sciendo

www.degruyter.com/view/j/remav

CATEGORICAL VARIABLE PROBLEM IN REAL ESTATE SUBMARKET DETERMINATION WITH GWR MODEL

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

Abstract

Real estate market analysis can involve many aspects. One of them is the study of the influence of various factors on prices and property values. For this type of issues, different kinds of measures and statistical models are often used. Many of them do not give unambiguous results. One of the reasons for this is the fact that the real estate market is characterized by the concept of local markets, which may be affected in different ways by economic, social, technical, environmental and other factors. Incorporating the influence of local markets, otherwise known as submarkets, into models often helps improve the precision of mass real estate valuation results. The delineation of submarket boundaries can be done in several different ways. One tool that is helpful in these types of situations are geographically weighted regression (GWR) models. The problem that may arise when using such models is related to the nature of some market factors, which may be of a qualitative nature. Because neighborhoods of individual properties may lack variability in terms of some variables, estimating GWR models is significantly difficult or impossible.

The study will present an approach in which the categorical variables are transformed into a single synthetic variable, and only this variable will constitute the explanatory variable in the model. Areas where the slope parameters of the GWR model are similar were considered a submarket.

The purpose of this paper is to determine the boundaries of submarkets in the study area and to compare the results of modeling the value of real estate using models that do not take local markets into account, as well as those that take into account local markets determined by experts and using the GWR model.

Key words: *property market segmentation, geographically weighted regression, property market analysis*.

JEL Classification: *C19, R30*.

Citation: Gnat, S. (2022). Categorical variable problem in real estate submarket determination with GWR model. *Real Estate Management and Valuation, 30*(4), 42-54.

DOI: https://doi.org/10.2478/remav-2022-0028.

1. Introduction

In the field of real estate valuation, two very different situations can be distinguished. The first is when a single property is subject to valuation. It is subjected to description, surveying and finally to the estimation of value. In the second situation, a large set of properties is subjected to valuation due to different needs: for example, in Poland, the legislator introduced three main objectives of mass property valuation: general property taxation (Bradbury et al., 2001), updating of perpetual usufruct fees and assessment of the economic effects of adopting and amending master plans. Mass valuations can be also useful for, e.g., banks (Korteweg & Sorensen, 2016; Tzioumis, 2017), which from time to time update the value of real estate, which are the basis for mortgage collateral. Mass appraisal

methods can also support investment decisions. When trying to estimate the value of a large number of properties, researchers often use multiple regression (MLR) models - hedonic models (Benjamin et al., 2004; Manjula et al., 2017; Sirmans et al., 2005). Mass real estate valuation is not a simple issue. It requires overcoming various types of difficulties(C. F. Chen & Rothschild, 2010; Santos et al., 2021). One such problem that arises when using multiple regression models is not considering spatial effects. MLR assumes that the relationship between individual characteristics and property prices (or values) is constant, which often turns out not to be a valid assumption. One way to account for space in regression models is to introduce model variables to the model which are responsible for the affiliation of individual properties to submarkets in which the relationships between attributes and prices are sufficiently stable. The application of such a procedure can be carried out in various ways (based on the existing division of the area into smaller administrative units, based on expert knowledge, or using various computational techniques). One of the approaches proposed in literature is based on a geographically weighted regression model. This assumes that the model is estimated for each property and then the structural parameters of the models are grouped into areas with similar values. Research in this area does not take into account a specific aspect of property description that is often used – the fact that some property attributes are qualitative in nature and variables are described on a nominal or ordinal scale. In such a situation, attributes are included in multiple regression models as dummy variables. When creating the area in GWR model from which observations on the basis of which the model is estimated for *i*-th property are taken, a situation when some of dummy variables have the same value may occur, thus making it impossible to estimate the model. The problem regarding the lack of variability for the dummy variables in the process of applying the GWR model to delineate real estate submarkets is not sufficiently addressed, which is a research gap that the proposal presented in this study attempts to fill.

The study has two main objectives. The first objective is to determine whether introducing variables responsible for accounting for the influence of submarkets into multiple regression models improves the accuracy of the models. The second objective of the study is to present a several-step computational process, the application of which enables the determination of property submarkets using a geographically weighted regression model, when the properties under study are described using nominal or ordinal scale data. The use of the analytical method, as opposed to expert methods, is characterized by an increased level of objectivity, which may contribute to more accurate conclusions from studies that take into account the subdivision of property markets.

