sciendo

ECONOMETRICS. EKONOMETRIA
Advances in Applied Data Analysis
Year 2019, Vol. 23, No. 2
ISSN 1507-3866; e-ISSN 2449-9994

PREDICTION OF CHANGES IN THE TAX BURDEN OF LAND PLOTS WITH THE USE OF MULTIVARIATE STATISTICAL ANALYSIS METHODS*

Krzysztof Dmytrów
University of Szczecin, Szczecin, Poland
e-mail: krzysztof.dmytrow@usz.edu.pl
ORCID: 0000-0001-7657-6063

Sebastian Gnat
University of Szczecin, Szczecin, Poland
e-mail: sebastian.gnat@usz.edu.pl
ORCID: 0000-0003-0310-4254

DOI: 10.15611/eada.2019.2.03
JEL Classification: C38, C51, H71, R30

Abstract: It is believed that the ad valorem tax will increase fiscal burdens. In order to verify this statement, with the use of the Szczecin Algorithm of Real Estates Mass Appraisal, the land plots were appraised and the ad valorem tax was calculated. Next, a training set was sampled, for which the composite variable was calculated by means of three approaches: the TOPSIS method, the Generalised Distance Measure as the composite measure of development (GDM2), and the quasi-TOPSIS. They were the explanatory variables in the logistic regression model. Next, for the test set, changes of tax burden were forecasted. The aim of the research was to check the effectiveness of the presented approach for the estimation of the consequences of introducing the ad valorem tax. The results showed that all three approaches yielded similar results, but GDM2 was the best one. The main finding is that these approaches can be used in the prediction of changes in the tax burden of land plots.

Keywords: logistic regression, classification, multivariate statistical analysis, real estate mass appraisal.

1. Introduction

The ad valorem tax, i.e. a tax on real estate value, is one of the methods of taxing real estate. It exists in many countries in Europe and throughout the world. In Poland the fiscal solution involves a different, competitive method of tax assessment based

* The article is financed by the National Centre of Science within the scope of project No. 2017/25/B/HS4/01813.

on area, however the predominant opinion among fiscal system experts is that the ad valorem tax is a better fiscal tool. Taxing a real estate area is criticised as being ineffective [Etel, Dowgier 2013]. What is more, real estate tax fails to perform non-fiscal functions. It is not used as an instrument of a rational space management policy [Wójtowicz 2007; Gnat, Skotarczak 2006]. Since 1989 discussions have been conducted in Poland in the spheres of science and politics aiming at the implementation of a real estate ad valorem tax. Owing to substantial concerns of a socio-political as well as an economic and organizational nature, the implementation of the ad valorem tax (also called a cadastral tax) has not taken place. In order to examine the results of introducing the cadastral tax, many studies and simulations have been conducted producing a variety of results depending on the adopted assumptions. Conducting such a simulation requires defining the current tax burden, adopting an ad valorem tax rate value (understood as a percentage rate on the real property value) and, most importantly, defining the value of the real estate being subject to taxation. This stage is of particular significance since the financial dimension of the tax will depend on the appraised value. In Poland, appraisal of a real estate value requires a qualification certified by a professional licence of a property appraisal expert. Although the process of universal real estate taxation does not mean that every real estate is to be subject to appraisal but only representative real estate, even such a set is going to be numerous, and therefore will require significant financial outlays and, significantly, a lot of time. That fact makes the simulations difficult if they are based on reliable stages of ad valorem tax implementation carried out in reality. In the article an approach was proposed which, in respect of time as well as financial aspects, is far more effective. The objective of the study is to verify whether it is equally effective with respect to the precise evaluation of the effects of replacing the real estate area with its value as the basis for real estate tax assessment. This approach comes in the following stages:

1. Defining the real estate properties that affect its value.
2. Assigning the attributes of those properties to the real properties located in the analysed area.
3. Real estate appraisal with the use of the Szczecin Algorithm of Real Estate Mass Appraisal.
4. Adoption of an ad valorem tax rate.
5. Comparison of the amount of tax assessed on the basis of the real estate area and on the real estate value.
6. Selecting a sample of real estate that will constitute a training data set.
7. Calculating a composite variable for a training data set and for other real estate.
8. Estimating logistic regression models that predict the growth of the tax burden. In the model the composite variable will be the explanatory variable.
9. Evaluation of the effectiveness of the predictions both for the training data set and for the test data set.

The studies conducted on the effects of replacing the property tax with the ad valorem tax [Gnat 2010; 2016] indicate that such a process may lead to significant changes in the structure of tax burdens. The presented results show that the amount of fees would be significantly different from the current one. This applies to both higher and lower burdens. Such pilot studies, while very important and enabling the assessment of the impact of property taxation reform, are nonetheless time--consuming and costly. The assumption underlying the study concerns the issue of whether it is possible to carry out an effective simulation of changes in tax burdens without the need for conducting a mass appraisal for the entire analysed area. An affirmative answer will open up possibilities for further research (e.g. concerning the search for more effective classifiers or the introduction of a greater number of classes into the explanatory variable describing the assignment of tax burden increases into the ranges) and it will enable to confirm or disprove more cheaply and quickly the existing belief in the universal increase of the tax burden. In that regard the research will touch upon the issue of the sample size, on the basis of which a change in tax burdens can effectively be predicted. At present, in the initial stage the training data set will amount to 25%. The remaining real estate will constitute a test data set, which is an important aspect of the study. The smallest, and at the same time the cheapest, sample size must be found on the basis of which the classification efficiency will be sufficient so that it is possible to assess the possible effects of real estate tax reform with the use of the applied model.

