.L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 60 (2), 271–293. DOI: 10.1111/1467-9868.00125.
- Isakson, H.R. (1998). The Review of Real Estate Appraisals Using Multiple Regression Analysis. *Journal of Real Estate Research*, 15 (2), 177–190.
- Jahanshiri, E., Buyong, T., Shariff, A.R.M. (2011). A review of Property Mass Valuation Models. *Pertanika Journal of Science Technology*, 19, 23–30.
- James, G., Witten, D., Hastie, T., Tibshirani, R. (2017). *An introduction to statistical learning: With applications in R*. New York, NY, USA: Springer New York Inc.
- Kauko, T., d'Amato, M. (eds.) (2008). *Mass Appraisal Methods. An international perspective for property valuers*. Wiley-Blackwell.
- Khalaf, G., Shukur, G. (2005). Choosing ridge parameter for regression problems. *Communications in Statistics – Theory and Methods*, 34 (5), 1177–1182. DOI: 10.1081/STA-200056836.
- Ligas, M. (2010). Metody statystyczne w wycenie nieruchomości. *Studia i Materiały Towarzystwa Naukowego Nieruchomości*, 18 (1), 49–64.
- Lipovetsky, S. (2010). Enhanced ridge regressions. *Mathematical and Computer Modelling*, 51 (5–6), 338–348. DOI: 10.1016/j.mcm.2009.12.028.
- McCluskey, W.J., McCord, M., Davis, P.T., Haran, M., McIlhatton, D. (2013). Prediction accuracy in mass appraisal: a comparison of modern approaches. *Journal of Property Research*, 30 (4), 239–265.
- Melkumova, L.E., Shatskikh, S.Y. (2017). Comparing Ridge and LASSO estimators for data analysis. *Procedia Engineering*, 201, 746–755. DOI: 10.1016/j.proeng.2017.09.615.
- Muczyński, A., Walacik, M. (2017). Neural Networks Modelling of Municipal Real Estate Market Rent Rates. *Folia Oeconomica Stetinensia*, 16 (2), 17–28. DOI: 10.1515/foli-2016-0022.
- Ohishi, M., Yanagihara, H., Fujikoshi, Y. (2020). A fast algorithm for optimizing ridge parameters in a generalized ridge regression by minimizing a model selection criterion. *Journal of Statistical Planning and Inference*, 204, 187–205. DOI: 10.1016/j.jspi.2019.04.010.
- Pagourtzi, E., Assimakopoulos, V., Hatzichristos, T., French, N. (2003). Real estate appraisal: a review of valuation methods. *Journal of Property Investment & Finance*, 21 (4), 383–401.

- Pawlukowicz, R. (2007). Użyteczność modeli ekonometrycznych w wycenie nieruchomości. *Zeszyty Naukowe Uniwersytetu Szczecińskiego. Prace Katedry Ekonometrii i Statystyki, Metody ilościowe w ekonomii*, 450, 453–466.
- Pereira, J.M., Basto, M., Silva, A.F. da. (2016). The Logistic Lasso and Ridge Regression in Predicting Corporate Failure. *Procedia Economics and Finance*, 39, 634–641. DOI: 10.1016/s2212-5671(16)30310-0.
- Rakesh, K., Suganthan, P.N. (2017). An Ensemble of Kernel Ridge Regression for Multi-class Classification. *Procedia Computer Science*, 108, 375–383. DOI: 10.1016/j.procs.2017.05.109.
- Toker, S., Kaçiranlar, S. (2013). On the performance of two parameter ridge estimator under the mean square error criterion. *Applied Mathematics and Computation*, 219 (9), 4718–4728. DOI: 10.1016/j.amc.2012.10.088.
- Zurada, J., Levitan, A.S., Guan, J. (2011). A Comparison of Regression and Artificial Intelligence Methods in a Mass Appraisal Context. *Journal of Real Estate Research*, 33 (3), 349–387.