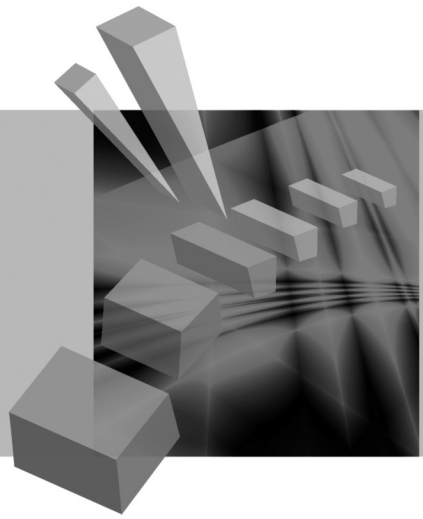
PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu
RESEARCH PAPERS
of Wrocław University of Economics

242

Taksonomia 19.

Klasyfikacja i analiza danych – teoria i zastosowania



Redaktorzy naukowi
Krzysztof Jajuga
Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2012

Recenzenci: Eugeniusz Gatnar, Elżbieta Gołata, Tadeusz Kufel, Józef Pociecha,
Mirosław Szreder, Feliks Wysocki

Redaktor Wydawnictwa: Aleksandra Śliwka

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Małgorzata Czupryńska

Projekt okładki: Beata Dębska

Tytuł sfinansowano ze środków Sekcji Klasyfikacji i Analizy Danych PTS
i Uniwersytetu Ekonomicznego we Wrocławiu

Publikacja jest dostępna na stronie www.ibuk.pl

Streszczenia opublikowanych artykułów są dostępne w międzynarodowej bazie danych
The Central European Journal of Social Sciences and Humanities http://cejsh.icm.edu.pl
oraz w The Central and Eastern European Online Library www.ceeol.com,
a także w adnotowanej bibliografii zagadnień ekonomicznych BazEkon http://kangur.uek.krakow.pl/
bazy_ae/bazekon/nowy/index.php

Informacje o naborze artykułów i zasadach recenzowania znajdują się
na stronie internetowej Wydawnictwa
www.wydawnictwo.ue.wroc.pl

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM
Nakład: 320 egz.

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 242
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS 2012

Taksonomia 19. Klasyfikacja i analiza danych – teoria i zastosowania ISSN 1899-3192

Jacek Kowalewski
Uniwersytet Ekonomiczny w Poznaniu,
Urząd Statystyczny w Poznaniu

ZINTEGROWANY MODEL OPTYMALIZACJI BADAŃ STATYSTYCZNYCH

Streszczenie: Artykuł stanowi próbę sprecyzowania teoretycznego modelu, który pozwalałby na przeprowadzenie optymalizacji procesu badań statystycznych. Złożoność zagadnienia powoduje, że prezentowany model ma charakter wieloetapowego zagadnienia wielokryterialnego z uwzględnieniem aspektów jakościowych, ekonomicznych oraz aktualności informacji. Część rozważań poświęcono przejściu z kryteriów jednostkowych dla poszczególnych informacji na globalne funkcje celu dla całego systemu badań statystycznych. Przedstawiono także warunki niezbędne do praktycznego zastosowania zaprezentowanego modelu.

Słowa kluczowe: badania statystyczne, optymalizacja, jakość informacji.

1. Wstęp

Dynamiczny rozwój gospodarki informacyjnej i społeczeństwa informacyjnego wiąże się z powstawaniem różnego rodzaju wyzwań. Jednym z nich jest szybko narastający popyt na informacje zgłaszany przez szeroki krąg różnych odbiorców. Zaspokajanie tego popytu wiąże się z podejmowaniem prób pozyskiwania danych różnymi drogami (np. przez badania reprezentacyjne, wykorzystanie rejestrów administracyjnych, metod Statystyki Małych Obszarów). Różnorodność dróg dotarcia do pożądanych informacji, konieczność zmniejszenia obciążeń respondentów wymaga całościowej koordynacji badań statystycznych i wiąże się z dylematem wyboru optymalnego sposobu ich pozyskiwania.

