AKADEMIA EKONOMICZNA W POZNANIU

STATYSTYKA REGIONALNA

w służbie samorządu lokalnego i biznesu

Redaktor naukowy: Jan Paradysz

INTERNETOWA OFICYNA WYDAWNICZA
CENTRUM STATYSTYKI REGIONALNEJ
POZNAŃ 2002

Projekt okładki
Eryk Grześkowiak

Redakcja i korekta
Marcin Szymkowiak

ISBN 83-907538-0-5

INTERNETOWA OFICYNA WYDAWNICZA
CENTRUM STATYSTYKI REGIONALNEJ
ul. Towarowa 53, 60-896 Poznań, tel. 0 prefix 61 854 39 43
e-mail: panorama@csr.pl

SPIS TREŚCI

Wstęp (**Jan Paradysz**) 5

Część pierwsza
TEORIA I PRAKTYKA BADAŃ REPREZENTACYJNYCH

Janusz Wywiał, On accuracy of mean estimation on the basis of two-phase sampling for stratification 15

Marcin Skibicki, Janusz Wywiał, On optimal sample allocation in strata 29

Jan Kordos, Nowy projekt zastosowania estymacji dla małych obszarów w krajach europejskich 38

Grażyna Dehnel, Rola średnich i dużych przedsiębiorstw w ekonomicznym rozwoju regionu w świetle statystyki małych obszarów 51

Jan Kubacki, Wybrane metody statystyki małych obszarów zastosowane w szacowaniu bezrobocia dla powiatów 58

Tomasz Jurkiewicz, Badanie efektywności estymatorów proporcji dla małych domen za pomocą eksperymentów symulacyjnych 75

Część druga
KLASYFIKACJA I ANALIZA DANYCH W UKŁADACH REGIONALNYCH

Jarosław Lira, Wiesław Wagner, Feliks Wysocki, Mediana w zagadnieniach porządkowania obiektów wielocechowych 87

Aleksandra Łuczak, Feliks Wysocki, Metody klasyfikacji w badaniu zróżnicowania funkcjonalnego obszarów wiejskich 100

Marek Walesiak, Propozycja uogólnionej miary odległości w statystycznej analizie wielowymiarowej 115

Michal Woźniak, Kazimierz Zając, Zbigniew Ziolo, Pozycja województwa małopolskiego w strukturze regionalnej Polski 122

Wiaczeslaw Sobieszczański, Zróżnicowanie warunków mieszkaniowych w województwie wielkopolskim 131

Część trzecia
PROMOCJA „MAŁYCH OJCZYZN"

Piotr Jankowski, GIS Technology in Support of Participatory Spatial Decision Making 145

Jan Paradysz, Internetowa panorama miast, gmin i regionów jako informacyjne wspomaganie samorządu lokalnego 153

Zbigniew Zwoliński, Internetowa wizualizacja statystyki regionalnej 169
Tomasz Klimanek, Zasoby mapy numerycznej Geo-Info w promocji „Małych Ojczyzn" 199
Piotr Błażczak, Przekształcenia ekonomiczne i ich znaczenie w rekultywacji wsi 211
Mieczysław Sobczyk, Regionalne zróżnicowanie degradacji gleb w Polsce..... 226

Część czwarta
NOWE I POTENCJALNE ŹRÓDŁA ZASILANIA INFORMACYJNEGO W STATYSTYCE REGIONALNEJ

Marek Obrębalski, Danuta Strahl, Potrzeby informacyjne biznesu w świetle badań ankietowych........ 241
Daria Marcinowicz, Ocena źródeł w statystyce ludności aglomeracji wielkomiejskiej 250
Reinhard Mummelthey, Umlandwanderungen in Berlin im Vergleich zu ausgewählten Deutschen und Polnischen Großstädten 257
Aleksandra Witkowska, Marek Witkowski, Propozycja wykorzystania danych z różnych źródeł w SMO podmiotów gospodarczych........ 276

