PRZEGLĄD STATYSTYCZNY
R. LVII – ZESZYT 2-3 – 2010

PAWEŁ CABAŁA

ZASTOSOWANIE WSPÓŁCZYNNIKA KONKORDANCJI W POMIARZE ZGODNOŚCI OCEN EKSPERTÓW

1. UWAGI WSTĘPNE

Według tzw. koherencyjnej teorii prawdy prawdziwe jest to, co jest wewnętrznie spójne, co do czego istnieje zgoda. Zgodność ocen wyrażanych przez ekspertów jest podstawą formułowania sądów ogólnych, które jednak – wbrew postulatom teorii koherencji – nie zawsze okazują się sądami prawdziwymi. Znane z historii biznesu przypadki pomyłek ekspertów doprowadzały do poważnych dysfunkcji wdrażanych rozwiązań, upadku przedsiębiorstw, a także kosztujących życie ludzi katastrof przemysłowych. Mimo ryzyka bezkrytycznego polegania na opiniach ekspertów, błędów jednomyślności i ich skutków społecznych, wydaje się, że nie ma lepszej alternatywy. Ryzyko wystawienia błędnej oceny bywa jeszcze większe, gdy polegamy na jednej tylko opinii.

Weryfikacja zgodności ocen odgrywa doniosłą rolę tak w zakresie badań diagnostycznych jak i w studiach prospektywnych (np. *foresighting*). Rzetelna ocena złożonych zjawisk społeczno-gospodarczych nie jest w praktyce możliwa bez odwoływania się do opinii ekspertów. W analizie i projektowaniu systemów zarządzania przedsiębiorstwem ekspertami są nie tylko konsultanci zewnętrzni, lecz przede wszystkim kadra menedżerska, która jest odpowiedzialna za przygotowanie i wdrażanie strategii.

Jeżeli zgodność wystawionych przez grupę ekspertów ocen jest na odpowiednio wysokim poziomie, to jesteśmy uprawnieni do formułowania sądów ogólnych. Sąd ogólny w tym przypadku jest średnią (lub sumą) wystawionych ocen przez wszystkich ekspertów. Odnotowana rozbieżność ocen nie uprawnia nas do wykonywania takich operacji. Stosowana praktyka odrzucania ocen skrajnych i przyjmowania średniej pozostałych ocen jest również niedopuszczalna z metodologicznego punktu widzenia. Sprawą otwartą pozostaje odpowiedź na pytanie, jakie działania należy podjąć w przypadku braku zgodności w ocenach.

Generalnie istnieją trzy źródła braku zgodności w opiniach na temat danego obiektu. Pierwszym jest niska kompetencja grupy osób oceniających (nie chodzi tyle o niskie kompetencje poszczególnych osób, lecz o brak kompetencji całej grupy oceniających). Drugi powód braku zgodności jest związany z niewłaściwie zorganizowanym procesem oceny. Trzecim źródłem niezgodności jest źle zdefiniowany obiekt oceny.

Stwierdzenie braku zgodności w ocenach umożliwia podjęcie działań mających na celu eliminację przyczyn niezgodności bądź – gdy przyczyny są niemożliwe do usunięcia – powstrzymanie się od formułowania oceny ogólnej.

W praktyce badawczej do oceny zgodności ocen (preferencji) często wykorzystuje się metody analizy korelacji rang. Proces oceny sprowadza się wtedy do realizacji trzech faz badawczych. W fazie pierwszej ustalany jest stopnień współzależności między ocenami ekspertów na temat względnej pozycji badanych obiektów. W fazie drugiej przeprowadzane są odpowiednie testy niezależności. Faza trzecia polega na sporządzeniu ostatecznego porządku rangowego badanych obiektów.

Artykuł niniejszy przedstawia zagadnienie analizy zgodności ocen z wykorzystaniem metod korelacji rang oraz dyskusję nad zaletami i ograniczeniami stosowania tych metod. Zwrócono uwagę przede wszystkim na dwa zagadnienia: pomiar zgodności między ocenami co najmniej trzech ekspertów oraz problematykę pomiaru zgodności w przypadkach występowania tzw. rang wiązanych.

2. POMIAR ZGODNOŚCI USZEREGOWAŃ MOCNYCH

Dwa czynniki mają rozstrzygające znacznie w pomiarze zgodności ocen ekspertów. Są to: ilość szeregów preferencyjnych oraz rodzaj skali porządkowej. Szereg preferencyjny odzwierciedla uporządkowanie zbioru obiektów (elementów) dokonane przez danego eksperta (obserwatora, sędziego, jurora). Każdemu obiektowi nadawana jest ranga określająca jego pozycję względem pozostałych obiektów na skali porządkowej. Rodzaj skali porządkowej (mocna lub słaba) ma natomiast wpływ zarówno na sposób obliczeń, jak i na dobór odpowiedniego testu istotności.

Do pomiaru zgodności uporządkowań między dwoma szeregami preferencyjnymi stosuje się współczynnik korelacji rang ρ Spearmana lub współczynnik korelacji rang τ Kendalla. Gdy mamy do czynienia z więcej niż dwoma szeregami rangowymi, wówczas najczęściej stosowanym miernikiem oceny zgodności preferencji jest wskaźnik konkordancji W Kendalla (*concordance coefficient*), zwany współczynnikiem zgodności [1] lub precyzyjniej: współczynnikiem zgodności uporządkowań wielokrotnych [5].

Poniżej skoncentrujemy się na problemie oceny zgodności między więcej niż dwoma uszeregowaniami ($m > 2$). Rozwinięte zostanie również zagadnienie uszeregowań słabych (tzw. rang wiązanych – *tied ranks*), które mimo że komplikuje obliczenia, ma niebagatelne znacznie praktyczne.

Pomiar stopnia zgodności uporządkowań sprowadza się do konstrukcji współczynnika, w którym licznik wyraża wartość odzwierciedlającą stopień rzeczywistych powiązań między szeregami preferencyjnymi (S), a mianownik analogiczną wartość, jednak dla sytuacji pełnej zgodności uporządkowań rangowych, czyli maksymalnie możliwą do uzyskania ($S_{\max}$).

