sciendo

TESTING THE EFFECTIVENESS OF OUTLIER DETECTING METHODS IN PROPERTY CLASSIFICATION

Sebastian Gnat
Faculty of Economics, Finance and Management
Department Econometrics and Statistics
University of Szczecin
e-mail: sebastian.gnat@usz.edu.pl, ORCID ID: 0000-0003-0310-4254

Abstract

The introduction of the property value tax in Poland may lead to an increase in the tax burden on real estate. Pilot studies may be carried out on samples and the results should feature a high degree of certainty as to the extrapolation of the results on populations (e.g. entire municipalities). Each study may, for various reasons, include outliers in the analyzed data sets. If their presence results from measurement errors or other reasons that cause such observations not to be the result of naturally occurring processes, they should be omitted in the calculations, because they interfere with the study of the occurring regularities.

The study presents the results of statistical modelling carried out to determine whether individual objects (land properties), due to their attributes, are at risk of increasing the tax burden as a result of the introduction of *ad valorem* tax. First, logistic regression model estimation was carried out for the entire set of analyzed properties. Next, several methods of outlier detection were applied, and model estimation was repeated without the observations, i.e. real estates, pointed out as abnormal.

The objective of the study is to verify the usefulness of outlier detecting methods in the context of improving the classification results of the analyzed properties.

Key words: *outlier detection, logistic regression, real estate market analysis, property taxation*.

JEL Classification: *C38, H71, R30*.

Citation: Gnat, S. (2020). Testing the effectiveness of outlier detecting methods in property classification. *Real Estate Management and Valuation*, 28(4), 81-92.

DOI: https://doi.org/10.1515/remav-2020-0033

1. Introduction

The predominant systems of land property taxation are those where the taxable amount of the property is based on its area or value. In Poland, taxation of land property is based on its area (Act on Local Taxes and Fees of 12 January 1991, Journal of Laws of 2019, item 1170). Such a system has been repeatedly criticized (Etel & Dowgier, 2013; Wójtowicz, 2006). For several decades there have been discussions on replacing property taxation based on area with taxation based on property value. Such a solution is considered to be superior from the point of view of entities obtaining income from property taxation, but worse - from the point of view of taxpayers. The main obstacle here is the public perception that an *ad valorem* tax will result in a significant increase in the tax burden, even leading, in some cases, to the insolvency of taxpayers. Various types of pilot and simulation studies are being carried out, which are confronting a number of issues related to property taxation reform (Trojanek & Kisiała, 2016), including the scale of changes in the tax burden (Głuszak & Marona, 2015; Gnat, 2009, 2018). The pilot studies can be carried out on samples and the results should be highly reliable in extrapolating the results to populations (e.g. entire municipalities). There are many factual,

technical and computational reasons that may reduce this reliability. One such reason is that research may be conducted on the basis of data sets containing outliers. If their presence is due to measurement errors or if there is a risk that such observations are not the result of naturally occurring processes, they should be omitted from the calculation, as they interfere with the study of the occurring regularities. The article presents the results of a study on the possibility of applying logistic regression to the classification of land properties into one of two categories - properties that are characterized by an increase and a decrease in the amount of tax as a result of area-based tax being replaced with a value-based tax. The subject of the study involved 524 land properties intended for residential purposes. These properties were described with the use of several attributes and valued. Such a data set may include properties whose value, with given attributes, differs from properties similar to them. Such atypical cases may interfere with the results of the classification and create a risk of misrepresentation of potential changes in the tax burden of the property. This risk should be as small as possible. Therefore, the objective of the study is to verify the hypothesis stating that removing outlier observations improves classification results. The second objective is to determine, in the case of positive verification of the research hypothesis, which of the methods works best in this respect, i.e. enables the highest accuracy of classification to be achieved. Simulation and pilot studies on the reform of property taxation in Poland are of broad significance. Firstly, they make it possible to create various scenarios and assess their effects. The results obtained may give a broader perspective on the replacement of area-based tax with value-based tax. The results obtained in such studies may confirm or contradict the common belief regarding the dangers of an *ad valorem* tax. Secondly, the introduction of a value-based tax on real estate will be a project that requires the use of quantitative methods. From this point of view, testing various solutions, methods and calculation algorithms increases the knowledge of their applicability in the area of real estate market analysis, mass valuation or general taxation process.

