GEOMATICS AND ENVIRONMENTAL ENGINEERING • Volume 8 • Number 1 • 2014

http://dx.doi.org/10.7494/geom.2014.8.1.33

Marek Kulczycki*, Marcin Ligas*

# Qualitative Similarity Coefficients in Real Estate Market Analysis**

## 1. Introduction

Often, for the purposes of real estate market analysis all real estate characteristics are treated equally as quantitative or ordinal ones, the latter after normalization also treated as quantitative. It is worth recalling that a quantitative attribute is the one which takes numerical values. The area expressed in square meters is a typical quantitative attribute.

It is difficult, however, not to agree that the area seen in this way, particularly for properties differing slightly in this respect, will not be directly relevant to the assessment of similarity, and moreover while assessing the difference in their mutual attractiveness, what in turn causes a lack of effect on their unit values.

Ordinal attributes are for instance those ranked like below:

- "attractive",
- "not attractive",
- "very attractive",
- "average".

In this case we can indicate without any problem which value is the best, which is the worst and e.g. that "average" is better than "not attractive" and worse than "attractive". On the other hand, we are not able to define either the difference or the fraction between consecutive values of the scale. Therefore, this type of attribute requires normalization in this way that after sorting the elements in ascending order (increase of attractiveness) we number them (we set the position on the scale) in sequence from 1 to n. Then each element of the scale is assigned a normalized

* AGH University of Science and Technology, Faculty of Mining Surveying and Environmental Engineering, Department of Geomatics, Kraków, Poland

** This paper is the result of research carried out within statutory research grant no. 11.11.150.006 in the Department of Geomatics, AGH University of Science and Technology, Krakow.

numerical value by dividing each element by the number of all elements wherein the nominator and denominator are decreased by one i.e. (*position* – 1)/(*n* – 1), and in the analyzed case taking the following values:

- "not attractive" – 0.00,
- "average" – $0.33_{(3)}$,
- "attractive" – $0.66_{(6)}$,
- "very attractive" – 1.00.

For objects described by means of vectors of quantitative attributes the literature provides certain measures of similarity: Euclidean distance coefficient in different versions (weighted or averaged), Canberra metric coefficient or cosine coefficient.

The purpose of this paper is among other things to present a procedure for assessing the similarity of properties when they are described by means of features which are not treated as quantitative attributes but as qualitative ones – binary or nominal. Such a case does not seem senseless for the real estate market and there are some who indicate the necessity of considering the characteristics of the property just as qualitative attributes, for the sake of subjectivity accompanying describing the property that causes the characteristics are only seemingly of ordinal nature.

## 2. Qualitative Similarity Coefficients

Feature of real estate being a qualitative attribute is the one that takes values from a limited catalog of possible values. If there are more than two possible values we deal with a nominal attribute. An example of such a feature may be "location" understood as the name of a district for the market territorially restricted to the administrative boundaries of the city or the name of the municipality for the regional market. Another, even more appropriate example could be a feature "basement" relating to the premises belonging to the dwelling. In this case we are dealing with the so-called binary attribute. This attribute can take only two mutually excluding values: Yes – there is a basement, No – there is no basement. Comparing two properties ($N_1$ and $N_2$) described by means of one feature ($C_1$) which is a binary attribute, we will be dealing with one of the four possible cases (similarity variants "a", "b", "c" and "d") (Tab. 1).

**Table 1.** Similarity variants of two properties described by one feature of binary nature

| Feature | $N_1$ | $N_2$ | Variant |
|---|---|---|---|
| $C_1$ | Yes | Yes | a |
| | Yes | No | b |
| | No | Yes | c |
| | No | No | d |

If properties being compared are described by more than one feature i.e. ($C_1$, $C_2$,..., $C_n$) all being binary attributes then in order to asses their similarity it will be necessary to count the number of occurrences of appropriate variants beforehand. And thus for instance for two properties ($N_1$ and $N_2$) described by means of six characteristics (form $C_1$ to $C_6$) (Tab. 2), we will obtain a similarity matrix (Tab. 3).

