World of Real Estate Journal
SWIAT NIERUCHOMOSCI

The Applicability of the Gini Coefficient for Analyses of Real Estate Prices

Justyna Brzezicka[1], Marta Gross[2], Katarzyna Kobylińska[3]

[1] *University of Warmia and Mazury in Olsztyn, ORCID: https://orcid.org/0000-0001-9500-1054, justyna.brzezicka@uwm.edu.pl*
[2] *University of Warmia and Mazury in Olsztyn, ORCID: https://orcid.org/0000-0002-8599-351X, marta.gross@uwm.edu.pl*
[3] *University of Warmia and Mazury in Olsztyn, ORCID: https://orcid.org/0000-0001-6183-3215, katarzyna.kobylinska@uwm.edu.pl*

ABSTRACT

Purpose - This article discusses the selection of outlying transactions for the purpose of developing a database for analyses of real estate prices. The study aimed to determine the applicability of the Gini coefficient as an auxiliary tool for evaluating the distribution of real estate prices.

Design/methodology/approach– Changes in the values of the Gini coefficient were analyzed in databases that were gradually reduced by extreme values. The primary database was composed of real estate prices in three districts of Warsaw in 2017-2018.

Findings - The results were also collated with other variability indicators and measures of location, asymmetry and skewness. The study demonstrated that the Gini coefficient produces similar results to a simple coefficient of variation, but is less sensitive to price changes.

Research implications - The main research limitation was a small database covering a single location; therefore, the article could be regarded as a case study. The study demonstrated that the Gini coefficient can be applied as an auxiliary tool in assessments of price dispersion on the real estate market.

Keywords:	Gini coefficient; coefficient of variation; lorenz curve; real estate marker analyses
JEL codes:	C10; C46; R21
Article type:	case study
DOI:	10.14659/WOREJ.2020.111.01

INTRODUCTION

Real estate prices are characterized by significant variation in time and space. The basic statistical tools for analyzing price dispersion include standard deviation, variance, coefficient of variation, and range. Apart from the measures of variability, one can also distinguish the measures of location, concentration, asymmetry and skewness. The Gini coefficient is a less popular measure of concentration, and it is used mainly to assess social inequality and to quantify the unevenness in variable distribution, including the distribution of income. This study aimed to evaluate the applicability of the Gini coefficient as an auxiliary tool for analyzing price variations. We applied the Gini Coefficient to our analysis, which is a measure of concentration. We did not apply a Gini's mean difference, which is a measure of distribution. The article discusses the dispersion of real estate prices, the heterogeneity of real estate, and statistical indicators for measuring variations in databases of real estate prices. Changes in the values of the Gini coefficient were examined in databases that were gradually reduced by extreme prices. The primary database was composed of real estate prices in three districts of Warsaw (Mokotów, Ochota, Wola) in 2017-2018. The results were also collated with other descriptive statistics. The study demonstrated that the Gini coefficient can be used as an auxiliary tool in heterogeneous databases to eliminate outliers. The Gini coefficient produces similar results to a simple coefficient of variation, but it is less sensitive to price changes.

The article has the following structure. The first chapter reviews the literature on the implications following from the heterogeneity of the real estate market; it discusses the problems associated with removing outlying transactions from databases and presents the standard applications of the Gini coefficient in research. The second chapter presents the methodology for calculating the Gini coefficient and describes the analyzed database. The results of the analysis are presented and discussed in the third chapter. The conclusions stemming from the results of the study are formulated in the last chapter.

LITERATURE REVIEW

The real estate market is a challenging object of study. These difficulties stem mainly from the heterogeneity of real estate – all properties differ from one another in specific characteristics (see: Barańska, 2016). Real estate is heterogeneous in terms of prices (Galati & Teppa, 2017), the parties involved in real estate transactions (Ozhegov & Sidorovykh, 2017; Qiu & Tu, 2018),

