Józef Czaja *, Edward Preweda *

ANALIZA ILOŚCIOWA RÓŻNYCH WSPÓŁCZYNNIKÓW KORELACJI NA PRZYKLADZIE SZEŚCIOWYMIAROWEJ ZMIENNEJ LOSOWEJ **

1. Charakterystyka analizowanej próby zmiennej losowej

Przedmiotem analizy są informacje rynkowe o nieruchomościach gruntowych, które zostały opisane za pomocą cen jednostkowych (zł/ m^2) oraz za pomocą pięciu atrybutów.Dla poszczególnych atrybutów przyjęto następujące skale:

1) strefa (miasta) - centralna (4), śródmiejska (3), pośrednia (2), peryferyjna (1);
2) komunikacja (dostęp) – b. korzystna (2), korzystna (1), przeciętna (0), niekorzystna (-1);
3) moda (na lokalizację) – najmodniejsza (5), b. modna (3), modna (1), przeciętna (0), niemodna (-1);
4) otoczenie - b. korzystne (2), korzystne (1), przeciętne (0), niekorzystne (-1);
5) pole powierzchni - liczba arów.

Zebrane informacje dla 15 nieruchomości gruntowych zostały przedstawione w tabeli 1.

Analizowane ceny nieruchomości, jako zmienna losowa jednowymiarowa, charakteryzują się następującymi parametrami:

– wartością przeciętną $\bar{c} = 77{,}93$ zł/m^2 ,
– odchyleniem standardowym cen rynkowych $\sigma_n[c] = 38{,}00$ zł/m^2,
– współczynnikiem dyspersji (rozproszenia) $v_c = 0{,}49$.

* Akademia Górniczo-Hutnicza, Katedra Informacji o Terenie
** Praca stanowi wynik realizacji badań statutowych w AGH Nr 11.11.150.316

Tabela 1. Zestawienie atrybutów i cen nieruchomości gruntowych

Lp.	Strefa	Komunikacja	Moda	Otoczenie	Pole	Cena
1	2	2	1	2	6	56
2	1	2	1	2	8	78
3	1	2	3	2	4	98
4	2	2	3	1	7	135
5	3	1	0	0	8	46
6	3	1	–1	–1	8	25
7	2	1	0	0	8	46
8	4	0	3	2	6	98
9	4	1	5	2	7	112
10	4	–1	5	2	6	145
11	1	0	5	1	3	115
12	1	2	3	2	4	98
13	3	1	0	0	8	46
14	3	1	–1	–1	15	25
15	2	1	0	0	8	46

Na podstawie współczynnika rozproszenia można wnioskować, że ceny nieruchomości są silnie uzależnione od innych zmiennych losowych, które w tym przypadku są reprezentowane przez atrybuty (cechy) rozważanych nieruchomości. Analiza rynku nieruchomości pozwoli określić ilościowy wpływ poszczególnych atrybutów na zmienność cen nieruchomości. Do analizy tej powinny być wykorzystane odpowiednie współczynniki korelacji oraz ściśle zdefiniowane nowe parametry, czyli funkcje na tych współczynnikach.

2. Formuły do analizy statystycznej rynku – zgodnie z [11]

Dla analizy statystycznej rynku zakładamy, że zmienną zależną jest cena, natomiast atrybuty nieruchomości reprezentują zmienne niezależne. Korelacje między ceną i tymi cechami nieruchomości można opisywać za pomocą jednego modelu wielowymiarowej zmiennej losowej lub za pomocą kilku niezależnych modeli dwuwymiarowej zmiennej losowej.

Kowariancja cov[X, Y] pomiędzy zmienną X i zmienną Y jest parametrem określającym wzajemną zależność zmiennych losowych w rozkładzie dwuwymiarowym. Wielkość tę można standaryzować za pomocą wartości odchyleń standardowych $\sigma[X]$ i $\sigma[Y]$ obu zmiennych losowych w rozkładach brzegowych. Standaryzowana cov[X, Y], czyli

$$r = \frac{\mathrm{cov}[X,Y]}{\sigma[X]\cdot\sigma[Y]} \tag{1}$$

nosi nazwę **współczynnika korelacji zupełnej (Pearsona)**.

