geosciences

MDPI

Article

Estimating Residential Property Values on the Basis of Clustering and Geostatistics

Beata Calka

Faculty of Civil Engineering and Geodesy, Military University of Technology, 00-908 Warsaw, Poland; beata.calka@wat.edu.pl; Tel.: +48-608-896-206

Received: 7 February 2019; Accepted: 21 March 2019; Published: 24 March 2019

Abstract: The article presents a two-stage model for estimating the value of residential property. The research is based on the application of a sequence of known methods in the process of developing property value maps. The market is divided into local submarkets using data mining, and, in particular, data clustering. This process takes into account only a property's non-spatial (structural) attributes. This is the first stage of the model, which isolates local property markets where properties have similar structural attributes. To estimate the impact of the spatial factor (location) on property value, the second stage involves performing an interpolation for each cluster separately using ordinary kriging. In this stage, the model is based on Tobler's first law of geography. The model results in property value maps, drawn up separately for each of the clusters. Experimental research carried out using the example of Siedlce, a city in eastern Poland, proves that the estimation error for a property's value using the proposed method, evaluated using the mean absolute percentage error, does not exceed 10%. The model that has been developed is universal and can be used to estimate the value of land, property, and buildings.

Keywords: clustering; k-means; geostatistics; value map; spatial location; mass appraisal

1. Introduction

Property value estimates are usually carried out to determine the value of individual or multiple properties. The valuation of many properties at once, called mass appraisal, is carried out using statistical methods. Mass appraisal aims to determine a property's market value for a general transaction or to draw up property value maps. Converting mass appraisals into algorithms requires the development of a model for estimating property value. These models are usually based on a comparative approach [1,2], which assumes that a property's market value is determined by comparing it with other properties with known transaction (purchase and sale) prices and the factors differentiating these properties and affecting their value. All the information needed for the appraisal comes from the market and has a probabilistic character.

There are many ways of applying statistical methods and creating multidimensional mathematical models to describe a given property market [3–7]. The most frequently used model is the hedonic one, which uses multiple regression [8] and the ordinary least square method (OLS). In hedonic models, all the attributes affecting the property's value are analyzed together; the size of each one's influence is also studied [9–11]. The main disadvantage of hedonic models is that they take into account location by parametrizing the distance from the city center using van Thunen's theory [12]; for the neighborhood, they look at the zoning of the site in the land-use plan [13].

The development of computer techniques and geographical information systems means that statistical analysis is increasingly supplemented with geostatistical analyses, and that models take into account the spatial auto-correlation between a property's transaction price and its geographical location. This type of research has been carried out by Basu and Thibodeau [14], Gillen et al. [15],

Genfald et al. [16], Tu et al. [17], Chica-Olmo et al. [18], Giannopoulou et al. [19], Zhang et al. [20,21], Palma et al. [22], and Renigier-Biłozor et al. [23]. Cichociński [24] carried out an attempt to use geostatistical methods, namely simple kriging, to interpolate property values, while Cellmer et al. [13] as well as Colaco and Vucetic [25] applied the regression–kriging model, called the hybrid model, to estimate the value of land. They all found that developing a model that takes into account the spatial auto-correlation between a property's price and its location is incredibly difficult, as the price is affected by many spatial factors, not just location. Kontrimas and Vericas [1] found that estimate accuracy is increased by the weight assigned to a property's attributes and the use of computational intelligence. Artificial neural networks (ANN) are increasingly used to valuate property. McCluskey et al. [26] and Worzala et al. [27] investigated the comparative performance of an ANN and several multiple regression techniques in terms of their predictive accuracy and potential for use in the mass appraisal industry. The results obtained were dissatisfying since the linear regression model had a higher predictive accuracy than the artificial neural network. The latest research by Peterson and Flanagan [28] found that a single layer ANN is functionally equivalent to OLS, while multiple layered ANNs are capable of modeling complex nonlinearities. In general, an ANN is better suited to hedonic models that typically use large numbers of dummy variables.

The models used to estimate property values are multidimensional. This is because many factors (structural, environmental, and economic) affect a property's value. The multidimensional nature of geographical space is usually ignored and the impact of a property's situation reduced to an analysis of its location and neighborhood, treated as environmental features. Wong et al. [29] set out to solve this problem by pointing out that geographical space is three-dimensional and that it is also important to take into account a property's vertical location when estimating its value. Their three-dimensional model featuring spatial correlation gives results comparable to OLS.

Property value is typically presented using land value maps, which were already being drawn up at the start of the 20th century [27]. Analyzing them allows the property market to be assessed visually and anomalies and mistakes in estimating property value to be spotted. It helps provide an economic evaluation of an investment and makes it easier to plan the development of a city or the use of a given area. Many scholars emphasize the usefulness of property value maps, including Cellmer [30], Gall [31], and Sayce et al. [32]. They note that apart from fiscal purposes, property value maps are used by real estate buyers and sellers, developers, planners, and urbanists, as well as politicians. Źróbek et al. [33] point to two main goals of property value maps. These are short-term goals, which are related to market restructuring efforts, and long-term goals, which involve regular monitoring of prices as an effective tool for managing land resources.

Moreover, Batt [34] found that public access to property value maps impedes corruption when determining the tax on property or issuing various administrative decisions. The rules for creating property value maps for purposes other than mass appraisal have not been established. The most frequently used cartographic methods are isolines or choropleth maps [13]. There are also maps drawn up using anamorphosis, e.g., showing the value of residential property in an area by purchasing power parity [35]. An analysis of the maps has revealed that the combination of cartographic presentation methods has a beneficial effect on the scope of information that can be extracted from them [36]. It is worth pointing out that the cartographic aspect of property value maps is incredibly important, as the maps are intended for a wide range of users. Medyńska-Gulij [37,38], Calka [39,40], and Horbiński et al. [41] draw attention to the importance of cartographic correctness in maps and cartographic presentation methods in their works.