**Table 2.** Exemplary values of characteristics of properties being compared

| Feature | $N_1$ | $N_2$ | Variant |
|---|---|---|---|
| $C_1$ | Yes | Yes | a |
| $C_2$ | Yes | No | b |
| $C_3$ | Yes | Yes | a |
| $C_4$ | No | Yes | c |
| $C_5$ | No | No | d |
| $C_6$ | Yes | No | b |

**Table 3.** Exemplary similarity matrix

| | | $N_2$ | |
|---|---|---|---|
| | | Yes | No |
| $N_1$ | Yes | 2 | 2 |
| | No | 1 | 1 |

Basing on thus constructed similarity matrix it is possible to compute many similarity coefficients known from the literature. In this study similarity coefficients that are not a function of "d" variants were omitted because there is a consistent view in the literature that they do not give satisfactory results for nominal attributes and as one knows the majority of property characteristics have this particular character.

In the subsequent part the following qualitative similarity coefficients (*QRC*) will be computed:

– Russel and Rao Coefficient:

$$QRC = \frac{a}{a+b+c+d} \tag{1}$$

– Simple Matching Coefficient:

$$QRC = \frac{a+d}{a+b+c+d} \tag{2}$$

– Sokal and Sneath Coefficient:

$$QRC = \frac{2\cdot(a+d)}{2\cdot(a+d)+b+c} \tag{3}$$

– Rogers and Tanimoto Coefficient:

$$QRC = \frac{a+d}{a+2\cdot(b+c)+d} \tag{4}$$

– Baroni-Urbani and Buser Coefficient:

$$QRC = \frac{a+\sqrt{a\cdot d}}{a+b+c+\sqrt{a\cdot d}} \tag{5}$$

– Sokal Binary Distance Coefficient:

$$QRC = \sqrt{\frac{b+c}{a+b+c+d}} \tag{6}$$

– Yule Coefficient:

$$QRC = \frac{a\cdot d-b\cdot c}{a\cdot d+b\cdot c} \tag{7}$$

– Hamann Coefficient:

$$QRC = \frac{a+d-b-c}{a+b+c+d} \tag{8}$$

– Phi Coefficient:

$$QRC = \frac{a\cdot d-b\cdot c}{\sqrt{(a+b)\cdot(a+c)\cdot(b+d)\cdot(c+d)}} \tag{9}$$

As pointed out above, the computation of qualitative similarity coefficients is based on binary attributes hence a necessary step in assessing the similarity of the properties is to the first transform characteristics of the property from nominal to binary attributes. The base for this transform is the use of binary attributes (so called dummy binary attributes) in place of one nominal attribute in the number corresponding to the unique values of the nominal attribute. And so, for example, for the nominal attribute ($C_1$) "acceptable type of building" taking values:

– "extensive single family detached housing",
– "intensive single family detached housing",
– "single family semidetached housing",
– "single family terraced housing",
– "low intensity multi family housing",

it will be necessary to introduce five dummy binary attributes (from $C_{1.1}$ to $C_{1.5}$) in place of $C_1$.

Comparing two properties ($N_1$ and $N_2$) where for the first one a terraced housing is authorized ($C_{1,1.4}$) and for the second one only intensive detached housing is acceptable ($C_{2,1.2}$) we will obtain the result of comparison as in Table 4.

**Table 4.** Dummy binary values corresponding to a nominal attribute

| Feature | $N_1$ | $N_2$ | Variant |
|---|---|---|---|
| $C_{1.1}$ | No | No | d |
| $C_{1.2}$ | No | Yes | c |
| $C_{1.3}$ | No | No | d |
| $C_{1.4}$ | Yes | No | b |
| $C_{1.5}$ | No | No | d |

Transformation of a nominal attribute into the corresponding set of dummy binary attributes, especially when considering the similarity of properties, causes the need of asking a few questions:

- Whether and if so, in what way, the similarity is affected by differentiating the number ($n_i$) of unique values of *i*-th nominal attribute which translates onto the differentiation of quantity of "d" variants and at the same time does not translate onto the number of "a", "b" and "c" variants?
- Whether for every market feature ($C_i$) it will be more correct to see it as a nominal attribute rather than an ordinary one and if not, how does it affect the similarity?
- Whether and if so, in what way, the similarity is affected by differentiating the importance of attributes ($k_i$ %), expressed as a percentage summing up to 100%, and representing the impact of individual market characteristics on the explanation of price variablility in the analyzed real estate market?