The second major objective of the study involved the evaluation of the employed composite variable as an explanatory variable in the logistic regression model. This is significant because the majority of real estate properties, in the appraisers' practice, are recorded in an ordinal scale, which causes a host of complications in the modelling process. Confirmation of the efficient switch from such variables to a composite variable will also constitute an additional value, increasing the applicability of the proposed procedure. The distance of each real estate from the pattern (denoted by GDM2), the TOPSIS and quasi-TOPSIS measures will be adopted as the composite variable.

2. Methodology

The classification of tax burden changes was performed with the use of a logistic regression model (cf. inter alia [Batóg, Foryś 2011; Hastie et al. 2009]). The explained variable is a dummy variable. It assumes the value of 1, when the tax burden of a given real estate has grown and the value of 0 when it has decreased or if it has remained unchanged. The composite variable serves as the explanatory variable, which has been obtained from the attributes describing the real property (the values of individual attributes are described in brackets):

x_1 – land plot area (1 – large, 2 – average, 3 – small);

x_2 – location (1 – unfavourable, 2 – average, 3 – favourable);

x_3 – utility infrastructure (1 – none, 2 – incomplete, 3 – complete);

x_4 – shape (1 – poor, 2 – average, 3 – favourable);

x_5 – land use (1 – farmyards, 2 – industrial, 3 – multi-family, 4 – one-family, 5 – commercial);

x_6 – location attractiveness zone (the variable assumes the values between 1 and 32, where 1 means an area of the lowest average price of 1 m^2, and 32 – an area of the highest average price of 1 m^2).

All the variables are measured in an ordinal scale and they are stimulators, i.e. the highest possible values are desirable. On account of the above, the pattern vector is as follows: [3 3 3 3 5 32], whereas the anti-pattern vector [1 1 1 1 1 1]. The fact that the variables are measured in the ordinal scale renders the GDM2 distance to be a very good measure of the distance of each structure (real estate) from the standard and anti-standard [Walesiak 2016, p. 52]. Therefore a Generalised Distance Measure (GDM) was used for determining a composite measure. This is based on the generalized correlation coefficient, comprising the Pearson correlation coefficient and the τ Kendall correlation coefficient. This is calculated in accordance with the following formula [Walesiak 2016, p. 42]:

$$d_{ik} = \frac{1}{2} - \frac{\sum_{j=1}^{m} a_{ikj} b_{kij} + \sum_{j=1}^{m} \sum_{l=1, l \neq i,k}^{n} a_{ilj} b_{klj}}{2\left[\sum_{j=1}^{m} \sum_{i=1}^{n} a_{ilj}^2 \cdot \sum_{j=1}^{m} \sum_{i=1}^{n} b_{klj}^2\right]^{\frac{1}{2}}}, \tag{1}$$

where: d_{ik} – distance measure, i, k, l = 1, 2, …, n – structure number, j = 1, 2, …, m – variable number.

If variables are measured in the ordinal scale, then the only possible describable relations include elevation relations. Thus, values a and b in formula (1) are computed in the following fashion [Walesiak 2016, p. 45]:

$$a_{ipj}(b_{krj}) = \begin{cases} 1 & \text{for} \quad x_{ij} > x_{pj}(x_{kj} > x_{rj}) \\ 0 & \text{for} \quad x_{ij} = x_{pj}(x_{kj} = x_{rj}), \\ -1 & \text{for} \quad x_{ij} < x_{pj}(x_{kj} < x_{rj}) \end{cases} \text{ for } p = k, l, r = i, l. \tag{2}$$

It needs to be mentioned that the variables used for calculating the generalised distance measure can be assigned with weights. However, in the study it was assumed that all the variables (the attributes describing a real estate) are of the same weight.

The generalised distance measure can be applied in a multivariate statistical analysis to [Walesiak 2003, p. 135]:

1) determine distance matrices in the process of structure classification,

2) construct a composite development measure in linear ordering methods.

In this study the second application of the generalised distance measure was employed. With the use of the GDM a composite variable was built to describe real

estate and it was used for evaluating the logistic equation. Three approaches were applied in determining the composite measure:

1. GDM2 was employed in the TOPSIS method.

2. The composite variable was determined taking the distances of each real estate from the pattern (the approach was marked GDM2).