Rangę j-tego obiektu (gdzie $j = 1, 2, \ldots, n$) nadaną przez i-tego eksperta ($i = 1, 2, \ldots, m$) oznaczmy a_{ij}. W wyniku oceny n obiektów dokonanej przez m obserwatorów otrzymujemy macierz uporządkowań, której wiersze przedstawiają szeregi preferencyjne, a kolumny rangi nadane obiektom przez kolejnych ekspertów:

$$\begin{matrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ . & . & \dots & . \\ a_{m1} & a_{m_2} & \dots & a_{mn} \end{matrix}$$

Wyróżniamy dwa typy uporządkowań rangowych (czyli szeregów preferencyjnych): mocne i słabe (o rangach niepowiązanych i rangach powiązanych). Najpierw omówimy pierwszy przypadek, czyli zagadnienie pomiaru zgodności między szeregami o uporządkowaniu mocnym.

Rozważmy przykład podany w tabeli 1. Pięciu obiektom A, B, C, D oraz E ($n = 5$) zostały nadane rangi przez trzech ekspertów ($m = 3$). Rangi te nie są powiązane, co oznacza że mamy do czynienia z uporządkowaniem mocnym.

Tabela 1

Przykład trzech szeregów preferencyjnych o uporządkowaniu mocnym

Wyszczególnienie	A	B	C	D	E	Suma szeregów
ekspert 1	4	2	5	1	3	**15**
ekspert 2	3	1	4	2	5	**15**
ekspert 3	2	3	5	1	4	**15**
Suma rang (R_j)	**9**	**6**	**14**	**4**	**12**	**45**

Źródło: opracowanie własne.

W uporządkowaniu mocnym wszystkie liczby oznaczające pozycje ocenianych obiektów są różne. Mówiąc inaczej ekspert jest w stanie rozróżnić ważność ocenianych obiektów (uszeregować je w kolejności ważności). Bez względu na porządek takiego uszeregowania, suma rang dla danego szeregu jest równa sumie n pierwszych liczb naturalnych, czyli $\frac{n(n+1)}{2}$. Jeżeli wszystkie szeregi preferencyjne będą miały porządek mocny, to suma sum tych rang będzie równa $\frac{mn(n+1)}{2}$, a średnia arytmetyczna sum rang $\frac{m(n+1)}{2}$. W zapisie formalnym:

a) suma rang R_j dla j-tego obiektu:

$$R_j = \sum_{i=1}^{m} a_{ij}, \tag{1}$$

b) suma sum rang wszystkich n obiektów:

$$\sum_{j=1}^{n} R_j = \frac{mn(n+1)}{2}, \tag{2}$$

c) średnia arytmetyczna sum rang dla wszystkich n obiektów:

$$\bar{R} = \frac{1}{n}\sum_{j=1}^{n} R_j = \frac{m(n+1)}{2}. \tag{3}$$

Jeżeli sumę kwadratów odchyleń R_j od średniej $\bar{R}$ oznaczymy jako S, to:

$$S = \sum_{j=1}^{n}\left(R_j - \bar{R}\right)^2 = \sum_{j=1}^{n}\left(R_j - \frac{m(n+1)}{2}\right)^2. \tag{4}$$

Wielkość S osiąga maksymalną wartość (którą oznaczymy S_{max}) tylko w przypadku pełnej zgodności rang między szeregami preferencyjnymi. W takiej sytuacji sumy rang R_j dla kolejnych obiektów będą równe: m, $2m$, …, jm, …, nm (lub w dowolnym innym porządku). Maksymalną wartość sumy kwadratów odchyleń R_j od średniej $\bar{R}$ można wyznaczyć w sposób następujący:

$$S_{\max} = \sum_{j=1}^{n}\left(jm - \frac{m(n+1)}{2}\right)^2 = m^2\sum_{j=1}^{n}\left(j^2 - 2j\frac{(n+1)}{2} + \frac{(n+1)^2}{4}\right), \tag{5}$$

a ponieważ suma kwadratów n pierwszych liczb naturalnych wynosi $\frac{n(n+1)(2n+1)}{6}$, więc

$$S_{\max} = m^2\left(\frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)^2}{2} + \frac{n(n+1)^2}{4}\right) = \frac{m^2\left(n^3-n\right)}{12}. \tag{6}$$

Współczynnik zgodności W jest ilorazem rzeczywistej wielkości S i wielkości S_{max}, czyli maksymalnej możliwej do uzyskania. Ostatecznie:

$$W = \frac{S}{S_{\max}} = \frac{\sum_{j=1}^{n}\left(R_j - \frac{m(n+1)}{2}\right)^2}{\frac{1}{12}m^2\left(n^3-n\right)}. \tag{7}$$

Współczynnik W osiąga wartości od 0 do 1. W przypadku całkowitej niezgodności uszeregowań S wyniesie zero bądź będzie relatywnie niskie w porównaniu z S_{max}.

Dla przykładu podanego w tabeli 1 stopień zgodności ocen trzech ekspertów wyrażony współczynnikiem konkordancji wynosi:

$$W = \frac{12}{3^2\left(5^3-5\right)}\sum_{j=1}^{n}\left(R_j - 9\right)^2 = \frac{12\cdot 68}{1080} = 0{,}756.$$

3. POMIAR ZGODNOŚCI USZEREGOWAŃ SŁABYCH

Rangi powiązane oznaczają sytuacje, w której eksperci oceniający obiekty uznają, że nie ma różnicy między niektórymi z tych obiektów. Brak różnicy w ocenie oznacza,

że jedna (lub więcej) ranga wiąże co najmniej dwa obiekty. Możemy przy tym rozważyć dwa szczególne przypadki:

a) obserwatorzy proszeni są o uszeregowanie n obiektów w skali od 1 do n i nie są w stanie rozróżnić istotności między niektórymi obiektami, w związku z tym nadają nierozróżnialnym obiektom tę samą rangę. Przykładowo szereg: 6, 2, 3, 1, 3, 4, 8, 7, 6 oznacza, że obserwator nadał rangi 9 obiektom, przy czym nie był w stanie dostrzec różnicy między obiektem 1 i 9 oraz 3 i 5 (pozycje te są związane tą samą rangą),

b) obserwatorzy proszeni są o uszeregowanie n przedmiotów w skali od 1 do k, gdzie k jest liczbą naturalną, ale mniejszą od n. Przykładowo obserwator jest proszony o ocenę 10 obiektów, przy czym ocena ma być dokonana w skali od 1 do 7. Naturalnym jest, że w takim przypadku część obiektów będzie musiała być związana tą samą rangą.