and the spatial distribution of housing prices (Wen, Jin & Zhang, 2017; Wu, Wei & Li, 2020). In the literature, the heterogeneity of real estate is often linked with price dispersion (Leung, Leong & Wong, 2006). Homogeneous products are rarely sold for the same price by different sellers, in the same market, under the same conditions of sale, and with the same geolocation. Therefore, even properties with similar characteristics can be sold at different prices, and the prices on a given market can be highly dispersed. Recent research has demonstrated that the variations in real estate prices and attributes are large enough to exert a significant impact on the market (Zyga, 2015; 2019). According to Renigier-Bilozor, Janowski and Walacik (2019, p. 3), this observation poses numerous challenges in research, in particular during the development of the optimal databases. Źróbek et al. (2020) noted that the complexity of market analyses can be attributed to a large number of real estate characteristics, including on the market of agricultural land. In contrast, Renigier-Biłozor, Janowski and d'Amato (2019) have argued that insufficient data pose the key problem in market analyses. They proposed an algorithm of non-deterministic relationships between real estate variables which were tested on small datasets of commercial real estate in Italy and residential real estate in Poland.

The elimination of outlying transactions also poses a considerable problem in database design. This problem is encountered when databases are developed with the use of microeconomic data. The selection of outlying observations is also often arbitrary. Brzezicka, Wisniewski and Figurska (2018, p. 518) eliminated 0.5% of the lowest prices and 0.5% of the highest prices from the database. In a study by Brzezicka et al. (2019, p. 5), 1% of the transactions that deviated considerably from the average market price were removed from every quarter at the top and bottom of the price range, and the eliminated transactions were uniformly distributed in time. The top 5% and the bottom 5% of the observations are eliminated arbitrarily from some databases. This approach was used by Case, Shiller and Thompson (2012) to analyze the surveyed respondents' expectations on home prices.

The problems associated with the elimination of outlying observations from the database can be resolved with the use of popular statistical and econometric tools. Econometric tools include residual analysis, whereas statistical tools involve prediction errors and measures of variability (such as standard deviation from the mean) which are described by the three-sigma rule. The three-sigma rule states that for normal distribution, 99.7% of the population lies within three standard deviations of the mean. The three-sigma rule was used by Renigier (2005), whereas Cichociński (2011) relied on descriptive statistics and estimation errors (mean squared error, MSE; mean

squared deviation ratio, MSDR) in their analyses. In the international literature, descriptive statistics (mean price, standard deviation) were used by Ben-Shahar and Golan (2019) to evaluate price dispersion. Chiang et al. (2019) additionally calculated the coefficient of variation in their study. Cook's distance, which is used in multiple regression models is an important method of eliminating outliers from the dataset. The method shows how much the residual ratios change when the observation is removed from the analysis. This method is used by Nalepka, Tomal (2016) in the context of the real estate market. There are more sophisticated methods of identifying outliers in the database in the literature (e.g. DFFITS, DFBETAS, and COVRATIO), but because of the basic and introductory nature of our research more advanced methods have not been described.

The most popular statistical measures of dispersion for analyzing price variability include the coefficient of variation which denotes the relationship between the mean value and standard deviation, as well as standard deviation, quartile deviation, range, interquartile range, and the interpercentile range. The Gini coefficient is less widely applied. The Gini coefficient is a measure of concentration that assesses the inequality of distribution of the analyzed categories on a scale of 0 to 1, where 0 (1) represents perfect equality (inequality). The higher the value of the Gini coefficient, the greater the inequality in the evaluated sample (the equation for calculating the Gini coefficient is presented in the next chapter). Gini coefficient is applied to assess the inequality of home values by Aladangady, Albouy and Zabek (2017). The authors also calculate the Gini coefficient to assess inequality over time in rents, housing consumption and household income. The Gini coefficient is widely used to assess income inequality (Gray, 2020), and it is applied in real estate market analyses in the same context. According to Zhang, Jia and Yang (2016), the Gini coefficient is positively correlated with the housing price-to-income ratio as well as the housing vacancy rate. The empirical results reported by Özmen, Kalafatcılar and Yılmaz (2019) based on panel data revealed that an increase in the Gini coefficient (higher income inequality) reduced the sensitivity of housing prices to income changes. The Gini coefficient is equal to twice the area between the Lorenz curve and the diagonal of the unit square. The Lorenz curve describes the concentration of one-dimensional distribution of the analyzed variable, and it is contained in a square with sides measuring 1 unit. The 45-degree line represents perfect income equality.