3. GDM2 was used for constructing a modified composite measure, called quasi--TOPSIS.

The stages of constructing the composite measure in the TOPSIS method (Technique for Order of Preference by Similarity to Ideal Solution) [Hwang, Yoon 1981] for GDM2 distances are as follows:

1. The starting point is provided by $[x_{ij}]$ matrix, containing the values of j-th variables (j = 1, 2, …, n, where n – number of variables) in i-th structures (i = 1, 2, …, m, where m – number of structures).

2. With the use of formula (1) with the substitution of a given formula (2) we determine the distance of each real estate from the pattern and we determine d_{i0}^{+} as well as from the anti-pattern, which we determine by d_{i0}^{-}.

3. We calculate the similarity of each structure (real estate) to the pattern with the following formula:

$$q_i = \frac{q_{i0}^{-}}{q_{i0}^{-}+q_{i0}^{+}}, \text{ where } i = 1, 2, \ldots, m. \tag{3}$$

Subsequently, the composite variable determined on the basis of equation (4) was used as an explanatory variable in the logistic model.

Even though the TOPSIS method is commonly applied both in multivariate statistical analysis as well as in multi-criterion decision-making, it appears that its construction has a certain defect. Namely, the distance of a structure from the anti--standard is featured both in the formula (3) numerator and denominator. Thus both the highest possible distance of the structure from the anti-pattern is desired (because then the formula numerator grows), and on the other hand, its smallest possible value is desirable as well (because then the denominator grows). That is why it was decided that a comparison would be made for the results obtained with the use of a classical TOPSIS method, in which the very distance of each real property from the pattern is a composite variable (a composite development measure). In such a case we substitute the values of d_{i0}^{+} to the logistic equation with an explanatory variable. For the purpose of the article this approach was termed GDM2. The approach may be considered as being equivalent to a classical composite measure of development proposed by Hellwig [1968]. Moreover, it was decided that the composite variable would be calculated with the following formula:

$$q_i^{*} = \frac{q_{i0}^{-}}{q_{i0}^{+}}, \text{ where } i = 1, 2, \ldots, m. \tag{4}$$

The above method of determining the composite variable was called a quasi--TOPSIS method. The statistical variable obtained on the basis of formula (4) was substituted in the logistic equation for the explanatory variable. It ought to be noted that the values of the composite variable values obtained with equation (3) and with the use of d_{i0}^{+} value are normalized within the range of ⟨0, 1⟩. The highest possible values of q_i^* and the lowest possible values of d_{i0}^{+} are desired. The values of the composite variable obtained with equation (4) are not normalized and it is best when they are as high as possible. Furthermore, in the case of equation (4) a situation is possible in which the value of q_i^* variable is undetermined, while the value of a variable provided with equation (3) is equal to 1. This is the case for an event when all the variables characterizing the analysed structure assume standard values. Then the distance from the pattern (d_{i0}^{+}) is equal to 0. In such a situation the following procedure was adopted:

- For a structure for which a variable provided with equation (4) cannot be calculated and for a structure which features its highest computable value, the value of a composite variable was determined for a classic TOPSIS method using equation (3) (for a structure for which q_i^* cannot be computed, q_i will be equal to 1, and the greatest computable value of q_i^* will be the greatest value of q_i). These will be the values of, in order: $q_{\max} = 1$ and $q_{\max-1}$.
- The following relation was calculated $\frac{q_{max}}{q_{max-1}} = \frac{1}{q_{max-1}}$.
- Then the maximum computable value of q_i^* variable was multiplied by the above--specified relation and the value of q_i^* was obtained for a structure featuring the same attribute values as the standard structure.

The values of q_i, q_{i0}^{+} and q_i^* are substituted in the logistic equation and three different models are estimated. If the theoretical value of the explained variable is higher than 0.5 then it is predicted that the tax burden will increase, otherwise a decrease of the tax burden is forecasted.

The estimated models will be subject to a two-stage evaluation. A confusion matrix will be calculated [Szeliga 2017] for a training data set, i.e. the one which served for determining the parameters of a logistic curve and for the test data set. Four measures will be used for the model evaluation:

1. Accuracy – defines what part of the predicted class labels is consistent with the real results:

$$ACC = \frac{TP + TN}{TP+TN+FP+FN}. \tag{5}$$

2. Sensitivity (fraction of true positive TPR) – defines the fraction of true positive classifications with respect to all positive cases:

$$TPR = \frac{TP}{TP+FN}. \tag{6}$$

3. Precision of positive prediction (PPV) – defines the fraction of true positive classifications with respect to all positive classifications:

$$PPV = \frac{TP}{TP+FP}. \tag{7}$$

4. The F_1 measure – the harmonic mean of the precision and sensitivity:

$$F_1 = \frac{2TP}{2TP+FP+FN}. \tag{8}$$

Additionally, in the case of a test data set, ROC curves will be determined [Raschka 2018; James et al. 2015] and Area Under the Curve (AUC) will be calculated for them, which also serve for the evaluation of classification models quality.