Jeżeli wyniki ocen ekspertów dają szeregi z rangami wiązanymi i jeżeli chcemy posłużyć się w pomiarze zgodności współczynnikiem konkordancji, konieczne jest zastosowanie metody rang uśrednionych (*mid-rank metod*). Metoda ta polega na uśrednieniu rang powiązanych w taki sposób, by dały szereg analogiczny do szeregu, w którym wszystkie obiekty są rozróżnialne. W wyniku takiego przekształcenia otrzymujemy uszeregowanie, którego suma rang jest równa $\frac{n(n+1)}{2}$, czyli równa sumie analogicznego szeregu o uporządkowaniu mocnym. W tabeli 2 pokazano przykład zamiany rang wiązanych na rangi uśrednione.

Tabela 2

Przykład przekształcania rang wiązanych

Numer obiektu	A	B	C	D	E	F	G	H	I	Suma rang
Rangi eksperta:	6	2	3	1	3	4	6	7	6	38
Rangi uśrednione:	7	2	3,5	1	3,5	5	7	9	7	45

Źródło: opracowanie własne.

Jak widać w tabeli 2 surowe wyniki rangowania dają szereg, którego suma jest równa 38. Obiekt D otrzymał rangę 1, a obiekt B rangę 2. Obiekty C oraz E są związane tą samą rangą 3. W celu zamiany na rangi uśrednione należy ich pozycje uśrednić, a ponieważ zajmują one w kolejności rangowania pozycje 3 i 4 (po D i B) więc obiekty te otrzymują rangę 3,5 $\left(\frac{3+4}{2}\right)$. Piątą pozycję w kolejności zajmuje obiekt F, który otrzymał rangę 4, lecz ze względu na dwie poprzedzające go rangi wiązane otrzymuje rangę 5 (przed nim znalazły się już 4 obiekty D, B, C i E). Obiekty A, G i I wiąże z kolei ranga 6, a ponieważ zajmują one kolejne pozycje 6, 7 i 8, to otrzymują rangę 7 $\left(\frac{6+7+8}{3}\right)$. Ostatnią pozycję (czyli rangę 9) zajmuje obiekt H, którego pierwotna ranga wyniosła 7. Suma rang w ten sposób uśrednionych wynosi 45 i jak można sprawdzić jest równa sumie pierwszych 9 liczb naturalnych.

Poza uśrednieniem rang należy wprowadzić korektę w mianowniku współczynnika konkordancji ($S_{\max}$). W tym celu dla szeregów, w których występują rangi powiązane oblicza się wartość T_i:

$$T_i = \frac{1}{12}\sum_{j=1}^{k}\left(t_j^3 - t_j\right), \tag{8}$$

gdzie k oznacza liczbę grup posiadających tą samą rangę ($j = 1, 2, \ldots, k$) w i-tym szeregu, a symbol t_j liczbę identycznych rang wiązanych w danej grupie. W szeregu pokazanym w tabeli 2 występują dwie grupy rang o identycznych rangach wiązanych ($k = 2$); w pierwszej grupie są dwie identyczne rangi wiązane (obiekty C i D otrzymały rangę 3,5, czyli $t_1 = 2$), a w drugiej – trzy takie same rangi wiązane (obiekty A, G, I mają rangę 7, czyli $t_2 = 3$). Stąd:

$$T_1 = \frac{1}{12}\left(\left(2^3 - 2\right) + \left(3^3 - 3\right)\right) = 2{,}5.$$

Dla szeregu o uporządkowaniu mocnym (bez rang związanych) suma kwadratów odchyleń od średniej wynosi $\frac{n^3 - n}{12}$, czyli w naszym przypadku jest równa 60 ($n = 9$). Wielkość $\frac{n^3 - n}{12} - T_1$ jest sumą kwadratów odchyleń od średniej szeregu z rangami wiązanymi (w przykładzie: 60 – 2,5 = 57,5). Można bezpośrednio sprawdzić, że suma kwadratów odchyleń od średniej dla szeregu z rangami uśrednionymi (drugi wiersz tabeli 2) jest równa 57,5.

Dla m szeregów z rangami wiązanymi wartość T:

$$T = \sum_{i=1}^{m} T_i \tag{9}$$

oznacza poprawkę na rangi wiązane występujące we wszystkich szeregach. W przypadku pełnej zgodności między m szeregami rangi wiązane dotyczą tych samych obiektów, dlatego wielkość T mnożymy przez m. Ostatecznie współczynnik konkordancji dla rang powiązanych jest równy:

$$W = \frac{S}{S_{\max} - mT} = \frac{\sum_{j=1}^{n}\left(R_j - \frac{m(n+1)}{2}\right)^2}{\frac{1}{12}m^2\left(n^3 - n\right) - mT}. \tag{10}$$

4. UOGÓLNIONA FORMUŁA WSPÓŁCZYNNIKA KONKORDANCJI I TESTY ISTOTNOŚCI

Poniżej przedstawimy propozycję dogodniejszego sposobu obliczania współczynnika konkordancji, którą warto stosować zwłaszcza w przypadku występowania rang wiązanych:

$$W = \frac{S}{mG}, \tag{11}$$

gdzie

$$G = \sum_{i,j=1}^{m,n} \left(a_{ji} - \frac{(n+1)}{2} \right)^2. \tag{12}$$

Wielkość G jest sumą kwadratów odchyleń wszystkich rang od średniej szeregu. Z definicji średnie wszystkich szeregów są równe i wynoszą $\frac{(n+1)}{2}$, bez względu na występowanie rang powiązanych (i pod warunkiem, że rangi powiązane zostaną uśrednione).

Suma kwadratów odchyleń od średniej szeregu, w którym nie występują rangi wiązane wynosi $\frac{n^3 - n}{12}$. Mając m takich szeregów otrzymujemy:

$$mG = \frac{m^2 (n^3 - n)}{12} = S_{\max}. \tag{13}$$

W przypadku występowania rang wiązanych prawdziwe jest z kolei równanie:

$$mG = S_{\max} - mT. \tag{14}$$

Równanie powyższe pozwala na wyznaczenie T, bez konieczności zliczania ilości rang powiązanych, wyznaczania dla każdego szeregu T_i, a następnie sumowania tych wartości. Warto więc w praktyce korzystać z poniższej formuły:

$$T = \frac{S_{\max} - mG}{m} = \frac{\frac{1}{12} m^2 (n^3 - n) - mG}{m} = \frac{m(n^3 - n)}{12} - G. \tag{15}$$

Wyznaczanie wielkości T jest niezbędne do przeprowadzenia testu istotności dla szeregów z rangami wiązanymi, które przedstawiamy poniżej.