Z formuły (1) widać, że współczynnik korelacji zupełnej jest bezwymiarowy i może się zawierać w przedziale od –1 do +1. Bezwzględna wartość współczynnika korelacji określa siłę współzależności liniowej, zaś znak współczynnika korelacji wyraża charakter współzależności. Ujemna wartość współczynnika korelacji dowodzi, iż współzależność rozpatrywanych zmiennych ma charakter negatywny, a to oznacza, że wzrastanie wartości jednej zmiennej powoduje zmniejszanie wartości drugiej zmiennej. Dodatnia wartość współczynnika korelacji świadczy o tym, że współzależność rozpatrywanych zmiennych ma charakter pozytywny.

Jeżeli wartości zmiennych losowych zaobserwowane w próbie (x_i, y_i), dla $i = 1, 2, ..., n$ będą posiadały równomierny rozkład prawdopodobieństwa

$$p_i = \frac{1}{n} \tag{2}$$

to estymatorem współczynnika korelacji liniowej będzie współczynnik $\hat{r}$:

$$\hat{r} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} = \frac{\frac{1}{n}\sum_{i=1}^{n} (x_i y_i) - \bar{x} \cdot \bar{y}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} (x_i)^2 - \bar{x}^2} \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i)^2 - \bar{y}^2}} \tag{3}$$

przy czym

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{4}$$

$$\sigma[x] = \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \qquad \sigma[y] = \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{5}$$

oznaczają wartość przeciętną oraz odchylenia standardowe rozpatrywanych zmiennych, określone na podstawie wyników z próby.

Jeżeli opis zmiennych losowych jest utrudniony pod względem ilościowym, wtedy elementom badanych zmiennych można nadać rangi, które są traktowane jako zaobserwowane wartości zmiennej losowej skokowej o równomiernym rozkładzie prawdopodobieństwa. W przypadku rozpatrywania dwóch cech (X, Y), można kolejnym obserwacjom każdej z cech z osobna nadać odpowiednie rangi, a następnie wyznaczyć współczynnik korelacji rang. W tym celu elementy (x_i, y_i) zaobserwowanej próby losowej porządkujemy według niemalejących wartości x_i, czyli

$$\begin{pmatrix} x_{(1)}, & x_{(2)}, & x_{(3)}, & \dots & x_{(n)} \\ y_1^* & y_2^* & y_3^* & \dots & y_n^* \end{pmatrix} \text{ gdzie: } x_{(1)} \le x_{(2)} \le x_{(3)} \le \dots \le x_{(n)} \tag{6}$$

i zastępujemy pary elementów $(x_{(i)}, y_i^*)$ odpowiadającymi im rangami. Dla zmiennej x_i rangami są liczby naturalne 1, 2, ..., n, zaś dla zmiennej y_i rangami są elementy $y_i^* \leftrightarrow s_i$. Zbiór wyników obserwacji (x_i, y_i) (6) zastępujemy zatem nowym zbiorem (i, s_i), którego elementy zawierają tylko numery kolejności uporządkowania wartości (x_i, y_i):

$$\begin{pmatrix} 1, & 2, & 3, & ..., & n \\ s_1 & s_2 & s_3 & ... & s_n \end{pmatrix} \tag{7}$$

Współczynnik korelacji rang (kolejności uporządkowania) rozpatrywanych cech, zdefiniowany na podstawie wzoru (3) (por. [11]), czyli

$$\hat{r}_s = 1 - \frac{6\sum_{i=1}^{n}(i - s_i)^2}{n^3 - n} \tag{8}$$

nosi nazwę **współczynnika korelacji rang (kolejności) Spearmana**.

