

ORIGINAL ARTICLE


Citation: Gnat, S. (2019). Spatial weight matrix impact on real estate hierarchical clustering in the process of mass valuation. *Oeconomia Copernicana*, 10(1), 131–151. doi: 10.24136/oc.2019.007

Contact: sebastian.gnat@usz.edu.pl; University of Szczecin, Faculty of Economics and Management, Institute of Econometrics and Statistics, ul. Mickiewicza 64, 71-101 Szczecin, Poland

Received: 28.12.2018; Revised: 2.02.2019; Accepted: 16.02.2019; Published online: 26.02.2019

Sebastian Gnat

University of Szczecin, Poland

 orcid.org/0000-0003-0310-4254

Spatial weight matrix impact on real estate hierarchical clustering in the process of mass valuation

JEL Classification: C38; R30

Keywords: *agglomerative clustering; entropy; property mass appraisal, market analysis*

Abstract

Research background: The value of the property can be determined on an individual or mass basis. There are a number of situations in which uniform and relatively fast results obtained by means of mass valuation undoubtedly outweigh the advantages of the individual approach. In literature and practice there are a number of different types of models of mass valuation of real estate. For some of them it is postulated or required to group the valued properties into homogeneous subset due to various criteria. One such model is Szczecin Algorithm of Real Estate Mass Appraisal (SAREMA). When using this algorithm, the area to be valued should be divided into the so-called location attractiveness areas (LAZ). Such division can be made in many ways. Regardless of the method of clustering, its result should be assessed, depending on the degree of implementation of the adopted criterion of division quality. A better division of real estate will translate into more accurate valuation results.

Purpose of the article: The aim of the article is to present an application of hierarchical clustering with a spatial constraints algorithm for the creation of LAZ. This method requires the specification of spatial weight matrix to carry out the clustering process. Due to the fact that such a matrix can be specified in a number of ways, the impact of the proposed types of matrices on the clustering process will be described. A modified measure of information entropy will be used to assess the clustering results.

Methods: The article utilises the algorithm of agglomerative clustering, which takes into account spatial constraints, which is particularly important in the context of real estate valuation. Homogeneity of clusters will be determined with the means of information entropy.

Findings & Value added: The main achievements of the study will be to assess whether the type of the distance matrix has a significant impact on the clustering of properties under valuation.

Introduction

There are two main approaches in real estate valuation: individual valuation and mass valuation. In an individual valuation, a valuer focuses on one or a small number of properties. In the case of mass valuation, the subject of valuation is a large number of properties of one type, which are appraised with a uniform approach yielding consistent results.

In practical implementations, and to a greater extent in the research field of mass valuation of real estate, many models and algorithms can be distinguished. Acceptance of the adopted model of mass valuation of real estate should be based on the reliability of the results, in particular in order to prevent complaints of the parties involved regarding the way the valuation has been carried out and the obtained results. One of the basic elements of many mass valuation models is the division of the valued area into sub-areas, which in the case of valuations for tax purposes are called tax zones. The specification of these zones constitutes one of the key problems from the point of view of the correctness of the obtained valuation results. A tax zone is an area in which a given number of appraisable properties has the same impact of location on their value. In other words, all properties located in a given tax zone at a certain level of generality don't differ from one another in terms of location influence on their value. In the case of methods that are based on the valuation of a sample of real estate in a given zone, the concept of representative real estate shall be introduced. This representative real estate is selected by taking into account the characteristic features of a given type of real estate in a given zone. The properties in a given zone should be similar to each other, thus the differences between them, due to the characteristics that describe them, should be as small as possible. The aim is to obtain a situation in which representative properties reflect as much as possible the collection of properties from which they originate, in order to allow the extrapolation of the value of the representatives to the entire zone with the highest possible degree of accuracy. The quality of subdivision of real estate into subzones is of great economic importance. If there are properties in a given zone exerting different influence of features on the value, the obtained valuation results will be inaccurate. In the case of valuations conducted for tax purposes, this will lead to a tax being charged on the under- or overestimated value. In the first case, this will reduce tax revenue. In the latter case, it will result in objections from the taxpayers. None of these cases should be accepted. Hence, the issue of

proper specification of valuation zones is an important economic and computational matter.

In the literature one can find many proposals for creating tax zones. One of the ways is to use statistical methods, namely various methods of objects clustering. Among the classic methods one can distinguish the k -means or hierarchical clustering. The latter method makes it possible to take into account not only the characteristics of objects, but also their spatial relations. These relationships take the form of spatial weights, which can be defined in at least several ways. This diversity in the creation of spatial weight matrices creates a research problem, which is the aim of this study. A question arises whether the use of different types of spatial weight matrices significantly changes the results of real estate clustering. The article will present the results of real estate clustering, taking into account different spatial weights. It will be assessed whether the results of the clusterings are significantly different from one another. Designation of valuation sub-areas (hereinafter referred to as Location Attractiveness Zones *LAZ*) is connected with the assessment of similarity of properties located within their boundaries.

The article will present an approach in which the measurement of entropy (e.g. Truffet, 2018) will be employed to determine the diversity of real estate in each *LAZ*. A modification of the classic entropy measure will be proposed, which will allow for a better reflection of the specificity of the real estate market. The geographical area of the survey encompasses the northern part of Szczecin (the largest city in north-western Poland). More than 1 600 plots of land are valued. As a part of the valuation process *LAZ* are created.

