www.degruyter.com/view/j/remav

THE USE OF STATISTICAL METHODS FOR DETERMINING ATTRIBUTE WEIGHTS AND THE INFLUENCE OF ATTRIBUTES ON PROPERTY VALUE[1]

Anna Gdakowicz
Uniwersytet Szczeciński
e-mail: anna.gdakowicz@usz.edu.pl

Ewa Putek-Szeląg
Uniwersytet Szczeciński
e-mail: ewa.putek-szelag@usz.edu.pl

Abstract

Determining the impact of individual attributes on the value or price of real estate in business practice poses many problems. One of the solutions to this problem is the use of statistical methods. The article proposes correlation coefficients (and their partial modifications) that can be used to determine the impact of selected features on the value of real estate. In addition, several procedures were taken into account for the factors in further calculations, using different methods for determining weights. Empirical verification of the proposed solutions was based on the mass valuation of land properties. The obtained results were compared with valuations developed by property appraisers and valuation errors were calculated. Based on valuation errors, the proposed methods of calculation procedures were ranged, indicating those which provide results closest to the individual valuations carried out by property appraisers.

Key words: *attribute weights, dependency coefficients, mass real estate valuation*.

JEL Classification: *C40, C89, R33*.

Citation: Gdakowicz, A., & Putek-Szeląg, E. (2020). The use of statistical methods for determining attribute weights and the influence of attributes on property value. *Real Estate Management and Valuation*, 28(4), 33-47.

DOI: https://doi.org/10.1515/remav-2020-0030

1. Introduction

Real estate market analysis, describing real estate and their attributes is often associated with operating qualitative or quasi-quantitative (ordinal) features. This type of feature is most often presented in verbal form, which describes the condition of the feature (e.g. the purpose of the property: housing, recreation, agricultural, industrial) or allows us to rank variants of the feature from the smallest to the largest, from the weakest to the strongest (e.g. neighborhood: unfavorable, average, favorable). The number of adopted variants and assessment of their impact on the analyzed properties depends on the experience of the analyst, the adopted research objective and the researcher's determination of the condition of individual properties. To ensure the standardization of the features selected for the description of the property, at the initial stage of the study, it is necessary to define the method of qualifying each of the states of individual features of the described properties, e.g. when

[1] The article financed by the National Science Center project, registration number 2017/25/B/HS4/01813.

the neighborhood should be considered favorable, when average, and in which cases the neighborhood will be unfavorable.

The use of order features obliges the researcher to use appropriate statistical methods. The use of statistical measures is tightened by the conditions regarding the type of characteristics available to the researcher and the form of the data (statistical series or table). In the case of real estate market analysis, real estate features (attributes) are most often ordinal features, and the collected data is ordered in statistical series. There are a number of statistical measures that can be used to analyze interdependence for such a set of data.

The article proposes seven dependency coefficients (including partial coefficients)[2] that can be used to calculate the relationship between attributes of a given property and its value (or price). The calculated coefficients were defined as the weights of the attributes (attributes). Then, four ways to include the received measures in further calculations and six ways to determine the impact of attributes on the value of the property (weighting) were proposed. The applied calculation procedures were used to estimate the value of real estate using the Szczecin Mass Real Estate Valuation Algorithm (SAMWN), and then the obtained results were compared with valuations of property appraisers by calculating valuation errors. The purpose of the article is to present ways of using statistical methods to calculate the weight of property characteristics and their impact on its value, and to indicate the best relationship factor and weighting method, so that the obtained property values (estimated based on SAMWN) are as close as possible to the value of real estate received in individual valuations of property appraisers.

2. Literature review

When assessing the market value of real estate, the property appraiser determines the impact of individual levels of attributes (features) of the property on its value. In accordance with applicable regulations, this can be done based on (*Powszechne Krajowe Zasady Wyceny*, 2008):

1. the results of analyses of data on prices and market features of similar real estate traded on the real estate market specified for the purpose of the valuation,
2. analogy to similar local markets in terms of type and area,
3. examination and /or observation of preferences of potential real estate buyers,
4. another reliable way.

The second and third of the proposed methods are relatively the simplest methods. They require the property appraiser to constantly monitor the literature on the subject and use available studies on the relevant market. The first way to determine the impact of individual attributes is more complicated. In this case, it is necessary to conduct analyses with an appropriate objective, spatial and temporal coverage. In addition, the requirements of this type of research, such as - appropriate database size, ensuring sample representativeness, etc. should be taken into account. Despite the fact that this method may give rise to many complications at the level of the test methodology, it simultaneously ensures objectivity of the results and often also the possibility of their generalizations.

The problem of using appropriate coefficients to calculate the relationship between property features (attributes) is increasingly discussed. The authors point out the correct use of coefficients depending on the scale on which the attributes are presented (Dmytrów, et al., 2019; Babatunde, 2018; Gaca, 2018; Doszyń, 2017; Foryś & Gaca, 2016; Gaca & Sawiłow, 2014). Attention is also paid to the type of relationship between variables - straight or curvilinear (Barańska, 2019a, 2019b) and to issues related to the collinearity of variables (Doszyń, 2019).

3. Methodology

Taking into account the specifics of the attributes that describe valued properties (ordinal features, presented on the ordinal scale), the following relationship factors have been proposed to calculate the impact of individual features on the value of the property:

- Spearman's rank correlation ρ coefficient,
- Kendall's correlation τ coefficient,
- gamma statistic G,
- Pearson's linear correlation r coefficient.

[2] The use of partial coefficients is aimed at eliminating collinearity of variables.

For selected coefficients - Spearman and Kendall - modifications were used that take into account the occurrence of bound ranks. Bundled ranks are observed when the number of feature categories is smaller than the number of analyzed objects. This situation occurs when analyzing the real estate market. The number of categories of real estate attributes usually ranges from three (e.g. unfavorable, average, favorable) to several (e.g. bad, moderately bad, average, moderately good, good), and the analyzed real estate base is much more numerous. The number of attribute categories is obviously not specified anywhere and results from the need to take into account all levels that differentiate a given feature and the experience of the property appraiser.

The Spearman rank correlation coefficient is calculated as follows (Kendall, 1948):

$$\rho_{xy} = \rho_{yx} = \frac{\frac{1}{6}(n^3 - n) - (\sum_{i=1}^{n} d_i^2) - T_x - T_y}{\sqrt{\left(\frac{1}{6}(n^3 - n) - 2T_X\right)\left(\frac{1}{6}(n^3 - n) - 2T_Y\right)}} \quad (1)$$

where:

$$d_i = \text{rank } x_i - \text{rank } y_i,$$
$$T_x = \frac{1}{12}\sum_j (t_j^3 - t_j),$$
$$T_y = \frac{1}{12}\sum_k (u_k^3 - u_k),$$

t_j – the number of observations in the sample that have the same j rank value of the feature x,
u_k – the number of observations in the sample that have the same k value of the rank of the y feature,
n – the number of observations.

