Ewa Wędrowska

Zarządzanie zasobami informacji statystycznej z wykorzystaniem miary ilości informacji strukturalnej

Ekonomiczne Problemy Usług nr 35, cz. 2, 289-306

2009

ZESZYTY NAUKOWE UNIWERSYTETU SZCZECIŃSKIEGO

NR 544 EKONOMICZNE PROBLEMY USŁUG NR 35 2009

EWA WĘDROWSKA
Uniwersytet Warmińsko-Mazurski w Olsztynie

ZARZĄDZANIE ZASOBAMI INFORMACJI STATYSTYCZNEJ Z WYKORZYSTANIEM MIARY ILOŚCI INFORMACJI STRUKTURALNEJ

Wprowadzenie

Rosnąca rola informacji służącej zarówno instytucjom publicznym, przedsiębiorstwom jak i pojedynczym jednostkom, sprawia, że wzrasta popyt na wiarygodną, aktualną i dobrą jakościowo informację statystyczną. Istotne jest odrzucenie szumów informacyjnych i wybór tych informacji, które z punktu widzenia finalnego użytkownika stanowią informacje niezbędne w procesie podejmowania decyzji, gdyż rosną koszty złych decyzji (lub zaniechania decyzji) jako konsekwencji niskiej jakości informacji otrzymanych przez decydentów. Nadrzędnym kryterium w selekcji informacji powinny być subiektywne poglądy i postawy poszczególnych odbiorców informacji. Jednakże w wielu procesach badawczych, szczególnie o charakterze eksploracyjnym, użytkownik informacji wynikowych nie jest do końca ściśle określony.

Jednocześnie badana rzeczywistość jest złożona, co prowadzi do tworzenia w opisie rezultatów badania znacznego zbioru tablic wynikowych. W takich sytuacjach kolejność prezentacji tablic wynikowych ma wpływ na kreowaną informację i w ten sposób na interpretację rozpoznawanych zjawisk. W artykule skupiono uwagę na kryterium służącym do uszeregowania tablic wynikowych, a celem artykułu jest zaprezentowanie algorytmu porządkującego tablice wyni-

kowych informacji statystycznych według zawartego w nich ładunku informacyjnego.

1. Informacje statystyczne

W Polsce, podobnie jak w innych krajach, istnieje duża liczba instytucji i organizacji wykonujących różnego rodzaju badania statystyczne. Ważną warstwą infrastruktury informacyjnej zarówno państwa, gospodarki jak i społeczeństwa są systemy statystyki publicznej finansowane ze środków publicznych. Systemy te dostarczają informacji statystycznych organom władzy i administracji państwowej, przedsiębiorstwom, organizacjom społecznym oraz poszczególnym obywatelom (Oleński 2006). Często podkreśla się niezbywalne prawo obywateli do prawdy, bez której nie ma wolności. W państwie demokratycznym, w społeczeństwie obywatelskim prawo to wyraża się w formie dostępu każdego podmiotu do rzetelnej, obiektywnej, pełnej i wiarygodnej informacji. Realizowanie tegoż prawa jest jednocześnie związane z nałożeniem na państwo obowiązku zapewnienia informacji dobrej jakościowo poprzez tworzenie odpowiedniej infrastruktury informacyjnej, częścią której jest miedzy innymi statystyka publiczna. Statystyka publiczna to system zbierania danych statystycznych, ich gromadzenia, przechowywania i opracowywania oraz ogłaszania i rozpowszechniania wyników dokonywanych obliczeń, opracowań i analiz, w tym podstawowych wielkości i wskaźników.

W 1918 roku Rada Królewska Królestwa Polskiego ustanowiła Główny Urząd Statystyczny (GUS), jako pierwszą instytucję statystyczną w kraju. Obecnie GUS zajmuje się zbieraniem danych statystycznych, przechowywaniem i przetwarzaniem zebranych danych, publikowaniem, interpretacją oraz rozpowszechnianiem rezultatów badań. Do informacji statystycznych opracowywanych przez GUS należą: roczniki statystyczne, serie wydawnicze, biuletyny oraz bazy danych. Szczegółowe funkcje statystyki publicznej w Polsce wyznacza ustawa z dnia 29 czerwca 1995 r. o statystyce publicznej (Dz. U. Nr 88, poz. 439, z 1996 r. Nr 156, poz. 775, z 1997 r. Nr 88, poz. 554 i Nr 121, poz. 769 oraz z 1998 r. Nr 99, poz. 632 i Nr 106, poz. 668). Tryb oraz formę udostępniania i rozpowszechniania wynikowych informacji statystycznych statystyki publicznej reguluje rozporządzenie Prezesa Rady Ministrów z dnia 10 września 1999 r. w sprawie trybu i form ogłaszania, udostępniania i rozpo-

wszechniania wynikowych informacji statystycznych (Dz. U. z dnia 17 września 1999 r.). Zasady, zgodnie z którymi powinny funkcjonować oficjalne publiczne systemy infrastrukturalne zostały sformułowane w Rezolucji EKD ONZ w sprawie fundamentalnych zasad statystyk oficjalnych w regionie europejskim przyjętej w dniu 15 kwietnia 1992 r. w Genewie (rezolucja EKD ONZ E/1992/32). W środowiskach statystycznych dokument ten nazywany jest dekalogiem statystyki oficjalnej (Oleński 2006), gdyż formułuje dziesięć fundamentalnych zasad, jakimi powinna kierować się statystyka oficjalna.