3. Real estate mass appraisal

The process of real estate mass appraisal (i.e. universal real estate taxation) which will need to precede the introduction of the cadastral tax in Poland, is regulated in the Real Estate Economy Act of 21 August 1997 and the regulation of the Council of Ministers dated 29 June 2005 on universal real estate taxation. Real estate mass appraisal is a process in which the market (cadastral) value of a number of real properties is determined simultaneously. Such an approach requires the use of a specific procedure based on mathematical models. The proposed real estate mass appraisal algorithms can be found in the works of many authors (cf. inter alia [Hozer et al. 1999; Czaja 2001; Sawiłow 2009]). The study was based on a mass appraisal procedure developed by J. Hozer:

$$\widehat{W}_{ji} = WWR_{ji} \cdot pow_i \cdot C_{baz} \prod_{k=1}^{K} \prod_{p=1}^{k_p} (1 + A_{kpi}), \tag{9}$$

$$\widehat{WWR}_i = \frac{C_{ri}}{C_{hi}}, \tag{10}$$

where: W_{ij} – market (or cadastral) value i-th real property in j-th location attractiveness zone, WWR_{ij} – market value coefficient in j-th location attractiveness zone (j = 1, 2, ..., J), J – number of location attractiveness zones, pow_i – surface of i-th real property, C_{baz} – price of 1m^2 of the cheapest land (without the utility infrastructure) in the appraised area, A_{kpi} – influence of p-th category of k-th attribute (k = 1, 2, ..., K; p = 1, 2, ..., k_p) in i-th property, K – number of attributes, k_p – number of categories of k-th attribute, C_{ri} – real property value determined by a real estate appraiser, C_{hi} – hypothetical value determined in line with the following formula:

$$C_{hi} = pow \cdot C_{baz} \cdot \prod_{p=1}^{k_p} \prod_{k=1}^{K} (1 + A_{kpi}). \tag{11}$$

The amounts of real estate taxes were determined for urbanized land located within the territory of one municipality of the Zachodniopomorskie (West Pomeranian) Voivodeship in line with a resolution of its Municipality Council in force in 2015. Hence, no real tax burden amounts were used, only their approximations. The

values obtained as a result of the algorithm application constituted the grounds for assessing an ad valorem tax amount. The subject of appraisal in the study involved 2,337 land plots.

In order to define the various effects of the real estate taxation reform, a percentage rate of an ad valorem tax was used, which will ensure income into a municipality's budget equal to the one determined for a real estate tax. Such a rate, on the one hand, prevents a situation in which there is a significant increase of inflows into the budget at the expense of the entities obligated to pay real estate tax, and on the other hand, it ensures a balanced share of real estate featuring the increase and decrease of tax burdens.

Figure 1 demonstrates a fragment of the area of the analysed municipality. The darker colour marks the land plots for which the tax burden increased with the assumed cadastral tax rate; the lighter colour designates the land plots which owing to their low real estate value feature a lower ad valorem tax amount than the current real estate tax amount. The land plots that are coloured white in the figure constitute non-urbanized plots and thereby they are not subject to a real estate tax, hence they are not the subject of the study.

Fig. 1. Part of the analysed municipality taking into account urbanised plots in terms of change in tax burden

Source: own work.

There are two classes designating land plots in this study, for which a change of fiscal burden after replacing the area-based tax with value-based tax is positive (fiscal burden increase 1) and for which it is negative (fiscal burden decrease 0). A majority of land plots feature a fiscal burden increase, but the difference in the percentage is not very significant. The percentage of land plots that feature a fiscal burden increase amounts to 54.6%. The land plots for which the simulations demonstrates a drop in the fiscal amount is equal to 45.5%.

4. Research results

As mentioned in Section 3, the number of urbanized land plots in the analysed municipality amounted to 2,337. At this juncture, it needs to be pointed out that the number of combinations of property variants and specified location attractiveness zones is lower, therefore there are numerous repetitions in the real estate data set. For further analyses the real estate data set was reduced so that at the same time it contained every combination of properties occurring in the set, thus the reduced data set contains 210 elements. In the so reduced data set a greater number of land plots features a fiscal burden decrease. The source of this situation stems from the fact that in the original data set there are numerous subsets of land plots, such as the ones presented in Figure 1, which have identical variants of properties and which feature a fiscal burden increase. The diversification of land plots featuring a fiscal burden decrease is greater, which translates into a greater percentage of such types of plots in the reduced data set.

Out of the set of 210 land plots, a training data set was selected numbering 25%, i.e. 52 elements and a test data set numbering 158 elements. For both sets composite measures were calculated, which served as composite variables in the logistic regression models. In Figure 2 confusion matrices were presented for the estimated models with the TOPSIS, GDM2 and quasi-TOPSIS variables. The modelling and classification evaluation in the study were conducted with the use of the Scikit-Learn library of Python programming language [Pedregosa et al. 2011]. On the grounds