Kendall's τ correlation coefficient occurs in several variants: τ_A, τ_B and τ_C. The τ_A coefficient is calculated assuming that there are no associated ranks. Kendall's τ_B coefficient assumes that there are bundled ranks, while the τ_C coefficient is most appropriate for data in the form of a contingency table. In the case of data used in the study, where the number of observations is much higher than the number of variations of the tested attributes, there will certainly be associated ranks. Therefore, the τ_B Kendall's coefficient was used, which was calculated as follows (Parker, et al., 2011):

- The observations were combined into all possible pairs: (x_i, y_i) and (x_j, y_j), $i \neq j$.
- If both $x_i > x_j$ and $y_i > y_j$ or $x_i < x_j$ and $y_i < y_j$, then such a pair is called *matching*. The number of such pairs is n_c.
- If, on the other hand, $x_i > x_j$ and $y_i < y_j$ or $x_i < x_j$ and $y_i > y_j$, then such a pair is called *incompatible*. The number of such pairs is n_d.
- If $x_i = x_j$ or $y_i = y_j$, then this pair is neither *compatible* nor *incompatible*. It is a *tied* pair.

Having the above values determined, the Kendall's τ_B coefficient is calculated according to the formula:

$$\tau_B = \frac{n_c - n_d}{\sqrt{(n_0 - n_1)(n_0 - n_2)}}, \quad (2)$$

where:

$n_1 = \frac{\sum_i t_i(t_i - 1)}{2}$,
$n_2 = \frac{\sum_j u_j(u_j - 1)}{2}$,
t_i – the number of observations in the sample having the same i rank value of the feature x,
u_j – the number of observations in the sample that have the same j value of the rank of the feature y.

The next factor is the statistic gamma G (Goodman & Kruskal, 1963). This is the measure indicated as more appropriate than the previous two coefficients for calculating the relationship between the features presented on the ordinal scale with many tied weights. The factor tests the orderliness of the selected observation pairs. Data must be presented in a contingency table. The gamma statistic G is calculated according to the formula (Siegel & Castellan, 1988):

$$G = \frac{\#\ agreements - \#\ disagreements}{\#\ agreements + \#\ disagreements} \quad (3)$$

where:

$$\#\ agreements = \#(+) = \sum_{i=1}^{r-1}\sum_{j=1}^{k-1} n_{ij}N_{ij}^{+} \tag{3a}$$

$$\#\ disagreements = \#(-) = \sum_{i=1}^{r-1}\sum_{j=1}^{k-1} n_{ij}N_{ij}^{-} \tag{3b}$$

N_{ij}^{+} - the sum of all of the frequencies *below and to the right* of the *ij*-th cell,
N_{ij}^{-} - the sum of all of the frequencies *below and to the left* of the *ij*-th cell.

In the case when there are no associated ranks, i.e. all partial numbers in a contingency table are equal to 1 or 0, then the statistic gamma *G* is equal to the Kendall's coefficient.

The last of the coefficients used is the Pearson's linear correlation coefficient *r*. As the name suggests, this is a factor that is dedicated to analyzing relationships characterized by straightness. When the effect of the features indicates a curvilinear relationship, Pearson's *r* correlation coefficient will falsely indicate the undervaluation. The additional limitation concerns the type of features for which this factor applies, seeing as how these should be the quantitative. A problem arises when the trait is a trait presented on the ordinal scale, but the variants are coded not in the verbal description, but digitally (e.g. 1 for the most favorable state and subsequent digits for the weaker states). The feature may give the impression of a feature represented on a quotient scale. In such a situation, the researcher's experience should exclude this factor from further analyses, however, due to the popularity of this measure and ease of calculation (the factor is programmed in most computer programs), it is very often used. The conducted studies took into account Pearson's coefficient due to its popularity and widespread use, but it should be emphasized that its use is not substantively justified in this case, because the attributes of real estate are the features presented on an ordinal scale (more about the popularity of this coefficient and its universality of use is found in ARMSTRONG 2019). The coefficient was calculated according to the formula:

$$r_{xy} = r_{yx} = \frac{cov(x,y)}{S_x\, S_y} \tag{4}$$

where:
cov(x,y) – covariance between the examined features,
S_x – standard deviation of the characteristic of *x*,
S_y – standard deviation of the characteristic of *y*.

All coefficients determined using formulas (1) – (4) have the following properties:
- they are symmetrical,
- they take values from the range ⟨-1; 1⟩ - measure both strength and direction of the relationship,
- if their value is negative, then the increase in one feature is accompanied by a decrease in the other, and vice versa,
- if their value is positive, then the growth of one characteristic is accompanied by the growth of the other, and vice versa,
- if their value is 0, then there is no relationship between the features, and if 1 or -1 – the relationship is functional. The closer their value is to 1 or -1, the stronger the relationship.

For coefficients (1) – (3), both variables must be measured at least on the ordinal scale. For coefficient (3), it is required to prepare a contingency table, while the remaining coefficients are calculated for data presented in a correlation series. The linearity of the relationship is assumed a priori for coefficient (4); other coefficients can be used for features with a curvilinear relationship.

Analyzing the relationship between the value of 1 m2 of real estate and individual features of these properties, one should eliminate the impact of other features that may disturb the examined interdependence. To this end, partial coefficients (except for statistic *G*) have also been calculated for the above coefficients, according to the formula:

$$v_{yx.z} = -\frac{R_{\mathrm{yx}}}{\sqrt{R_{\mathrm{yy}} \cdot R_{\mathrm{xx}}}} \tag{5}$$

where:
v – appropriate relationship factor,
y – vector of the explained variable value,
x – vector of the explaining variable value,
z – vector (or matrix) of the remaining variables,

R_{yx} – determinant of a matrix cofactor obtained by removing the row corresponding to the y variable and the column corresponding to the x variable,

R_{yy} – determinant of a matrix cofactor obtained by removing the row and column corresponding to the y variable,

R_{xx} – determinant of a matrix cofactor obtained by removing the row and column corresponding to the x variable.

In addition, four variants were introduced to take into account the impact of the factors on the value of the property. Further calculations took into account: only statistically significant coefficients or the square of these coefficients and all coefficients or their square. Detailed variants of the use of dependency coefficients are presented in Table 1. Each proposed variant was described by the letter of the dependency coefficient. The subscript indicates the use of partial coefficients (letter c) and taking into account individual measures of dependence (only significant - i, square of significant coefficients - i2, all coefficients - w, square of all coefficients - w2).