Wyniki badań prowadzonych przez statystykę publiczną mają charakter oficjalny, do których ma dostęp każdy obywatel w ramach prawa do informacji. W gospodarkach rozwiniętych rośnie znaczenie systemów informacji statystycznej będących systemami publicznymi. Zakres informacji statystycznej jest upowszechniony jako dobro publiczne i stanowi integralną część wiedzy ogólnej społeczeństwa.

Wynikowe informacje statystyczne określane są w statystyce publicznej jako „wyniki obliczeń, opracowań i analiz dokonanych na podstawie zebranych w badaniach statystycznych statystyki publicznej danych statystycznych" (Dz. U. Nr 88, poz. 439, 1996 r). Pojecie informacji należy jednak do kategorii pojęć rozmaicie definiowanych i rozumianych ze względu na wykorzystanie w różnych dziedzinach poczynając od filozofii, poprzez cybernetykę, ekonomię czy statystykę. Konieczność przystosowania tego pojęcia do wymagań rzeczywistości powoduje, że często jest ono wieloznaczne.

Jako klasyczną teorię informacji traktuje się teorię matematyczną. Taką bowiem postać nadał jej Shannon, uważany za twórcę ilościowej teorii informacji. Sam Shannon nie zdefiniował jednak pojęcia informacji definiując jedynie pojęcie jej ilości. W literaturze najczęściej jednak rozwijane są następujące koncepcje informacji: syntaktyczna, semantyczna oraz pragmatyczna. Najbardziej interesujące, zdaniem Autorki, podejście do zagadnienia informacji przedstawili B. Langeforse oraz B. Sundgren. Jest to koncepcja infologiczna informacji, która zakłada, że działalność człowieka wymaga wiedzy. Wiedza zaś powstaje dzięki informacjom, które reprezentowane są przez dane. W oparciu o infologiczną interpretację informacji prowadzone będą dalsze rozważania. Formalna istota tej koncepcji wymaga zdefiniowania pojęcia komunikatu, który można zapisać następująco:

$$K:=(O, X, x, t, q), \tag{1}$$

gdzie:
O - obiekt, X - cecha (atrybut) obiektu O, x - wartość cechy X, t - czas, w którym cecha X obiektu O ma wartość x, q - wektor dodatkowych charakterystyk związanych z obiektem O, cechą X i (lub) czasem t.

Układ K jest komunikatem infologicznym (Langefors 1979), (Stefanowicz 1996). Komunikat K pełni rolę nośnika informacji i stanowi minimalny wystarczający zestaw danych do przekazania jednoznacznej treści. Treść zawarta w elementarnym komunikacie opisanym formułą (1.1) jest informacją elementarną.

Wykorzystując koncepcję infologiczną informacji informację statystyczną zdefiniować można jako informację opisującą pewien złożony obiekt O, będący zbiorem jednostek statystycznych scharakteryzowanych cechą X, przyjmującą w czasie t wartość x. W dalszej części artykułu pojecie informacji statystycznej rozumiane będzie zgodnie z koncepcją infologiczną.

W wyniku przetwarzania zebrane dane opracowane zostają do poziomu przewidywanego w informacjach wynikowych i zawartego w projekcie tablic. Zagregowane i opracowane informacje przedstawione są zazwyczaj w formie tablic, które mogą być dopuszczone do publikacji. Działania podejmowane w ostatniej fazie badania, czyli publikowaniu, nie są już nakierowane na dodanie nowych danych lub poprawienie ich wiarygodności. Część z tych działań ma na celu wręcz ograniczenie i selekcję informacji wynikowych możliwych do rozpowszechniania. Pierwszym ograniczeniem jest zastosowanie środków zapobiegających ujawnieniu poufnych danych (Wędrowska 2002). Na szczególne podkreślenie zasługuje w tym miejscu kwestia tajemnicy statystycznej. Zgodnie z ustawą o statystyce publicznej (Dz. U. Nr 88, poz. 439, 1996 r., art. 10), zbierane i gromadzone w badaniach statystycznych statystyki publicznej dane indywidualne i dane osobowe są poufne i podlegają szczególnej ochronie, tzn. mogą być wykorzystywane wyłącznie do opracowań zestawień i analiz statystycznych. Problem tajemnicy statystycznej jest obecnie często podejmowany i stał się przedmiotem wielu polemik (zob. Szreder 2008).

Drugim etapem służącym ograniczeniu informacji wynikowych możliwych do rozpowszechniania jest wyselekcjonowanie zbiorów informacji o maksymalnej użyteczności.