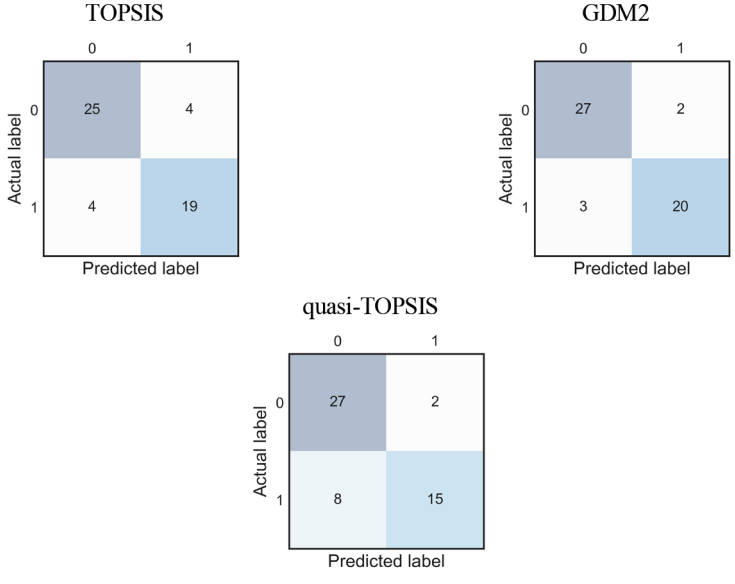

Fig. 2. Confusion matrices (training data set)

Source: own work.

Table 1. Measures of classification assessment (training data set)

Classification measure	Composite variable		
	TOPSIS	GDM2	quasi--TOPSIS
accuracy	84.62%	90.38%	80.77%
sensitivity	82.61%	86.96%	65.22%
precision	82.61%	90.91%	88.24%
F_1	82.61%	88.89%	75.00%

Source: own work.

of the matrices the previously indicated classification measures were calculated, which are presented in Table 1.

At the stage of the analysis of the results obtained for the sample of real estate amounting to 25% of the population, the most important results were obtained for the model in which the GDM2 measure constituted the explanatory variable. In that case, only 5 real properties out of 52 were erroneously classified. The worst results (apart from classification accuracy) occurred in the third model, i.e. for the quasi-TOPSIS explanatory variable. Overall, the obtained results ought to be assessed highly, particularly in the case of the model containing the GDM2 variable, in which all the classification measures exceeded 85%.

The question of which classification measure will be recognized as the most important one depends on the nature of the explained variable. In the event of a fiscal burden increase, sensitivity was recognized as the most significant measure. This is important for a model to correctly indicate the greatest possible percentage of the actual taxation increases, and occurs at the expense of precision, which causes a greater percentage of "false alerts", i.e. the indications of fiscal burden increase, whereas in reality the tax amount decreases. The search for a compromise between the degree of precision and sensitivity may be achieved through a change in the level of threshold probability above which the observation is assigned the value of 1 for the explained variable. In the study a typical level of threshold probability was assumed, namely 0.5.

The theoretical values of logistic regression models in comparison to the explained variable in the training data set are presented in Figure 3. The figure also demonstrates the level of threshold probability (equal to 0.5) and the resulting classification limit. Decreasing or increasing the threshold changes the classification results.

The most important element of the study is to define the effectiveness of the predicted fiscal burden changes for the test data set. It is only at this stage that it becomes evident whether the models constructed on the set constituting a quarter of the population enable an effective evaluation of whether individual real property types can expect a tax amount increase or decrease as a result of the real estate tax being replaced with the ad valorem tax. The assessment will be conducted on the basis of the measures analogous to the training data set and with the application of ROC curves. The analysis of the results presented in Table 2 demonstrates that the model in which the quasi-TOPSIS measure constituted an explanatory variable proved to be the best model. This is interesting because the model for that variable