The literature review includes topics related to the existing methods of mass valuation of real estate and the specification of sub-areas of valuation. References to the use of different clustering methods in the real estate market will also be presented. The next section will discuss the utilised research methods. After the presentation of the results obtained, the conclusions of the study will be presented.

Literature review

There are many models, algorithms and procedures in the area of mass property valuation (see Jahanshiri *et al.*, 2011; Kauko & d'Amato, 2008). Furthermore, the work of Pagourtzi *et al.* (2003) provides good insight of possible approaches in property valuation. Current studies regarding mass valuation and automated valuation models (AVM's) present application of

machine learning algorithms in this field (e.g. Zurada *et al.*, 2011). Some studies present remarks regarding implementation of mass valuation (Grover, 2016). It is emphasised that quality of data and other requirements are sometimes omitted. Particular group of the existing models and algorithms, in order to obtain more precise results, are based on the division of the valued area into possibly homogeneous subzones. The issue of their proper designation should be considered as one of the key problems from the standpoint of the correctness of the obtained valuation results. This correctness is a major issue, since high valuation errors cause problems of economic nature. They can lead to a flawed decision of any party involved in the valuation process. The introduction of real estate submarkets is widely discussed in literature (e.g. Palm, 1978; Bourassa *et al.*, 1999; Keskin & Watkins, 2016). Submarkets are identified for analytical purposes, real estate valuation or for tax purposes. The approaches used to determine tax zones indicate the use of techniques based on zoning plans, as well as the use of aerial and satellite imagery (e.g. Dąbrowski & Latos, 2015). The borders of districts and housing estates, plot lines, streets, roads, rivers, railway routes and other artificial and natural objects can be used for this purpose as well (Dedkova & Polyakova, 2018). There are also postulates to take into account not only the functional features of the areas, but also their so-called physiognomic or even legal features. A separate group of methods that can be used to create tax zones are clustering methods (e.g. Hastie *et al.*, 2009). One of the available approaches is agglomeration clustering (e.g. Kantardzic, 2003). This method makes it possible to group similar objects (in this case real estate) on the basis of many features describing them. Clustering is a very common concept in scientific research. The issue of determining an optimal number of clusters is discussed (Kolesnikov *et al.*, 2015; Fang & Wang, 2012). Researchers also present proposals for the improvement of already known methods (see Arguelles *et al.*, 2014) or they propose the use of an ensemble clustering method (Wu *et al.*, 2018; Boon-Goen & Iam-On, 2018). From the point of view of the real estate market, the possibility of introducing spatial constraints into the algorithm constitutes a particularly important element of this way of clustering (Guo, 2008; Davidson & Ravi, 2005), which allows one to take into account the adjacency of objects. The dominant form of taking into account spatial relations between objects entails spatial weights matrices. One of the basic functions of spatial weights matrices is the identification of spatial effects. These matrices can be created in a number of ways. One of the divisions of the spatial matrices can be found in the paper by Getis and Aldstadt (2004, pp. 147–163). This division includes, among others, the following:

- neighbourhood matrices,
- k -nearest neighbours' matrices,
- reverse distance matrices,
- reverse distance matrices limited to k -nearest neighbours,
- matrices in which i and j are neighbours if the distance between them is less than or equal to a predetermined value.

This flexibility is one of the main objections to spatial analysis (LeSage & Pace, 2014), because even within a single study it is possible to formulate more than one matrix of spatial weights. The issue of the impact of the applied matrix of spatial weights was discussed mainly within the studies related to the estimation of spatial regression models (Cellmer, 2013; Zhang & Yu, 2018). Algebraic properties related to distance matrices are also studied (Bapat, 2006). Furthermore, the influence of the distance matrix on clustering results was analysed (Mimmack *et al.*, 2000). The authors point out the fact that the results of the clustering are sensitive to the distance matrix used in the process. This sensitivity of the choice of spatial weight matrix has been given the main attention in this study.

From the point of view of mass valuation, it is important that clusters should contain real estate similar to each other. A modified measure of entropy was used to assess the homogeneity of location attractiveness zones. In the mid-20th century, an American mathematician, C. E. Shannon, laid the foundations of the theory of information transmission basing it on the concept of entropy. Entropy of the distribution of the analysed variable allows specifying the degree of determination (definiteness) of this distribution on account of the analysed variable (e.g. Raschka & Mirjalili, 2017, p. 90). Shannon's entropy coefficient is standardised and it reaches the values within the range of (0,1). A high value of the coefficient indicates a high indefiniteness of the tested system. A low value of the entropy measure indicates a significant determination of the system (the system demonstrates an inclination). In the publications related to spatial issues in their broadest sense, entropy explores, for example, land use changes (Bai & Wang, 2012), ecosystem development (Ludovisi, 2014), or it evaluates geological models (Wellman & Regenauer-Lieb, 2012). In this study, entropy will also be used for evaluation. In this case, the evaluation of clustering of properties.