Główny Urząd Statystyczny prowadzi prace nad podniesieniem jakości informacji statystycznych w celu lepszego zaspokojenia potrzeb użytkowników, zmniejszenia obciążeń respondentów oraz obniżenia kosztów tworzenia informacji wynikowych. Przyjęta definicja jakości w statystyce odwołuje się w swoich aspektach do satysfakcji użytkowników, a więc ich subiektywnych oczekiwań i potrzeb informacyjnych. Jakość w statystyce publicznej oparta jest na definicji jakości Europejskiego Systemu Statystycznego i określona na podstawie pożądanych cech informacji statystycznej: użyteczności, dokładności, terminowości i punktualności, dostępności i przejrzystości, porównywalności, spójności[1].

Jednakże kwestią słabo rozpoznaną w standardach dotyczących prezentacji zestawień tabelarycznych są zasady porządkowania tablic informacji wynikowych. Kolejność prezentowania wyników w dużej mierze zależy od toku prowadzonego wywodu, czyli jest podporządkowana koncepcji badania. Jednakże w wielu procesach badawczych użytkownik wyników nie jest do końca ściśle określony, zaś badana rzeczywistość jest złożona, co prowadzi do tworzenia w opisie rezultatów badania znacznego zbioru tablic wynikowych. W takich sytuacjach kolejność prezentacji tablic wynikowych ma wpływ na kreowaną informację statystyczną i w ten sposób na interpretację rozpoznawanych zjawisk przez użytkownika informacji. W artykule uwagę skupiono na kryterium służącym do uszeregowania tablic wynikowych według zawartego w nich ładunku informacyjnego. Proponowany mechanizm (Wędrowska 2003) oparty na obiektywnym (datalogicznym) kryterium, może przyczynić się do racjonalizacji procesów przetwarzania wynikowych informacji statystycznych.

2. Ilość informacji strukturalnej dostarczanej przez tablice wynikowe

Rozważmy obiekty O_j ze zbioru **O** będące przedmiotem opisu tablicy T scharakteryzowane przez wektory $[n_{jk}]$ lub $[x_{jk}]$ ($j = 1,...,m$; $k = 1,...,n$), gdzie n_{ij} oznacza liczbę występujących k-tych wariantów cechy X w j-tym obiekcie badania. Dla każdego obiektu można wyznaczyć odpowiednio współczynniki struktury lub współczynniki udziału, oznaczone α_{jk} tworzące odpowiedni wektor $S_j = [\alpha_{j1}, \alpha_{j2},..., \alpha_{jn}]$. Dysponując wskaźnikami α_{jk} dla wszystkich m obiek-

[1] www.stat.gov.pl, marzec 2009.

tów będących przedmiotem opisu tablicy *T* zbudować można następującą macierz wskaźników α_{jk}:

$$[\alpha_{ij}] \tag{2}$$

Wyznaczanie wektora S_j jest zasadne tylko wtedy, gdy cecha *X* podlegająca badaniu spełnia własność addytywności, to znaczy wtedy, gdy suma wartości poszczególnych wariantów cechy przejawia sens ekonomiczny. W szczególności w analizach ekonomicznych klasę najczęściej spotykanych struktur stanowią addytywne struktury ekonomiczne, dla których suma wartości przyporządkowanych określonym elementom procesu gospodarczego ma sens ekonomiczny (Kukuła 1996).

Wskaźniki struktury oraz wskaźniki udziału spełniają następujące warunki:

$$0 \le \alpha_{jk} \le 1 \quad (j = 1,..., m;\ k = 1,..., n), \tag{3}$$

$$\sum_{k=1}^{n} \alpha_{jk} = 1 \quad (j = 1,...,m;\ k = 1,...,n) \tag{4}$$

Sumę wszystkich wskaźników struktury α_{jk} dla *m* obiektów O_j (j= 1,2,...,m) można wyrazić:

$$\sum_{j=1}^{m} \sum_{k=1}^{n} \alpha_{jk} = m \tag{5}$$

Załóżmy dalej, że w tablicy *T* opisane zostały obiekty O_j (j = 1,2,...,m) scharakteryzowane cechą *X* spełniającą warunek addytywności. Zatem dla każdego obiektu, dysponując pełnymi danymi o współrzędnych wektora S_j spełniających warunek normy (4), można wyznaczyć entropię rzeczywistą (empiryczną) obiektu O_j, przyjmując za podstawę logarytmu liczbę 2:

$$H(O_j) = -\sum_{k=1}^{n} \alpha_{jk} \log_2 \alpha_{jk} \tag{6}$$

Entropia $H(O_j)$ zależy wyłącznie od częstości występowania *k*–tego wariantu cechy *X* w *j*–tej strukturze S_j, a więc od wskaźników struktury (lub udziału) charakteryzujących dany obiekt O_j. Wielkość $H(O_j)$ przedstawia miarę nieokreśloności i charakteryzuje średni poziom nieokreśloności rozkładu cechy *X* (Kuriata 2001).