2. Literature review

Outlier detection, also referred to as anomaly detection, is a process of identifying unexpected items or events in data sets, which differ from the norm. Hawkins (1980) defines an outlier as an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism. Johnson (1992) defines an outlier as an observation in a data set which appears to be inconsistent with the remainder of that set of data. The detection of anomalies is based on two basic assumptions: the anomalies rarely appear in data; their characteristics differ significantly from normal cases. Anomalies in data appear for various reasons. They can be connected to some kind of a problem or a rare event, such as, e.g., bank fraud, medical problems, structural defects, malfunctioning equipment. Outlier observations are detected for two reasons. The first reason is to try to identify events that require specific action. Such a situation occurs in the banking or insurance industry when attempts to detect fraud are made. The second reason is an attempt to remove such observations from the analyzed sets, such observations that bear the signs of measurement errors or, for any other reason, cannot be considered to have been made by the same mechanisms as the majority. In such a case, these observations should be removed because they disrupt the observed accuracy and their inclusion may lead to inaccurate conclusions. Research on the detection of outlier observations has been conducted for several decades (e.g. Grubbs, 1969; Stefansky, 1972; Cook, 1977). These and other numerous studies have resulted in the creation of at least a dozen or so procedures for detecting anomalies in data (e.g. Maimon & Rokach, 2005; Aggarwal, 2017). Initial research was the basis on which many more studies were conducted, both theoretical (e.g. ZHU, et al., 2011) and application-related. Statistical methods for detecting outliers are used in medicine (Prasta, et al., 2004), computer science (Liu, et al., 2004) and other fields (Worden, et al., 2000). Many scientific papers direct research towards the comparison of the results of outlier detection methods (Pimentel, et al., 2014; Domingues, et al., 2018). The detection and analysis of atypical objects may lead to finding unique investment opportunities (Ng & Khor, 2016). Authors proved that their research had successfully proposed an investment tool that rapidly identifies outstanding REIT stocks using two outlier detection algorithms. Detection of market opportunities has also been of interest in the Spanish real estate market (Baldominos, et al., 2018). The authors explored the application of diverse machine learning techniques with the objective of identifying real estate opportunities for investment. Identifying atypical properties may also mean searching for similar and dissimilar properties, which is of particular importance in the valuation process (Zyga, 2016; Doszyń, et al., 2017). Kontrimas and

Verikas (2006) used methods of detection of unusual real estate transactions, recognizing that they may result from non-market conditions or from attempts to understate taxes paid in connection with the transaction. They concluded that outliers must be detected and removed, or parameters of valuation models must be estimated using robust techniques. Such techniques on real estate data are tested by Śpiewak (2018). Among methods tested in this article, Baarda's algorithm was deemed to be better. The need to remove outliers in econometric modelling is also advocated by Morano, De Mare and Tajani (2013). They stated that "the presence of even a single outlier in a sample estimate can have strong repercussions on the regression models obtained with the method of least squares, nullifying its reliability". Their study revealed "that the regression model, which was initially to be rejected, showed instead excellent performance once all the outliers (...) were removed from the sample". A review of the literature, especially the publications connected with the real estate market, shows that most applications of outlier detection techniques are related to the preparation of data for estimation of regression models. Few studies are conducted in the area of real estate classification. This gap, to some extent, is filled by this study.

3. Data and methods

As mentioned in the introduction, 524 land properties were the subject of the study. These plots of land are part of a larger base which was used to simulate the introduction of the cadastral tax (Gnat, 2009). In this study, only a part of this base was used to select properties for housing purposes. The properties were described in previous studies (no changes were made in this respect for the purposes of possible confrontation of results) using the following attributes:

- area - large (above 5000 m^2), medium (between 1000 and 5000 m^2), small (below 1000 m^2),
- location - unfavorable (plots located in the smallest towns and on the outskirts of larger towns), average (plots located in larger towns), favorable (plots located in areas considered attractive by potential market participants),
- utility access - incomplete (usually without a sewage system), full,
- land plot shape - bad (difficult to build on, narrow, irregularly shaped plots), good (similar to a square or rectangle, easy to build on and use).

These properties were also valued in earlier studies using one of the mass valuation methods. For the purpose of this study, which is to evaluate methods for detecting unusual properties, the results of valuation were randomly changed. For 10% of the properties, the result obtained in the process of mass valuation was distorted in the range from -50% to 200% of the originally obtained value. In this way, a group of properties with too low or too high value was created in relation to the level of the attributes describing them. This procedure was carried out because mass valuation using a regression model with a specific mathematical form will not yield outlier results. This is due to the fact that the values obtained from the model result from the structural parameters of the model and there is no possibility that properties with the same states of market characteristics have different values.

The current tax burden was determined for the analyzed real estate in accordance with the applicable regulations on real estate tax in the municipality under study.

The correct approach in modelling - using qualitative variables - requires the transformation of their individual levels into dummy variables. Such a procedure results in the creation of a variable space with as many dimensions as there are levels of all attributes together. Such a situation would not allow for visualizing the performed calculations. That is why, using the influences of particular attributes on the value of the real estate, a single composite variable was created. This variable synthetically describes the attributes of the real estate. It is a stimulant variable, i.e. the higher the value, the higher the total value of the property attributes. Thanks to this, the study concerns two dimensions - the composite variable and the property value. This variable was created on the basis of the value of structural parameters of the mass valuation model, which is described in the paper (Doszyń, 2020).Using this model, the influence of particular states of market characteristics on the property value was determined (Gnat, 2010). A composite variable was calculated as the product of coefficients describing increases in value of the property in relation to the base property, which was the cheapest property in the analyzed area and was characterized by the worst levels of market attributes. Table 1 presents a description of several examples of analyzed properties with the indication of their characteristics and the value of the composite variable.

Table 1

Sample properties from analyzed dataset

No.	area	location	utility access	land plot shape	composite variable
1	small	unfavorable	incomplete	good	8,28
2	average	unfavorable	incomplete	bad	6,60
3	average	unfavorable	incomplete	bad	6,60
4	small	unfavorable	incomplete	bad	7,20
5	average	average	incomplete	good	9,11
6	large	average	incomplete	good	7,45
7	small	favorable	full	good	14,49
8	average	favorable	full	good	13,28
9	average	favorable	full	bad	11,55
10	large	favorable	full	good	7,61

Source: Gnat (2010).