Research methodology

As mentioned above, there are a number of mass property valuation methods. One example of such methods is Szczecin Algorithm of Real Estate

Mass Appraisal (*SAREMA*). One of the stages of the algorithm is the specification of elementary areas (location attractiveness zones) and it will be this algorithm that will be used to determine the value of the properties, whereas the proposal to modify the entropy measure will be used to assess the homogeneity of the property in designated areas. Although the present study uses as an example one of the available solutions in the field of mass valuation of real estate, it should be noted that in the literature there are more models requiring a subdivision of the valued area. The procedure analyzed in the study does not refer only to the described algorithm of valuation. Hierarchical clustering is a valid procedure for all procedures with this kind of requirement. Szczecin algorithm of mass valuation of real estate assumes the following form (based on Hozer *et al.*, 2002):

$$w_{ji} = wwr_j \cdot pow_i \cdot c_{baz} \prod_{k=1}^K \prod_{p=1}^{k_p} (1 + a_{kp}) \quad (1)$$

where:

w_{ji} – market (or cadastral) value i -th real property in j -th LAZ,

wwr_j – market value coefficient in j -th LAZ ($j = 1, 2, \dots, J$),

pow_i – area of i -th real property,

c_{baz} – price of 1m^2 of the cheapest land (without the utility infrastructure) in the appraised area,

a_{kp} – influence of p -th category of k -th attribute ($k = 1, 2, \dots, K; p = 1, 2, \dots, k_p$),

K – number of attributes,

k_p – number of categories of k -th attribute.

Algorithm (1) has a multiplicative form. The point of reference for appraisal with the use of the algorithm is provided by the base price. It is the price of 1m^2 of the cheapest land without the utility infrastructure in the appraised municipality. It may be assumed that it is a unitary price of a real property of the worst categories of attributes, which include the impact of attributes of the appraised real properties. The impact (a_{kp}) may be defined with an expert approach, by property valuers.

For each LAZ, coefficients of market value (wwr_j) are determined, which reflect the impact of a widely understood location.

The market value coefficient for j -th LAZ is a geometric mean of the quotients of real properties actual values and their hypothetical values:

$$WWR_j = \sqrt[n_j]{\prod_{i=1}^{n_j} \frac{w_{ji}^{rZ}}{w_{ji}^h}} \quad (2)$$

where:

w_{ji}^{rZ} – value of i -th real property in j -th LAZ defined by a real property appraiser,

w_{ji}^h – hypothetical value of i -th real property in j -th LAZ,

n_j – the number of representative real properties valued by real property appraisers in j -th LAZ.

Hypothetical values (w_{ji}^h) are calculated on the basis of formula (1), omitting the market value coefficients:

$$w_{ji}^h = p \cdot o w_i \cdot c_{baz} \prod_{k=1}^K \prod_{p=1}^{k_p} (1 + a_{kp}) \quad (3)$$

If the values of the drawn representative real properties (w_{ji}^{rZ}), the categories of attributes and their impact are known, the base price (c_{baz}) and areas are known, then for each LAZ the market value coefficients may be evaluated as a geometric mean from the quotients of actual and hypothetical real estate values.

As mentioned above, *SAREMA* requires the valued area to be divided into homogeneous sub-areas within which representative properties are drawn and individually assessed. This procedure of mass valuation of real estate has already been used several times in practice and the experience from these applications shows that the appropriate representativeness of the location expressed in the algorithm by the *WWR* ensures a random choice of:

- one representative property from LAZs, where up to 10 properties are subject to valuation,
- two representative properties from LAZs, where 11 to 50 properties are subject to valuation,
- three representative properties from LAZs, where 51 to 100 properties are subject to valuation,
- four representative properties from LAZs, where 101 to 500 properties are subject to valuation,

- five representative properties from *LAZs*, where 501 to 1000 properties are subject to valuation,
- six representative properties from *LAZs*, where more than 1,000 properties are valued.

The study will be carried out in the following way:

1. Clustering – using agglomeration clustering with spatial constraints, real estate will be classified into sub-areas (*LAZ*). Two types of spatial weight matrices were used as constraints. The first one is based on *k*-nearest neighbours (*KNN*) and the second one on the distance band (*DB*). In both types of matrices several variants were used. For the *KNN* matrix, values 3, 5, 10, 20, 50 were taken as *k*. In the case of *DB* matrix, it was 50m, 75m, 100m, 250m and 400 meters. Examples of neighbourhood graphs based on two *KNN* matrices and two *DB* matrices are shown in Figures 1 and 2. It can be seen that taking a larger *k* and a larger distance band (*r*) increases the number of neighbours for particular objects. The influence of a larger number of neighbours on the outcome of the clustering is the primary objective of the survey. Thus, 10 different matrices of spatial weights were obtained. For each of these matrices a clustering process was carried out with different number of clusters (*LAZ*). In the study, the number of clusters ranged from 5 to 60. This gave a total of 560 clusterings.
2. For each of these clusterings, on the basis of the size of each *LAZ*, the required number of representative properties and the average entropy (\bar{H}_2) were determined. The parameters of each clustering obtained in this way will be compared both with the reference division and with one another. Existing cadastral districts have been used as the reference clustering.
3. The conclusions will be determined on the basis of whether agglomeration clusterings have different entropy depending on the adopted spatial weight matrix, and whether the procedure used in the study produces better results than the benchmark (cadastral districts).

