MARRAKECH, May 18-22, 2011

Qualitative and quantitative methods for assessing the similarity of real estate

PhD Anna Barańska
Department of Geomatics
Faculty of Mining Surveying and Environmental Engineering
University of Science and Technology
Krakow, POLAND

INTRODUCTION

The assessment of the real estate similarity is a problem constantly topical in the everyday work of estate experts and real estate market analysts. One of the essential difficulties here is the selection in a, very numerous often, database containing so called price-making real estate features – these ones that really form its price on a given market. The point is that the assessment of similarity, made on the grounds of such a selection, remained objective, i.e. was reliable. The problem becomes particularly important in the case of mass valuations with a huge database, where it is difficult to pick out the objects similar.

QUALITIVE AND QUANTITIVE VARIABLES

A real estate attributes can be divided generally into obligatory and facultative. Obligatory will be the information on the real estate permitting to identify it explicitly in documents and in site. These are, among the others, address data, number of building plot, number of real estate register, number of registering unit, number and name of the district and the like. Whereas, facultative are the features, which can, potentially, influence the real estate prices. They describe the real estate quality, in broad terms. We distinguish among them so called price-making attributes, really shaping the prices.

Facultative attributes belong usually to the qualitative variables. Most of them answer the question "what kind?" not "how much?". For this reason many methods of similarity assessment was adapted to this qualitative character of variables. The similarity often comes down to the identity of a determined number among all analysed features or it is based on a qualitative comparison of the real estate attributes, aiming only to notice differences, without considering how great they are.

CORRELATION COEFFICIENTS

In order to make objective the similarity assessment procedures, the application of different correlations types is proposed to select from a large database containing the variables determining a real estate attraction - the real price-making attributes:

- Pearson's correlation;
- Spearman's correlation;
- Kendall's correlation;
- Gamma correlation;
- partial correlation;
- nonlinear correlation.

To determine the correlation coefficients, it is necessary to make a preliminary transformation of qualitative features into the quantitative ones by assigning to them definite numerical scales. The scales result from the intensity of the examined feature and they function as ranks.

Pearson's correlation: $$r_p = \frac{\sum_{i=1}^{n}(x_i - \hat{x})(y_i - \hat{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \hat{x})^2 \sum_{i=1}^{n}(y_i - \hat{y})^2}} \quad (1)$$

(x_i, y_i) – values of a two-dimensional random variable,
$\hat{x}, \hat{y}$ – mean values of variables *X* and *Y*,
n – random sample size.

Spearman's correlation: $$r_S = 1 - \frac{6 \cdot \sum_{i=1}^{n}(i - s_i)^2}{n \cdot (n^2 - 1)} \quad (2)$$

s_i – rank assigned to the position i after the pairs (x_i, y_i) are arranged in series in relation to one component, for example x,
n – size of a random sample.

In order to determine r_s, a ranking is made first, i.e. every observed value is replaced with its subsequent number resulting from its item in the database sorted in growing order. Next, the ordinary Pearson's coefficient of linear correlation is calculated. The ranking approaches possible divergent observations to the rest, levelling thus their influence disturbing the result. A monotonic nonlinear relationship is transformed by ranking into a linear one. In consequence, the linear correlation Pearson's coefficient, applied to ranks, measures the nonlinear relation force.

Kendall's correlation: $$r_K = 2 \cdot \frac{z - m}{n(n-1)} \quad (3)$$

z – number of compatible pairs (variables compared within two observations change in the same direction, i.e. they are both larger or smaller in the first observation than in the second),
m – number of incompatible pairs,
n – size of random sample.

Gamma correlation: Gamma statistics is recommended in the cases, when data contain many combined observations (one of variables has equal values in both observations).

Partial correlation: The coefficient of partial correlation is the measure of the relationship of two random variables, considering the influence of all other variables, analysed in parallel.

$$r_{ij} = \frac{-Ad(K_{ij})}{\sqrt{Ad(K_{ii}) \cdot Ad(K_{jj})}} \quad (4)$$

$Ad(K_{tl})$ – algebraic complement of the element tl of the correlation matrix K, containing coefficients of Pearson's correlation for all pairs created in the analysed system of random variables

Nonlinear correlation:

$$\rho = \sqrt{1 - \frac{\sum_{i=1}^{n}(y_i - f(x_i))^2}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}} \quad (5)$$

- f – model dependence of the variable Y from the variable X,
- y_i – empirical values of random variable Y,
- $f(x_i)$ – model values of dependent variable Y,
- $\hat{y}$ – mean value of variable Y,
- n – size of the random sample.

All correlation coefficients take the values within the interval [-1,1]. Rank coefficients r_S, r_K, r_G are measure of the monotonic relationship. They are all resistant to the diverging cases.

In order to estimate the magnitude of the differences between individual coefficients of correlation, we can create the confidence interval on the determined confidence level *p*, for example *p*=0.95. Then, as approached, we will consider the correlations, which values belong to the confidence interval.