Table 1

The use of dependency coefficients - variants

Correlation coefficients	Calculation variants			
	significant coefficients	significant square coefficients	all coefficients	square of all coefficients
ρ Spearman's	$S_{i.}$	$S_{i2.}$	$S_{w.}$	$S_{w2.}$
partial ρ Spearman's	$S_{ci.}$	$S_{ci2.}$	$S_{cw.}$	$S_{cw2.}$
τ_B Kendall's	$T_{i.}$	$T_{i2.}$	$T_{w.}$	$T_{w2.}$
partial τ_B Kendall's	$T_{ci.}$	$T_{ci2.}$	$T_{cw.}$	$T_{cw2.}$
statistic gamma G	$G_{i.}$	$G_{i2.}$	$G_{w.}$	$G_{w2.}$
r Pearson's	$r_{i.}$	$r_{i2.}$	$r_{w.}$	$r_{w2.}$
partial r Pearson's	$r_{ci.}$	$r_{ci2.}$	$r_{cw.}$	$r_{cw2.}$

Source: own study.

In the index of coefficient variants used, after the dot (Table 1), the number of the weighing method for individual cases was specified. Six ways were proposed:

1) Ratio of the average value of real estate with the best attribute states to the average value of real estate with the weakest attribute states:

$$1 + \alpha_{kp} = e^{\frac{w_{śmaxk}}{w_{śmink}}} \tag{6}$$

2) Ratio of the median value of the property with the best attribute states to the median value of the property with the weakest attribute states:

$$1 + \alpha_{kp} = e^{\frac{w_{Mmaxk}}{w_{Mmink}}} \tag{7}$$

3) Ratio of the average value of real estate with the best attribute states to the average value of real estate with the weakest attribute states, corrected by the factor $\ln(\frac{w_{max}}{w_{min}})$:

$$1 + \alpha_{kp} = e^{ln\left(\frac{w_{max}}{w_{min}}\right).\frac{w_{śmaxk}}{w_{śmink}}} \tag{8}$$

4) Ratio of the median value of the property with the best attribute states to the median value of the weakest attribute states, corrected by the factor $\ln(\frac{w_{max}}{w_{min}})$:

$$1 + \alpha_{kp} = e^{ln\left(\frac{w_{max}}{w_{min}}\right).\frac{w_{Mmaxk}}{w_{Mmink}}} \tag{9}$$

5) Ratio of the coefficient of variation of the value of real estate with the best attribute states to the coefficient of variation of the value of real estate with the weakest attribute states:

$$1 + \alpha_{kp} = e^{\frac{w_{Vmaxk}}{w_{Vmink}}} \quad (10)$$

6) Ratio of the coefficient of variation of the value of the property with the best attribute states to the coefficient of variation of the value of the property with the weakest attribute states, corrected by the coefficient $\ln(\frac{w_{max}}{w_{min}})$:

$$1 + \alpha_{kp} = e^{ln\left(\frac{w_{max}}{w_{min}}\right)\frac{w_{Vmaxk}}{w_{Vmink}}} \quad (11)$$

where:
$w_{śmaxk}$, w_{Mmaxk}, w_{Vmaxk}– respectively: average, median, coefficient of variation of the property value for the highest category of the k-th attribute,
$w_{śmink}$, w_{Mmink}, w_{Vmink} – respectively: average, median, coefficient of variation of the property value for the lowest category of the k-th attribute,
w_{max} – maximum value of a property in the data set,
w_{min} –minimum value of a property in the data set.

Whether the proposed method of calculating the impact of attributes on the value of real estate can be used in professional practice was verified empirically for land properties that were valued using the Szczecin Mass Real Estate Valuation Algorithm (SAMWN). The specificity of mass valuation is that the value of many properties is estimated at the same time using a uniform approach - preferably an objective calculation algorithm. The mass valuation procedure should be organized in such a way as to minimize the contribution of the human factor. Estimating the value of many properties in an objective manner required many steps to be taken (Hozer, et al., 1999):

- determining the collection of real estate for valuation,
- specifying the market features (attributes) adequately to the type of real estate to be assessed, along with assigning them a numerical impact on the value of the real estate,
- determining location attractiveness zones (SALs),
- randomly selecting a real estate sample for valuation of a representative sample of real estate for individual valuation,
- individual valuation of real estate (by property appraisers) from a representative sample,
- constructing market features databases for the collection of real estate for valuation,
- calculating the value of hypothetical real estate from a representative sample according to the formula:

$$w_{ji}^{h} = pow_i \cdot w_{baz} \cdot \prod_{k=1}^{K} \prod_{p=1}^{k_p} (1 + A_{kpi}) \quad (12)$$

where:
w_{ji}^{h} – the hypothetical value of the i-th property in the j-th area of location attractiveness,
pow_i – the area of the i-th real estate,
w_{baz} – estimated value of 1 m² of the property with the worst attribute states in the worst location attractiveness zone,
A_{kpi} – the impact of the p-th category of the k-th attribute for the i-th property $(k = 1, 2, \dots, K;\ p = 1,2, \dots, k_p)$,
K – the number of attributes,
k_p – the number of categories of the k-th attribute,
N – the number of real estate valued $(i = 1, 2, \dots, N)$,
n_j –the number of representative properties in the j-th zone of location attractiveness,

- calculating market value coefficients for location attractiveness zones according to the formula:

$$wwr_j = \sqrt[n_j]{\prod_{i=1}^{n_j} \frac{w_{ji}^{rz}}{w_{ji}^{h}}} \quad (13)$$

where:
wwr_j – market value coefficient in the j-th zone of location attractiveness $(j = 1, 2, \dots, J)$,
J – the number of attractiveness zones for the location,
w_{ji}^{rz} – the value of the i–th property in the j-th zone of location attractiveness determined by the property appraiser,

– calculating the market values of valued properties, according to the formula:

$$w_{ji} = wwr_j \cdot pow_i \cdot w_{baz} \cdot \prod_{k=1}^{K} \prod_{p=1}^{k_p} (1 + A_{kpi}) \tag{14}$$

where:

w_{ji} – market value (or cadastral value) of the *i*-th property in the *j*-th zone of location attractiveness.

When estimating the value of real estate based on the Formula (14), its market or cadastral value is determined. The algorithm is multiplicative and the reference point is the value of the underlying property, which was defined as the property with the worst attribute states, in the worst SAL. According to the formula, the base value is multiplied by the area of the property, market value coefficient and the effect of attribute states estimated on the basis of the approach proposed above. Detailed issues related to the construction and use of SAMWN are described in *System kalibracji macierzy wpływu atrybutów* (*The system for calibrating the matrices of attribute influence*) … (2020).