The subject of mass valuation, for which the procedure of clustering to homogeneous *LAZ* is carried out, includes a collection of 1 630 plots of land located in Szczecin — the capital of the West Pomeranian Voivodeship (one of the 17 Polish Voivodeships). These plots are a collection for which annual fees for perpetual usufruct of land were updated. The important point is that they did not constitute a single, coherent area. They were located randomly in the entire north part of the city. These plots of land are described by means of several attributes:

- area, in the following variants: large, medium, small,
- utility access, in variants: none, incomplete, full,

- surroundings, in the following variants: onerous, unfavourable, average, favourable,
- communication accessibility, in the following variants: unfavourable, average, good,
- shape of the plot, in the following variants: unfavourable, average, favourable.

The clustering procedure was carried out on the basis of characteristics directly related to the location of the property: utility access, surroundings and transportation accessibility. These features, being on the ordinal scale, were transformed into dummy variables. The use of variables on the ordinal scale is not coincidental. Such a way of describing the features determining the value of real estate is typical in individual valuations of real estate in Polish conditions. The algorithm used in the study, one of the assumptions of which is mimicking expert's behavior, is also based on the variables on this measurement scale.

In order to assess the homogeneity of the obtained *LAZ*, an entropy measure was used. However, this classical measure was modified due to the different number of possible types of plots (classes of plots). Transformed information on variants of the three above-mentioned plots' characteristics was used to assess entropy. The variants of these characteristics were encoded in the form of natural numbers (the worst state 0, intermediate state 1, the best state 2 and in the case of transportation accessibility, which was a feature of four states — 3) and they were combined into a three-digit codes. Each code value is a combination of variants of characteristics, which was understood as a class. The entropy of the *LAZ* was calculated using the following measure:

$$H_z = \frac{(-\sum_{i=1}^k p_i \cdot \log_k p_i) \cdot k}{L}, \quad (4)$$

where:

p_i – share of real estate belonging to the i -th class,

k – number of classes (combination of variants of market characteristics present in a given *LAZ*),

L – number of classes present in the analysed set.

A modification of the classical entropy measure involves changing the assessment depending on the number of classes present in particular areas. For example, in the classic approach a *LAZ* with two or ten classes and an even share of these classes will be characterised by total entropy. However, in the case of a real estate market analysis, these two situations should be

assessed differently. Two classes of plots' characteristics, when the specified area contains several dozen or more properties, should be described as highly homogeneous, even if the shares of both classes are 50%. The H_z measure will distinguish between the level of entropy depending both on the shares of classes and on their number.

The calculations were carried out in the Python programming environment (i.e. Müller & Guido, 2016; Unpingco, 2016), using mainly Numpy, Pandas, Geopandas, PySAL, SciPy and scikit-learn packages.

Results

The plots of land subject to analysis were qualified into location attractiveness zones (*LAZ*) by means of agglomeration clustering with spatial constraints. The number of proposed *LAZ* ranges between 5 and 60. Each of the divisions was carried out taking into account 10 different spatial weights matrices. Figure 3 presents a fragment of the research area with an assignment of the valued plots of land to two *LAZ*. There is a clear demarcation between the two zones. With the increase in the number of *LAZ*, their average entropy was decreasing. Such a result was to be expected, because with a larger number of sub-areas they contain fewer dissimilar plots. A more important observation is that, depending on the adopted spatial weight matrix, the average entropy for the same number of *LAZ* is different. This effect is similar for both types of utilised matrices (*KNN* and *DB*), as shown in Figures 4 and 7. Initially, with the increase of the number of *LAZ*, the decrease in average entropy is more rapid. After exceeding 20–30 *LAZ*, this decrease is much slower. In the case of *KNN* matrix, the lowest mean entropy in 55 out of 56 cases was obtained for $k = 50$. In turn, the highest mean entropy was recorded for $k = 3$ and 5. In total, it was 50 cases out of 56. In the case of *DB* matrix, the lowest mean entropy was obtained in 50 out of 56 cases for $r = 400$. The highest mean entropy occurred at the shortest distance range of 50 meters. This was the case 28 times, so for the half of the clusterings. In addition to the analysis of the mean entropy for each clustering, the study also drew attention to the dispersion of entropy. Figures 5 and 8 show the standard deviations of entropy for all analysed divisions and types of spatial weight matrices. $S(H_z)$ values differed significantly depending on the variant of the matrix and the number of *LAZ*. The greatest dispersion of *LAZ* entropy was recorded for their smallest numbers. For *KNN* matrix the highest values of standard deviations of entropy were for $k = 10$. The lowest variability was observed, similarly to the lowest mean entropy, for $k = 50$. In case of the second type of analysed spatial

weights matrices, the changes in standard deviations of entropy were more heterogeneous. Standard deviations of entropy in the case of r of 50 and 75 meters decreased at first, then increased and then decreased again. For other distance bands there was also no clear downward trend for the dispersion of entropy.