Część piąta
AKTYWNOŚĆ EKONOMICZNA LUDNOŚCI W UJĘCIU REGIONALNYM

Bogumiła Pleskowicz, Źródła informacji o aktywności zawodowej ludności w ujęciu regionalnym........ 289
Elżbieta Gołata, Estymacja bezrobocia w przekroju powiatów województwa wielkopolskiego........ 300
Edyta Mazurek, Stanisława Ostasiewicz, Nieparametryczny model trwania bezrobocia 326
Dorota Ziemba, Metody weryfikacji hipotezy rynków 338

Literatura 353

Indeks osobowy 366

Indeks rzeczowy........ 369

Uczestnicy konferencji........ 371

Marek Walesiak

PROPOZYCJA UOGÓLNIONEJ MIARY ODLEGŁOŚCI W STATYSTYCZNEJ ANALIZIE WIELOWYMIAROWEJ[1]

Abstract

A generalised correlation coefficient is well known in statistical literature (M.G. Kendall and W.R Buckland, 1986). This general definition includes Kendall's and Pearson product-moment correlation as particular cases. In the article a proposal of generalised distance measure is shown. The construction of this distance measure based on the idea of generalised correlation coefficient and Kendall's coefficient.

Keywords: Measurement Scales, Distance Measures, Data Analysis

1. Wprowadzenie

Stosowanie konkretnych konstrukcji miar korelacji i odległości w statystycznej analizie wielowymiarowej jest uzależnione od skal pomiaru zmiennych. W teorii pomiaru rozróżnia się 4 podstawowe skale pomiaru, wprowadzone przez Stevensa (por. S.S. Stevens, 1959), uporządkowane od najsłabszej do najmocniejszej: nominalna, porządkowa (rangowa), przedziałowa (interwałowa), ilorazowa (stosunkowa). Podstawowe własności skal pomiaru zawiera tab. 1.

Typ skali, ze względu na dopuszczalne przekształcenia, determinuje stosowalność rozmaitych technik statystyczno-ekonometrycznych. Technikami statystycznymi dopuszczalnymi dla danego typu skali są takie techniki, które dostarczają wyników (w sensie relacji) niezmiennych względem dopuszczalnych przekształceń (por. K. Walenta, 1971). W artykule Handa (D.J. Hand, 1996) dyskutowany jest problem relacji między skalami pomiaru a dopuszczalnymi dla nich technikami statystycznymi. Pokazano w nim przykłady, które są źródłem kontrowersji w przypadku ścisłego stosowania reguł pomiaru.

---

[1] Pracę wykonano w ramach projektu badawczego Nr 1 H02B 011 16 finansowanego przez Komitet Badań Naukowych w latach 1999-2000.

2. Uogólniony współczynnik korelacji

Uogólniony współczynnik korelacji między dwiema zmiennymi przyjmuje postać (por. M.G. Kendall i W.R. Buckland, 1986; M.G. Kendall, 1955):

$$\Gamma_{jh} = \frac{\sum_{i=2}^{n}\sum_{k=1}^{i-1} a_{ik} b_{ik}}{\left[\sum_{i=2}^{n}\sum_{k=1}^{i-1} a_{ik}^2 \sum_{i=2}^{n}\sum_{k=1}^{i-1} b_{ik}^2\right]^{\frac{1}{2}}}, \tag{1}$$

gdzie: $i, k = 1, ..., n$ – numer obiektu; j, h – numer zmiennej.

Tab. 1. Podstawowe własności skal pomiaru

Typ skali	Dozwolone przekształcenia matematyczne	Dopuszczalne relacje	Dopuszczalne operacje arytmetyczne
Nominalna	$z = f(x), f(x)$ – dowolne przekształcenie wzajemnie jednoznaczne	równości $(x_A = x_B)$, różności $(x_A \neq x_B)$	zliczanie zdarzeń (liczba relacji równości, różności)
Porządkowa	$z = f(x), f(x)$ – dowolna ściśle monotonicznie rosnąca funkcja	powyższe oraz większości $(x_A > x_B)$ i mniejszości $(x_A < x_B)$	zliczanie zdarzeń (liczba relacji równości, różności, większości, mniejszości)
Przedziałowa	$z = bx + a$ $(b>0)$, $z \in R$ dla wszystkich x zawartych w R. Wartość zerowa na tej skali jest zwykle przyjmowana arbitralnie lub na podstawie konwencji (por. Ackoff [1969], s. 240).	powyższe oraz równości różnic i przedziałów $(x_A - x_B = x_C - x_D)$	powyższe oraz dodawanie i odejmowanie
Ilorazowa	$z = bx + a$ $(b>0)$, $z \in R_+$ dla wszystkich x zawartych w R_+. Naturalnym początkiem skali ilorazowej jest wartość zerowa (zero lewostronnie ogranicza zakres skali).	powyższe oraz równości ilorazów $\left(\frac{x_A}{x_B} = \frac{x_C}{x_D}\right)$	powyższe oraz mnożenie i dzielenie