In order to classify the real estate in view of a possible increase or decrease in the tax burden resulting from the tax reform, a percentage rate of the *ad valorem* tax, which will provide revenue to the municipal budget equal to the revenue set for the current property tax, has been applied. The assumption of equal revenues is purely technically motivated. It allowed for a "safe" application of the classification model, as the share of real estate with an increase in the burden (1) is close to the share of real estate with an expected decrease in the burden (0). This results in a situation that does not require any additional calculation procedures. The application of potential tax rates at a different level causes new challenges for the models due to the disproportions between classes, which will be the subject of further research, in which it is planned to test the methods minimizing the effects of class imbalance.

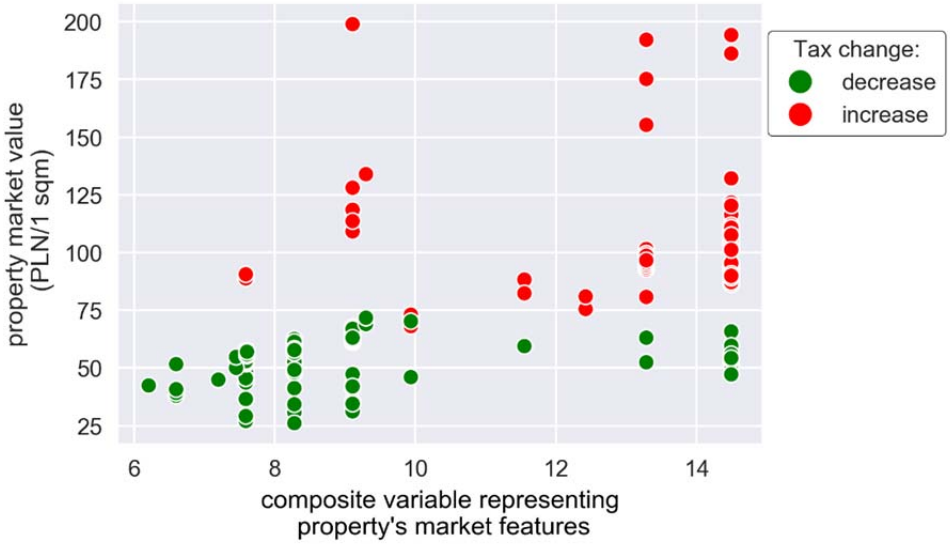

Fig. 1 Expected change of tax burden in relation to both composite variables representing the level of properties' attributes and their value. *Source*: own elaboration.

Figure 1 presents information about the values of individual properties in relation to the composite variable describing their attributes and the change of tax burden (according to the study assumptions). Intuitively, it is possible to indicate a group of properties characterized by a composite variable at the level of about 9 and values exceeding 100 PLN/m², which seems to be unusual. There is also a group of properties with the highest levels of the composite variable and values at the level corresponding to properties with significantly worse composite variable values. Perhaps the properties with the highest levels of the composite variable and the highest values should also be considered atypical. However, the issue of considering individual properties as an outlier, in the context of the relationship between

their attributes and values, should be subject to analytical procedures, which is one of the stages of this study.

Two classes of statistical methods were used in the study. The first one is a tool to classify objects, which in this case were land properties. 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). Logistic regression allows the use of linear regression for classification problems. The logistic regression equation takes the following form:

$$p(X) = \frac{e^{\beta_0 + B_1 X}}{1 + e^{\beta_0 + B_1 X}}, \quad (1)$$

where:
P(X) – probability of occurrence of the analyzed event,
X – explanatory variable,
β_0, β_1 – the structural parameters of the model.

A two-class classification using logistic regression consists, in its simplest form, of comparing the probability obtained from the model with the assumed threshold value of 0.5 according to the following rule: ŷ=1 (there will be an increase in tax burden) if p(X)>0.5 and ŷ=0 (there will be a decrease in tax burden if p(X) ≤0.5). After converting the probability to y=0 and y=1, it is possible to assess the quality of the classification using a confusion matrix or other measure assessing the accuracy of classification carried out. In this study, a confusion matrix and classification accuracy (2) were used to assess the classification. Confusion matrix summarizes the classification performance of a model with respect to some test data. It is a two-dimensional matrix, indexed in one dimension by the true class of an object and, in the other, by the class that the classifier (model) assigns (SAMMUT, WEBB 2017, p. 277). Accuracy defines which part of the predicted class labels is consistent with real results:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}, \quad (2)$$

where:
ACC – classification accuracy,
TP – number of true positive predictions (both the actual and predicted values indicated an increase in tax burden),
TN – number of true negative predictions (both the actual and predicted values indicated a decrease in tax burden),
FP – number of false positive predictions (the prediction indicated an increase in tax burden, but actual tax burden was lower),
FN – number of false negative predictions (the prediction indicated a decrease in tax burden, but actual tax burden was higher).

The second group of statistical tools involves algorithms for detecting unusual objects. There are many ways to detect atypical objects. A strong general division indicates the following groups of methods:

- statistical,
- based on clustering,
- based on the nearest neighbors,
- classification,
- spectral,
- based on sampling,
- deep learning methods.