An analysis of the literature clearly shows that one of the main problems when developing models for estimating the value of property is taking into account geographical location. The existing solutions based on auto-correlation are dissatisfying, mainly due to the considerable variation within the property market. One solution is to divide the market into local submarkets, based on similarities in the use of land. Yet this solution is not always satisfying either. This article presents an application of sequences of methods in the process of developing property value maps. This approach is based

on isolating submarkets, using a property's physical attributes, and cluster analysis. The submarkets formed this way are characterized by relative homogeneity but are not continuous in the spatial sense. In the next step, the model performs geostatistical interpolation and designates a continuous area of property value, which is then used to draw an isoline map. The number of maps created this way corresponds to the number of clusters the properties are divided into. With this set of property value maps it is possible to estimate the value of any given property described by a set of concrete structural attributes. It is also possible to estimate the potential price of a property in a particular location, depending on its standard, the type of market, the number of rooms, and the story it is on; in other words, depending on its structural attributes.

The proposed maps of residential property prices can be used as a source of information for different purposes like property management and administration, where rapid access to objective values and prices is crucial. They can be also be used as a base for any analysis of the residential property market, in particular for analyses dealing with investment advice.

The article is structured as follows: the second section describes the research methods used, stating the assumptions of the model, the choice of attributes affecting the price, the way of grouping using k-means, and interpolation using ordinary kriging. The third section describes the research experiment, carried out using the example of individual properties in Siedlce, a city in eastern Poland. The article closes with a summary and conclusions.

2. Methods

2.1. The Two-Stage Model, General Assumptions

The two-stage model for estimating property values is based on the assumption that the non-spatial and spatial attributes of a property should be analyzed separately. As a result, a location-insensitive model is developed in the first stage, based only on the properties' structural characteristics. Using data mining techniques, in particular k-means clustering, clusters of similar properties creating submarkets are formed. Each cluster is characterized by defined attribute values, expressed on a rank scale. In this model, the independent variables are: the floor area of the property, the number of rooms, additional rooms, which story it is on, its standard, the year the building was built, and the type of market.

The second stage develops a pure spatial model, based on the assumption that the value of a property in each of the clusters depends exclusively on the distance between properties with known prices and the property being evaluated. Each property is spatially unique, in a horizontal and vertical sense, and location is always an intrinsic attribute that directly determines the quality of the mass appraisal model. This is consistent with Tobler's first law of geography [42], which states that "everything is related to everything else, but near things are more related than distant things" and Pearson's [28] well-known statement "location, location, location". To estimate the value of a property in any given place, ordinary kriging—one of the geostatistical methods—is used. The independent variable in this model is the geographical location (x, y coordinates); the dependent one is the property's value. The model results in isoline maps of property values, drawn up separately for each of the clusters. The modification of the kriging depends on its purpose. In the process of property value estimation, discontinuity lines and areas are not taken into account [13,24], but they are important in environmental and geological research [43,44].

2.2. The Choice of Representative Attributes and Updating of Transaction Prices

In the mass appraisal of properties, the choice of attributes depends on the sources of information available. In Poland, information on the sale price of properties is collected by the public administration in the Register of Property Prices and Values. Apart from the price, it also includes data on the technical state of the property and its location (in the form of an address). Finding out more on a mass scale is practically impossible. For land property, additional information—such as on how it is being used,

the quality of the soil, or the site's purpose in the land-use plan—can be obtained from the cadastre. However, this only relates to the technical attributes of the property and its owners [45]. This study analyzes the following attributes: the floor area of the property, the number of rooms, additional rooms, the value, the year the building was built, the type of building, the story the property is on, and the type of market.

In statistical procedures the selection of an appropriate analytical function that would include a set of information on market prices is crucial. A linear regression can be applied if individual prices indicate uniform distribution over time. However, rapid falls and increases of prices together with periods of stability are more frequent, which excludes the use of linear estimation. Instead the weighted linear regression can be used with a division into time periods [46–48]. According to Czaja [45] the general formula of weighted linear regression is:

$$c = a_0 + a_1 \times X_1 + a_2 \times X_2 + a_3 \times X_3 + \ldots + a_n \times X_n, \quad (1)$$

where:

$X_1, X_2, \ldots, X_n$—all property attributes taken into account at the market (including the time of the transaction);
c—individual property prices at the market; and
$a_1, a_2, \ldots, a_n$—regression coefficient of variable c with respect to variable X.

2.3. Clustering—Spatially Insensitive Model

The k-means algorithm is a well-known method for partitioning n points that lie in the d-dimensional space into k clusters [49]. Given a set of observations $(x_1, x_2, \ldots, x_n)$, where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k ($\leq$n) sets $S = \{S_1, S_2, \ldots, S_k\}$ so as to minimize the within-cluster sum of squares. Its objective is to find:

$$\operatorname{argmin}(s) \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2, \quad (2)$$

where μ_i is the mean of points in S_i.

When grouping individual properties, the spatial dimension (d) is defined by the number of property attributes being analyzed.

As the k-means algorithm is sensitive to the number of clusters adopted a priori, the grouping was carried out using hierarchical agglomerative clustering, using the Euclidian distance merged with Ward's method, described in Ward [50].

2.4. Interpolation of Property Values Employing the Ordinary Kriging Method—Pure Spatial Model

In ordinary kriging, as in other kriging procedures, the interpolated value takes the form of a weighted average:

$$Z(x_0) = \sum_{i=1}^{n} w_i Z(x_i) \quad (3)$$

where:

w_i—is the weighting factor assigned to a single observation;
Z_i—is the value of the parameter being studied at a single point, and;
n—is the amount of data taken into account when estimating the value of the parameter.

The value of the weighting factors w_i assigned to individual observations is calculated based on the variability of the complex parameter, depending on the distance from the measuring point provided by the semivariogram. The semivariograms present the variation in the parameter values depending on the distance between the measuring points, and therefore the character of their variability; indirectly, they characterize the auto-correlation between the observations.

An advantage of the geostatistical method is that it computes the interpolation error, describing the reliability of the result. Cross-validation, a process estimating the value of the studied parameter for each point based on the user-chosen semivariogram model, is used to analyze the accuracy of the model. To calculate the models' error, this estimate is then compared to the input value at a given point. The mean error (ME), mean square standard residual (MSSR), and root-mean-square error (RMSE) are the most commonly used [51]. The average standardized error ought to tend towards 0; a positive or negative value indicates that the values being studied have been overestimated or underestimated. An MSSR close to 1 points to a very close fit between the theoretical and empirical model.