The final stage of the study was to compare the results of the clustering of land plots into *LAZ* with the reference division. Two criteria were taken into account: average entropy and the required number of representative plots. On the basis of the number of plots in each of *LAZ*, the required number of representatives was determined in accordance with the guidelines set out earlier. It was necessary to assess whether the agglomeration clustering allowed the achievement of a clustering of land plots with a lower entropy than the reference division and whether it allowed to obtain a smaller number of representative plots. The latter criterion has a particular economic rationale. Each representative property must be valued individually, which involves the use of time and money. Figures 6 and 9 show the average entropy and the required number of representative plots in each clustering and for both types of spatial weights matrices. The observed regularity indicates that the smaller the average entropy is (i.e. the greater number of *LAZ*), the more representative properties should be drawn from a given zone. This regularity (which should have been expected) occurred in all 10 cases analysed. However, there were quite significant differences as to how many of the 56 clusterings carried out for a given spatial weight matrix type meet both the required criteria. Tables 1 and 2 present selected results of clustering evaluation. In the case of *KNN* matrices, the number of clusterings, for which the average entropy was lower than the average entropy of the reference division and for which the number of representatives was also lower than in the case of cadastral districts, was highly diversified. For $k = 5$ out of 56 divisions only 5 met both criteria. Whereas for $k = 50$ it was 26. This means that, at best, less than half of the clusterings turned out to be more favourable than the reference breakdown. The valued plots of land are located in the area of 39 cadastral districts. For each number of neighbours used in the study, the smallest number of *LAZ* whose average entropy was lower than the average entropy of cadastral districts was lower than 39 and ranged from 17 to 34. For such numbers of *LAZ*, the number of representative plots was also lower than the number of representatives for cadastral division and ranged from 50 to 77 depending on k , compared to 84 representatives established for the reference division. On the other hand, the results obtained for the matrix based on the distance band indicate that the differentiation of the number of clusterings meeting the criteria of entropy and the number of representative plots was much smaller. The lowest

number of clusterings meeting both criteria is 22 (for $r = 100$ m) and the highest is 38 (for $r = 50$ m). What is more, the smallest number of *LAZ* with an average entropy below the reference one was less differentiated for the *DB* matrix and ranged from 14 to 16, i.e. much less than for the *KNN* matrix. Which means that with just 14 to 16 *LAZ* one can obtain average entropy equal to the benchmark. As a result, this translated into a smaller (and less diversified, depending on r) number of required representative plots, whose number ranged from 41 to 48.

Conclusions

The article presents the results of the study, the aim of which was to determine what influence the applied spatial weights matrix exerts on the clustering. The study covered over 1600 plots of land, which were subject to mass valuation with the use of the Szczecin Algorithm of Mass Property Valuation. The effect of different matrices was assessed by changes in the average entropy of *LAZ* for a given clustering and by the required number of representative plots. The obtained results indicate that the decision regarding the applied matrix has a large impact on the level of entropy of location attractiveness zones and the number of plots to be valued (according to the *SAREMA* assumptions) in the individual approach. The results of agglomeration clustering were compared with entropy and a fixed number of representatives for the benchmark division. Depending on the number of *LAZ*, their average entropy was either lower or higher than the reference clustering. There were also significant differences in the number of clusterings that met the thresholds of a smaller number of representative plots than the reference division, as well as smaller than the reference average entropy. These conclusions were analogous for both types of the distance matrix. From both types of spatial weights matrices better results were obtained for the matrix based on the distance band. For the divisions carried out with the use of this type of matrix, lower than the reference average entropy and a smaller number of representative plots were more often obtained. The results for the different distance bands obtained with the *DB* matrix differed from one another significantly less than those for the *KNN* matrix. This means that clustering with spatial constraints in the form of a *DB* matrix is less sensitive to input parameters. The above results lead to a conclusion that the stage of selection of spatial weight matrix is an important element of the described mass valuation procedure. Before making a final choice, a preliminary assessment of the clusterings should be carried out with various matrices in order to obtain more precise valuations.

Further study will examine various available ways of taking into account the spatial constraints and other methods of clustering. Special attention will be paid to expert methods of creating sub-zones of valuation subjects. Methods that will not depend on arbitrarily selected distance matrices. The application of different approaches will enable, at subsequent stages of research, a comparison of the results of mass valuation and a verification of the impact of the clustering of valued properties on the accuracy of valuations.

References

- Arguelles, M., Benavides, C., & Fernandez, I. (2014). A new approach to the identification of regional clusters: hierarchical clustering on principal components. *Applied Economics*, 46(21). doi: 10.1080/00036846.2014.904491.
- Bai, Y. P., & Wang, B. H. (2012). Study on regional land use structure change characteristics in Baolan-Lanqing-Qingzang urban belt based on information entropy and regional entropy. *Advanced Materials Research*, 518-523. doi: 10.4028/www.scientific.net/AMR.518-523.6024.
- Bapat, R. B. (2006). Determinant of the distance matrix of a tree with matrix weights. *Linear Algebra and its Applications*, 416.
- Boongoen, T., & Iam-On, N. (2018). Cluster ensembles: a survey of approaches with recent extensions and applications. *Computer Science Review*, 28. doi: 10.1016/j.cosrev.2018.01.003.
- Bourassa, S. C., Hamelink, F., Hoesli, M., & Macgregor, B. D. (1999). Defining housing submarkets. *Journal of Housing Economics*, 8(2). doi: 10.1006/jhhec.1999.0246.
- Cellmer, R. (2013). Use of spatial autocorrelation to build regression models of transaction prices. *Real Estate Management and Valuation*, 21(4). doi: 10.2478/remav-2013-0038.
- Dąbrowski, R., & Latos, D. (2015). Possibilities of practical application of the remote sensing data in the real property appraisal. *Real Estate Management and Valuation*, 23(2). doi: 10.1515/remav-2015-0016.
- Davidson, I., & Ravi, S.S. (2005). Agglomerative hierarchical clustering with constraints: theoretical and empirical results. In A. M. Jorge, L. Torgo, P. Brazdil, R. Camacho, & J. Gama (Eds.). *Knowledge discovery in databases: PKDD 2005. PKDD 2005. Lecture notes in computer science. vol 3721*. Berlin, Heidelberg: Springer. doi:10.1007/11564126_1.
- Dedkova, O., & Polyakova, I. (2018). Development of mass valuation in Republic of Belarus. *Geomatics And Environmental Engineering*, 12(3). doi: 10.7494/geom.2018.12.3.29.
- Fang, Y. X., & Wang, Y. H. (2012). Selection of the number of clusters via the bootstrap method. *Computational Statistics & Data Analysis*, 56. doi: 10.1016/j.csda.2011.09.003.