The proposed calculation procedure was finally carried out 168 times. The compliance of the estimated valuations with unit valuations of property appraisers was checked using the matching measures: the root mean square error RMSE and the mean absolute percentage error MAPE. Because, in business practice, it is important not only to obtain the smallest error for all observations together, but also to minimize unit deviations of *in plus* and *in minus* valuations, additional measures were proposed: the share of valuations above B^+ and below B^- the checked value. The following formulas were used:

– *RMSE* root square error:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (w_i - w_{t,i})^2}{n}} \tag{15}$$

where:

w_i – real unit value of real estate determined by a real estate appraiser,

$w_{t,i}$ – theoretical, unit value of the property determined by SAMWN,

n – the number of observations,

– mean absolute *MAPE* percentage error:

$$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \frac{|w_i - w_{t,i}|}{w_i} \tag{16}$$

– based on the percentage error of PE:

$$PE_i = \frac{w_i - w_{t,i}}{w_i} \cdot 100\% \tag{17}$$

– the share of valuations deviating *in plus* were calculated $PE_i > 0$:

$$B^+ = \frac{\sum_{i=1}^{n^+} PE_i^+}{n^+} \cdot 100\% \tag{18}$$

– and the share of deviations *in minus* for which $PE_i < 0$:

$$B^- = \frac{\sum_{i=1}^{n^-} PE_i^-}{n^-} \cdot 100\% \tag{19}$$

where:

n^+ – the number of observations for which $PE_i > 0$,

n^- – the number of observations for which $PE_i < 0$.

4. Empirical study

Empirical analysis was carried out for properties located in the northern part of Szczecin. This area was divided into 17 SAL (Hozer, et al., 2019) – areas where real estate was characterized by the same or similar impact of location on their value (Fig. 1).

Of the identified properties, 405 housing properties were drawn at random. These were properties that were individually valued by property appraisers. The set in question constituted the basis for which the calculations were carried out. It was divided into two parts: 45 properties were drawn (in a

simple draw) - creating a test sample, and the remaining part of the set was treated as a learning sample - the dependence coefficients and market value coefficients were calculated on its basis (see Fig. 2). In the next step, based on the results obtained from the analysis of data from the teaching sample, real estate from the testing sample was measured using SAMWN and the measures of matching the valuations obtained with individual valuations of property appraisers were calculated.

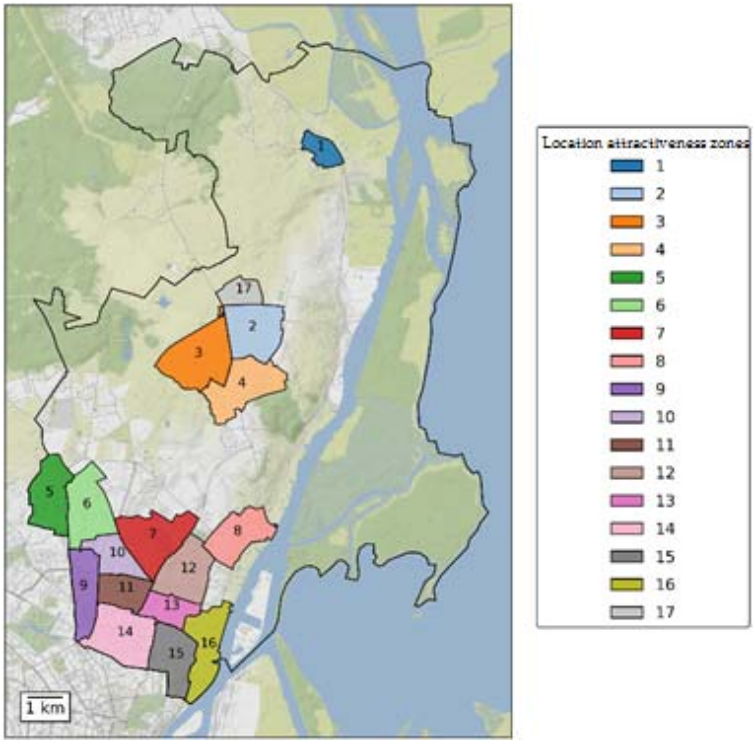

Fig. 1. Location attractiveness zones designated for the valuation. *Source*: own study.

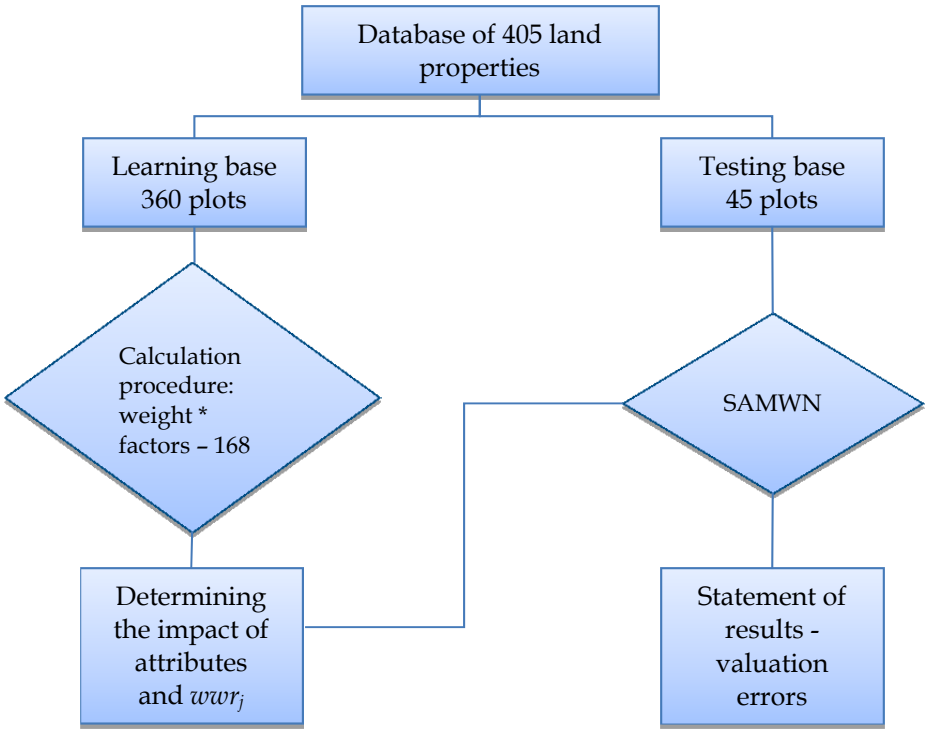

Fig. 2. Diagram of the procedure in the proposed statistical procedure to consider the impact of attribute attributes on the value of real estate. *Source*: own study.

Real estate was described by the following set of attributes: area, utilities, transport accessibility, surroundings and physical features. All attributes have been coded on an ordinal scale, with categories from 0 (the least desirable and the least valued state) to 2 or 3 (the most desirable state):

- x_1 – area: 0 - large, 1 - medium, 2 - small;
- x_2 – technical infrastructure: 0 - none, 1 - incomplete, 2 - full;
- x_3 – access: 0 - unfavorable, 1 - average, 2 - good;
- x_4 – neighborhood: 0 - troublesome, 1 - unfavorable, 2 - average, 3 - favorable;
- x_5 – physical properties: 0 - unfavorable, 1 - average, 2 - favorable.

