land

MDPI

Article

Aggregated Housing Price Predictions with No Information About Structural Attributes—Hedonic Models: Linear Regression and a Machine Learning Approach

Joanna Jaroszewicz * and **Hubert Horynek**

Department of Spatial Planning and Environmental Sciences, Warsaw University of Technology, Plac Politechniki 1, 00-661 Warsaw, Poland; hubert.horynek@pw.edu.pl
* Correspondence: joanna.jaroszewicz@pw.edu.pl

Abstract: A number of studies have shown that, in hedonic models, the structural attributes of real property have a greater influence on price than external attributes related to location and the immediate neighbourhood. This makes it necessary to include detailed information about structural attributes when predicting prices using regression models and machine learning algorithms and makes it difficult to study the influence of external attributes. In our study of asking prices on the primary residential market in Warsaw (Poland), we used a methodology we developed to determine price indices aggregated to micro-markets, which we further treated as a dependent variable. The analysed database consisted of 10,135 records relating to 2444 residential developments existing as offers on the market at the end of each quarter in the period 2017–2021. Based on these data, aggregated price level indices were determined for 503 micro-markets in which primary market offers were documented. Using the analysed example, we showed that it is possible to predict the value of aggregated price indices based only on aggregated external attributes—location and neighbourhood. Depending on the model, we obtained an R^2 value of 75.8% to 82.9% for the prediction in the set of control observations excluded from building the model.

Keywords: asking prices; data aggregation; external attributes; hedonic models; linear regressions; machine learning

check for updates

Citation: Jaroszewicz, J.; Horynek, H. Aggregated Housing Price Predictions with No Information About Structural Attributes—Hedonic Models: Linear Regression and a Machine Learning Approach. *Land* **2024**, *13*, 1881. https://doi.org/10.3390/land13111881

Academic Editor: Piyush Tiwari

Received: 10 October 2024
Revised: 1 November 2024
Accepted: 6 November 2024
Published: 11 November 2024

1. Introduction

The influence of an apartment's neighbourhood on its price has been proven many times [1–5]. Understanding what affects market price allows prices to be modelled more effectively [6]. Such models are successfully created based on hedonic methods, with their concept based on the assumption that the value of certain goods is determined by a certain set of attributes (characteristics) [7]. Each of these attributes is assigned a certain utility, where the value of a good is a function of the sub-utilities of the individual characteristics [1,3,4,8–10]. Hedonic models of real property prices are used both to predict prices and to determine the strength of the impact of individual real property attributes on prices [11]. In a formal sense, hedonic models take the form of econometric models in which the value of the good is the dependent variable (p), and groups of attributes of the good (for example: S, L, N, . . .) are groups (vectors) of independent variables with a relationship between them [4,12] (f): $p = \mathrm{f}(\mathrm{S, L, N}, \ldots) + \varepsilon$.

In the case of the housing market, groups of attributes can be divided into structural (internal) and external attributes [13]. Internal attributes are variables describing the structural characteristics of houses or apartments (S), while external attributes are variables describing place—location (L) and neighbourhood (N) [3,13–21].

Recent literature has analysed issues related to external attributes concerning, among other things, (1) the correct determination of spatial relationships for location attributes (L) and (2) the correct definition of a neighbourhood in terms of neighbourhood attributes (N).

The values of location and neighbourhood attributes are mainly determined on the basis of spatial data, using spatial analyses in geographic information systems (GIS). The method used when determining accessibility is crucial for the location attributes group [18–20], as well as the method to define a neighbourhood [21]. The size of the reference units or the boundary distance adopted defines the geographical scale of the analysis. Hedonic models are sensitive to changes in geographical scale [22]. All this means that a single attribute describing the location or neighbourhood of a real property (for example, the availability of green space) can be expressed through attribute values defined in various ways.

In hedonic models, internal attributes affect the coefficient of determination (R^2) to a greater extent than external attributes [13,17,23]. This makes it difficult to analyse the impact of external attributes, because even within one construction project (in the same location) internal (structural) attributes will differ. An additional difficulty when determining the impact of external attributes on housing prices is the phenomenon of the spatial autocorrelation of property values [24–26]. Under market conditions, the effect of nearby housing prices on the price of housing entering the market is assumed. One way to account for spatial autocorrelation and to decrease the negative impact of data heterogeneity on model validity is to define sub-markets and work out model prices separately for each of these sub-markets [2,27,28]. However, it may be difficult in the case of primary market analysis due to the insufficient number of observations within a sub-market.

For appropriately small units, referred to here as micro-markets, external attributes (location and neighbourhood) can be considered similar. In our research, we have assumed that each micro-market can be characterised by an average price level. Instead of considering the prices per square metre of individual apartments, which are so strongly influenced by internal (structural) attributes, we analysed the average price levels of the micro-markets. This allowed us to include only a group of external attributes (location and neighbourhood) in the hedonic models we built. We obtained the average price levels through a two-stage aggregation: (1) asking prices for each residential development and (2) upon determining the price indices at the end of each quarter for small spatial units (micro-markets). Larger housing projects executed by developers are generally divided into stages. One stage is a single development, a single construction project, carried out on the basis of a single building permit. We used the boundaries of cadastral sections as the boundaries of micro-markets. A cadastral section is an area unit of the country's division created for the purposes of the land and building register (real property cadastre) (Regulation of the Minister of Development, Labour and Technology of 27 July 2021 on the land and building register). These small spatial units are clearly separated from each other by facilities such as streets, railways, rivers and escarpments. They are recognisable, distinctive and internally coherent, thus fitting into Lynch's proposed definition of a "district", in the context of the elements of the city structure [29]. Aggregated offers to cadastral sections are therefore similar to each other in terms of neighbourhood (N) and location (L) characteristics. With this in mind, we treat these units as individual observations.

In our research, we analysed asking prices from the primary market for apartments in Warsaw offered for sale at the end of each quarter (Q) over a five-year period (2017–2021). Between 1Q17 and 4Q21, the average asking price per square metre of new apartments in Warsaw increased by almost 65%. The price variability over time in hedonic models is taken into account by introducing the attribute of time as an independent variable (categorical or dummy variable) into the model [3,13,30–32]. Another solution is to determine price indices. An empirical study showed that the highest R^2 in the developed hedonic model is obtained for price indices based on the aggregation of house prices in quarterly periods [33]. In our study, we eliminated the trend of rising prices by relating the observed prices per m^2 of apartment for each development to the average quarterly price calculated for the entire market.

In our study, we test various regression models, including both OLS linear regression and ML algorithms. However, our analyses do not aim to compare and evaluate them, but to demonstrate that, when aggregated to micro-markets (cadastral sections), asking price indicators can be modelled efficiently while taking into account only external attributes (location and environment) determined by spatial relationships, while ignoring information on internal attributes (structural characteristics of individual apartments). This can serve as important support for projecting asking prices for newly marketed residential developments (allowing for appropriate asking prices accepted by potential buyers), with limited access to information on the structural characteristics of individual apartments.

This study is unique and innovative in that (1) the observations constituting the dependent (explanatory) variable in the linear regression models and models based on ML algorithms are, in our study, aggregated indicators of asking prices to small, internally consistent, spatial units (micro-markets) defining the average level of achievable prices; (2) we only analyse asking prices from the primary market for apartments; and (3) our research provides support for forecasting the price, which allows us to propose an appropriate price level, acceptable to potential buyers, for newly marketed residential developments based only on external attributes (location and environment).

2. Literature Review

In the case of the housing market, groups of attributes can be divided into structural (internal) and external attributes [13]. Internal attributes describing the structural characteristics of an apartment include fit-out standard, area, layout, number and height of rooms, floor on which the apartment is located, presence of a balcony or loggia, orientation and view from windows [14,34–36].

External attributes are variables describing place: neighbourhood (N) and location (L) [3,13–17]. Neighbourhood attributes (N) describe the quality of the environment [37,38]. Some studies have focused on environmental aspects such as noise or air pollution [4,13,37], as well as on environmental quality expressed by the proportion of open space, parks, surface water and forests [21]. Variables describing the immediate neighbourhood (N) define its perception, state of development, density of services and environmental characteristics such as noise and air pollution levels, proportion of greenery and open areas. Variables describing location (L) are expressed through spatial relationships (proximity, accessibility) of important facilities for residents, related to meeting their needs. Location attributes describe the location of a building in relation to the city centre (CBD, bona fide city centre) and local centres [2,37,39]), determine the accessibility of public transport [18,30,36], basic services such as schools, shops and recreational areas (surface water, parks, forests and sports facilities) [18,31,40]. Herat and Maier (2010) conducted a systematic literature review on location attributes [41].

The values of external attributes are mainly determined on the basis of spatial data, using spatial analyses in geographic information systems (GIS). The method used when determining accessibility is crucial for the location attributes group [18–20] as well as the method to define a neighbourhood [21], with two approaches typically used: (1) tessellation—a complete and separate division of an area into adjacent reference units, and (2) buffering—determining a neighbourhood around each location defined by a boundary distance (Euclidean or grid). The tessellation can be done into geometric units, such as hexagons [18] or units related to administrative division, such as census tracts or other defined boundaries [42]. Neighbourhoods defined by a buffer around each location can be determined by Euclidean distance [21,32] or network distance calculated along roads or pedestrian routes [43,44]. The size of the reference units or the boundary distance adopted defines the geographical scale of the analysis. Hedonic models are sensitive to changes in geographical scale [22]. To avoid problems regarding analyses' sensitivity they may be conducted on different geographical scales, representing both the immediate neighbourhood in the walking access zone as well as the wider neighbourhood [18,21,40]. For example, the neighbourhood of each transaction in buffer zones with a radius of 100 m and 1 km [21], in squares with a certain side length, as well as in neighbouring squares and second-degree neighbours [40] or in hexagons with edge lengths of 66 m and 174 m, respectively [18]. All this means that a single attribute describing the location or neighbourhood of a real property (for example, the availability of green space) can be expressed through attribute values defined in various ways.

Hedonic models often take the form of regression models. Depending on the version of the regression model adopted, a number of conditions must be met. For example, the frequently used OLS (Ordinary Least Squares) linear regression is a parametric method requiring the explicit modelling of nonlinearities and interactions, including low-model multicollinearity, a nonsignificant spatial autocorrelation of standard residuals and a normal distribution of residuals [21]. The way the independent variables are defined and selected is crucial here [21,45]. The occurrence of the spatial autocorrelation of regression residuals is considered one of the tools for verifying the validity of the hedonic model [1,46]. Regression analysis can also be performed using machine-learning (ML) algorithms [47] in which patterns of data relationships are detected from a training data set. Algorithms include Random Forest regression or the eXtreme Gradient Boosting algorithm (XGBoost) [18,21,47–51]. ML algorithms can easily deal with nonlinear relations [36]. ML-Regression algorithms are used to estimate housing prices because of their ability to learn nonlinear relationships between incoming and outgoing variables, which may correspond more closely to the real situation than linear models [36]. In the literature, the advantage of estimating housing prices using ML-Regression algorithms over traditionally used linear models is usually demonstrated (for example: in an analysis of house transaction data from Onondaga County, NY, USA [21]; for real property data in the district of Gangnam, South Korea [31]; for real property data from the city of Ljubljana [52]; for data from Lisbon [53]; for transactions for apartments in Nicosia District, Cyprus [54]; for a database of housing (asking) prices in Alicante City, Spain [36]). On the other hand, with the use of ML algorithms for mass appraisal of real estate, some problems are indicated, such as the lack of transparency of models and poor repetition of results using machine learning techniques [54,55]. Like the OLS model, ML algorithms allow the significance of the influence of individual independent variables to be assessed [9,26]. Recently, the aggregation of Shapley values for computing feature importance was used [55].

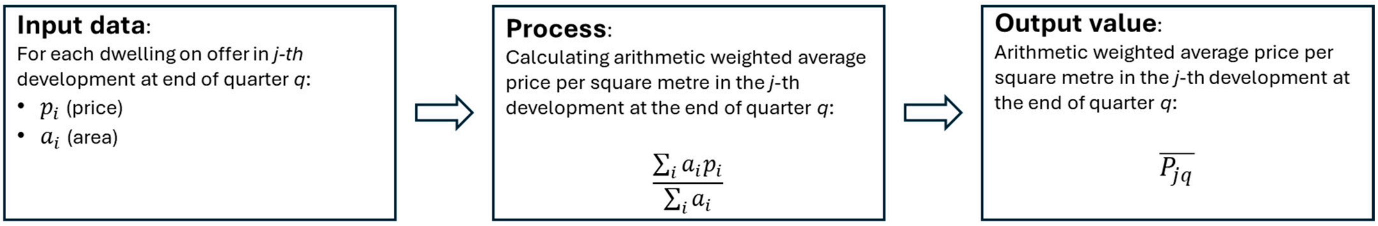
3. Materials and Methods

3.1. The Aggregated Price Level Indices

In our research, we used a database of asking prices on the primary market in Warsaw. In this database, the data are immediately aggregated to entire residential developments (S_{jq}). Large residential investments are divided into smaller developments, and in many cases a single development means a single building. A single residential development constitutes a single record in the extracted asking price database. Each residential development remaining on offer at the end of any quarter in the period 2017–2021 is described by the following data: (1) longitude and latitude, (2) number of apartments remaining on offer, (3) total number of apartments, (4) average price per square metre of apartments on offer and (5) quality segment to which the residential development is assigned. A single record thus contains pre-aggregated data for each j-th development remaining on sale at the end of a given q-th quarter. For information on asking prices, aggregation consists of determining weighted average prices of apartments available for sale at the end of each q-th quarter, where averages are prices per square metre in individual apartments (p_i), and the weights are the usable floor area of these apartments (a_i) (Equation (1), Figure 1).

$$\overline{P_{jq}} = \frac{\sum_i a_i p_i}{\sum_i a_i};\ for\ i:\ d_i \epsilon S_{jq}, \tag{1}$$

where $\overline{P_{jq}}$ is the weighted average price per square metre of the j-th residential development at the end of q-th quarter, a_i is the area of the i-th apartment expressed in square metres, p_i is the price per square metre of the i-th apartment, d_i is the i-th apartment and S_{jq} is the j-th development at the end of the q-th quarter.