Entropia $H(O_j)$ osiągnie maksimum dla struktury S_j o wskaźnikach α_{jk}, takich, że

$$\alpha_{j1} = \alpha_{j2} = ... = \alpha_{jn} \tag{7}$$

Maksimum to wynosi $\log_2 n = H_{\max}$

Dla struktury S_j (j=1,...,m) zdefiniowana zostanie dalej miara dekoncentracji struktur jako stosunek entropii rzeczywistej H(O_j) do maksymalnej wartości entropii H_{max}:

$$DC_{S_j} = \frac{H(O_j)}{H_{\max}} \tag{8}$$

Wskaźnik dekoncentracji struktury S_j jest miarą dekoncentracji rozkładu wartości cechy X dla badanego obiektu O_j a także rozkładu cechy X w czasie. Przyjmuje wartości z przedziału [0,1], gdyż entropia rzeczywista H(S_j) jest wartością nieujemną i osiąga wartość najmniejszą równą 0 oraz wartość największą równą $\log_2 n = H_{max}$.

Wskaźnik dekoncentracji DC_{S_j} osiąga wartość równą 0, gdy jeden ze wskaźników α_{jk} struktury S_j osiąga wartość 1 dla pewnego k (k = 1,2,...,n), a pozostałe $\alpha_{jr} = 0$ dla $r \neq k$ (r = 1,2,...,n). Oznacza to, iż dla badanego obiektu O_j wartości cechy X jest skoncentrowana tylko w jednym z n wariantów, czyli następuje całkowita koncentracja. Jeśli natomiast $DC_{S_j} = 1$ **,** to oznacza, że rozkład cechy X dla j–tego obiektu (lub w badanym okresie czasu) jest równomierny, a wskaźniki α_{jk} struktury S_j (j =1,...,m; k = 1,..,n) spełniają relację (7). Występująca wtedy całkowita dekoncentracja towarzyszy sytuacji, gdy H(S_j) osiąga maksimum (Rószkiewicz, Wędrowska 2004).

W m–elementowym zbiorze **O** obiektów obiekty zostaną połączone w pary (O_i,O_j), gdzie i, j są numerami obiektów oraz $i \neq j$ (i ,j = 1,...,m). Dla m opisanych w tablicy T obiektów, można wyróżnić

$$\frac{m!}{(m-2)!} \tag{9}$$

par obiektów. Liczbę (9) nazywać będziemy liczbą spotkań obiektów (Wędrowska 2003).

Dla każdej pary (O_i,O_j) ($i \neq j$; i, j = 1,..., m) określony zostanie wskaźnik struktury pary obiektów jako stosunek k–tej wartości cechy X dla i – tego obiektu badania, do sumy wartości cechy X dla pary obiektów (O_i,O_j):

$$\alpha_{ijk} = \frac{n_{ik}}{\sum_{k=1}^{n} n_{ik} + \sum_{k=1}^{n} n_{jk}} \tag{10}$$

($i, j = 1, 2,..., m$; $i \neq j$, $k = 1, 2,..., n$), gdzie n_{ik} jest liczbą jednostek o k – tym wariancie cechy w i – tym obiekcie badania, n_{jk} - liczbą jednostek o k – tym wariancie cechy w j – tym obiekcie badania.

Wskaźniki struktury α_{ijk} wyrażają częstość występowania k–tej realizacji cechy X obiektu O_i w łącznej sumie realizacji cechy X dla pary obiektów (O_i O_j).

Wskaźniki α_{ijk} oraz α_{jik} spełniają warunek normy:

$$\sum_{k=1}^{n} \alpha_{ijk} + \sum_{k=1}^{n} \alpha_{jik} = 1 \ ; \ i, j = 1,..., m; i \neq j, k = 1,..., n. \tag{11}$$

Dysponując wskaźnikami struktury par obiektów (O_i,O_j) dla wszystkich $i, j = 1,...,m$, takich, że $i \neq j$, można zbudować $\frac{m!}{(m-2)!}$ tablic współczynników struktury par obiektów. Przykładem takiej tablicy dla i – tego obiektu O_i ($i = 1,..., m$) jest tablica A.

Tablica A

Wskaźniki struktury par obiektów

Para obiektów	Wariant cechy X			
	1	2	...	n
(O_i,O_1)	α_{i11}	α_{i12}	...	α_{i1n}
(O_i,O_2)	α_{i21}	α_{i22}	...	α_{i2n}
....	...	...	...	...
(O_i,O_{i-1})	$\alpha_{i,i-1,1}$	$\alpha_{i,i-1,2}$	...	$\alpha_{i,i-1,n}$
(O_i,O_{i+1})	$\alpha_{i,i+1,1}$	$\alpha_{i,i+1,2}$	...	$\alpha_{i,i+1,n}$
...	...	...	...	...
(O_i,O_m)	α_{im1}	α_{im2}	...	α_{imn}

Źródło: Opracowanie własne.

Znając rozkłady wariantów cechy X odpowiednio dla obiektów O_i oraz O_j można zbadać poziom nieokreśloności rozkładu dla obiektu O_j, która pozostaje w wyniku nieokreśloności rozkładu dla obiektu O_i (Kuriata 2001). Wskaźniki α_{ijk} będą podstawą do wyznaczenia entropii warunkowej pary obiektów.

