# Spatial Segmentation of Urban Housing Markets: A Case Study of Minsk

Milvari Alieva [1*], Natallia Zhukouskaya [2]

[1]Belarusian State University, Faculty of Geography and Geoinformatics, Minsk, Belarus - m.alieva5030@gmail.com
[2]Adam Mickiewicz University, Faculty of Human Geography and Planning, Poznan, Poland - natazhuk@gmail.com

**Keywords:** spatial econometrics, Moran's I, Geographic Weighted Regression, submarkets, Minsk housing market

**Abstract**

This study carried out a spatial segmentation of the urban housing market based on a combination of sophisticated data-driven techniques. Such delineation is essential for enhancing the accuracy of mass appraisal and guiding effective urban planning policies. The residential real estate market of Minsk serves as the case study. The analysis relied on a dataset of approximately 4,600 offer prices for secondary market apartments in 2017, sourced from the real estate platform Realt.by. A multi-stage methodological workflow was proposed. An initial evaluation using the Global Moran's I index (I = 0.39) and Local Indicators of Spatial Association (LISA) confirmed significant price clustering. To address dimensionality and multicollinearity among spatial predictors, Spatial Principal Component Analysis (sPCA) was employed, reducing ten infrastructure variables to four interpretable latent components representing centrality, environmental quality, social-industrial balance, and transport accessibility. Subsequently, these derived spatial factors as well as structural property attributes (such as area, floor level, and room count) were used as inputs for a Geographically Weighted Regression (GWR) model. This specification demonstrated substantial explanatory power ($R^2 = 0.58$) and successfully accounted for spatial heterogeneity, eliminating residual autocorrelation. Finally, the local GWR coefficients were grouped using k-means clustering, delineating three distinct submarkets with unique pricing mechanisms: a Central Urbanized zone, driven primarily by factors such as centrality and the number of floors; a Developed Middle-Ring submarket, influenced mainly by property attributes including room count and construction year; and a Modern Peripheral submarket shaped strongly by construction year and the "Centrality and Prestige" component.

## 1. Introduction

Urban housing markets are inherently complex and heterogeneous, influenced by a wide array of interconnected factors, including the different types of properties available, the diverse socio-economic profiles of its inhabitants, and the unique physical and environmental characteristics of locations within a city. The concept that housing markets are not unitary but are composed of distinct submarkets is well-established in the literature (Fletcher et al., 2000; Bourassa et al, 2003). Methodological approaches to delineating these submarkets have evolved from traditional a priori classifications based on predefined boundaries to sophisticated data-driven techniques (Chen et al., 2023). These modern approaches employ empirically derived segmentation using statistical tools such as K-means clustering, neural networks, and geostatistical models (Kauko et al., 2002; Wu and Sharma, 2012). Understanding the distinct structures of their submarkets is crucial for enhancing the accuracy of property valuation and market prediction (Morawakage et al., 2021). Such insights are vital for urban planning and guiding effective taxation policies.

The research addresses this need by segmenting the residential real estate market of Minsk. The formation of the private housing market in the city began in the early 1990s, evolving through distinct phases. Today, the city's residential sector represents a distinct environment where this historical context intersects with contemporary market forces, creating complex spatial price patterns. Traditional global hedonic models often fail to capture local variations in price determinants, a phenomenon known as spatial non-stationarity. Geographically Weighted Regression (GWR) effectively overcomes this limitation by allowing regression coefficients to vary across space, thereby capturing local heterogeneity. Some research confirms that GWR provides higher levels of accuracy in comparison with other models for mass appraisal (Lockwood and Rossini, 2011; Dimopoulos and Moulas, 2016) and is effective for market segmentation (McCluskey and Borst, 2011). Consequently, we employ an integrated framework combining Geographically Weighted Regression with cluster analysis, a prominent approach for identifying spatially coherent submarkets (Bidanset et al., 2024). Our primary innovation extends these established methods (Kopczewska and Ćwiakowski, 2021) through the prior application of a spatial Principal Component Analysis (sPCA) to the predictor variables to create a more robust and interpretable model.