Figure 3 shows the distribution of the analyzed attributes.

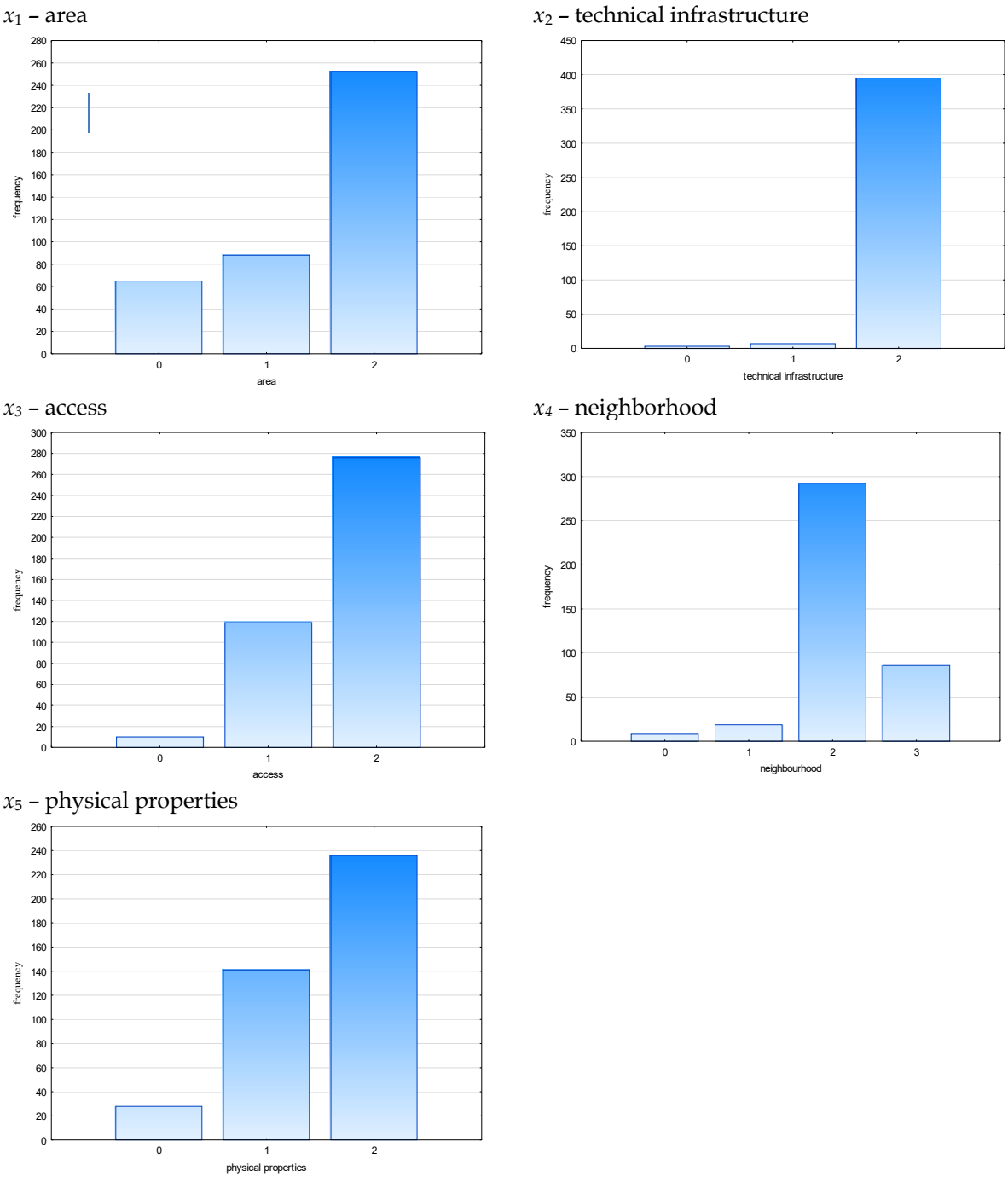

Fig. 3. Distribution of individual property attributes. *Source*: own study.

All distributions of the analyzed attributes were characterized by extremely left-hand asymmetry - the most favorable attribute states dominated (the exception was the neighborhood whose distribution was strongly left-hand asymmetrical, i.e. properties with the desired state of this attribute prevailed). The smallest differentiation of individual attribute states was observed for the following items: technical infrastructure and neighborhood. Because the analysis concerned residential real estate located in a part of Szczecin which is well developed, individual real estate properties were characterized, in the majority of cases, by fully developed technical infrastructure. Hence, only isolated cases of real estate with no or incomplete utilities were observed. Also, the neighborhood of the properties in question was described as favorable rather than unfavorable, although an "average" state prevailed. For individual properties, the level of the neighborhood attribute was defined as "onerous" or "unfavorable".

When preparing the real estate learning database, ensure that all states of each attribute are represented. If one of the states is missing, the effect of this attribute level on the property value cannot be calculated.

The relationship factors (Formulas (1) - (5)) between the value of 1 m² of the property and the defined attributes were calculated. The results are summarized in Table 2.

Table 2

Relationship coefficients between the value of 1 m² and individual attributes

Dependency coefficients	Attributes				
	area	technical infrastructure	access	neighborhood	physical properties
ρ Spearman's	-0.001	0.098	**0.524**	**0.297**	0.042
Partial ρ Spearman's	0.058	**0.143**	**0.534**	**0.304**	0.051
τ_B Kendall's	-0.001	**0.081**	**0.434**	**0.244**	0.033
Partial τ_B Kendall's	0.043	**0.109**	**0.435**	**0.236**	0.039
statistic gamma G	-0.001	**0.407**	**0.640**	**0.370**	0.046
r Pearson's	-0.012	**0.105**	**0.415**	**0.217**	0.056
Partial r Pearson's	0.029	**0.142**	**0.421**	**0.219**	0.060

Significant coefficients at the significance level 0.05 were determined in bold.

Source: own calculation.

For most coefficients (with the exception of ρ Spearman's), the relationship between the value of 1 m² and three attributes was significant: technical infrastructure, access, and neighborhood (Table 2). Both the area and physical properties of the relationship were statistically insignificant. The attributes having the greatest impact on the value of 1 m² were: accessibility followed by neighborhood, with the weakest (though still significant) exhibited by technical infrastructure. Such a system of relationships between the studied variables occurred for each type of coefficient

The different strength of dependence observed for the various coefficients results from the slightly different construction of individual measures. The strongest relationships were recorded for statistic gamma G, and the weakest for r Pearson's (except for the physical properties attribute – the lowest relationship value was for τ_B Kendall's). The differences in the values of the coefficients for individual attributes (the weakest dependencies were indicated by Pearson's coefficients) result from the special conditions that must be met in order for this coefficient to be used (e.g. straight-line dependence).

