sustainability

MDPI

Article

Statistical Modelling of the Market Value of Dwellings, on the Example of the City of Kraków

Elżbieta Jasińska and Edward Preweda *

Department of Integrated Geodesy and Cartography, AGH University of Science and Technology, 30-059 Krakow, Poland; elzbieta.jasinska@agh.edu.pl
* Correspondence: edward.preweda@agh.edu.pl; Tel.: +48-126172251

Abstract: The analysis of a city's spatial development, in terms of a location that meets the needs of its inhabitants, requires many approaches. The preliminary assessment of the collected material showed that there was real estate in the database whose price did not have market characteristics. For the correct formulation of the valuation model, it is necessary to detect and eliminate or reduce the impact of these properties on the valuation results. In this study, multivariate analysis was used and three methods of detecting outliers were verified. The database of 8812 residential premises traded on the primary market in Kraków was analyzed. In order to detect outliers, the following indices were determined: projection matrix, Mahalanobis distances, standardized chi test and Cook distances. Critical values were calculated based on the formulas proposed in the publication. The probability level was P = 0.95. The article shows that the selected methods of eliminating outliers—the methods of standardized residuals and the Cook's distance method give similar regression models. Further analysis (with the use of classification tree methods) made it possible to distinguish zones that are homogeneous in terms of price dispersion. In these zones, a set of features influencing real estate prices were determined.

Keywords: outlier observations; Cook's distance; statistical analysis; classification tree; real estate valuation model

Citation: Jasińska, E.; Preweda, E. Statistical Modelling of the Market Value of Dwellings, on the Example of the City of Kraków. *Sustainability* **2021**, *13*, 9339. https://doi.org/10.3390/su13169339

Academic Editors: Pierfrancesco De Paola and Brian Deal

Received: 3 July 2021
Accepted: 16 August 2021
Published: 20 August 2021

1. Introduction

Kraków is the second largest city in Poland by population and is undergoing a housing boom, particularly in apartment development. It is an important centre in Poland for many sectors of the economy, including higher education, start-ups, outsourcing, business, tourism and culture. The motive of some of the purchasers of apartments is to derive income from renting, attracted by its high profitability compared to low-interest bank deposits. The average rental yield was over three times higher than that gained from bank deposits. Today, the real estate market is facing the COVID-19 (coronavirus) pandemic that has been ongoing for several months [1]. The tourism industry has been particularly affected by the pandemic, which has meant that private short-term rentals have suffered [2]. Long-term rentals have also deteriorated because of the increasing popularity of remote learning and working. Rental prices in large agglomerations have decreased significantly. Most of the apartments for rent are empty so their owners often consider selling. Thus, the supply on the secondary market is increasing [1].

The housing market provides random information. Random factors disaggregate into the type of characteristics that shape the market value, which obliges the application of statistical rules. One of these features is the location of the property [3]. The work in the present paper characterizes the original market of residential real estate in Kraków. The research, covering the period 2015–2019, aims to determine the market processes in 18 districts of the city (Figure 1).

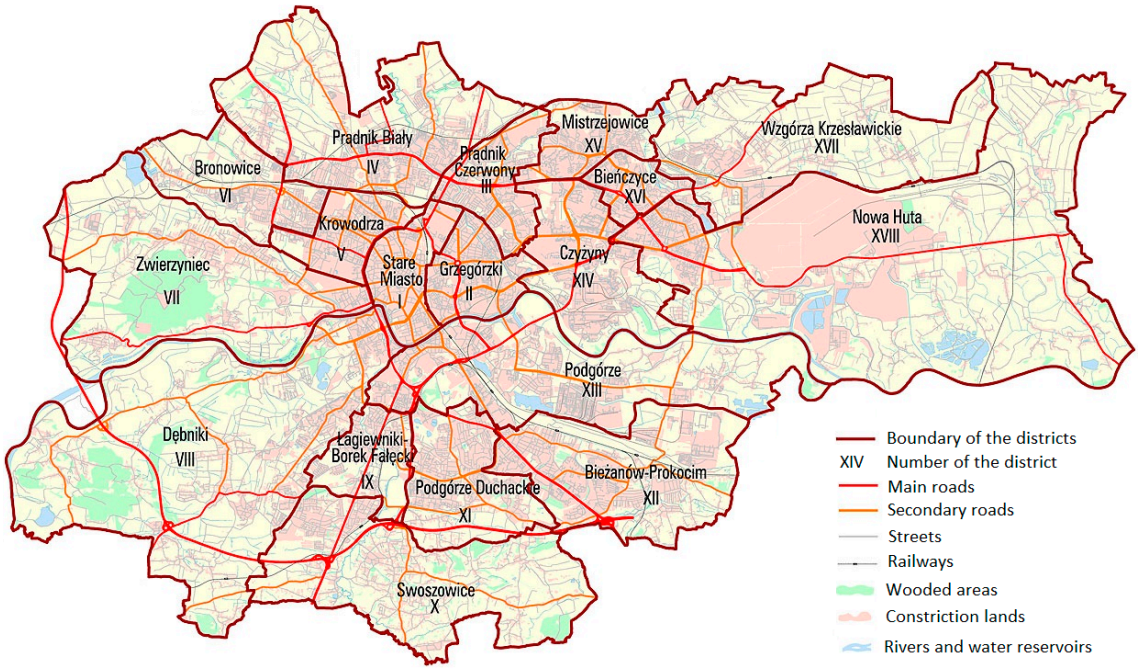

Figure 1. District of Kraków [4].

One of the most important market process is to estimate the value of real estate in individual city zones. Short descriptions of the zones are written below (based on [5] and subpages linked in the text on this website).

District I, Stare Miasto (English name: the Old Town) is the oldest district of Krakow, the heart of the city. Because of the large number of monuments, there is the greatest tourist traffic here. The district has been inscribed on the UNESCO World Heritage List. The Old Town is also the center of entertainment and cultural events in Krakow. There are many clubs, pubs, cafes and restaurants around the Main Square. In the Old Town, there is the Regional Bus Station and the Main Railway Station. Apartments in this district are among the most expensive in the Krakow real estate market.

District II Grzegórzki (English name is not used)—a small district next to the Old Town from the east. The construction of Grzegórzki is extremely diverse. There are many historic buildings on their premises. The area in the center of Krakow attracts more residents and developers who introduce modern buildings to the district. The biggest disadvantages of the district, apart from high prices, include too fast development and traffic jams at rush hour in some regions. However, this does not adversely affect the popularity of Grzegórzki. The prices of apartments in this district are in third place, right after the Old Town and Zwierzyniec.

III Prądnik Czerwony—next to Grzegórzki and the Old Town, it is in Śródmieście, the former district of the city. It is located close to the center, but away from the hustle and bustle of the city. High residential buildings and modern office buildings dominate here, with lower buildings in several places. There are many green and recreational areas in Prądnik Czerwony. The district is very well connected to the center of Krakow. The biggest disadvantages of the district are traffic jams and air quality. Despite this, Prądnik Czerwony is very popular among both people looking for a flat and developers. Prices in the district are lower compared to other districts of Krakow, and the location is exceptionally favorable.

IV Prądnik Biały—the most populous district of Krakow in the city's north. It is one of the most dynamically developing districts of Krakow. Every year there are more and more housing investments in its area. Prądnik Biały used to be the bedroom of Krakow, now it is becoming one of the most modern districts. There are many green areas in the district where the residents like to spend their time. Unfortunately, there are also exit streets to other cities, which increase the volume of traffic. This has a negative impact on the air

quality, which frequently exceeds the dust norms. For some, Prądnik Biały may also be too quiet a district, as there is a shortage of pubs and nightclubs.

V Krowodrza—because of the proximity of the Old Town and the AGH campus, Krowodrza is quite a busy district of Krakow. The nearby Błonia (large grassy area surrounded by bicycle paths) and the beautiful Jordan Park are an ideal place for walks and sports. The district is characterized by a very large number of new residential investments. It is a kind of compromise between the bustling Old Town and the quieter surroundings that are ideal for the elderly and families with children. As a result, real estate prices in Krowodrza are much higher than the average.

VI Bronowice—a sparsely populated district on the outskirts of the city. It is currently one of the most popular districts of Krakow. Residents appreciate this district, it is a more intimate neighborhood, adapted mainly to families with children and people looking for peace. There are many green and recreational places. Unfortunately, the district suffers from an insufficient number of nurseries, kindergartens and cultural centers. In Bronowice, there are both modern apartments in apartment buildings, slightly larger and cheaper premises in blocks of flats, as well as single-family houses. The district is very well connected with the rest of Krakow. VII Zwierzyniec—is one of the greenest, most prestigious and the most beautiful districts of Krakow. In its area there are, among others, two mounds: Kościuszko and Piłsudski, the monastery complex in Salwator, the Zoo, Lasek Wolski and Błonia. The district can boast excellent medical and educational facilities, interesting premises and interesting events. The Zwierzyniec housing estates are very diverse. There are both compact urban buildings, modern blocks and villas. Typically, large-city construction is only found in a small area. Zwierzyniec is the least populated district in Krakow. Apartment prices are second only to the Old Town.

VIII Dębniki—one of the prettier and quieter districts of Krakow. Dębniki is on the southern side of the Vistula River, on which it borders and which separates it from Zwierzyniec and the Old Town. The proximity to the city center makes it an interesting location for living. The air here is much cleaner than in the center of Krakow, which makes it possible to stay outdoors more often. Currently, many new investments, housing estates, educational, entertainment, service and commercial facilities are being built here. The district has a wonderful connection with other districts through many tram and bus lines. Dębniki is one of the most dynamically developing districts of Krakow. Developers currently offer thousands of apartments for sale here.

IX Łagiewniki—Borek Fałęcki—a district known mainly for the Sanctuary of Divine Mercy, to which Catholics make pilgrimages. It is characterized by vast green areas and a recently launched suburban railway. The buildings are diversified, but the farther from the city center, the faster it moves from multi-story apartment blocks to single-family houses.

X Swoszowice—a district built up with single-family houses, and is rather quiet. It is not of particular interest to real estate market participants rather than tourists, although there is a health resort in this district and council flats have been built there in recent years (because of cheap land); it is a district on the border of Krakow.

XIII Podgórze—one of the oldest districts of Krakow, although it was connected to Krakow in 1915. The area of mainly pre-war buildings, modernized over time, supplemented with new facilities in recent years. Next to Kazimierz, it is a painful memory of the Second World War (the forced labor camp in Płaszów, the Kraków Ghetto). It is a well-urbanized area, enabling the use of local transport; it is full of parks (Bednarski Park, Jordan Park) and green areas (the former Korona stadium, area next to Vistula River) and is one of the pleasant parts of Krakow, both for residents and tourists. XI Podgórze Duchackie, XII Bieżanów—Prokocim—two districts built in a similar period (1950–1980 of the 20th century), with the assumption of providing housing facilities for the developing Krakow. They have good public transport and green areas. However, it is mainly about housing estates, not parks. Bieżanów is connected to the city center by a suburban railway, which significantly shortens the transport time.

XIV Czyżyny—there are 5–15-storey multi-family apartment blocks with pure social infrastructure (kindergarten, school, playground, medical clinic, park), as well as the Polish Aviation Museum, which has been open since 2003. The aviation park is a large green area enjoying great popularity by residents. However, recent housing investments are assessed negatively—there is a lack of green areas and parking spaces.

XV Mistrzejowice, XVI Bieńczyce, XVII Wzgórza Krzesławickie—districts of single-family houses with a small share in the Krakow real estate market.

XVIII Nowa Huta—the largest district of Krakow in terms of area. Designed in the 1950s, it is well thought-out for housing. Low (4-storey) blocks are surrounded by advantages of the district, including greenery, very good city communication, parks, a lagoon, green areas. It was once intended to serve the employees of Huta im. Tadeusz Lenin (later Sędzimir). Although it is at a distance from the center of Krakow, it is practically self-sufficient and now and then you hear voices about the proposal to disconnect it from Krakow.