Artykuł stanowi próbę sprecyzowania teoretycznego modelu, który pozwalałby na przeprowadzenie całościowej optymalizacji badań statystycznych[1], z uwzględnieniem aspektów ekonomicznych, jakościowych oraz aktualności informacji.

[1] Nie jest intencją autora podejmowanie w artykule dyskusji na temat, bardzo obszerny sam w sobie, definicji badań statystycznych czy też organizacji procesu informacyjnego. Interesujące rozważania w tej kwestii zawierają np. prace: [Sundgren 2003; Oleński 2003; *Information Systems Architecture*…1999].

2. Model systemu badań statystycznych

Pojedyncze badanie statystyczne można przedstawić jako pewien proces transformacji pozyskanych obserwacji i danych w określone informacje wynikowe. Rozpatrując wiele takich procesów, możemy mówić o systemie, w którym różne badania statystyczne pokrywają te same obszary tematyczne, a tym samym umożliwiają pozyskanie tych samych (lub zbliżonych) informacji jako wynik różnych badań.

Podstawowe założenia modelu badań można przedstawić w następujący sposób:

1. Możliwe jest precyzyjne zdefiniowanie wektora *Y*, zawierającego zmienne, których wartości są efektem całego procesu badań statystycznych. Wektor ten będzie nazywany **wektorem wyjściowym**. Tym samym zasadniczym celem procesu badań statystycznych jest oszacowanie wartości *n* różnych docelowych zmiennych

$$Y^T = \left[y_1, y_2, y_3, ..., y_n \right]. \tag{1}$$

2. Wartości zmiennych wyjściowych Y są bezpośrednio oszacowywane na podstawie **zmiennych wejściowych** (pierwotnych), które będą reprezentowane przez *m*-elementowy wektor *X*

$$X^T = \left[x_1, x_2, x_3, ..., x_m \right]. \tag{2}$$

3. Proces przekształcania zmiennych wejściowych *X* w zmienne wyjściowe *Y* będzie nazywany **procesem transformacji** danych i oznaczany jako Ω.

$$\Omega : X \times ... \times X \to Y. \tag{3}$$

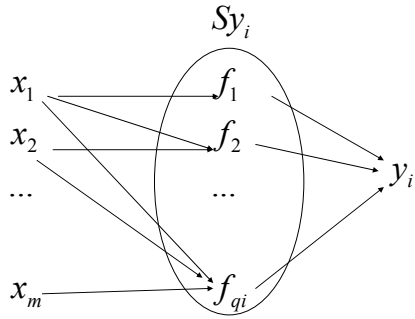
Transformacje mogą mieć różny charakter, a oszacowanie wartości określonej zmiennej wyjściowej y_i uzyskiwane jest drogą przekształcenia jednej lub kilku zmiennych wejściowych

$$y_i = f(X). \tag{4}$$

4. Może występować wiele sposobów oszacowania tej samej zmiennej wyjściowej y_i na podstawie różnych zmiennych wejściowych. Założono, że istnieje skończona liczba transformacji (*qi*), które prowadzą do oszacowania *i*-tej zmiennej wejściowej, które można opisać w zbiorze Sy_i (rys. 1)

$$Sy_i = \left\{ f_1^i(X), .., f_{qi}^i(X) \right\}, \tag{5}$$

gdzie $f_k^i(X)$, oznacza *k*-ty sposób oszacowywania zmiennej wyjściowej y_i na podstawie wektora zmiennych wejściowych *X*.



Rys. 1. Przykład możliwości szacowania *i*-tej zmiennej wyjściowej

Źródło: opracowanie własne.