Entropia warunkowa $H(O_i/O_j)$ pary obiektów (O_i,O_j) przedstawiona będzie w postaci:

$$H(O_i/O_j) = -\sum_{k=1}^{n} \alpha_{ijk} \log_2 \alpha_{ijk} \quad (i,j = 1,...,m,\ i \neq j;\ k = 1,...,n). \tag{12}$$

Jeśli obliczona zostanie entropia warunkowa (12) dla każdej pary (O_i,O_j) takiej, że $i \neq j$ $(i, j = 1,...,m)$, to liczność zbioru wartości $H(O_i/O_j)$ otrzymanych wartości entropii warunkowej par obiektów będzie równa liczbie spotkań obiektów (9). Wartości te można zapisać w tablicy, w której diagonale zostają pominięte (tablica B).

Tablica B

Schemat partnerstwa par obiektów

	O_1	O_2	...	O_m
O_1		$H(O_1/O_2)$	...	$H(O_1/O_m)$
O_2	$H(O_2/O_1)$		....	$H(O_2/O_m)$
....	...	...	...	...
O_m	$H(O_m/O_1)$	$H(O_m/O_2)$	...	

Źródło: Opracowanie własne.

Ponieważ $\sum_{k=1}^{n} \alpha_{ijk} \neq \sum_{k=1}^{n} \alpha_{jik}$ (równość zachodzi tylko wtedy, gdy wektory charakteryzujące struktury S_i oraz S_j są sobie równe, to znaczy $\alpha_{ijk} = \alpha_{jik}$; $i, j = 1,...,m$ oraz $i \neq j$; $k = 1,...,n$), entropia $H(O_i/O_j)$ nie spełnia warunku symetrii, stąd tablica 2 nie jest symetryczna.

Jeśli **O** jest zbiorem obiektów badania opisanych w tablicy T, można zdefiniować wskaźnik struktury obiektu O_j w całym zbiorze **O** jako stosunek sumy wartości zmiennej X dla obiektu O_j do sumy wszystkich realizacji zmiennej X w całym zbiorze **O**:

$$\alpha_j = \frac{\sum_{k=1}^{n} n_{jk}}{\sum_{j=1}^{m}\sum_{k=1}^{n} n_{jk}}, \quad j = 1,...,\ m;\ k = 1,...,\ n. \tag{13}$$

Wskaźnik struktury obiektu O_j w zbiorze **O** opisuje częstość występowania wszystkich elementów występujących w obiekcie O_j jakie wystąpiły w całym

zbiorze **O**. Wskaźnik udziału obiektu O_j w zbiorze **O**, wyznaczany jest analogicznie.

Uwzględnienie realizacji cechy *X* dla wszystkich obiektów jednocześnie zmniejsza entropię warunkową będącą wartością oczekiwaną informacji. Poniższy wzór definiuje średnią entropię warunkową:

$$H(O_j / \mathbf{O}) = \sum_{i=1}^{m} H(O_j/O_i) \cdot \alpha_i \ , \ i \neq j;\ i, j = 1, 2,..., m. \qquad (14)$$

Znajomość jednocześnie entropii rzeczywistej H(O_j) oraz średniej entropii warunkowej H(O_j/**O**) pozwala na zastosowanie wzoru Shannona wyznaczającego ilość informacji jako różnicę pomiędzy entropią rzeczywistą obiektu $H(O_j)$ oraz średnią entropią warunkową $H(O_j/\mathbf{O})$:

$$I(O_j/\mathbf{O}) = H(O_j) - H(O_j/\mathbf{O}) \qquad (15)$$

Ilość informacji $I(O_j/\mathbf{O})$ stanowi ilość informacji strukturalnej wyrażającej ilość informacji o strukturze obiektu O_j w zbiorze obiektów **O** opisanych za pomocą komunikatów *K*. Wielkość (15) zależy nie tylko od struktury obiektu O_j, ale również od wzajemnych relacji i powiązań pomiędzy strukturą tego obiektu, a strukturami pozostałych obiektów ze zbioru **O**, uwzględnionych w schemacie partnerstwa.

Tablica *T*, rozumiana w sensie datalogicznym jako komunikat, niesie treść o zjawiskach uwzględnionych w badaniach statystycznych. Treść tę rozumiemy jako informację statystyczną. W szczególności treść wynikająca ze struktury obiektów ze zbioru **O** jest informacją strukturalną. Ilość informacji strukturalnej określona zostanie w następujący sposób:

$$\boldsymbol{E}\ (\boldsymbol{T}) = \sqrt[m]{\prod_{j=1}^{m} I(O_j/\mathbf{O})} \qquad (20)$$

gdzie $O_j \in \mathbf{O}$ $(j = 1, ...m)$.

Ilość informacji strukturalnej $\boldsymbol{E}(\boldsymbol{T})$ wyraża wielkość ładunku informacyjnego dostarczanego przez tablicę *T,* wynikającego ze struktury obiektów w niej uwzględnionych. Przedstawiona miara jest propozycją teoretyczną, opartą na datalogicznej interpretacji informacji wynikającej wyłącznie z rozkładów statystycznych cechy *X* (Wędrowska 2003).

Ilość informacji strukturalnej $\boldsymbol{E(T)}$ jest miarą jednocześnie zróżnicowania dekoncentracji rozkładów jednostek statystycznych na wszystkie warianty cechy X dla poszczególnych obiektów O_j uwzględnionych w tablicy oraz różnic pomiędzy bezwzględnymi wartościami cechy charakteryzującymi obiekty O_j.