The basic source of data on real estate in Poland is data from the Price Register, but the information contained often requires appropriate analysis before starting the study of statistical relationships [6–8]. The analysis is based on data from the Register of Prices. This is a public register which contains data on the prices of immovable property specified in the notarial deeds and the value of the property as estimated by the property valuers. Over 9000 (exactly 9312) residential properties from the primary market were surveyed. 500 were eliminated because of incomplete data (8812 left). The reason for removing the property from the database was the lack of information about the location (25%), transaction date (25%), usable area (20%), and number of rooms or storeys (30%). The preliminary analysis of the collected material showed that the database registered observed values since 2015. Outliers may be caused by errors in the data or may be present because the set contains unusual observations, for example, motions intended for speculative purposes, with very small or very large areas compared to the cut value of the set [9]. A large number of outliers may also indicate incorrect model selection. Methods and estimators based on the assumption of a normal distribution and linear relationships are particularly resistant to outliers, so it is necessary to remove them from the set or minimize their impact [10]. The problem of rejecting outlier properties is not limited only to the study of price volatility, it also finds its way into the analyses of consolidation and replacement of land [11,12]. The identification of outliers was carried out based on the rest of the model [13], using the Mahalanobis distance metric [14,15] and by determining Cook's distance [16,17]. The Mahalanobis distances were used for real estate market analyses, mainly as a verification of bank portfolios [18,19], however, the authors did not find the application presented in this publication [20,21]. The method of the smallest squares was determined by the regression model for individual districts of the city, before and after the elimination of outlier observations. The regression and classification measures used in the further part of the work confirmed the results of analyses carried out by multiple regression. Hedonic modeling of real estate values is the subject of many studies [22–26]. Some of them concern the detection of outliers with selected methods, e.g., [18–20], others use solutions minimizing the impact of outliers [27,28]. A certain gap in the analyzed works is the lack of a precise definition of the criterion allowing for the recognition of the real estate as an outlier. This applies especially to Cook's distances in real estate market analysis. Based on the Fisher–Snedecor distribution, the authors precisely defined the Cook's distances criterion.

The article uses 4 methods of detecting outliers; we present them in points 2.2–2.5. Section 2.6 presents classification and regression trees. In Section 3, outliers were identified and C&RT and CHAID tree models were constructed after eliminating outliers from the base. The conclusions can be found in Section 4.

2. Materials and Methods

2.1. Multidimensional Regression

Forecasting the market value of a property is often carried out according to the classic multi-regression model. According to the classic linear regression model, explanatory variables should be correlated with the explained variable and not correlated with each other. For a simple model with one y-explained variable and two explanatory variables, the following dependencies should occur: (x_1, x_2)

$$\operatorname{cov}(y, x_1) \neq 0 \qquad \operatorname{cov}(y, x_2) \neq 0 \qquad \operatorname{cov}(x_1, x_2) = 0 \tag{1}$$

In practice, these assumptions are very rarely met. Real estate data are always correlated to some extent, so regressors are colinear. The variance of each model estimator can be saved as:

$$V(b_j) = \frac{\sigma^2}{(1 - r_{12}^2) \times \sum_{i=1}^{n} (x_{i,j} - \overline{x}_j)^2} = \frac{\sigma^2}{(1 - r_{12}^2) \times V_j} \qquad (j = 1 \ldots 2) \tag{2}$$

If the explanatory variables of the model are strongly correlated (the correlation coefficient tends to 1), then the estimator variance tends to infinity. Equation (2) can be generalized to multiple explanatory variables. If it is a vector of explanatory variables and a correlation coefficient of the *k*th regressor with the others, then the variance of the estimator can be saved as: $(x_1, x_2, \ldots, x_k) r_k^2 b_k$

$$V(b_j) = \frac{\sigma^2}{\left(1 - r_{j.}^2\right) \times \sum_{i=1}^{n} (x_{i,j} - \overline{x}_j)^2} = \frac{\sigma^2}{\left(1 - r_{j.}^2\right) \times V_j} \qquad (j = 1 \ldots k) \tag{3}$$

It follows that the variance of the parameter estimator increases with the correlation between the *j*th regressor and the others and decreases with the variance of the *j*th variable. In practice, there are often cases where there is a relationship between explanatory variables, but this is not a strictly linear relationship. Then, although the assumptions of the classic linear regression model are met (as we are dealing with a linear relationship, the number of observations must be greater than or equal to the number of parameters derived from the regression analysis, the variance of the residuals, the random component is the same for all observations, there is no autocorrelation of residuals, residuals have a distribution close to the normal distribution, and there is no predictor collinearity) [13,29,30]. There are other problems, the most important of which are two:

1. small changes in the database result in large changes in the value of estimators;
2. regression equation coefficients have large standard deviations, thus they may be statistically insignificant, despite even a high R^2 determination factor (together they are relevant).

Both of these problems may be due to outliers in the database. In regression analysis, we mean atypical values of explanatory (independent) variables, unusual values of a dependent variable (explained), or unusual values for both variables. Outliers can be caused by data errors, such as mistakes when entering information in the property price and value register. They may also exist because the database contains unusual observations, for example, properties with very small or very large areas compared to the average value of the set. Methods and estimators based on the assumption of a normal distribution and linear dependencies are particularly resistant to outliers, so it is necessary to remove them from the set or minimize their impact. In the case of linear multidimensional regression, diagnostic tests to detect outliers are most common: standardized model residual analysis, Mahalanobis distance and Cook's distance. It has been noted that on different databases, these tests detect observations as outliers slightly differently, although all are based on a similar principle.

To detect outliers, descriptive statistics were set out in the first stage, broken down by the district of Krakow (Table 1).

Table 1. Descriptive statistics of the database.

	Number of Properties	Average Unit Price [PLN/m^2]	Median [PLN/m^2]	Min. Unit Price [PLN/m^2]	Max. Unit Price [PLN/m^2]	Standard Deviation [PLN/m^2]
Bieńczyce	70	4662	4535	3733	6383	474
Bieżanów	711	5273	5210	3400	6805	478
Bronowice	106	6606	6718	5054	8321	605
Czyżyny	1029	5356	5311	3999	8279	813
Dębniki	937	6570	6173	2790	18,043	2107
Grzegórzki	1281	7634	7395	2166	15,688	1459
Krowodrza	464	7527	7389	4736	11,629	1015
Łagiewniki	99	6002	6246	2990	7516	992
Mistrzejowice	352	5116	5096	3999	6799	490
Nowa Huta	94	4507	4466	2735	5921	432
Podgórze	996	6787	6712	2510	11,979	1302
P. Duchackie	360	5855	5826	3367	7511	593
Prądnik Biały	1276	6204	6221	3585	12,006	825
Prądnik Czerwony	439	6111	5962	3276	9100	841
Stare Miasto	424	10,051	9473	2467	20,446	3030
Swoszowice	42	4907	4951	4298	6170	338
Wzgórza K.	35	4755	4490	2853	6258	890
Zwierzyniec	97	8709	8888	3100	13,393	1926

Figure 2 presents the boxplot, presenting information on the location, dispersion and shape of the distribution of data from individual districts of the city.

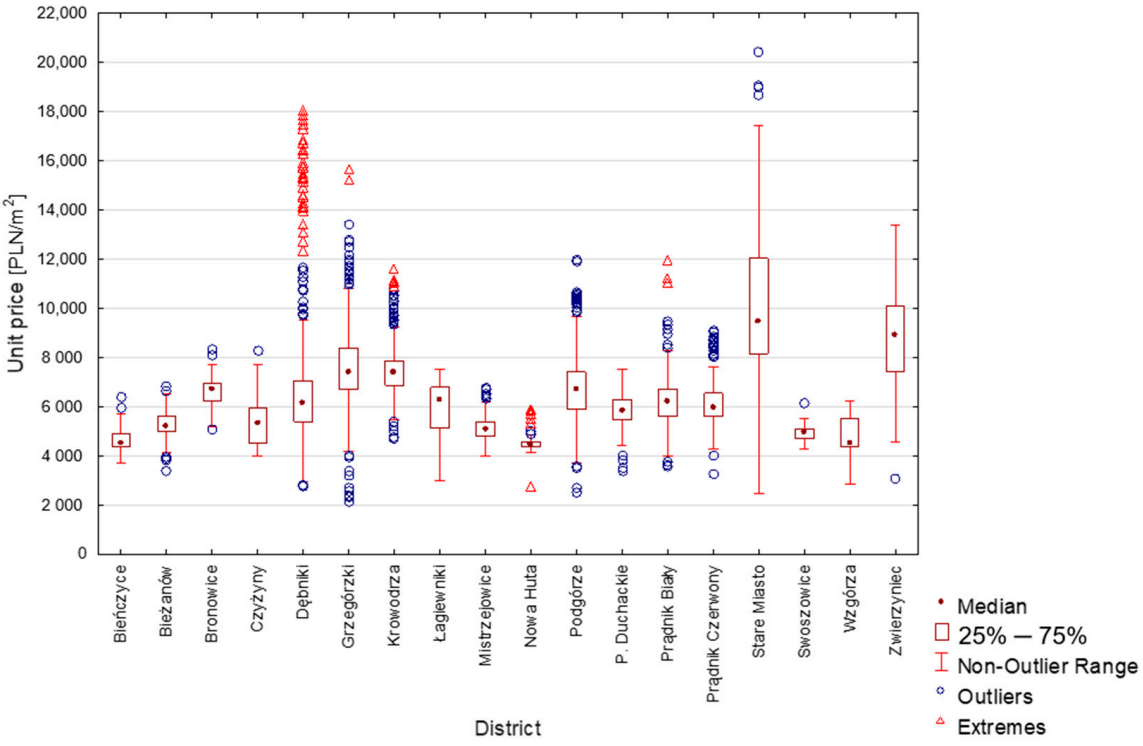

Figure 2. Box plot of unit price grouped by district—8812 properties.

Extreme and outlier unit prices are noted. These values must be verified and removed from the database so that they do not interfere with the further modelling process. It should be noted, however, that these values may be due to the characteristics of the property in question, which are more or less favourable than the average characteristic. In the first place, it is necessary to verify the statistic distributions of property prices in the following specific districts. The authors verified the hypothesis of normal distribution using the Shapiro–Wilk (SW-W) test, because of its eminent power compared to other tests, low sensitivity to autocorrelation and variance of variance. The hypothesis was verified for the entire city and separately for individual districts. Selected results of the analyses are presented in Figures 3–21.

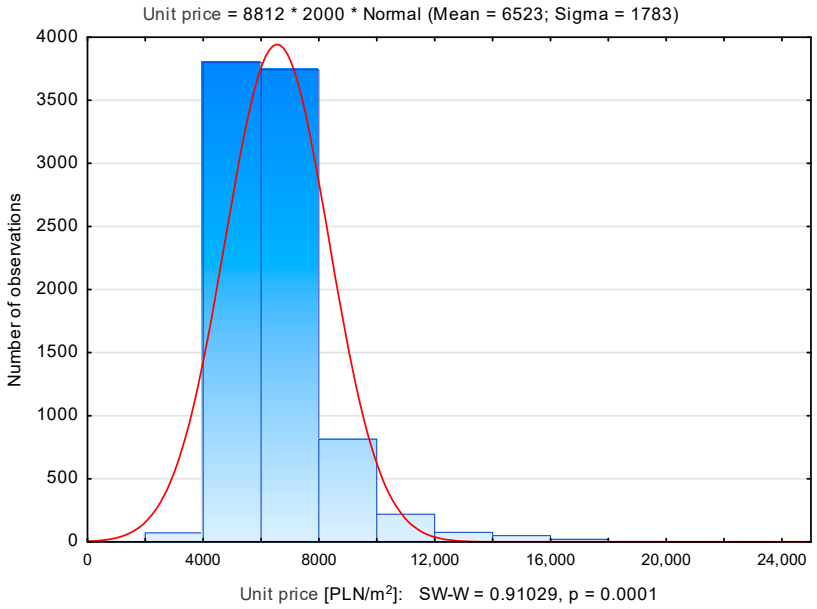

Figure 3. Histogram of unit price—Krakow, the whole city.