2. Literature review

The topic of determining and analyzing real estate submarkets has been present in scientific literature for several decades. Seminal papers on this topic are (Schnare & Struyk, 1976; Watkins, 2001). Thibodeau and Goodman (2007) state that a housing submarket may be defined as a collectivity of buyers and sellers with a distinct pattern of price-attribute valuations, which results in geographical areas with constant marginal prices. In other words, submarkets can be defined as specific types of properties, located close together, that are close substitutes for potential buyers. Keskin and Watkins (2016) argue that there are three main benefits of dividing real estate markets into submarkets. They state that statistical models will exhibit greater predictive accuracy if housing units have been assigned to submarkets as a prior step in the estimation procedure. Secondly, submarkets offer a useful framework for policymakers and planners to explore dynamic changes in the housing system. Lastly, an understanding of submarket structures can help improve the decision making of a variety of real estate market stakeholders. Specifically, this can assist housing consumers to understand and minimize search costs. It is also worth noting that real estate segmentation can take place at different levels. Micro and macro levels can be distinguished here (Tomal, 2021). As the author indicates, the division of the housing market is also extremely important for real estate entrepreneurs, both when it comes to the selection of investment location and identifying similar housing markets to ensure the accuracy of property valuation. Analyzing real estate markets for submarkets is a dynamic issue. Over the course of time, new factors emerge that may influence changes in real estate market segmentation. Such factors may be related to various economic and social aspects, and, in recent years, even in terms of public health. Tomal and Helbich (2022) evaluate the impact of the COVID-19 pandemic on real estate submarkets in Cracow, Poland. They state that the emergence of the coronavirus reshaped the residential rental market in three ways: rents were decreased, the underlying rental price-determining factors changed, and the spatiotemporal submarket structure was altered.

Research distinguishes different approaches in delineating real estate submarkets. The most common classification divides the methods of classifying properties into submarkets into *a priori* and data driven methods. Usman et al., (2020) also add a mixed approach. to these two entirely different approaches. *A priori* methods are based on existing divisions of analyzed areas. Such existing areas may be aggregated census blocks, zip codes, local government boundaries or physical features (Wu & Sharma, 2012). Examples of such studies can be found in many works. In one such study, Hwang (2015) analyzes real estate markets in St. Louis and Cincinnati. The most commonly analyzed markets in real estate market segmentation studies are residential properties. However, other types of real estate have also been studied. The application of the *a prori* approach to the commercial real estate market has been presented by (Chen & Hao, 2010; Usman et al., 2021). They proved that accounting for the submarket effect in a market-wide model improves model fit and reduces model errors. In turn Beracha et al. (2018) attempted to segment the hotel market. However, it is pointed out that, despite some advantages, such methods do not take into account the influence of socio-economic factors on the formation of real estate submarkets; hence, the increasingly popular use of data driven methods (Helbich et al., 2013). In the case of data driven methods, their use is usually based on two types of property data. The first group is data about individual properties and the qualities of their surroundings. The second group is based on data that indirectly indicates the existence of submarkets. An example of such data aretransaction prices. The use of this type of data to create submarkets has been implemented in Chile (Cox & Hurtubia, 2020). In addition to the dichotomy of data on which the indication of real estate submarkets is based, there is also a different view as to whether submarkets should be spatially integrated and spatially continuous; some researchers take this stance(Bourassa et al., 2007, 2010). A popular tool used in grouping properties into submarkets is the k-means method. An application of this method, along with a proposal to incorporate fuzzy set theory, is presented in the paper (Gabrielli et al., 2017). Another solution proposed for regionalization, i.e. combining areas into larger similar ones, is the SKATER algorithm (Assunção et al., 2006). According to authors - Spatial 'K'luster Analysis by Tree Edge Removal is an efficient method for regionalization of socio-economic units represented as spatial objects, which combines the use of a minimum spanning tree with combinational optimization techniques. It is worth mentioning here that the notions of clusters and submarkets are treated as the same in research on real estate market.

As Kopczewska (2021) indicates, a popular approach to clustering points is the use of β coefficients of geographically weighted regression models. These types of models are widely used both in the real estate market (Cellmer et al., 2020; Kestens et al., 2006) as well as other fields, e.g. the hotel services market (Soler & Gemar, 2018; Zhang et al., 2011). A similar study scheme to the one used in this study was conducted by Kopczewska and Ćwiakowski (2021). In their study, they focused on examining the temporal and spatial stability of the housing submarkets in Warsaw. Some difficulty in using geographically weighted regression models for real estate market segmentation is the fact that properties are usually described by many characteristics, which results in a series of β coefficients for each explanatory variable after the model is estimated. To determine submarkets, such sets of coefficients are clustered by various methods - k-means, hierarchical clustering and others. In turn, dimensionality reduction, for example, is used to visualize the submarkets. However, each of these techniques assumes that there are no problems with the estimation of GWR equations.

There are, however, situations in which real estate market is described by qualitative characteristics, which are introduced to models as dummy variables. The specificity of GWR models is such that, for each analyzed point, a "window" is created and local model is estimated on the basis of only those points, which are located in this window. If the window contains points for which some of the variables do not show variability, then the model cannot be estimated. Converting categorical variables to a single synthetic variable in the pre-estimation stage of the GWR model makes the model estimable for any window. A second advantage of the proposed approach is that there is no need for β-parameter clustering methods.

3. Data and methods

The study encompassed 252 land plots located in one of the largest cities of Poland - Szczecin. The location of the city of Szczecin in Poland is presented in Figure 1. The data set used here was the subject of a study on the problem of mass valuation of real estate (Dmytrów et al., 2020; Doszyń, 2020).