The answer to these questions is a hint for accepting the appropriate qualitative measure of similarity, for example, by selecting one of the nine coefficients listed above or by a modification of the method of calculating the elements of a similarity matrix.

The authors suggest that such a modification could rely on replacing counting (incrementing) individual variants with increasing their values by $\Delta$, the value of which would be determined by taking into account the number of (and possibly the order) unique values of a nominal attribute and its importance, for example, by the formulas:

$$\Delta_{\mathrm{a}} = \Delta_{\mathrm{b}} = \Delta_{\mathrm{c}} = \Delta_{\mathrm{d}} = k_i\% \tag{10}$$

for all variants without taking into account the order of unique values of the attribute, and:

$$\Delta_{\mathrm{a}} = \Delta_{\mathrm{d}} = k_i\% \quad \mathrm{i} \quad \Delta_{\mathrm{b}} = \Delta_{\mathrm{c}} = \frac{\left|C_{1,i} - C_{2,i}\right|}{n_i} \cdot k_i\% \tag{11}$$

when the order of unique values of the attribute is considered.

## 3. Study of the Usefulness of Qualitative Similarity Coefficients in the Real Estate Market Analysis

An algorithm for computing qualitative similarity coefficients is not very complicated. Implementation of the algorithm, however, consists of a large number of operations that require concentration and time hence the authors coded the algorithm in a programming language.

While examining the usefulness of the coefficients in the market analysis at the first step it was assumed that properties are described by four characteristics which may take different numbers of unique values, this is shown in Table 5.

**Table 5.** Features and range of their values

| Feature | Possibile values | | | | |
|---|---|---|---|---|---|
| $C_1$ | c1_A | c1_B | | | |
| $C_2$ | c2_A | c2_B | c2_C | | |
| $C_3$ | c3_A | c3_B | c3_C | c3_D | |
| $C_4$ | c4_A | c4_B | c4_C | c4_D | c4_E |

In such a catalog of characteristics we are dealing with 120 properties with the unique arrangement of individual characteristics. For all these properties coefficients of similarity (given in the section 3) to the properties with the following characteristics have been computed:

- c1_A, c2_A, c3_A, c4_A,
- c1_B, c2_B, c3_C, c4_C,
- c1_B, c2_C, c3_D, c4_E.

The computations confirmed that the use of any of the coefficients given in the section 3 gives in every case:

- the same ranking of 120 properties (form the most to the least similar),
- the same grouping of 120 properties into properties identical with the one being compared (group 1) and also those properties differing from the one being compared with: one feature (group 2), two features (group 3), three features (group 4) and four features (group 5).

Table 6 presents values of the similarity coefficients for some selected properties.

It is worth noting that:

- the value of the coefficient $QRC_6$ is increasing while the values of the remaining ones are decreasing, it is understandable when we realize that this particular coefficient is a measure of distance thus the greater the distance the less the similarity;
- coefficients $QRC_7$, $QRC_8$, $QRC_9$ took negative values because these three coefficients may take values from –1 to 1 while the others take values from 0 to 1;

– properties numbered 100 and 115 are equally similar to the subject property what confirms that applied methods do not differentiate the importance of attributes;
– properties numbered 94 and 61 are equally similar to the subject property what confirms that we are dealing with features seen as nominal attributes.