Statistical methods assume that the data are consistent with a certain distribution. The simplest approach to detecting anomalies in data would be to mark observations that deviate from selected distribution statistics (e.g. Tuckey's fence).

Clustering methods assume that similar points belong to one group. This is determined by the distance from the cluster centroid (e.g. k-means algorithm).

Outlier detection based on the nearest neighborhood usually assumes that normal observations appear in areas that seem dense, while the anomalies are far from the nearest neighbors (e.g. KNN, LOF).

The detection of anomalies based on classifiers consists of two stages. In the learning phase, the classifier is taught on the basis of available training data with a label (supervised learning). Then, test observations are classified as normal or abnormal using the classifier built in the initial stage. The single-class carrier vector machine (OCSVM) and neural network methods constitute examples of such methods for the detection of unusual objects.

The detection of anomalies based on sampling consists of drawing from the analyzed data subsets of observations and features and detecting unusual objects in these subsets. Outlier values are points, which are most often indicated as such in smaller subspaces. Isolation forest (a variant of random forests) is an algorithm that divides multidimensional space into smaller dimensions and tries to find untypical objects in less-dimensional space.

In this study four procedures were used to detect outlier observations:

- k nearest neighbors (KNN) (e.g. Angiulli & Pizzuti, 2002).
- Isolation forest (Liu et al. 2008).
- Local outlier factor (LOF) (Breunig, et al., 2000; He, et al., 2003).
- Principal component analysis (PCA) (Jolliffe, 2002).

4. Empirical results

The study was conducted in several stages. In the first stage, the logistic regression model was estimated on the basis of the entire analyzed set of real estate data. In this model the aforementioned composite variable served as an explanatory variable. The results of property classification were presented by means of a confusion matrix and an accuracy coefficient. The estimated logistic function, along with the indications of correct and incorrect classification for the increase (class 1) and decrease (class 0) of the property tax burden is demonstrated in Figure 2. As can be seen in Figure 3, the total number of plots incorrectly classified in both classes amounted to 22, which translated into a 95.8% accuracy ratio of the classification. This value can be considered very high, but the question arises whether even higher accuracy can be achieved.

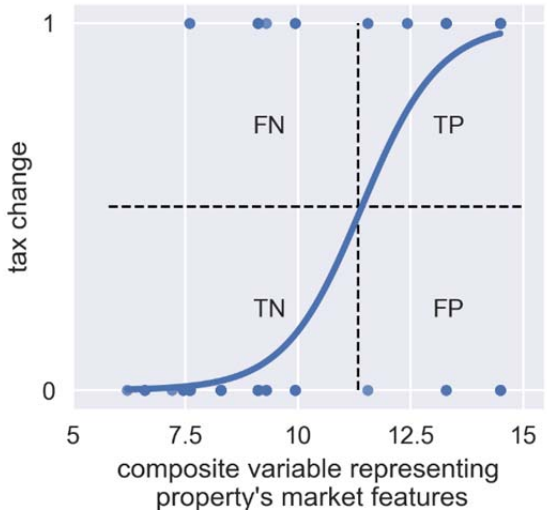

Fig. 2. Logistic regression model based on initial data set. *Source*: own elaboration.

The second stage of the study was conducted in order to answer this question. The previously indicated methods of detecting outliers were used and logistic regression models were once again built. It is worth remembering that different types of statistical calculation methods require the researcher to indicate certain threshold levels on the basis of which the results of these procedures are appropriately evaluated. An example of such arbitrarily accepted threshold levels is the p-value. A similar situation occurs when objects are classified as atypical. In the algorithms used in the study, it should be specified how high of a percentage of observations will be indicated as anomalous. In the software used in this study, i.e. the PyOD package for Python programming language (Zhao, et al.,

2019), the default level is 10%, and the same value is also assumed in the calculations. The data sets on the basis of which the models were estimated no longer contained properties considered atypical. Using mlxtend package for Python programming language (Raschka, 2018), the decision areas indicated by each method were shown in Figure 4. Each of them operates on the basis of different assumptions and, therefore the, green areas, i.e. normal and the red ones, i.e. outliers, differ from each other.

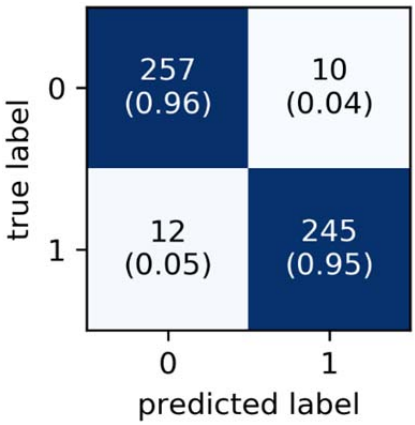

Fig. 3. Confusion matrix for logistic regression model based on initial data set. *Source*: own elaboration.