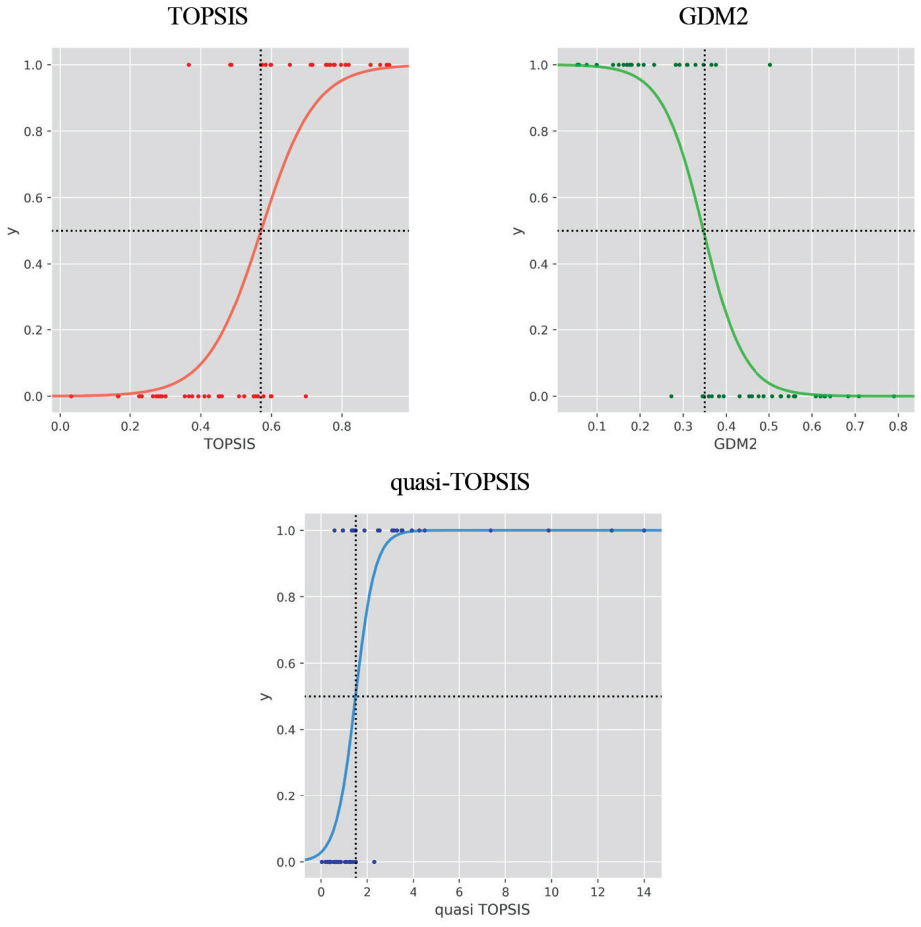

Fig. 3. The course of logistic function for individual methods of determining a composite variable (training data set)

Source: own work.

had the poorest results in the training data set. The remaining models in the test data set yielded worse results than in the training data set. Such a situation means that the models were overfitted. They fitted well to the data, on the basis of which the models were estimated, but the effect did not reoccur in the test data set.

Despite the fact that all the models produced good results, it is worth focusing on classification errors. The most significant parameters measuring classification quality is sensitivity and precision. From the standpoint of the fiscal burden, what seems to be most important is the fact that the model predicted their decrease, while in fact they increased. For the TOPSIS method and for the composite measure of development, determined on the basis of GDM2 distance, this was the case for five real properties

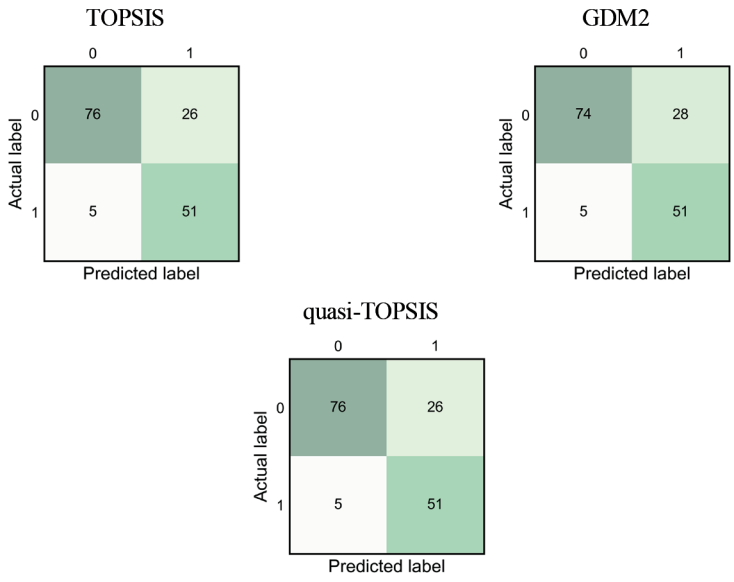

Fig. 4. Confusion matrices (test data set)

Source: own work.

and for the quasi-TOPSIS method for six of them (Figure 4). During the analysis of the differences it was found that this occurred for the real properties located in the location attractiveness zones of relatively low average prices of 1 m^2. In turn, as far as precision is concerned, i.e. the percentage of correctly classified increases in fiscal burden in all predicted increases, the converse is true. The increases in fiscal burdens that were incorrectly classified chiefly referred to the land plots that were located in the location attractiveness zones of relatively high average price of 1 m^2.

Table 2. Measures of classification assessment (test data set)

Classification measure	Composite variable		
	TOPSIS	GDM2	quasi-TOPSIS
accuracy	80.38%	79.11%	82.91%
sensitivity	91.07%	91.07%	89.29%
precision	66.23%	64.56%	70.42%
F_1	76.69%	75.56%	78.74%

Source: own work.

In each of the employed methods a significantly greater classification sensitivity was achieved in relation to precision. As far as sensitivity goes, all of the methods yielded similar results at the level of approximately 90% (the TOPSIS and GDM2

methods produced slightly better results than the quasi-TOPSIS method). In turn, the classification precision featured far greater variability and the best results were produced by the quasi-TOPSIS method, which correctly classified 70% of the fiscal burden increases. The worst result came with the use of the GDM2 method (which correctly classified 64.5% of the fiscal burden increases). An improvement of the sensitivity is achieved at the expense of precision, calculating F_1, one may take into consideration both of the above-mentioned measures at the same time. The F_1 measure demonstrates that the best classification results were produced by the quasi--TOPSIS method. The TOPSIS method was the second best and GDM2 yielded the poorest results. The same classification order was demonstrated by the coefficient measuring accuracy.