3. Ilustracja empiryczna

Coraz większym problemem staje się czasochłonne przetwarzanie dużej ilości danych. W rozwiązywaniu tych problemów poszukuje się między innymi rozwiązań informatycznych. Przykładem jest wykorzystanie przez Główny Urząd Statystyczny w 2002 r. rozwiązań bazujących na produkcie Citrix MetaFrame XP przy analizie danych pochodzących z dwóch spisów powszechnych: Narodowego Spisu Powszechnego Ludności i Mieszkań oraz Powszechnego Spisu Rolnego. Spis ludności zawierał dane dotyczące ponad 38 milionów osób, a spis rolny o około 3 milionach gospodarstw rolnych. Rozwiązanie to umożliwiło zdalne przetwarzanie zgromadzonych danych i w ten sposób efektywne wykorzystanie zespołów pracujących na terenie całego kraju. Dodatkowo skrócił się czas przetwarzania danych, zminimalizowały koszty przesyłu danych a ponadto dostęp do aplikacji był szybki i bezpieczny z różnych lokalizacji na terenie całej Polski.

Wciąż problem pozostaje jednak kwestia przygotowania do publikacji już opracowanych informacji wynikowych. Wynikowe informacje statystyczne zgromadzone są w bazach danych, a możliwości wykorzystania tych informacji przez użytkowników maleją wraz ze wzrostem rozmiarów zebranych baz danych. Specjaliści – statystycy zajmujący się analizą tablic, oceniają przydatność tablic wynikowych w badaniach statystycznych i podejmują decyzje, które z tablic powinny ukazać się w publikacjach. Tablice wynikowe, które zostają wybrane ze względu na ich przydatność, mogą zostać uporządkowane wg ilości informacji strukturalnej dostarczanej przez te tablice zgodnie z zaproponowanym przez autorkę algorytmem porządkującym tablice wynikowe.

Aby zilustrować wykorzystanie algorytmu porządkującego tablice informacji statystycznych rozpatrzono cztery tablice informacji wynikowych $T_1,...,$ T_4 pochodzących z bazy danych regionalnych Głównego Urzędu Statystyczne-

go[2]. Dane statystyczne w rozbiciu na sześć regionów administracyjnych Polski opisują powierzchnię użytkowania gruntów (tabela 1) oraz dotyczą liczby udzielonych noclegów turystom zagranicznym (tabela 2), liczby rezydentów korzystających z turystycznych obiektów zakwaterowania (tabela 3) i miejsc noclegowych w turystycznych obiektach zbiorowego zakwaterowania (tabela 4).

Tabela 1

Użytkowanie gruntów (2006, stan w czerwcu)

REGIONY	Powierzchnia ogólna	użytki rolne			lasy i grunty leśne	grunty ugorowane łącznie z nawozami zielonymi [(a) (b)]
		grunty orne	sady	łąki i pastwiska		
	w ha					
POLSKA	31 268 315	12 357 372	292 356	3 215 648	9 200 448	1 025 407
Region centralny	5 483 536	2 417 715	125 840	676 489	1 186 566	212 209
Region południowy	2 760 629	808 369	17 416	323 872	837 477	125 370
Region wschodni	7 463 491	2 849 795	98 732	956 661	2 178 218	240 789
Region północno-zachodni	6 660 071	2 678 098	27 528	517 844	2 302 365	195 572
Region południowo-zachodni	2 924 458	1 306 823	7 286	222 615	850 730	98 105
Region północny	5 976 130	2 296 572	15 554	518 167	1 845 092	153 362

(a) Według siedziby użytkownika.

(b) W 2006 ugory i odłogi - łącznie z nawozami zielonymi

[2] www.stat.gov.pl, marzec 2009.

Tabela 2

Udzielone noclegi turystom zagranicznym w turystycznych obiektach zbiorowego zakwaterowania (2006)

REGIONY	Obiekty hotelowe	Kempingi i pola biwakowe	Zespoły ogólnodostępnych domków turystycznych	Pozostałe obiekty
POLSKA	7 910 690	211 152	46 207	2 387 070
Region centralny	1 626 924	3 618	1 366	57 882
Region południowy	2 260 595	38 493	7 502	505 474
Region wschodni	378 410	4 856	4 195	89 644
Region północno-zachodni	1 491 826	77 156	24 089	1 263 726
Region południowo-zachodni	1 071 167	17 171	2 315	239 557
Region północny	1 081 768	69 858	6 740	230 787

Tabela 3

Rezydenci korzystający z turystycznych obiektów zbiorowego zakwaterowania (2006)

REGIONY	Obiekty hotelowe	Kempingi i pola biwakowe	Zespoły ogólnodostępnych domków turystycznych	Pozostałe obiekty
POLSKA	7 563 895	245 050	230 926	5 158 666
Region centralny	1 684 107	9 831	52 888	405 613
Region południowy	1 573 112	23 207	18 782	1 389 012
Region wschodni	921 377	24 357	36 862	607 089
Region północno-zachodni	1 344 473	81 316	60 196	1 056 960
Region południowo-zachodni	812 466	13 425	15 719	683 742
Region północny	1 228 360	92 914	46 479	1 016 250