Źródło: Opracowano na podstawie: (S.S Stevens, 1959; E.W. Adams, R.F. Fagot i R.E. Robinson, 1965; M. Walesiak, 1995).

Dla obserwacji na zmiennych $(x_{1j}, ..., x_{nj})$, $(x_{1k}, ..., x_{nk})$ mierzonych na skali ilorazowej i (lub) przedziałowej stosując we wzorze (1) podstawienie:

$$\begin{aligned} a_{ik} &= (x_{ij} - x_{kj}) \\ b_{ik} &= (x_{ih} - x_{kh}) \end{aligned}, \tag{2}$$

otrzymuje się współczynnik korelacji liniowej Pearsona (gdzie: x_{ij}, x_{kj} (x_{ih}, x_{kh}) – i-ta, k-ta obserwacja na j-tej (h-tej) zmiennej). Współczynnik korelacji liniowej Pearsona przybiera wartości z przedziału [–1; 1]. Wartość 0 oznacza, że między zmiennymi nie występuje korelacja, natomiast wartości graniczne odpowiadają doskonałej korelacji ujemnej lub dodatniej.

Wzór na współczynnik korelacji liniowej Pearsona można przedstawić inaczej jako:

$$r_{jh} = \frac{n\sum_{i=1}^{n} x_{ij}x_{ih} - \sum_{i=1}^{n} x_{ij}\sum_{i=1}^{n} x_{ih}}{\left\{\left[n\sum_{i=1}^{n} x_{ij}^2 - \left(\sum_{i=1}^{n} x_{ij}\right)^2\right]\left[n\sum_{i=1}^{n} x_{ih}^2 - \left(\sum_{i=1}^{n} x_{ih}\right)^2\right]\right\}^{0,5}}. \tag{3}$$

Dowód. (zob. M.G. Kendall, 1955; M. Walesiak, 1993)

$$\sum_{i=2}^{n}\sum_{k=1}^{i-1}(x_{ij}-x_{kj})(x_{ih}-x_{kh}) = \sum_{i=2}^{n}\sum_{k=1}^{i-1} x_{ij}x_{ih} - \sum_{i=2}^{n}\sum_{k=1}^{i-1} x_{ij}x_{kh} - \sum_{i=2}^{n}\sum_{k=1}^{i-1} x_{kj}x_{ih} + \sum_{i=2}^{n}\sum_{k=1}^{i-1} x_{kj}x_{kh} =$$

$$= (n-1)\sum_{k=1}^{n} x_{ij}x_{ih} - \left(\sum_{k=1}^{n}\sum_{i=1}^{n} x_{ij}x_{kh} - \sum_{i=1}^{n} x_{ij}x_{ih}\right) = n\sum_{i=1}^{n} x_{ij}x_{ih} - \sum_{i=1}^{n} x_{ih}\sum_{i=1}^{n} x_{ij};$$

$$\sum_{i=2}^{n}\sum_{k=1}^{i-1}(x_{ij}-x_{kj})^2 = \sum_{i=2}^{n}\sum_{k=1}^{i-1} x_{ij}^2 - \sum_{i=2}^{n}\sum_{k=1}^{i-1} x_{ij}x_{kj} - \sum_{i=2}^{n}\sum_{k=1}^{i-1} x_{kj}x_{ij} + \sum_{i=2}^{n}\sum_{k=1}^{i-1} x_{kj}^2 =$$

$$= (n-1)\sum_{i=1}^{n} x_{ij}^2 - \left(\sum_{i=1}^{n}\sum_{k=1}^{n} x_{ij}x_{kj} - \sum_{i=1}^{n} x_{ij}^2\right) = n\sum_{i=1}^{n} x_{ij}^2 - \sum_{i=1}^{n}\sum_{k=1}^{n} x_{ij}x_{kj} = n\sum_{i=1}^{n} x_{ij}^2 - \left(\sum_{i=1}^{n} x_{ij}\right)^2.$$

