

PRACE NAUKOWE

Uniwersytetu Ekonomicznego we Wrocławiu

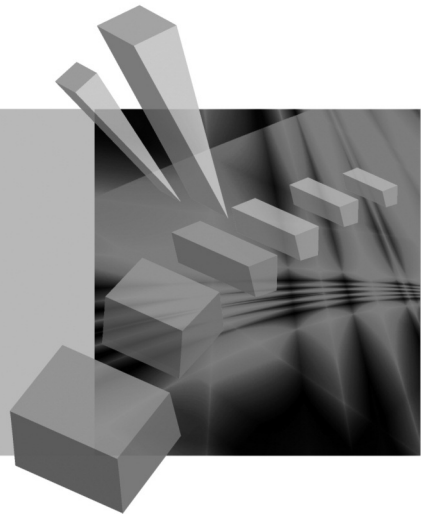
RESEARCH PAPERS

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 Pocięcha,
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](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

Kopowanie i powielanie w jakiegokolwiek formie
wymaga pisemnej zgody Wydawcy

© Copyright by Uniwersytet Ekonomiczny we Wrocławiu
Wrocław 2012

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

Wstęp	13
Stanisława Bartosiewicz , Jeszcze raz o skutkach subiektywizmu w analizie wielowymiarowej	17
Andrzej Sokolowski , Q uniwersalna miara odległości	22
Eugeniusz Gatnar , Jakość danych w systemach statystycznych banków centralnych (na przykładzie NBP)	31
Marek Walesiak , Pomiar odległości obiektów opisanych zmiennymi mierzonymi na skali porządkowej – strategię postępowania.....	39
Krzysztof Jajuga, Marek Walesiak , XXV lat konferencji taksonomicznych – fakty i refleksje	47
Józef Pocięcha, Barbara Pawelek , Model SEM w analizie zagrożenia bankructwem przedsiębiorstw w świetle koniunktury gospodarczej – problemy teoretyczne i praktyczne	50
Paweł Lula , Uczące się systemy pozyskiwania informacji z dokumentów tekstowych	58
Ewa Roszkowska , Zastosowanie metody TOPSIS do wspomaganie procesu negocjacji.....	68
Andrzej Młodak , Sąsiedztwo obszarów przestrzennych w ujęciu fizycznym oraz społeczno-ekonomicznym – podejście taksonomiczne	76
Andrzej Bąk , Modele kategorii nieuporządkowanych w badaniach preferencji	86
Jacek Kowalewski , Zintegrowany model optymalizacji badań statystycznych.....	96
Jan Paradysz, Karolina Paradysz , Obszary bezrobocia w Polsce – problem benchmarkowy.....	106
Tomasz Szubert , W co grać, aby jak najmniej przegrać? Próba klasyfikacji systemów gry w zakładach bukmacherskich.....	116
Izabela Szamrej-Baran , Klasyfikacja krajów UE ze względu na ubóstwo energetyczne	126
Sylwia Filas-Przybył, Tomasz Klimanek, Jacek Kowalewski , Analiza dojazdów do pracy za pomocą modelu grawitacji.....	135
Marta Dziechciarz-Duda, Anna Król, Klaudia Przybysz , Minimum egzystencji a czynniki warunkujące skłonność do korzystania z pomocy społecznej. Klasyfikacja gospodarstw domowych	144
Hanna Dudek , Subiektywne skale ekwiwalentności – analiza na podstawie danych o satysfakcji z osiągniętych dochodów	153