Na bazie zbioru rang (7) można również konstruować **współczynnik korelacji rang Kendalla** ($\hat{r}_k$), który oblicza się według wzoru (por. [11]):

$$\hat{r}_k = \frac{\textit{Suma przyporządkowanych not}}{\textit{Liczba wszystkich par}} = \frac{2U}{n(n-1)} \tag{9}$$

Jeżeli chcemy jednocześnie analizować wszystkie zmienne, to dalsze parametry wzajemnej zależności (korelacji) mogą być zdefiniowane na podstawie macierzy korelacyjnej $\{K\}$. Elementy tej macierzy stanowią współczynniki korelacji zupełnej (Pearsona) r_{ij} pomiędzy poszczególnymi zmiennymi Y, X_i, X_j, przy czym wskaźnik „0" będzie odpowiadał zmiennej zależnej Y. Uwzględniając powyższe oznaczenia, macierz korelacyjna przyjmuje następującą postać

$$\{K\} = \begin{Bmatrix} 1 & r_{01} & r_{02} & \cdots & r_{0k} \\ r_{10} & 1 & r_{12} & \cdots & r_{1k} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ r_{k0} & r_{k1} & r_{k2} & \cdots & 1 \end{Bmatrix} \tag{10}$$

Pary atrybutów, które wykazują współczynnik korelacji wysoki ($|r_{ij}| > 0{,}6$), mają podobny wpływ na kształtowanie się cen nieruchomości, a to oznacza, że oba atrybuty wyjaśniają analogiczną część zmienności zmiennej Y (cen).

Na podstawie macierzy korelacyjnej (10) można zdefiniować **współczynniki korelacji cząstkowej (częściowej)**, opisujące współzależność dwóch ustalonych zmiennych X_i, X_j, ale przy uwzględnieniu wpływu pozostałych zmiennych losowych, czyli

$$r_{ij\cdot 1,2,\ldots,k} = \frac{-Ad[K_{ij}]}{\sqrt{Ad[K_{ii}]\cdot Ad[K_{jj}]}} \tag{11}$$

gdzie:

$Ad[K_{ij}]$ – algebraiczne dopełnienie elementu k_{ij} macierzy $\{K\}$,
$Ad[K_{ii}]$ – algebraiczne dopełnienie elementu k_{ii} macierzy $\{K\}$,
$Ad[K_{jj}]$ – algebraiczne dopełnienie elementu k_{jj} macierzy $\{K\}$.

Dla każdej wyróżnionej zmiennej (i-tej) można określić **współczynnik liniowej korelacji wielorakiej *R***, który opisuje jednoczesną współzależność wszystkich pozostałych zmiennych

$$R_i = \sqrt{1 - \frac{det\{K\}}{det\{K_i\}}} \tag{12}$$

gdzie:

$det\{K\}$ – wyznacznik z całej macierzy $\{K\}$,
$det\{K_i\}$ – wyznacznik podmacierzy $\{K_i\}$, która powstaje ze skreślenia i-tego wiersza i i-tej kolumny macierzy $\{K\}$, czyli elementów macierzy dotyczących współczynników korelacji dla zmiennej wyróżnionej.

Standaryzowane współczynniki regresji (wagi β) oblicza się według następującego wzoru

$$\beta_i = \frac{Ad[K_{0i}]}{Ad[K_{00}]}\,\frac{\sigma[x_i]}{\sigma[y]} \tag{13}$$

przy czym:

$Ad[K_{0i}]$, $Ad[K_{00}]$ – algebraiczne dopełnienia odpowiednich elementów macierzy $\{K\}$,
$\sigma[y]$, $\sigma[x_i]$ – odchylenia standardowe dla zmiennej zależnej i zmiennych niezależnych liczone z próby.

Standaryzowane współczynniki regresji (wagi β) wskazują, z jaką siłą zmienia się standaryzowana zmienna zależna w odniesieniu do wyróżnionej zmiennej zależnej standaryzowanej, przy równoczesnym uwzględnieniu wpływów pozostałych zmiennych.

3. Wyniki analizy różnych współczynników korelacji cen i atrybutów nieruchomości

Realizując wzory (3), (8) i (9) dla atrybutów i cen nieruchomości zamieszczonych w tabeli 1 otrzymano poszczególne rodzaje współczynników korelacji, których wartości ilustruje tabela 2. Zamieszczone w tabeli 2 wyniki obliczeń dowodzą, że wartości poszczególnych rodzajów współczynników korelacji mogą różnić się między sobą w zakresie istotnym (wielokrotnym), a dla niektórych zmiennych słabo skorelowanych współczynniki korelacji mogą się różnić nawet znakiem. Na podstawie tych analiz można sformułować wniosek, że do analizy rynku nieruchomości powinny być uwzględniane tylko współczynniki korelacji zupełnej Pearsona.