Consequently, this study makes two primary contributions:

- Development of an integrated sPCA - GWR framework to create statistically valid valuation zones;
- Identification of three statistically validated submarkets of Minsk via the clustering of GWR coefficients, revealing distinct mechanisms of price formation.

## 2. Data and Methodology

### 2.1. Real Estate Data and Pre-Processing

The study utilized data obtained from Realt.by (Realt.by, 2017), one of the largest real estate listing platforms in Belarus. The dataset contains offer prices for secondary-market multi-apartment residential properties in Minsk recorded in 2017. The year 2017 was selected because it represents the last stable pre-pandemic period of the Minsk housing market, unaffected by

* Corresponding author

later shocks. This makes it a reliable baseline for analysing structural spatial regularities without distortions introduced in later years.

The initial dataset consisted of approximately 4,600 housing listings, each described by a set of attributes including geographic coordinates, normalized price (USD/m²), and physical characteristics such as total area, number of rooms, apartment floor level, year of construction, and wall material.

A multi-stage data cleaning process was carried out to ensure analytical consistency. First, observations with missing key values were removed. Next, duplicate records—identified based on a combination of primary attributes and geometric location—were excluded. Extreme outliers in the dependent variable (normalised price per m²) were filtered out using distribution-based thresholds. The categorical attribute *"wall material"* was transformed into a numerical variable based on its physical indicator of thermal conductivity, ensuring comparability across building types.

### 2.2. Spatial Data Enrichment

To construct the necessary locational and environmental variables for spatial analysis each residential property was further enriched with attributes describing its accessibility to social and transport infrastructure as well as characteristics of the surrounding urban environment.

OpenStreetMap (OpenStreetMap Contributors, 2017), an open and collaboratively maintained spatial database, served as the primary source of all urban infrastructure objects.

Spatial integration was performed using SQL query within a PostgreSQL database containing OpenStreetMap (OSM) data for the city of Minsk (Figure 1). Using the PostGIS extension, a table of minimum Euclidean distances between each residential property and selected types of urban amenities was generated and subsequently joined to the main dataset via a unique property identifier. As a result, every property was assigned a set of distance-based indicators reflecting its accessibility to essential facilities.

```
SELECT  st_distance(pr.geometry,c.geometry) as centre_dist
FROM centre c
ORDER BY pr.geometry <-> c.geometry
LIMIT 1
```

Figure 1. Part of SQL query used to compute Euclidean distances in PostGIS (pr is an alias for table "Properies")

In addition to accessibility metrics, each property was enriched with several urban-environment attributes, including population density, building density, and road network density.

Population estimates were sourced from the Kontur Population Dataset, an openly accessible model published on the Humanitarian Data Exchange platform (Kontur Inc., 2019). Building and road densities were computed from OpenStreetMap data.

All continuous spatial variables were aggregated using a regular H3 hexagonal grid (resolution 9), which provides a consistent spatial scale for population, build environment, and road network indicators (Uber Technologies Inc., 2018).

The final enriched dataset contains approximately 4,635 observations, each comprising *18* structural, locational, and environmental attributes together with corresponding geographical coordinates.

### 2.3 Exploratory Diagnostics

To quantitatively assess multicollinearity among the potential predictors, Variance Inflation Factors (VIFs) were calculated for all explanatory variables.

None of the variables exhibited a VIF value greater than 3, indicating the absence of problematic multicollinearity. Therefore, all predictors were retained for further modelling.

### 2.4 Application of Spatial Econometric

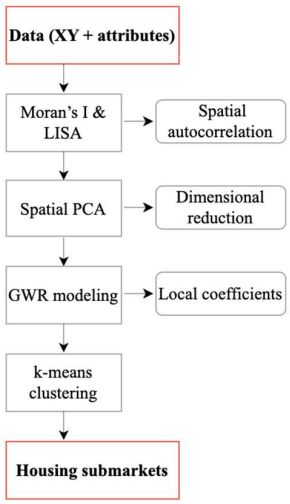


Figure 2. Workflow of the proposed spatial segmentation method (sPCA + GWR + clustering).