Joanicjusz Nazarko, Ewa Chodakowska, Marta Jaročka , Segmentacja szkół wyższych metodą analizy skupień <i>versus</i> konkurencja technologiczna ustalona metodą DEA – studium komparatywne.....	163
Ewa Chodakowska , Wybrane metody klasyfikacji w konstrukcji ratingu szkół.....	173
Bartosz Soliński , Sektor energetyki odnawialnej w krajach Unii Europejskiej – klasyfikacja w świetle strategii zarządzania zmianą.....	182
Krzysztof Szwarz , Klasyfikacja powiatów województwa wielkopolskiego ze względu na sytuację demograficzną.....	192
Elżbieta Gołata, Grażyna Dehnel , Rejestry administracyjne w analizie przedsiębiorczości.....	202
Katarzyna Chudy, Marek Sobolewski, Kinga Stępień , Wykorzystanie metod taksonomicznych w prognozowaniu wskaźników rentowności banków giełdowych w Polsce.....	212
Katarzyna Dębowska , Modelowanie upadłości przedsiębiorstw przy wykorzystaniu metod dyskryminacji i regresji.....	222
Alina Bojan , Wykorzystanie metod wielowymiarowej analizy danych do identyfikacji zmiennych wpływających na atrakcyjność wybranych inwestycji.....	231
Justyna Brzezińska , Analiza logarytmiczno-liniowa w badaniu przyczyn umieralności w krajach UE.....	240
Aneta Rybicka, Bartłomiej Jefmański, Marcin Pelka , Analiza klas ukrytych w badaniach satysfakcji studentów.....	247
Bartłomiej Jefmański , Pomiar opinii respondentów z wykorzystaniem elementów teorii zbiorów rozmytych i środowiska R.....	256
Julita Stańczuk , Porównanie rezultatów wielostanowej klasyfikacji obiektów ekonomicznych z wykorzystaniem analizy dyskryminacyjnej oraz sieci neuronowych.....	265
Jerzy Krawczuk , Skuteczność metod klasyfikacji w prognozowaniu kierunku zmian indeksu giełdowego S&P500.....	275
Anna Czapkiewicz, Beata Basiura , Symulacyjne badanie wpływu zaburzeń na grupowanie szeregów czasowych na podstawie modelu Copula-GARCH.....	283
Radosław Pietrzyk , Ocena efektywności inwestycji funduszy inwestycyjnych z tytułu doboru papierów wartościowych i umiejętności wykorzystania trendów rynkowych.....	291
Aleksandra Witkowska, Marek Witkowski , Zastosowanie metody Panzara-Rosse’a do pomiaru poziomu konkurencji w sektorze banków spółdzielczych.....	306
Marcin Pelka , Podejście wielomodelowe z wykorzystaniem metody <i>boosting</i> w analizie danych symbolicznych.....	315
Justyna Wilk , Analiza porównawcza oprogramowania komputerowego w klasyfikacji danych symbolicznych.....	323

Tomasz Bartłomowicz, Justyna Wilk , Zastosowanie metod analizy danych symbolicznych w przeszukiwaniu dziedzinowych baz danych.....	333
Kamila Migdał-Najman , Propozycja hybrydowej metody grupowania opartej na sieciach samouczących	342
Dorota Rozmus , Porównanie dokładności taksonomii spektralnej oraz zagregowanych algorytmów taksonomicznych opartych na idei metody <i>bagging</i>	352
Krzysztof Najman , Grupowanie dynamiczne z wykorzystaniem samouczących się sieci GNG	361
Małgorzata Misztal , Wpływ wybranych metod uzupełniania brakujących danych na wyniki klasyfikacji obiektów z wykorzystaniem drzew klasyfikacyjnych w przypadku zbiorów danych o niewielkiej liczebności – ocena symulacyjna	370
Mariusz Kubus , Zastosowanie wstępnego uwarunkowania zmiennej objaśnianej do selekcji zmiennych.....	380
Barbara Batóg, Jacek Batóg , Wykorzystanie analizy dyskryminacyjnej do identyfikacji czynników determinujących stopę zwrotu z inwestycji na rynku kapitałowym	387
Katarzyna Wójcik, Janusz Tuchowski , Analiza porównawcza miar podobieństwa tekstów opartych na macierzy częstości i tekstów opartych na wiedzy dziedzinowej	396
Iwona Staniec , Analiza czynnikowa w identyfikacji obszarów determinujących doskonalenie systemów zarządzania w polskich organizacjach	406
Marek Lubicz, Maciej Zięba, Adam Rzechonek, Konrad Pawełczyk, Jerzy Kołodziej, Jerzy Błaszczyk , Analiza porównawcza wybranych technik eksploracji danych do klasyfikacji danych medycznych z brakującymi obserwacjami	416
Iwona Foryś , Wykorzystanie analizy log-liniowej do wyboru czynników determinujących atrakcyjność cenową mieszkań w obrocie wtórnym na przykładzie lokalnego rynku mieszkaniowego.....	426
Ewa Genge , Analiza skupień oparta na mieszankach uciętych rozkładów normalnych.....	436
Jerzy Korzeniewski , Ocena efektywności metody uśredniania zmiennych i metody Ichino selekcji zmiennych w analizie skupień	444
Andrzej Dudek , SMS – propozycja nowego algorytmu analizy skupień	451
Artur Mikulec , Metody oceny wyniku grupowania w analizie skupień.....	460
Małgorzata Machowska-Szewczyk , Algorytm klasyfikacji rozmytej dla obiektów opisanych za pomocą zmiennych symbolicznych oraz rozmytych	469
Artur Zaborski , Analiza PROFIT i jej wykorzystanie w badaniu preferencji	479
Karolina Bartos , Analiza skupień wybranych państw ze względu na strukturę wydatków konsumpcyjnych obywateli – zastosowanie sieci Kohonena	488