The classification assessment based on the ROC curves and the areas under the curves presented in Figure 5 further demonstrates that classification assessments do not differ significantly from one another. The highest value of the area under the ROC curve occurred for the classification based on the GDM2 variable. Owing to the fact that for the classification conducted in the study the correct identification of the land plots with increased fiscal burden is particularly important, sensitivity was deemed to be the most important measure of classification assessment. It reached the highest value in the case of GDM2 and TOPSIS methods.

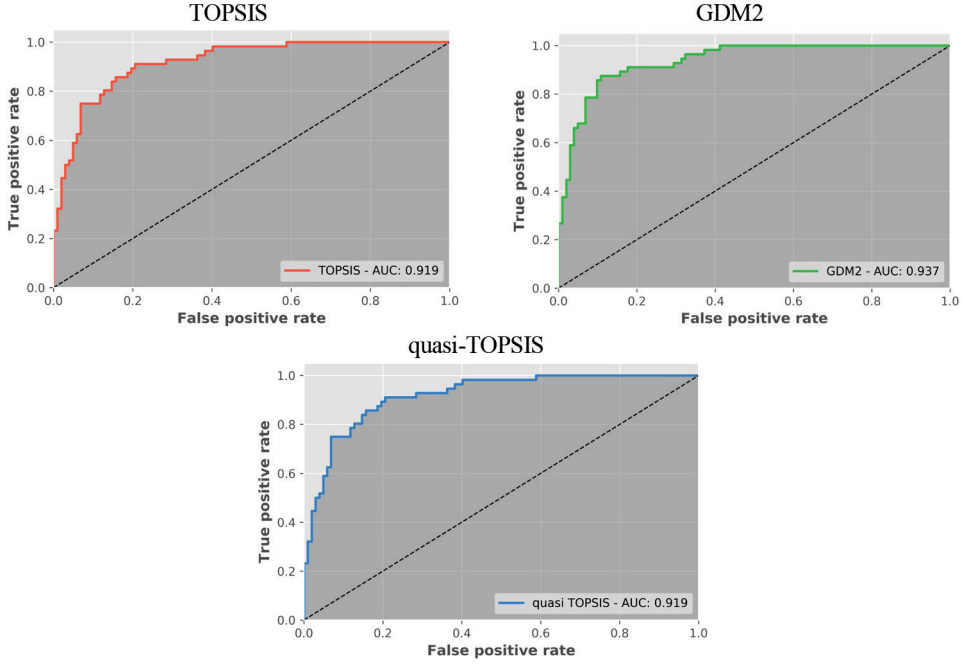

Fig. 5. ROC curves (test data set)

Source: own work.

5. Conclusion

The study focused on examining the quality of real property classification in terms of changes in fiscal burdens. This was achieved through the logistic regression, and the composite measure, which was created out of six attributes describing real properties, used as the explained variable. The attributes included the following properties: surface area, location, utility infrastructure, shape, intended use, location attractiveness zone. Because all of the attributes were measured in the ordinal scale, the generalized distance measure for the variables measured in the ordinal scale (GDM2) was used in the creation of the composite measure. Three approaches to the creation of a composite measure were compared. In the first one the TOPSIS method was applied, in the second one the distances of each real property from the standard created a composite measure of development (GDM2), and in the third one the distances of each real property from the standard and anti-standard were used and on their basis a composite measure (quasi-TOPSIS) were used. A logistic function was created on the grounds of a training data set, numbering 25% of all the real properties, and then the classification quality was examined on a test data set. Typical applications of classification algorithms are built for the largest possible training data sets and smaller test data sets. Here a different approach was adopted. In order to minimize the costs of conducting the analysis of a potential increase of fiscal burdens the sample ought to be as small as possible, but sufficiently numerous to ensure high measure values of the classification assessment. All three approaches yielded good results (accuracy at the level of 80%, sensitivity on average 90%, precision on average approximately 67%, F_1 at the level of 77%), the differences between them were not significant. The best prediction of fiscal burden changes was obtained with the use of the GDM2 approach. This was also confirmed through the analysis of the area under the ROC curve. The logistic regression model has already been used in forecasting changes in real estate tax burdens [Gnat 2018]. The results of this study indicate that this model allows to obtain accurate predictions. The survey was conducted on the basis of the entire population of land plots, and the results of the accuracy of the predictions were only slightly better than those obtained on the basis of the 25% sample used in this survey. The conducted research proved that the idea of the creation of a composite variable from the values of attributes and its application as an explanatory variable in the model of logistic regression can be successfully used in the prediction of changes in the tax burden of land plots.

The next stage of the research will involve analysing the classification quality by adopting a different level of threshold probability.

Bibliography

Batóg B., Foryś I., 2011, *Modele logitowe w analizie transakcji na warszawskim rynku mieszkaniowym*, Studia i Materiały TNN, 19(3), pp. 33-48.

Czaja J., 2001, *Metody szacowania wartości rynkowej i katastralnej nieruchomości*, Komp-system, Kraków.