The most surprising was the insignificant relationship between the surface and the value of 1 m². Analyses carried out in other sub-markets show that these two features are correlated with each other - the unit price (or value) decreases along with a larger area of the property (Foryś & Gdakowicz, 2004). Because the analyzed features were coded in such a way that the lowest level corresponds to the lowest-valued level of the feature, and the highest – to the best valued, a positive, significant relationship was expected. It turned out, however, that the relationship is almost non-existent and has a negative sign. Such values of the area versus 1 m² dependence ratios result from the specificity of the

data analyzed – large properties were valued higher, whereas small ones lower (per 1 m²). This is undoubtedly related to the type of investing entity; these tended to be persons (satisfying their housing needs) for smaller ones, and business entities (developers), possessing greater financial resources and buying real estate for investment purposes.

The relationships between the value of 1 m² and the area improved in the case of partial coefficients - although they were still at a very low level (and not statistically significant), but the direction of the relationship was consistent with that observed in other markets. The use of partial coefficients strengthened the relationships for all coefficients

Based on the measures of the coefficients presented in Table 2, and using 6 ways of taking into account the weights (Formulas (6) - (11)), the influence of individual attributes on the value of the property was determined by calculating the hypothetical values of the property according to formula (12) to finally determine the value coefficients market wwr_j (Formula (13)). The calculated measures were then used to estimate the value of the property from the test base (45 properties) using SAMWN (formula (14)). For each of the 6 ways of taking weights into account, the matching measures: *RMSE*, *MAPE*, B^+ and B^- are summarized. The parameters of the distribution of measures obtained for individual weighing methods are shown in Figure 4.

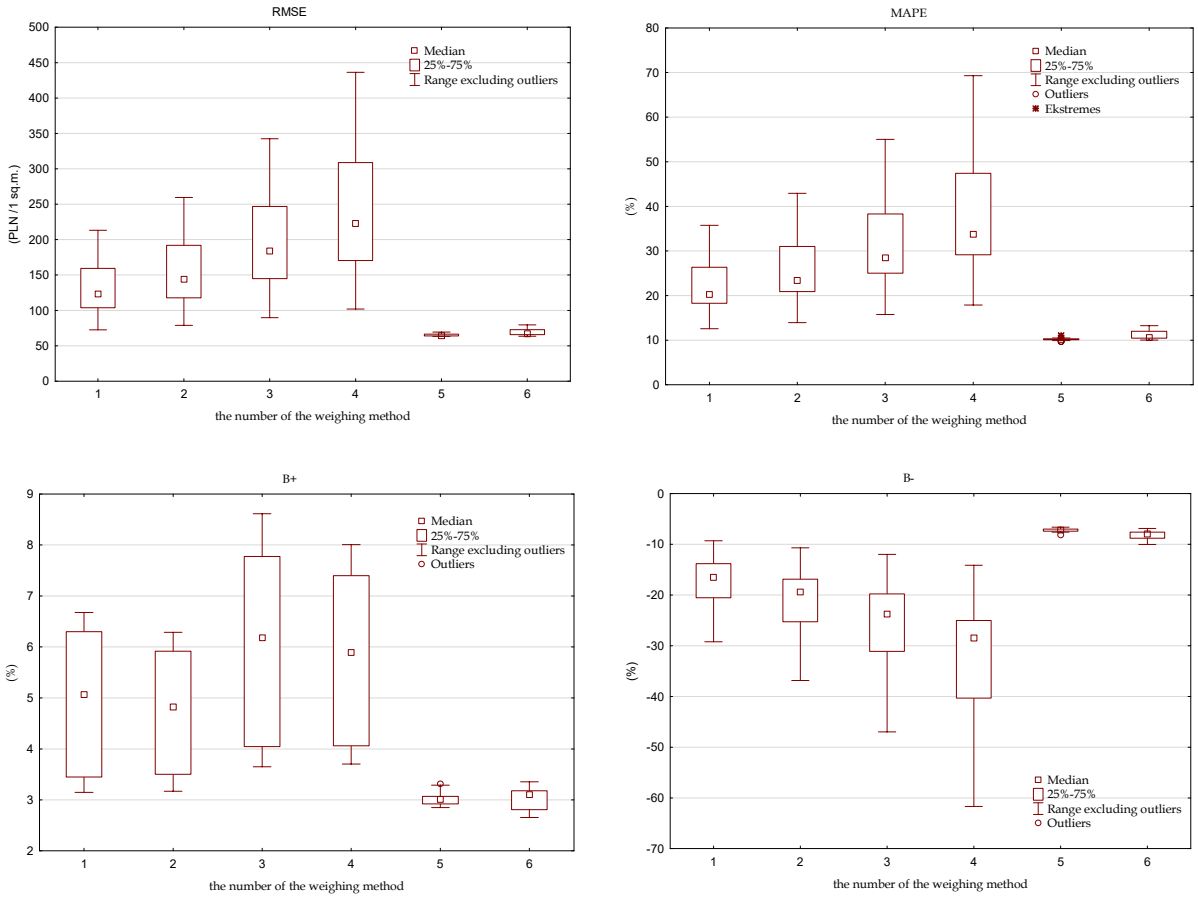

Fig. 4. Distribution parameters of matching values of real estate values estimated by real estate appraisers and real estate values calculated by the SAMWN algorithm according to weighting methods. *Source*: own study.

Depending on the method used to take into account weights, different results of fit measurements were obtained (Fig. 4). The smallest differentiation, for all measures, was observed for the 5th and 6th weighing methods (in this case, variation coefficients were used), with the largest for the 4th. The variation scheme of individual weights for *RMSE*, *MAPE* and B^- measures was very similar. The smallest values of positional measures (median, quartiles, minimum and maximum values) were recorded in the 5th and 6th weighing methods. The next values of positional measures – significantly higher – were observed for method 1, and then for 2, 3 and 4. In the case of the B^+ measure, the

smallest values of the coefficients were also observed for methods 5 and 6, taking into account weights, though the order of the following methods is as follows: 2, 1, 4 and 3.

The *RMSE* measure indicates how much the unit values of real estate valued by property appraisers differ from the values estimated by SAMWN. The differences were significant. For example, the median of the *RMSE* error for method 5 was PLN 64/1 m², while for method 4 – PLN 223/1 m² (almost 4 times more). The lowest values of this measure were recorded for the 5th method – 63.29 PLN /1 m² (for the $G_{i2.6}$ variant), while the highest for the 4th method - 436.43 PLN /1 m² (the $S_{i.4}$ variant).