2.5. Assessing the Accuracy of the Estimate

Estimating property values, like any other statistical modeling, produces a certain degree of error associated with estimation. The error measurements proposed for evaluating the model are studied in depth in several works, such as Bielecka and Bober [52], Isaaks and Srivastava [53], and Willmott [54]. Commonly used error measurements include: the mean error (ME), mean absolute error (MAE), mean absolute percentage error (MAPE), mean square error (MSE), and root mean square error (RMSE). RMSE and MAE are considered among the best overall measures of model performance as they summarize the mean difference between observed and estimated values [48]. MAPE is useful because it is expressed in generic percentage terms.

To verify the accuracy of the estimated property values, MAE and MAPE were used. The value of both errors was assigned based on a randomly chosen 10% sample of the data, not included in the interpolation. The mean absolute error (MAE) is calculated as an absolute mean difference between observed and estimated values of a sample in an n location, according to the formula:

$$MAE = \frac{1}{n}\sum_{\tau}^{n}\left|y_\tau - y_\tau^p\right|, \quad (4)$$

where:

$\left|y_\tau - y_\tau^p\right|$—difference between the estimated value and the observed value for the time period τ.

The mean absolute error (MAE) provides information how much on average observed variables and the estimated variables will deviate from the absolute value for a given time period. It is also possible to designate the mean absolute percentage error (MAPE).

$$MAPE = \frac{1}{n}\sum_{\tau}^{n}\left|\frac{y_\tau - y_\tau^p}{y_\tau}\right|, \quad (5)$$

The mean absolute percentage error provides information on the mean error value for the time periods of $\tau = 1, 2, \ldots, m$, in the real percentage of observed variables. The designation of the above errors will make it possible to validate the developed model and evaluate the accuracy of the prediction of property value.

3. Experimental Investigation

3.1. Study Area and Data

The study area was Siedlce, a city in eastern Poland, in the Masovian province (Figure 1). According to the Polish statistical data, the total housing stock in Siedlce amounts to 32,254 housing properties. Therefore, there are 418 housing properties for every 1000 inhabitants. It is a value comparable to the value for the Mazowieckie Province and much higher than the average for the whole of Poland.

The data on 1873 transaction prices (sales) of housing property in Siedlce and property attributes were obtained from the Register of Prices and Values belonging to Siedlce's city administration.

The properties being bought or sold belonged to both the primary (911 properties) and secondary (962) markets (Figure 1). The average indicators of living conditions in Siedlce are somewhat below the Polish average. The average flat in Siedlce has an area of 54.33 m^2 (national average: 55.80 m^2) and is home to 3.17 people (national average: 2.98). The average number of rooms is 3.48 (national average: 3.37). In Siedlce, the average flat contains 17.13 m^2 of space per person, compared to the national average of 18.89 m^2. Over the past few years, around 350 new flats have appeared each year.

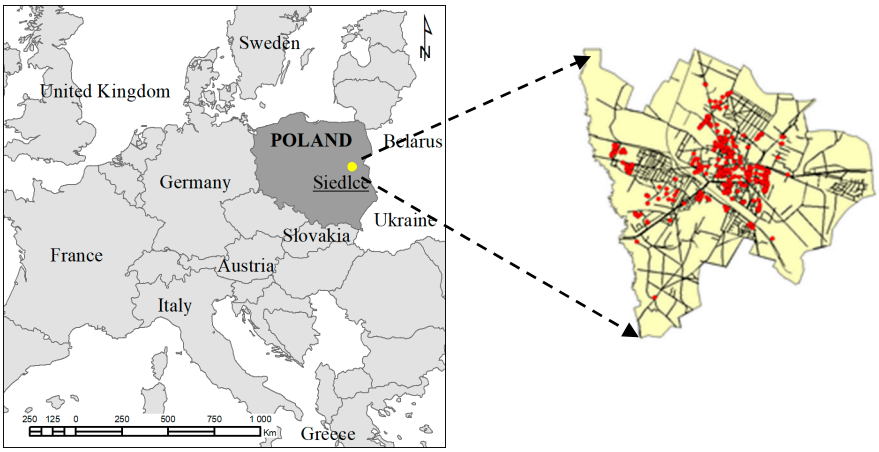

Figure 1. Location of study area and properties being bought and sold in Siedlce.

The research used data on property purchases or sale transactions between 2007 and 2011. To analyze the trend of price changes over time, a weighted linear regression was used. Based on the analysis a clear linear upward trend can be noticed, with different values during three time periods. In the first time period, from January 2007 to February 2008, the monthly average property price increased by 39 Polish zloty (PLN)/month. In the second time period the increase was 77 PLN/month. During that time property prices in Siedlce increased both on the primary and secondary markets. In mid-2009 there was some price stabilization. Despite that, the increasing trend continued in the third time period, from May 2009 to March 2011, but it was of a lower value, 13 PLN/month. The prices were updated for the date of the last transaction, which was March 2011. The coefficient of determination R^2 was 0.67.

Table 1 presents the descriptive statistics for property prices. After a preliminary analysis of the correlation, four attributes were chosen: the type of market, the standard of the flat, the year it was built, and the story it is on. Table 2 present the attributes taken into account in the analysis together with their domains and rank. The others were dismissed as statistically insignificant.

Table 1. Descriptive statistics for selling price * data.

Statistics	Values
Number of properties	1873
Minimum price	2518
Maximum price	4830
Average price (arithmetical)	3577
Standard deviation	470
Asymmetry coefficient (skewness)	0.0621
Kurtosis	−0.5300
First quartile	3237
Median	3575
Third quartile (PLN)	3887
Shapiro–Wilk test	W = 0.99217; p = 0.0000

* prices in Polish zloty (PLN).

Table 2. Property attributes analyzed.