Tablica: 4

Miejsca noclegowe w turystycznych obiektach zbiorowego zakwaterowania (2006, stan w dniu 31 lipca)

REGIONY	Obiekty hotelowe	Kempingi i pola biwakowe	Zespoły ogólnodostępnych domków turystycznych	Pozostałe obiekty
POLSKA	178 056	54 430	22 074	320 052
Region centralny	31 748	2 513	2 392	16 011
Region południowy	39 061	4 851	1 453	52 561
Region wschodni	19 221	4 422	3 225	31 945
Region północno-zachodni	33 325	19 829	7 514	104 330
Region południowo-zachodni	21 178	3 629	1 142	26 560
Region północny	33 523	19 186	6 348	88 645

Każda z tabel zawiera dane dotyczące sześciu regionów (obiektów O_j) scharakteryzowanych cechami addytywnymi. Dlatego też można było w oparciu o wartości analizowanych cechy wyznaczyć wektory S_j dla każdego z sześciu regionów. Wektory struktur posłużyły wyznaczeniu entropii warunkowej $H(O_j)$ oraz wskaźnika dekoncentracji DC_{Sj}. Kolejny krok algorytmu to wyznaczenie entropii warunkowej $H(O_i/O_j)$ oraz średniej entropii warunkowej $H(O_j/\mathbf{O})$ i ostatecznie ilości informacji strukturalnej $I(O_j/\mathbf{O})$ (tabela 5).

Tabela 5

Wartości wskaźnika dekoncentracji, średniej entropii warunkowej oraz miar ilości informacji strukturalnej

REGIONY	**Tabela 1**			**Tabela 2**		
	Wskaźnik dekoncentracji DC_{Sj}	Średnia entropia warunkowa $H(O_j/\mathbf{O})$	Ilość informacji strukturalnej $I(O_j/\mathbf{O})$	Wskaźnik dekoncentracji DC_{Sj}	Średnia entropia warunkowa $H(O_j/\mathbf{O})$	Ilość informacji strukturalnej $I(O_j/\mathbf{O})$
Region centralny	0,7512	1,1153	0,62892	0,1233	0,0891	0,15763
Region południowo-zachodni	0,7637	0,9602	0,81298	0,4029	0,6711	0,13470
Region południowy	0,7465	1,1143	0,61898	0,4229	0,5752	0,27058
Region północno-zachodni	0,4504	0,6491	0,39661	0,6044	0,4690	0,73985
Region północny	0,4522	0,5414	0,50853	0,3970	0,7167	0,07728
Region wschodni	0,6764	1,0465	0,52408	0,4827	0,7779	0,18748
REGIONY	**Tabela 3**			**Tabela 4**		
	Wskaźnik dekoncentracji DC_{Sj}	Średnia entropia warunkowa $H(O_j/\mathbf{O})$	Ilość informacji strukturalnej $I(O_j/\mathbf{O})$	Wskaźnik dekoncentracji DC_{Sj}	Średnia entropia warunkowa $H(O_j/\mathbf{O})$	Ilość informacji strukturalnej $I(O_j/\mathbf{O})$
Region centralny	0,4488	0,3431	0,55451	0,6872	0,4546	0,91980
Region południowo-zachodni	0,5516	0,8474	0,25583	0,6578	0,9219	0,39371
Region południowy	0,6024	0,8868	0,31793	0,7580	0,9264	0,58956
Region północno-zachodni	0,6496	0,5331	0,76614	0,7273	0,6795	0,77511
Region północny	0,5655	0,8562	0,27482	0,7061	0,8708	0,54151
Region wschodni	0,6552	0,9527	0,35777	0,7526	0,9885	0,51681

Źródło: Obliczenia własne.

Znajomość ilości ***E(T)*** informacji strukturalnej zawartej w każdej z tablic $T_1,..., T_4$ posłużyła do ustalenia odpowiedniego porządku tych tablic (tabela 6).

Poniższa tablica może być traktowana jako wskazówka w sprawie kolejności prezentacji tablic $T_1,..., T_4$ w sytuacji, gdyby odbiorca informacji statystycz-

nych nie określił żadnego innego kryterium lub gdyby owi odbiorcy nie byli znani w momencie drukowania tablic. Na poziomie datalogicznym w pierwszej kolejności na uwagę zasługuje tabela 4 zawierająca dane o liczbie miejsc noclegowych w turystycznych obiektach zbiorowego zakwaterowania.

Tabela 6

Porządek tabel zgodnie z kryterium ilości informacji strukturalnej E(T)

Ranga	Tytuł tabeli	*E(T)*
1	Tabela 4: Miejsca noclegowe w turystycznych obiektach zbiorowego zakwaterowania	**0,66895**
2	Tabela 1: Użytkowanie gruntów	**0,50686**
3	Tabela 3: Rezydenci korzystający z turystycznych obiektów zbiorowego zakwaterowania	**0,46018**
4	Tabela 2: Udzielone noclegi turystom zagranicznym w turystycznych obiektach zbiorowego zakwaterowania	**0,26268**

Źródło: Obliczenia własne.