5. Oszacowania zmiennych wejściowych X uzyskiwane są z **badań**. Założono, że istnieje skończony zbiór r możliwych do przeprowadzenia różnych badań, które można przedstawić w postaci wektora

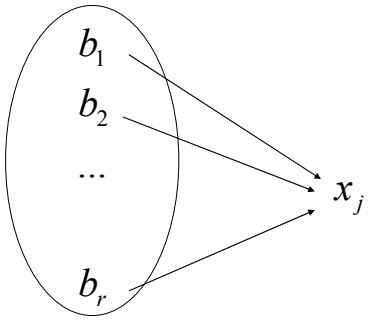
$$B^T = [b_1, b_2, b_3, ..., b_r]. \tag{6}$$

Przez badanie b_l rozumie się wyodrębniony, jednolity proces pozyskiwania danych. Może nim być badanie pełne, reprezentatywne, wykorzystanie rejestrów administracyjnych czy też innych źródeł.

6. Relacja Φ oszacowywania zmiennych wejściowych X na podstawie danych z badań B będzie nazywana **procesem pozyskiwania** danych, co można zapisać jako

$$\Phi : B \to X. \tag{7}$$

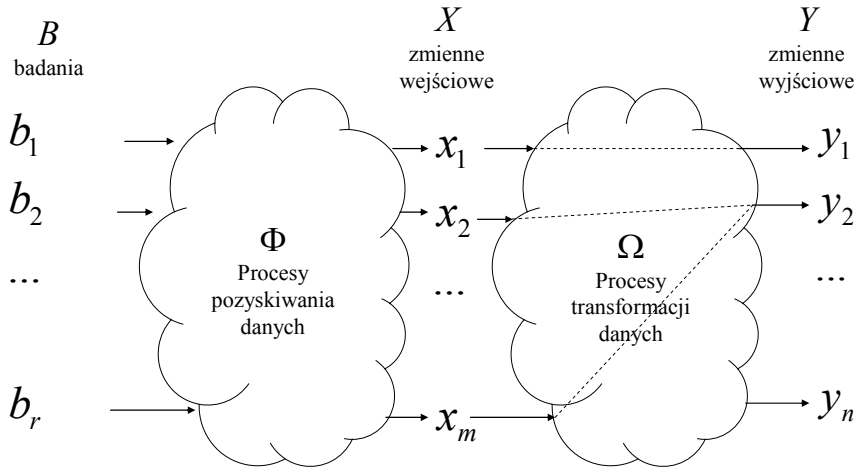
Przyjęto, że oszacowania wartości zmiennej x_j dokonuje się bezpośrednio w badaniu (bez dalszych transformacji). Założono także możliwość szacowania konkretnej, tej samej zmiennej x_j w różnych badaniach (rys. 2).



Rys. 2. Możliwość szacowania *j*-tej zmiennej wejściowej na podstawie różnych badań

Źródło: opracowanie własne.

Na podstawie powyższych założeń można zaproponować model procesu badań statystycznych w postaci zadania dwuetapowego, co przedstawia schematycznie rys. 3.



Rys. 3. Model badań statystycznych

Źródło: opracowanie własne.

Rozwiązanie zadania sprowadza się do ustalenia organizacji procesów pozyskiwania danych (Φ) oraz procesów transformacji danych (Ω) tak, aby skutecznie oszacować wektor zmiennych wejściowych (*Y*). Warto zauważyć, że szacowanie części założonych zmiennych wejściowych (*X*) może okazać się zbędne oraz że w rozwiązaniu optymalnym nie wszystkie planowane badania (*B*) muszą zostać przeprowadzone.

3. Zmienne modelu

Rozwiązanie ogólnego modelu zaproponowanego na rys. 3 wymaga wprowadzenia dodatkowych założeń technicznych umożliwiających określenie typu zadań matematycznych do rozwiązania. Przyjęto zatem, że dla każdego *l*-tego badania b_l można przyporządkować bx^l – *m*-elementowy wektor definiujący, jakie zmienne pozyskiwane są w *l*-tym badaniu, gdzie

$$bx_j^l = \begin{cases} 1, & \text{gdy wartość zmiennej wejściowej } x_j \text{ szacowana jest w badaniu } b_l \\ 0, & \text{w przeciwnym przypadku.} \end{cases} \tag{8}$$