The *MAPE* ratio indicates by what percent the value of property appraisers' valuations differ from those using SAMWN. The minimum median *MAPE* error was 10.2% (5th weighing method) and the maximum median 33.7% (4th weighing method). The other two measures (B^+ and B^-) indicate the percentage by which the variant used overestimates or underestimates the unit values of real estate. Larger deviations were observed for the B^- measure, which indicates SAMWN's tendencies to underestimate the valued properties. The deviations in plus of property values reach a maximum of 8.6% (for $G_{i.3}$), while undervaluation can be as high as 61.7% in an extreme case (for $S_{i.4}$).

The choice of the best way to determine the impact of attributes on the value of the property (starting from the selection of the dependency coefficient, its inclusion in further calculations, up to the selection of the weighting method) in SAMWN should take into account all 4 matching measures (Formulas (15) - (16) and (18) - (19)). For this purpose, the linear ordering method was used (Walesiak, 2016). It was found that it was desirable for all fitting measures to be close to 0. *RMSE, MAPE* and B^+ coefficients were determined as destimulants (lower value desired), while *B-* as a stimulant (the higher the value, the better). A linear ordering procedure[3] was carried out for each of the ways of considering weights and for all variants together. Table 3 lists the three best and three worst variants of considering the impact of attributes.

Table 3

The best and worst variants of considering the impact of attributes on the value of real estate by weighing methods

Position in the classification	Variant of weight impact					
	1	2	3	4	5	6
1	$G_{w2.1}$	$G_{w2.2}$	$G_{w2.3}$	$G_{w2.4}$	$G_{i.5}$	$S_{ci2.6}$
2	$G_{i2.1}$	$G_{i2.2}$	$G_{i2.3}$	$G_{i2.4}$	$G_{w.5}$	$r_{w2.6}$
3	$r_{cw2.1}$	$r_{cw2.2}$	$r_{cw2.3}$	$r_{cw2.4}$	$r_{cw.5}$	$S_{w2.6}$
…	…	…	…	…	…	…
26	$T_{w.1}$	$T_{w.2}$	$T_{w.3}$	$T_{w.4}$	$r_{i.5}$	$r_{i.6}$
27	$T_{i.1}$	$T_{i.2}$	$T_{i.3}$	$T_{i.4}$	$T_{i.5}$	$T_{i.6}$
28	$S_{i.1}$	$S_{i.2}$	$S_{i.3}$	$S_{i.4}$	$S_{i.5}$	$S_{i.6}$

Source: own calculation.

Depending on how the weights are taken into account, different variants of the coefficients for determining the effect of attributes on the value of the property proved to be the best (Table 3). In the six linear ordering procedures, the variant whose application gave the best results was the variant based on statistic gamma *G*, taking into account the squares of all or significant relationships between attributes and value - G_{w2} or G_{i2} (for weighing methods 1 to 4). In the method of considering weights, calculations based on statistic gamma *G* had the best results (but in further calculations, significant or all coefficients were taken into account - not their squares). Different results were obtained in the last way of considering weights - 6. The highest positions in the linear ordering classification were occupied by variants based on the Spearman coefficient (partial and square of significant dependencies or square of all dependencies). In all methods of considering weights, satisfactory results were obtained using the Pearson partial correlation coefficient (for the square of all relationships).

The procedure that gave by far the worst results (regardless of the method of weighing) was the procedure based on the use of the Spearman coefficient and significant relationships (in this case only

[3] Due to the large variation in variables, fit measures have been standardized. The linear ordering procedure was carried out for the Euclidean distance, with the upper development pattern.

two relationships were statistically significant) and the Kendall's coefficient and significant or all relationships. The weakest results also included those for which Pearson's coefficient and significant dependencies were used.

In the last step of the analysis, all results obtained from the 168 valuations performed were ranked (also using the linear ordering method) (Figure 5). Due to the large number of cases, every other label is displayed on the x axis.

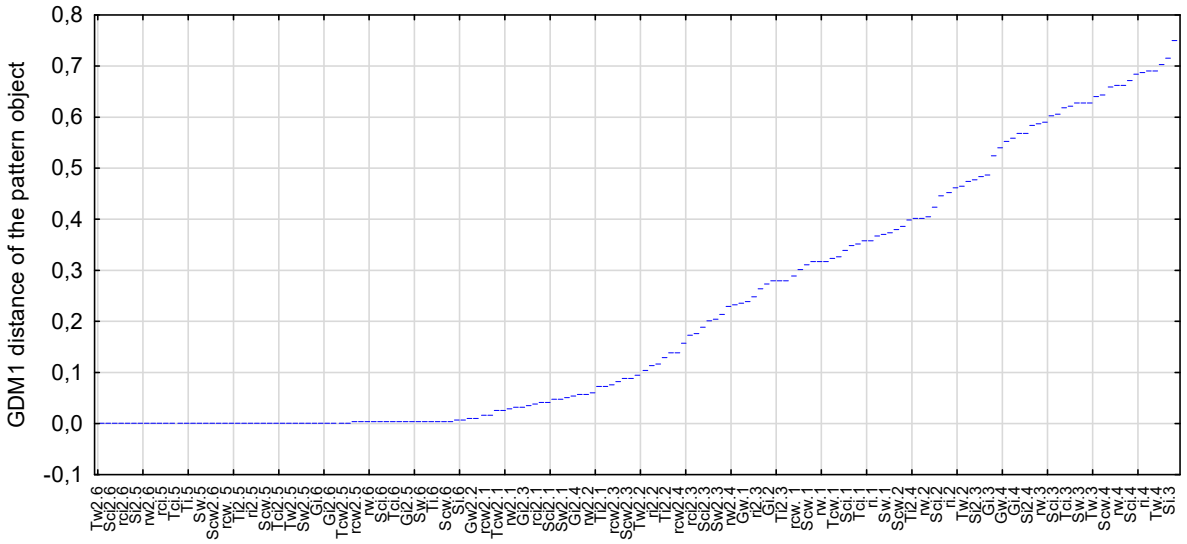

Fig. 5. The ordering of the variants used to take into account the impact of attributes on the value of the property and weighting methods from best to worst due to valuation errors received using SAMWN. *Source*: own calculation.

Sorting from the smallest to the largest aggregate value calculated based on valuation errors allows the best real estate valuation procedures to be determined (Figure 5). The best results were obtained for procedures using the 6th and 5th methods of considering weights. Both of these methods were based on the coefficient of variability of the property value. The weakest results were obtained for the 3rd and 4th methods of considering weights - mean logarithms and median.

The element that significantly affected the accuracy of valuations was the method of weighting. The method of including the coefficients in further calculations was also important - better results were obtained using the square of dependencies, although it was not important whether all coefficients or only statistically significant ones were taken into account. The use of a specific dependency coefficient and the way it was taken into account in further calculations (significant-all) was less important.