The analytical framework relied on spatial econometric methods enabling the examination of spatial phenomena, such as *spatial heterogeneity* and *spatial dependence* (Anselin, 1988), within the urban territory and the identification of price formation patterns characteristic of different housing submarkets.

The study's methodological framework consisted of a sequence of interrelated stages (Figure 2).

The first step is testing for spatial autocorrelation. The dependent variable – normalized price (USD/m²) – was tested for the global spatial autocorrelation using the Moran's I. The results indicated a positive and statistically significant spatial autocorrelation of offer prices ($I = 0.39$; $p = 0.01$), suggesting clear market clustering and justifying the application of spatial variants of statistical methods in subsequent analysis.

The Local Indicators of Spatial Association (LISA) were computed to analyze the spatial structure of the housing supply and to identify spatial clusters representing distinct market segments.

To further account for the spatial context, a Geographically Weighted Regression (GWR) model was constructed using the full set of explanatory factors.

For each observation, the model can be expressed as follows:

$$\log_e(P_i) = \beta_0 + \sum_{k=1}^{17} \beta_k X_{ki} + \varepsilon_i, \qquad (1)$$

where $\log_e(P_i)$ = natural logarithm of property price/m²
$\varepsilon_i$= random error term
$X_i$= ( $X_{1i}$, … $X_{17i}$) = explanatory variables including: number of rooms, total area, year of construction,

wall material, number of floors, population density, building density, road network density, distance to city center, distance to parks, distance to metro stations, distance to hospitals, distance to schools, distance to public transport stops, distance to highways, distance to industrial zones, and distance to entertainment facilities.

The model performance was evaluated using three metrics: The Akaike Information Criterion (AIC), the coefficient of determination ($R^2$), and the spatial autocorrelation of residuals.
The residuals exhibited statistically significant spatial autocorrelation ($p < 0.05$), indicating specification insufficiency or redundancy in the initial variables set.
To address the issue, a spatial Principal Component Analysis (sPCA) was applied on a subset of independent variables describing accessibility to urban infrastructure. Unlike standard PCA, sPCA accounts for the spatial arrangement of observations, ensuring that the resulting components capture not only variance in the variables themselves but also spatial patterns across the analyzed territory.
The subsequent GWR modeling stage incorporated a reduced set of predictors derived from the sPCA results:

$$\log_e(P_i) = \beta_0 + \sum_{k=1}^{12} \beta_k X_{ki} + \varepsilon_i, \qquad (2)$$

where $\log_e(P_i)$ = natural logarithm of property price/m$^2$
$\varepsilon_i$= random error term
$X_i$= ( $X_{1i}$, … $X_{12i}$) - explanatory variables including: number of rooms, total area, year of construction, house type, number of floors, population density, building density, road network density, and four spatial principal components (sPC1–sPC4) derived from spatial accessibility and environmental indicators.

Residuals from this reduced model no longer exhibited statistically significant spatial autocorrelation, indicating that factor compression using sPCA effectively captured the spatial structure present in the data. A comparison of the two GWR models – one estimated with the full set of predictors and another based on the sPCA-reduced components – shows that dimensionality reduction did not affect explanatory power, while improving model stability and eliminating residual spatial autocorrelation.
Finally, the resulting local coefficients from the second GWR model were clustered via the k-means algorithm to delineate homogeneous spatial groups reflecting similar patterns of price determinants.

### 2.5 Technical basis

The raw dataset was retrieved from the Realt.by platform using a web-scraping approach implemented in Python via *Beautiful Soup* library.
Analysis was conducted primarily in Python using the following frameworks: *pandas*, *geopandas*, and *numpy* for working with tabular and geospatial datasets; *osmnx* for retrieving and processing OpenStreetMap data; *SQLAlchemy* for database connectivity; *libpysal*, *esda*, *mgwr*, and *spreg* for spatial statistical analysis; and *matplotlib* and *seaborn* for visualization.
Prior to spatial analysis, OpenStreetMap data for the city of Minsk and the scraped listing dataset were imported into a PostgreSQL database with the PostGIS extension via *SQLAlchemy* enabling efficient spatial querying and distance calculations. Minimum Euclidean distances were computed through SQL query executed directly in this geographic database. All calculations, spatial operations, and data transformations were performed within a single coordinate reference system – WGS 84 / UTM zone 35N (EPSG:32635) – a projected CRS suitable for the central part of Belarus, where Minsk is located.