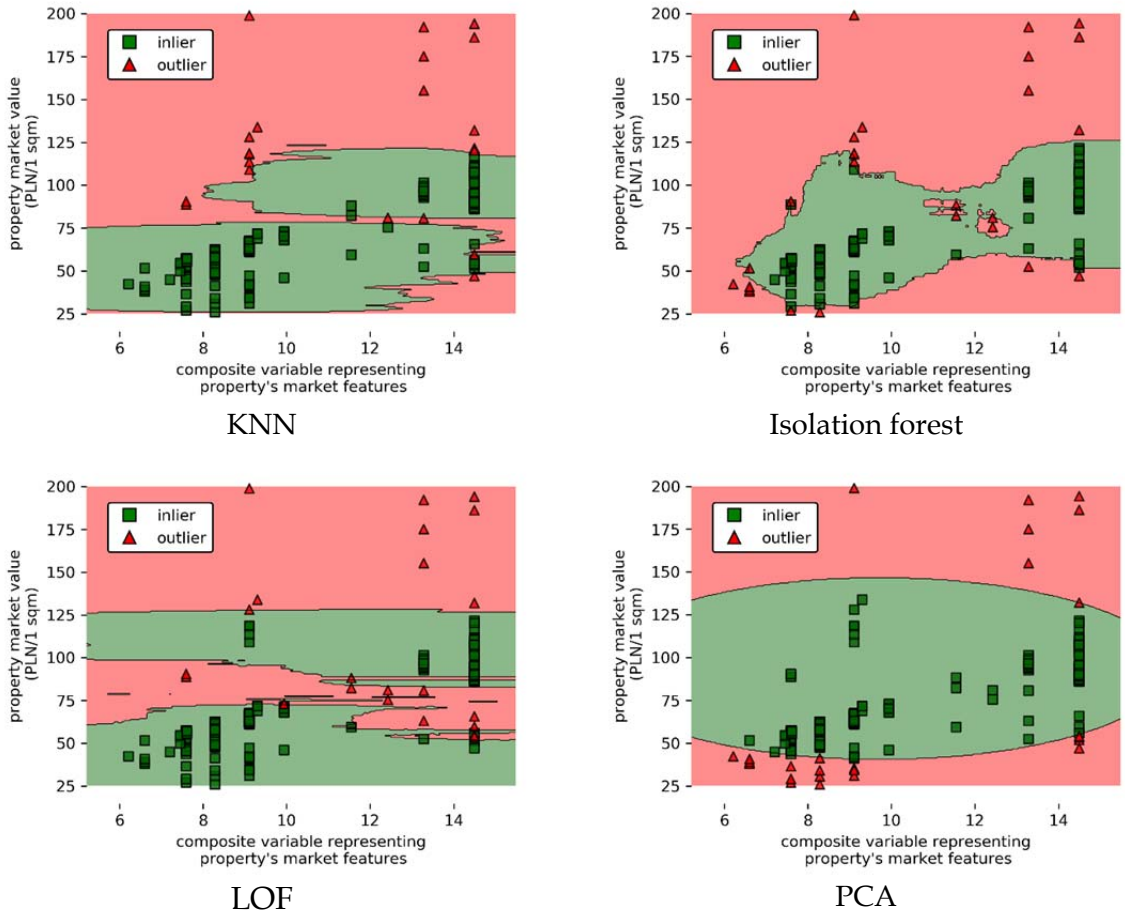

Fig. 4. Decision regions for analyzed outlier detection methods. *Source*: own elaboration.

The results of the classification also differ. The number of incorrectly indicated properties ranged, depending on the method, from 9 to 17, as shown in Figure 5. These numbers are smaller than the 22

properties indicated by the model in which atypical properties were not rejected, but the values cannot be directly compared, due to omission of outliers, data sets used in the second stage were smaller than the initial set. Therefore, the comparison of the models was conducted by means of accuracy coefficients, which are presented in Figure 6.

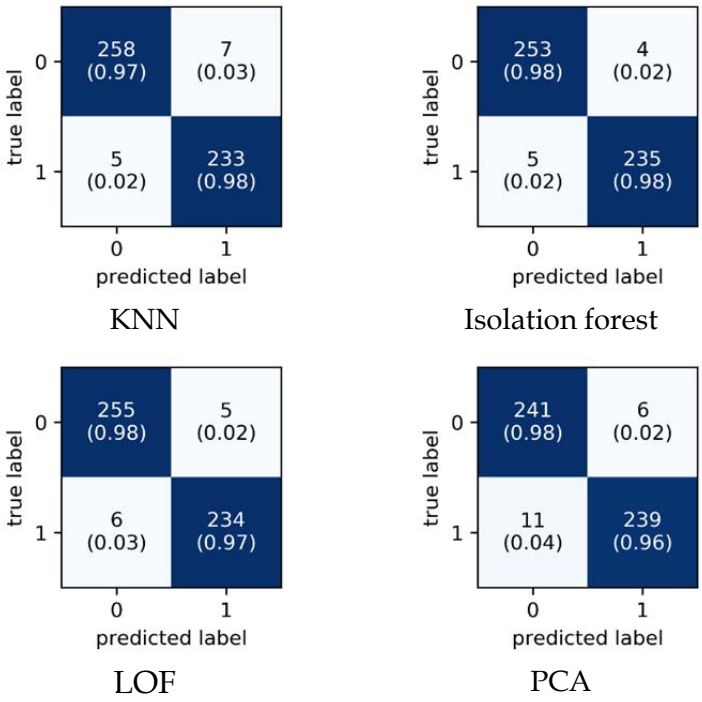

Fig. 5. Confusion matrix for logistic regression models based on data sets without outliers. *Source*: own elaboration.