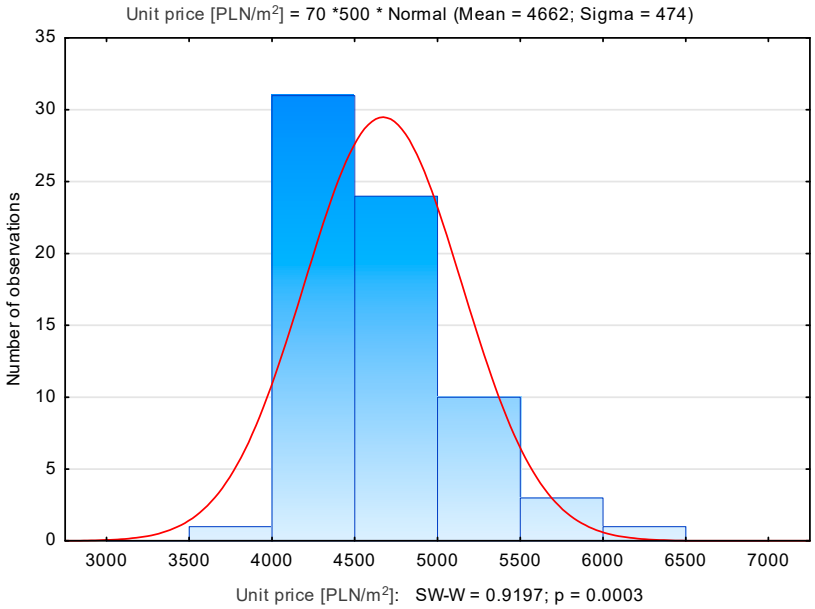

Figure 4. Histogram of unit price—district Bieńczyce.

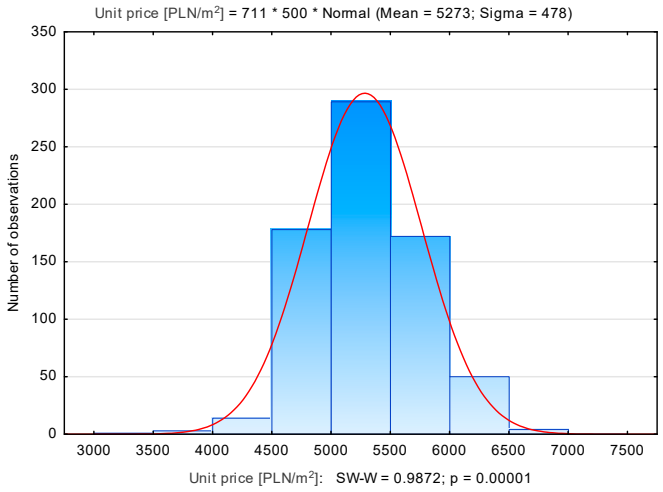

Figure 5. Histogram of unit price—district Bieżanów.

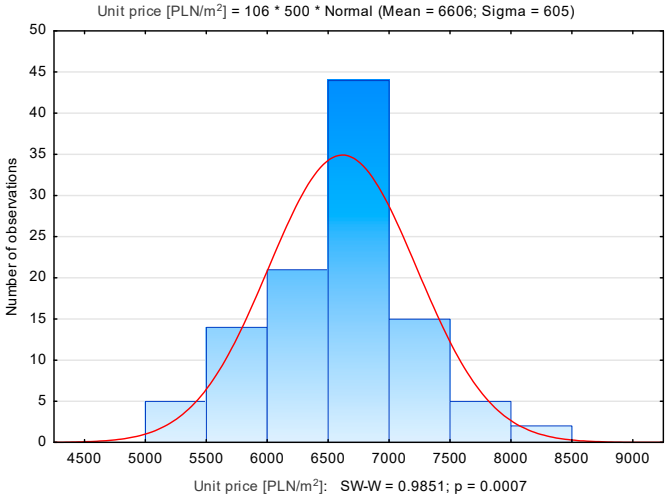

Figure 6. Histogram of unit price—district Bronowice.

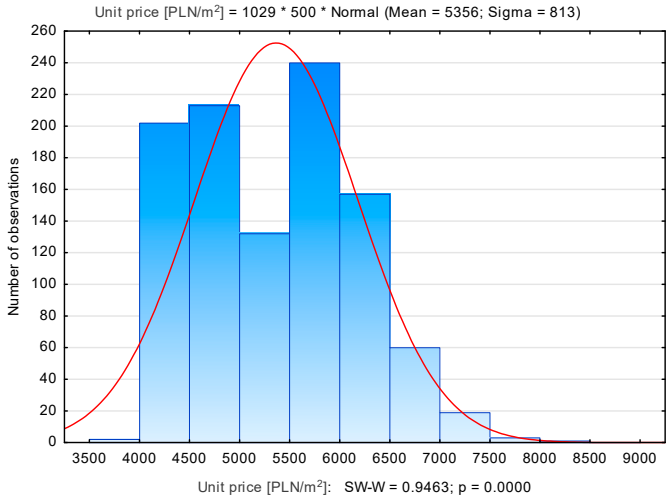

Figure 7. Histogram of unit price—district Czyżyny.

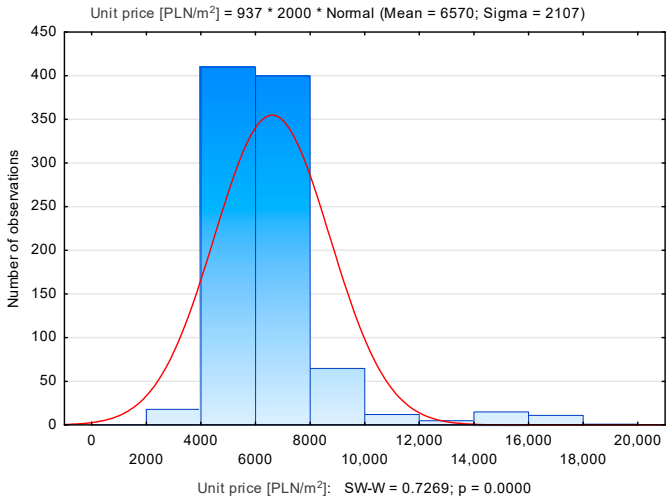

Figure 8. Histogram of unit price—district Dębniki.

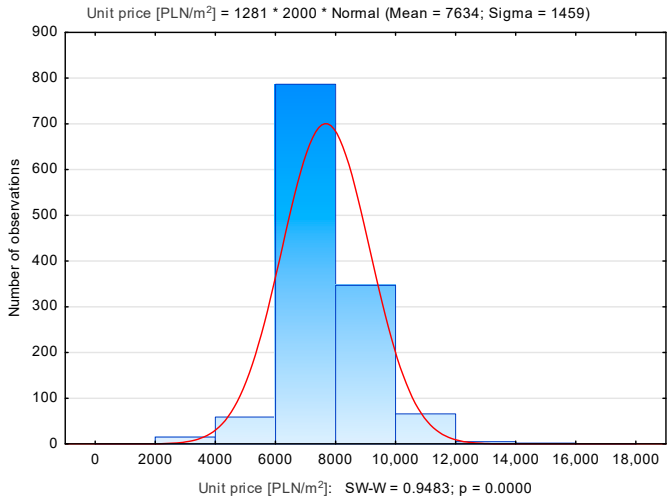

Figure 9. Histogram of unit price—district Grzegórzki.

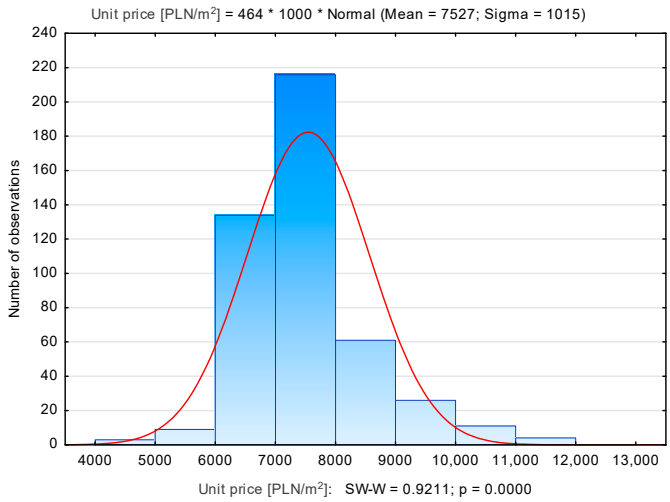

Figure 10. Histogram of unit price—district Krowodrza.

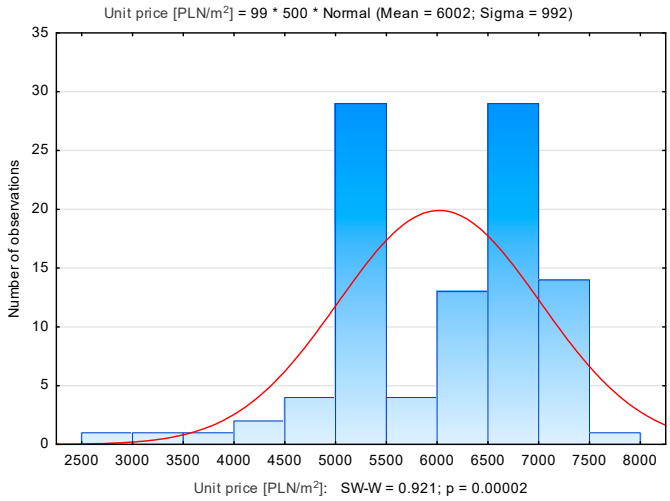

Figure 11. Histogram of unit price—district Łagiewniki.

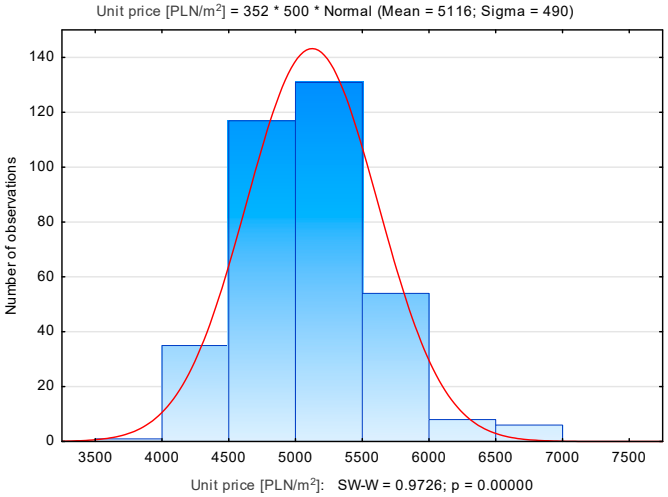

Figure 12. Histogram of unit price—district Mistrzejowice.