- Getis, A., & Aldstadt, J. (2004). Constructing the spatial weights matrix using a local statistic. *Geographical Analysis*, 36(2). doi: 10.1111/j.1538-4632.2004.tb01127.x.
- Grover, R. (2016). Mass valuations. *Journal of Property Investment & Finance*, 34(2). doi: 10.1108/JPIF-01-2016-0001.
- Guo, G. (2008). Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP). *International Journal of Geographical Information Science*, 22(7). doi: 10.1080/13658810701674970.
- Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*. New York: Springer.
- Hozer, J., Kokot, S., & Kuźmiński, W. (2002). *Methods of statistical analysis of the market in real estate appraisal*. Warsaw: PFSRM.
- Jahanshiri, E., Buyong, T., & Shariff, A. R. M. (2011). A review of property mass valuation models. *Pertanika Journal of Science & Technology*, 19.
- Kantardzic, M. (2003). *Data mining. Concepts, models, methods, and algorithms*. Wiley-IEEE Press.
- Kauko, T., & d'Amato, M. (Eds.) (2008). *Mass appraisal methods. An international perspective for property valuers*. Wiley-Blackwell.
- Keskin, B., & Watkins, C. (2016). Defining spatial housing submarkets: exploring the case for expert delineated boundaries. *Urban Studies*, 54(6). doi: 10.1177/0042098015620351.
- Kolesnikov, A., Trichina, E., & Kauranne, T. (2015). Estimating the number of clusters in a numerical data set via quantization error modeling. *Pattern Recognition*, 48(3). doi: 10.1016/j.patcog.2014.09.017.
- LeSage, J. P., & Pace, R. K. (2014). The biggest myth in spatial econometrics. *Econometrics*, 2. doi: 10.3390/econometrics2040217.
- Ludovisi, A. (2014). Effectiveness of entropy-based functions in the analysis of ecosystem state and development. *Ecological Indicators*, 36. doi: 10.1016/j.ecolind.2013.09.020.
- Mimmack, G. M., Mason, S. J., & Galpin, J. S. (2000). Choice of distance matrices in cluster analysis: defining regions. *Journal of Climate*, 14. doi: 10.1175/1520-0442(2001)014<2790:CODMIC>2.0.CO;2.
- Müller, A. C., & Guido, S. (2016). *Introduction to machine learning with python*. Sebastopol: O'Reilly.
- Pagourtzi, E., Assimakopoulos, V., Hatzichristos, T., & French, N. (2003). Real estate appraisal: a review of valuation methods. *Journal of Property Investment & Finance*, 21(4). doi: 10.1108/14635780310483656.
- Palm, R. (1978). Spatial segmentation of the urban housing market. *Economic Geography*, 54(3).
- Raschka, S., & Mirjalili, V. (2017). *Python machine learning*. Birmingham-Mumbai: Packt Publishing.
- Truffet, L. (2018). Shannon entropy reinterpreted. *Reports on Mathematical Physics*, 81(3). doi:10.1016/S0034-4877(18)30050-8.
- Unpingco, J. (2016). *Python for probability, statistics, and machine learning*. Springer International Publishing.

- Wellman, J. F., & Regenauer-Lieb, K. (2012). Uncertainties have a meaning: information entropy as a quality measure for 3-D geological models. *Tectonophysics*, 526–529. doi:10.1016/j.tecto.2011.05.001.
- Wu, X., Ma, T., Cao, J., Tian, Y., & Alabdulkarim, A. (2018). A comparative study of clustering ensemble algorithms. *Computers and Electrical Engineering*, 68. doi:10.1016/j.compeleceng.2018.05.005.
- Zhang, X., & Yu, Y. (2018). Spatial weights matrix selection and model averaging for spatial autoregressive models. *Journal of Econometrics*, 203. doi:10.1016/j.jeconom.2017.05.021.
- Zurada, J., Levitan, A., & Guan, J. (2011). A comparison of regression and artificial intelligence methods in a mass appraisal context. *Journal of Real Estate Research*, 33(3).