## 3. Results and Discussion

### 3.1 Descriptive statistics of the Dependent Variable

To characterize the fundamental properties of the studied phenomenon, descriptive statistics were calculated for the dependent variable — normalized price (USD/m²). The distribution of normalized prices (Figure 3) exhibits right-skewness, with the upper tail formed by high-priced listings.

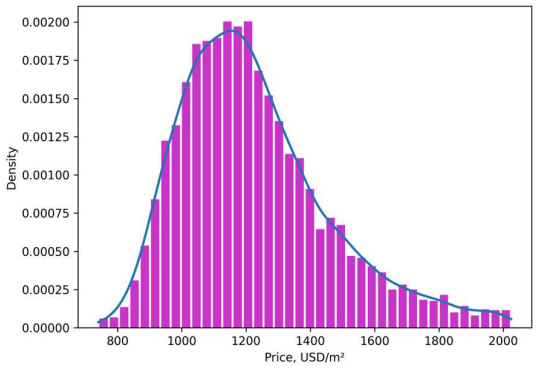


Figure 3. Distribution of normalized price per square meter.

Prices range approximately from 741 to 2,024 USD/m², with a mean of 1,227 USD/m² and a median of 1,189 USD/m². The close values of the mean and median indicate that extreme high-priced listings have some influence but do not dominate the central tendency of the distribution.
Spatial distribution of prices (Figure 4) demonstrates a definite territorial heterogeneity: higher values concentrate in central districts, extending toward the north-western part of the city, while peripheral areas are characterized by lower price levels.

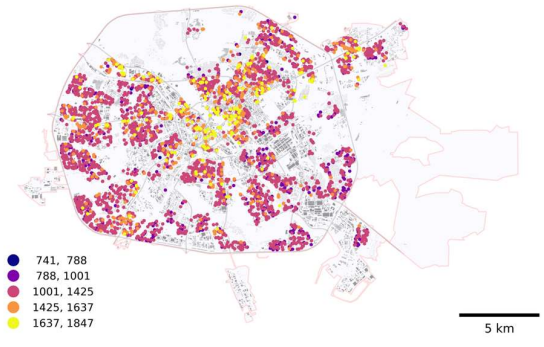


Figure 4. Distribution of normalized price per square meter (method Standard Deviation).

Nevertheless, positive right skewness (skewness=0.89) justifies the use of a logarithmic transformation at subsequent modelling

stages, as it reduces the influence of extreme values and makes the distribution closer to normal.

### 3.2 Spatial Autocorrelation (Moran's I and LISA)

Moran's I indicated a positive and statistically significant spatial autocorrelation of offer prices ($I = 0.39$; $p = 0.01$), suggesting a *pronounced* spatial structure and thereby validating the use of spatially explicit methods.
The LISA-analysis identified four types of local spatial structures: stable clusters of high-prices properties (*High-High*) concentrated in central part of the city, extensive zones of low-cost mass housing (*Low-Low*) on the periphery, as well as two forms of spatial outliers (*High-Low and Low-High*).
Local indicators segmentation not only delineates the price clusters but makes it possible to identify potentially undervalued or overvalued properties. According to the analysis, the number of overvalued listings (n = 338) exceeds that of undervalued ones (n = 251) (Figure 5).

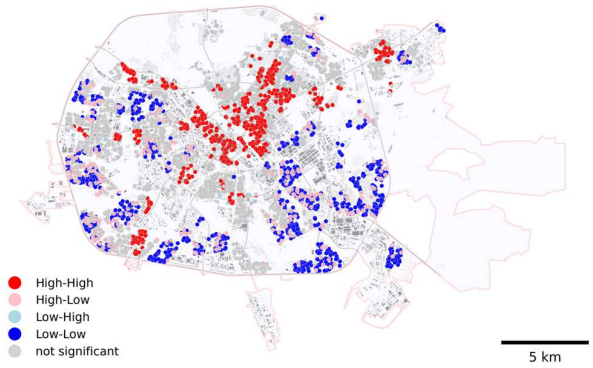


Figure 5. LISA clusters.