Rozwiązując zadanie, należy szczegółowo ustalić

– dla procesów pozyskiwania danych (Ω) macierz *BX*, zdefiniowaną jako

$$BX = \begin{bmatrix} bx_1^1 \ldots bx_1^l \ldots bx_1^r \\ \ldots \quad bx_j^l \quad \ldots \\ \ldots \\ bx_m^1 \ldots bx_m^l \ldots bx_m^r \end{bmatrix}, \tag{9}$$

gdzie: bx_j^l – binarna zmienna określającą, czy w *l*-tym badaniu szacowana jest wartość zmiennej x_j (zob. (8)),

r – liczba rozpatrywanych badań,

m – liczba potencjalnych zmiennych wejściowych;

– dla procesów transformacji danych (Ω) zbiór sposobów naliczania poszczególnych zmiennych wyjściowych (*Y*), który można zapisać jako

$$\begin{aligned} Sy^* &= \left\{ Sy_1^*, .., Sy_i^*, .., Sy_n^* \right\} \\ Sy_i^* &= \left\{ f_k^i(X) \in Sy_i \right\}, \end{aligned} \tag{10}$$

gdzie: $f_k^i(X)$ – oznacza *k*-ty sposób oszacowywania zmiennej wyjściowej y_i na podstawie wektora zmiennych wejściowych *X*,

Sy_i^* – zastosowany sposób naliczenia zmiennej wyjściowej y_i.

Można zatem przyjąć, że rozwiązanie zadania polega na ustaleniu zbioru (*Z*), składającego się z wektora zmiennych wyjściowych (*Y*), macierzy opisującej pozyskiwanie danych w poszczególnych badaniach (*BX*) oraz zbioru opisującego procesy transformacji danych (Sy^*) opisanego w sposób następujący

$$Z = \left\{ Y, BX, Sy^* \right\}. \tag{11}$$

4. Funkcje kryterium

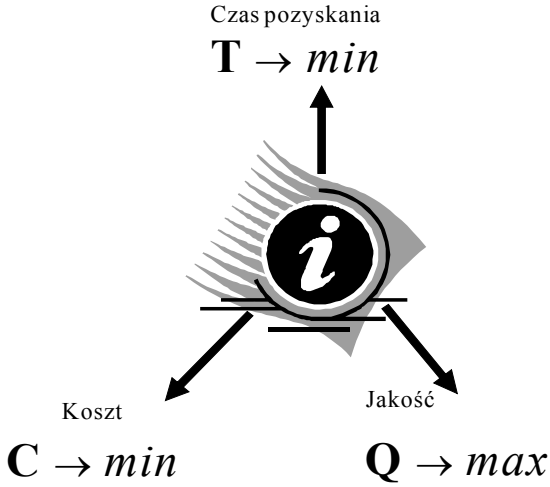
Znalezienie optymalnego sposobu przeprowadzania badań statystycznych wymaga wprowadzenia stosownych funkcji celu. Zakładając, że celem całego procesu badań statystycznych jest uzyskanie odpowiednich zmiennych wyjściowych *Y*, które można traktować jako informacje będące efektem działania systemu, niezbędne jest ustalenie podstawowych cech, które powinny charakteryzować dobrą informację[2]. Przyjmując pewne uproszczenia, można założyć, że należy uwzględnić aspekty ekono-

[2] Ze względu na różnorodność procesów zarządzania informacją jest wiele definicji dobrej informacji. Szerzej np. [Platek, Sarndal 2001; Dillman 1996; College 1996; Kordos 1988].

miczne, jakościowe i aktualność danych. Tym samym badanie statystyczne można opisać jako proces, którego celem jest ustalenie wartości zmiennej $y_i \in Y$ w sposób, który zapewni niskie koszty ($C(y)$), szybki dostęp do informacji[3] ($T(y)$) oraz wysoką jakość informacji[4] ($Q(y)$), co można zapisać jako:

$$\begin{aligned} &C(y) \to \min \\ &T(y) \to \min \\ &Q(y) \to \max \\ &y \in Y \end{aligned} \tag{12}$$

Przyjęte do modelu zarządzania informacją funkcje kryterium można przedstawić jak na rys. 4.