**Table 6.** Similarity coefficients for selected properties being compared with the property of characteristics c1_B, c2_C, c3_D, c4_E

| Property (characteristics) | Position in ranking | Values of similarity coefficients | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $QRC_1$ | $QRC_2$ | $QRC_3$ | $QRC_4$ | $QRC_5$ | $QRC_6$ | $QRC_7$ | $QRC_8$ | $QRC_9$ |
| 120 (B C D E) | 1 | 0.29 | 1.00 | 1.00 | 1.00 | 1.00 | 0.00 | 1.00 | 1.00 | 1.00 |
| 100 (B B D E) | 2 | 0.21 | 0.86 | 0.92 | 0.75 | 0.80 | 0.38 | 0.93 | 0.71 | 0.65 |
| 115 (B C C E) | 2 | 0.21 | 0.86 | 0.92 | 0.75 | 0.80 | 0.38 | 0.93 | 0.71 | 0.65 |
| 102 (B C A B) | 3 | 0.14 | 0.71 | 0.83 | 0.56 | 0.60 | 0.53 | 0.60 | 0.43 | 0.30 |
| 94 (B B C D) | 4 | 0.07 | 0.57 | 0.73 | 0.40 | 0.38 | 0.65 | –0.12 | 0.14 | –0.05 |
| 61 (B A A A) | 4 | 0.07 | 0.57 | 0.73 | 0.40 | 0.38 | 0.65 | –0.12 | 0.14 | –0.05 |
| 1 (A A A A) | 5 | 0.00 | 0.43 | 0.60 | 0.27 | 0.00 | 0.76 | –1.00 | –0.14 | –0.40 |

Similar computations (obtaining parallel results) were conducted for different sets of property characteristics e.g. for five features from which everyone may get one of the five unique values.

The next analysis concerned the case with differentiating the significance of individual characteristics. This time properties were described by four features defined like in Table 5, but now the feature number four was recognized as the most significant and was assigned $k_4\% = 40\%$, whilst for the remaining features there were $k_1\% = 10\%$, $k_2\% = k_3\% = 25\%$.

The obtained result is consistent with common sense. All the similarity coefficients still allow for the identical ranking and grouping of properties. This time 11 groups of properties with identical values of the similarity coefficients were obtained. Table 7 presents the results for some selected properties.

It is worth noting that:

– feature $C_4$ is more important than the feature $C_3$ hence properties identical with the subject property with respect to the feature $C_4$ obtain always not less similarity with the subject property than properties identical with respect to the feature $C_3$; this happens of course under the condition that in both cases values of the features $C_1$ and $C_2$ are alike equal or different from values of these characteristics of the subject property; this may be observed in Table 7 for properties numbered 110 and 119 as well as 65 and 98;

– properties that differ from the subject property with the same characteristics, never mind what values of these characteristics they get, still obtain identical values of similarity coefficients what is seen in Table 7 for properties numbered 76 and 98;
– features $C_2$ and $C_3$ are equally important hence the fact that the evaluated property differs from the subject property with respect to the second or the third feature has no impact on the mutual similarity; this happens of course under the condition that values of the features $C_1$ and $C_4$ will be alike equal or different from values of these characteristics of the property being compared; this may be observed in Table 7 for properties numbered 20 and 55 as well as 16 and 43.

**Table 7.** Similarity coefficients that take into account the importance of attributes for some selected properties being compared with the property of characteristics c1_B, c2_C, c3_D, c4_E