Input data:
For each dwelling on offer in *j-th* development at end of quarter *q*:
- p_i (price)
- a_i (area)

Process:
Calculating arithmetic weighted average price per square metre in the *j*-th development at the end of quarter *q*:

$$\frac{\sum_i a_i p_i}{\sum_i a_i}$$

Output value:
Arithmetic weighted average price per square metre in the *j*-th development at the end of quarter *q*:

$$\overline{P_{jq}}$$

Figure 1. Procedure illustration of the $\overline{P_{jq}}$ calculation (own elaboration).

The obtained database consisted of records relating to residential developments that remained on offer at the end of a quarter during the period 2017–2021. The number of records exceeded the number of developments. The reason for this is that the commercialisation process of a residential development is stretched out over time. In the vast majority of cases, residential developments remained on offer for more than one quarter. The aggregation of data in the asking price database was the first stage of aggregation. The price data for individual apartments were aggregated to the entire residential development. In the second stage, there was an aggregation of the developments then located within the boundaries of one cadastral section (Figure 2). The cadastral sections were then treated as single observations. We assumed that the entire area of the cadastral section, analogous to Lynch's definition of an internally coherent district [29], is characterised by the same attributes of location and neighbourhood.

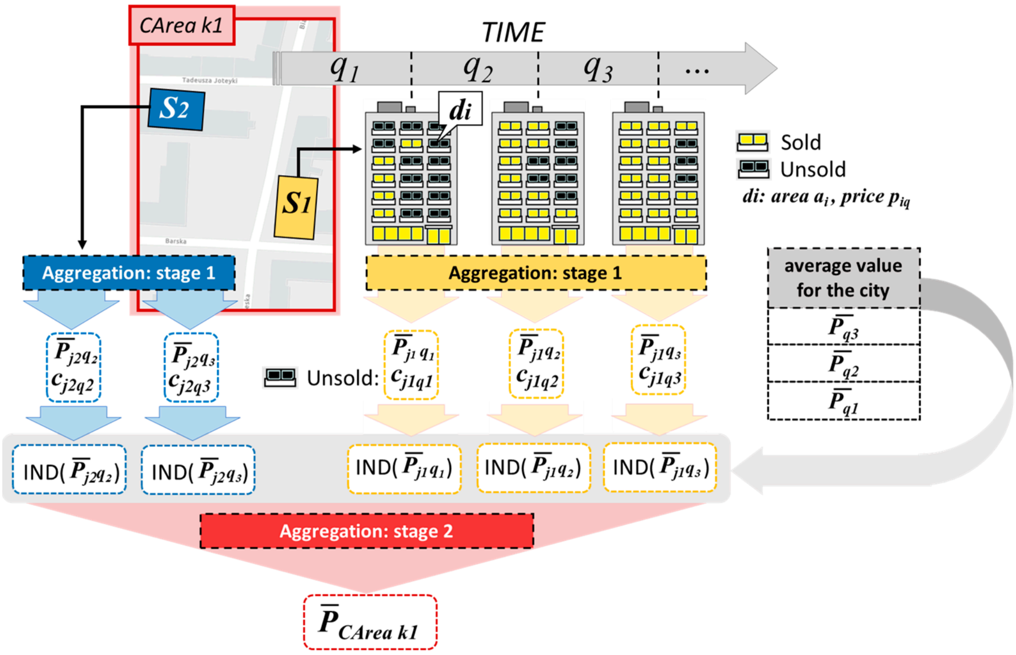
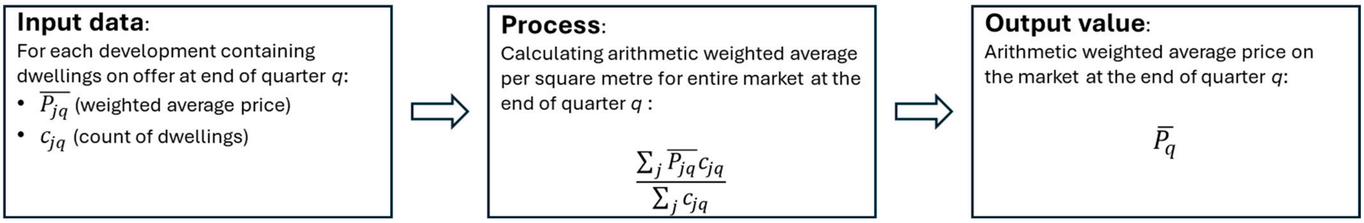

CArea k1
TIME
q_1
q_2
q_3
...
S2
S1
di
Sold
Unsold
di: area a_i, price p_{iq}
Aggregation: stage 1
Aggregation: stage 1
average value for the city
$\overline{P}_{q3}$
$\overline{P}_{q2}$
$\overline{P}_{q1}$
$\overline{P}_{j2q_2}$ c_{j2q2}
$\overline{P}_{j2q_3}$ c_{j2q3}
Unsold:
$\overline{P}_{j1q_1}$ c_{j1q1}
$\overline{P}_{j1q_2}$ c_{j1q2}
$\overline{P}_{j1q_3}$ c_{j1q3}
IND($\overline{P}_{j2q_2}$)
IND($\overline{P}_{j2q_3}$)
IND($\overline{P}_{j1q_1}$)
IND($\overline{P}_{j1q_2}$)
IND($\overline{P}_{j1q_3}$)
Aggregation: stage 2
$\overline{P}_{CArea\ k1}$
Tadeusza Joteyki
Barska

Figure 2. Procedure illustration of the two stages of aggregation of asking prices (own elaboration).

In the analysed period, average prices in Warsaw increased by approximately 65%. As the aggregation to cadastral sections involved data from many quarters, before the second stage of aggregation (to cadastral sections), the price growth trend was eliminated by determining price level indices $IND\left(\overline{P_{jq}}\right)$. Price level indices were calculated by relating the average price of a residential development remaining for sale at the end of the q-th quarter $\left(\overline{P_{jq}}\right)$ to the weighted average, calculated for the entire market (the city of Warsaw) for the same q-th quarter $\left(\overline{P_q}\right)$. The weightings, however, were the number of apartments remaining for sale at the end of the q-th quarter in the j-th development (c_{jq}) (Equation (2), Figures 3 and 4).

$$IND\left(\overline{P_{jq}}\right) = \frac{\overline{P_{jq}}}{\overline{P_q}};\ where:\ \overline{P_q} = \frac{\sum\limits_j \overline{P_{jq}} c_{jq}}{\sum\limits_j c_{jq}}, \tag{2}$$

where $IND\left(\overline{P_{jq}}\right)$ is the price level index: the indexed price per square metre of an apartment in the j-th development at the end of the q-th quarter, $\overline{P_q}$ is the weighted arithmetic average calculated for the given q-th quarter for the whole market; c_{jq} is the number of apartments remaining at the j-th development at the end of the q-th quarter.



Input data:
For each development containing dwellings on offer at end of quarter q:
- $\overline{P_{jq}}$ (weighted average price)
- c_{jq} (count of dwellings)

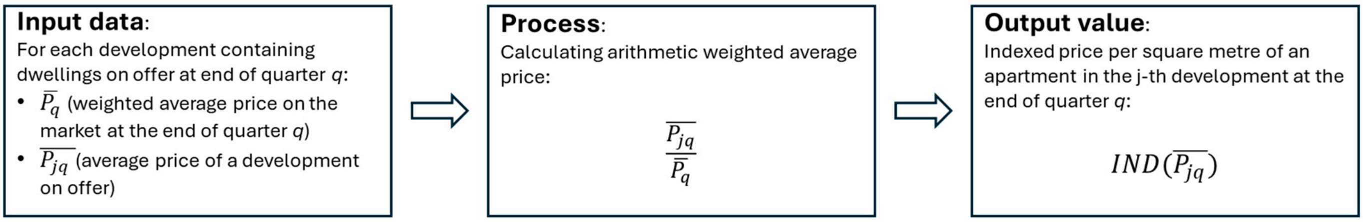
Process:
Calculating arithmetic weighted average per square metre for entire market at the end of quarter q :

$$\frac{\sum_j \overline{P_{jq}} c_{jq}}{\sum_j c_{jq}}$$

Output value:
Arithmetic weighted average price on the market at the end of quarter q:

$$\overline{P_q}$$

Figure 3. Procedure illustration of the $\overline{P_q}$ calculation (own elaboration).



Input data:
For each development containing dwellings on offer at end of quarter q:
- $\overline{P_q}$ (weighted average price on the market at the end of quarter q)
- $\overline{P_{jq}}$ (average price of a development on offer)

Process:
Calculating arithmetic weighted average price:

$$\frac{\overline{P_{jq}}}{\overline{P_q}}$$

Output value:
Indexed price per square metre of an apartment in the j-th development at the end of quarter q:

$$IND\left(\overline{P_{jq}}\right)$$

Figure 4. Procedure illustration of the $IND\left(\overline{P_{jq}}\right)$ calculation (own elaboration).

After determining the price level indices $IND\left(\overline{P_{jq}}\right)$, outlier observations were removed. Outlier observations were considered to be values less than 1.5 × IQR below the first quartile and values greater than 1.5 × IQR above the third quartile. The inter-quartile range (IQR) is the difference between the values that correspond to the first and third quartiles. A second stage of aggregation (to cadastral sections) was carried out for the price level indices $IND\left(\overline{P_{jq}}\right)$ of all developments S_{jq} for any quarter in 2017–2021, which were located within the boundaries of a given k-th cadastral section ($for\ j:\ S_{jq} \in CArea_k$). The aggregated price level indices for each cadastral section $\overline{P}_{CArea_k}$ were calculated as a weighted geometric mean, taking the number of available apartments as weightings c_{jq} (Equation (3), Figure 5). This avoided overly large price impacts of developments with individual apartments that were subject to discounts.

$$\overline{P}_{CArea_k} = exp\left(\frac{\sum\limits_q \sum\limits_j c_{jq} ln\left[IND\left(\overline{P_{jq}}\right)\right]}{\sum\limits_q \sum\limits_j c_{jq}}\right);\ for\ j:\ S_{jq} \in CArea_k, \tag{3}$$

where $\overline{P}_{CArea_k}$ is the price level index aggregated to the k-th cadastral section, $IND\left(\overline{P_{jq}}\right)$ is the price level index for the j-th development at the end of the q-th quarter, c_{jq} is the number of apartments at the j-th development remaining for sale at the end of the q-th quarter and $S_{jq} \in CArea_k$ is such a development that is located in the k-th cadastral section.

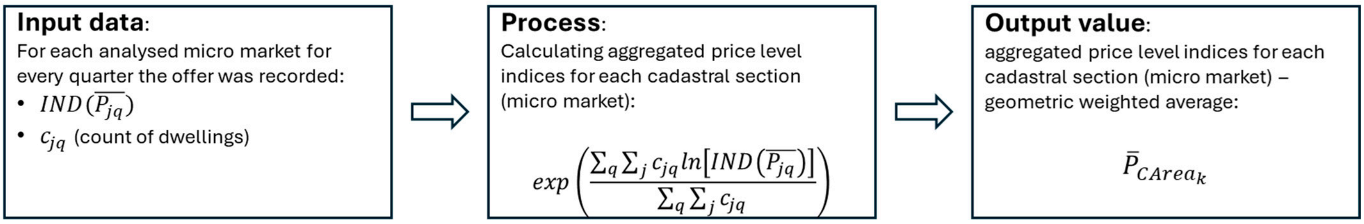

Input data:
For each analysed micro market for every quarter the offer was recorded:
• $IND(\overline{P_{jq}})$
• c_{jq} (count of dwellings)
Process:
Calculating aggregated price level indices for each cadastral section (micro market):
$exp\left(\frac{\sum_q \sum_j c_{jq} ln[IND(\overline{P_{jq}})]}{\sum_q \sum_j c_{jq}}\right)$
Output value:
aggregated price level indices for each cadastral section (micro market) – geometric weighted average:
$\bar{P}_{CArea_k}$

Figure 5. Procedure illustration of the $\overline{P}_{CArea_k}$ calculation (own elaboration).

The price level indices aggregated to the cadastral section $\overline{P}_{CArea_k}$ were logarithmised, transforming the distribution of the values of the dependent variable to be closer to normal. These values ultimately constituted the set of dependent (explanatory) values. For a set of observations on real property prices, a right-skewed value distribution is a fairly common phenomenon, and prices in the literature are generally transformed using the natural logarithm [3,38,39,56,57]. The data, prepared in this manner, formed the basis for further analyses. The set of observations, processed as described above, constituted the values of the explanatory (dependent) variable in the regression models. The two-stage aggregation described above yielded a set of observations consisting of N = 503 cadastral sections.