### 3.4 Full-Variable GWR

According to specification (1), the initial model incorporated 17 structural, locational and environmental predictors representing property attributes, distances to key infrastructure, and urban density indicators. The full-variable GWR model demonstrated a solid performance:

$$AIC = -5746.8;$$
$$R^2 = 0.574.$$

These results indicate that the model explains a substantial share of the spatial variation in housing prices across the urban area and that the estimated coefficients exhibit meaningful spatial variation, capturing localized market effects.
However, despite the satisfactory overall fit, Moran's statistic computed on the residuals revealed statistically significant spatial autocorrelation ($p < 0.05$) suggesting that the initial specification did not fully account for spatial structure of the price-forming process.

### 3.5 Spatial PCA (sPCA) Reduction

Spatial Principal Component Analysis (sPCA) was applied to reduce the dimensionality of the set of spatially referenced variables describing accessibility to urban infrastructure.

Four principal components were retained based on multiple criteria, including: the proportion of explained variance, the stability of the eigenspectrum (scree-plot behaviour), and the interpretability of the resulting latent factors (Table 1).
Collectively, the first four components account for 66.41% of the total variance, which is sufficient for capturing the major spatial patterns embedded in the original structure of indicators.
To enhance interpretability, a loading matrix was examined to access the distribution of each original variable to the latent components (Table 1). Loadings with statistically significant values are shown in **bold**.

| Distances to | sPC1 | sPC2 | sPC3 | sPC4 |
|---|---|---|---|---|
| City center | **−0.52** | **0.29** | −0.02 | −0.08 |
| Parks | −0.10 | **0.43** | −0.41 | 0.44 |
| Metro | **−0.41** | −0.27 | 0.14 | 0.22 |
| Hospitals | **−0.32** | **−0.50** | 0.18 | −0.19 |
| Schools | −0.14 | −0.19 | **−0.48** | 0.09 |
| Transport | **−0.31** | **0.56** | 0.44 | −0.24 |
| Highways | **−0.31** | −0.08 | 0.17 | **0.68** |
| Industrial zones | **−0.36** | 0.01 | **−0.55** | −0.40 |
| Entertainment facilities | **−0.31** | **−0.17** | 0.04 | −0.08 |
| Explained variance, % | 30.50 | 14.92 | 10.72 | 10.26 |

Table 1. Principal component analysis (PCA). Loadings of variables.

Based on the loading patterns, the four components can be interpreted as follows:

1. *Centrality and prestige (sPC1)* − primarily determined by the influence of all distances: to city center, building density and proximity to highways, etc.
2. *Quality of environment and accessibility of green infrastructure (sPC2)* − reflecting the influence of the presence of parks and recreational areas and the same time social amenity accessibility.
3. *Balance of social infrastructure and industrial facilities (sPC3)* − capturing the spatial relationship between educational and healthcare institutions versus industrial zones;

4. *Transport accessibility and street network connectivity (sPC4)* − reflecting accessibility of major transport corridors.

These latent dimensions summarize the complex spatial configuration of urban amenities and infrastructure, reducing the number of predictors while preserving the key spatial information. This process not only simplified the model but also provided deeper insights into the underlying spatial structure of urban environmental factors, ensuring capturing of spatial patterns across the city. For example, the "Centrality and Prestige" component captures the classic premium associated with the urban core.