| Property (characteristics) | Position in ranking | Values of similarity coefficients | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $QRC_1$ | $QRC_2$ | $QRC_3$ | $QRC_4$ | $QRC_5$ | $QRC_6$ | $QRC_7$ | $QRC_8$ | $QRC_9$ |
| 120 (B C D E) | 1 | 0.25 | 1.00 | 1.00 | 1.00 | 1.00 | 0.00 | 1.00 | 1.00 | 1.00 |
| 60 (A C D E) | 2 | 0.23 | 0.95 | 0.97 | 0.90 | 0.93 | 0.23 | 0.99 | 0.90 | 0.87 |
| 110 (B C B E) | 3 | 0.19 | 0.87 | 0.93 | 0.78 | 0.81 | 0.36 | 0.94 | 0.75 | 0.67 |
| 20 (A A D E) | 4 | 0.16 | 0.82 | 0.90 | 0.70 | 0.74 | 0.42 | 0.86 | 0.65 | 0.53 |
| 55 (A C C E) | 4 | 0.16 | 0.82 | 0.90 | 0.70 | 0.74 | 0.42 | 0.86 | 0.65 | 0.53 |
| 119 (B C D D) | 5 | 0.15 | 0.80 | 0.89 | 0.66 | 0.70 | 0.45 | 0.81 | 0.59 | 0.46 |
| 65 (B A A E) | 6 | 0.13 | 0.75 | 0.86 | 0.60 | 0.62 | 0.50 | 0.66 | 0.49 | 0.33 |
| 76 (B A D A) | 8 | 0.09 | 0.67 | 0.80 | 0.50 | 0.49 | 0.57 | 0.31 | 0.34 | 0.13 |
| 98 (B B D C) | 8 | 0.09 | 0.67 | 0.80 | 0.50 | 0.49 | 0.57 | 0.31 | 0.34 | 0.13 |
| 16 (A A D A) | 9 | 0.06 | 0.62 | 0.77 | 0.45 | 0.40 | 0.62 | –0.01 | 0.24 | 0.00 |
| 43 (A C A C) | 9 | 0.06 | 0.62 | 0.77 | 0.45 | 0.40 | 0.62 | –0.01 | 0.24 | 0.00 |
| 66 (B A B A) | 10 | 0.03 | 0.54 | 0.70 | 0.37 | 0.23 | 0.68 | –0.60 | 0.09 | –0.21 |
| 1 (A A A A) | 11 | 0.00 | 0.49 | 0.66 | 0.33 | 0.00 | 0.71 | –1.00 | –0.01 | –0.34 |

The last test concerned the case not only differentiating the importance of individual features but also arranging (in alphabetical order) unique values of individual characteristics. This time, not all the similarity coefficients led to the identical ranking and grouping of the properties. But if the analysis is limited to the coefficients $QRC_2$, $QRC_3$, $QRC_4$, $QRC_6$ and $QRC_8$ the compatibility is ideal again. Understandably, this time, 109 real estate groups with identical values of the similarity coefficients have been obtained. As one could guess, the real estate assigned to the same group are those with extremely different features of $C_2$ and $C_3$ but identical with respect to the remaining characteristics.

## 4. Conclusions

The authors recognize both ways of seeing the characteristics of the property mentioned in the introduction as equally acceptable, indicating at the same time that the perception of features of the property as qualitative characteristics may be particularly applicable in a situation where we have no certainty about the scaling (setting an order) of individual characteristics.

The determination of the similarity of properties in this way allows for eliminating the error resulting from the subjectivity of an appraiser when describing the property. No one needs to be persuaded that by calculating the similarity coefficient it is possible to indicate several or a dozen properties quickly and unequivocally. And later on, these properties will be the basis for inference about the value of the property in the sales comparison approach.

Moreover, similarity coefficients can be used in properties' grouping algorithms and for indicating the representative property for each group.

Given the results, presented and discussed partially in the section 3, the authors draw a careful conclusion that the coefficients $QRC_2$, $QRC_3$, $QRC_4$, $QRC_6$ i $QRC_8$ will find the best application in the real estate market analysis.

## References

[1] Czaja J., Kulczycki M.: *Parametry oceny wiarygodności i spójności informacji rynkowych o nieruchomościach*. Geodezja. Półrocznik Akademii Górniczo-Hutniczej im. Stanisława Staszica w Krakowie, t. 8, z. 1, 2002, pp. 137–144.

[2] Jaruga K.: *Statystyczna analiza wielowymiarowa.* Wydawnictwo Naukowe PWN, Warszawa1993.

[3] Kulczycki M., Ligas M.: *Statystyczna metoda wyznaczania nieruchomości reprezentatywnych*. Kraków, Geodezja. Półrocznik Akademii Górniczo-Hutniczej im. Stanisława Staszica w Krakowie, t. 9, z. 2/2, 2003, pp. 631–637.

[4] Romesburg H.C.: *Cluster analysis for researchers*. Lulu Press, North Carolina 2004.

[5] Zyga J.: *Identyfikacja podobieństwa nieruchomości*. Olsztyn, Studia i Materiały Towarzystwa Naukowego Nieruchomości, vol. 19, nr 4, 2011, pp. 141–158.