Acknowledgments

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

Annex

Table 1. Summary of selected results of plots of land clustering (k -nearest neighbours spatial weights matrices)

Number of neighbours (k)	Number of clusterings meeting both thresholds	Minimum number of LAZ with entropy below the threshold	Required number of representative plots for first number of LAZ meeting entropy threshold
3	10	32	71
5	5	34	77
10	6	29	74
20	14	23	59
50	26	17	50

Table 2. Summary of selected results of plots of land clustering (distance band spatial weights matrices)

Distance band (r)	Number of clusterings meeting both thresholds	Minimum number of LAZ with entropy below the threshold	Required number of representative plots for first number of LAZ meeting entropy threshold
50 m	38	16	46
75 m	37	14	41
100 m	22	15	47
250 m	32	16	48
400 m	34	14	44

Figure 1. Example of neighbour graph for $k = 3$ and $k = 5$ nearest neighbours

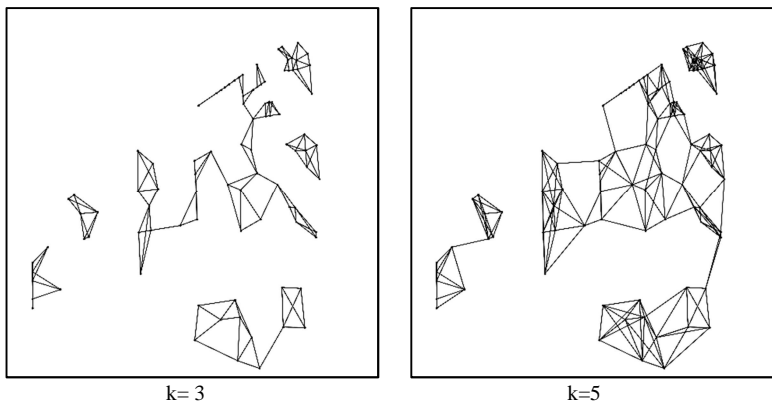


Figure 2. Example of neighbour graph for $r = 50$ meters and $r = 75$ meters distance band

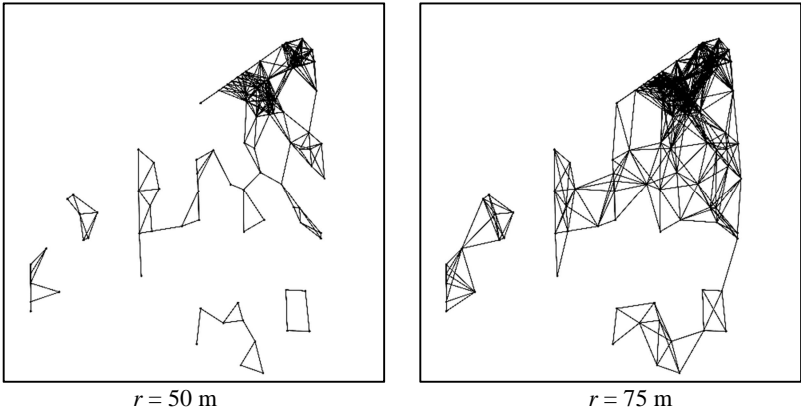


Figure 3. Example of plots of land clustering



Figure 4. Average LAZs' entropy for selected k nearest neighbours spatial weights matrices and different number of LAZs

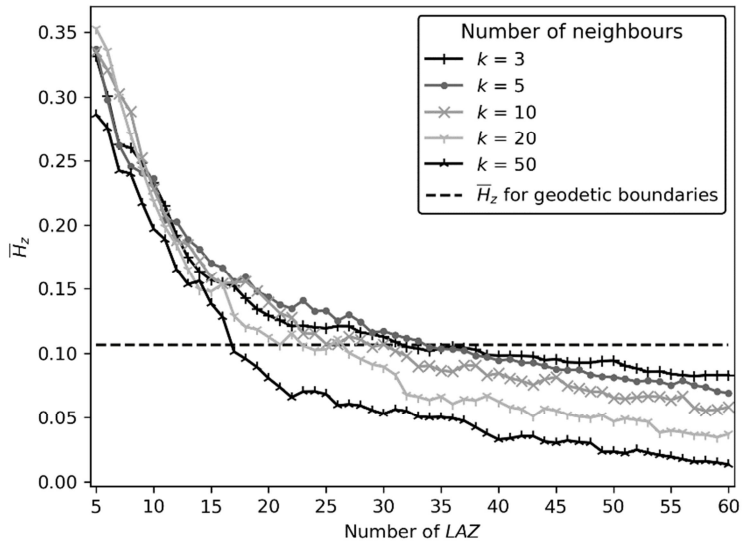


Figure 5. Heatmap of standard deviations for all examined clusterings (k -nearest neighbours spatial weights matrices)

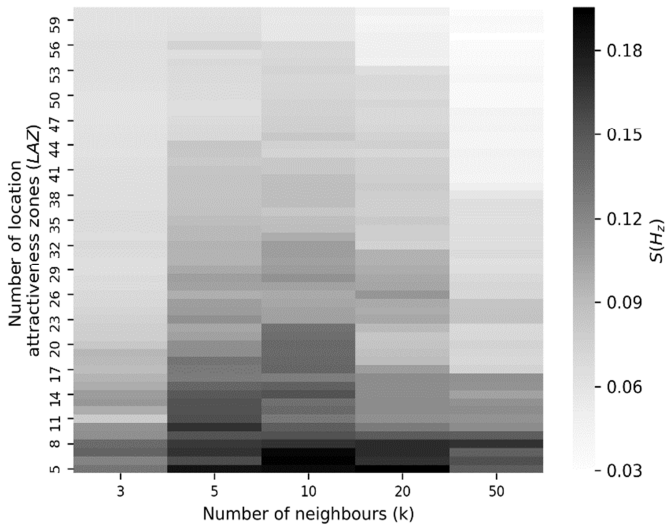


Figure 6. Scatterplot of average LAZs' entropy and required number of representative plots (k -nearest neighbours spatial weights matrices)