Name of Attribute	Domains of Attributes	Rank	Cramer's Correlation Coefficient (*p*-Value)
Type of market	primary secondary	1 2	0.28 (0.000)
Standard of the flat	high average low	1 2 3	0.62 (0.000)
Year of construction	since 2001 1985–2000 older	1 2 3	0.23 (0.000)
Story	first floor middle floor ground and top floors	1 2 3	0.18 (0.000)

3.2. Results

Separating similar properties based on their non-spatial properties was carried out in two stages. To establish the optimal number of clusters, a hierarchical taxonomy using the agglomerative method was performed, followed by appropriate grouping using the k-means method. An analysis of the dendrogram (Figure 2), the result of agglomerative clustering, led to all the transactions being divided into five clusters. The k-means clustering was carried out iteratively, assuming that the preliminary centers of the clusters would be designated using the Euclidean distance sorting method.

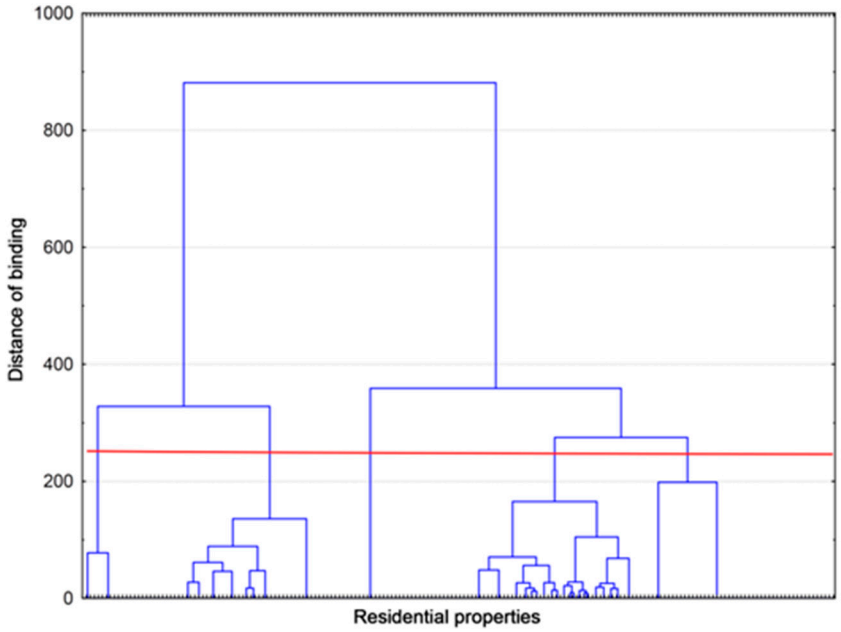

Figure 2. Dendrogram of agglomerative clustering.

As a result, the analyzed properties were divided into five clusters: two consisting of properties acquired directly from the developer (on the primary market) and three with properties bought on the secondary market. The general characteristics of the properties in each cluster are presented in Table 3.

The most expensive properties are in cluster B. These flats are found on the best stories (the middle ones), and they are of a high standard and in new buildings on the secondary market. Their high standard has a significant impact on their average price, which is the highest. The cheapest properties are in cluster C. They are on the top or bottom story, in old buildings of a low standard. These flats were also sold on the secondary market.

Table 3. Characteristics of the separated property clusters.

Attribute	Cluster A	Cluster B	Cluster C	Cluster D	Cluster E
Type of market	secondary	secondary	secondary	primary	primary
Standard of the property	average	high	low	average	high
Year of construction	before 2006	after 1990	before 1990	after 2006	after 2006
Story	middle	middle	top	middle or top	middle
Average price (PLN/m^2)	3671	3906	3195	3510	3790
Number of properties	441	199	307	668	258

For geostatistics to be used to estimate average property values, a few boundary conditions need to be met. These include: a lack of local extrema, a normal distribution of data, and stationary data. Extreme local values were identified by analyzing graphs of normal percentiles and then removed from the analyzed data. The existence of a normal distribution was examined by setting the Shapiro–Wilk coefficient p; its values are presented in Table 4. For all the clusters the value of coefficient $p < 0.05$, so at a significance level of 0.05 a normal distribution cannot be assumed.

Table 4. Descriptive statistics for property prices in different clusters.

Descriptive Statistics	Cluster A	Cluster B	Cluster C	Cluster D	Cluster E
Number of properties	441	199	307	668	258
Minimum value (PLN)	2582	2747	2518	2560	2807
Maximum value (PLN)	4830	4764	4370	4655	4691
Arithmetic mean (PLN)	3671	3906	3195	3510	3790
Standard deviation (PLN)	496	499	410	350	388
Asymmetry coefficient (skewness)	0.1283	−0.5856	0.5865	−0.3654	−0.2630
Kurtosis	−0.8164	−0.6665	−0.1457	0.1811	−0.5209
First quartile (PLN)	3303	3570	2851	3292	3506
Median (PLN)	3624	4020	3167	3559	3829
Third quartile (PLN)	4070	4282	3433	3764	4072
Shapiro–Wilk test	W = 0.98046 p = 0.00001	W = 0.94194 p = 0.00000	W = 0.96287 p = 0.00000	W = 0.98091 p = 0.00000	W = 0.98621 p = 0.01388

In order to obtain a normal distribution, property prices from cluster A were subjected to logarithmic transformation and those from cluster B to Box–Cox transformation with a coefficient of 2.73. In the other clusters, a normal distribution of prices was obtained by removing outlying data. The occurrence of extrema was examined in global terms by creating histograms of property prices and graphs of normal percentiles for each of the groups. Extreme values were also observed in local terms, which means that they can be applied only to a specific region by creating Voronoi maps. Global and local extrema can have an adverse effect on the estimated area modifying the semivariogram model, which is why it is so important to indentify and remove outliers.

To draw up an isoline map of the variability of property values, spherical semivariograms were used. The parameters of the semivariograms and cross-validation errors are set out in Table 5. The results of the interpolation are presented in Figure 3.

Table 5. Semivariogram parameters and errors of pure spatial model cross-validation.