Rys. 4. Kryteria celu w procesie pozyskiwania informacji

Źródło: opracowanie własne.

Wychodząc z powyższych założeń o przypisaniu kryteriów dla pojedynczej zmiennej wyjściowej y przy przejściu na podejście holistyczne optymalizacji całego

[3] Funkcja opisująca to kryterium może np. przyjąć formę różnicy pomiędzy czasem pozyskania (udostępniania) informacji a momentem zajścia opisywanego zjawiska.

[4] Samo zagadnienie jakości danych jest problemem niezwykle złożonym, wymagającym odrębnego rozpatrzenia, który wykracza poza ramy prezentowanego artykułu. Świadczy o tym chociażby niejednokrotne traktowanie aktualności informacji jako elementu jej jakości. Szerzej o jakości danych np. w [Kordos 1988; Oleński 2003; Zarkovich 1966]. Próbując kwantyfikować to kryterium, przy oczywistych uproszczeniach, jakość informacji można traktować np. jako założenie odwrotności dokładności (błędu szacunku) oraz spójności informacji.

systemu badań statystycznych, należy stwierdzić, że w rozważaniach pojawiają się dalsze dwa kluczowe problemy.

Pierwszy dotyczy przyjęcia do modelu założeń, które umożliwią stosowanie przyjętych funkcji kryterium. Ponieważ założono, że wartości zmiennych wyjściowych Y są uzyskiwane drogą transformacji zmiennych wejściowych X, które z kolei pochodzą z badań B, zasadne jest przyjęcie dodatkowo, że:

7. Dla każdego l-tego badania b_l można przyporządkować:
- całkowity koszt przeprowadzenia badania $CB(b_l)$,
- czas pozyskania danych w badaniu $TB(b_l)$, jednakowy dla wszystkich zmiennych wejściowych uzyskiwanych w danym badaniu,
- ocenę jakości dla każdej potencjalnej zmiennej x_j uzyskanej w danym badaniu, którą można oznaczyć jako xq_j^l.

8. Na podstawie wartości ustalonych dla badania b_l istnieje możliwość przyporządkowania wartości kosztu (xc) oraz czasu pozyskania danych (xt) dla zmiennych wejściowych X

$$\begin{aligned} xc_j^l &= f(CB(b_l)) \\ xt_j^l &= f(TB(b_l)) \end{aligned}. \tag{13}$$

9. Koszty procesów transformacji danych (Ω) są znikome, a przez to pomijalne w modelu. Tym samym koszty pozyskania zmiennej wynikowej y_i będą wypadkową kosztów zmiennych wejściowych użytych do wyliczenia zmiennej y_i oraz częstości stosowania poszczególnych zmiennych wejściowych x_j w całym systemie.

10. Podobnie przyjęto, że znikome są czasy przetwarzania danych w procesach transformacji (Ω), a zatem także są pomijalne w modelu. Tym samym czas pozyskania $T(y_i)$ zmiennej wyjściowej y_i uzyskanej metodą Sy_i jest określany przez najdłuższy czas pozyskania dla zmiennych wejściowych użytych do wyliczenia zmiennej wyjściowej y_i

$$T(y_i) = \max_{x_j \in Sy_i} xt_j. \tag{14}$$

11. Założono, że możliwe jest ustalenie jakości oszacowania dowolnej zmiennej wyjściowej y_i, niezależnie od sposobu transformacji. Schematycznie można to przedstawić następująco

$$Q(y_i) = f(f_k^i(X), XQ), \tag{15}$$

gdzie: $Q(y_i)$ – współczynnik jakości oszacowania zmiennej wyjściowej y_i,
$f_k^i(X)$ – k-ty sposób transformacji wektora zmiennych wejściowych X, stosowany do szacowania zmiennej y_i,
XQ – macierz współczynników jakości oszacowania zmiennych wejściowych w poszczególnych badaniach.