Największym ładunkiem ***E(T)*** informacji strukturalnej charakteryzują się tablice zawierające obiekty opisane przez cechę, której:

- realizacje dla poszczególnych obiektów wykazują równomierny rozdział łącznego funduszu cechy,
- bezwzględne wartości cechy są wysoce zróżnicowane dla wszystkich obiektów uwzględnionych w tablicy.

Przedstawiony porządek tablic pozwala na bliższe poznanie właściwości rozkładów wartości liczbowych zamieszczonych w tych tablicach, rozdziału tychże wartości na odpowiednie warianty cechy oraz ich wzajemnych relacji. Przedstawiony porządek stanowi jedna z nielicznych prób przedstawienia niektórych aspektów datalogicznego ujęcia informacji i zastosowania go w procesach analizy wynikowych informacji statystycznych (Wędrowska 2003).

E-miara ilości informacji strukturalnej zawartej w tablicach, a w konsekwencji ustalony porządek tych tablic, jest obiektywną konsekwencją istnienia elementów składowych zbioru wartości realizacji cechy zamieszczonych w tablicy. Istnienie informacji strukturalnej w tablicy można uznać za fakt obiektywny, niezależny od podmiotu, jaki ją odbiera, czyli niezależny od subiektywnego odbioru wynikowych informacji statystycznych przez jej od-

biorcę, którym może być człowiek lub dowolny system. Dlatego też zaproponowane kryterium rangowania tablic wynikowych pozwala na uporządkowanie tablic jedynie ze względu na wybraną właściwość obiektów rozpatrywanych w tablicy – rozkłady wartości liczbowych charakteryzujących stan badanych obiektów (Rószkiewicz, Wędrowska 2004).

Podsumowanie

Zaprezentowane kryterium porządkujące tablice wynikowe stanowi propozycję ustalenia datalogicznego porządku, który może być wstępnym usystematyzowaniem tablic. Porządek ten może poprzedzać wykorzystanie informacji wynikowych przez użytkownika zależnie od jego potrzeb i dotychczas posiadanych informacji. Prezentowany algorytm przyczynić się może zatem do usprawnienia i racjonalizacji procesów przetwarzania informacji statystycznych.

Literatura

1. Kukuła K. 1996: Statystyczne metody analizy struktur ekonomicznych, Wyd. Edukacyjne; Kraków.
2. Kuriata E. 2001: Teoria informacji i kodowania, Oficyna Wyd. Politechniki Zielonogórskiej, Zielona Góra.
3. Oleński J. 2006: Infrastruktura informacyjna państwa w globalnej gospodarce, Uniwersytet Warszawski, wyd. Nowy Dziennik, Warszawa.
4. Przybyszewski R., Wędrowska E. 2005: Algorytmiczna teoria entropii, Przegląd Statystyczny nr 2, tom 52, s. 85-102, Warszawa.
5. Rószkiewicz M., Wędrowska E. 2004: Datalogiczna koncepcja ilości informacji strukturalnej w analizie zależności, Monografie i Opracowania Szkoły Głównej Handlowej w Warszawie, nr 533, s. 49-62, Warszawa.
6. Stefanowicz B. 1995: Infologiczne aspekty systemów informacyjnych; Roczniki Kolegium Analiz Ekonomicznych SGH, zeszyt 2.
7. Stefanowicz B.1996: Różnorodność informacji; Wiadomości Statystyczne nr 4, GUS, Warszawa.

8. Stefanowicz B.2001: Informatyka statystyczna; Wiadomości Statystyczne nr 6, GUS, Warszawa.
9. Sundgren B. 1973: An infological approach to data bases; Skriftserie Statistiska Centralbyran, Lund, Sztokholm.
10. Szreder M. 2008: O znaczeniu tajemnicy statystycznej dla jakości badań ilościowych, http://www.stat.gov.pl/gus/5840_4340_PLK_HTML.htm
11. Wędrowska E. 2003: Datalogiczna miara ilości informacji strukturalnej jako instrument zarządzania zasobami informacji statystycznej, Prace Naukowe AE Wrocław nr 975.
12. Wędrowska E., Forkiewicz M. 2005: Algorytm porządkowania tablic wynikowych informacji statystycznych; Zeszyty Naukowe Wydziału Elektrotechniki i Automatyki Politechniki Gdańskiej, nr 21, Gdańsk.

THE MANAGEMENT OF STATISTIC INFORMACION RESOURCES WITH THE USE OF THE MEASUREMENT OF QUANTITY OF INFORMACION

Summary

The proposed criterion for table ranking allows for table arrangement by the numeric values that describe the state of the examined objects. The established arrangement may be an initial step in table ranking that provides information in an objectively established hierarchy. This allows analysis and utilisation of result information in a systematic way. The E - measure can also be used to adequately "portion" information in the process of information transfer to the end users. The criterion for table arrangement will contribute to the enhancement and rationalisation of the data processing and interpretation processes by public statistical offices responsible for information publishing**.** The main aim of the article was to formulate a measure (E-measure) to determine an amount of information.

Translated by Ewa Wędrowska