Table 1

Descriptive statistics in unit values (in PLN - Polish zlotys) of real properties and their attributes defined for a set of 252 real properties

Statistics	Values of 1m²	Neighborhood	Transport availability	Physical properties	Plot area
Min	502.11	1	1	1	1
Q_1	569.37	3	2	2	2
M	591.29	3	3	3	3
Q_3	621.73	3	3	3	3
Max	701.43	4	3	3	3
Q	26.18	0	0.5	0.5	0.5
V_Q (%)	4.43	0	16.667	16.667	16.667

Source: own elaboration.

Fig. 1. Location of the city of Szczecin within Poland. *Source*: own elaboration.

Table 2 shows the property attributes listed in the dataset used and their respective categories. All property attributes have been described on an ordinal scale, which is the same as the property description used in the valuation process by appraisers in Poland. Even such an attribute as the area of the real estate is transferred from the quotient scale to the ordinal scale.

Table 2

Real estate attributes and their variants

No.	Attribute	Attribute category (state)
1	Neighborhood	Onerous, Unfavorable Average Favorable
2	Transport availability	Unfavorable Average Favorable

3	Physical plot properties	Unfavorable Average Favorable
4	Plot size	Large (>1200 m²) Average (500 - 1200 m²) Small (<500 m²)

Source: own elaboration

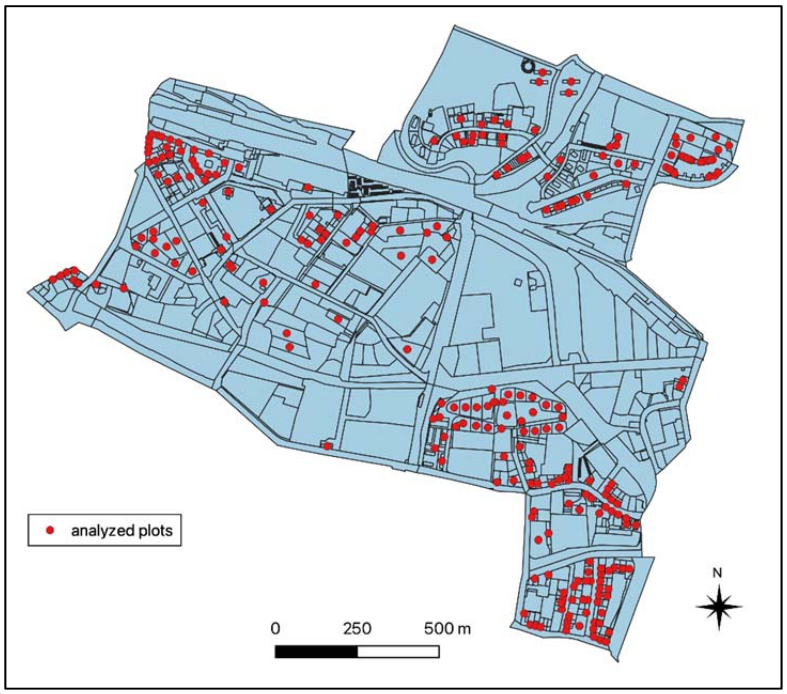

Fig. 2 Study area with indication of land parcels subject to valuation. *Source*: own elaboration.

Conducting the study required several computational steps. In the first stage, categorical variables were transformed into a single synthetic variable, which represents the characteristics of real estate in a combined manner. In the second stage the GWR model was estimated for synthetic variable and real estate value. The values of beta parameter were interpolated by means of the IDW technique to the whole analyzed area after which it was divided into three sub-markets. Sub-market membership was assigned to individual properties as a dummy variable in model (5). The model was estimated three times. Once without variables indicating submarket membership. The second time the model was estimated with variables for submarkets whose boundaries were determined by the real estate appraiser. The last estimation concerned the model in which submarkets were determined by the GWR model.

Because of the fact that variables describing properties in this study are qualitative in nature, their transformation into synthetic variable is difficult. In order to do that, GDM2 distance measure was used (Walesiak & Dudek, 2010), which was designed for qualitative variables. It is based on the generalized correlation coefficient, comprising the Pearson correlation coefficient and τ Kendall correlation coefficient. It is calculated in accordance with the following formula:

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

where:

d_{ik} – distance measure,

$i, k, l = 1, 2, \ldots, n$ – structure number,

$j = 1, 2, \ldots, m$ – variable number.

If variables are measured in the ordinal scale, then the only possible describable relations include elevation relations. Thus, values a and b in Formula (1) are computed in the following fashion:

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

It needs to be mentioned that the variables used for calculating the generalized distance measure can be assigned with weights. However, in the study it was assumed that all the variables (the attributes describing a real estate) are of the same weight. The generalized distance measure can be applied in a multivariate statistical analysis to construct a composite measure in linear ordering methods.

The values of the synthetic variable were defined as the opposite of the distance from the benchmark, i.e. the property with the best variants of all the property characteristics considered. The better the combination of features, the smaller the distance from the benchmark and the higher the value of the synthetic variable:

$$s_{ik} = -d_{ik}, \tag{3}$$

where:
s_{ik}–synthetic variable.