For example, the confidence interval for the coefficient of Pearson's correlation is determined using the following formula:

$$r_p \in \left(\frac{e^{2\cdot z_1} - 1}{e^{2\cdot z_1} + 1}, \frac{e^{2\cdot z_2} - 1}{e^{2\cdot z_2} + 1}\right) \quad (6)$$

where:

$$z_1 = \frac{1}{2}\cdot ln\left(\frac{1+\hat{r}_p}{1-\hat{r}_p}\right) - \frac{u\left(1-\frac{\alpha}{2}\right)}{\sqrt{n-3}} \qquad z_2 = \frac{1}{2}\cdot ln\left(\frac{1+\hat{r}_p}{1-\hat{r}_p}\right) + \frac{u\left(1-\frac{\alpha}{2}\right)}{\sqrt{n-3}} \quad (7)$$

- $\hat{r}_p$ – estimator of Pearson's correlation coefficient,
- $u\left(1-\frac{\alpha}{2}\right)$ – quantile of the normal distribution for the confidence level $1\text{-}\alpha$.

ATTRIBUTES PARTS IN EXPLAINING REAL ESTATE PRICES

Based on the correlation relationship, the measure of which can be the square of the correlation coefficient, we can determine weight parts of individual features of real estate in creating their prices. To estimate relative parts in full space of random events creating probability space, we standardize the correlation square:

$$w_i = \frac{r_i^2}{\sum_{i=1}^{m} r_i^2} \quad (8)$$

r_i – correlation coefficient of attribute *i* with the real estate price.

On the basis of the weight parts, determined from different correlation types, we could select from a large database of the features describing real estates in a database, the features, which significantly shape market prices. The degree of diversification between these features can be an estimation criterion of the similarity between real estates.

So-called <u>*beta weights*</u> have a character similar to the weight parts. They are standardized coefficients of multiple regression and they can be calculated, like the partial correlations, on the basis of the correlation matrix K:

$$\beta_i = \frac{Ad(K_{0i})}{Ad(K_{00})} = a_i \cdot \frac{\sigma(x_i)}{\sigma(c)} \qquad (9)$$

where:

$Ad(K_{0i})$ $Ad(K_{00})$ – algebraic complements of the appropriate elements of the correlation matrix K, concerning real estate price, to which corresponds the index '0',

a_i – regression coefficient in the model of multiple regression, standing at the variable X_i,

$\sigma(x_i), \sigma(c)$ – standard deviations of the independent variable X_i and of the price.

Beta weights are a good measure of the estimation of relative degree of real estate prices explaining by individual attributes, on condition however that it is the case of a homogeneous market, where the multiple regression model can be well adjusted to the market tendencies.

EXAMPLE

Table 1. Different correlation coefficients of attribute with the premises price

Attribute	r_P	confidence interval for r_P		r_S	r_K	r_G	r_{ij}
Z	0,432	0,29	0,56	0,335	0,249	0,294	0,065
C	0,490	0,35	0,61	0,369	0,273	0,323	0,133
BS	0,336	0,18	0,47	0,330	0,246	0,284	0,082
PF	0,442	0,30	0,57	0,426	0,328	0,395	0,061
T	0,234	0,07	0,38	0,259	0,187	0,212	-0,108
BC	0,439	0,30	0,56	0,419	0,328	0,372	0,165
PC	0,142	-0,02	0,30	0,123	0,101	0,385	0,145
S	0,066	-0,10	0,23	0,049	0,032	0,038	0,049
FF	0,087	-0,08	0,25	0,010	-0,003	-0,003	0,191
UR	-0,117	-0,28	0,05	-0,117	-0,093	-0,128	-0,042
P	0,264	0,10	0,41	0,273	0,207	0,245	0,044
HL	0,034	-0,13	0,20	-0,010	-0,008	-0,012	0,047
LL	0,071	-0,10	0,23	0,065	0,054	0,262	0,111
SA	-0,043	-0,21	0,12	-0,014	-0,009	-0,009	-0,195
NR	0,184	0,02	0,34	0,190	0,146	0,177	0,147
percent of different				0,00	0,13	0,20	0,40

Table 2. Weight parts of the attributes in premises price, **n=142, R^2=0,43**

Atrybut	$w(r_P)$	$w(r_S)$	$w(r_K)$	$w(r_G)$	*BETA*
Z	0,16	0,12	0,12	0,09	0,12
C	0,21	0,15	0,14	0,11	0,23
BS	0,10	0,12	0,11	0,09	0,15
PF	0,17	0,20	0,20	0,17	0,11
T	0,05	0,07	0,07	0,05	-0,14
BC	0,17	0,19	0,20	0,15	0,20
PC	0,02	0,02	0,02	**0,16**	0,11
S	0,00	0,00	0,00	0,00	0,04
FF	0,01	0,00	0,00	0,00	0,19
UR	0,01	0,01	0,02	0,02	-0,04
P	0,06	0,08	0,08	0,06	0,05
HL	0,00	0,00	0,00	0,00	0,04
LL	0,00	0,00	0,01	**0,07**	0,09
SA	0,00	0,00	0,00	0,00	-0,29
NR	0,03	0,04	0,04	0,03	0,20

If we assume a symbolic limit for the significant value of the attribute weight part in the explanation of dwelling prices on the level of 3%, it turns out that the parts determined on the grounds of three different types of correlation detail exactly the same premises features as the features of significance for creating their prices. These are town zone, communication access, building surroundings, access to the public facilities, building technology, technical condition of the building, parking place and number of rooms. Only the Gamma correlation resulted in isolating additionally two features as essential for shaping prices: technical condition of the premises, legal loads.