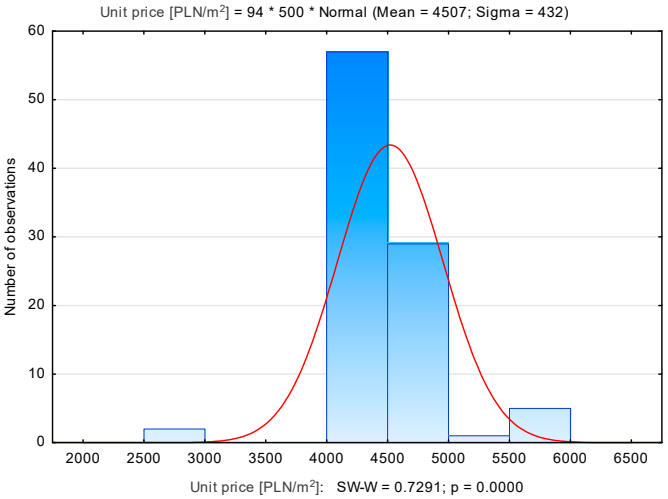

Figure 13. Histogram of unit price—district Nowa Huta.

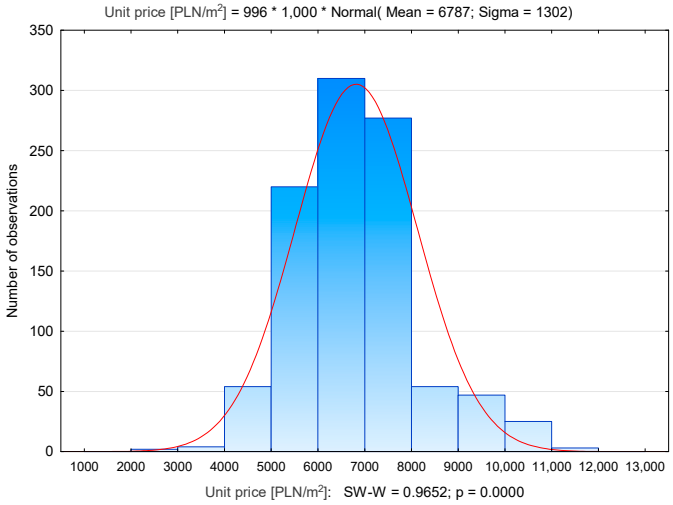

Figure 14. Histogram of unit price—district Podgórze.

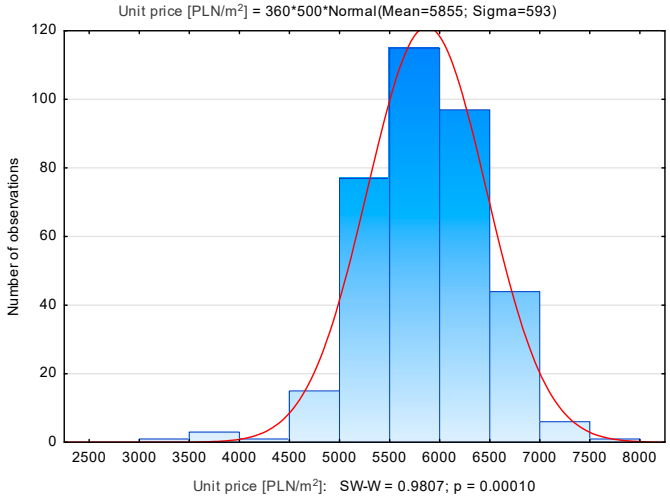

Figure 15. Histogram of unit price—district Podgórze Duchackie.

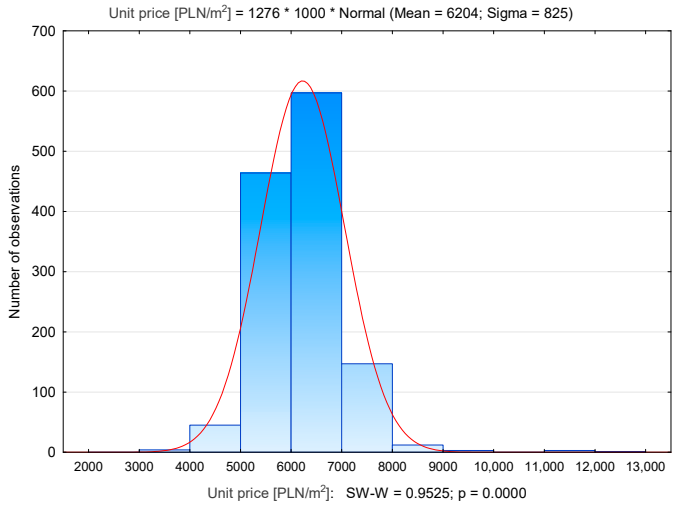

Figure 16. Histogram of unit price—district Prądnik Biały.

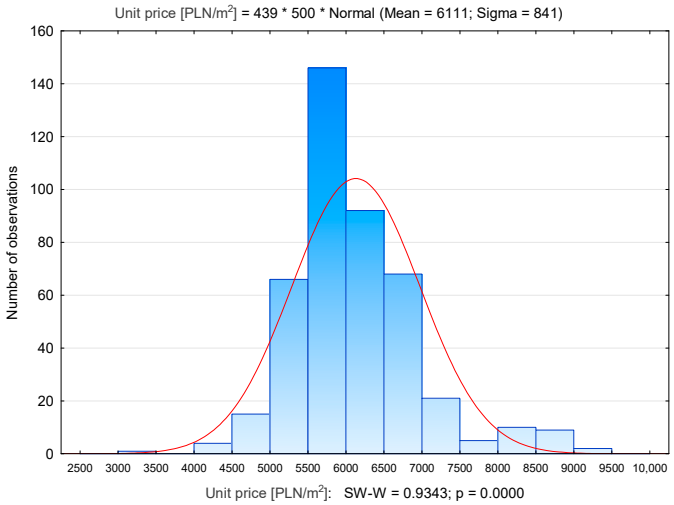

Figure 17. Histogram of unit price—district Prądnik Czerwony.

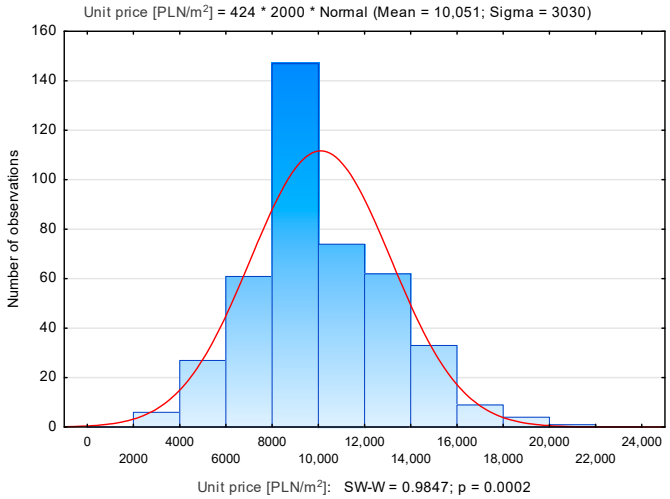

Figure 18. Histogram of unit price—district Stare Miasto.

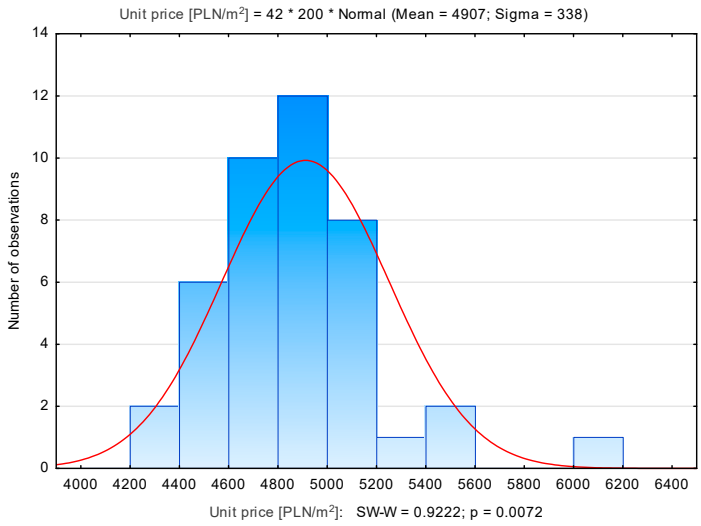

Figure 19. Histogram of unit price—district Swoszowice.

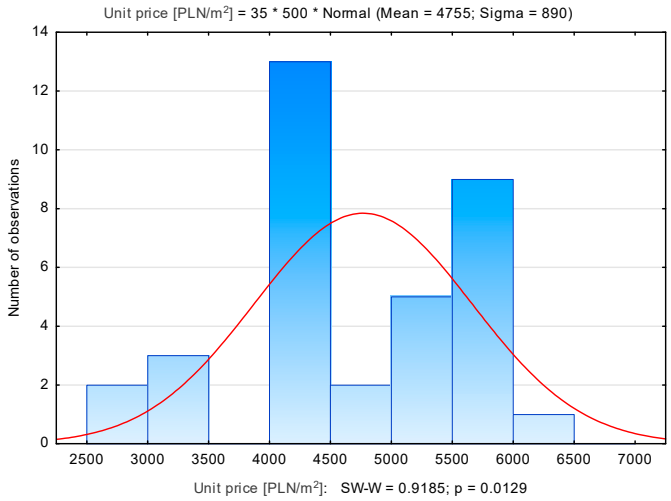

Figure 20. Histogram of unit price—district Wzgórza Krzesławickie.

The analysis of the results allows the conclusion that with the assumed significance level (P = 0.05) there are no grounds to reject the hypothesis of a normal distribution of unit prices only for the Zwierzyniec district. This proves the occurrence of outliers. The aim of further research is to eliminate observations outliers from the database.

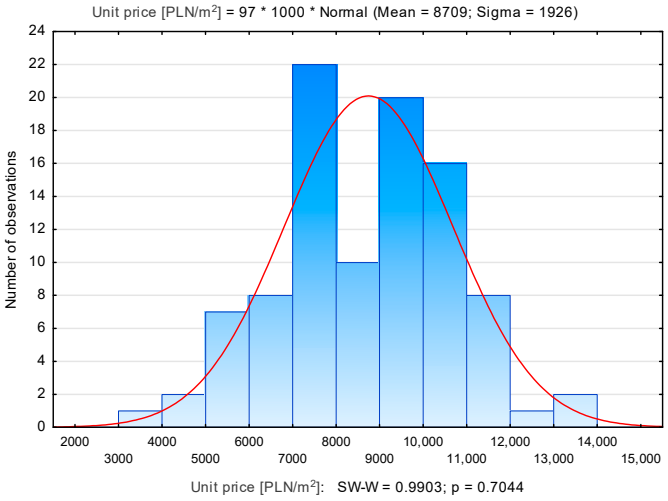

Figure 21. Histogram of unit price—district Zwierzyniec.

2.2. Rule of Thumb

It is necessary to examine the characteristics of the set of data to take any further action based on inference. Even if the model does not have a co-linearity problem or a data problem (for example, data shortages), it is prudent to see which observations have a big impact on regression results. Impact observation diagnostics provide information on the reliability of conclusions drawn from an estimated model. To pre-detect influential observations in a property database, an R-projection matrix can be used, as follows:

$$\mathbf{R} = \mathbf{X}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)^{-1}\mathbf{X}^{\mathrm{T}} \tag{4}$$

where $\mathbf{X}$ is a matrix of coefficients of multiple regression model equations. Elements from the diagonal $\mathbf{R}$ matrix determine the effect of the i-th observation on model parameter estimates. The $\mathbf{R}$ matrix is idempotent, therefore these elements will always be contained in the range [0, 1]. The well-known 'rule of thumb' principle states that if $\mathbf{R}_{i,i} > 0.5$ is observed, it should be considered influential. Based on analyses of many lands and residential

property bases [31,32], the authors conclude that to detect influential observations in property bases, the value should be $\mathbf{R}_{i,i}$ reduced to around $\mathbf{R}_{i,i} > 0.07$. This is because this parameter depends solely on the values of explanatory variables, the number of which is small relative to the number of observations obtained from the register. It should be emphasized that the matrix does not depend on the transaction prices of the property, therefore, it is possible to determine only which properties in terms of characteristics (attributes) have a significant impact on the regression model.