	Cluster A	Cluster B	Cluster C	Cluster D	Cluster E
Lag size (m)	360	164	98	184	180
Number of steps	12	12	12	12	12
Smooth factor	0.2	0.2	0.2	0.2	0.2
ME (PLN)	−0.22	2.19	3.59	1.44	2.68
RMSE (PLN)	342.71	310.21	318.69	194.53	218.34
MSSR	1.016	0.917	1.009	0.935	0.920

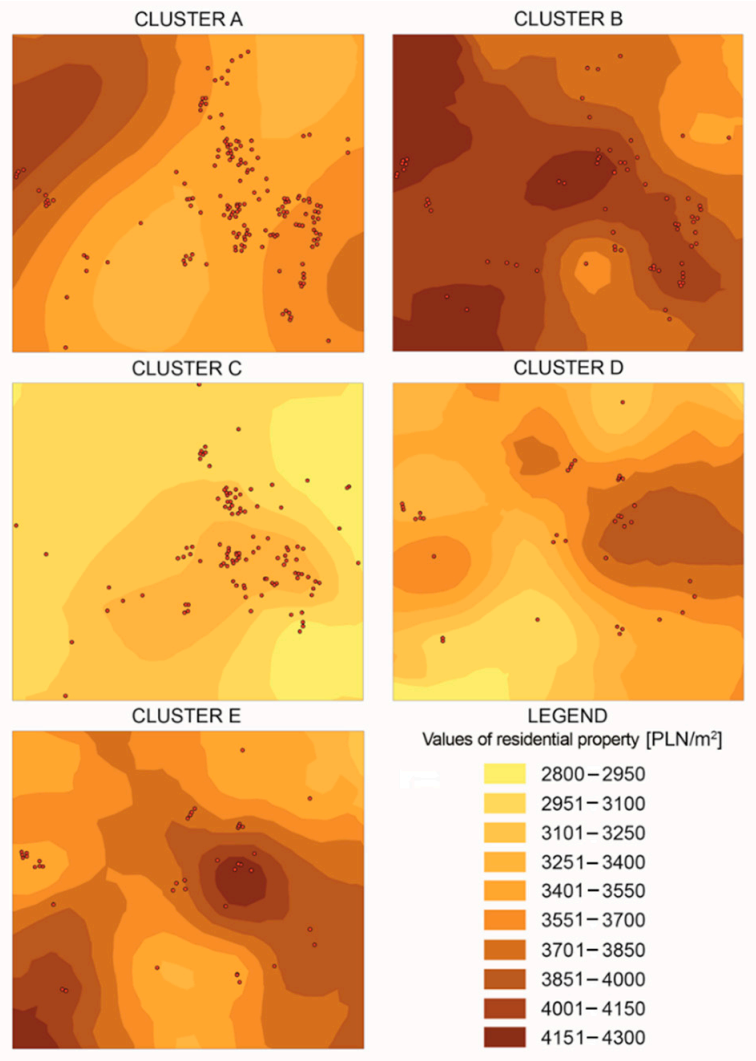

Figure 3. Kriging interpolation of residential property values.

When developing a map of residential property mean values, it was assumed that the background map would feature selected topographic content (transport networks, water areas, woods) and the land which in accordance with the local development plan may not be intended for housing. A sample map with low-price properties on the primary market is presented in Figure 4.

The MSSR of the pure spatial model is close to 1, which means that the model is credible. A relatively large RMSE was observed for property on the secondary market (clusters A, B, and C). This is linked to a much higher variation in the property attributes and, as a result, a greater variation in property prices (see Tables 3 and 4). The size of the errors in these clusters is also undoubtedly affected by a widely-known shortcoming of ordinary kriging: underestimating low values and overestimating high ones.

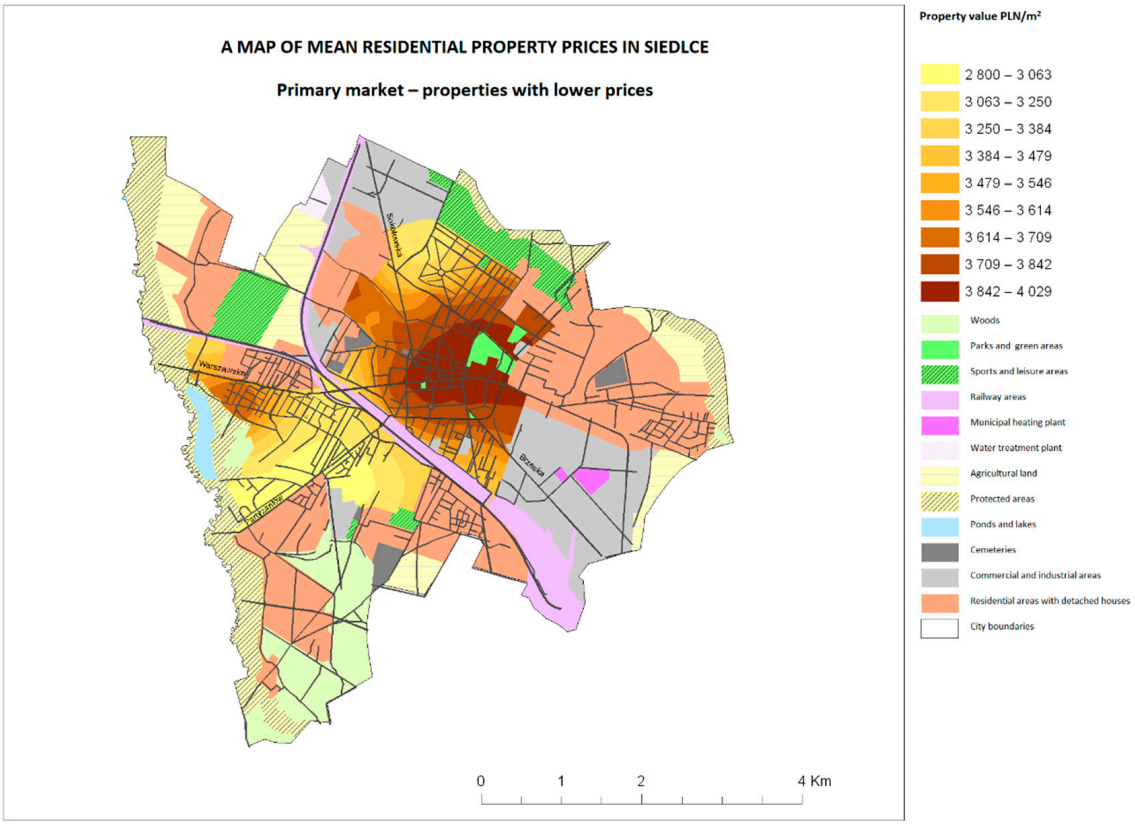

Figure 4. Map of residential property values.