Barbara Batóg, Magdalena Mojsiewicz, Katarzyna Wawrzyniak , Klasyfikacja gospodarstw domowych ze względu na bodźce do zawierania umowy o ubezpieczenie z wykorzystaniem modeli zmiennych jakościowych .	496
Izabela Kurzawa , Zastosowanie modelu LA/AIDS do badania elastyczności cenowych popytu konsumpcyjnego w gospodarstwach domowych w relacji miasto–wieś	505
Aleksandra Łuczak, Feliks Wysocki , Metody porządkowania liniowego obiektów opisanych za pomocą cech metrycznych i porządkowych	513
Agnieszka Sompolska-Rzechuła , Porównanie klasycznej i pozycyjnej taksonomicznej analizy zróżnicowania jakości życia w województwie zachodniopomorskim	523
Joanna Banaś, Małgorzata Machowska-Szewczyk , Ocena intensywności wykorzystania skrzynek poczty elektronicznej za pomocą uporządkowanego modelu probitowego	532
Iwona Bąk , Segmentacja gospodarstw domowych emerytów i rencistów pod względem wydatków na rekreację i kulturę	541
Aneta Becker , Zastosowanie metody ANP do porządkowania województw Polski pod względem dynamiki wykorzystania ICT w latach 2008-2010	552
Katarzyna Dębowska , Klasyfikacja sektorów ze względu na ich kondycję finansową przy użyciu metod wielowymiarowej analizy statystycznej	562
Anna Domagała , Propozycja metody doboru zmiennych do modeli DEA (procedura kombinowanego doboru w przód).....	571
Henryk Gierszal, Karina Pawlina, Maria Urbańska , Analiza statystyczna w badaniach zapotrzebowania na usługi teleinformatyczne sieci łączności ruchomej	580
Hanna Gruchociak , Konstrukcja estymatora regresyjnego dla danych o strukturze dwupoziomowej.....	590
Tomasz Klimanek, Marcin Szymkowiak , Zastosowanie estymacji pośredniej uwzględniającej korelację przestrzenną w opisie niektórych charakterystyk rynku pracy	601
Jarosław Lira , Prognozowanie opłacalności produkcji żywca wieprzowego w Polsce	610
Christian Lis , Wykorzystanie metody klasyfikacji w ocenie konkurencyjności portów południowego Bałtyku	619
Beata Bieszk-Stolorz, Iwona Markowicz , Wykorzystanie wielomianowego modelu logitowego do oceny szansy podjęcia pracy przez bezrobotnych .	628
Lucyna Przezbórska-Skobiej, Jarosław Lira , Przestrzeń agroturystyczna Polski i ocena jej atrakcyjności.....	637
Paweł Ulman , Model rozkładu wydatków a funkcje popytu.....	646
Maria Urbańska, Tadeusz Mizera, Henryk Gierszal , Zastosowanie metod analizy statystycznej w badaniach mięczaków	655