2.3. Mahalonobis Distance

Methods for identifying outliers based on Mahalonobis distances use the following criteria [14,33]:

$$MD_i^2 = (\mathbf{x}_i - \bar{\mathbf{x}})\mathbf{Cov}(\mathbf{x})^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})^{\mathrm{T}} \quad (5)$$

MD_i—Mahalonobis Distance,
$\mathbf{x}_i$ —a vector containing the ith explanatory variables,
$\bar{\mathbf{x}}$—vector of the average explanatory variables,
$\mathbf{Cov}(\mathbf{x})$—covariance matrix for explanatory variables.

The Mahalanobis distance can also be determined by means of levers (leverage) h_{and}:

$$MD_i^2 = (n-1)\left(h_i - \frac{1}{n}\right) \quad (6)$$

where

n—number of observations,
h_i —the value of leverage for the first observation.

This measures the distance of a given observation from the mean of the independent variables. In practice, it is difficult to determine the cut-off point for influential cases. In the literature, it is difficult to find a clear answer to the question of how this criterion should be determined. This approach also has the disadvantage that the value of the criteria itself (6) is very sensitive to the occurrence of outliers. To determine the cut-off point of influential observations, it should be noted that these criteria should depend on the number of model parameters and the number of observations. Based on the theory of square forms, it is concluded that Equation (6) has a distribution of o $(n - u)$ degrees of freedom, where u is the number of parameters of the model. Observations with high statistics, i.e., a square of the Mahalanobis distance compared to the critical values of the distribution, can be considered as influential observations. In this case, the data need to be checked and the appropriate limit value selected. In the case of a large number of degrees of freedom, the authors propose to set the cut-off criterion at:

$$k\_MD = \frac{n/u}{\sqrt{\chi^2(n-u,\alpha)}} \quad (7)$$

2.4. Analysis of Standardized Model Residuals

If each residual is divided by its standard deviation, i.e., as follows:

$$RS_i = \frac{v_i}{\sigma_i} \quad (8)$$

then it will generate a statistic that points to the impact of observations. It is customary to consider that if the absolute value of t is greater than 2, then it is influential. However, this is an approximate criterion. In practical applications, it is worth applying stricter or milder conditions. This is because the model is exuded by the least-squares method, the residuals have a normal distribution and the statistics (8) of the Student's t-distribution are o $(n - u)$ degrees of freedom, where n is the number of observations and the number of parameters of the model is estimated. By assuming any level of materiality α, the residuals

can be considered to be inputs to the model and thus the observation may be considered as an outlier by comparing the statistic (8) with the Student's *t*-distribution.

$$RS_i > k\_RS_i;\ k\_RS_i = t_S\left(1 - \frac{\alpha}{2}, n - u\right) \tag{9}$$

Figure 22 shows graphs of critical values for selected probability levels of value P, depending on the number of degrees of freedom.

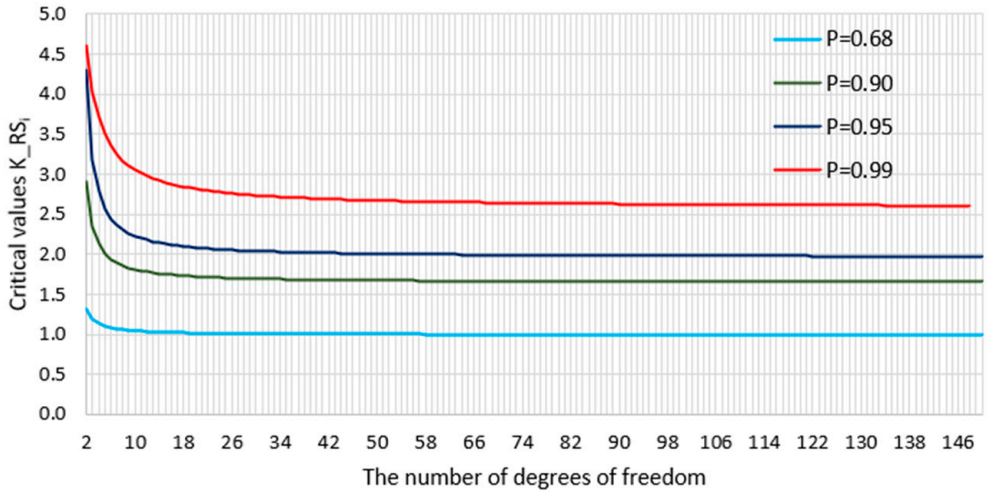

Figure 22. Critical values K_RS_{and}.

It follows from the above statements that at the set probability level P = 0.95 it can be approximately assumed that observations for which standardized residuals exceed twice the quantum value of the Student's *t*-distribution, but with at least 25 to 30 degrees of freedom, can be considered outliers from the model. For P = 0.99 and with degrees of freedom above 30, this value is on the order of 2.7–2.6. The exact value of the criterion can be calculated on a case-by-case basis according to the given algorithm. It should be noted that increasing the probability level results in fewer observations being detected as outliers, while a decrease in the probability level makes the criterion more stringent. Figure 22 indicates that for P = 0.68 outliers will be considered as observations whose residuals exceed the standard deviations directly determined by the least-squares method. Based on several experiments conducted by the authors, it is concluded that the elimination of observations from the model should be carried out individually, even if a larger group of outliers is detected in a given iteration. This is particularly important in the case of a more stringent criterion. This is because in a regression model even a single outlier can significantly change the form of the model and observations that did not match the primary model can meet the criteria of the second model after eliminating even one observation.

2.5. Cook's Distance

Cook's distance measures the change in regression coefficient values when a single observation is eliminated from the model. In the case of the Mahalanobis distance [33,34], the distance of the case from the centre of gravity determined by the independent variables is measured. Standardized tests determine the distance from the regression line. Cook's distances combine these two distances and are a cumulative measure of the effect of individual observations on the regression line [16]. To determine whether the vector of independent variables x and for the ith observation is unusual against the background of the other x, the lever hand can be determined by the following form:

$$h_i = \boldsymbol{\delta}_i^{\mathrm{T}} \mathbf{X} (\mathbf{X}^{\mathrm{T}} \mathbf{X})^{-1} \mathbf{X}^{\mathrm{T}} \boldsymbol{\delta}_i = \boldsymbol{\delta}_i^{\mathrm{T}} \mathbf{P}_X \boldsymbol{\delta}_i = \mathbf{x}_i (\mathbf{X}^{\mathrm{T}} \mathbf{X})^{-1} \mathbf{x}_i^{\mathrm{T}} \tag{10}$$

where

$$\mathbf{P}_X = \mathbf{X} (\mathbf{X}^{\mathrm{T}} \mathbf{X})^{-1} \mathbf{X}^{\mathrm{T}}, . \boldsymbol{\delta}_i = [0, \ldots, 0, 1, 0, \ldots 0]$$

There is a dependency for each model. Informal rules say that 'if, then' observations can be considered unusual. The fact that the observation is unusual does not yet indicate that it does not fit the model. However, if unusual observations, i.e., large levers, have high residual values at the same time, then this may indicate that they are outliers from the model. Since the variances have chi-squared distributions of $(n - u)$ and $(n - u - 1)$ degrees of freedom, respectively, the statistics are as follows:

$$0 \leq h_i \leq 1 \quad h_i \geq \frac{2u}{n} \hat{\sigma}^2, \hat{\sigma}_i^2$$

$$CD_i = \frac{1}{u} \frac{\hat{\sigma}_i^2}{\hat{\sigma}^2} \frac{h_i}{(1 - h_i)} \quad \sim \quad F_{(1, n-u)} \tag{11}$$

where

CD_i—Cook's distance,
$\hat{\sigma}^2$—variance estimator calculated based on all observations,
$\hat{\sigma}_i^2$—an estimator of variance calculated after elimination of *the first* observation,
n—number of observations,
u—number of model parameters.

In practice, Cook's distance with modified residues is determined in the form of:

$$CD_i = \frac{\sum_{j=1}^{n} \left(\hat{y}_{j(i)} - \hat{y}_j \right)^2}{u \hat{\sigma}^2} \tag{12}$$

where

$\hat{y}_j$—the value predicted by the model for the jth observation determined in the full model,
$\hat{y}_{j(i)}$—the value predicted by the model for the jth observation determined based on the model from which the ith observation was removed.

It is generally assumed that if CD $\geq 4/n$ then you should look at such observation because it can be an outlier observation. This approach is correct if the significance level is set at 0.05 and the number of observations is large. Based on (11), by employing the Fisher–Snedecor distribution, the critical value can be specified precisely. The authors propose that the criterion be determined based on dependencies:

$$k\_CD_i = \frac{F(\alpha, 1, n - u)}{n} \tag{13}$$

Figures 23 and 24 show Cook's distance values for three probability levels, depending on the number of observations and the number of degrees of freedom.

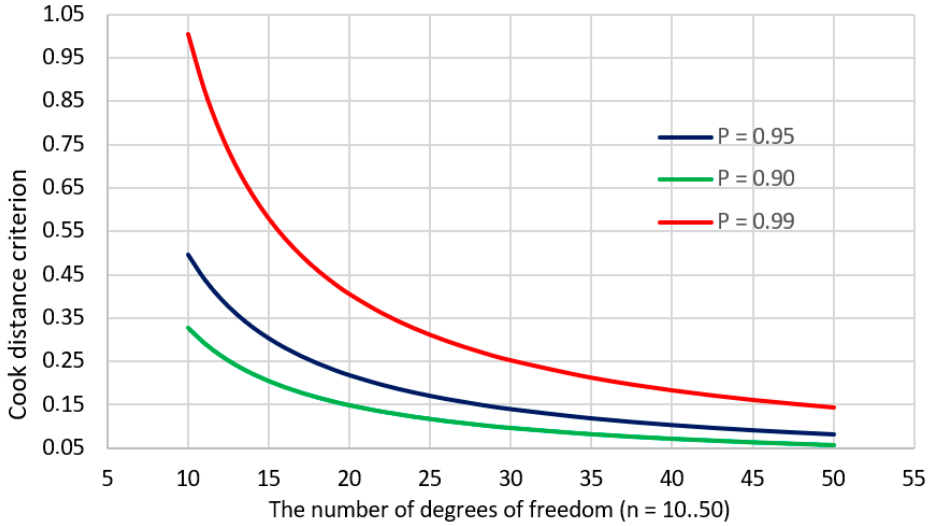

Figure 23. Cook distance criterion for 5 model parameters depending on the number of observations (n = 10 to 50) and the probability level.

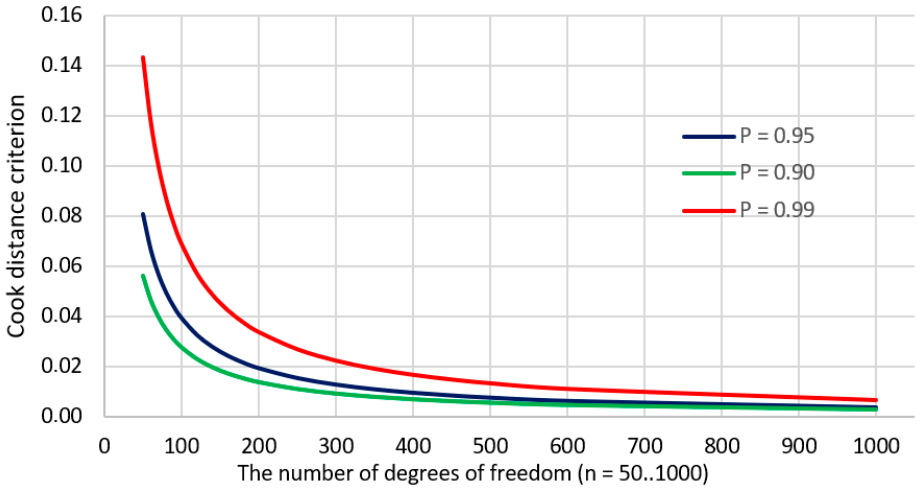

Figure 24. Cook distance criterion for 5 model parameters depending on the number of observations (n = 50 to 1000) and a fixed level of probability.