GWR is an adaptation of the local regression model in spatial econometrics, first formulated in the famous work of (S. A. Fotheringham et al., 2002). The idea of local regression is to estimate value at a given point based on its immediate surroundings (nearest neighbors). The closest neighbors are determined based on the distance from the point at which the estimation is carried out. While in the classical approach, the distance between points is determined on the basis of the values of explanatory variables (X matrix), a more intuitive approach can be applied in geographical weighted regression, and the distance between the location of two points in geographical space is calculated. The most important advantages of GWR are: flexibility (can adjust the individual model even to small clusters of points, although this has some negative consequences, which will be discussed later), the lack of a need to define segment affiliation in advance (the model itself chooses the optimal number of nearest neighbors) and the simplicity of the functional form - all equations are linear, meaning that the interpretation of each of the effects is possible and straightforward (Kopczewska, 2020). In this model, a separate equation and distinct coefficients are estimated for each observation. In addition, each observation enters the equation with the weight, calculated on the basis of a certain function assuming the inversely proportional relationship of the weights and distance from the point. In this simple way, the a-spatial model can be used to model spatial heterogeneity, that is, spatial non-stationarity of parameters:

$$Y_i = \beta_{i0} + \sum_{j=1}^{k} \beta_{ij} x_{ij} + u_i, \quad i = 1, \ldots, n \tag{4}$$

where Y_i is the target variable, β_{i0} is the constant term intercept coefficient at location i, x_{ij} *is the j-th* explanatory variable at location i, β_{ij} is the *j-th* local regression coefficient for the *j-th* explanatory variable at location i, and u_i is the random error term associated with location i. (Oshan et al., 2019) point out that i is typically indexed by two-dimensional geographic coordinates, (u_i , v_i), indicating the location of the regression point. As Kopczewska states, the GWR model requires that several problems be solved:

- selecting the functional form to weigh the observation,
- determining by cross-validation or information criteria the optimal range (bandwidth) for the weighing function,
- solving the problem of collinearity of variables in local equations;
- determining methods for calculating the distance between observations.

In order to estimate structural coefficients of the GWR model it is necessary to select a distance-weighting scheme. This involves first selecting a kernel function and kernel type. Next, the bandwidth parameter that controls the intensity of the weighting performed by the kernel must be preselected. Many implementations of GWR estimation, i.e., Python MGWR library, offer tools for bandwidth optimization. This library, as default, utilizes Akaike information criterion (*AIC*) as a model fit criterion. The optimal bandwidth is selected through trials (A. S. Fotheringham et al., 2017). In each trial, a bandwidth is selected, after which GWR is fitted using the bandwidth, and finally a goodness-of-fit measure such as *AIC* is calculated. The optimal bandwidth is the one that minimizes *AIC*. Finally, the model parameters can be estimated along with several diagnostics.

In the survey, a non-linear multiple regression model constitutes a point of reference:

$$ln(w_{ji}) = \alpha_0 + \sum_{k=1}^{K}\sum_{p=2}^{k_p}\alpha_{kp}x_{kpi} + \sum_{j=2}^{J}\alpha_j laz_{ji} + u_i, \quad (5)$$

where:
w_{ji} - unit market value of *i-th* real estate in *j-th* location attractiveness zone,
N - number of real estate properties ($i = 1, 2, \dots, N$),
J - number of location attractiveness zones ($j = 2, 3, \dots, J$),
α_0 - constant term,
K - number of real estate attributes,
k_p - number of states of *k-th* attribute,
α_{kp} - impact of *p-th* state of attribute *k*,
x_{kpi} - dummy variable for *p-th* state of attribute *k*,
α_j - market value coefficient for *j-th* location attractiveness zone,
laz_{ji} - dummy variable equal one for *j-th* location attractiveness zone,
u_i - random component.

The dependent variable is a natural logarithm of a real estate unit value. Real estate values are determined by certified appraisers in individual appraisals. Real estate attributes are qualitative characteristics measured on an ordinal scale, thus they are introduced into the model (5) through dummy variables for each state of an attribute.

In model (5), there is a constant term. In order to avoid strict collinearity of the explanatory variables, each dummy variable for the worst attribute state is skipped. Hence, the summation of $p = 2, \dots, k_p$ in Formula (5). In the interpretation, the ignored state of an attribute serves as a point of reference for the remaining states.

The most important part of the study involves comparing valuation errors obtained with the use of model (5) and other models. In each case, once model valuations (obtained with the application of a model) have been computed, their error was determined by comparing property appraisers' valuations with the results achieved with regression models. The error is a relational (relative) root mean square error (*rRMSE*):

$$RMSE = \sqrt{\frac{\sum_{j=1}^{n}(\hat{w}_j - w_j)^2}{n}} \quad (6)$$

$$rRMSE = \frac{RMSE}{\bar{w}_j}, \quad (7)$$

where:
w_j - actual property value defined by a property surveyor,
$\hat{w}_j$ - theoretical property value,
$\bar{w}_j$ - mean actual property value,
n - number of real properties,
RMSE - root mean square error,
rRMSE - relational root mean square error.

The error in percentage terms indicates how much valuations obtained from a model differ, on average, from the valuations carried out by property appraisers.