Then, it can be concluded that about a half of considered dwelling features influences significantly their prices, and their selection is possible both on the grounds of Pearson's correlation and of Spearman or Kendall rank correlations. Gamma correlations lead to the results a bit different. For the selected in such a way price-making real estate features, we can apply one of the methods of similarity assessment to choose the most similar real estate.

Table 3. Different correlation coefficients of attribute with the premises price
***n*=125, R^2=0,63**

Attribute	r_P	confidence interval for r_P		r_S	r_K	r_G	r_{ij}
Z	0,517	0,38	0,63	0,396	0,295	0,347	0,238
C	0,547	0,41	0,66	0,388	0,286	0,334	0,072
BS	0,377	0,22	0,52	0,357	0,265	0,305	0,163
PF	0,490	0,34	0,61	0,453	0,350	0,419	-0,043
T	0,328	0,16	0,48	0,323	0,230	0,262	-0,004
BC	0,524	0,38	0,64	0,483	0,374	0,425	0,238
PC	0,226	0,05	0,39	0,197	0,162	0,647	0,362
S	0,115	-0,06	0,29	0,085	0,059	0,070	0,203
FF	0,017	-0,16	0,19	-0,071	-0,068	-0,082	0,154
UR	-0,084	-0,26	0,09	-0,094	-0,074	-0,101	0,119
P	0,291	0,12	0,44	0,287	0,215	0,254	-0,076
HL	0,112	-0,06	0,28	0,049	0,039	0,055	0,121
LL	0,076	-0,10	0,25	0,069	0,057	0,263	0,202
SA	-0,118	-0,29	0,06	-0,052	-0,038	-0,038	-0,396
NR	0,244	0,07	0,40	0,237	0,181	0,221	0,304
percent of different				0,07	0,20	0,27	0,60

Table 4. Weight parts of the attributes in premises price, ***n*=125, R^2=0,63**

Attribute	$w(r_P)$	$w(r_S)$	$w(r_K)$	$w(r_G)$	*BETA*
Z	0,17	0,13	0,13	0,09	0,36
C	0,19	0,13	0,12	0,08	0,10
BS	0,09	0,11	0,10	0,07	0,24
PF	0,15	0,17	0,18	0,13	-0,07
T	0,07	0,09	0,08	0,05	0,00
BC	0,17	0,20	0,21	0,13	0,25
PC	0,03	0,03	0,04	**0,30**	0,25
S	0,01	0,01	0,01	0,00	0,13
FF	0,00	0,00	0,01	0,00	0,12
UR	0,00	0,01	0,01	0,01	0,09
P	0,05	0,07	0,07	0,05	-0,07
HL	0,01	0,00	0,00	0,00	0,08
LL	0,00	0,00	0,00	0,05	0,13
SA	0,01	0,00	0,00	0,00	-0,48
NR	0,04	0,05	0,05	0,04	0,34

Table 4. is an equivalent of the table 2 for a full database. The weight parts included here, calculated on the grounds of the correlations from the table 3, can be compared with *BETA weights*, as the multiple regression model sufficiently explains the relationships on the analysed real estate market. So, larger than previously number of pale blue weight parts of attributes in a price proves that not necessarily the same features of a real estate considered as factors independently shaping their prices, shape significantly a dependent variable (by price) in a multidimensional regression model.

Assuming, like previously, that the criterion of considering the weight part as significant on the level of 3% - also in this case, the same attributes turned out to have a significant influence on the price in the event of applying three first types of correlation: r_p, r_S, r_K. Gamma correlation, as previously, gives slightly different results. The same 8 attributes mentioned above and additionally the technical condition of the premises significantly shape the prices. These results do not coincide with the indications of *beta weights*. As mentioned above, it can be due to a different character of the influence of individual independent variables on the dependent variable, treated as a system of variables from the situation, when we consider each variable separately.

CONCLUSIONS

Among the correlation coefficients applied to assessing the influence of facultative attributes on real estate market prices, Pearson's correlations and Spearman's or Kendall's rank correlations lead to the convergent results. It allows applying interchangeably these parameters depending on a quantity or quality character of the variables used to describe the real estates.

As it turned, the Gamma correlation leads to slightly different conclusions. However, the differences in relation to the results achieved for the other types of correlation lie in an only slight extension of the database of real estate features recognized as the ones having a price-making character. Then, we can suppose that Gamma correlation is a less restrictive indicator of the dependence degree between variables.

Beta weights assessing the relative influence of independent variables on the values of a dependent variable in a multidimensional regression model, do not lead to distinguish the same significant variables, which generate the weight parts determined independently for each variable.

Thank you very much for your attention