2.6. Classification and Regression Tree Models

Classification and Regression Tree (C&RT) models provide a significant solution to regression and classification problems. The method is described in detail elsewhere [35–38]. In the C&RT method, there are several basic steps:

- Tree building: the process occurs through the recursive division of nodes,
- Stopping the construction of the tree: at this stage, the tree is as extensive as possible, usually containing redundant information,
- Pruning of the tree consists of removing redundant branches,
- Choosing the right tree: some branches are restored to increase the effectiveness of the method.

3. Results

3.1. Identification of Influential and Outliers in the Kraków Database

The database of 8812 premises traded on the primary market in Kraków was analysed. Initially, based on correlation analysis, five characteristics were found that shaped property prices and at the same time represented variables explaining the multiple regression model.

The regression model was determined by the least-squares method. Based on the original base, the following model parameters were obtained, drawn up for the entire city and separately for each district (presented in Table 2):

Table 2. Regression model parameters—base.

Object	R^2	σ	Distance	Usable Area	Storey	Rooms	Transaction Date
Bieńczyce	0.04	471	-	−0.090	−0.550	−0.120	−0.330
Bieżanów	0.11	451	0.234	−0.030	0.279	−0.160	−0.180
Bronowice	0.28	522	−0.450	0.150	0.150	−0.310	0.119
Czyżyny	0.29	612	−0.380	0.457	0.061	−0.680	−0.160
Dębniki	0.45	1577	−0.550	0.387	0.071	−0.150	0.061
Grzegórzki	0.32	1203	−0.510	0.087	0.149	−0.190	0.055
Krowodrza	0.18	925	−0.050	0.406	−0.020	−0.690	−0.030
Łagiewniki	0.24	885	−0.020	0.140	−0.170	−0.490	−0.160
Mistrzejowice	0.28	417	−0.480	0.034	0.172	−0.270	−0.310
Nowa Huta	0.12	414	−0.014	0.013	0.043	−0.280	0.212
Podgórze	0.36	1009	−0.460	−0.170	0.155	−0.220	0.002
P. Duchackie	0.19	514	−0.270	−0.420	0.146	0.085	−0.110
Prądnik Biały	0.05	804	−0.010	−0.050	0.094	−0.140	0.118
Prądnik Cz.	0.24	736	−0.400	0.345	0.147	−0.330	0.032
Stare Miasto	0.27	2552	−0.420	0.274	0.161	−0.330	0.214
Swoszowice	0.03	350	−0.010	0.295	−0.080	−0.340	0.120
Wzgórza K.	0.48	685	−0.150	−0.080	0.029	−0.630	−0.080
Zwierzyniec	0.11	1863	−0.140	−0.420	0.086	0.321	−0.300

The statistically significant parameters are marked in red. The results of the analysis show that it is not possible to use a multiferroic regression model based on a raw database. The coefficients of determination R^2 are unsatisfactory for each of the analysed districts of Kraków. Outliers should therefore be eliminated. To detect outliers, for the entire city and each district separately, the following indicators were set: projection matrix, Mahalanobis distance, standardized residuals of Ri and Cook's distance (Table 3). Critical values are calculated based on Formulas (7), (9) and (13). The probability level of P = 0.95 is assumed. Table 4 highlights selected observations that have been identified as influential or outliers for the example district of Dębniki. This district has the highest number of outliers (Figure 2). The number of observations in the Dębniki district is 935. Based on Formula (7), the criterion for outliers determined by the Mahalanobis distance is k_MD_{and} = 6.45. As proposed by (13), Cook's distance criterion was set at

$$k_{CD_i} = \frac{F(0.05,\ 1930)}{935} = \frac{3006}{935} = 0.00321$$

The number of detected influential observations in the analysed district varies, depending on the method used. For Cook's distances it is 6.5%, the standardized residual of the model is 4.9% and Mahalanobis and 'rule of thumb' distances are about 6.2%. None of the 18 districts exceeded 8% of the total number of properties. The elimination of outliers was carried out based on Cook's distance. The rule of thumb method and Mahalanobis distance should only be considered as supporting the decision to treat observations as an outlier. The standardized residual method overlaps approximately 90% with the Cook's distance method. Figure 25 shows the plot box after eliminating outliers.

Table 3. Influential properties identified by the Mahalanobis distance, projection matrix, standardized model rest and Cook distance.

Case	Cook's Distance	Standard Residual	Mahalanobis Distance	R_{ii}
873	0.134967	−3.65	40.45	0.14
567	0.098323	4.04	24.66	0.11
565	0.079709	3.99	20.52	0.10
845	0.064456	1.47	104.04	0.09
563	0.057482	3.37	20.66	0.14
372	0.056316	3.84	15.6	0.12
562	0.055229	3.33	20.41	0.10
566	0.044671	5.22	6.28	0.11
371	0.043808	3.94	11.39	0.08
165	0.043595	3.83	12.04	0.14
352	0.041662	2.48	27.65	0.12
373	0.040856	5.08	6.02	0.08
...	...	...	...	....
834	0.007834	−2.19	6.27	0.08
809	0.005884	−2.14	4.71	0.10
159	0.004302	1.67	5.86	0.08

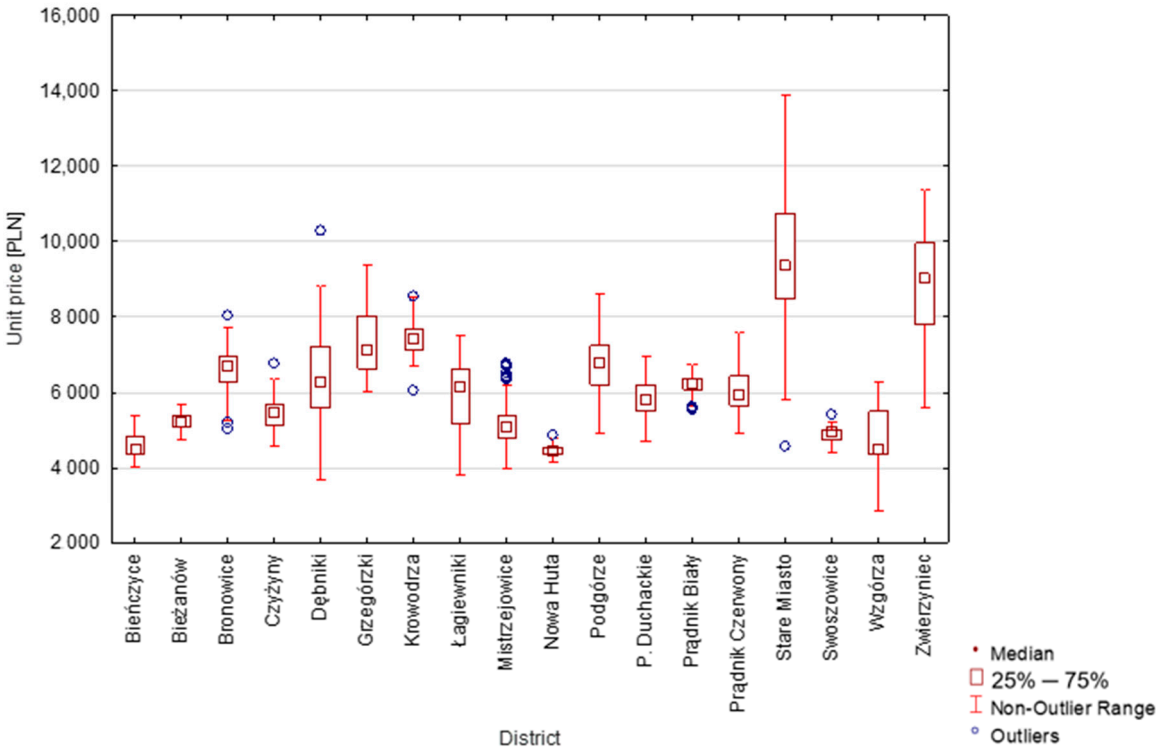

Figure 25. Box plot of unit price grouped by District—8290 properties.

3.2. Multidimensional Regression Models for Kraków Databases

In the case of six districts where the value of R^2 is statistically insignificant: Bieńczyce, Podgórze D., Prądnik Biały, Prądnik Czerwony, Swoszowice, Zwierzyniec (Table 4), the regression model is not suitable for predicting the market value of the property. In these cases, subsequent iterations should eliminate subsequent outliers resulting from the change in the regression model, or another predictive solution should be used.

Table 4. Regression model parameters—database after removal of outlier observations.

Object	R^2	σ	Distance	Usable Area	Storey	Rooms	Transaction Date
Bieńczyce	0.47	317	-	−0.100	−0.550	0.914	−0.330
Bieżanów	0.82	350	−0.583	−0.110	0153	−0.280	0.210
Bronowice	0.76	493	−0.440	−0.080	0.026	−0.390	0.135
Czyżyny	0.79	162	0.026	0.911	−0.010	−1.300	0.003
Dębniki	0.92	286	−0.960	0.497	0.084	−0.320	0.017
Grzegórzki	0.92	226	−0.880	0.063	0.028	−0.250	0.030
Krowodrza	0.78	220	0.056	0.818	−0.070	−0.140	−0.040
Łagiewniki	0.72	445	0.127	0.203	−0.260	−0.800	−0.050
Mistrzejowice	0.78	238	−0.480	−0.050	0.191	−0.230	0.270
Nowa Huta	0.74	353	−0.390	0.006	−0.024	0.310	−0.040
Podgórze	0.84	297	−0.740	−0.300	0.206	−0.350	0.011
Podgórze D.	0.47	337	−0.510	−0.580	0.169	0.244	0.120
Prądnik Biały	0.56	144	−0.130	−0.440	0.298	−0.230	0.010
Prądnik Cz.	0.49	393	−0.540	0.336	0.330	−0.380	0.065
Stare Miasto	0.79	781	−0.780	0.175	0.318	−0.250	0.020
Swoszowice	0.18	228	−0.270	−0.550	0.160	0.622	0.184
Wzgórza K.	0.79	492	−0.580	−0.250	0.057	−0.590	0.211
Zwierzyniec	0.43	969	−0.470	−0.410	0.081	0.397	0.123

3.3. C&RT Trees

When the C&RT tree schema is created, the following parameters are assumed [38,39]:

- variable dependent—unit price,
- quality predictors—district,
- quantitative predictors—distance, area, floor, transaction date,
- minimum number in the end node: 20.

For the above parameters, considering the original database (8812 properties), more than 250 trees can be created. The characteristics of the selected sample tree are shown in Figure 26.

The average unit price at the first node is PLN 6625/m^2 ± 1795 PLN/m^2. Its number is equal to the number of the base, i.e., 8812 properties. The division of the first node was made based on the distance attribute, dividing the entire base into two subsets, above and below a distance of 3.5 km from the city centre. In the case of base analysis, after eliminating outliers by Cook's distance, the corresponding tree is presented Figure 27.

In this case, it is worth noting a significant decrease in the value of variance. However, one of the most important factors determining the value of the property is still the distance from the city centre.