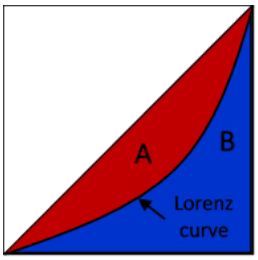


Figure 1. Lorenz curve
Source: own elaboration.

RESEARCH METHODOLOGY

The study was performed on a database of real estate prices, developed based on data from the Register of Real Estate Prices and Values kept by the Office of Geodetic and Cadastral Data in Warsaw. The analysis was limited to residential real estate traded on the secondary market in three districts of Warsaw (Mokotów, Ochota and Wola) in 2017-2018. The analyzed districts are situated in the central part of Warsaw, and together with the district of Żoliborz, they constitute a larger macro-district (cf. Brzezicka et al., 2019). These districts were selected for the study due to their central location in the city, similar price levels, low volatility of volume of transactions, homogeneous research area and data availability. The analysis covered apartments with a floor area of 40 to 60 m^2. The database was composed of 2341 transactions. There was no individual information about the property attributes in the database and the transactions for the study were not selected due to the similarity of properties. For this reason, the obtained results may be biased and should be accepted with caution.

The study aimed to determine the extent to which data reduction affects variability parameters in the database and to calculate the Gini coefficient for successive databases which were obtained by reducing the primary database. The Gini coefficient was calculated with equation (1). The calculations were performed in the R Studio software package.

$$G(y) = \frac{\sum_{i=1}^{n}(2i - n - 1)yi}{n^2\overline{y}} \quad (1)$$

where:

y_i - every successive price in the database,

$\overline{y}$ - mean price in the database.

The database was reduced in 15 iterations. In each step, the database was reduced by eliminating 1% of the lowest prices and 1% of the highest prices. This approach produced 15 databases, where each successive database contained 2% fewer transactions than the previous database (base_0 – database without data reduction; base_1 – database obtained by eliminating 1% of the highest prices and 1% of the lowest prices; base_2 – database obtained by eliminating 2% of the highest prices and 2% of the lowest prices from base_0; base_15 – database obtained by eliminating 15% of the highest prices and 15% of the lowest prices from base_0).

The Gini coefficient and other descriptive statistics were calculated for each database to determine the applicability of the Gini coefficient for evaluating the dispersion of prices on the real estate market. Lorenz curves were plotted for selected databases. The analyzed measures of dispersion were the coefficient of variation, standard deviation, quartile deviation, minimum, maximum, range, lower quartile, upper quartile, interquartile range, 10th percentile, 90th percentile, and the interpercentile range. Measures of location, concentration and asymmetry were also calculated for other groups, including the mean, median, dominant, amount, kurtosis and skewness.

RESULTS AND DISCUSSION

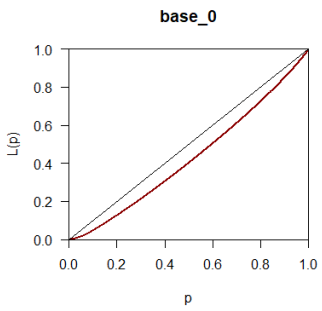
The basic descriptive statistics for selected databases are presented in Table 1. Special attention was paid to the values of the Gini coefficient and the shape of the Lorenz curve. Lorenz curves and the distribution of prices in ascending order for base_0, base_1, base_5 and base_15 are presented in Figure 2. The values of the Gini coefficient and the coefficient of variation were compared in Figure 3.