The data set used in the study contains information not on transaction prices, but on real estate values, which were determined by property appraisers in individual valuations. In a short period, transactions may refer to the real properties having attributes that differ very little. Low variability of attributes (explanatory variables) translates into e.g. low effectiveness of econometric model estimators. This problem can be avoided when commissioning the appraisal of real properties of various attribute states since the variance of explanatory variables (attributes) is greater.

4. Empirical results

The study consisted of several stages. In the first stage, a multiple regression model (5) excluding submarket variables (*laz*) was estimated based on the analyzed real estate dataset, which included data on their four characteristics and the value estimated by real estate appraisers. The evaluation of this model provides a benchmark for the results obtained in subsequent stages. In the second stage, a subdivision of the study area into submarkets was conducted. This division was made by real estate appraisers who are familiar with the local specifics of the real estate market. They distinguished three

submarkets. The division indicated by them was incorporated into the model (5) in the form of variables laz_i. In the third step, an alternative way of determining the boundaries of the submarkets was introduced. This proposal involved the following steps:

1. Using the formulas (1-3), a synthetic variable was calculated for each property.
2. Based on the synthetic variable, a GWR model (4) was estimated, whose structural β parameter estimates at each location will be used to indicate property submarkets.
3. *Using IDW* interpolation, a parameter map describing the relationship between the synthetic variable and property value was created.
4. Using the division into three value ranges, submarket boundaries were extracted. The number of submarkets is derived from the number indicated by the experts in stage 2 of the study.
5. As in stage two, the affiliation of properties to particular submarkets was entered into the model (5) in the form of variables laz_i .

The final fourth stage of the study is to compare performance measures (7) of the models estimated in stages one (model without submarkets), stage two (submarkets indicated by experts) and stage three (submarkets indicated according to the proposed method).

Figure 3 shows the boundaries of the three submarkets, which were determined by experts. Three clearly separated areas can be distinguished. In drawing these boundaries, the experts used the existing administrative division.

Fig. 3 Expert division of valued area into submarkets. *Source*: own elaboration.

Figure 4shows the values of the GWR model-estimated estimates of the structural parameters describing the dependence of property values on the synthetic variable. When looking at the parameter values visualized using shades of the purple color, one can see clusters of similar values and it is this fact that underlies the creation of submarkets by the proposed analytical method. An important step in verifying the estimated model is to assess the statistical significance of the slope (β) parameters of the model. The mgwr package (Oshan et al., 2019) of the Python programming language used in the study allows for testing the significance of model parameters. The results of the test conducted indicate that, for each point, the estimated slope parameter was found to be statistically significant.

In order to determine the submarket boundaries, the values of β parameters were interpolated to the whole analyzed area (see Figure 5). Visual analysis shows that it is possible to distinguish sub-areas, in which the dependence of the synthetic variable, describing all property features taken into account, and property values is similar.

The interpolated values of β coefficients were divided into three ranges to correspond in number to the submarkets determined by the experts. This discrete division of the study area into three parts is presented in Figure 6. It shows the boundaries of the separated areas.

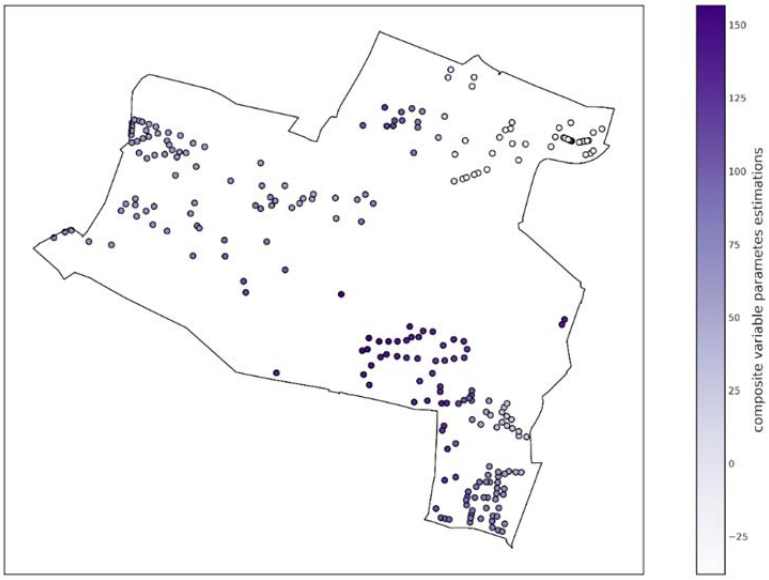

Fig. 4. Values of structural β coefficients estimation of GWR model. *Source*: own elaboration.