3.4. Chi-Square Automatic Interaction Detector (CHAID) Trees

Figures 28 and 29 show the CHAID decision tree model, which confirms the conclusions. The designated decision tree is a statistical classification procedure in this case. The nodes correspond to the statistical tests carried out on the values of property attributes, the branches are the potential results of the tests carried out and the leaves of these trees present the decision-making, that is, the dependent variable—in this case, the market value of the residential property in Kraków. Decision trees are straightforward to interpret and allow, among other things, estimation of the value of the property.

Division on nodes is most often done by a variable district, then by distance. It should be noted that these two predictors are interdependent. The relationship, based on the correlation of Spearman's and Kendall's ranks, is 0.70 and is statistically significant. The division only occurs twice because of the area of the apartment. The minimum unit prices of the property can be found in the end nodes. The maximum unit price is characterized by properties from the Old Town.

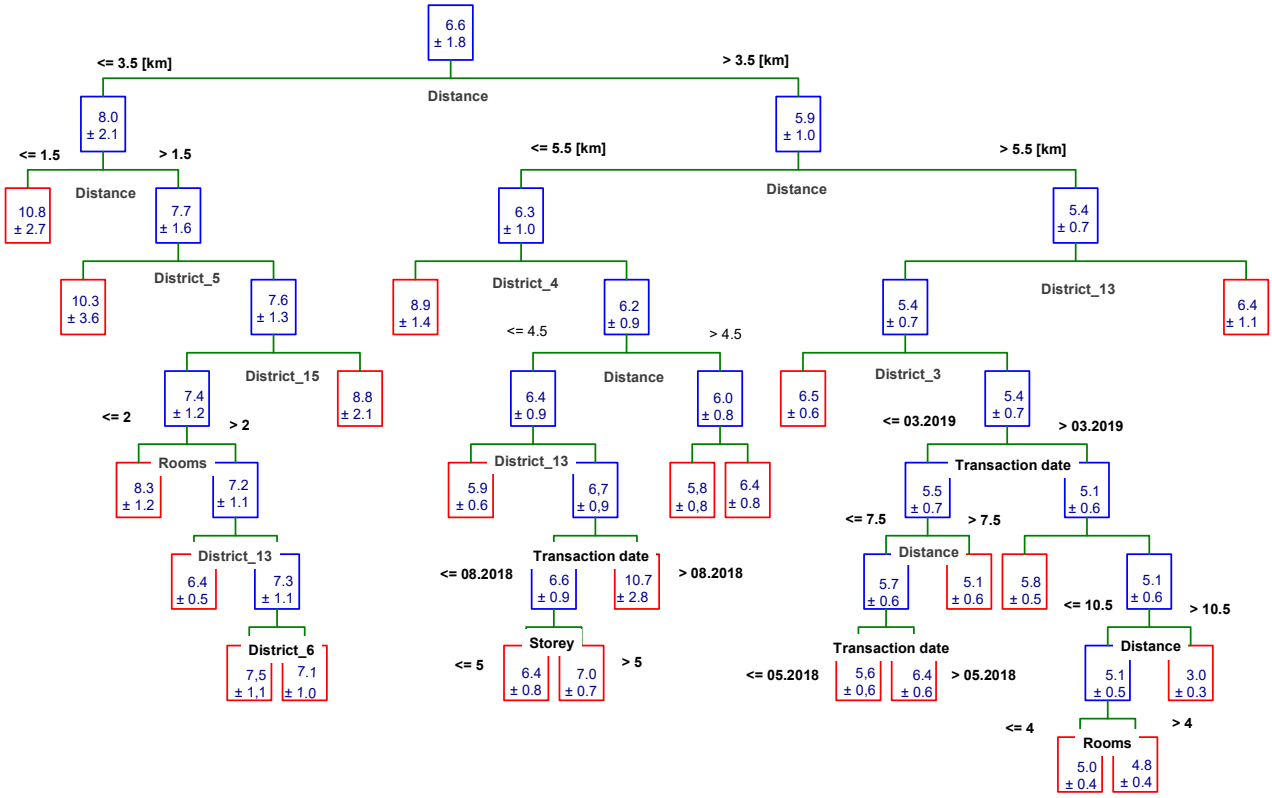

Figure 26. C&RT tree model, 20 end nodes. Database before elimination of outliers: 8812 properties.

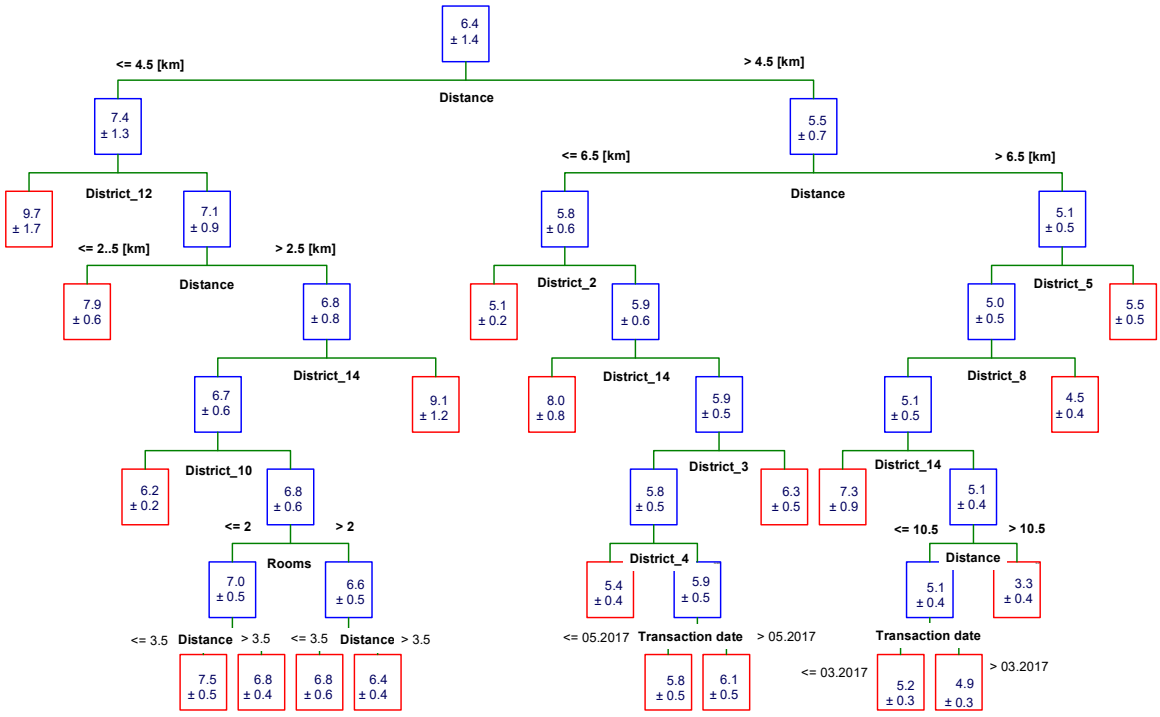

Figure 27. C&RT tree model, 17 end nodes. Database after elimination of outliers: 8290 properties.

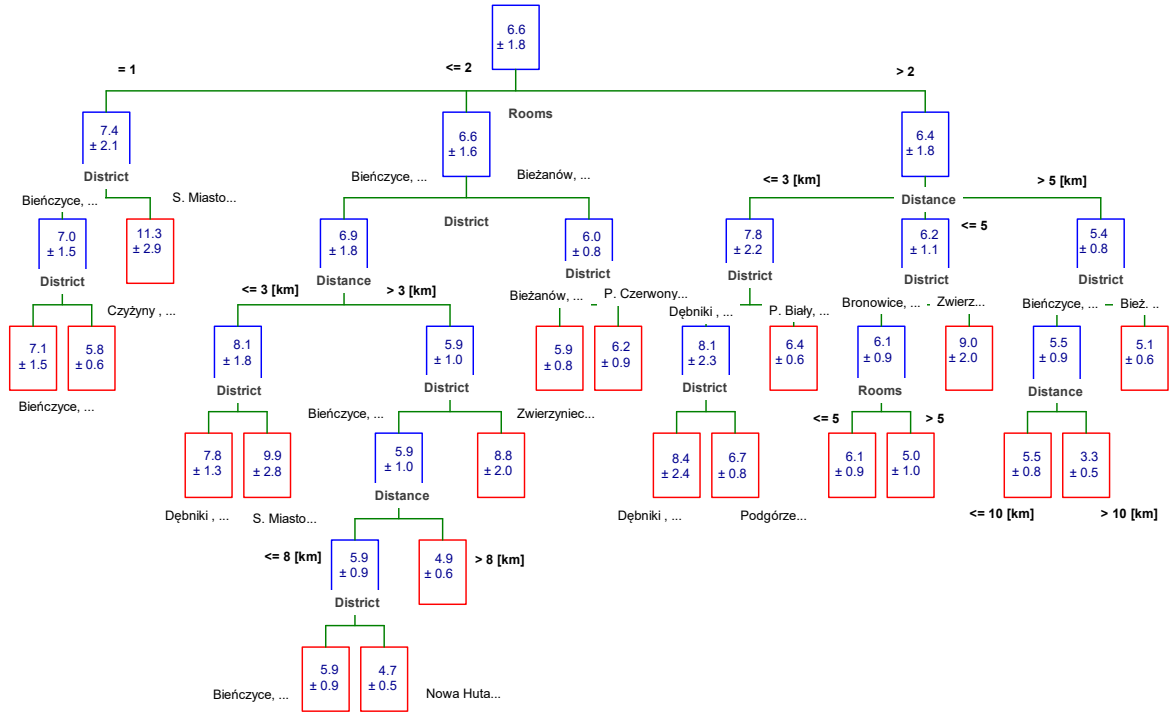

Figure 28. CHAID tree model, 20 end nodes. Database before elimination of outliers: 8812 properties.

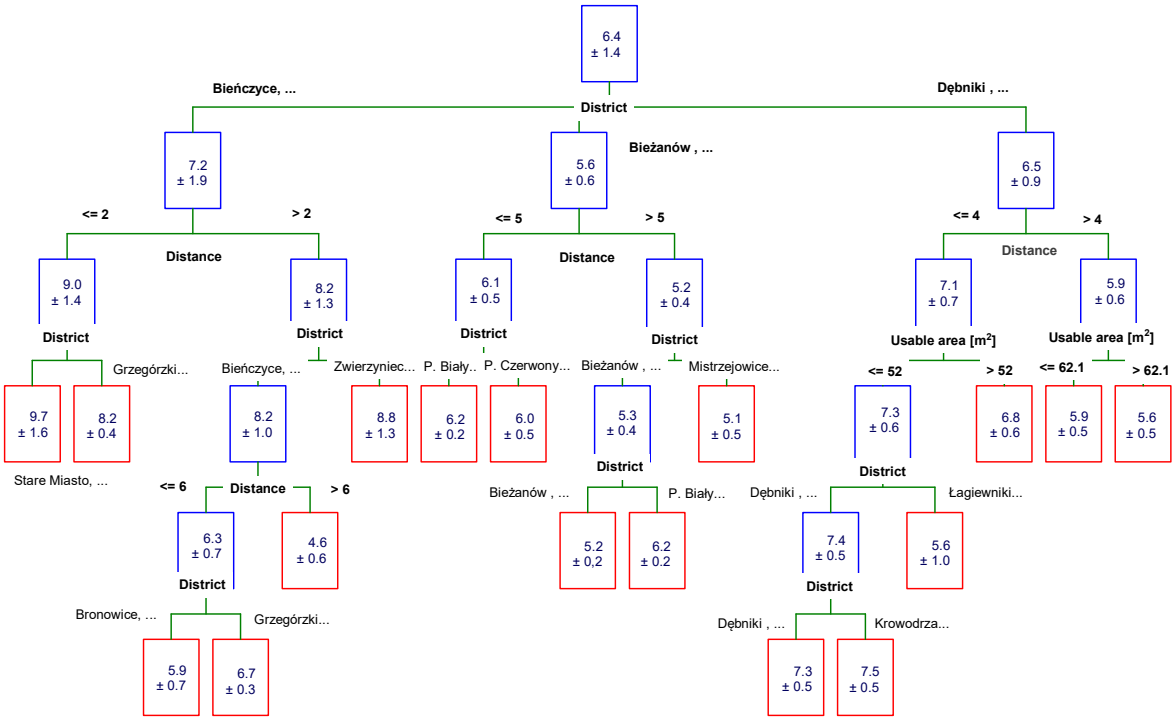

Figure 29. CHAID tree model, 17 end nodes. Database after elimination of outliers: 8290 properties.