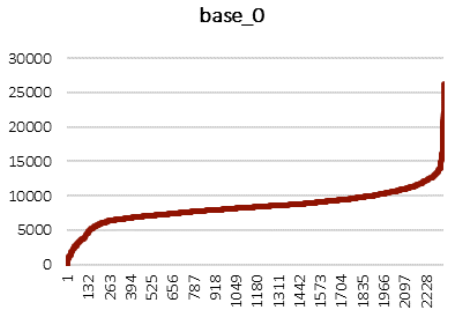
Table 1. Descriptive statistics

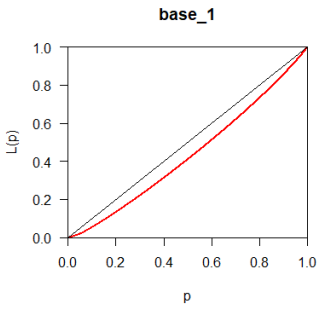
	base_0	base_1	base_2	base_3	base_5	base_10	base_15
Measures of concentration and asymmetry							
Gini coefficient	0.145	0.131	0.123	0.115	0.102	0.081	0.068
Kurtosis	4.04	1.08	0.72	0.42	-0.16	-0.67	-0.80
Skewness	0.16	-0.31	-0.17	-0.08	0.13	0.26	0.23
Measures of dispersion							
Coefficient of variation	0.28	0.24	0.22	0.21	0.18	0.14	0.12
Standard deviation	2 323	2 014	1 878	1 741	1 520	1 206	1 012
Quartile deviation	1 134	1 107	1 080	1 067	1 017	889	759
Minimum	55	1 427	2 404	3 178	4 237	6 151	6 661
Maximum	26 292	13 839	13 337	12 764	12 176	11 222	10 686
Range	26 237	12 412	10 933	9 587	7 939	5 071	4 025

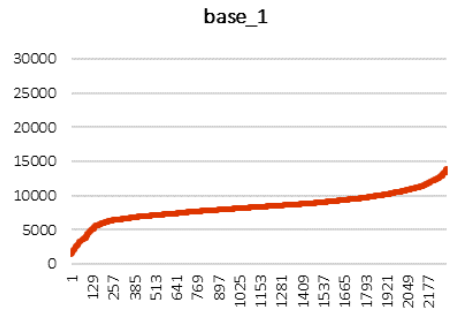
Lower Quartile	7 267	7 276	7 313	7 329	7 384	7 536	7 684
Upper Quartile	9 534	9 489	9 474	9 464	9 418	9 314	9 203
Interquartile range	2 267	2 213	2 161	2 135	2 034	1 778	1 519
Percentile_10	6 141	6 277	6 388	6 460	6 574	6 900	7 125
Percentile_90	11 068	10 935	10 852	10 734	10 570	10 225	9 949
Interpercentile range	4 927	4 658	4 464	4 274	3 996	3 325	2 825
Measures of location							
Mean	8 402	8 375	8 406	8 418	8 442	8 459	8 468
Median	8 366	8 365	8 366	8 367	8 372	8 386	8 402
Dominant	8 610	8 610	8 610	8 610	8 610	8 610	8 610
Amount	2 341	2 288	2 250	2 203	2 113	1 896	1 689

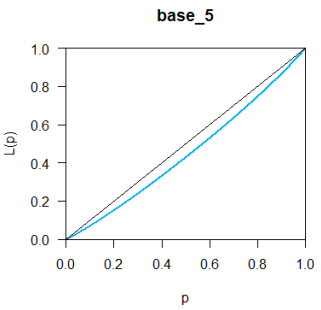
Source: own elaboration.

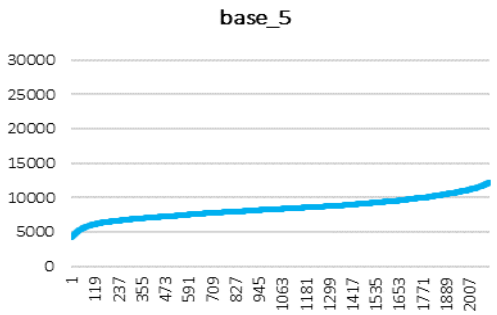












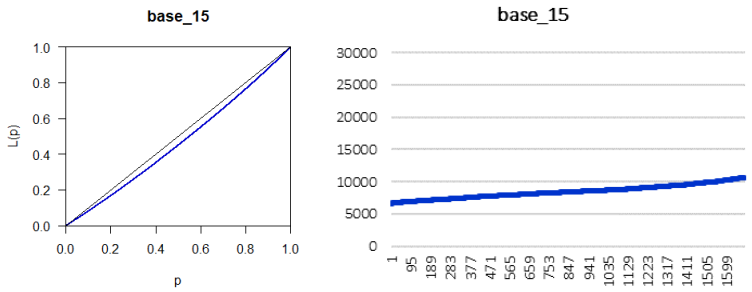


Figure 2. Lorenz curve and price distribution in successive databases: base_0, base_1, base_5, base_15
Source: own elaboration.