The accuracy of the property value estimates was checked using a test sample of 10% of the properties not taken into account in the interpolation. The values of the MAE and MAPE errors are presented in Table 6.

Table 6. Errors of estimating property values.

Statistics	Cluster A	Cluster B	Cluster C	Cluster D	Cluster E
Number of points in test sample	34	18	24	46	20
MAE (PLN)	346	263	246	302	165
MAPE (%)	9.5	6.3	7.8	8.5	4.5

The estimation error was lowest for cluster E, consisting of high-standard flats bought directly from the developer. The value of the mean absolute error was 165 PLN, while the mean absolute percentage error was 4.5%. The chart of predictions in Figure 5 shows the discrepancies between the prices in the Register of Property Prices and Values and the predicted values for that cluster. The fact that the blue line (tailored to the points) is above the black one in the early phases of the chart shows that, when estimating property values, the kriging method overestimates low values and underestimates high ones.

An analysis of the error values (Table 5) allows us to state that the two-stage modeling of property values enables property values to be estimated without an error exceeding 10%. To obtain credible results, the number and distribution of points with known values—in this case flats in each of the clusters whose price is known—is relevant. The research carried out shows that the minimum number of points in a cluster should not be below 30, but that to obtain an estimation error of less than 10% it should be around 200.

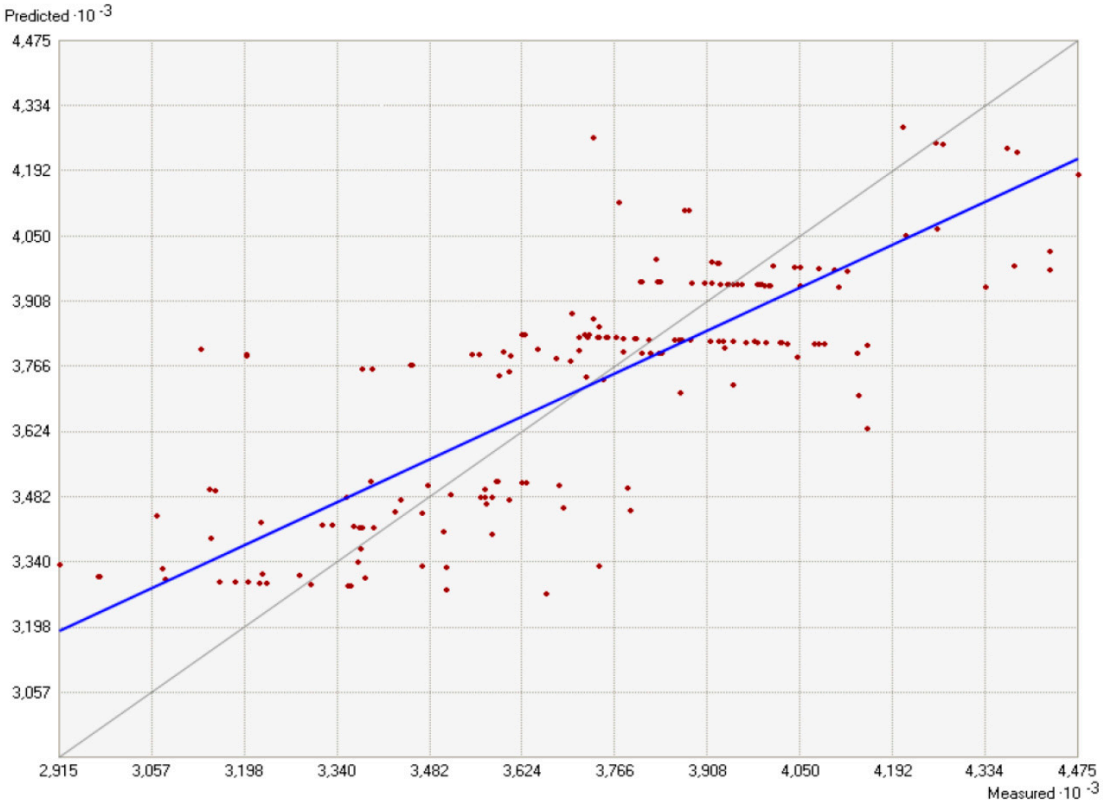

Figure 5. Chart of predictions of average property values for high-standard flats on the primary market (cluster E).

4. Discussion

In the hedonic model, the most commonly-used one for estimating property values, spatial position (location) is taken into account indirectly, by determining the accessibility and neighborhood. Accessibility is measured as distance from the city center, in line with the location theory developed by von Thunen [55], and neighborhood is usually understood as the property's purpose in the land development plan. The proposed model takes into account location, analyzed in accordance with the rules of geostatistics and interpolation using the kriging model.

Although the method used in the article guarantees that within each of the zones reliability in the estimated property value is greater than 90%, it has some limitations. First of all, geostatistical analysis manifests a high sensitivity to the presence of a normal distribution in an analyzed data sample. This can have the effect of serious errors for extreme values. That is why an important part of the research is to prepare appropriate data before analysis, and in particular to detect global and local extrema. Of the many types of kriging, ordinary kriging was chosen to estimate the value of a property at any given point. This method limits the stationary nature of the average price to the local neighborhood with its center at the point of estimation, which is particularly significant in the case of individual residential properties.

There are some limitations to the k-means method since it requires a declaration of the number of groups. If the number of groups is too large or too small then the clusters of data are heterogeneous and inseparable. To choose an appropriate number of groups the agglomeration method was applied. It presents the aggregation of individual elements into groups in the form of a dendrogram. It is possible to observe the distance between newly created clusters, which in a clear way allows interpretation of the grouping process and correct decision making.

The accurateness of the grouping is largely determined by establishing a scale to express the values of property attributes and the range of adopted values. The number of attributes determines

the spatial dimension (d) and the range of attribute values the correctness of assigning a particular property to a cluster. Unlike hedonic models, grouping does not use a dummy variable. Choosing a rank scale enables us to minimize the number of properties with extreme attribute values (outliers), which considerably disrupt the correctness of the grouping.