3.2. Attribute Determination

Information on topographic objects was obtained from the Topographic Objects Database (BDOT10k)—a vector database containing the spatial locations of topographic features along with basic descriptions of their properties. The content and level of detail of the BDOT10k database generally correspond to a traditional topographic map at the scale of 1:10,000. BDOT10k contains information concerning networks of watercourses, roads and railways and utility lines, land cover, protected areas, administrative units, buildings, structures and equipment, land development complexes and other objects. The boundaries of cadastral sections were obtained from the National Register of Boundaries (PRG)—an official reference database providing the basis for other spatial information systems, using data concerning the country's administrative units. Information on the noise level was determined based on the acoustic map of Warsaw, available on the city's mapping service (https://mapa.um.warszawa.pl/ (accessed on 6 April 2024)). The map shows noise levels in urban space, depending on the noise source. In our study, we used daytime road noise data. The district ranking values were obtained from the study by Statistics Poland [58]. In the Statistics Poland study, the districts were compared among themselves using a number of indicators. The result of the comparison was expressed in the form of a quantitative evaluation index.

External attributes are divided into two main categories: location attributes (L) and neighbourhood attributes (N). Location attributes relate to the location in relation to the city centre and local centres, as well as the accessibility of public transport and basic services expressed in terms of distance. Neighbourhood attributes describe the availability of daily needs services, the quality of space and the state of the environment. Regarding the idea of a 15-min city, it is important that a variety of services should be available within the distance of a 15-min walk, including grocery shops, banks and ATMs, restaurants, places of work, sports, recreation and leisure, as well as health, education and cultural services [59,60].

Within a certain group, attribute series were defined. An attribute series is composed of attributes describing the same real property characteristic but determined (a) using a different spatial analysis model, and (b) determined for a different neighbourhood granularity. An example that illustrates the process well is the variables from the social services group concerning the availability of educational facilities. The accessibility of kindergartens, primary schools and secondary schools was calculated using three methods: (1) by calculating the average distance in a straight line to the k nearest facilities, (2) by calculating the number of facilities per unit area and (3) by calculating the relative kernel density for a given type of facility. In our study, we assumed attribute values aggregated to three degrees of granularity: (1) cadastral sections, (2) cadastral sections and contiguous first-order cadastral sections and (3) cadastral sections, contiguous first-order cadastral sections and contiguous second-order cadastral sections (Figure 6).

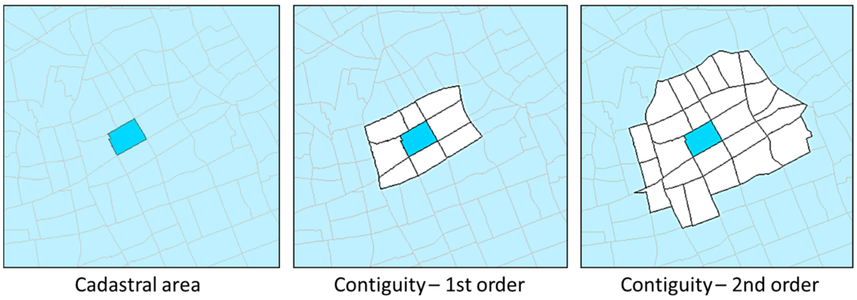


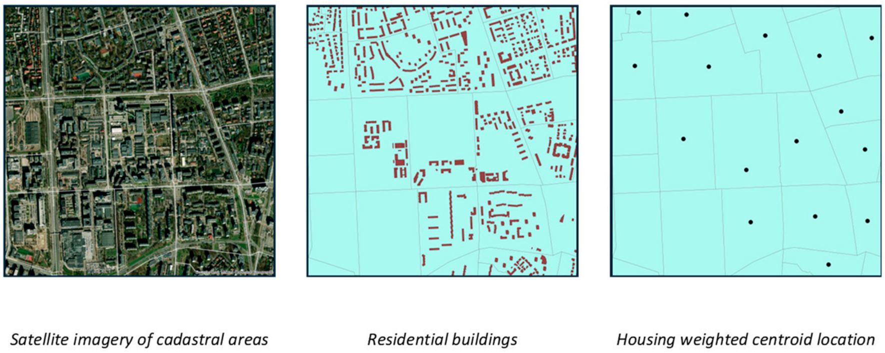
Figure 6. Three degrees of granulation delineating the neighbourhoods adopted for the analyses (own elaboration).

A similar solution of several degrees of granularity can be found, for instance, in [26,28], but they use regular geometric units. In our study, values determined based on a first-order or second-order adjacency relationship were always assigned to the cadastral section record. Attribute values were determined using spatial analysis in GIS software (ArcGIS Pro 3.2.1 ESRI), while various methods were used to determine them, including the distance from the weighted centroid of the cadastral section to the nearest object, the average distance from the weighted centroid of the cadastral section to k nearest objects, the density of objects per square km, the number of objects in the unit, the percentage shares of area/type in the unit, the average values of the attribute in the unit, in the case of the distance from the city centre, the network distances along the roads, the travel time by public transport with and without congestion during rush hours.

Variables that characterise entire development rather than the single product (apartment) are also analysed. These include the number of apartments in a building [31,36,40]. Attributes describing the characteristics of the development combine both structural attributes (S) and neighbourhood attributes (N) related to the prestige and character of the built environment; in our study, we referred to them as semi-structural attributes. We determined two variables that are semi-structural attributes. The first is the average number of apartments in new residential buildings. As this value relates to the entire micro-market (cadastral section), we consider it a characteristic of the built environment rather than a structural variable. The second variable derived from the structural variables relates to the prestige of the neighbourhood in a broad sense. The database of asking prices on the primary market in Warsaw contains data on quality segments to which particular developments are assigned. The values of the prestige variables were determined, taking into account the composition of the quality segments in a given spatial unit. The method of determining the values of these variables is described in Appendix B. In the conducted analyses, we performed calculations with and without semi-structural attributes. Similarly, variables related to neighbourhood prestige have been used [37,38].

For the groups of location attributes (L), based mainly on distance relationships, the reference for the spatial analyses was the housing weighted centroids (Figure 7) designated within each cadastral section. The weightings were determined as the total area of each residential building located within the cadastral section, calculated as the product of the number of storeys and the area of the building's footprint.

As a result, for each issue, we obtained a series of attributes, differing in the way they were determined and the degree of aggregation of values.



Satellite imagery of cadastral areas

Residential buildings

Housing weighted centroid location

Figure 7. Housing weighted centroid as reference for attributes determined by the distance relationship (own elaboration).

We divided the external attributes into two categories: location (L) and neighbourhood (N). The attributes were also divided into groups: location in the city structure, public transport and individual transport, buildings, commercial services, social services, nuisances, greenery and recreation, as well as prestige. Each group included several to a dozen attributes, for which values were then calculated for the previously adopted degrees of granularity. For infrequent sites, such as theatres, some of the methods of determining values returned a value of 0 for many cadastral sections. Those methods that returned values equal to 0 for more than 25% of the cadastral sections after calculating attribute values were abandoned. Once the attribute values had been calculated, they were analysed for the value of the coefficient of determination R^2 for the correlation with the dependent variable. After the initial selection of attributes, the distributions of their values were analysed, the coefficient of skewness was assessed and the method of normalisation was selected. One of four types of normalisations was used: natural logarithm, exponential function (exp), arcsine function or Box–Cox transformation, depending on the observed distribution of values of the attribute in question. Table A1 in Appendix A lists the attributes determined, the category they have been assigned to (location/neighbourhood), the group, how they were determined, the degree of granularity and the R^2 coefficient for the correlation with the dependent variable.

Since in OLS linear regression models the explanatory (independent) variables should not be mutually collinear, a second stage of attribute selection was necessary. The following rationales were followed: (1) obtaining a set of non-redundant variables, (2) guaranteeing a high level of explained variance (R^2), (3) meeting, if possible, the other diagnostic tests of OLS analysis and (4) establishing variables representing all attribute groups. For this purpose, we used the Exploratory Regression tool in ArcGIS Pro 10.3.1, which analyses the diagnostics of OLS models with increasing number of variables. In addition, correlation coefficients between pairs of variables were analysed. Finally, two sets of attributes were selected for further analysis, allowing the analysis of two OLS models: OLS_I—for a set of variables including semi-structural attributes, and OLS_II—without variables representing semi-structural attributes (see Table A2 in Appendix C).

All the variables obtained after the initial selection stage, representing groups of attributes, were accepted into ML-Regression models. Analyses were performed for models containing variables from the structurally derived attribute group and for models without these variables. In addition, ML models were analysed for the set of variables adopted for the linear regression models OLS_I and OLS_II.

In our study, we tested ML-Regression models: the Random Forest regression algorithm [61] and the Extreme Gradient Boosting (XGBoost) regression algorithm [48]. The aim of the tests was to verify the effectiveness of ML-Regression methods for predicting aggregate price indices based on external attributes only, related to the location and neighbourhood of certain stages of housing developments. For the Random Forest (RF) model, we carried out an optimisation (tuning) of hyperparameters using Random Search (Robust) to optimize R^2. Random Search (Robust) is a stratified random sampling algorithm used to select the search points. Each search is run 10 times using a different random seed. The result of each search is the median best run determined by the R^2 value. Optimising the hyperparameters of the model reduces variability in their prediction accuracy [62,63]. Some papers have offered a detailed description of hyperparameters, including [55,64].

3.3. Model Evaluation

The models' results underwent a two-stage evaluation: (1) an internal evaluation of the model carried out on the basis of diagnostics with the data used to build the models (adopting the ML nomenclature—on the basis of the training dataset) and (2) an external evaluation of the model (prediction) carried out on the basis of diagnostics using data excluded from model building—based on the control dataset. For each of the eighteen districts of Warsaw, a control dataset representing approximately 10% of the observations was selected from the cadastral sections in which new apartments were offered in the period 2017–2021. For the internal evaluation of the OLS model, the following was adopted: Adjusted R-Squared (R^2), AICc—corrected Akaike Information Criteria [65], JB (Jarque–Bera) test for normality of regression residuals [66], a studentised Breusch–Pagan *p*-value (BP) [67], Variance Inflation Factor (VIF) and the autocorrelation of standardised residuals: Global Moran's I *p*-value (SA). The best model returns the highest possible R^2 value, the lowest possible AICc value, a VIF value less than 7.5 and statistically insignificant ($p > 0.05$) JB, BP and SA scores. The internal evaluation of the OLS models allowed the selection of a set of attributes generating the best results. Two OLS models were adopted for further analysis: OLS_I containing variables representing semi-structural attributes, and OLS_II without these variables—see Table A2 in Appendix C.

The following values were taken for the internal evaluation of the ML-Regression algorithms: R^2, RMSE (Root Mean Square Error), MAPE (Mean Absolute Percentage Error) determined on the basis of k-fold cross-validation and Global Moran's I *p*-value (SA) of standardised residuals. In addition, error values for "out of the bag" (OOB) observations are given for the Random Forest model. In the Random Forest ensemble algorithm, the model is created using a bagging technique. In this technique, there is a random sampling of observations with repetitions. Some of the observations are not used in any of the individual trees—these observations are considered "out of the bag" (OOB). The OOB data are used to estimate the mean square error of the Random Forest predictions and to assess the significance of the variables [67]. The prediction of values in the control set was used, and R^2 and RMSE were determined [68] for the external evaluation of the OLS model and ML-Regression models. The formulae of the diagnostic measures are summarised in Table 1.

Table 1. Diagnostic measures adopted.

Diagnostic Measures	Equations
Root Mean Square Error (RMSE)	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$
Mean Absolute Percentage Error (MAPE)	$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left\lvert\frac{y_i - \hat{y}_i}{y_i}\right\rvert$
R-Square (R^2)	$R2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$
Adjusted R-Square (adjR^2)	$adjR2 = 1 - \left[\frac{(1-R^2)(n-1)}{n-k-1}\right]$

Where: y_i is the observed value; $\hat{y}_i$ is predicted value; $\overline{y}$ is the average value of the observed; n is the number of observations; k is the number of independent variables (attributes) in the model.

3.4. The Study Area

The study area was within the administrative boundaries of Warsaw (Figure 8). Warsaw covers an area of approximately 517 km^2, with approximately 23.9% of its area consisting of residential areas. In 2021, Warsaw had 1,863,056 residents who lived in a housing stock of 1,046,864 housing units. The housing stock in Warsaw has continued to increase in recent years, as evidenced, for example, by the fact that 107,431 housing units (apartments and houses) were put into use between 2017 and 2021. The vast majority of these—95.6%—were constructed by developers. In this period, developers built 98,424 new apartments.