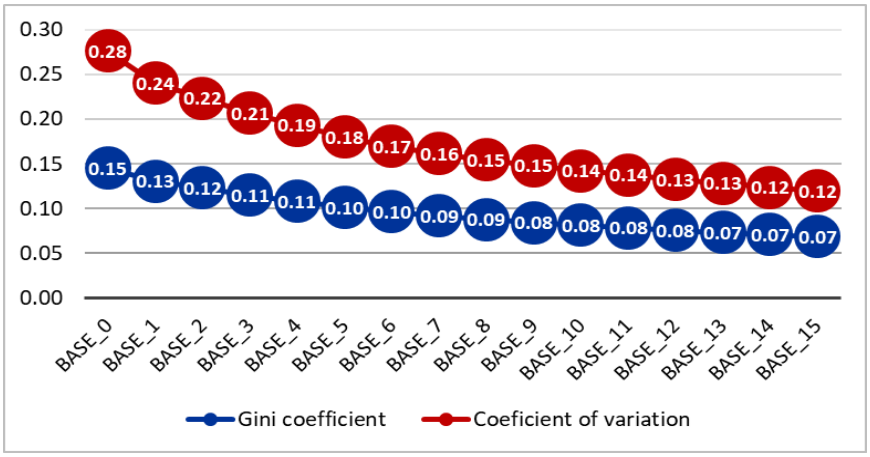


Figure 3. Gini coefficient and coefficient of variation
Source: own elaboration.

Every successive database was characterized by lower price variation due to the criteria and rules adopted in the design process. In the primary database (base_0), the minimum price per one square meter of the apartment area was PLN 55/m^2, and the maximum price was PLN 26,292/m^2. In base_15, the corresponding values were PLN 6,661/m^2 and PLN 10,686/m^2. The measures of variability denoting price distribution (range, interquartile range and interpercentile range) decreased with every reduction. The measures of location (mean and median) displayed a growing trend, whereas skewness (excluding the primary database) increased and decreased respectively.

The values of the Gini coefficient and the coefficient of variation decreased, but the Gini coefficient decreased at a slower rate. The initial reductions, where 1-5% of extreme values were eliminated from the database, played a key role. Considerable differences in price distribution can

be observed on the right side of Figure 2. However, the changes illustrated by the Lorenz curve in successive iterations were less extensive. The Gini coefficient was less sensitive to the elimination of outlying transactions than the coefficient of variation, and a minor flattening was observed in the shape of the Lorenz curve. These results suggest that the elimination of equal percentages of extreme transactions from the database is not an optimal solution. Special attention should be paid to the range of vertical axes (expressing prices) on the right side of Figure 2. Although an identical number of transactions were removed from the top and bottom of the distribution, the resulting database was not "symmetrical" due to individual variations.

In the next step, correlogram depicting the strength and direction of the correlations between descriptive statistics and the Gini coefficient was developed for successive databases (see: Fig. 4). The Gini coefficient was bound by strong positive correlations with the coefficient of variation, standard deviation, quartile deviation, interquartile range, and the interpercentile, and it was partially correlated with kurtosis. The Gini coefficient was negatively correlated with the mean, median and the coefficient of kurtosis.