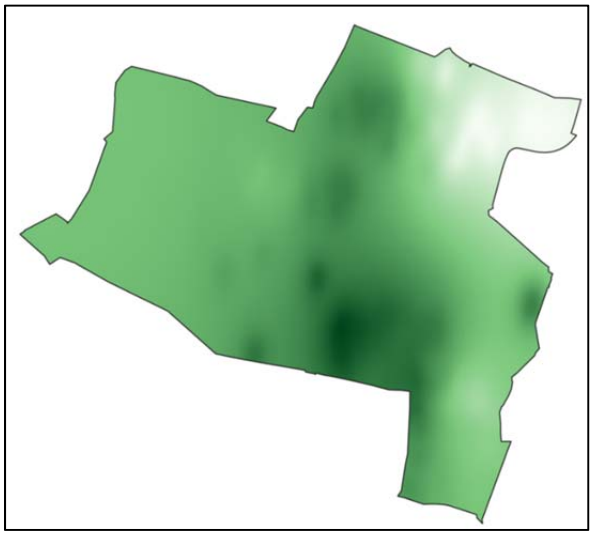

Fig. 5.IDW interpolation of GWR model β coefficients. *Source*: own elaboration

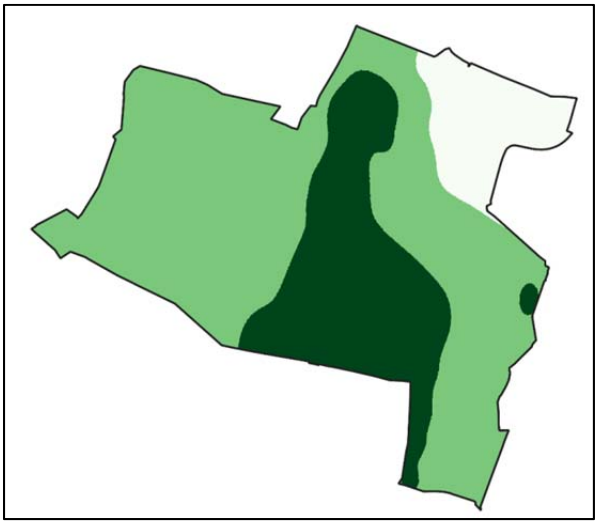

Fig. 6. Sub-markets extracted using GWR model. *Source*: own elaboration

The main differences between such a division and the expert division come down to two aspects. First, the new subdivision is not based on existing administrative divisions. Second, the delineated areas are not fully geographically consistent. One of the submarkets points to two separate locations.
In this way, the necessary data were obtained to estimate models (5) in three variants and compare them. All models were estimated using the least squares method and compared using the Akaike information criterion, the adjusted coefficient of determination and the relative root mean square error (7). The results of the comparison are presented in Table 3.

Table 3

Models performance measures

Model.	measure		
	AIC	*Adj. R^2*	*rRMSE*
No submarkets	-797.5	50.4%	4.82%
Expert submarkets	-872.1	63.6%	4.06%
GWR submarkets	-889.9	66.3%	3.90%

Source: own elaboration.

Table 3 presents selected measures of the quality of the estimated models. All of them show that the introduction of variables accounting for the introduction of submarkets improves the modeling results. The model fit increased from 50.4% to 63.6% using expert disaggregation and to 66.3% with the submarkets singled out by the proposed procedure based on the GDM2 measure and the GWR model. The improvement was thus nearly one third. Analogous conclusions are drawn from the comparison of *AIC* and relative *RMSE* - the rRMSE dropped from 4.82% to 3.90%.

5. Discussion and conclusions

There is no doubt that real estate market segmentation is an issue whose use in research improves the results of modeling and analysis in various areas - price modeling, estimating the effects of property taxation, urban planning and others. Segmenting submarkets can be done in a variety of ways. This study compares the results of property value modeling in three approaches. The first used a multiple regression model that did not account for submarkets. The second approach involved delineating the boundaries of smaller areas by experts whose knowledge of the study area allowed them to indicate where the submarket boundaries should be and the number of submarkets. The last approach was based on a several-step computational procedure that used a geographically weighted regression model and interpolation. An additional element that filled a research gap was the use of the GDM2 distance measure in the GWR model to determine the value of a synthetic variable representing the combined state of all property characteristics included in the study. This synthetic variable acted as an explanatory variable in the model. This approach solved the problem of the lack of variability of some features for some of the points modeled with GWR. The results of the estimated models indicate two facts. First, the introduction of variables in the model that account for the belonging of properties to submarkets improved the models, which meant a decrease in rRMSE. Moreover, the submarkets whose boundaries were determined based on the GWR model and interpolation of its structural parameters gave better value modeling results than the model in which the expert division of the study area was used. The use of a synthetic variable eliminated the problem of the lack of variability of property characteristics. The results of the study confirmed the need to consider submarkets postulated by researchers (i.e. J. Chen & Hao, 2010; Tomal, 2021).

The results presented here are the first step in exploring the feasibility of using a synthetic variable in the GWR model and interpolating its parameters to determine real estate submarkets. Several issues remain to be considered and investigated. Further steps should be taken to study what the optimal number of submarkets should be, whether the β coefficients of the GWR model or its residuals should be interpolated, what interpolation technique works best, and how much data on either transactions or valuations of individual properties is needed for the proposed procedure to still be superior to expert apportionment. This series of issues presented indicates that market segmentation procedures are an important and not yet fully solved research problem.