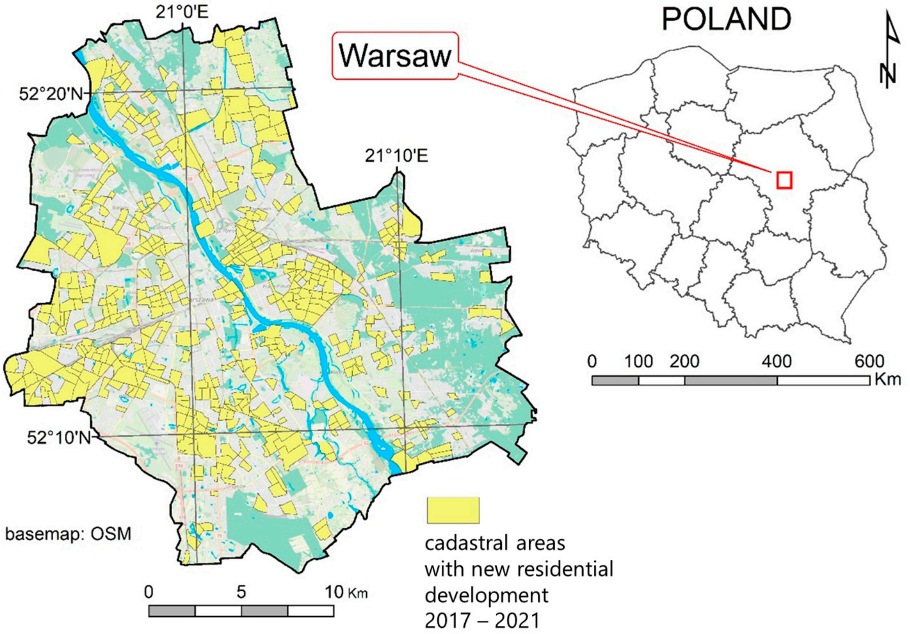
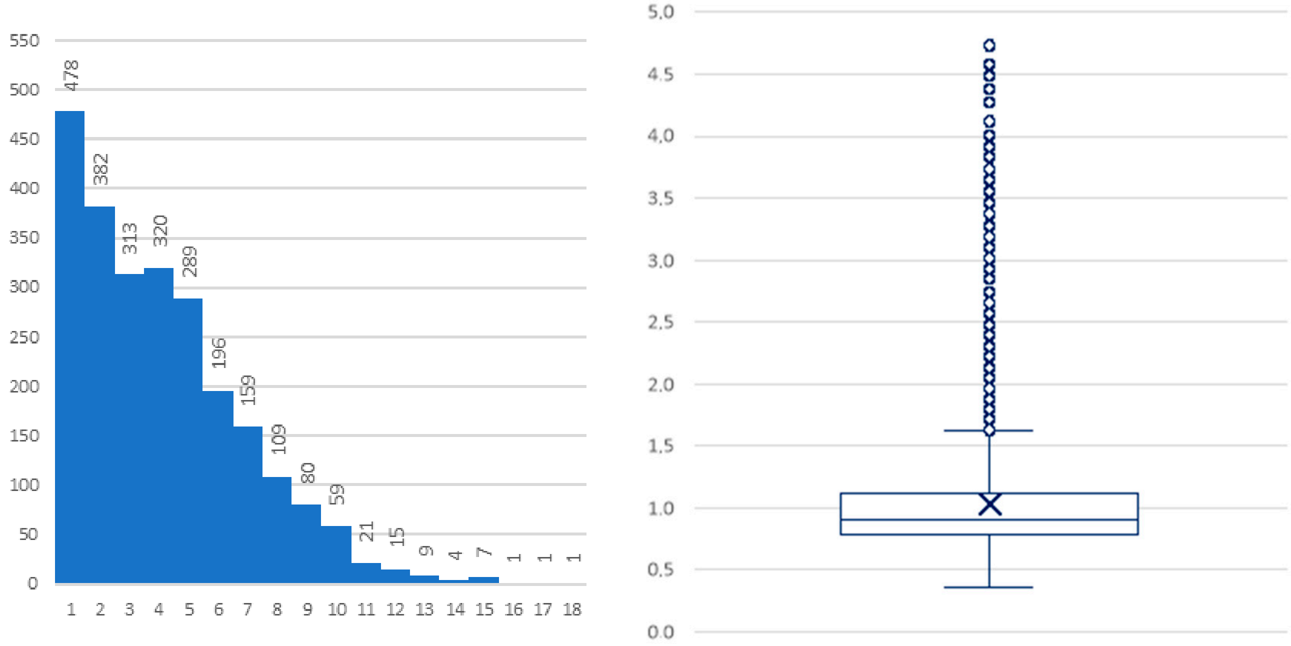


Figure 8. The analysis area: own study based on the National Register of Boundaries (PRG) and Open Street Map (OSM).

4. Results

4.1. Aggregated Average Price Indices

The average value of the housing projects price indices $IND\left(\overline{P_{jq}}\right)$ was 1.0298, median: 0.9097, IQR: 0.3401. The outlier observations were developments whose price level indices $IND\left(\overline{P_{jq}}\right) > 1.63$, meaning that the weighted average prices of apartments in a given development were 1.63 times higher than the weighted average price for the entire market. Outlier values accounted for 748 records out of 10,135 (7.38%) (Figure 9b). After rejecting the outlier observations, aggregate price level indices $\overline{P}_{CArea_k}$ were determined (Figure 10). In Figure 10, the values of the aggregated price level index $\overline{P}_{CArea_k} > 1.00$ indicate values above the average for the entire market (city), and values below the $\overline{P}_{CArea_k} < 1.00$ indicate values below the average. The spatial distribution of values $\overline{P}_{CArea_k}$ deviates from a random distribution, showing significant spatial autocorrelation (for Moran's I, the pseudo-*p* value is 0.001 for 999 permutations) (Figure 11). Clusters of high values of aggregate price level indices $\overline{P}_{CArea_k}$ are observed in the city centre and along metro lines, particularly the M1 line in a north–south direction. The location on the left bank part of the city also seems to be significant. On the right bank, clusters of low values of the aggregate price level index are observed in districts farther from the city centre $\overline{P}_{CArea_k}$. It seems, therefore, that the attributes related to the accessibility of the city centre (distance from the centre, accessibility of metro stations, distances to facilities located mainly in the centre) will have a significant impact on the analysed values of the aggregated price level indices.



(**a**) (**b**)

Figure 9. (**a**) Number of housing project (vertical axis) that remained on sale at the end of a certain number of quarters (horizontal axis); (**b**) *Box plot* showing the distribution of the value of indexed prices of housing project stages $IND\left(\overline{P_{jq}}\right)$.

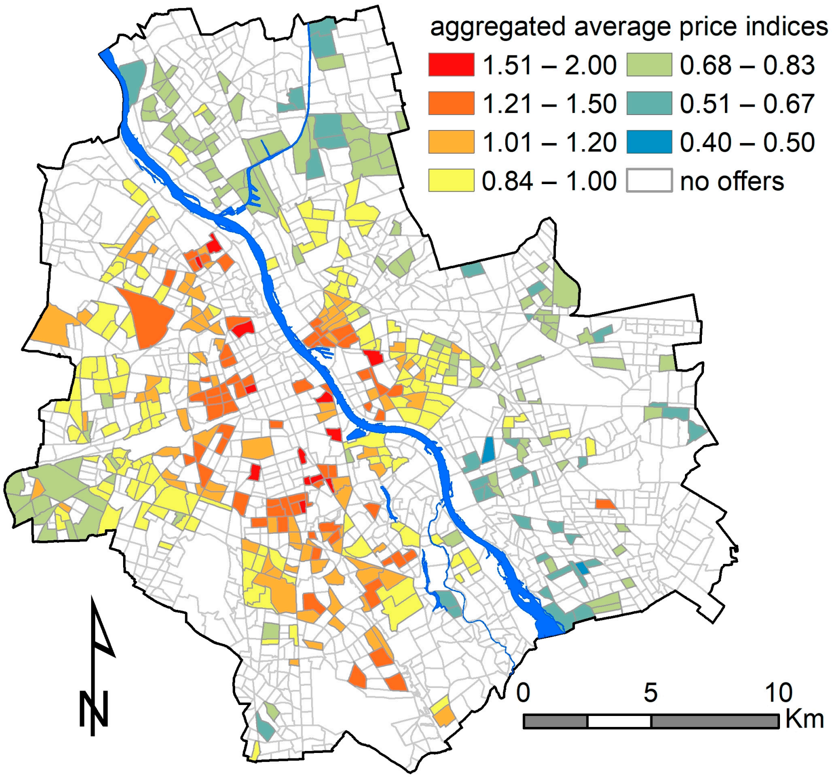


Figure 10. Aggregated average price indices $\overline{P}_{CArea_k}$—dependent variable.

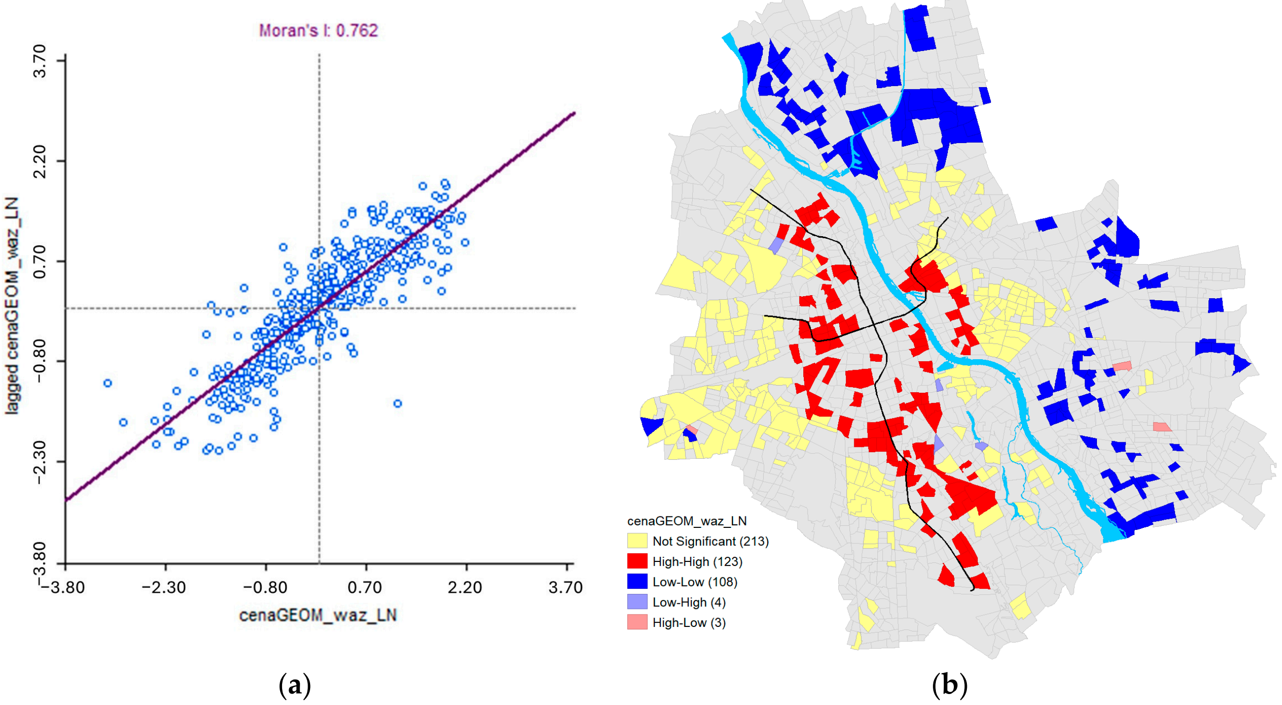


Figure 11. Spatial autocorrelation analysis of the determined indicators of aggregate price levels $\overline{P}_{CArea_k}$: (**a**) Moran's I scatter plot, pseudo-*p* value for 999 permutations = 0.001; (**b**) Map of the local Moran's I coefficient—Anselin statistic. Spatial weightings included: type k-NN (to eight neighbours), inverse distance. Black lines—metro lines existing in 2021 (own work; calculations and visualisation using GeoDa 1.20.0.20).

4.2. Regression Models Results

The dataset was divided into a training set (451 observations, 90%) and a control set (52 observations, 10%). The cadastral sections of the control set were randomly selected from each district of the city. Figure A1 in Appendix D shows their location.

Analyses were carried out using ArcGIS Pro 10.3.1 software. The analysis was carried out in two variants: (1) MODEL OLS_I—for selected variables including semi-structural attributes; (2) MODEL OLS_II—for selected variables without semi-structural attributes. The variables used in the OLS_I and OLS_II models are summarised in Table A2 in Appendix C. The obtained diagnostic results are presented in Table 2.

Table 2. Diagnostic results of OLS linear regression models for the training data from 451 cadastral sections.

Diagnostic Values	OLS_I	OLS_II
Number of variables	19	18
Multiple R-Squared (R^2)	0.794	0.782
Adjusted R-Squared (adjR^2)	0.785	0.773
Akaike's Information Criterion (AiCc)	−626.00	−602.74
Jarque–Bera Statistic (JB) [1]	0.0000	0.0000
Autocorrelation of standardized residuals [2]: Moran's I	0.002	0.001
MSE	0.013	0.014
RMSE	0.115	0.118
MAPE	2.102	2.382

[1] Prob(>chi-squared), (2) degrees of freedom; [2] pseudo-*p* value for 999 permutations.

Table 2 shows that, for both linear regression models, a high coefficient of determination R^2 was obtained, but the residuals obtained do not have a normal distribution (statistically significant Jarque–Bera Statistic (JB) value) and are not randomly distributed in geographical space (statistically significant spatial autocorrelation). Due to the collinearity of the variable values, the set was reduced from 66 variables to 19 for OLS_I and 18 for OLS_II. The inclusion of semi-structural variables in the OLS_I linear regression model increased the amount of explained variance (increase in R^2 and adjR^2) and reduced model uncertainty (reduction in MSE, RMSE and MAPE errors).

The prediction results for the 52 cadastral sections of the control set in the OLS_I and OLS_II models are shown in Table 3. In the OLS_I model, the prediction was based on 18 selected external attributes (location L and neighbourhood N), with the addition of one semi-structural attribute (prestige_f2). In the OLS_II model, the prediction was based on 18 external attributes without a semi-structural one. For the aggregated price level coefficients $\overline{P}_{CArea_k}$, a properly selected set of external attributes allowed 84.1% and 82.9% of the variance to be explained, respectively. In districts where no investment stages were recorded in the analysed period, prediction is possible according to the OLS_II model, while if asking prices of developments are already recorded in the database, it is possible to determine semi-structural attributes and make predictions using the OLS_I model, characterised by a higher R^2 coefficient and lower uncertainty of the result (lower RMSE, MAPE).