Tabela 2. Współczynniki korelacji zupełnej Pearsona, korelacji Spearmana i korelacji Kendalla

Oznaczenie zmiennej	Rodzaj korelacji	Strefa [1]	Komunikacja [2]	Moda [3]	Otoczenie [4]	Pole [5]	Cena [6]
Strefa [1]	**Pearsona**	1	**–0,6055**	**0,0351**	**–0,1424**	**0,3596**	**–0,0074**
	Spearmana	1	–0,6160	–0,0500	–0,1415	0,3034	–0,0691
	Kendalla	1	–0,5747	–0,1163	–0,1752	0,2376	–0,1119
Komunikacja [2]	**Pearsona**		1	**–0,3266**	**0,0741**	**0,0273**	**–0,2382**
	Spearmana		1	–0,1388	0,2185	–0,0120	–0,0760
	Kendalla		1	–0,0624	0,1745	–0,0254	–0,0360
Moda [3]	**Pearsona**			**1**	**0,7558**	**–0,6741**	**0,9396**
	Spearmana			1	0,7931	–0,8234	0,9639
	Kendalla			1	0,6970	–0,7053	0,9187
Otoczenie [4]	**Pearsona**				**1**	**–0,6652**	**0,7479**
	Spearmana				1	–0,7184	0,7416
	Kendalla				1	–0,6221	0,6120
Pole [5]	**Pearsona**					**1**	**–0,6163**
	Spearmana					1	–0,7643
	Kendalla					1	–0,6232
Cena [6]	**Pearsona**						**1**
	Spearmana						1
	Kendalla						1

W dalszej analizie określono wartości współczynników korelacji cząstkowej (11) dla zmiennej zależnej (ceny) względem pozostałych zmiennych, czyli atrybutów oraz wartości standaryzowanych współczynników regresji (wag β) liczonych według wzoru (13). Wartości tych współczynników zestawiono w tabeli 3.

Tabela 3. Współczynniki korelacji cząstkowej (Pearsona) oraz wagi dla zmiennej zależnej (ceny)

Rodzaj parametru	Strefa [1]	Komunikacja [2]	Moda [3]	Otoczenie [4]	Pole [5]
Korelacja cząstkowa	**–0,0742**	**0,0894**	**0,8235**	**0,1167**	**0,1695**
Korelacja zupełna	–0,0074	–0,2382	0,9396	0,7479	–0,6163
Waga	**–0,0365**	**0,0467**	**0,9646**	**0,0728**	**0,0942**

Z porównania wartości współczynników korelacji cząstkowej i korelacji zupełnej widać, że tylko dla atrybutu „moda" wpływ innych zmiennych jest minimalny na wyjaśnianie zmienności cen nieruchomości. Zmienność pozostałych atrybutów względem cen jest mocno uzależniona od wpływu innych atrybutów, gdyż wartości współczynników korelacji cząstkowej i korelacji zupełnej różnią się w istotnym zakresie.

Otrzymane wartości wag 13 świadczą o tym, że w liniowych modelach zmienności cen nieruchomości największy wpływ będzie miał atrybut „moda", zaś pozostałe atrybuty będą mieć minimalny wpływ.

Na podstawie zamieszczonej w tabeli 2 macierzy korelacyjnej (współczynników Pearsona) obliczono według wzoru (12) wartości współczynnika korelacji wielorakiej oraz jego kwadrat, czyli

$$R = 0{,}9444,$$

$$R^2 = 0{,}8919.$$

Otrzymane wartości dowodzą, że razem wszystkie atrybuty są skorelowane liniowo z cenami nieruchomości z siłą 0,9444 oraz, że w liniowym modelu wielorakiej regresji rozważane atrybuty mogą wyjaśniać 89,19% zmienności cen nieruchomości zawartych w bazie.