5. Conclusions

Statistical methods can be used in the real estate valuation process. Based on the relationship between the attributes of the property and their price (or value), one can determine the effect of individual attributes on the value. However, the use of these statistical measures must be preceded by an analysis of the types of characteristics and the nature of the relationships between variables.

To sum up the conducted research, it should be noted on what scale the property features (attributes) are coded. The ordinal scale will indicate measures based on rank or probability, the quotient scale - on Pearson's ratio. The nature of the dependence also affects the choice of the coefficient - the rectilinear relationship can be described by all the measures proposed, whereas in the curvilinear relationship - Pearson's factor should be excluded from the analyses. Due to the collinearity of real estate features, partial coefficients may be used in the analyses. The study showed that partial factors did not give the best results (although also not the worst). The best results were obtained in variants in which Spearman's or Kendall's dependence coefficients were calculated and their squares were taken into account in further calculations, regardless of whether all or only significant dependencies were taken into account.

The use of one of six ways of accounting for weights in further calculations was the most important for the obtained results was. The smallest differences in matching measures and the best results were

obtained for methods based on the coefficient of variation (5 and 6 weighting methods). The worst results were recorded for mean and median logarithms (weighing methods 3 and 4).

In addition, it was noted that the analysis of all 168 variants combined gave different results than the analysis of each of the methods of taking weights into account. In the latter case, the best results were recorded for the gamma *G* statistic (square of all or significant coefficients), whereas the worst for Spearman's and Kendall's coefficients (significant or all coefficients).

6. References

Armstrong, R. A. (2019). Should Pearson's correlation coefficient be avoided? *Ophthalmic and Physiological Optics I, 39*, 316–327. https://doi.org/10.1111/opo.12636 PMID:31423624

Babatunde, I. O. (2018). Examining Heuristics for Building – Work-In-Progress Valuations in Niger State Nigeria. *Real Estate Management and Valuation, 26*(2), 92–103. https://doi.org/10.2478/remav-2018-0019

Barańska, A. M. (2019a). Correlation Analysis in the Process Of Weighting Real Property Attributes. *Real Estate Management and Valuation, 27*(4), 74–84. https://doi.org/10.2478/remav-2019-0037

Barańska, A. M. (2019b). Linear and Nonlinear Weighing of Property Features. *Real Estate Management and Valuation, 27*(1), 59–68. https://doi.org/10.2478/remav-2019-0006

Dmytrów, K., Gdakowicz, A., & Putek-Szeląg, E. (2019). Statistical Relations of the Qualitative Attributes of Real Properties Subject oo Mass Appraisal. *Folia Oeconomica Stetinensia, 19*(2), 25–37. https://doi.org/10.2478/foli-2019-0011

Doszyń, M. (2017). Statistical Determination of Impact of Property Attributes for Weak Measurement Scales. *Real Estate Management and Valuation, 25*(4), 75–84. https://doi.org/10.1515/remav-2017-0031

Doszyń, M. (2019). Individual Capacities of Hellwig's Information Carriers and the Impact of Attributes in the Szczecin Algorithm of Real Estate Mass Appraisal. *Real Estate Management and Valuation, 27*(1), 15–24. https://doi.org/10.2478/remav-2019-0002

Gaca, R., & Sawiłow, E. (2014). Zastosowanie współczynników korelacji rang Spearmana do ustalenia wag cech rynkowych nieruchomości, Rzeczoznawca Majątkowy, nr 82, pp. 24-30 (Application of Spearman's rank correlation coefficient for establishing ranks of real estate characteristics. *Real Estate Appraiser,* (82), 24–30.

Goodman, L. A., & Kruskal, W. H. (1963). Measures of association for cross classification. III: Approximate sampling theory. *Journal of the American Statistical Association, 58*, 310–364. https://doi.org/10.1080/01621459.1963.10500850

Gaca, R. (2018). Parametric and Non-Parametric Statistical Methods in the Assessment of the Effect of Property Attributes on Prices. *Real Estate Management and Valuation, 26*(2), 83–91. https://doi.org/10.2478/remav-2018-0018

Foryś, I., & Gaca, R. (2016). Application of the Likert and Osgood Scales to Quantify the Qualitative Features of Real Estate Properties. *Folia Oeconomica Stetinensia, 16*(2), 7–16. https://doi.org/10.1515/foli-2016-0021

Foryś, I., & Gdakowicz, A. (2004). Wykorzystanie metod ilościowych do badania rynku nieruchomości (The application of quantitative methods for real estate market analysis). *Studia i Materiały Towarzystwa Naukowego Nieruchomości, 12*(1).

Hozer, J., Foryś, I., Zwolankowska, M., Kokot, S., & Kuźmiński, W. (1999). Ekonometryczny algorytm masowej wyceny nieruchomości gruntowych (An Econometric Algorithm for Land Mass Appraisal). Katedra Ekonometrii i Statystyki Uniwersytetu Szczecińskiego, Stowarzyszenie Pomoc i Rozwój.

Hozer, J., Gnat, S., Kokot, S., & Kuźmiński, W. (2019). The Problem of Designating Elementary Terrains for the Purpose of Szczecin Algorithm of Real Estate Mass Appraisal. *Real Estate Management and Valuation, 27*(3), 42–58. https://doi.org/10.2478/remav-2019-0024

Kendall, M. G. (1948). Rank Correlation Methods. Charles Griffin & Company Limited.

Parker, R. I., Vannest, K. J., Davis, J. L., & Sauber, S. B. (2011). Combining nonoverlap and trend for single-case research: Tau-U. *Behavior Therapy, 42*(2), 284–299. https://doi.org/10.1016/j.beth.2010.08.006 PMID:21496513

Powszechne Krajowe Zasady Wyceny. (2008). Nota Interpretacyjna, Zastosowanie podejścia porównawczego w wycenie nieruchomości (Common National Valuation Rules, Interpretative Note, Application of the Comparative Approach to Property Valuation), Polska Federacja

Stowarzyszeń Rzeczoznawców Majątkowych. http://wycena.net.pl/standardy/NI-Zastosowanie.podejscia.porownawczego.pdf.

Siegel, S., & Castellan, N. J., Jr. (1988). Nonparametrics statistics for the behavioral sciences (2nd ed.). McGraw-Hill.

Doszyń, M. (Ed.). (2020). System kalibracji macierzy wpływu atrybutów w szczecińskim algorytmie masowej wyceny nieruchomości (Calibration system of attribute impact matrix in Szczecin algorithm of real estate mass appraisal). Wydawnictwo Naukowe Uniwersytetu Szczecińskiego.

Walesiak, M. (2016). Uogólniona miara odległości GDM w statystycznej analizie wielowymiarowej z wykorzystaniem programu R. Wydanie drugie poprawione i rozszerzone (A General Dissimilarity Measure in Multi-variate Statistical Analysis with Application of R Program. Second revised and extended edition). Wydawnictwo Uniwersytetu Ekonomicznego.