Table 3. OLS model diagnostics for the prediction of aggregate price level coefficients $\overline{P}_{CArea_k}$ in the 52 cadastral sections of the control set.

Diagnostic Values	OLS_I	OLS_II
Multiple R-Squared (R^2)	0.8413	0.8287
Adjusted R-Squared (adjR^2)	0.8249	0.7353
MSE	0.0089	0.0098
RMSE	0.0946	0.0992
MAPE	0.6847	0.8711

Analyses were carried out using ArcGIS Pro 10.3.1 software, and model variants were analysed using the following algorithms: Random Forest regression (RF-Regression) with hyperparameter optimisation and XGBoost-Regression. A summary of the models is presented in Table 4. The diagnostic results for the test data, determined using the k-folds method, of the RF-Regression and XGBoost-Regression models are presented in Tables 5 and 6, respectively.

Table 4. Analysed ML-Regression models.

Attributes	RF-Regression	XGBoost-Regression
All attributes (66)	MODEL10	MODEL20
All attributes excluding semi-structural (61)	MODEL11	MODEL21
OLS_I attributes (19)	MODEL12	MODEL22
OLS_II attributes (18)	MODEL13	MODEL23

Table 5. Random Forest regression model diagnostics. Predictions for the test data (excluded from model training) compared to the observed values for those test features.

Diagnostic Values	MODEL10	MODEL11	MODEL12	MODEL13
OOB: Errors	0.014	0.016	0.016	0.017
OOB: % of variation explained	78.324	74.457	75.143	73.686
R2	0.867	0.793	0.825	0.829
MAE	0.076	0.090	0.078	0.088
MAPE	1.743	3.466	1.803	0.834
RMSE	0.099	0.131	0.100	0.110
stResid SA (*p* value)	0.910	0.069	0.456	0.636
stResid SA (pseudo *p* value for 999 permutations)	0.443	0.046	0.213	0.282

Table 6. XGBoost regression model diagnostics. Predictions for the test data (excluded from model training) compared to the observed values for those test features.

Diagnostic Values	MODEL20	MODEL21	MODEL22	MODEL23
R2	0.831	0.801	0.711	0.753
MAE	0.089	0.095	0.086	0.107
MAPE	0.861	2.015	1.223	1.089
RMSE	0.116	0.119	0.123	0.136
stResid SA (*p* value)	0.946	0.291	0.676	0.654
stResid SA (pseudo *p* value for 999 permutations)	0.447	0.116	0.298	0.113

In addition, for the ML-Regression models, predictions were performed for 52 control set cadastral sections in the same manner as for the OLS models. The diagnostics of the obtained control results for the models using the RF-Regression algorithm and the XGBoost-Regression algorithm are set out in Tables 7 and 8, respectively.

Table 7. External diagnostics of Random Forest regression models for the prediction of aggregate price level coefficients $\overline{P}_{CArea_k}$ in 52 control set cadastral sections.

Diagnostic Values	MODEL10	MODEL11	MODEL12	MODEL13
R^2	0.801	0.743	0.787	0.758
MSE	0.011	0.014	0.012	0.014
RMSE	0.106	0.120	0.110	0.117
MAPE	1.134	1.022	1.153	1.581

Table 8. External diagnostics of XGBoost regression models for the prediction of aggregate price level coefficients $\overline{P}_{CArea_k}$ in 52 control set cadastral sections.

Diagnostic Values	MODEL20	MODEL21	MODEL22	MODEL23
R^2	0.785	0.739	0.760	0.762
MSE	0.012	0.015	0.014	0.013
RMSE	0.110	0.121	0.116	0.116
MAPE	1.536	1.309	1.234	1.191

Upon comparing the prediction results of the aggregate price level coefficients $\overline{P}_{CArea_k}$ obtained in the test set (451 cadastral sections) for linear regression models (Table 2) and for ML-Regression models (Tables 5 and 6), it can be observed that, with the exception of MODEL22 and MODEL23 (XGBoost algorithm, with a small number of attributes), the ML models provided higher values of the coefficient of determination R^2 than the OLS linear regression models. The OLS models achieved R^2 of 79.4% and 78.2%, respectively, while the RF-Regression models achieved R^2 values of 79.3–86.7%. However, for predictions in the control set (52 cadastral sections excluded from model building), the OLS linear regression models were the ones that performed better: they achieved R^2 = 82.9% (OLS_II) and R^2 = 84.1% (OLS_I) of variance, while the RF-Regression models achieved R^2 = 74.3–80.1% and the XGBoost-Regression models achieved R^2 = 73.9–78.5%. In the control set, predictions using OLS linear regression also showed lower uncertainty (lower RMSE and MAPE values) compared to the ML models: RF-Regression (MODEL10–MODEL13) and XGBoost-Regression (MODEL20–MODEL23).

4.3. Attributes/Variables Results

In the majority of the analysed models, the inclusion of semi-structural variables improved the predictions in both the test and control datasets. For the OLS linear regression models, the inclusion of semi-structural variables increased the R^2 by 1.2% for the z-prediction in the test and control datasets, while reducing the RMSE and MAPE uncertainty values. The results achieved for ML models are less explicit, especially for the XGBoost-Regression models. Figure 12 compares the R^2 values obtained by each model for prediction in the test and control sets.

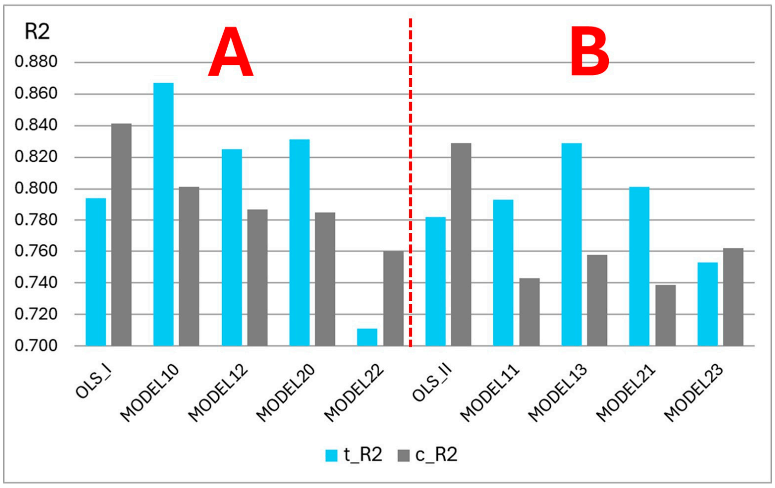


Figure 12. R^2 values obtained in regression models (**A**) with semi-structural attributes and (**B**) without semi-structural attributes; t_R^2 is the R^2 coefficient obtained for the prediction in the test set (451 cadastral sections); c_R^2 is the R^2 coefficient in the control set (52 cadastral sections).

The TOP-10 most relevant attributes were selected for each model analysed (Table 9). The main goal is to demonstrate that if the so-called semi-structural attributes are included in the model, they are highly ranked for both the linear and ML models. Table 9 presents a summary of the TOP-10 attributes for the OLS and Random Forest regression models.

Table 9. A list of TOP-10 attributes for OLS and Random Forest regression models.

Models Including Semi-Structural Attributes		
MODEL OLS_I (19 attributes)	MODEL12 (RF) (19 attributes)	MODEL10 (RF) (66 attributes)
prestige_f2 metro_2nd_closest noise1_dist tennis_MeanDist7 elementary_MeanDist7 sport_MeanDist7 theatre MF2_share_builtUp district_rank retail_krnl2	prestige_f2 museum_MeanDist4 bank_MeanDist7 MF2_share_builtUp theatre noise1_dist metro_2nd_closest tennis_MeanDist7 sport_MeanDist7 city_center_time_public	bank_krnl2_dist prestige_mid1 prestige_mid2 prestige_f1 prestige_f2 pharmacy_krnl2_dist bank_MeanDist7 highschool_krnl2_dist tennis_MeanDist7 MeanCountDwellings_ca
Models Excluding Semi-Structural Attributes		
MODEL OLS_II (18 attributes)	MODEL13 (RF) (18 attributes)	MODEL11 (RF) (61 attributes)
bank_MeanDist7 tennis_MeanDist7 noise1_dist district_rank metro_2nd_closest theatre park2_area city_center_time_public elementary_MeanDist7 MF2_share_builtUp bank_MeanDist7	bank_MeanDist7 MF2_share_builtUp noise1_dist metro_2nd_closest tennis_MeanDist7 city_center_time_public theatre district_rank landmarks_MeanDist7 elementary_MeanDist7 bank_MeanDist7	bank_krnl2_dist city_center_dist_roads tennis_MeanDist7 bank_MeanDist7 museum_MeanDist4 metro_3rd_closest MF2_share_builtUp park_meanDist_ca SF2_share_builtUp landmarks_krnl2_dist bank_krnl2_dist

Table 9 shows that, if semi-structural attributes are included, they are very significant. The variable prestige_f2 came first in the MODEL OLS_I and MODEL12 rankings. If all the variables were taken into account (MODEL10), all five semi-structural attributes were in the TOP-10 of the relevance ranking, with quite an apparent advantage of prestige_mid over prestige_f and with the medium granularity of reference units (1st order adj.) over low granularity (2nd order adj.). For models excluding semi-structural attributes in the TOP-10, the most relevant attributes were:

1. Attributes related to the location of banks (in each of the three models, this was the most significant variable; more banking service points are located in the city centre and in local centres; in the OLS linear regression models, the attribute "bank_MeanDist7"—determined as the average distance to the seven closest banking service points—was considered; when all 61 attributes were considered in MODEL11, the most significant attribute was bank_krnl2_dist, i.e., the relative kernel density calculated for banks within the second grade cadastral section neighbours divided by distance to the city centre);
2. Attributes related to facilities such as museums, theatres and historical buildings that are mainly found in the city centre;
3. Attributes related to the accessibility of the city centre: city_center_time_public, i.e., travel time to the centre by public transport, and, with all variables included, city_center_dist_roads, i.e., distance to the centre calculated by roads;
4. Attributes related to the accessibility of public transport, i.e., metro_2nd_closest and, in the model for all attributes, metro_3rd_closest—defining the distance to the second and third nearest metro stations respectively—these distances are small when metro stations occur close to each other, such is the case for intersecting metro lines (transfer convenience) or for densely distributed metro stations (central districts);
5. Attributes related to the accessibility of sport and recreation in the neighbourhood, i.e., park2_area (parks area share in the second-grade cadastral area neighbour's area) in the OLS_II model and in MODEL11: park_meanDist_ca (average distance to the closest park in the cadastral area) and tennis_MeanDist7 (average distance to the closest tennis courts);
6. Attributes describing neighbourhood quality: MF2_share_builtUp, SF2_share_builtUp (percentage of MF multifamily and SF single-family development designated at low granularity (second-order adjacency to the cadastral area), noise1_dist (average minimum level of daytime traffic noise within the first-grade cadastral area neighbours divided by the distance to the city centre) and district_rank (composite index of districts ranking determined by Statistics Poland), though the latter two were in the TOP-10 only for models based on 18 attributes selected for OLS_II;
7. Attributes determining the availability of elementary schools in the neighbourhood, i.e., elementary_MeanDist7 (average distance to the nearest elementary schools), though, like noise1_dist and district_rank, the attribute elementary_MeanDist7 was not in the TOP 10 for the model including more variables.

In summary, for the location variables (L), the most relevant attributes were those describing the relationship (directly or indirectly) between the cadastral section's location to the city centre or local centres and the possibilities of reaching the centre—public transport commuting time or distances to the nearest metro stations. On the other hand, for the neighbourhood variables (N), the most relevant attributes were those describing the structure of development in the larger neighbourhood (second-order adjacency) and those related to the availability of parks and tennis courts. Additionally, the average noise level in the medium granulation (first-order adjacency) and the availability of primary schools were significant in the models with 18 variables selected.

5. Discussion

In hedonic models, prices or price indices are analysed on a non-aggregate basis. This requires the inclusion of real property structural variables. In a number of studies (e.g., [13,17,23]), location and neighbourhood variables showed negligible significance in the regression models created (both linear OLS and regression using ML algorithms) compared to the significance of structural variables. The analysis of location (L) and neighbourhood (N) variables, therefore, required searching for transactions with similar structural variables, differing only in external attributes. However, when analysing asking prices, this is difficult due to the smaller number of new properties on the market compared to the number of properties traded on the secondary market. In our research, we used price level indicators aggregated to micro-markets (cadastral sections). We showed that, on the basis of only the external attributes, namely location (L) and neighbourhood (N), it is possible to create models to predict these values. We obtained models with an R^2 of 71.1–86.7%, depending on the tools and algorithms used, with a prediction uncertainty RMSE of 0.099–0.136 for the test data, and with an R^2 of 73.9–84.1% and an RMSE of 0.095–0.121 for the prediction in the control dataset. To the best of our knowledge, no analysis has yet been carried out on prediction models for aggregate price level indicators on the primary market.