### 3.6 Improved GWR Model (with sPCA Components)

According to specification (2), a second GWR model was estimated using the latent components derived from the spatial PCA, which replaced the original proximity-based variables.
The model retained a high level of performance, comparable to the full-variable specification:

$$AIC = -5702.1;$$
$$R^2 = 0.577.$$

Importantly, the Moran's I statistic computed on the residuals indicated no statistically significant spatial autocorrelation ($p < 0.05$). That confirms solidly that factor compression through sPCA effectively captured the spatial structure present in the initial predictors and removed the residual spatial dependence observed on the full-variable model.
The modeling produced a set of local coefficients representing the influence of each predictor on the price of individual properties (Table 2). Based on the GWR predictor estimates the most influential variables are sPC1, which exhibits considerable spatial variation, and Year of Construction, which shows a consistently strong positive effect across the study area.

| Predictor | Min | Median | Max |
|---|---|---|---|
| Intersept | −32.1 | −31.92 | −31.5 |
| Number of rooms | −0.12 | −0.03 | 0.07 |
| Total area | −0.21 | −0.03 | 0.08 |
| Year of construction | −0.05 | 0.08 | 0.19 |
| Wall material | −0.09 | −0.02 | 0.04 |
| Number of floors | −0.04 | 0.008 | 0.134 |
| Population Density | −0.08 | 0.004 | 0.06 |
| Building Density | −0.06 | −0.003 | 0.029 |
| Road Network Density | −0.049 | 0.002 | 0.09 |
| sPC1 | −0.11 | −0.06 | 0.275 |
| sPC2 | −0.19 | −0.006 | 0.197 |
| sPC3 | −0.13 | −0.02 | 0.06 |
| sPC4 | −0.18 | −0.009 | 0.105 |

Table 2. Summary statistics for GWR estimates

### 3.7 Spatial Segmentation via k-Means

Clustering of the local regression coefficients obtained from the improved GWR model was performed using the k-means algorithm (Figure 6).
This procedure revealed three clearly distinguishable housing submarkets, reflecting the *functional* and *historical radial structure* of the urban environment of Minsk:

1. Central Urbanized Submarket (CBD) − prices are determined by centrality mainly, a positive influence of the number of floors is also observed;
2. Developed middle-ring residential submarket of Soviet and post-Soviet construction − prices are positively affected by the number of rooms and year of construction;
3. Modern Peripheral Submarket − price levels are strongly influenced by the construction year and moderately by the "Centrality and Prestige" component.

Instead of traditional microzones, cluster-based submarkets derived from GWR coefficients provide statistically justified valuation areas. These zones reflect not only the price level but also the underlying mechanisms of price formation − for example, whether prices are driven by centrality, environmental amenities, or properties characteristics.

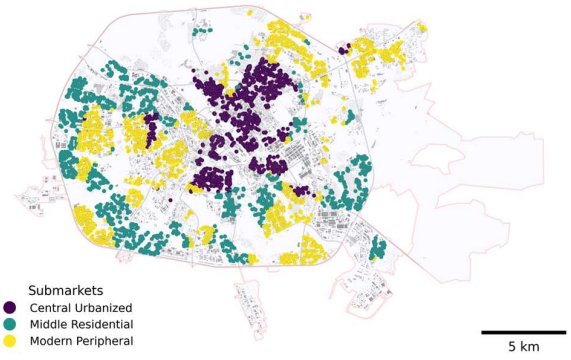


Figure 6. Spatial delineation of housing submarkets in Minsk.

## 4. Conclusions

This study applied an integrated spatial econometric framework to segment the urban housing market of Minsk, demonstrating that residential prices exhibit pronounced spatial structure and heterogeneous price-forming mechanisms. Significant spatial clustering of offer prices (Moran's I = 0.39) confirmed the presence of spatial dependence and justified the use of spatially explicit methods.

The initial GWR model revealed substantial spatial variation in the influence of structural and locational characteristics. At the same time, statistically significant spatial autocorrelation in the residuals indicated that the full set of variables did not fully represent the spatial structure of the market. To address this limitation, a spatial Principal Component Analysis (sPCA) was applied to the accessibility-related predictors. This procedure reduced dimensionality and produced four spatial components that captured the dominant patterns of the urban environment.