As can be seen, for each method of detecting outliers, the logistic regression model built without the indicated properties was more accurate. Accuracy turned out to be the highest in the case of the isolation forest method. The smallest advantage over the accuracy for the initial data set occurred in the model based on the set, in which the atypical properties indicated by the PCA method were omitted.

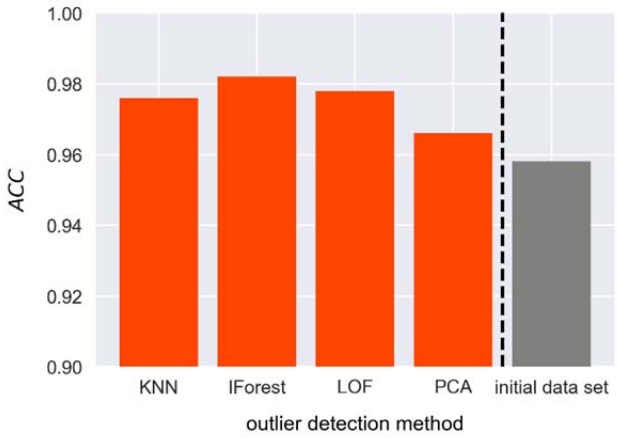

Fig. 6. Comparison of the accuracy of logistic regression models based on initial data set and data sets without outliers indicated with the analyzed methods. *Source*: own elaboration.

Such results should not come as a surprise. The rejection of objects that do not conform to the existing regularity must lead to better modelling results. However, another issue arises here. Pilot or simulation studies can be carried out on samples, not on entire populations. The fact that models without outliers are better in samples does not mean that they will prove more effective when applied to populations. In order to check how the individual models cope with the new data, the next stage of

the study was conducted. In order to avoid results obtained by pure chance, this stage was carried out in the following manner. The initial set of real estate data was randomly divided into a training set (used to build logistic regression models) and a test set, which was used to assess the effectiveness of the model confronting new information. The proportion of these sets was determined at 75/25, where 75% are training observations and 25% are test observations. Such a draw was conducted 1000 times. Therefore, 5000 logistic regression models were estimated. Figure 7 shows the classification accuracy in 1000 repetitions by means of box charts. As can be seen, the classifications without outliers were, on average, more accurate than in the case of models for which outliers were not removed from training sets. Only in the small number of draws the accuracy for models with whole training sets came close to the median accuracy for models in which outliers were removed. This result is basically a repetition of the previous stage. It was confirmed that, for training data, removing outliers is an effective way to improve classification results. And what did the accuracy of the property classification in the test collections look like? The results are shown in Figure 8.

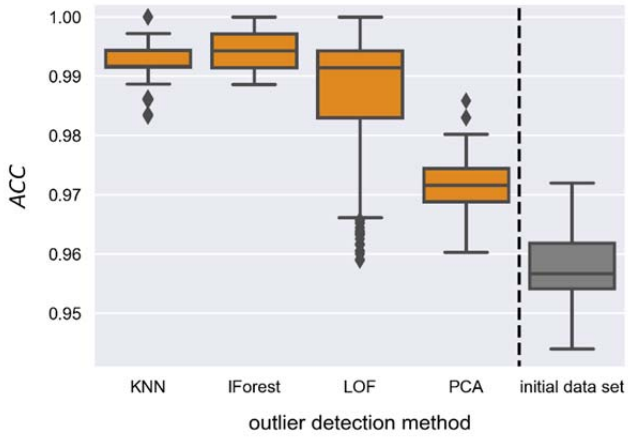

Fig. 7. Boxplots of 1000 training sets accuracy for logistic regression models based on initial data set and data sets without outliers indicated with the analyzed methods. *Source*: own elaboration.

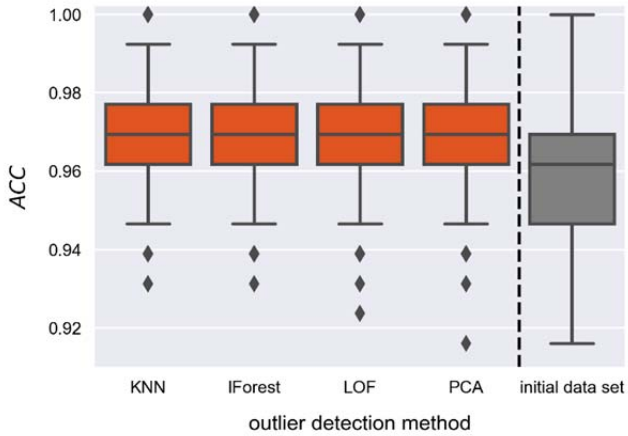

Fig. 8. Boxplots of 1000 test sets accuracy for logistic regression models based on initial data set and data sets without outliers indicated with analyzed methods. *Source*: own elaboration.

Apart from single draws, the accuracy range was narrower and higher in cases where outliers were omitted than in models built with initial data, i.e. without removing the atypical properties. In addition to the statement that models built on the basis of a corrected training data demonstrated better accuracy in the test sets than models built on initial data, something equally important is worth noting. In the test sets, models based on adjusted data obtained, on average, very similar accuracy results to one another. This means that the choice of the method for removing atypical properties is not a determinant of the final results. In many studies the use of different methods leads to different results and conclusions, which makes it difficult to make a clear assessment of the study. In this case, the conclusions are unambiguous. First of all, removing observations from training sets (samples)

increases the accuracy of classification. Secondly, in test sets, this accuracy does not depend on the choice of method.