Etel L., Dowgier R., 2013, *Podatki i opłaty lokalne – czas na zmiany*, Temida 2, Białystok.

Gnat S., 2010, *Analysis of the effects of replacing current property tax with ad valorem property tax in a sample municipality*, Folia Oeconomica Stetinensia, no. 8(16), pp. 82-98.

Gnat S., 2016, *Powierzchniowy a katastralny system opodatkowania nieruchomości – symulacja wybranych skutków fiskalnych*, Annales Universitatis Mariae Curie-Skłodowska Sectio H. Oeconomia, vol. 50 (1), pp. 371-380.

Gnat S., 2018, *Model logitowy jako narzędzie prognozowania obciążeń podatkowych działek gruntu w wyniku wprowadzenia podatku ad valorem*, Studia i Prace WNEIZ US, nr 54/3, pp. 173-183.

Gnat S., Skotarczak M., 2006, *Analiza rozkładów stawek podatków lokalnych w gminach województwa zachodniopomorskiego w latach 2002–2004*, [in:] J. Hozer (ed.), *Koniunktura gospodarcza a rynek nieruchomości*, Uniwersytet Szczeciński, Instytut Analiz, Diagnoz i Prognoz Gospodarczych, Szczecin, pp. 74-82.

Hastie T., Tibshirani R., Friedman J., 2009, *The Elements of Statistical Learning*, Springer, New York.

Hellwig Z., 1968, *Zastosowanie metody taksonomicznej do typologicznego podziału krajów ze względu na poziom rozwoju oraz zasoby i strukturę wykwalifikowanych kadr*, Przegląd Statystyczny, no. 15(4), pp. 307-326.

Hozer J., Foryś I., Zwolankowska M., Kokot S., Kuźmiński W., 1999, *Ekonometryczny algorytm masowej wyceny nieruchomości gruntowych*, Uniwersytet Szczeciński, Stowarzyszenie Pomoc i Rozwój, Szczecin.

Hwang C.L., Yoon K., 1981, *Multiple Attribute Decision Making: Methods and Applications*, Springer-Verlag, New York.

James G., Witten D., Hastie T., Tibshirani R., 2015, *An Introduction to Statistical Learning*, Springer, New York.

Pedregosa F. et al., 2011, *Scikit-learn: machine learning in Python*, Journal of Machine Learning Research, no. 12, pp. 2825-2830.

Raschka S., 2018, *Python. Uczenie maszynowe*, Wydawnictwo Helion, Gliwice.

Sawiłow E., 2009, *Zastosowanie metod wielowymiarowej analizy porównawczej dla potrzeb ustalania wartości katastralnych*, Studia i Materiały TNN, vol. 17, no. 1, pp. 105-115.

Szeliga M., 2017, *Data science i uczenie maszynowe*, PWN, Warszawa.

Walesiak M., 2003, *Uogólniona Miara Odległości GDM jako syntetyczny miernik rozwoju w metodach porządkowania liniowego*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu, nr 988, Taksonomia 10, *Klasyfikacja i analiza danych – teoria i zastosowania*, Wrocław, pp. 134-144.

Walesiak M., 2016, *Uogólniona miara odległości w statystycznej analizie wielowymiarowej z wykorzystaniem programu R*, Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, Wrocław.

Wójtowicz K., 2007, *System opodatkowania nieruchomości w Polsce*, *Finanse publiczne,* Wyd. UMCS, Lublin.

PROGNOZOWANIE ZMIAN OBCIĄŻEŃ PODATKOWYCH GRUNTÓW Z WYKORZYSTANIEM METOD WIELOWYMIAROWEJ ANALIZY STATYSTYCZNEJ

Streszczenie: W powszechnie panującej opinii podatek *ad valorem* ma doprowadzić do wzrostu obciążeń podatkowych. W celu weryfikacji tego stwierdzenia, przy zastosowaniu szczecińskiego algorytmu masowej wyceny nieruchomości, wyceniono wartość gruntów oraz wysokość podatku *ad valorem*. Następnie wylosowano próbę uczącą, dla której utworzono zmienne syntetyczne, użyte jako zmienne objaśniające w modelu regresji logistycznej. Zastosowano trzy podejścia w wyznaczeniu zmiennej syntetycznej: metodę TOPSIS, uogólnioną miarę odległości jako syntetyczny miernik rozwoju (GDM2) i *quasi*-TOPSIS. Następnie dla próby testowej zaprognozowano zmianę obciążeń podatkowych. Celem badania była weryfikacja hipotezy o skuteczności zastosowanego podejścia do oceny skutków wprowadzenia podatku *ad valorem*. Otrzymane wyniki wykazały, że wszystkie trzy podejścia dały podobne rezultaty (najlepsze dla GDM2). Głównym wnioskiem jest to, że zaprezentowane podejścia mogą być stosowane w prognozowaniu zmian obciążeń podatkowych gruntów.

Słowa kluczowe: regresja logistyczna, klasyfikacja, wielowymiarowe analiza statystyczna, masowa wycena nieruchomości.