The updated GWR model incorporating these components maintained the explanatory performance of the initial specification while eliminating spatial autocorrelation in the residuals, confirming the adequacy of the compressed predictor set. Based on the local coefficients of this improved model, a k-means clustering procedure was used to identify three spatially coherent housing submarkets. These submarkets reflect the functional and historical differentiation of the city: a central zone influenced primarily by centrality-related factors; a middle-ring area where structural dwelling attributes dominate; and a peripheral zone where construction age and "Centrality and Prestige" component plays a decisive role.

The combined application of sPCA, GWR, and cluster analysis resulted in a clear and statistically validated segmentation of the Minsk housing market. The proposed framework is reproducible, relies on openly available spatial data, and can be adapted for valuation zoning, market monitoring, and applied urban-planning tasks in other cities. All analytical stages can be implemented entirely with open-source software, ensuring transparency and low implementation costs.

## References

Anselin, L., 1988. *Spatial Econometrics: Methods and Models.* Dordrecht: Kluwer Academic Publishers.

Bidanset, P., McCord, M., Davis, P., & Hermans, L., 2024. Delineating Market Areas Used for Mass Valuation Using Geographically Weighted Regression (GWR) and Hierarchical Cluster Analysis (HCA). *Journal of Property Tax Assessment and Administration*, 22(1), 5–26. https://doi.org/10.63642/1357-1419.1272

Bourassa, S.C., Hoesli, M., Peng, V.S., 2003. Do housing submarkets really matter? *Journal of Housing Economics*, 12(1), 12–28.

Chen, M., Chun, Y., Griffith, D. A., 2023. Delineating housing submarkets using space–time house sales data: spatially constrained data-driven approaches. *Journal of Risk and Financial Management*, 16(6), 291. https://doi.org/10.3390/jrfm16060291

Dimopoulos, T., Moulas, A., 2016. A Proposal of a Mass Appraisal System in Greece with CAMA System: Evaluating GWR and MRA techniques in Thessaloniki Municipality. *Open Geosciences*, 8, 675–693. https://doi.org/10.1515/geo-2016-0064

Fletcher, M., Gallimore, P., Mangan, J., 2000. The modelling of housing submarkets. *Property Investment & Finance*, 18(4), 473-487. https://doi.org/10.1108/14635780010345436

Uber Technologies Inc., 2018. H3: Hexagonal Hierarchical Spatial Index. Software documentation. https://www.uber.com/en-FR/blog/h3/ (accessed 10 November 2025).

Kauko, T., Hooimeijer, P., Hakfoort, J., 2002. Capturing housing market segmentation: an alternative approach based on neural network modelling. *Hous. Stud.,* 17(6), 875–894. https://doi.org/10.1080/02673030215999

Kontur Inc., 2019. Kontur Population Dataset. Research dataset, Humanitarian Data Exchange (HDX). https://data.humdata.org/dataset/kontur-population-dataset (accessed 10 November 2025).

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, Article 105292. https://doi.org/10.1016/j.landusepol.2021.105292

Lockwood, T., Rossini, P., 2011. Efficacy in Modelling Location within the Mass Appraisal Process. *Pacific Rim Property Research Journal*, 17, 418–442. https://doi.org/10.1080/14445921.2011.11104335

McCluskey, W.J., Borst, R.A., 2011. Detecting and validating residential housing submarkets A geostatistical approach for use in mass appraisal. *International Journal of Housing Markets and Analysis*, 4, 290–318. https://doi.org/10.1108/17538271111153040

Morawakage, P., Earl, G., Liu, B., Roca, E., Omura, A., 2022. Housing risk and returns in submarkets with spatial dependence and heterogeneity. Journal of *Real Estate Finance and Economics*, 65(4), 1–40. https://doi.org/10.1007/s11146-021-09877-7

OpenStreetMap contributors, 2017. OpenStreetMap. Geospatial database. https://www.openstreetmap.org (accessed 10 November 2025).

Realt.by, 2017. Realt.by – Real estate portal. Web portal. https://realt.by (accessed 12 October 2017).

Wu, C., Sharma, R., 2012. Housing submarket classification: the role of spatial contiguity. *Appl. Geogr.* 32, 746–756 https://doi.org/10.1016/j.apgeog.2011.08.011