Zakładając, że eksperci w swych opiniach są niezależni, możemy uznać, że pojawienie się konkretnego układu uszeregowań jest równie prawdopodobne jak każdego innego. Na tej podstawie można zidentyfikować rozkład S. Dla danego m i n istnieje $(n!)^m$ wszystkich możliwych uszeregowań rangowych. Dla niskich wartości m i n opracowano tablice rzeczywistego rozkładu prawdopodobieństwa uzyskania określonej wartości S. I tak M. Kendall podaje rozkłady prawdopodobieństwa uzyskania wartości S dla następujących wielkości: a) $n = 3$; $2 \leq m \leq 10$ b) $n = 4$; $2 \leq m \leq 6$, c) $n = 5$, $m = 3$. Bardziej podręczne są tablice rozkładu S na poziomach istotności $\alpha = 0{,}05$ i 0,01, dla $3 \leq n \leq 7$ oraz $3 \leq m \leq 20$ ([9], a w polskiej literaturze: [12]).

Dla liczby obiektów większej od siedmiu ($n > 7$) zadowalającym przybliżeniem rzeczywistego rozkładu S jest rozkład chi-kwadrat. Dla uszeregowań mocnych (bez rang wiązanych) wartość statystyki:

$$\chi_r^2 = m(n-1)W = \frac{S}{\frac{1}{12}mn(n+1)}, \tag{16}$$

porównujemy z wartością χ_α^2 odczytaną z tablic rozkładu chi-kwadrat dla $n-1$ stopni swobody (*df*) oraz dla założonego poziomu istotności α. Jeżeli w szeregach preferencyjnych występują rangi wiązane, to wartość chi-kwadrat obliczamy na podstawie statystyki:

$$\chi_r^2 = \frac{S}{\frac{1}{12}mn(n+1) - \frac{1}{n-1}T}. \tag{17}$$

Testowanie istotności statystycznej współczynnika konkordancji polega na postawieniu hipotezy zerowej (H_0), która głosi że badane szeregi rangowe nie są ze sobą powiązane. Hipotezę zerową odrzucamy, gdy obliczona wartość χ_r^2 (lub wartość *S*, jeżeli korzystamy z tablic rzeczywistego rozkładu *S* dla $n \leq 7$) jest równa lub wyższa niż χ_α^2, czyli wartości odczytanej z tablic rozkładu chi-kwadrat dla $df = n-1$ stopni swobody przy danym poziomie istotności α.

Przykładowo dla szeregów podanych w tab. 1 wartość współczynnika *W* wyniosła 0,756. Mieliśmy tam do czynienia z oceną pięciu obiektów dokonaną przez trzech ekspertów i wszystkie szeregi miały porządek mocny. W takim wypadku najlepiej jest skorzystać z opracowanych tablic rzeczywistego rozkładu *S*, z których możemy odczytać, że dla $n = 5$ i $m = 3$ oraz dla $\alpha = 0{,}05$ wartość krytyczna *S* wynosi 64,4. Ponieważ rzeczywista wartość *S* w tym przykładzie wyniosła 68, więc przy $\alpha = 0{,}05$ hipotezę zerową należy odrzucić na korzyść stwierdzenia, że między badanymi szeregami istnieje statystycznie istotna zależność. Dla porównania z tablic podanych przez M. Kendalla można odczytać, że prawdopodobieństwo uzyskania wartości $S \geq 64$ wynosi 0,045.

5. INTERPRETACJA WYNIKÓW POMIARU

Celem oceny zgodności między szeregami preferencyjnymi jest ustalenie ostatecznego porządku rangowego. Ostateczne rangi są wyznaczane na podstawie średniej (lub sumy) rang dla badanych obiektów. Porządek taki możemy jednak ustalić dopiero po stwierdzeniu satysfakcjonującego poziomu zgodności między ekspertami (obserwatorami). Stąd właśnie wynika potrzeba pomiaru stopnia zgodności.

Odpowiedź na pytanie, jaka wartość współczynnika konkordancji świadczy o zgodności uporządkowań nie jest prosta. Wartość współczynnika *W* kształtuje się w przedziale od 0 do 1, gdzie 1 oznacza pełną zgodność uporządkowań, a 0 brak zgodności. Bardziej przemawiającym do intuicji są współczynniki korelacji, których wartości kształtują się od -1 do 1, gdzie wartości ujemne odzwierciedlają siłę niezgodności, wartości dodatnie siłę zgodności. Taka interpretacja jest prosta, gdy mamy do czynienia z dwoma szeregami (mówiąc szerzej dwiema zmiennymi). W miarę wzrostu liczby szeregów rośnie jednak ilość rang, które zostaną przypisane temu samemu obiektowi, co sprawia że maleje poziom maksymalnego możliwego nieuporządkowania między szeregami.

Zjawisko to jest wyraźnie widoczne, gdy prześledzimy zależność miedzy współczynnikiem konkordancji a współczynnikiem korelacji rang Spearmana (ρ). Ogólną ocenę zgodności między $m > 2$ szeregami preferencyjnymi można bowiem wyrazić poprzez obliczenie średniej współczynników korelacji rang ρ dla wszystkich możliwych par uporządkowań rangowych.

Powróćmy do przykładu (tab. 1), w którym trzech ekspertów oceniało pięć obiektów na skali porządkowej, a współczynniki konkordancji wyniósł 0,756. Wyrazimy teraz ogólną zgodność ekspertów za pomocą średniej arytmetycznej współczynników ρ. Jak wiadomo współczynniki korelacji rang Spearmana służy do oceny siły związku między dwiema cechami i jest obliczany wg wzoru:

$$\rho = 1 - \frac{6\sum_{j=1}^{n}(d_j)^2}{n^3 - n}, \tag{18}$$

gdzie d_j jest różnicą rang dla j-tego obiektu. W przykładzie z tab. 1 mamy 3 możliwe pary uszeregowań. Obliczmy $\rho_{1,2}$ (współczynnik korelacji miedzy szeregami eksperta 1 i 2), $\rho_{1,3}$ (między szeregiem eksperta 1 i 3) oraz $\rho_{2,3}$ (czyli między szeregiem 2 i 3), a następnie uśredniamy otrzymane wyniki. Średnia wszystkich możliwych par $\rho_{\alpha v}$ wynosi dla naszego przykładu:

$$\rho_{\alpha v} = \frac{\rho_{1,2} + \rho_{1,3} + \rho_{2,3}}{3} = \frac{0{,}6 + 0{,}7 + 0{,}6}{3} = 0{,}633.$$

Jak widać uzyskany wynik (0,633) jest wartością niższą od obliczonego wcześniej współczynnika W (0,756).