Drugi z kluczowych problemów dotyczący funkcji kryterium związany jest ze sposobem przejścia od kryteriów jednostkowych do kryteriów ogólnych, dla całego systemu badań. Ze względu na specyfikę rozważanych kryteriów kosztu, czasu i jakości każde z nich wymaga odrębnego rozpatrzenia.

Stosunkowo najprostsza wydaje się kwestia oszacowania całkowitych kosztów uzyskania pożądanych zmiennych wyjściowych z wektora Y, którą można przedstawić jako całkowity koszt przeprowadzenia wszystkich badań (CY):

$$\begin{aligned} CY &= \sum_{l=1}^{r} CB(b_l) \cdot \operatorname{sgn}\left(n_l\right) \\ n_l &= \sum_{j=1}^{m} bx_j^l, \forall l = \overline{1,..,r} \end{aligned}, \tag{16}$$

gdzie: n_l – parametr oznaczający, ile zmiennych wejściowych jest ustalanych na podstawie badania,

bx_j^l – binarna zmienna określającą, czy w l-tym badaniu szacowana jest wartość zmiennej x_j (zob. (8)),

$CB(b_l)$ – koszt badania b_l,

r – liczba możliwych badań,

m – liczba potencjalnych zmiennych wejściowych,

sgn – funkcja zwracająca znak liczby (dla liczb dodatnich jest to 1, w przypadku zera – 0).

Globalna funkcja określająca czas uzyskania zmiennych wyjściowych ze względu na założenie (10) sprowadza się do analizy czasów pozyskania wektora zmiennych wejściowych X. Tym samym czas uzyskania wszystkich pożądanych zmiennych (TY) wyniesie

$$TY = \max_{i=1,..n} T(y_i) = \max_{j=1,..m} xt_j, \tag{17}$$

gdzie: $T(y_i)$ – czas pozyskania zmiennej wyjściowej y_i,

xt_j – czas pozyskania zmiennej wejściowej x_j.

Najbardziej skomplikowaną kwestią jest sposób szacowania globalnej funkcji jakości (QY). Na użytek artykułu przyjęto, że jest możliwe jej oszacowanie w postaci złożenia funkcji jakości poszczególnych zmiennych wyjściowych[5] ($Q(y_i)$)

$$QY = f(Q(y_1),...,Q(y_n)). \tag{18}$$

[5] Ze wstępnego rozeznania wynika, że przy posiadaniu miary jakości przypisanej do poszczególnych zmiennych wyjściowych y_i przejście na kryterium globalne QY dla całego systemu badań wymaga w gruncie rzeczy spełnienia nie jednego, ale wielu warunków. Zaliczyć do nich można: oczekiwanie wysokiej, średniej jakości, uwzględnienie „najgorszego" oszacowania, określenie minimalnej akceptowalnej jakości itp. Można wręcz stwierdzić, że jest to odrębne zagadnienie wielokryterialne, którego rozpatrzenie wykracza poza ramy prezentowanego artykułu.

5. Podsumowanie

Zaprezentowany model może stanowić narzędzie do ustalania optymalnego sposobu pozyskiwania pożądanych informacji, które reprezentuje wektor *Y*. Proces optymalizacji wymaga rozwiązania pewnego dwuetapowego zadania wielokryterialnego, w którym pierwszy etap – ustalenia organizacji procesów pozyskiwania danych (Φ) – sprowadza się do rozwiązania zadania binarnego, a drugi – ustalania procesu transformacji danych (Ω) – do rozwiązania zadania kombinatorycznego.