When comparing the results for housing price prediction, examples of R^2 values can be given. For instance, Ref. [47] developed a model based mainly on structural attributes and one location attribute, taking into account commuting time to the city centre for a Random Forest regression model for a housing price database in Hong Kong covering more than 90,000 records over 18 years, and obtained R^2 = 90.3%. In an analysis of apartment transaction data from the period of 2006 to 2017 in the district of Gangnam (South Korea) [31], achieved for RF-Regression R^2 = 97.6%, and for OLS regression R^2 = 72.6%. Ref. [65], in an analysis of a random sample of 200 houses in Christchurch, New Zealand, obtained for a hedonic price model for out-of-sample forecast evaluation R^2 = 38.1–75.0%, and for the neural network model R^2 = 69.1–90.0%. In an analysis of dwelling prices using Ordinary Least Squares Regression and Geographically Weighted Regression in Poznań, Poland, Ref. [69] obtained R^2 = 54.9%. In most of these studies, the authors point out the better results achieved with ML algorithms. In our study, concerning aggregate price level indices, for test data diagnosed with the k-folds method, ML models also performed better. On the other hand, OLS models, developed for carefully selected attributes, performed better for predictions in the control set, containing observations completely excluded from the model development stage.

On the basis of incomplete information about the structural attributes—the number of residential units in the development and to which of the four quality segments it was assigned—we determined the attributes referred to as semi-structural, namely the prestige and MeanCountDwellings. The attributes prestige_mid and prestige_f were calculated on the basis of assigning the development to a quality segment, which in turn is derived from the structural variables, among other things. A detailed description of the determination of the prestige_mid and prestige_f variables is provided in Appendix B. These variables showed a high correlation with aggregate price level indices (R^2: 62.8–65.7%). In a study [36], the number of apartments in a building was among the top 20 characteristics with the highest impact on prices, both using the Random Forest algorithm and linear regression. Similarly, in a study [31], the number of apartments in a building was the second most significant characteristic of the apartments analysed. In our study, we included a semi-structural attribute indicating the average number of apartments in new residential buildings located in the neighbourhood of a given cadastral section (MeanCountDwellings). This variable did not show a high correlation with the aggregate price level index analysed (R^2 = 7.5%). It should be noted that the way this variable is captured is quite different. We use an averaged value aggregated to the cadastral section. It does not characterise a single building, and we treat it as a neighbourhood attribute—indicating the direction of the developments in the neighbourhood. Other relevant structural variables include apartment area, type and age of development. However, when considering the area of the apartment, in our research, we had data aggregated to whole housing developments; when considering the type of development, we only analysed apartments in multifamily buildings; and for the age of the building, we used data from the primary market, so some housing developments were still under construction.

6. Conclusions

The inclusion of semi-structural variables improved the prediction results in both the test set and the control set for the OLS models and for most of the ML models developed. These variables are also highly ranked in terms of their significance. If there are no new housing projects in a certain area, adjacent first-order cadastral sections and adjacent second-order cadastral sections within the analysed time period, no information is available for the determination of semi-structural attribute values. In the developed OLS linear regression models, not including semi-structural variables resulted in a small difference in R^2 (1.2%), while in models using the Random Forest (RF) algorithm, the difference in R^2 was 5.8% for models developed with all attributes and 2.9% for models developed using selected attributes from OLS models. The prediction of aggregate asking price levels in the cadastral sections for which information to determine semi-structural attributes is missing is less certain, but still the level of explained variance for the prediction of the control set is about 75% for the ML models and over 82% in the OLS linear regression model. In the diagnostics of the models for the control set prediction, observations that were completely excluded from the model building process performed in favour of the OLS linear regression models. However, creating an OLS linear regression model that meets the assumptions is difficult due to collinearity occurring between many variables. By restricting the set to significant variables that have a VIF < 7.5 (i.e., to avoid increasing the variance with collinearity), there is a risk of omitting key variables. This led to flawed results when maximising R^2: the deviations did not have a normal distribution and there was spatial autocorrelation of the standardised deviations. Many studies [70–72] have shown that there are interdependencies between attributes belonging to groups such as greenery, location in the city structure and transport accessibility, and that their impact on real property prices is complex. This nonlinear nature of the relationship has been demonstrated on several occasions [73]. This phenomenon seems to be of particular relevance for the variables belonging to the location (L) and neighbourhood (N) attributes determined using GIS spatial analyses, which naturally exhibit relationships in geographical space. The use of ML algorithms allowed the lack of linearity to be taken into account. The standardised deviations obtained in the regression models based on ML regression had a random distribution over the area of analysis (no statistically significant spatial autocorrelation). However, for predictions in the control set, the results of the ML algorithms were subject to greater uncertainty.

The developed methodology can serve as the basis for developing predictive aggregate price level indices. These indices can be used to determine the level of asking prices in newly marketed developments. The accurate determination of price indices will allow development companies to pre-plan their budgets for new residential developments and site selection. It also allows the analysis of the impact of location (L) and neighbourhood (N) variables, without the need to include complex information on structural variables for which there is no detailed information available in the asking price database used.

Our research has some shortcomings. The aim of this paper was not to delve deeply into ML methods, nor was it to prove that ML methods outperform linear regression. We wanted to prove that external attributes (themselves) allow us to assess the price levels for the micro market. To do so, we decided to use different methods: linear regression and two ML methods (Random Forest and XGBoost—both available in ArcGIS Pro 3.3.1, which we used). We have not compared the various methods for tuning hyperparameters (such as grid search, random search, or Bayesian optimization). We have not explored spatial thresholds and heterogeneity of accessibility's impact on aggregate price indices. All these shortcomings indicate potential further directions of our research. Moreover, further research should focus on the application of moving window technique analysis (which can provide independence from a priori defined micro-markets), more detailed analysis of the impact of individual groups of location (L) and environment (N) attributes at a wider range of granularity, and the ongoing verification of selected models based on incoming data from the primary property market.

Author Contributions: Conceptualization, J.J. and H.H.; methodology, J.J. and H.H.; software, J.J. and H.H.; validation, J.J. and H.H.; formal analysis, J.J. and H.H.; investigation, J.J.; resources, H.H.; data curation, J.J. and H.H.; writing—original draft preparation, J.J. and H.H.; writing—review and editing, J.J. and H.H.; visualization, J.J.; supervision, J.J.; project administration, J.J.; funding acquisition, J.J. All authors have read and agreed to the published version of the manuscript.

Funding: This paper was financed under the research grant of the Warsaw University of Technology supporting the scientific activity in the discipline of Civil Engineering, Geodesy and Transport: 8/ILGiT/2023.

Data Availability Statement: Data derived from public domain resources: the data used for attribute determination were derived from the resources available in the public domain: Topographic data: https://www.geoportal.gov.pl/ (accessed on 6 April 2024), Statistical data: https://bdl.stat.gov.pl/bdl (accessed on 6 April 2024); 3rd Party Data: For the asking price database—data were made available for research by Jones Lang LaSalle Sp. z o.o. under an agreement concluded on 9 February 2023 between the Faculty of Geodesy and Cartography at Warsaw University of Technology and Jones Lang LaSalle Sp. z o.o. Restrictions apply to the availability of these data. Data were obtained from Jones Lang LaSalle Sp. z o.o and are available with the permission of Jones Lang LaSalle Sp. z o.o.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Variables Selected for Analysis After Initial Selection

Table A1. Variables selected for analysis after initial selection.

Name PL	Name EN	Category	Group	Determination Method	Granulation	R^2
park1_pow	park1_area	neighbourhood	green area	Parks area share in 1st grade neighbour cadastral areas	Contiguity—1st order	14.3
park2_pow	park2_area	neighbourhood	green area	Parks area share in 2nd grade neighbour cadastral areas	Contiguity—2nd order	27.1
zielen0_pow	green0_area	neighbourhood	green area	Greenery area share in cadastral area	Cadastral area	10.3
zielen1_pow	green1_area	neighbourhood	green area	Greenery area share in 1st grade neighbour cadastral areas	Contiguity—1st order	16.9
zielen2_pow	green2_area	neighbourhood	green area	Greenery area share in 2nd grade neighbour cadastral areas	Contiguity—2nd order	20.7
OBR2_park_las_zadrz_liczba	green2_count	neighbourhood	green area	Numer of greenery spots in 2nd grade neighbour cadastral areas	Contiguity—2nd order	11.5
INTENS_Zabud	density	neighbourhood	buildings	Buildings density	Cadastral area	35.6
zabytki_krnl2	landmarks_krnl2	neighbourhood	buildings	Relative kernel density calculated for monuments within 2nd grade neighbour cadastral areas divided by distance to the city center	Contiguity—2nd order	54.2

Table A1. *Cont.*

Name PL	Name EN	Category	Group	Determination Method	Granulation	R^2
BUBDshare	builtup_share	neighbourhood	buildings	Built-up area share in 2nd grade neighbour cadastral areas	Cadastral area	18.8
OSR	OSR	neighbourhood	buildings	Spacemate component value (spare outdoor space per person)	Cadastral area	24.0
MW2_udział _obreb	multifam_ share_ca	neighbourhood	buildings	Multifamily bult-up area share within 2nd grade neighbour cadastral areas	Contiguity—2nd order	37.7
MW2_udział_ zabudowa	multifam_ share_be	neighbourhood	buildings	Multifamily bult-up area share within 2nd grade neighbour cadastral built-up areas	Contiguity—2nd order	43.3
MN2_udział_ zabudowa	singlefam_ share_be	neighbourhood	buildings	Single-family bult-up area share within 2nd grade neighbour cadastral built-up areas	Contiguity—2nd order	45.8
Clm *	total_no_units *	neighbourhood	buildings	Average number of dwellings in new developments in cadastral area	Cadastral area	7.5
przedsz_krnl2	kindergarten _krnl2	neighbourhood	community services	Relative kernel density calculated for kindergartens within 2nd grade cadastral area neighbours	Contiguity—2nd order	30.4
podst_krnl2	elementary_krnl2	neighbourhood	community services	Relative kernel density calculated for elementary schools within 2nd grade cadastral area neighbours	Contiguity—2nd order	49.2
sredn_krnl2	highschool_krnl2	neighbourhood	community services	Relative kernel density calculated for high schools within 2nd grade cadastral area neighbours divided by distance to the city center	Contiguity—2nd order	56.6
eduindex2	eduindex2	neighbourhood	community services	Index calculated using number of different school facilities and number of types of schools	Contiguity—2nd order	34.9
przedsz2_pow	kindergarten2 _area	neighbourhood	community services	Number of kindergartens within 2nd grade cadastral area neighbours divided by area	Contiguity—2nd order	18.5
podst2_pow	elementary2_area	neighbourhood	community services	Number of elementary schools within 2nd grade cadastral area neighbours divided by area	Contiguity—2nd order	33.1
eduindex2_pow	eduindex2_area	neighbourhood	community services	Index calculated using number of different school facilities and number of types of schools divided by area	Contiguity—2nd order	39.0

Table A1. *Cont.*

Name PL	Name EN	Category	Group	Determination Method	Granulation	R^2
zdrowie_krnl2	healthcare_krnl2	neighbourhood	community services	Relative kernel density calculated for healthcare facilities within 2nd grade cadastral area neighbours divided by distance to the city center	Contiguity—2nd order	55.8
apteka_krnl2	pharmacy_krnl2	neighbourhood	community services	Relative kernel density calculated for pharmacies within 2nd grade cadastral area neighbours divided by distance to the city center	Contiguity—2nd order	54.7
bank_krnl2	bank_krnl2	neighbourhood	commercial services	Relative kernel density calculated for banks within 2nd grade cadastral area neighbours divided by distance to the city center	Contiguity—2nd order	59.6
poczta_krnl2	postoffice_krnl2	neighbourhood	commercial services	Relative kernel density calculated for post offices within 2nd grade cadastral area neighbours divided by distance to the city center	Contiguity—2nd order	54.7
rest_krnl2	restaurant_krnl2	neighbourhood	commercial services	Relative kernel density calculated for restaurants within 2nd grade cadastral area neighbours divided by distance to the city center	Contiguity—2nd order	53.4
ph_norm	retail_krnl2	neighbourhood	commercial services	Relative kernel density calculated for retail within 2nd grade cadastral area neighbours	Contiguity—2nd order	27.7
halas1_dist	noise1_dist	neighbourhood	disadvantages	Average minimum level of day-time traffic noise within 1st grade cadastral area neighbours divided by distance to city centre	Contiguity—1st order	49.5
dzielnica_ rankingGUS	district_rank	neighbourhood	prestige	Statistics Poland districts ranking	Cadastral area	17.6
seg_mid1 *	prestige_mid1 *	neighbourhood	prestige	Prestige index calculated within 1st grade cadastral area neighbours	Contiguity—1st order	62.8
seg_mid2 *	prestige_mid2 *	neighbourhood	prestige	Prestige index calculated within 2nd grade cadastral area neighbours	Contiguity—2nd order	65.3
prestiz_f1 *	prestige_f1 *	neighbourhood	prestige	Prestige index calculated within 1st grade cadastral area neighbours	Contiguity—1st order	64.0
prestiz_f2 *	prestige_f2 *	neighbourhood	prestige	Prestige index calculated within 2nd grade cadastral area neighbours	Contiguity—2nd order	65.7