4. Conclusions

This study analysed a database of 8812 dwellings that were traded on the primary market in Kraków. The basic characteristics that shaped property prices were established, while at the same time providing variables explaining the regression model. Beta (bi) weights were designated for these variables. Regression models for individual districts

of the city were determined by the least-squares method. The results show that it is not possible to use a multi regression model based on a raw database. The coefficients of determination R^2 are unsatisfactory for each of the analysed districts of Kraków.

To detect outliers, the following indicators were set: projection matrix, Mahalanobis distances, standardized chi test, and Cook's distances. Critical values were calculated based on the proposed Formulas (7), (9) and (13). The probability level of P = 0.95 was assumed. Mahalanobis distances only consider explanatory variables, so for the present issue, in which prices are the most common cause of outliers, they only provide information on influential observations. Similar regression models were obtained by eliminating outliers by standardized residuals and the Cook's distance method. In the case of 12 districts of Kraków, the regression model can be considered satisfactory, while in six cases it cannot be used to predict the market value of the property because of the very low coefficient of determination. For these six districts, it is advisable to supplement the database with new observations and then re-eliminate outliers. Analysis of the results compiled in Table 2 shows that in most districts a negative correlation with the price per m^2 has the attributes of distance from the centre and the number of rooms. The usable area affects property prices differently in different districts. On the other hand, a higher floor generally shows a positive correlation with the unit price.

The second part of the research was aimed at determining the suitability of C&RT trees to determine the effect of property attributes on their prices. Trees created using C&RT and CHAID have shown that the district attribute has a key influence on the unit price. The study was conducted for the entire database (8812 properties) and the database created after the outliers were eliminated by the Cook's distance method. Regression and classification studies confirmed the results of analyses carried out by multiple regression. The market for residential real estate in Kraków is not uniform. The individual districts create separate price zones. The apartments with the highest unit price are located in the Old Town and Zwierzyniec districts, located at a distance of up to 1.5 km from the city centre and located on higher floors.

From all tables we present in the publication two of them show how useful the presented solution is: Tables 2 and 4. Automating deleting outstanding data, based on clearly defined principles, significantly improves the accuracy parameters of the model describing the local real estate market. This relationship is especially beneficial when working on large data sets (several thousand). The lack of a precise definition of the criteria allowing for the recognition of real estate as an outlier is a significant obstacle here. This is especially true of Cook's distance in real estate analysis. Based on the Fisher–Snedecor distribution, the authors precisely defined the Cook distance criterion for the analyzed data set. Further studies will include the separation of sub-zones in individual districts. The number of attributes will be expanded with features such as street, noise, distance from green areas, window exposure, bathroom area, balcony area and window view. Preliminary analyses carried out for individual districts of Kraków showed that these are important factors influencing the market value of the residential real estate.

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

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are available on request from the corresponding author.

Acknowledgments: This paper was created as part of statutory research 16.16.150.545. The authors express sincere gratitude to the Journal Editor and the anonymous reviewers who spent their valued time to provide constructive comments and assistance to improve the quality of this paper.

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

References

1. Marona, B.; Tomal, M. The COVID-19 pandemic impact upon housing brokers' workflow and their clients' attitude: Real estate market in Krakow. *Entrep. Bus. Econ. Rev.* **2020**, *8*, 221–232.
2. Kowalczyk-Anioł, J.; Grochowicz, M.; Pawlusiński, R. How a Tourism City Responds to COVID-19: A CEE Perspective (Kraków Case Study). *Sustainability* **2021**, *13*, 7914. [CrossRef]
3. Kilpatrick, J.A. The future of real estate information. *Real Estate Issues* **2001**, *26*, 7–14.
4. Romańczyk, K. *Krakow—The City Profile Revisited*; Elsevier: New York, NY, USA, 2018; Volume 73.
5. Available online: https://pl.wikipedia.org/wiki/Podzia%C5%82_administracyjny_Krakowa (accessed on 23 July 2021).
6. Zyga, J. Evaluation of usefulness of real estate data contained in the register of prices and values of real estates. *Infrastrukt. Ekol. Teren. Wiej.* **2017**. [CrossRef]
7. Halik, Ł. Analysis of County Geoportals in Terms of Opportunities to Purchase Data of the Register of Real Estate Prices and Values Online. *Real Estate Manag. Valuat.* **2019**, *27*, 69–78. [CrossRef]
8. Halik, Ł. Information and Communication Systems Used for Keeping the Register of Real Estate Prices and Values (Rrepv) in Poland. *Real Estate Manag. Valuat.* **2018**, *26*, 45–53. [CrossRef]
9. Kannan, K.S.; Manoj, K. Outlier Detection in Multivariate Data. *Appl. Math. Sci.* **2015**, *9*, 2317–2324. [CrossRef]
10. Preweda, E. Outlier detection in surveying networks. In Proceedings of the 14th International Multidisciplinary Scientific Geoconference (SGEM), Albena, Bulgaria, 17–26 June 2014; Volume 2, pp. 365–372.
11. Crecente, R.; Alvarez, C.; Fra, U. Economic, social and environmental impact of land consolidation in Galicia. *Land Use Policy* **2002**, *19*, 135–147. [CrossRef]
12. Wójcik-Leń, J.; Leń, P.; Mika, M.; Kryszk, H.; Kotlarz, P. Studies regarding correct selection of statistical methods for the needs of increasing the efficiency of identification of land for consolidation—A case study in Poland. *Land Use Policy* **2019**, *87*, 104064. [CrossRef]
13. Greene, W.H. *Econometric Analysis*, 5th ed.; Prentice Hall: Upper Saddle River, NJ, USA, 2003.
14. Mahalanobis, P.C. On the Generalised Distance in Statistics. *Proc. Natl. Inst. Sci. India* **1936**, *2*, 49–55.
15. Ghorbani, H. Mahalanobis Distance and Its Application For Detecting Multivariate Outliers. *Facta Univ. Ser. Math. Inform.* **2019**, *34*, 583–595. [CrossRef]
16. Cook, R.D. Detection of Influential Observations in Linear Regression. *Technometrics* **1977**, *19*, 15–18.
17. Zhu, H.; Ibrahim, J.G.; Cho, H. Perturbation and scaled Cook's distance. *Ann. Stat.* **2012**, *40*, 785–811. [CrossRef]
18. Vukovic, O. Analysing bank real estate portfolio management by using impulse response function, Mahalanobis distance and financial turbulence. *Procedia Econ. Financ.* **2015**, *30*, 932–938. [CrossRef]
19. Stöckl, S.; Hanke, M. Financial applications of the Mahalanobis distance. *Appl. Econ. Financ.* **2014**, *1*, 78–84. [CrossRef]
20. Jung, E.; Yoon, H. Is Flood Risk Capitalized into Real Estate Market Value? A Mahalanobis-Metric Matching Approach to the Housing Market in Gyeonggi, South Korea. *Sustainability* **2018**, *10*, 4008. [CrossRef]
21. Isakson, H.R. Valuation analysis of commercial real estate using the nearest neighbors appraisal technique. *Growth Chang.* **1988**, *19*, 11–24. [CrossRef]
22. Chen, W.; Xie, X.; Wang, J.; Pradhan, B.; Hong, H.; Bui, D.T.; Duan, Z.; Ma, J. A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility. *Catena* **2017**, *151*, 147–160. [CrossRef]
23. Herath, S.; Maier, G. The Hedonic Price Method in Real Estate and Housing Market Research: A Review of the Literature; SRE-Discussion Papers, 2010/03; WU Vienna University of Economics and Business, Vienna, Austria, 31 March 2010. Available online: https://ideas.repec.org/p/wiw/wus009/588.html (accessed on 1 August 2021).
24. Janssen, C.; Söderberg, B.; Zhou, J. Robust estimation of hedonic models of price and income for investment property. *J. Prop. Invest. Financ.* **2001**, *19*, 342–360. [CrossRef]
25. Bourassa, S.C.; Cantoni, E.; Hoesli, M. Robust hedonic price indexes. *Int. J. Hous. Mark. Anal.* **2016**, *9*, 47–65. [CrossRef]
26. Mok, H.M.K.; Chan, P.P.K.; Cho, Y.S. A hedonic price model for private properties in Hong Kong. *J. Real Estate Financ. Econ.* **1995**, *10*, 37–48. [CrossRef]
27. Chau, K.W.; Chin, T.L. A Critical Review of Literature on the Hedonic Price Model (12 June 2002). *Int. J. Hous. Sci. Appl.* **2003**, *27*, 145–165.
28. Scott, D.W. Partial Mixture Estimation and Outlier Detection in Data and Regression. In *Theory and Applications of Recent Robust Methods*; Birkhauser: Basel, Switzerland, 2014; pp. 297–306. [CrossRef]
29. Rao, C.R.; Toutenburg, H. Linear models. In *Linear Models*; Springer: New York, NY, USA, 1995; pp. 3–18.
30. Casson, R.J.; Farmer, L.D. Understanding and checking the assumptions of linear regression: A primer for medical researchers. *Clin. Exp. Ophthalmol.* **2014**, *42*, 590–596. [CrossRef] [PubMed]

31. Frukacz, M.; Popieluch, M.; Preweda, E. *Real Estate Price Adjustment Due to Time in the Case of Large Databases*; Infrastructure and Ecology of Rural Areas; Committee on Technical Rural Infrastructure: Krakow, Poland, 2011; pp. 213–226. ISBN 1732-5587.
32. Jasińska, E. Real estate due diligence on the example of the polish market. In Proceedings of the 14th International Multidisciplinary Scientific Geoconference (SGEM), Albena, Bulgaria, 17–26 June 2014; Volume 2, pp. 419–426.
33. ScienceDirect Homepage. Available online: https://www.sciencedirect.com/topics/engineering/mahalanobis-distance (accessed on 20 November 2020).
34. Algur, S.P.; Biradar, J.G. Cooks Distance and Mahanabolis Distance Outlier Detection Methods to identify Review Spam. *Int. J. Eng. Comput. Sci.* **2017**, *6*, 21638–21649. [CrossRef]
35. Jasińska, E. *Chosen Statistical Method in Real Estate Market Analysis*; Wydawnictwa Akademii Górniczo-Hutniczej im; Stanisława Staszica w Krakowie: Krakow, Poland, 2012, ISBN 978-83-7464-471-6.
36. Ho, W.K.O.; Tang, B.-S.; Wong, S.W. Predicting property prices with machine learning algorithms. *J. Prop. Res.* **2020**, *38*, 48–78. [CrossRef]
37. Levantesi, S.; Piscopo, G. The Importance of Economic Variables on London Real Estate Market: A Random Forest Approach. *Risks* **2020**, *8*, 112. [CrossRef]
38. Jasińska, E.; Preweda, E. The use of regression trees to the analysis of real estate market of housing. In Proceedings of the 13th International Multidisciplinary Scientific Geoconference (SGEM), Albena, Bulgaria, 16–22 June 2013; Volume 2, pp. 503–508.
39. Buntine, W. Learning classification trees. *Stat. Comput.* **1992**, *2*, 63–73. [CrossRef]