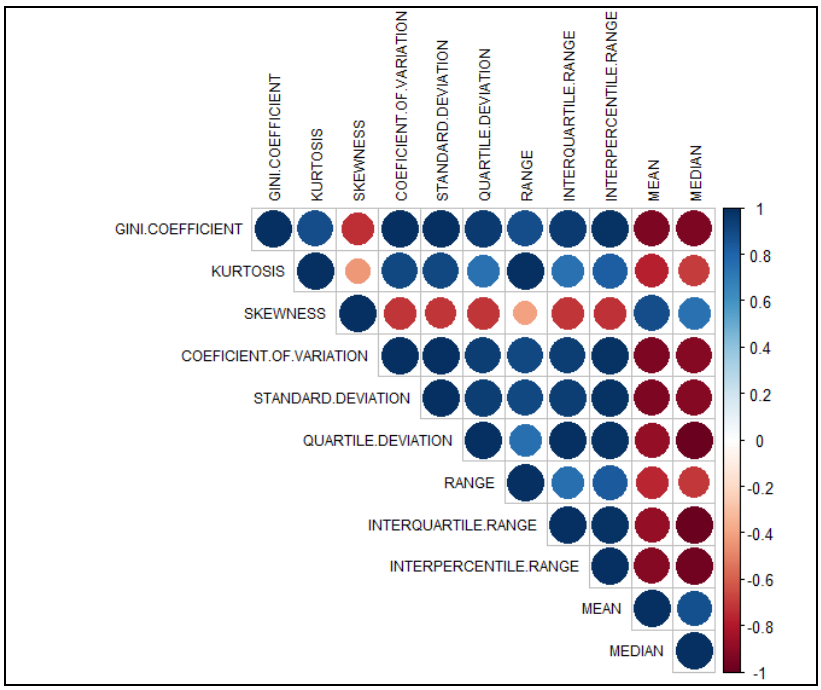


Figure 4. Gini coefficient and descriptive statistics - correlogram
Source: own elaboration.

CONCLUSION

In the current study, the Gini coefficient produced similar, but somewhat lower results than a simple coefficient of variation. The results of the analysis indicate that the Gini coefficient is less sensitive to price changes than the coefficient of variation, which is particularly visible in the first reduction step (1% of the highest and lowest prices) and several following steps. Further database reductions (by more than 5% of the highest and lowest prices) did not induce significant differences in the values of the analyzed indicators or their change rates. It should be noted that both the Gini coefficient and the coefficient of variability make a reference to the mean, but unlike the coefficient of variation, the Gini coefficient does not make a reference to standard deviation, which could explain the lower values of this indicator. The results of this study indicate that the Gini coefficient can be applied as an auxiliary tool in analyses of price dispersion on the real estate market, even though it is typically used for different purposes in research.

The main research limitation was the size of the analyzed database, and further research could be undertaken to test the applicability of the Gini coefficient in analyses of smaller and larger datasets. The prices of apartments in three Warsaw districts were analyzed, and the developed database consisted of 2341 transactions. The Gini coefficient is often used to determine income inequality in smaller datasets of several dozen or several hundred observations. Therefore, the present findings may not apply to databases containing a significantly lower or higher number of real estate prices. In future research, we intend to calculate the Gini's mean difference and assess the possibility of using this measure as a tool for measuring price variability in the housing market. Moreover, we plan to conduct the research with the Gini coefficient with a greater degree of details.

REFERENCES

Aladangady, A., Albouy, D., & Zabek, M. (2017). Housing Inequality. *NBER Working Paper Series*, no. 21916.

Barańska, A. (2016). The Significance of Database Size in Modelling the Market of Nonresidential Premises. *Real Estate Management and Valuation*, 24(2), 47-56. https://doi.org/10.1515/remav-2016-0013.

Ben-Shahar, D., & Golan, R. (2019). Improved Information Shock and Price Dispersion: A Natural Experiment in the Housing Market. *Journal of Urban Economics*, 112, 70-84. https://doi.org/10.1016/j.jue.2019.05.008.

Brzezicka, J., Łaszek, J., Olszewski, K., & Waszczuk, J. (2019). Analysis of the Filtering Process and the Ripple Effect on the Primary and Secondary Housing Market in Warsaw, Poland. *Land Use Policy*, 88, 104098. https://doi.org/10.1016/j.landusepol.2019.104098.

Brzezicka, J., Wisniewski, R., & Figurska, M. (2018). Disequilibrium in the Real Estate Market: Evidence from Poland. *Land Use Policy*, 78, 515-531. https://doi.org/10.1016/j.landusepol.2018.06.013.

Case, K. E., Shiller, R. J., & Thompson, A. (2012). What Have they Been Thinking? Home Buyer Behavior in Hot and Cold Markets (No. w18400). *National Bureau of Economic Research.*

Chiang, Y. H., Ku, Y., Liu, F., & Chang, C. O. (2019). House Price Dispersion in Taipei Residential Communities. *International Real Estate Review*, 22(1), 109-129.