References

Assunção, R. M., Neves, M. C., Câmara, G., & da Costa Freitas, C. (2006). Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. *International Journal of Geographical Information Science*, *20*(7), 797–811. https://doi.org/10.1080/13658810600665111

Benjamin, J., Guttery, R., & Sirmans, C. (2004). Mass Appraisal: An Introduction to Multiple Regression Analysis for Real Estate Valuation. *Journal of Real Estate Practice and Education*, *7*. https://doi.org/10.1080/10835547.2004.12091602

Beracha, E., Hardin III, W. G., & Skiba, H. M. (2018). Real Estate Market Segmentation: Hotels as Exemplar. *JOURNAL OF REAL ESTATE FINANCE AND ECONOMICS*, *56*(2), 252–273. https://doi.org/10.1007/s11146-017-9598-z

Bourassa, S., Cantoni, E., & Hoesli, M. (2007). Spatial Dependence, Housing Submarkets, and House Price Prediction. *The Journal of Real Estate Finance and Economics*, *35*, 143–160. https://doi.org/10.1007/s11146-007-9036-8

Bourassa, S., Cantoni, E., &Hoesli, M. (2010). Predicting House Prices with Spatial Dependence: A Comparison of Alternative Methods. *Journal of Real Estate Research*, *32*, 139–160. https://doi.org/10.1080/10835547.2010.12091276

Bradbury, K. L., Mayer, C. J., & Case, K. E. (2001). Property tax limits, local fiscal behavior, and property values: evidence from Massachusetts under Proposition 212. *Journal of Public Economics*, *80*(2), 287–311. https://doi.org/10.1016/S0047-2727(00)00081-5

Cellmer, R., Cichulska, A., & Belej, M. (2020). Spatial Analysis of Housing Prices and Market Activity with the Geographically Weighted Regression. *ISPRS International Journal of Geo-Information*, *9*, 380. https://doi.org/10.3390/ijgi9060380

Chen, C. F., & Rothschild, R. (2010). An application of hedonic pricing analysis to the case of hotel rooms in Taipei. *Tourism Economics*, *16*(3), 685–694. https://doi.org/10.5367/000000010792278310

Chen, J., & Hao, Q. (2010). Submarket, Heterogeneity and Hedonic Prediction Accuracy of Real Estate Prices: Evidence from Shanghai. *International Real Estate Review*, *13*, 190–217. https://doi.org/10.53383/100125

Cox, T., &Hurtubia, R. (2020). Subdividing the sprawl: Endogenous segmentation of housing submarkets in expansion areas of Santiago, Chile. *Environment and Planning B: Urban Analytics and City Science*, *48*(7), 1770–1786. https://doi.org/10.1177/2399808320947728

Dmytrów, K., Gdakowicz, A., &Putek-Szeląg, E. (2020). Methods of Analyzing Qualitative Variable Correlation on the Real Estate Market. *Real Estate Management and Valuation*, *28*(1), 80–90. https://doi.org/doi:10.2478/remav-2020-0007

Doszyń, M. (2020). Econometric Support of a Mass Valuation Process. *Folia Oeconomica Stetinensia*, *20*(1), 81–94. https://doi.org/10.2478/foli-2020-0005

Fotheringham, A. S., Yang, W., & Kang, W. (2017). Multiscale Geographically Weighted Regression (MGWR). *Annals of the American Association of Geographers*, *107*(6), 1247–1265. https://doi.org/10.1080/24694452.2017.1352480

Fotheringham, S. A., Brunsdon, C., & Charlton, M. (2002). *Geographically weighted regression*. JOHN WILEY & SONS, LTD.

Gabrielli, L., Giuffrida, S., & Trovato, M. R. (2017). Gaps and Overlaps of Urban Housing Sub-market: Hard Clustering and Fuzzy Clustering Approaches. In S. Stanghellini, P. Morano, M. Bottero, & A. Oppio (Eds.), *APPRAISAL: FROM THEORY TO PRACTICE* (pp. 203–219). SPRINGER INTERNATIONAL PUBLISHING AG. https://doi.org/10.1007/978-3-319-49676-4_15

Helbich, M., Brunauer, W., Hagenauer, J., & Leitner, M. (2013). Data-Driven Regionalization of Housing Markets. *Annals of the Association of American Geographers*, *103*(4), 871–889. https://doi.org/10.1080/00045608.2012.707587

Hwang, S. (2015). Residential Segregation, Housing Submarkets, and Spatial Analysis: St. Louis and Cincinnati as a Case Study. *Housing Policy Debate*, *25*(1), 91–115. https://doi.org/10.1080/10511482.2014.934703

Keskin, B., & Watkins, C. (2016). Defining spatial housing submarkets: Exploring the case for expert delineated boundaries. *Urban Studies*, *54*(6), 1446–1462. https://doi.org/10.1177/0042098015620351

Kestens, Y., Thériault, M., &Rosiers, F. (2006). Heterogeneity in hedonic modelling of house prices: Looking at buyers' household profiles. *Journal of Geographical Systems*, *8*, 61–96. https://doi.org/10.1007/s10109-005-0011-8