Dla obserwacji na zmiennych (x_{1j}, ..., x_{nj}), (x_{1h}, ..., x_{nk}) mierzonych na skali porządkowej stosując we wzorze (1) podstawienie:

$$a_{ik}(b_{ik}) = \begin{cases} 1 & \text{jeżeli} \quad x_{ij} > x_{kj}\ (x_{ih} > x_{kh}) \\ 0 & \text{jeżeli} \quad x_{ij} = x_{kj}\ (x_{ih} = x_{kh}) \\ -1 & \text{jeżeli} \quad x_{ij} < x_{kj}\ (x_{ih} < x_{kh}) \end{cases} \tag{4}$$

otrzymuje się współczynnik τ Kendalla (por. M.G. Kendall, 1955). Współczynnik korelacji τ Kendalla przybiera wartości z przedziału [–1; 1]. Wartość 1 oznacza pełną zgodność uporządkowań, natomiast wartość −1 pełną ich przeciwstawność. Współczynnik korelacji τ Kendalla jest więc dla wyników pomiaru porządkowego szczególną postacią współczynnika korelacji liniowej Pearsona (por. M. Walesiak, 1991).

2. Propozycja uogólnionej miary odległości

Wykorzystanie niektórych metod statystycznej analizy wielowymiarowej (metody klasyfikacji, skalowanie wielowymiarowe, metody porządkowania liniowego) wymaga sformalizowania pojęcia odległości obiektów. Funkcja $d: A \times A \to R$ (A – zbiór obiektów badania, R – zbiór liczb rzeczywistych) będzie nazywana miarą odległości wtedy i tylko wtedy, gdy spełnione są warunki:

- nieujemności: $d_{ik} \geq 0$ dla $i, k = 1, \ldots, n$;
- zwrotności: $d_{ik} = 0 \Leftrightarrow i = k$ dla $i, k = 1, \ldots, n$;
- symetryczności: $d_{ik} = d_{ki}$ dla $i, k = 1, \ldots, n$.

Wykorzystanie jako miary odległości uogólnionego współczynnika korelacji (obejmującego współczynnik korelacji liniowej Pearsona i współczynnik korelacji τ Kendalla) nie jest możliwe z uwagi na nie spełnienie warunku nieujemności i zwrotności. Przekształcenie $d_{ik} = (1 - \Gamma_{ik})/2$ wprawdzie zapewnia spełnienie warunku nieujemności (wartości d_{ik} zawarte są w przedziale [0; 1]), jednak nadal nie jest spełniony warunek zwrotności.

W konstrukcji uogólnionej miary odległości wykorzystano ideę współczynnika korelacji zmiennych porządkowych τ Kendalla oraz uogólnionego współczynnika korelacji. Proponowana uogólniona miara odległości przyjmuje postać: (5)

$$d_{ik} = (1 - s_{ik})/2 = \frac{1}{2} - \frac{\sum_{j=1}^{m} a_{ikj} b_{kij} + \sum_{j=1}^{m} \sum_{\substack{l=1 \\ l \neq i,k}}^{n} a_{ilj} b_{klj}}{2\left[\left(\sum_{j=1}^{m} a_{ikj}^2 + \sum_{j=1}^{m} \sum_{\substack{l=1 \\ l \neq i,k}}^{n} a_{ilj}^2\right)\left(\sum_{j=1}^{m} b_{kij}^2 + \sum_{j=1}^{m} \sum_{\substack{l=1 \\ l \neq i,k}}^{n} b_{klj}^2\right)\right]^{\frac{1}{2}}},$$

gdzie: $i, k, l = 1, \ldots, n$ – numer obiektu,
$j = 1, \ldots, m$ – numer zmiennej,
x_{ij} (x_{kj}, x_{lj}) – i-ta (k-ta, l-ta) obserwacja na j-tej zmiennej.