Cichociński, P. (2011). Porównanie metod interpolacji przestrzennej w odniesieniu do wartości nieruchomości. *Studia i Materiały Towarzystwa Naukowego Nieruchomości*, 19(3), 120-125.

Galati, G., & Teppa, F. (2017). Heterogeneity in House Price Dynamics. *De Nedrelandsche Bank Working Paper*, no. 564.

Gray, R. (2020). Inequality in Nineteenth Century Manhattan: Evidence from the Housing Market (No. 2020-02). QUCEH Working Paper Series.

Leung, C. K. Y., Leong, Y. C. F., & Wong, S. K. (2006). Housing Price Dispersion: an Empirical Investigation. *The Journal of Real Estate Finance and Economics*, 32(3), 357-385. https://doi.org/10.1007/s11146-006-6806-7.

Nalepka, A., & Tomal, M. (2016). Identyfikacja czynników kształtujących ceny ofertowe deweloperskich lokali mieszkalnych na obszarze jednostki ewidencyjnej Nowa Huta. *Świat Nieruchomości*, 96, 11-18. https://doi.org/10.14659/worej.2016.96.02.

Ozhegov, E. M., & Sidorovykh, A. S. (2017). Heterogeneity of Sellers in Housing Market: Difference in Pricing Strategies. *Journal of Housing Economics*, 37, 42-51. https://doi.org/10.1016/j.jhe.2017.03.002.

Özmen, M. U., Kalafatcılar, M. K., & Yılmaz, E. (2019). The Impact of Income Distribution on House Prices. *Central Bank Review*, 19(2), 45-58. https://doi.org/10.1016/j.cbrev.2019.05.001.

Qiu, L., & Tu, Y. (2018). Homebuyers' Heterogeneity and Housing Prices. Available at SSRN 3187233 https://ssrn.com/abstract=3187233.

Renigier, M. (2005). Budowa modelu geostatystycznego z wykorzystaniem reszt. *Acta Scientiarum Polonorum. Administratio Locorum*, 4(1-2), 83-96.

Renigier-Biłozor, M., Janowski, A., & d'Amato, M. (2019). Automated Valuation Model Based on Fuzzy and Rough Set Theory for Real Estate Market with Insufficient Source Data. *Land Use Policy*, 87, 104021. https://doi.org/10.1016/j.landusepol.2019.104021.

Renigier-Biłozor, M., Janowski, A., & Walacik, M. (2019). Geoscience Methods in Real Estate Market Analyses Subjectivity Decrease. *Geosciences*, 9(3), 130. https://doi.org/10.3390/geosciences9030130.

Wen, H., Jin, Y., & Zhang, L. (2017). Spatial Heterogeneity in Implicit Housing Prices: Evidence from Hangzhou, China. *International Journal of Strategic Property Management*, 21(1), 15-28. https://doi.org/10.3846/1648715X.2016.1247021.

Wu, Y., Wei, Y. D., & Li, H. (2020). Analyzing Spatial Heterogeneity of Housing Prices Using Large Datasets. *Applied Spatial Analysis and Policy*, 13(1), 223-256. https://doi.org/10.1007/s12061-019-09301-x.

Zhang, C., Jia, S., & Yang, R. (2016). Housing Affordability and Housing Vacancy in China: The Role of Income Inequality. *Journal of Housing Economics*, 33, 4-14. https://doi.org/10.1016/j.jhe.2016.05.005.

Zyga, J. (2015). Search for Dissimilarity Factors for Nominally Indiscernible Facilities. *Real Estate Management and Valuation*, 23(3), 65-72. https://doi.org/10.1515/remav-2015-0026.

Zyga, J. (2019). Dissimilarity as a Component of the Property Price Model. *Real Estate Management and Valuation*, 27(3), 124-132. https://doi.org/10.2478/remav-2019-0030.

Źróbek, S., Kovalyshyn, O., Renigier-Biłozor, M., Kovalyshyn, S., & Kovalyshyn, O. (2020). Fuzzy Logic Method of Valuation Supporting Sustainable Development of the Agricultural Land Market. *Sustainable Development*, 1, 1-12, https://doi.org/10.1002/sd.2061.