Kopczewska, K. (2020). *Applied Spatial Statistics and Econometrics: Data Analysis in R*. https://doi.org/10.4324/9781003033219

Kopczewska, K. (2021). Spatial machine learning: new opportunities for regional science. *Annals of Regional Science*. https://doi.org/10.1007/s00168-021-01101-x

Kopczewska, K., &Ćwiakowski, P. (2021). Spatio-temporal stability of housing submarkets. Tracking spatial location of clusters of geographically weighted regression estimates of price determinants. *Land Use Policy*, *103*. https://doi.org/10.1016/j.landusepol.2021.105292

Korteweg, A., & Sorensen, M. (2016). Estimating Loan-to-Value Distributions. *Real Estate Economics*, *44*(1), 41–86. https://doi.org/https://doi.org/10.1111/1540-6229.12086

Manjula, R., Jain, S., Srivastava, S., & Rajiv Kher, P. (2017). Real estate value prediction using multivariate regression models. *IOP Conference Series: Materials Science and Engineering*, *263*(4). https://doi.org/10.1088/1757-899X/263/4/042098

Oshan, T. M., Li, Z., Kang, W., Wolf, L. J., & Stewart Fotheringham, A. (2019). MGWR: A python implementation of multiscale geographically weighted regression for investigating process spatial heterogeneity and scale. *ISPRS International Journal of Geo-Information*, *8*(6). https://doi.org/10.3390/ijgi8060269

Santos, J. A. C., Fernández-Gámez, M., Solano-Sánchez, M. Á., Rey-Carmona, F. J., del Rio, L., & Caridad, L. (2021). Valuation models for holiday rentals' daily rates: Price composition based on booking.com. *Sustainability (Switzerland)*, *13*(1), 1–16. https://doi.org/10.3390/su13010292

Schnare, A. B., &Struyk, R. J. (1976). Segmentation in urban housing markets. *Journal of Urban Economics*, *3*(2), 146–166. https://doi.org/https://doi.org/10.1016/0094-1190(76)90050-4

Sirmans, G. S., Macpherson, D. A., &Zietz, E. N. (2005). The Composition of Hedonic Pricing Models. *Journal of Real Estate Literature*, *13*(1), 3–43. http://www.jstor.org/stable/44103506

Soler, I. P., & Gemar, G. (2018). Hedonic price models with geographically weighted regression: An application to hospitality. *Journal of Destination Marketing and Management*, *9*, 126–137. https://doi.org/10.1016/j.jdmm.2017.12.001

Thibodeau, T., & Goodman, A. (2007). The Spatial Proximity of Metropolitan Area Housing Submarkets. *Real Estate Economics*, *35*, 209–232. https://doi.org/10.1111/j.1540-6229.2007.00188.x

Tomal, M. (2021). Housing market heterogeneity and cluster formation: evidence from Poland. *International Journal of Housing Markets and Analysis*, *14*(5), 1166–1185. https://doi.org/10.1108/IJHMA-09-2020-0114

Tomal, M., & Helbich, M. (n.d.). The private rental housing market before and during the COVID-19 pandemic: A submarket analysis in Cracow, Poland. *Environment and Planning B: Urban Analytics and City Science*, *0*(0), 23998083211062908. https://doi.org/10.1177/23998083211062907

Tzioumis, K. (2017). Appraisers and Valuation Bias: An Empirical Analysis. *Real Estate Economics*, *45*(3), 679–712. https://doi.org/https://doi.org/10.1111/1540-6229.12133

Usman, H., Lizam, M., & Adekunle, M. U. (2020). PROPERTY PRICE MODELLING, MARKET SEGMENTATION AND SUBMARKET CLASSIFICATIONS: A REVIEW. *REAL ESTATE MANAGEMENT AND VALUATION*, *28*(3), 24–35. https://doi.org/10.1515/remav-2020-0021

Usman, H., Lizam, M., & Burhan, B. (2021). A PRIORI SPATIAL SEGMENTATION OF COMMERCIAL PROPERTY MARKET USING HEDONIC PRICE MODELLING. *REAL ESTATE MANAGEMENT AND VALUATION*, *29*(2), 16–28. https://doi.org/10.2478/remav-2021-0010

Walesiak, M., & Dudek, A. (2010). Finding Groups in Ordinal Data: An Examination of Some Clustering Procedures. In *Studies in Classification, Data Analysis, and Knowledge Organization* (pp. 185–192). https://doi.org/10.1007/978-3-642-10745-0_19

Watkins, C. A. (2001). The Definition and Identification of Housing Submarkets. *Environment and Planning A: Economy and Space*, *33*(12), 2235–2253. https://doi.org/10.1068/a34162

Wu, C., & Sharma, R. (2012). Housing submarket classification: The role of spatial contiguity. *Applied Geography*, *32*(2), 746–756. https://doi.org/10.1016/J.APGEOG.2011.08.011

Zhang, H., Zhang, J., Lu, S., Cheng, S., & Zhang, J. (2011). Modeling hotel room price with geographically weighted regression. *International Journal of Hospitality Management*, *30*(4), 1036–1043. https://doi.org/10.1016/j.ijhm.2011.03.010