Na podstawie ilościowej analizy formuł określających współczynnik korelacji rang Spearmana i współczynnik korelacji rang Kendalla można wyrazić opinię, że współczynniki korelacji rang powinny być stosowane do opisywania współzależności takich cech, dla których są ograniczone możliwości ich skalowania (wartościowania). W procesie opisywania zmiennych za pomocą rang zaniedbywany jest wpływ rozpiętości skal i wpływ uwzględniania różnic występujących pomiędzy wartościami sąsiednich zmiennych na wartość odpowiednich współczynników regresji tych zmiennych. W przypadku, gdy w rozpatrywanej próbie wartości cech powtarzają się, wtedy ich uszeregowanie (rangowanie) może być wykonane na wiele sposobów, a to dowodzi, że współczynniki korelacji rang mogą przyjmować różne wartości.

Z analizy tej można wnioskować, że dla zmiennych losowych kwantyfikowanych zawsze powinien być stosowany współczynnik korelacji zupełnej (Pearsona), który posiada ścisłą interpretację analityczną i geometryczną. Współczynniki korelacji Pearsona definiują elementy macierzy korelacyjnej dla zmiennej losowej wielowymiarowej, która jest podstawą do wszelkich analiz statystycznych tych zmiennych. Na podstawie elementów macierzy korelacyjnej można obliczyć współczynnik korelacji wielorakiej, współczynniki korelacji cząstkowej oraz współczynniki regresji i wagi beta.

Literatura

[1] Cegielski P.: *Przykład alternatywnej techniki analizy statystycznej rynku - część 2.* Warszawa, Rzeczoznawca Majątkowy, nr 2, 1999

[2] Czaja J.: *Estymacja uogólnionych modeli liniowych.* Warszawa, PAN, Geodezja i Kartografia, t. XLIII, z. 3, 1994

[3] Czaja J. i in.: *Wycena nieruchomości majątkowych metodą cenowo-porównawczą.* Warszawa, Przegląd Geodezyjny, 12, 1996
[4] Czaja J.: *Zamiast techniki porównywania parami – technika interpolacyjna – wyceny nieruchomości.* Katowice, Kwartalnik Nieruchomość, nr 3, 1996
[5] Czaja J.: *Modele statystyczne w informacji o terenie.* Kraków, Wydawnictwa AGH 1997
[6] Czaja J.: *Interval Estimation of Generalized Linear Modeels.* Warszawa, PAN, Geodezja i Kartografia, t. XLVI,z. 1, 1997
[7] Czaja J.: *Praktyczna weryfikacja różnych technik wyceny w podejściu porównawczym.* Kraków, Rzeczoznawca Małopolski, nr 2, 1997
[8] Czaja J.: *Metody i systemy szacowania nieruchomości.* Kraków, podręcznik dla rzeczoznawców majątkowych, 1999
[9] Czaja J.: *Podejście porównawcze wyceny nieruchomości w aspekcie standardów.* Warszawa, Rzeczoznawca Majątkowy, nr 4, 1999
[10] Czaja J.: *Metody i systemy określania wartości nieruchomości.* Kraków,UWND AGH 1999
[11] Czaja J., Preweda E.: *Analiza statystyczna zmiennej losowej wielowymiarowej w aspekcie korelacji i predykcji.* Kraków, UWND, AGH, Geodezja, t. 6, 2000
[12] Mazurkiewicz E.: *Uwagi do procedur określania wartości rynkowej nieruchomości przy zastosowaniu analizy statystycznej rynku.* Warszawa, Rzeczoznawca Majątkowy, nr 2, 1999
[13] Prystupa M.: *Wycena nieruchomości metodą cenowo-porównawczą.* Warszawa, Biblioteczka rzeczoznawcy majątkowego 1997
[14] Prystupa M.: *Jaki standard dla podejścia porównawczego?* Warszawa, Rzeczoznawca Majątkowy, nr 4, 1999
[15] Rao C.R.: *Modele liniowe statystyki matematycznej.* Warszawa, PWN 1982

Recenzent
prof. dr hab. inż. Stanisław Latoś