Dla zmiennych mierzonych na skali ilorazowej i (lub) przedziałowej w formule (5) stosowane jest podstawienie:

$$\begin{aligned} a_{ipj} &= x_{ij} - x_{pj} \quad dla \quad p = k,l \\ b_{krj} &= x_{kj} - x_{rj} \quad dla \quad r = i,l \end{aligned} . \tag{6}$$

Zasób informacji skali porządkowej jest nieporównanie mniejszy. Jedyną dopuszczalną operacją empiryczną na skali porządkowej jest zliczanie zdarzeń (tzn. wyznaczanie liczby relacji większości, mniejszości i równości). W konstrukcji miernika odległości musi być wykorzystana informacja o relacjach w jakich pozostają porównywane obiekty w stosunku do pozostałych obiektów ze zbioru A. Dla zmiennych mierzonych na skali porządkowej w formule (5) stosuje się podstawienie (por. M. Walesiak, 1993; M. Walesiak, J. Dziechciarz i A. Bąk, 1998):

$$a_{ipj}(b_{krj}) = \begin{cases} 1 & jeżeli \quad x_{ij} > x_{pj} \; (x_{kj} > x_{rj}) \\ 0 & jeżeli \quad x_{ij} = x_{pj} \; (x_{kj} = x_{rj}) \\ -1 & jeżeli \quad x_{ij} < x_{pj} \; (x_{kj} < x_{rj}) \end{cases} \text{ dla } p = k,\ l;\ r = i,\ l; \tag{7}$$

Wtedy w mianowniku wzoru (5) pierwszy czynnik oznacza liczbę relacji większości i mniejszości określoną dla obiektu *i*, czynnik drugi zaś liczbę relacji większości i mniejszości określoną dla obiektu *k*.

Miara odległości d_{ik} :

- może być stosowana w sytuacji, gdy obiekty opisane są zmiennymi mierzonymi na skali ilorazowej, przedziałowej lub porządkowej,
- przybiera wartości z przedziału [0; 1]. Wartość 0 oznacza, że dla porównywanych obiektów *i*, *k* między odpowiadającymi sobie obserwacjami na zmiennych zachodzą tylko relacje równości. W przypadku podstawienia (7) wartość 1 oznacza, że gdy dla porównywanych obiektów *i*, *k* między odpowiadającymi sobie obserwacjami na zmiennych porządkowych zachodzą tylko relacje większości (mniejszości) lub relacje większości (mniejszości) oraz relacje równości jeżeli relacje te są zachowane w stosunku do pozostałych obiektów (a więc obiektów o numerach $l = 1, ..., n$; gdzie $l \neq i, k$);
- spełnia warunki: nieujemności $d_{ik} \geq 0$, zwrotności $d_{ii} = 0$, symetryczności $d_{ik} = d_{ki}$ (dla wszystkich $i, k = 1,..., n$),
- nie zawsze spełnia warunek nierówności trójkąta (potwierdziły ten wniosek przeprowadzone analizy symulacyjne),

– istnieje przynajmniej jedna para obiektów w zbiorze badanych obiektów A, dla której obserwacje na zmiennych nie są identyczne (dla uniknięcia zera w mianowniku d_{ik}),
– nie zmienia wartości w wyniku transformacji wartości zmiennych za pomocą dozwolonego na danej skali przekształcenia matematycznego (zob. tab. 1).

Uogólniona postać miary odległości, w której uwzględnia się wagi zmiennych, określona jest wzorem (por. M. Walesiak, 1999): (8)

$$d_{ik} = \frac{1}{2} - \frac{\sum_{j=1}^{m} w_j a_{ikj} b_{kij} + \sum_{j=1}^{m} \sum_{\substack{l=1 \\ l \neq i,k}}^{n} w_j a_{ilj} b_{klj}}{2\left[\left(\sum_{j=1}^{m} w_j a_{ikj}^2 + \sum_{j=1}^{m} \sum_{\substack{l=1 \\ l \neq i,k}}^{n} w_j a_{ilj}^2\right)\left(\sum_{j=1}^{m} w_j b_{kij}^2 + \sum_{j=1}^{m} \sum_{\substack{l=1 \\ l \neq i,k}}^{n} w_j b_{klj}^2\right)\right]^{\frac{1}{2}}},$$

gdzie: w_j – waga j-tej zmiennej spełniająca warunki: $w_j \in (0;\ m)$, $\sum_{j=1}^{m} w_j = m$.