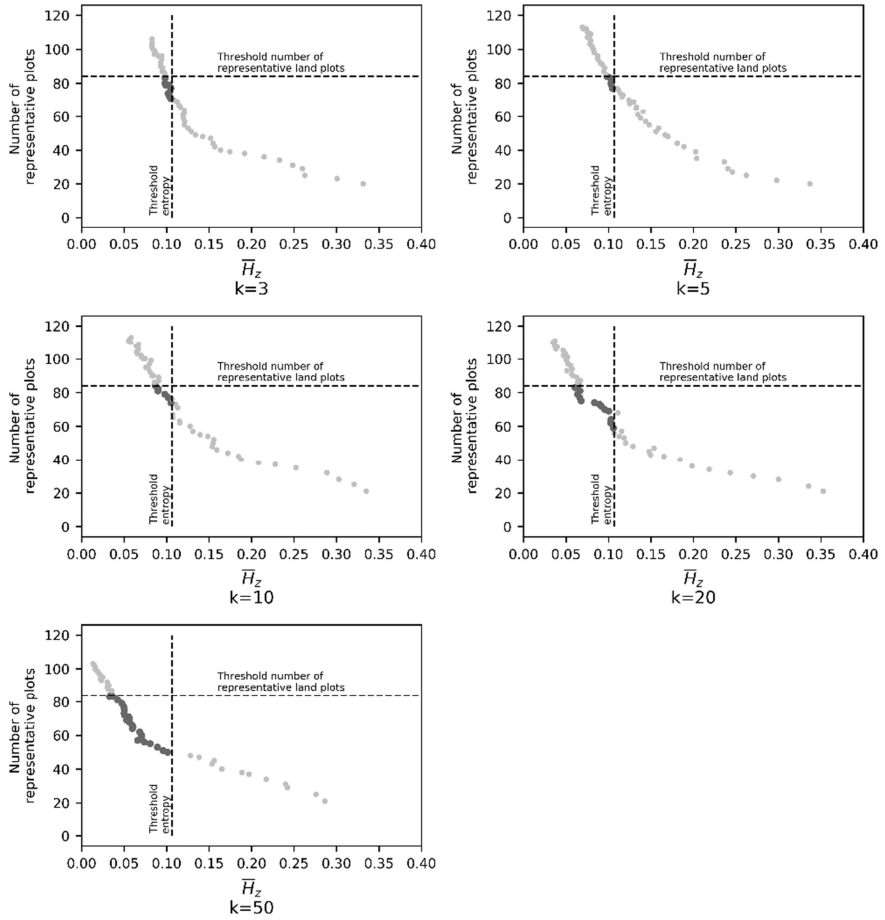


Figure 7. Average LAZs' entropy for selected distance band spatial weights matrices and different number of LAZs.

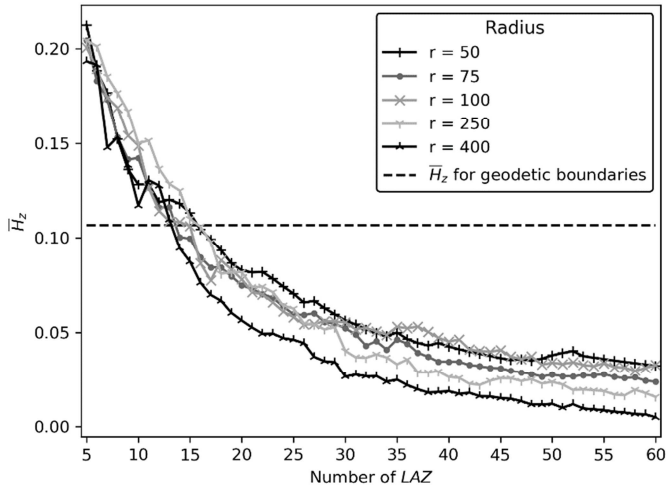


Figure 8. Heatmap of standard deviations for all examined clusterings (distance band spatial weights matrices)

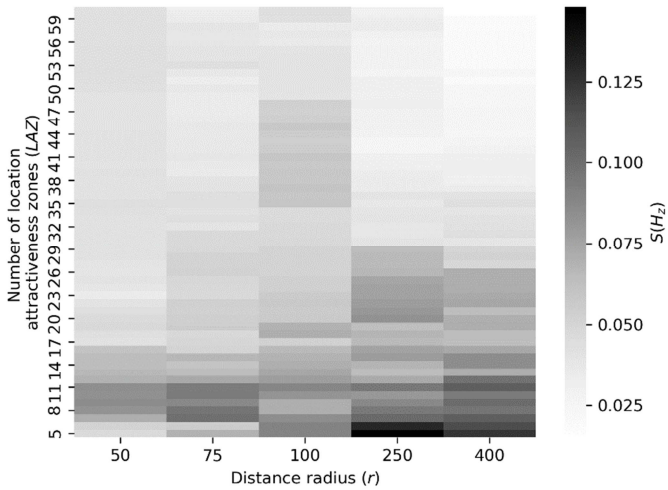


Figure 9. Scatterplot of average LAZs' entropy and required number of representative plots (distance band spatial weights matrices)