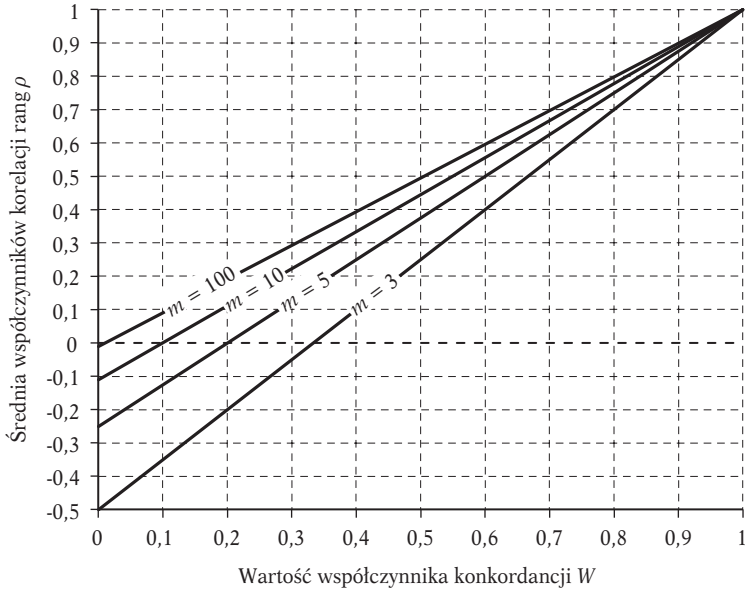
Istnieje relacja liniowa między współczynnikiem konkordancji W a średnią współczynników korelacji rang ρ Spearmana. Dla $\binom{m}{2}$ wszystkich par obserwatorów, a więc i szeregów preferencyjnych można wyznaczyć wartości ρ, a następnie obliczyć średnią dla wszystkich par[1].

Średnia współczynników korelacji rang Spearmana ($\rho_{\alpha v}$) jest powiązana ze współczynnikiem W następująco:

$$\rho_{\alpha v} = \frac{mW - 1}{m - 1}. \tag{19}$$

Gdy współczynnik zgodności $W = 0$, to średnia współczynników korelacji rang Spearmana wyniesie $\rho_{\alpha v} = -\frac{1}{m-1}$. Jest to najmniejsza wartość jaką może osiągnąć $\rho_{\alpha v}$ (w miarę wzrostu liczby szeregów, będzie się zbliżać do zera). Gdy $W = 1$ to $\rho_{\alpha v} = 1$. Z kolei dla $\rho_{\alpha v}$ równego zero, współczynnik W wyniesie $\frac{1}{m}$. Zależność między tymi współczynnikami dla $m = 3$, 5, 10 i 100 pokazano na rysunku 1. W przypadku analizy więcej niż 100 szeregów wartość $\rho_{\alpha v}$ można uznać za równą W (dla $m = 100$, wartości $W = 0$ odpowiada wartość $\rho_{\alpha v} = -0{,}01$).

[1] Jest to przewaga ρ nad τ, dla którego nie ma prostej metody interpretacji w kategoriach średniej z par $m > 2$ uszeregowań.



Rysunek 1. Relacje między W a ρ_{av} dla wybranych wartości m

Źródło: opracowanie własne.

Przedstawiona wyżej zależność, pokazuje problemy z właściwą interpretacją współczynnika konkordancji. Rozważmy następujący przykład. Grupa trzech ekspertów ustaliła rangi dla 20 czynników ryzyka w związku z realizacją inwestycji i współczynnik konkordancji wyniósł 0,6. Te same 20 czynników ryzyka dla tej samej inwestycji oceniła inna grupa, tym razem 7 ekspertów i współczynnik W wyniósł 0,8. Widać więc wyraźnie, że druga grupa jest bardziej zgodna w ocenie niż grupa pierwsza. Nie ma jednak jednoznacznej odpowiedzi na pytanie, czy różnica ta jest istotna statystycznie.

6. PRZYKŁAD POMIARU ZGODNOŚCI SZEREGÓW Z RANGAMI WIĄZANYMI

W badaniach mających na celu ustalenie wpływu tzw. gospodarki opartej na wiedzy (GOW) na funkcjonowanie polskich przedsiębiorstw jednym z etapów badawczych było opracowanie zestawu kryteriów oceny pozwalających na ocenę stopnia tego wpływu [4]. Z uwagi na fakt, iż pojęcie GOW jest nieostre uznano, że należy dobrać nie jedno, lecz zestaw kryteriów oceny stopnia takiego wpływu. Przejawem tego wpływu miał być poziom zaawansowania systemu zarządzania wiedzą w przedsiębiorstwie.

Na podstawie studiów literatury przedmiotu zidentyfikowano kilkadziesiąt najczęściej podawanych właściwości sytemu zarządzania wiedzą. Ostatecznie uznano, iż najczęściej opisywanymi cechami, które mogłyby stać się kryteriami oceny systemu zarządzania wiedzą w przedsiębiorstwie są: K1 – grupowe rozwiązywanie problemów, K2 – bariery w dzieleniu się wiedzą, K3 – częstotliwość aktualizacji baz danych, K4

– dzielenie się wiedzą z kooperantami, K5 – użyteczność systemów informatycznych, K6 – stopień zaawansowania systemów informacyjnych, K7 – narzędzia wspomagające zarządzanie wiedzą, K8 – znajomość technologii informatycznych, K9 – działalność badawczo-rozwojowa, K10 – współpraca w zakresie rozwoju z innymi firmami, K11 – poziom wykształcenia pracowników, K12 – szkolenia pracowników, K13 – stopień komputeryzacji stanowisk pracy.

Grupę 14 pracowników naukowo-dydaktycznych Uniwersytetu Ekonomicznego w Krakowie poproszono o dokonanie oceny stopnia istotności wymienionych kryteriów z uwagi na ocenę wpływu GOW na przedsiębiorstwo. Przyjęto, że kryteria będą oceniane w skali od 1 do 3, gdzie ocena 1 oznaczała umiarkowaną istotność, ocena 2 wysoką istotność, a ocena 3 bardzo wysoką istotność danego kryterium.