5. Summary and Conclusions

The research carried out shows that it is possible to model the spatial variation in property values using geostatistical methods after first eliminating the influence of non-spatial attributes on a property's price. The model developed to estimate property values is a predictive one. In the first (spatially insensitive) stage the model selects similar properties, characterized by close non-spatial attributes. Due to the large number of analyzed properties, data clustering is used to group them, including the k-means algorithm. An analysis of the number of property clusters needs to precede the grouping. The authors recommend an agglomeration model that allows the process of group formation to be observed on a dendrogram. Designating zones of property values using geostatistical interpolation (kriging) is carried out individually for each of the clusters, and the number of zones, their price range, and spatial distribution can vary significantly in each cluster.

The methodology described in the article brings new opportunities to explore a local property market, which may result in providing its thorough image. It allows the observation and detailed analysis of the dynamic processes taking place on property markets. Obviously, it is easier to model the primary market because it is more stable.

The methodology presented in this paper does not depend on the kind of property market and can be used to develop property maps of villages, towns, and cities, no matter whether the market is big or small. The methodology is scalable and can be adapted to develop property maps and land-use maps. Some modifications need to be made, like the selection of property attributes, or the proper determination of the number of groups. In further research it is planned to test the method in a different research area.

Funding: The research was carried out as part of the statutory work conducted in the years 2015–2018 at the Military University of Technology in Warsaw, Poland, grant number PBS 933/2018.

Conflicts of Interest: The author declares no conflict of interest. The funders had no role in the design of the study, in the collection, analyses, or interpretation of data, in the writing of the manuscript, or in the decision to publish the results.

References

1. Kontrimas, V.; Verikas, A. The mass appraisal of the real estate by computational intelligence. *Appl. Soft Comput.* **2011**, *11*, 443–448. [CrossRef]
2. Linne, M.R.; Kane, S.M.; Dell, G. *A Guide to Appraisal Valuation*; Appraisal Institute: Chicago, IL, USA, 2000.
3. Bielecka, E.; Calka, B. Taxonomy of real estate properties with the use of k-means method. In Proceedings of the 14th International Multidiscipli-nary Scientific GeoConference SGEM 2014, Albena, Bulgaria, 17–26 June 2014.
4. Ciuna, M.; Milazzo, L.; Salvo, F. A Mass Appraisal Model Based on Market Segment Parameters. *Buildings* **2017**, *7*, 34. [CrossRef]
5. Maleta, M.; Calka, B. Examining spatial autocorrelation of real estate features using moran statistics. In Proceedings of the 15th International Multidisciplinary Scientific GeoConference SGEM, Albena, Bulgaria, 18–24 June 2015.
6. Maleta, M.; Mościcka, A. Selection and significance evaluation of agricultural parcels determinants. *Geod. Cartogr.* **2018**, *67*, 239–253. [CrossRef]
7. Maclennan, D.; Tu, Y. Economic perspectives on the structure of local housing systems. *Hous. Stud.* **1996**, *11*, 387–406. [CrossRef]
8. Pi-ying, L. Analysis of mass appraisal model. In Proceedings of the 23rd Pan Pacyfic Congress of Appraisal, Valuers and Consumers, San Francisko, CA, USA, 16–19 September 2006.

9. Cebula, R.J. The hedonic pricing model applied to the housing market of the City of Savannah and Its Savannah Historic Landmark District. *Rev. Reg. Stud.* **2009**, *39*, 9–22.
10. Canavarro, C.; Caridad, J.M.; Ceular, N. Hedonic Methodologies in the Real Estates Valuation. 2010. Available online: http://repositorio.ipcb.pt/handle/10400.11/412 (accessed on 3 January 2019).
11. Monson, M. Valuation using hedonic pricing model. *Cornell Real Estate Rev.* **2009**, *7*, 62–73.
12. Stevens, B.H. Location theory and programming models: The Von Thünen case. *Pap. Reg. Sci.* **1968**, *21*, 19–34. [CrossRef]
13. Cellmer, R.; Belej, M.; Zrobek, S.; Kovac, M.S. Urban land value maps—A methodological approach. *Geod. Vestn.* **2014**, *58*, 535–551. [CrossRef]
14. Basu, S.; Thibodeau, T. Analysis of spatial autocorrelation in house prices. *J. Real Estate Financ. Econ.* **1998**, *17*, 61–85. [CrossRef]
15. Gillen, K.; Thibodeau, T.G.; Wachter, S. Anisotropic autocorrelation in house prices. *J. Real Estate Financ. Econ.* **2001**, *23*, 5–30. [CrossRef]
16. Gelfand, A.E.; Ecker, M.D.; Knight, J.R.; Sirmans, C.F. The Dynamics of Location in Home Price. *J. Real Estate Financ. Econ.* **2004**, *29*, 149–166. [CrossRef]
17. Tu, Y.; Sun, H.; Yu, S. Spatial autocorrelations and urban housing market segmentation. *J. Real Estate Financ. Econ.* **2007**, *34*, 385–406. [CrossRef]
18. Chica-Olmo, J.; Cano-Guervos, R.; Chica-Olmo, M. A coregionalized model to predict housing prices. *Urban Geogr.* **2013**, *34*, 395–412. [CrossRef]
19. Giannopoulou, M.; Vamvatsikos, V.; Lykostratis, K. A Process for Defining Relations between Urban Integration and Residential Market Prices. *Procedia—Soc. Behav. Sci.* **2016**, *223*, 153–159. [CrossRef]
20. Zhang, L.; Wang, H.; Song, Y.; Wen, H. Spatial Spillover of House Prices: An Empirical Study of the Yangtze Delta Urban Agglomeration in China. *Sustainability* **2019**, *11*, 544. [CrossRef]
21. Zhang, Z.; Lu, X.; Zhou, M.; Song, Y.; Luo, X.; Kuang, B. Complex spatial morphology of urban housing price based on digital elevation model: A case study of Wuhan city, China. *Sustainability* **2019**, *11*, 348. [CrossRef]
22. Palma, M.; Cappello, C.; De Iaco, S.; Pellegrino, D. The residential real estate market in Italy: A spatio-temporal analysis. *Qual. Quant.* **2018**, *53*, 1–22. [CrossRef]
23. Renigier-Biłozor, M.; Janowski, A.; Walacik, M. Geoscience Methods in Real Estate Market Analyses Subjectivity Decrease. *Geosciences* **2019**, *9*, 130. [CrossRef]
24. Cichociński, P. An attempt to apply geostatistical methods to real estate valuation. *Ann. Geomat.* **2009**, *VII*, 17–24.
25. Colakovic, M.; Vucetic, D. Possibility of Using GIS and Geostatistic for Modelling the Influence of Location on the Value of Residential Properties. 2012. Available online: http://www.fig.net/pub/fig2012 (accessed on 4 January 2019).
26. McCluskey, W.; Davis, P.; Haran, M.; McCord, M.; McIlhatton, D. The potential of artificial neural networks in mass appraisal: The case revisited. *J. Financ. Manag. Prop. Constr.* **2012**, *17*, 274–292. [CrossRef]
27. Worzala, E.; Lenk, M.; Silva, A. An exploration of neural networks and its application to Real estate valuation. *J. Real Estate Res.* **1995**, *10*, 185–202.
28. Peterson, P.S.; Flanagan, A.B. Neural network hedonic pricing models in mass real eestate appraisal. *J. Real Estate Res. (JRER)* **2015**. Available online: http://ssrn.com/abstract=1086702 (accessed on 3 January 2019).
29. Wong, S.K.; Yiu, C.Y.; Chau, K.W. Trading volume-induced spatial autocorrelation in real estate prices. *J. Real Estate Financ. Econ.* **2013**, *46*, 596–608. [CrossRef]
30. Cellmer, R. The possibilities and limitations of geostatistical methods in real estate market analyses. *Real Estate Manag. Valuat.* **2014**, *22*, 54–62. [CrossRef]
31. Gall, J. Future of value maps in European context. In Proceedings of the XXIII FIG Congress, TS-17 Land Value Maps and Taxation, Munich, Germany, 8–13 October 2006.
32. Sayce, S.; Vickers, T.; Morad, M.; Connellan, O. Value map the next utility. In Proceedings of the XXIII FIG Congress, TS-17 Land Value Maps and Taxation, Munich, Germany, 8–13 October 2006.
33. Źróbek, S.; Cellmer, R.; Kuryj, J. Land value map as a source of information about local real estate market. *Geodezja* **2005**, *11*, 63–74.
34. Batt, H.W. Tax regimes that don't invite corruption. *Int. J. Transdiscipl. Res.* **2012**, *6*, 65–82.
35. Amster, D. Housing Prices. Map 194. Available online: http://www.worldmapper.org/posters/worldmapper_map194_ver5.pdf (accessed on 4 January 2015).