Jego praktyczne zastosowanie wymaga nie tylko dalszego rozwinięcia modelu, ale także spełnienia kilku dodatkowych warunków. Jednym z nich jest diametralne odejście od dotychczasowego sposobu organizacji badań statystycznych tworzonych na bazie szczegółowej analizy zmiennych w poszczególnych formularzach. Wymaga to zmiany z podejścia jednostkowego na podejście holistyczne, które oznacza całkowite odwrócenie sposobu działania. Prace nad organizacją badań powinny zacząć się od precyzyjnego określenia „koszyka" niezbędnych zmiennych wyjściowych (jako efektu działania systemu), ustalenia potencjalnych wag istotności dla poszczególnych informacji oraz określenia niezbędnych lub oczekiwanych momentów czasowych udostępniania określonych zmiennych wyjściowych. Dalszym efektem powinna być integracja dotychczasowych badań.

Praktyczne zastosowanie zaprezentowanego modelu wymaga także stworzenia stosownej mapy systemu badań statystycznych przez ścisłe określenie wszystkich potencjalnych źródeł pozyskiwania danych, możliwych sposobów transformacji oraz precyzyjne zdefiniowanie całego zestawu informacji uzyskiwanych na wyjściu. Jest to zadanie wysoce pracochłonne, gdyż np. na podstawie tylko jednego krótkookresowego badania przedsiębiorstw(DG-1), którego formularz zawiera ok. 30 pytań, na wyjściu szacowane są wartości 460 udostępnianych zmiennych.

Istotnym wyzwaniem w sferze modelowej jest precyzyjne ustalenie wszystkich funkcji kryteriów, ze szczególnym uwzględnieniem funkcji jakości. Z dotychczasowych doświadczeń wynika, że mierniki jakości badań są ściśle powiązane ze stosowaną metodą badań, a przeprowadzony przegląd literatury wskazuje na brak uogólnionego modelu oceny jakości wyników badań.

Odrębnym problemem jest także praktyczne rozwiązanie zadania wielokryterialnego, jakim jest przedstawiony w artykule model. Rozbieżność kryteriów i ograniczoność zasobów oznacza, że nie jest możliwe pozyskanie informacji jednocześnie taniej, bardzo szybkiej i wysokiej jakości. Możliwość optymalizacji całego systemu badań statystycznych wymaga określenia dodatkowych założeń (np. sprowadzenia części kryteriów do poziomu warunków ograniczających).

Literatura

College M., *Comment on Why innovation is difficult in government surveys*, "Journal of Official Statistics" 1996, nr 12.

Dillman D., *Why innovation is difficult in government surveys*, "Journal of Official Statistics" 1996, nr 12.

Information Systems Architecture for National and International Statistical Offices. Guidelines and Recommendations, United Nations, Geneva 1999.

Kordos J., *Jakość danych statystycznych*, PWE, Warszawa 1988.

Kowalewski J., *Model optymalizacji badań statystycznych,* [w:] Prace Statystyczne i Demograficzne, red. I. Roeske-Słomka, ZN 133, Wydawnictwo UE, Poznań 2010.

Oleński J., *Ekonomika informacji. Metody,* PWE, Warszawa 2003.

Platek R., Sarndal C.-E., *Czy statystyk może dostarczać dane wysokiej jakości?,* „Wiadomości Statystyczne" 2001, nr 4.

Sundgren B., *Developing and Implementing Statistical Metadata Systems. A Network of Excellence for Harmonising and Synthesising the Development of Statistical Metadata*, EPROS Project Number IST-1999-29093, 2003, dostępny w Internecie: www.epros.ed.ac.uk/metanet/deliverables/D6/IST-1999-29093-D6.

Zarkovich F., *Quality of Statistical Data*, FAO, Rome 1966.

AN INTEGRATED MODEL OF OPTIMIZING STATISTICAL SURVEYS

Summary: The paper is an attempt to specify the characteristics of a theoretical model that could be used to optimize the process of statistical surveys. The complexity of this problem results in the fact that the presented model has a form of multi-stage and multi-criterion issue combining economic, quality and timeliness aspects. Part of the discussion is devoted to turning from individual criteria for particular types of information to global objective functions for the whole system of statistical surveys. The article also presents the prerequisites for the implementation of the model.

Keywords: statistical surveys, optimization, information quality.