Wyniki przeprowadzonej oceny 13 kryteriów (K1 do K13) dokonanej przez 14 ekspertów (E1 do E14) zestawiono w tabeli 3.

Tabela 3

Surowe wyniki oceny istotności dobranych kryteriów

	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10	K11	K12	K13
E1	3	3	2	3	1	2	1	2	2	1	1	3	1
E2	1	1	1	1	3	3	3	2	2	2	3	2	1
E3	3	2	3	1	3	1	2	3	2	1	2	1	2
E4	2	3	1	2	2	2	3	1	2	1	3	3	1
E5	1	1	3	3	2	2	2	1	3	2	1	3	1
E6	2	2	2	2	2	3	2	2	3	3	2	3	2
E7	3	2	3	2	3	2	2	1	2	3	3	2	2
E8	3	2	2	2	2	3	3	2	3	2	1	1	1
E9	3	2	2	3	3	2	3	2	2	3	1	2	1
E10	3	3	2	2	1	2	2	2	3	2	2	1	2
E11	3	2	2	1	3	2	2	3	2	1	3	2	2
E12	1	3	3	1	2	3	3	2	3	2	3	2	2
E13	1	1	2	1	3	2	3	3	2	2	3	3	1
E14	2	1	2	1	2	2	3	2	3	1	3	2	2

Źródło: opracowanie własne.

Jak widać z zaprezentowanych w tabeli 3 danych mamy do czynienia z oceną 13 obiektów dokonaną przez 14 obserwatorów ($n = 13$, $m = 14$). Ze względu na przyjętą skalę oceny wszystkie szeregi mają rangi wiązane. W celu obliczenia współczynnika konkordancji konieczne jest przeprowadzenie procedury uśredniania rang wiązanych. I tak w szeregu pierwszym (E1) z tabeli 3 kryteria K1, K2, K4 oraz K12 otrzymały

najwyższą ocenę w przyjętej skali (3). Oznacza to, że pozycje tych kryteriów (obiektów) są sobie równe i jednocześnie najwyższe. Te cztery pierwsze miejsca uśredniamy: $\frac{1+2+3+4}{4}$ i otrzymujemy dla tych obiektów rangę uśrednioną 2,5. Dalej, kryteria K3, K6, K8 i K9 otrzymały tę samą ocenę (2), a zatem je uśredniamy biorąc pod uwagę ich kolejne pozycje w szeregu: $\frac{5+6+7+8}{4} = 6{,}5$. Ekspert 1 najniżej ocenił (1) kryteria K5, K7, K10, K11 oraz K13. Stąd rangi wiązane dla tych kryteriów wyniosą: $\frac{9+10+11+12+13}{5} = 11$. Podobnie obliczenia przebiegają dla pozostałych szeregów. Wyniki obliczeń pokazano w tabeli 4.

Tabela 4

Rangi uśrednione dobranych kryteriów

	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10	K11	K12	K13	t_i
E1	2,5	2,5	6,5	2,5	11	6,5	11	6,5	6,5	11	11	2,5	11	20
E2	11	11	11	11	2,5	2,5	2,5	6,5	6,5	6,5	2,5	6,5	11	20
E3	2,5	7	2,5	11,5	2,5	11,5	7	2,5	7	11,5	7	11,5	7	20
E4	7	2,5	11,5	7	7	7	2,5	11,5	7	11,5	2,5	2,5	11,5	20
E5	11	11	2,5	2,5	6,5	6,5	6,5	11	2,5	6,5	11	2,5	11	20
E6	9	9	9	9	9	2,5	9	9	2,5	2,5	9	2,5	9	65
E7	3	9	3	9	3	9	9	13	9	3	3	9	9	38
E8	2,5	7,5	7,5	7,5	7,5	2,5	2,5	7,5	2,5	7,5	12	12	12	24,5
E9	3	8,5	8,5	3	3	8,5	3	8,5	8,5	3	12,5	8,5	12,5	28
E10	2	2	7,5	7,5	12,5	7,5	7,5	7,5	2	7,5	7,5	12,5	7,5	44,5
E11	2,5	8	8	12,5	2,5	8	8	2,5	8	12,5	2,5	8	8	33,5
E12	12,5	3,5	3,5	12,5	9	3,5	3,5	9	3,5	9	3,5	9	9	28
E13	11,5	11,5	7,5	11,5	3	7,5	3	3	7,5	7,5	3	3	11,5	20
E14	7	12	7	12	7	7	2	7	2	12	2	7	7	32
Suma	**87**	**105**	**95,5**	**119**	**86**	**90**	**77**	**105**	**75**	**111,5**	**89**	**97**	**137**	**413,5**

Źródło: opracowanie własne.

W ostatnim wierszu tabeli 4 podano sumę rang dla poszczególnych kryteriów (czyli R_j) oraz sumę T_i, czyli poprawkę na rangi wiązane (T, wzór 9). Suma kwadratów odchyleń R_j od średniej $\bar{R} = \frac{m(n+1)}{2}$ (czyli S) wynosi 3629,5. Tak przygotowane dane pozwalają na obliczenie współczynnika konkordancji (wzór 10):

$$W = \frac{\sum_{j=1}^{n}\left(R_j - \frac{14(13+1)}{2}\right)^2}{\frac{1}{12}\cdot 14^2(13^3-13) - 14\cdot 413{,}5} = \frac{3629{,}5}{3572-5789} = 0{,}121$$

W praktyce warto stosować zaproponowaną modyfikację w obliczaniu współczynnika konkordancji, dzięki której unikamy konieczności wyznaczania wartości poprawek na rangi wiązane (ostatnia kolumna tab. 4). Na podstawie tablicy rang uśrednionych (tab. 4) obliczmy:

$$G = \sum_{i=1,j=1}^{m,n}\left(a_{ji} - \frac{(n+1)}{2}\right)^2 = \sum_{i=1,j=1}^{m,n}(a_{ji}-7)^2 = 2134{,}5,$$

a stąd

$$W = \frac{S}{mG} = \frac{3629{,}5}{14\cdot 2134{,}5} = 0{,}121.$$

Korzystając z formuły (15) wyznaczamy wartość T, czyli sumę poprawek na rangi:

$$T = \frac{m(n^3-n)}{12} - G = \frac{14(13^3-13)}{12} - 2134{,}5 = 413{,}5.$$

Średnia współczynników korelacji rang Spearmana dla omawianego przypadku wynosi:

$$\rho_{\alpha v} = \frac{mW-1}{m-1} = \frac{14\cdot 0{,}121-1}{13} = 0{,}053.$$

W celach porównawczych obliczono średnią współczynników korelacji liniowej Pearsona ($r_{\alpha v}$). Dla wszystkich 91 par szeregów średnia ta wyniosła $r_{\alpha v} = 0{,}051$. Jak widać różnica między $r_{\alpha v}$ a $\rho_{\alpha v}$ jest nieznaczna. Wartości współczynników na podstawie których obliczono średnią zawiera tabela 5.