36. Sosnowska, M.; Karsznia, I. Methodology for mapping the average transaction prices of residential premises using GIS. *Pol. Cartogr. Rev.* **2016**, *48*, 161–171. [CrossRef]
37. Medynska-Gulij, B. Cartographic sign as a core of multimedia map prepared by non-cartographers in free map services. *Geod. Cartogr.* **2014**, *63*, 55–64. [CrossRef]
38. Medynska-Gulij, B. How the black line, dash and dot created the rules of cartographic design 400 years ago. *Cartogr. J.* **2013**, *50*, 356–368. [CrossRef]
39. Calka, B.; Cahan, B. Interactive map of refugee movement in Europe. *Geod. Cartogr.* **2016**, *65*, 139–148. [CrossRef]
40. Calka, B. Comparing continuity and compactness of choropleth map classes. *Geod. Cartogr.* **2018**, *67*, 21–34. [CrossRef]
41. Horbiński, T.; Medyńska-Gulij, B. Geovisualisation as a process of creating complementary visualisations: Static two-dimensional, surface three-dimensional, and interactive. *Geod. Cartogr.* **2017**, *66*, 45–58. [CrossRef]
42. Tobler, W. A computer movie simulating urban growth in the Detroit region. *Econ. Geogr.* **1970**, *46*, 234–240. [CrossRef]
43. Meyer, T.H. The Discontinuous Nature of Kriging Interpolation for Digital Terrain Modelling. *Cartogr. Geogr. Inf. Sci.* **2004**. [CrossRef]
44. Jimenez-Martinez, N.; Ramirez, M.; Diaz-Hernandez, R.; Rodriguez-Gomez, G. Fluvial Transport Model from Spatial Distribution Analysis of Libyan Desert Glass Mass on the Great Sand Sea (Southwest Egypt): Clues to Primary Glass Distribution. *Geosciences* **2015**, *5*, 95–116. [CrossRef]
45. Bielecka, E. The possibility of use the spatial data stored in state geodetic and cartographic resource for state property management. *J. Pol. Real Estate Sci. Soc.* **2012**, *20*, 19–30. (In Polish)
46. Maleta, M. Methods for Determining the Impact of the Temporal Trend in the Valuation of Land Property. *Real Estate Manag. Valuat.* **2013**, *21*. [CrossRef]
47. Czaja, J. *Methods of Estimating of the Market Value and Cadastral Value of the Properties*; KOMP-SYSTEM: Kraków, Poland, 2001.
48. Barańska, A.; Michalik, S. Variants of Modeling Dwelling Market Value. *Real Estate Manag. Valuat.* **2014**, *22*, 28–35. [CrossRef]
49. Vattani, A. K-means requires exponentially many iterations even in the plane. *Discret. Comput. Geom.* **2011**, *45*, 596–616. [CrossRef]
50. Ward, J.H. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **1963**, *58*, 236–244. [CrossRef]
51. Willmott, C.J. On the validation of models. *Phys. Geogr.* **1981**, *2*, 184–194. [CrossRef]
52. Bielecka, E.; Bober, A. Reliability analysis of interpolation methods in travel time maps-the case of Warsaw. *Geod. Vestn.* **2013**, *57*, 299–312. [CrossRef]
53. Isaaks, E.H.; Srivastava, R.M. *Applied Geostatistics*; Oxford University Press: New York, NY, USA, 1989.
54. Willmott, C.J. Some comments on the evaluation of model performance. *Bull. Am. Meteorol. Soc.* **1982**, *63*, 1309–1313. [CrossRef]
55. Dickinson, H.O. Von Thünen's Economics. *Econ. J.* **1969**, *79*, 894–902. [CrossRef]