5. Discussion and conclusions

The survey was conducted focusing on two main issues. The first one was an attempt to classify real estate as a group threatened by an increase in the tax burden or a group for which there may be a chance that such a burden may decrease as a result of a reform of real estate taxation involving the substitution of its area as the basis for calculating the tax based on its value. The variables on the basis of which such a classification was carried out were specified real estate characteristics. As a result of applying the logistic regression model, high classification accuracy coefficients were obtained. In the case of pilot studies on the property tax reform, owing to the cost and time of the studies, information about samples rather than entire populations is often considered. In such cases, an important aspect is to avoid model overfitting. As an effect of this phenomenon, results are fitted too much to random or incorrect data. The model should determine the main regularity in order to be also effective when dealing with new data, i.e. data from outside the training set. Avoiding overfitting to "information noise" can be achieved by removing outlier observations at the pre-construction stage. This constituted the second part of the presented study. Four procedures were used to detect unusual objects (in this case - real estate). After their detection, they were removed from the training sets. The results of the study made it possible to draw two main conclusions. Firstly, removing atypical properties improved the accuracy of classification. Secondly, in order to minimize the randomness of the obtained results, the analyzed real estate data set was divided into training and test sets 1000 times. As a result of this procedure, one can indicate that all applied methods of outlier detection function in a way that allows for a similar improvement of classification accuracy. This is an important statement because it leads to the conclusion that the results obtained do not depend on the method used. To put it briefly, the rejection of abnormal observations, regardless of the selected method, improves the accuracy of classification in test sets in a similar way. The above can be stated as the main scientific contribution of the study. The fact that similar results have been obtained using a variety of methods is a valuable observation. It means that there is little risk of obtaining results depending on the method used, which happens all too often in many studies. What is more, there are much more studies regarding regression problems on real estate market. Classification problems in this area are less frequent and it is therefore also important to add new ideas or case studies in this field.

In the study, the logistic regression model was used as a classifier. In subsequent studies, it is planned to use other classification methods in order to assess which of them allows the highest accuracy to be obtained and which of them is the most robust to the occurrence of outlier properties. It is also planned to conduct a study on the sensitivity of the obtained results to the threshold of outlier objects. Classification methods give the best results with balanced datasets, and this was the assumption made in the initial study. Therefore, the most important objective for future research, in order to increase the practical value of the research, is to adopt a property value tax rate at a level that makes the share of decreased and increased tax burden unbalanced.

6 References

Aggarwal, C. C. (2017). Outlier Analysis (2nd ed.). Springer Publishing Company. https://doi.org/10.1007/978-3-319-47578-3

Angiulli, F., & Pizzuti, C. (2002). Fast Outlier Detection in High Dimensional Spaces, In: Elomaa T., Mannila H., Toivonen H. (eds) Principles of Data Mining and Knowledge Discovery. PKDD 2002. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence), vol. 2431. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-45681-3_2

Baldominos, A., Blanco, I., Moreno, A. J., Iturrarte, R., Bernárdez, Ó., & Afonso, C. (2018). Identifying Real Estate Opportunities Using Machine Learning. *Applied Sciences (Basel, Switzerland), 8*(11), 2321. https://doi.org/10.3390/app8112321

Batóg, B., & Foryś, I. (2011). Logit models in the analysis of transactions on the Warsaw residential market in Polish: Modele logitowe w analizie transakcji na warszawskim rynku mieszkaniowym. *Studia i Materiały Towarzystwa Naukowego Nieruchomości, 19*(3), 33–48.

Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. *SIGMOD Record, 29*(2), 93–104. https://doi.org/10.1145/335191.335388

Cook, R. D. (1977). Detection of Influential Observation in Linear Regression. *Technometrics, 19*(1), 15–18.

Domingues, R., Filippone, M., Michiardi, P., & Zouaoui, J. (2018). A comparative evaluation of outlier detection algorithms: Experiments and analyses. *Pattern Recognition, 74*, 406–421. https://doi.org/10.1016/j.patcog.2017.09.037

Doszyń, M. (Ed.). (2020). Attribute influence matrix calibration system in Szczecin's algorithm of mass property valuation. University of Szczecin.

Doszyń, M., Gnat, S., & Bas, M. (2017). The Econometric Procedures of Specific Transactions Identification. *Folia Oeconomica Stetinensia, 17*(1), 20–30. https://doi.org/10.1515/foli-2017-0002

Etel, L., & Dowgier, R. (2013). Local taxes and charges – time for a change in Polish: Podatki i opłaty lokalne – czas na zmiany, Białystok. *Temida : Casopis o Viktimizaciji, Ljudskim Pravima i Rodu*, 2.

Głuszak, M., & Marona, B. (2015). Cadastral tax. Economic conditions of the property taxation reform in Polish Podatek katastralny. Ekonomiczne uwarunkowania reformy opodatkowania nieruchomości. Poltext.

Gnat, S. (2009). Analysis of the effects of replacing current property tax with ad valorem property tax in a sample municipality. *Folia Oeconomica Stetinensia, 8*(16), 82-98.