Table A1. *Cont.*

Name PL	Name EN	Category	Group	Determination Method	Granulation	R^2
przystanki_ index_1	stop_index1	neighbourhood	public transport	Average distance to public transport stop divided by average distance to them within 1st grade cadastral area neighbours	Contiguity—1st order	21.4
przystanki_ gestosc_1	stops_density1	neighbourhood	public transport	Number of public transportation stops divided by 1st grade cadastral area neighbours area	Contiguity—1st order	26.2
MEAN_Kernel1_ Przyst_OBR	MEAN_Kernel1_ BusStop_dist	neighbourhood	public transport	Relative kernel density calculated for public transportation stops within 1st grade cadastral area neighbours	Contiguity—1st order	35.5
ParkMeanOBR	park_mean_ca	location	green area	Average distance to closest park in cadastral area	Distance	35.6
zabytki_srednio7	landmarks_dist	location	buildings	Average distance to closest monuments	Distance	22.8
przedsz_srednio7	kindergartens_dist	location	community services	Average distance to closest kindergarden facilities	Distance	17.5
podst_srednio7	elementary_dist	location	community services	Average distance to closest elementary school facilities	Distance	41.9
sredn_srednio7	highschool_dist	location	community services	Average distance to closest high school facilities	Distance	54.4
muzeum	museum	location	community services	Distance to closest museum	Distance	41.1
teatr	theatre	location	community services	Distance to closest theatre	Distance	46.8
zdrowie_srednio7	healthcare_dist	location	community services	Average distance to closest healthcare facilities	Distance	37.7
apteka_srednio7	pharmacy_dist	location	community services	Average distance to closest pharmacies	Distance	40.0
muzeum_ mediana4	museum_median	location	community services	Average distance to closest museums	Distance	54.3
teatr_srednio7	theatre_dist	location	community services	Average distance to closest theatres	Distance	54.0
bank_srednio7	bank_dist	location	commercial services	Average distance to closest bank facilities	Distance	52.3
poczta_srednio7	postoffice_dist	location	commercial services	Average distance to closest post office facilities	Distance	49.7
transformatory	transformers	location	disadvantages	Distance to closest energy transformers facility	Distance	7.6
slupy_7najb	transmission_dist	location	disadvantages	Distance to seventh closest transmission tower	Distance	0.7
slupy_7sredn	transmission_dist	location	disadvantages	Average distance to closest transmission towers	Distance	0.9
scieki	sewage_dist	location	disadvantages	Distance to closest sewege treatment plant facility	Distance	7.7

Table A1. *Cont.*

Name PL	Name EN	Category	Group	Determination Method	Granulation	R^2
zajezdnie	busdepot_dist	location	disadvantages	Distance to closest bus/tram depot	Distance	23.3
autostrady	highway_dist	location	transport	Distance to closest highway	Distance	15.7
plac_sportowy _srednio7	sport_dist	location	recreation facilities	Average distance to closest sport facilities	Distance	29.6
plac_gier_i_zabaw _srednio7	playground_dist	location	recreation facilities	Average distance to closest playground facilities	Distance	15.0
korty_tenisowe _srednio7	tennis_dist	location	recreation facilities	Average distance to closest tennis courts	Distance	30.1
ODL_kmDr	city_center_dist_roads	location	location in the city	Distance to city centre measured along roads	Distance	58.6
poziom1	local_center_dist	location	location in the city	Distance to local centre	Distance	6.0
metro_min2	metro_2nd_closest	location	public transport	Distance to second closest metro station	Distance	49.1
stacja_min2	railway_2nd_closest	location	public transport	Distance to second closest railway station (either metro or suburban)	Distance	21.1
przyst_mediana7	stops_median	location	public transport	Average distance to closest public transportation stops	Distance	16.0
MEAN_Kernel1 _Przyst_OBR	MEAN_Kernel1 _BusStop_dist	neighbourhood	public transport	Relative kernel density calculated for public transportation stops within 1st grade cadastral area neighbours	Contiguity—1st order	35.5
czas_Kom	city_center_time_public	location	public transport	Commuting time to city centre	Distance	54.6
metro3	metro_3rd_closest	location	public transport	Distance to third closest metro station	Distance	51.8
stacja3	railway_3rd_closest	location	public transport	Distance to third closest railway station (either metro or suburban)	Distance	29.6

* Variables that we refer to as semi-structural attributes are marked in grey.

Appendix B. The Method of Determination of Prestige Indexes

In the analysed database, investment stages are assigned to one of four quality segments: 1, 2, 3 and 4. Segment 1 includes the most luxurious residential developments, usually intimate, located in quiet, peaceful neighbourhoods or slightly larger, located in the most prestigious locations. They have additional services such as an in-house gym or concierge, and the façade and common areas are finished using the best materials. Sometimes turnkey finishing of such an apartment is on offer. Segment 4 can be called popular; these are investments always sold in a developer's state, thus requiring financial outlays from the buyer to finish the apartments. The façade and common areas are finished using cheap, readily available materials. The values of the variables were calculated with reference to the surface reference unit formed by the following:

- Cadastral sections and 1st order contiguous cadastral sections (prestige_f1, prestige_mid1);
- Cadastral sections, 1st order contiguous cadastral districts and 2nd order contiguous cadastral sections (prestige_f2, prestige_mid2).

They were calculated as follows:

Variables: prestige_mid. This is the share of the total number of apartments contained in housing developments assigned to one of the two middle-quality segments (lower-middle and upper-middle) in the total number of apartments contained in the new housing developments included in the database analysed (Equation (A1)).

$$prestige_{mid} = \frac{c_k(seg2 + seg3)}{c_k(seg1 + seg2 + seg3 + seg4)}, \tag{A1}$$

where c_k is the total number of apartments in the development stages in k of this surface reference unit; and $c_k(seg1)$, $c_k(seg2)$, $c_k(seg3)$, $c_k(seg4)$ are the number of apartments in the development stages in k—this surface reference unit, allocated to segment 1, segment 2, segment 3 and segment 4, respectively.

Variables: prestige_f is calculated as follows (Equation (A2)):

$$\begin{aligned}
&if\ C_k(seg1) > 0,\ than\ prestigef = 3 + \tfrac{C_k(seg1)+C_k(seg2)}{C_k},\\
&if\ C_k(seg1) = 0\ AND\ C_k(seg2) > 0\ than\ prestigef = 2 + \tfrac{C_k(seg2)+C_k(seg3)}{C_k},\\
&if\ C_k(seg1) = 0\ AND\ C_k(seg2) = 0\ AND\ C_k(seg3) > 0\ than\ prestigef = 1 + \tfrac{C_k(seg3)}{C_k},\\
&if\ C_k(seg1) = 0\ AND\ C_k(seg2) = 0\ AND\ C_k(seg3) = 0\ AND\ C_k(seg4) > 0\ than\ prestigef = 1,
\end{aligned} \tag{A2}$$

where C_k is the total number of apartments in the development stages in k of this surface reference unit; and $C_k(seg1)$, $C_k(seg2)$, $C_k(seg3)$, $C_k(seg4)$ are the number of apartments in the development stages in k—this surface reference unit, allocated to segment 1, segment 2, segment 3 and segment 4, respectively.

If the unit has investment stages belonging to segment 1, the variable prestige_f takes values from 3 to 4. If the highest segment is segment 2, the variable prestige_f takes values from 2 to 3. If the highest segment is segment 3, the variable prestige_f takes values from 1 to 2. If the unit has only investment stages assigned to segment 4, the variable prestige_f takes the value 1.

Appendix C. Variables Selected for OLS_I and OLS_II Models

Table A2. List of variables selected to create OLS_I and OLS_II models.

OLS_I	OLS_II
BANK_MEANDIST	BANK_MEANDIST
LOCAL_CENTER_DIST	LOCAL_CENTER_DIST
STOP_INDEX1	STOP_INDEX1
METRO_2ND_CLOSEST	METRO_2ND_CLOSEST
CITY_CENTER_TIME_PUBLIC	CITY_CENTER_TIME_PUBLIC
ELEMENTARY_MEANDIST7	ELEMENTARY_MEANDIST7
NOISE1_DIST	NOISE1_DIST
SEWAGE_DIST	SEWAGE_DIST
THEATRE	THEATRE
MUSEUM_MEANDIST4	PARK2_AREA
PARK2_AREA	GREEN2_AREA
GREEN2_AREA	SPORT_MEANDIST7
SPORT_MEANDIST7	TENNIS_MEANDIST7
TENNIS_MEANDIST7	LANDMARKS_MEANDIST7
LANDMARKS_MEANDIST7	DISTRICT_RANK
DISTRICT_RANK	MF2_SHARE_BUILTUP
MF2_SHARE_BUILTUP	RETAIL_KRNL2
RETAIL_KRNL2	HIGHWAY_DIST
PRESTIGE_F2 *	

* semi-structural variable.

Appendix D. Collection of Training and Control Observations

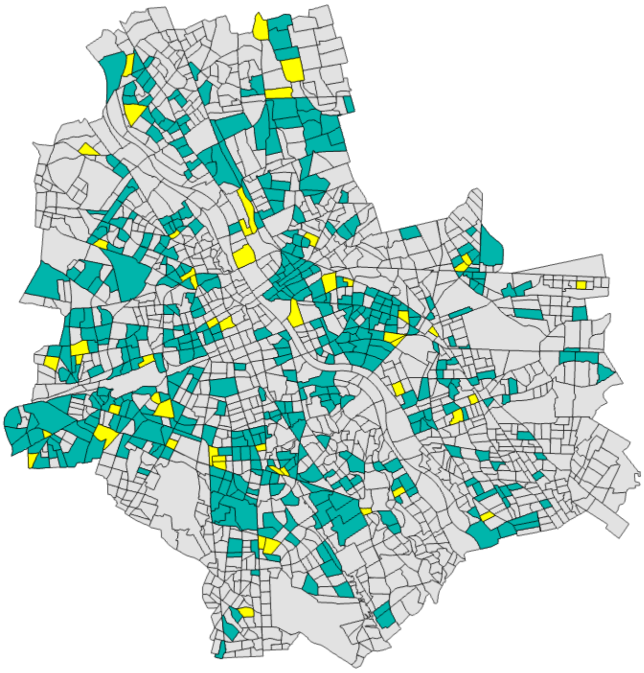


Figure A1. Training data and control data for the extrinsic validation of models. The number of cadastral districts: 1742—grey colour, the number of cadastral districts in training data: 451 (90%)—green colour, the number of control data: 52 (10%)—yellow colour.

References

1. Basu, S.; Thibodeau, T.G. Analysis of Spatial Autocorrelation in House Prices. *J. Real Estate Finance Econ.* **1998**, *17*, 61–85. [CrossRef]
2. Bourassa, S.C.; Hamelink, F.; Hoesli, M.; MacGregor, B.D. Defining Housing Submarkets. *J. Hous. Econ.* **1999**, *8*, 160–183. [CrossRef]
3. Helbich, M.; Brunauer, W.; Hagenauer, J.; Leitner, M. Data-Driven Regionalization of Housing Markets. *Ann. Assoc. Am. Geogr.* **2013**, *103*, 871–889. [CrossRef]
4. Ligus, M.; Peternek, P. Measuring Structural, Location and Environmental Effects: A Hedonic Analysis of Housing Market in Wroclaw, Poland. Poland. *Procedia Soc. Behav. Sci.* **2016**, *220*, 251–260. [CrossRef]
5. Peng, Z.; Inoue, R. Identifying Multiple Scales of Spatial Heterogeneity in Housing Prices Based on Eigenvector Spatial Filtering Approaches. *ISPRS Int. J. Geo-Inform.* **2022**, *11*, 283. [CrossRef]
6. Hoesli, M.; Malle, R. Commercial Real Estate Prices and COVID-19. *J. Eur. Real Estate Res.* **2022**, *15*, 295–306. [CrossRef]
7. Rosen, S. Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *J. Politi. Econ.* **1974**, *82*, 34–55. [CrossRef]
8. Tomczyk, E.; Widłak, M. Construction and properties of the hedonic housing price index for Warsaw. *Bank Kredyt* **2010**, *41*, 99–128. Available online: https://bazekon.uek.krakow.pl/gospodarka/163845744 (accessed on 10 May 2024). (In Polish).
9. Usman, H.; Lizam, M.; Burhan, B. Review of issues in the conventional hedonic property pricing model. In Proceedings of the 2nd African International Conference on Industrial Engineering and Operations Management, Harare, Zimbabwe, 7–10 December 2020; pp. 2806–2816. Available online: http://www.ieomsociety.org/harare2020/papers/631.pdf (accessed on 8 June 2024).
10. Vergara-Perucich, F. Testing Housing Price Drivers in Santiago de Chile: A Hedonic Price Approach. *Crit. Hous. Anal.* **2023**, *10*, 44–57. [CrossRef]
11. Aziz, A.; Anwar, M.M.; Abdo, H.G.; Almohamad, H.; Al Dughairi, A.A.; Al-Mutiry, M. Proximity to Neighborhood Services and Property Values in Urban Area: An Evaluation through the Hedonic Pricing Model. *Land* **2023**, *12*, 859. [CrossRef]
12. Król, A. Application of Hedonic Methods in Modelling Real Estate Prices in Poland. In *Data Science, Learning by Latent Structures, and Knowledge Discovery*; Lausen, B., Krolak-Schwerdt, S., Böhmer, M., Eds.; Springer: Berlin/Heidelberg, Germany, 2015; pp. 501–511. [CrossRef]
13. Din, A.; Hoesli, M.; Bender, A. Environmental Variables and Real Estate Prices. *Urban Stud.* **2001**, *38*, 1989–2000. [CrossRef]
14. Shi, D.; Guan, J.; Zurada, J.; Levitan, A.S. An Innovative Clustering Approach to Market Segmentation for Improved Price Prediction. *Int. Technol. Inf. Manag.* **2015**, *24*, 2. [CrossRef]