Tabela 5

Współczynniki korelacji *r* Pearsona dla wszystkich par szeregów

	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13
E2	–0,57												
E3	–0,12	–0,12											
E4	0,24	0,36	–0,25										
E5	0,22	–0,01	–0,36	0,00									
E6	0,06	0,26	–0,64	0,00	0,46								

cd. tabeli 5

	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13
E7	–0,26	0,05	0,16	0,00	0,05	–0,06							
E8	0,14	0,14	0,13	0,00	0,14	0,16	–0,05						
E9	0,16	0,03	0,00	0,00	0,30	0,02	0,20	0,57					
E10	0,31	–0,44	0,16	0,00	–0,29	–0,08	–0,06	0,50	–0,04				
E11	–0,12	0,30	0,74	0,15	–0,54	–0,41	0,07	–0,02	–0,24	–0,03			
E12	–0,35	0,43	0,00	0,27	0,04	0,18	–0,04	0,10	–0,45	0,12	0,06		
E13	–0,44	0,79	0,12	0,24	0,12	0,14	–0,05	–0,14	–0,03	–0,62	0,40	0,35	
E14	–0,41	0,55	0,29	0,29	0,00	0,00	0,00	0,16	–0,33	0,00	0,51	0,47	0,55

Źródło: opracowanie własne.

Uzyskanie wyniki świadczą o niskiej zgodność między opiniami ekspertów co do relatywnej istotności rang dla badanych kryteriów. Czy uzyskany wynik jest istotny statystycznie? W przykładzie mamy do czynienia z oceną 13 obiektów dokonaną przez 14 ekspertów. Wszystkie szeregi posiadają rangi wiązane. Dla takiej liczby obiektów i szeregów nie ma tablic rzeczywistego rozkładu S, więc musimy posłużyć się szacunkiem korzystając z rozkładu chi-kwadrat. Obliczamy wartość statystyki (wzór 17):

$$\chi_r^2 = \frac{3629{,}5}{\frac{14 \cdot 13(13+1)}{12} - \frac{413{,}5}{13-1}} = 24{,}599.$$

Z tabel rozkładu chi-kwadrat odczytujemy, że dla $df = 12$ (czyli $n - 1$) i poziomu istotności $\alpha = 0{,}05$ wartość $\chi_\alpha^2 = 21{,}0261$, a zatem odrzucamy hipotezę zerową mówiącą o braku zależności między badanymi uporządkowaniami $\left(\chi_r^2 > \chi_\alpha^2\right)$.

Uwzględnianie wartości T dla szeregów z dużą ilością rang wiązanych jest konieczne w testach istotności. Dla rozpatrywanego przykładu wartość χ_r^2 bez wprowadzenia poprawki na rangi wiązane $\left(\frac{T}{n-1}\right)$ wnosi 19,9423.

8. OGRANICZENIA METOD KORELACJI RANG W POMIARZE ZGODNOŚCI OCEN

W przypadku większej liczby obiektów ($n > 9$) rangowanie w sensie dosłownym stwarza istotne trudności, ponieważ oceniający nie jest w stanie jednocześnie porównywać ze sobą tak dużej liczby obiektów. Sensownym rozwiązaniem jest wtedy odwołanie się do metody porównań parami i na tej podstawie tworzenie porządku rangowego[2].

[2] Takie rozwiązanie sugeruje m.in. M. Kendall. Proponowany przez tego autora miernik zgodności u (*coefficient of agreement*) wyraża stopień zgodności między ocenami obserwatorów, którzy rozważane obiekty porównują parami. W przypadku złożonych i trudnych do oceny obiektów podejście takie wydaje się bardzo wskazane [7].

Zamiast tego w praktyce stosuje się skale oceny o mniejszej rozpiętości, które służą do oceny każdego obiektu z osobna, tak jak to zrobiono w powyższym przykładzie.

Pytanie, jakie w związku z tym się nasuwa to, czy mamy prawo przekształcania ocen wyrażanych w skali punktowej (tutaj od 1 do 3) w szeregi z rangami wiązanymi, zwłaszcza gdy liczba ocenianych obiektów jest większa do przyjętej rozpiętości skali punktowej (w naszym przypadku liczba ocenianych obiektów była równa 13, czyli znacznie większa od 3).

Rozważmy inny przykład. Czterech ekspertów ocenia stopień wpływu na działalność danego przedsiębiorstwa siedmiu określonych zagrożeń. Czy proces oceny polegający na uszeregowaniu tych zagrożeń w kolejności od zagrożenia o największym wpływie aż do zagrożenia o wpływie najmniejszym (czyli rangowanie), będzie równoznaczny z procesem oceny każdego z tych zagrożeń oddzielnie w skali od 1 do 7? Jaki wpływ na proces oceny będzie miało zwiększenie rozpiętości skali punktowej (np. skala od 1 do 9 zamiast od 1 do 7)? Jaki wpływ na proces oceny będzie miało zmniejszenie rozpiętości skali (np. skala od 1 do 3 zamiast od 1 do 7)? Odpowiedzi na powyższe pytania mają fundamentalne znaczenie dla pomiaru zgodności ocen.

Gdy zgodność ocen chcemy wyrazić za pomocą współczynników korelacji rang (dla $m = 2$) lub współczynnika konkordancji (dla $m > 2$), to musimy mieć świadomość, że adekwatny pomiar zgodności będzie możliwy wówczas, gdy mamy do czynienia z rangowaniem obiektów w ścisłym tego słowa znaczeniu. Wydaje się, że rangowanie jest procesem psychologicznie i merytorycznie odmiennym od niezależnej oceny każdego obiektu z osobna – nawet gdy te obiekty należą do jednej klasy (kategorii).