W literaturze można spotkać trzy sposoby ustalania wag zmiennych. Wagi ustala się albo metodą ekspertów (metoda *a priori*), albo z użyciem algorytmów obliczeniowych opierających się na informacjach zawartych w danych pierwotnych (surowych). Można też wykorzystać metodę opartą na obu tych ujęciach. (por. T. Borys, 1984; M. Abrahamowicz i K. Zając, 1986; G.W. Milligan, 1989; T. Grabiński, 1992; A. Bąk, 1999). Problem „ważenia" zmiennych nie został dotychczas zadowalająco rozwiązany. Williams stwierdza nawet, że ważenie zmiennych jest manipulowaniem wartościami zmiennych (por. M.S. Aldenderfer i R.K. Blashfield, 1984). Z tego względu często w badaniach empirycznych zakłada się, że zmienne są jednakowo ważne z punktu widzenia badanego problemu (takie stanowisko przyjmują m.in. P.H.A. Sneath i R.R. Sokal, 1973).

Zaproponowana uogólniona miara odległości o postaci (5) i (8) może być wykorzystana jako formuła wzorcowa syntetycznego miernika rozwoju w metodach porządkowania liniowego. Formuły wzorcowe są różnego rodzaju odległościami poszczególnych obiektów od obiektu wzorcowego, którym w badaniach empirycznych jest na ogół tzw. dolny bądź górny biegun rozwoju (por. M. Walesiak, 1996). Metody porządkowania liniowego można wykorzystać w badaniach marketingowych m.in. w analizie atrakcyjności rynku polegającej na ocenie możliwości istniejących (nowych) produktów w stosunku do produktów konkurencyjnych. Pozwalają więc one określić pozycję produktu na rynku na tle produktów konkurencyjnych. Badania tego typu mogą być wykonywane również w odniesieniu do innych obiektów (np. przedsiębiorstw, krajów), ponieważ umożliwiają wyznaczenie pozycji badanego obiektu (obiektów) na tle obiektów konkurencyjnych.

4. Podsumowanie

W artykule zaproponowano uogólnioną miarę odległości o postaci (5), która dla zmiennych mierzonych na skali ilorazowej i (lub) przedziałowej stosowana jest z podstawieniem (6), a dla zmiennych porządkowych z podstawieniem (7). W konstrukcji uogólnionej miary odległości wykorzystano ideę uogólnionego współczynnika korelacji o postaci (1), który obejmuje współczynnik korelacji liniowej Pearsona i współczynnika korelacji zmiennych porządkowych τ Kendalla.

Dodatkowym rezultatem opracowania jest program komputerowy ułatwiający stosowanie uogólnionej miary odległości o postaci (8). Program ten jest dostępny w Katedrze Ekonometrii i Informatyki Wydziału Gospodarki Regionalnej i Turystyki Akademii Ekonomicznej we Wrocławiu (e-mail: abak@oscar.ae.jgora.pl). Ponadto wersja programu miary odległości (8) z podstawieniem (7) została opublikowana w pracy Bąka (por. A. Bąk,1999). Umożliwia on obliczanie odległości między obiektami (rezultatem jest symetryczna macierz odległości) oraz obliczanie odległości obiektów od wzorca (rezultatem jest wektor odległości). Macierz odległości można wykorzystać w hierarchicznych aglomeracyjnych metodach klasyfikacji do podziału zbioru obiektów na klasy, np. w programie *SPSS for Windows.* W programie komputerowym wykorzystywane są pliki formatu DBF, które służą zarówno do dostarczania danych do obliczeń, jak i do przechowywania otrzymanych wyników.