15. Sopranzetti, B.J. Hedonic Regression Models. In *Handbook of Financial Econometrics and Statistics*; Lee, C.-F., Lee, J.C., Eds.; Springer: New York, NY, USA, 2015; pp. 2119–2134. ISBN 978-1-4614-7750-1. [CrossRef]
16. Xiao, Y. Urban Configuration and House Price. In *Urban Morphology and Housing Market*; Xiao, Y., Ed.; Springer: Singapore, 2017; pp. 63–94. ISBN 978-981-10-2762-8.
17. Chikhmous, A.; Rahman, M.T. Examining the Effect of Apartment Attributes on Their Sale Prices in Riyadh, Saudi Arabia. *Spat. Inf. Res.* **2024**, *32*, 411–424. [CrossRef]
18. Rey-Blanco, D.; Zofío, J.L.; González-Arias, J. Improving Hedonic Housing Price Models by Integrating Optimal Accessibility Indices into Regression and Random Forest Analyses. *Expert Syst. Appl.* **2024**, *235*, 121059. [CrossRef]
19. Heyman, A.V.; Law, S.; Berghauser Pont, M. How Is Location Measured in Housing Valuation? A Systematic Review of Accessibility Specifications in Hedonic Price Models. *Urban Sci.* **2018**, *3*, 3. [CrossRef]
20. Geerts, M.; De Weerdt, J. A Survey of Methods and Input Data Types for House Price Prediction. *ISPRS Int. J. Geo-Inf.* **2023**, *12*, 200. [CrossRef]
21. Yoo, S.; Im, J.; Wagner, J.E. Variable Selection for Hedonic Model Using Machine Learning Approaches: A Case Study in Onondaga County, NY. *Landsc. Urban Plan.* **2012**, *107*, 293–306. [CrossRef]
22. von Graevenitz, K.; Panduro, T.E. An Alternative to the Standard Spatial Econometric Approaches in Hedonic House Price Models. *Land Econ.* **2015**, *91*, 386–409. [CrossRef]
23. Ottensmann, J.R.; Payton, S.; Man, J. Urban Location and Housing Prices within a Hedonic Model. *J. Reg. Anal. Policy* **2008**, *38*, 19–35.
24. Anselin, L. Spatial Externalities, Spatial Multipliers, And Spatial Econometrics. *Int. Reg. Sci. Rev.* **2003**, *26*, 153–166. [CrossRef]
25. Beron, K.J.; Hanson, Y.; Murdoch, J.C.; Thayer, M.A. Hedonic Price Functions and Spatial Dependence: Implications for the Demand for Urban Air Quality. In *Advances in Spatial Econometrics*; Anselin, L., Florax, R.J.G.M., Rey, S.J., Eds.; Advances in Spatial Science; Springer: Berlin/Heidelberg, Germany, 2004; pp. 267–281. ISBN 978-3-642-07838-5.
26. Lo, D.; Chau, K.W.; Wong, S.K.; McCord, M.; Haran, M. Factors Affecting Spatial Autocorrelation in Residential Property Prices. *Land* **2022**, *11*, 931. [CrossRef]
27. Barreca, A.; Curto, R.; Rolando, D. Housing Vulnerability and Property Prices: Spatial Analyses in the Turin Real Estate Market. *Sustainability* **2018**, *10*, 3068. [CrossRef]
28. Cellmer, R.; Cichulska, A.; Bełej, M. Spatial Analysis of Housing Prices and Market Activity with the Geographically Weighted Regression. *SPRS Int. J. Geo-Inf.* **2020**, *9*, 380. [CrossRef]
29. Lynch, K. *The Image of the City*; The MIT Press: Cambridge, MA, USA, 1960; ISBN 978-0-262-62001-7.
30. Páez, A.; Long, F.; Farber, S. Moving Window Approaches for Hedonic Price Estimation: An Empirical Comparison of Modelling Techniques. *Urban Stud.* **2008**, *45*, 1565–1581. [CrossRef]
31. Hong, J.; Choi, H.; Kim, W.-S. A House Price Valuation Based on the Random Forest Approach: The Mass Appraisal of Residential Property in South Korea. *Int. J. Strat. Prop. Manag.* **2020**, *24*, 140–152. [CrossRef]
32. Arslanlı, K.Y. Analysis of House Prices: A Hedonic Model Proposal for Istanbul Metropolitan Area. *J. Des. Resil. Arch. Plan.* **2020**, *1*, 57–68. [CrossRef]
33. Burhan, B.; Kazunori, H.; Diah, M.L.M. Temporal Aggregate Effects in Hedonic Price Analysis. In Proceedings of the 19th Annual PRRES Conference, Melbourne, Australia, 13–16 January 2013.
34. Bourassa, S.C.; Cantoni, E.; Hoesli, M. Spatial Dependence, Housing Submarkets, and House Price Prediction. *J. Real Estate Finance Econ.* **2007**, *35*, 143–160. [CrossRef]
35. Helbich, M.; Brunauer, W.; Vaz, E.; Nijkamp, P. Spatial Heterogeneity in Hedonic House Price Models: The Case of Austria. *Urban Stud.* **2014**, *51*, 390–411. [CrossRef]
36. Rico-Juan, J.R.; de La Paz, P.T. Machine Learning with Explainability or Spatial Hedonics Tools? An Analysis of the Asking Prices in the Housing Market in Alicante, Spain. *Expert Syst. Appl.* **2021**, *171*, 114590. [CrossRef]
37. Kryvobokov, M.; Wilhelmsson, M. Analysing Location Attributes with a Hedonic Model for Apartment Prices in Donetsk, Ukraine. *Int. J. Strat. Prop. Manag.* **2007**, *11*, 157–178. [CrossRef]
38. Choi, K.; Park, H.J.; Uribe, F.A. The Impact of Light Rail Transit Station Area Development on Residential Property Values in Calgary, Canada: Focus on Land Use Diversity and Activity Opportunities. *Case Stud. Transp. Policy* **2023**, *12*, 100924. [CrossRef]
39. Diewert, W.E.; Shimizu, C. Residential Property Price Indexes: Spatial Coordinates Versus Neighborhood Dummy Variables. *Rev. Income Wealth* **2022**, *68*, 770–796. [CrossRef]
40. Hjort, A.; Pensar, J.; Scheel, I.; Sommervoll, D.E. House Price Prediction with Gradient Boosted Trees under Different Loss Functions. *J. Prop. Res.* **2022**, *39*, 338–364. [CrossRef]
41. Herath, S.; Maier, G. The Hedonic Price Method in Real Estate and Housing Market Research: A Review of the Literature. Faculty of Business—Papers (Archive). 2010, pp. 1–21. Available online: https://ro.uow.edu.au/buspapers/971/ (accessed on 15 August 2024).
42. Guo, B.; Li, K.; Fu, C. Utilizing Multilevel Modeling to Measure Neighborhood Dynamics and Their Impact on House Prices. *Appl. Sci.* **2023**, *13*, 5180. [CrossRef]
43. Welch, T.F.; Gehrke, S.R.; Wang, F. Long-Term Impact of Network Access to Bike Facilities and Public Transit Stations on Housing Sales Prices in Portland, Oregon. *J. Transp. Geogr.* **2016**, *54*, 264–272. [CrossRef]

44. Berawi, M.A.; Miraj, P.; Saroji, G.; Sari, M. Impact of Rail Transit Station Proximity to Commercial Property Prices: Utilizing Big Data in Urban Real Estate. *J. Big Data* **2020**, *7*, 71. [CrossRef]
45. Osland, L. The Importance of Unobserved Attributes in Hedonic House Price Models. *Int. J. Hous. Mark. Anal.* **2013**, *6*, 63–78. [CrossRef]
46. Liu, X.; Kounadi, O.; Zurita-Milla, R. Incorporating Spatial Autocorrelation in Machine Learning Models Using Spatial Lag and Eigenvector Spatial Filtering Features. *ISPRS Int. J. Geo-Inform.* **2022**, *11*, 242. [CrossRef]
47. Ho, W.K.O.; Tang, B.-S.; Wong, S.W. Predicting Property Prices with Machine Learning Algorithms. *J. Prop. Res.* **2021**, *38*, 48–70. [CrossRef]
48. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; Association for Computing Machinery: New York, NY, USA, 2016; pp. 785–794.
49. Beimer, J.; Francke, M. Out-of-Sample House Price Prediction by Hedonic Price Models and Machine Learning Algorithms. *Real Estate Res. Q.* **2019**, *18*, 13–20.
50. Siwicki, D. The Application of Machine Learning Algorithms for Spatial Analysis: Predicting of Real Estate Prices in Warsaw. Working Papers. 2021. Available online: https://ideas.repec.org/p/war/wpaper/2021-05.html (accessed on 10 June 2023).
51. Zaki, J.; Nayyar, A.; Dalal, S.; Ali, Z.H. House Price Prediction Using Hedonic Pricing Model and Machine Learning Techniques. *Concurr. Comput. Pract. Exp.* **2022**, *34*, e7342. [CrossRef]
52. Čeh, M.; Kilibarda, M.; Lisec, A.; Bajat, B. Estimating the Performance of Random Forest versus Multiple Regression for Predicting Prices of the Apartments. *ISPRS Int. J. Geo-Inform.* **2018**, *7*, 168. [CrossRef]
53. Fonseca, P.; Jardim, B.; de Castro Neto, M. Hedonic Model, Random Forest and Artificial Neural Network: Comparison for Real Estate Price Prediction in Lisbon. Available online: https://doi.org/10.2139/ssrn.4363508 (accessed on 5 November 2024).
54. Dimopoulos, T.; Bakas, N. Sensitivity Analysis of Machine Learning Models for the Mass Appraisal of Real Estate. Case Study of Residential Units in Nicosia, Cyprus. *Remote. Sens.* **2019**, *11*, 3047. [CrossRef]
55. Tchuente, D. Real Estate Automated Valuation Model with Explainable Artificial Intelligence Based on Shapley Values. *J. Real Estate Finance Econ.* **2024**, 1–39. [CrossRef]
56. García Pozo, A. A Nested Housing Market Structure: Additional Evidence. *Hous. Stud.* **2009**, *24*, 373–395. [CrossRef]
57. Zhou, Z.; Chen, H.; Han, L.; Zhang, A. The Effect of a Subway on House Prices: Evidence from Shanghai. *Real Estate Econ.* **2021**, *49*, 199–234. [CrossRef]
58. GUS. Ranking of Warsaw Districts According to the Attractiveness of Living Conditions. Available online: https://warszawa.stat.gov.pl/en/publications/others/ranking-of-warsaw-districts-according-to-the-attractiveness-of-living-conditions,2,1.html (accessed on 3 October 2023).
59. Moreno, C.; Allam, Z.; Chabaud, D.; Gall, C.; Pratlong, F. Introducing the "15-Minute City": Sustainability, Resilience and Place Identity in Future Post-Pandemic Cities. *Smart Cities* **2021**, *4*, 93–111. [CrossRef]
60. Jaroszewicz, J.; Denis, M.; Fijałkowska, A.; Graszka, O.; Pluto-Kossakowska, J.; Krzysztofowicz, S. Spatially Explicit Mixed-Use Indicators to Measure Life Quality across the City—A Conceptual Framework and Case Study: Piaseczno—A Medium Sized City in the Peri-Urban Zone of Warsaw, Poland. *Cities* **2023**, *137*, 104296. [CrossRef]
61. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
62. Probst, P.; Boulesteix, A.-L.; Bischl, B. Tunability: Importance of Hyperparameters of Machine Learning Algorithms. *J. Mach. Learn. Res.* **2019**, *20*, 1–32.
63. Boehmke, B.; Greenwell, B.M. *Hands-On Machine Learning with R*; Chapman and Hall/CRC: New York, NY, USA, 2019; ISBN 978-0-367-81637-7.
64. Greenwell, B.; Wu, Q. A Review of Methods Used in Machine Learning and Data Analysis. In Proceedings of the 2019 11th International Conference on Machine Learning and Computing, Zhuhai, China, 22–24 February 2019; Association for Computing Machinery: New York, NY, USA, 2019; pp. 43–51.
65. Sugiura, N. Further Analysis of the Data by Akaike's Information Criterion and the Finite Corrections: Further Analysis of the Data by Akaike's. *Commun. Stat. Theory Methods* **1978**, *7*, 13–26. [CrossRef]
66. Jarque, C.M.; Bera, A.K. A Test for Normality of Observations and Regression Residuals. *Int. Stat. Rev.* **1987**, *55*, 163–172. [CrossRef]
67. Koenker, R. A Note on Studentizing a Test for Heteroscedasticity. *J. Econ.* **1981**, *17*, 107–112. [CrossRef]
68. Grömping, U. Variable Importance Assessment in Regression: Linear Regression versus Random Forest. *Am. Stat.* **2009**, *63*, 308–319. [CrossRef]
69. Chwiałkowski, C.; Zydroń, A.; Kayzer, D. Assessing the Impact of Selected Attributes on Dwelling Prices Using Ordinary Least Squares Regression and Geographically Weighted Regression: A Case Study in Poznań, Poland. *Land* **2022**, *12*, 125. [CrossRef]
70. Liebelt, V.; Bartke, S.; Schwarz, N. Urban Green Spaces and Housing Prices: An Alternative Perspective. *Sustainability* **2019**, *11*, 3707. [CrossRef]
71. Guan, C.; Tan, M.J.; Peiser, R. Spatiotemporal Effects of Proximity to Metro Extension on Housing Price Dynamics in Manhattan, New York City. *J. Transp. Land Use* **2021**, *14*, 1295–1315. [CrossRef]

72. Keeler, Z.T.; Stephens, H.M. The Capitalization of Metro Rail Access in Urban Housing Markets. *Real Estate Econ.* **2023**, *51*, 686–720. [CrossRef]

73. Song, Y.; Zhang, S.; Deng, W. Nonlinear Hierarchical Effects of Housing Prices and Built Environment Based on Multiscale Life Circle—A Case Study of Chengdu. *ISPRS Int. J. Geo-Inform.* **2023**, *12*, 371. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.