Gnat, S. (2010). Use of operational research methods in modelling the impact of cadastral tax on the financial situation of the municipality in Polish: Wykorzystanie metod badań operacyjnych w modelowaniu wpływu podatku katastralnego na sytuację finansową gminy. *Doctoral Dissertation*, Univesrity of Szczecin, Szczecin.

Gnat, S. (2018). Analysis of Communes' Potential Fall in Revenue Following Introduction of ad Valorem Property Tax. *Real Estate Management and Valuation, 26*(1), 63–72. https://doi.org/10.2478/remav-2018-0006

Grubbs, F. E. (1969). Procedures for Detecting Outlying Observations in Samples. *Technometrics, 11*(1), 1–21. https://doi.org/10.1080/00401706.1969.10490657

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning. Springer. https://doi.org/10.1007/978-0-387-84858-7

Hawkins, D. (1980). Identification of Outliers. Chapman and Hall. https://doi.org/10.1007/978-94-015-3994-4

He, Z., Xu, X., & Deng, S. (2003). Discovering cluster-based local outliers. *Pattern Recognition Letters, 24*(9-10), 1641–1650. https://doi.org/10.1016/S0167-8655(03)00003-5

Johnson, R. (1992). Applied Multivariate Statistical Analysis. Prentice Hall.

Jolliffe, I. (2002). Principal Component Analysis (2nd ed.). Springer.

Kontrimas, V., & Verikas, A. (2006). Tracking of doubtful real estate transaction by outlier detection methods: A comparative study. *Information Technology and Control, 35*(2), 94–105.

Liu, F. T., Ting, K. M., & Zhou, Z. 2008, Isolation Forest, Eighth IEEE International Conference on Data Mining, Pisa, 2008, 413-422.

Liu, H., Shah, S., & Jiang, W. (2004). On-line outlier detection and data cleaning. *Computers & Chemical Engineering, 28*(9), 1635–1647. https://doi.org/10.1016/j.compchemeng.2004.01.009

Maimon, O. (ed.), & Rokach, L. (ed.). (2005). Data Mining and Knowledge Discovery Handbook. Springer-Verlag. https://doi.org/10.1007/b107408

Morano, P., De Mare, G., & Tajani, F. (2013). LMS for Outliers Detection in the Analysis of a Real Estate Segment of Bari. In B. Murgante, . . . (Eds.), Lecture Notes in Computer Science: *Vol. 7974*. Computational Science and Its Applications – ICCSA 2013. ICCSA 2013. Springer. https://doi.org/10.1007/978-3-642-39649-6_33

Ng, K. H., & Khor, K. (2016). Rapid identification of outstanding real estate investment trusts with outlier detection algorithms. *Journal of Theoretical and Applied Information Technology, 88*(2), 321–330.

Pimentel, M. A. F., Clifton, D. A., Clifton, L., & Tarassenko, L. (2014). A review of novelty detection. *Signal Processing, 99*, 215–249. https://doi.org/10.1016/j.sigpro.2013.12.026

Prastawa, M., Bullitt, E., Ho, S., & Gerig, G. (2004). A brain tumor segmentation framework based on outlier detection. *Medical Image Analysis, 8*(3), 275–283. https://doi.org/10.1016/j.media.2004.06.007 PMID:15450222

Raschka, S. (2018). MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. *Journal of Open Source Software, 3*(24), 638. https://doi.org/10.21105/joss.00638

Sammut, C. (ed.), & Webb, G. I. (ed.). (2017). Encyclopedia of Machine Learning and Data Mining (2nd ed.). Springer Publishing Company. https://doi.org/10.1007/978-1-4899-7687-1

Stefansky, W. (1972). Rejecting Outliers in Factorial Designs. *Technometrics, 14*(2), 469–479. https://doi.org/10.1080/00401706.1972.10488930

Śpiewak, B. (2018). Application of Chosen Methods of Robust Estimation: Baarda's and Huber's in Search for Outliers in the Real Estate Market Modeling. *Folia Oeconomica Stetinensia, 18*(1), 27–38. https://doi.org/10.2478/foli-2018-0003

Trojanek, M., & Kisiała, W. (2016). The Diversification of Communes' Revenue from Real Estate Across Provinces. *Real Estate Management and Valuation, 24*(2), 36–49. https://doi.org/10.1515/remav-2016-0012

Wójtowicz, K. (2006). Analysis of potential effects of real estate tax system reform in Poland in Polish: Analiza potencjalnych skutków reformy systemu opodatkowania nieruchomości w Polsce. Finanse Publiczne. UMCS.

Worden, K., Manson, G., & Fieller, N. R. J. (2000). Damage detection using outlier analysis. *Journal of Sound and Vibration, 229*(3), 647–667. https://doi.org/10.1006/jsvi.1999.2514

Zhao, Y., Nasrullah, Z., & Li, Z. (2019). PyOD: A Python Toolbox for Scalable Outlier Detection. *Journal of Machine Learning Research, 20*(96), 1–7.

Zhu, C., Kitagawa, H., Papadimitriou, S., & Faloutsos, C. (2011). Outlier detection by example. *Journal of Intelligent Information Systems, 36*, 217–247. https://doi.org/10.1007/s10844-010-0128-1

Zyga, J. (2016). Connection Between Similarity and Estimation Results of Property Values Obtained by Statistical Methods. *Real Estate Management and Valuation, 24*(3), 5–15. https://doi.org/10.1515/remav-2016-0017