Współczynniki korelacji rang (ρ, τ, czy współczynnik konkordancji) wyrażają stopień ogólnego powiązania między preferencjami grupy ekspertów. Jednak potrzeba pomiaru stopnia zgodności ocen dotyczy nie tylko sytuacji, w których kilku ekspertów wypowiada się na temat kilku obiektów łącznie. Często spotykamy się z praktyką, kiedy oceny są przyznawane niezależnie każdemu z badanych obiektów w określonej skali punktowej. Nie mamy wtedy do czynienia z rangowaniem w sensie ścisłym, ponieważ oceniając każdy z obiektów osobno eksperci mają do dyspozycji każdorazowo skalę o tej samej rozpiętości.

W przypadku współczynnika konkordancji badamy stopień powiązań między rangami. Ranga oznacza miejsce danego obiektu wśród innych ocenianych obiektów, a zatem stosowana w tym przypadku rozpiętość skali porządkowej jest równa $n - 1$. Wyznaczając rangę danego obiektu, każdorazowo zmniejszamy tak rozumianą rozpiętość skali, aż do jej wyczerpania. Dla szeregów słabych, czyli gdy dopuszczamy możliwość nadania tej samej rangi niektórym obiektom, rozpiętość skali będzie mniejsza niż $n - 1$.

Wydaje się, że skala punktowa, za pomocą której n obiektów jest ocenianych niezależnie i która ma rozpiętość wyższą od $n - 1$, jest skalą mocniejszą od skali wynikającej z procesu rangowania. Zależność odwrotna jest mniej oczywista, lecz można w pewnym sensie powiedzieć, że gdy rozpiętość skali punktowej ocen jest mniejsza do $n - 1$, to rangowanie – o ile nie stosuje się rang wiązanych – daje wyniki „mocniejsze" od ocen punktowych.

Uniwersytet Ekonomiczny w Krakowie

LITERATURA

[1] Brzeziński J., Maruszewski T., [1978], *Metoda sędziów kompetentnych i jej zastosowanie w badaniach pedagogicznych*, „Kwartalnik Pedagogiczny", Nr 1(87).
[2] Damodaran A., [2009], *Ryzyko strategiczne*, Wydawnictwa Akademickie i Profesjonalne, Warszawa.
[3] Day S.G., Schoemaker P.J.H., [2006], *Peripheral Vision. Detecting Weak Signals That Will Make or Break Your Company*, Harvard Business School Press, Boston.
[4] *Doskonalenie struktur organizacyjnych przedsiębiorstw w gospodarce opartej na wiedzy*, [2009], pod red. A. Stabryły, Wydawnictwo C.H. Beck, Warszawa.
[5] Góralski A., [1976], *Metody opisu i wnioskowania statystycznego w psychologii*, PWN, Warszawa.
[6] Grouard B., Meston F., [1997], *Kierowanie zmianami w przedsiębiorstwie*, Poltex, Warszawa.
[7] Kendall M., [1962], *Rank correlation methods*, Charles Griffin & Company, London.
[8] Sharpe N.R., Veaux De R.D., Valleman P.F., [2009], *Business Statistics*, Pearson, New York.
[9] Siegel S., [1956], *Nonparametric statistics for the behavioral sciences*, McGraw-Hill Company Inc., New York.
[10] Simons R., [2005], *Czy wiesz jak duże ryzyko ukryte jest w twojej firmie?*, „Harvard Business Review Polska", kwiecień.
[11] Stabryła A., [2000], *Zarządzanie strategiczne w teorii i praktyce firmy*, Wydawnictwo Naukowe PWN, Warszawa
[12] Steczkowski J, Zeliaś A., [1997], *Metody statystyczne w badaniach zjawisk jakościowych*, Wydawnictwo Akademii Ekonomicznej w Krakowie, Kraków.
[13] *The Delphi Method. Techniques and Applications*, [1975], pod red. H. Linstonea I M. Turoffa, Addiosn-Wesley Publishing Company Inc. London.
[14] Winer B.J., [1962], *Statistical principles in experimental design*, McGraw-Hill Company Inc., New York.

Praca wpłynęła do redakcji w marcu 2010 r.

ZASTOSOWANIE WSPÓŁCZYNNIKA KONKORDANCJI W POMIARZE ZGODNOŚCI OCEN EKSPERTÓW

Streszczenie

Celem artykułu jest przedstawienie zagadnienia pomiaru zgodności ocen wyrażonych na skali porządkowej. Istnieje szereg sposobów sprawdzenia stopnia zgodności opinii ekspertów. Z reguły do tego celu stosuje się współczynniki korelacji rang. W artykule skoncentrowano się na przypadku gdzie $m > 2$ ekspertów ocenia n obiektów, z dopuszczeniem rang wiązanych. Dla tak sformułowanego problemu jako miarę zgodności przyjęło się stosować współczynnik konkordancji W Kendalla. W artykule zaproponowano uogólnioną wersję obliczania tego współczynnika, która upraszcza obliczenia w przypadku występowania rang wiązanych. Oprócz przedstawienia procedury badania istotności statystycznej, podniesiono także problem interpretacji otrzymywanych wartości współczynnika konkordancji. W nawiązaniu do zaprezentowanego przypadku podjęto również dyskusję na temat ograniczeń zastosowań współczynników korelacji w ocenie zgodności ekspertów.

Słowa kluczowe: opnie ekspertów, rangi, rangi wiązane, metoda uśredniania rang, korelacja rang, współczynnik konkordancji

USING THE CONCORDANCE COEFFICIENT IN THE MEASUREMENT OF AGREEMENT AMONG EXPERTS

Summary

In this article the problem of the measurement of agreement among experts is discussed. There are several tools to verify whether opinions expressed by an ordinal scale are reliable. If the objects are ranked, one of correlation coefficients is chosen. At the beginning, the case of $m > 2$ experts ranking n objects is presented; first, using untied ranks and second, tied ranks. To this end Kendall's coefficient of concordance is applied. The method of examination of the statistical significance is also presented. Additionally the issue of the interpretation of the concordance coefficient in reference to the average of Spearman's ρ coefficients is raised. The paper provides an example of using concordance coefficient when ranks are tied and discusses limitations of using rank correlation methods as a tool for evaluating the degree of agreement among experts' opinions.

Key words: experts' opinions, ranks, tied ranks, mid-rank method, rank correlation, concordance coefficient,