Summaries

Stanisława Bartosiewicz , The effects of subjectivism in multivariate analysis revisited.....	21
Andrzej Sokółowski , Q universal distance measure	30
Eugeniusz Gatnar , Data quality in central banks' statistical systems (NBP example)	38
Marek Walesiak , Distance measures for ordinal data – strategies of proceedings.....	46
Krzysztof Jajuga, Marek Walesiak , XXV years of taxonomic conferences – some facts and remarks.....	49
Józef Pocięcha, Barbara Pawelek , General SEM model in researching corporate bankruptcy and business cycles – theoretical and practical problems.....	57
Paweł Lula , Learning-based systems of information extraction from textual resources	67
Ewa Roszkowska , The application of the TOPSIS method to support the negotiation process	75
Andrzej Młodak , Neighborhood of spatial areas in the physical and socio-economic context – a taxonomic approach.....	85
Andrzej Bąk , Models for unordered categories in preference analysis.....	95
Kowalewski Jacek , An integrated model of optimizing statistical surveys	105
Jan Paradysz, Karolina Paradysz , Areas of unemployment in Poland – benchmark problem	115
Tomasz Szubert , How to play to lose the least? Classification of systems in sports bets	125
Izabela Szamrej-Baran , Classification of EU member states in view of fuel poverty	134
Sylvia Filas-Przybył, Tomasz Klimanek, Jacek Kowalewski , An attempt to use the gravity model in the analysis of commuters.....	143
Marta Dziechciarz-Duda, Anna Król, Klaudia Przybysz , Subsistence minimum versus factors influencing tendency to benefit from social care. Classification of households	152
Hanna Dudek , Subjective equivalence scales – analysis based on data about satisfaction with incomes.....	162
Joanicjusz Nazarko, Ewa Chodakowska, Marta Jarocka , Segmentation of universities using cluster analysis versus technological competitors determined by the DEA method – a comparative study	172
Ewa Chodakowska , Selected methods of classification in schools' rating.....	181
Bartosz Soliński , Renewable energy sector in the European Union – classification in the light of change management strategy	191
Krzysztof Szwarc , Classification of Wielkopolska voivodeship due to the demographic situation	201

Elżbieta Gołata, Grażyna Dehnel , Administrative registers in business analysis.....	211
Katarzyna Chudy, Marek Sobolewski, Kinga Stępień , Application of taxonomic methods in forecasting the profitability ratios of listed banks in Poland.....	221
Katarzyna Dębowska , Modeling bankruptcy of firms by using discrimination and regression methods.....	230
Alina Bojan , Identification of variables which influence attractiveness of given investments with the usage of multivariate analysis.....	239
Justyna Brzezińska , Log-linear analysis in the study of mortality in EU.....	246
Aneta Rybicka, Bartłomiej Jefmański, Marcin Pelka , Latent class analysis in student satisfaction surveys.....	254
Bartłomiej Jefmański , The respondent's opinions measurement in the R program with an application of fuzzy sets theory.....	264
Julita Stańczuk , A comparison of the results of multistate classification of economic objects using discriminant analysis and artificial neural networks.....	274
Jerzy Krawczuk , Effectiveness of classification methods in S&P500 stock index direction changes forecasting.....	282
Anna Czapkiewicz, Beata Basiura , The simulation study of the utility of the Copula-GARCH models for clustering financial time series.....	290
Radosław Pietrzyk , Timing and selectivity in mutual funds performance measurement.....	305
Aleksandra Witkowska, Marek Witkowski , Use of the Panzar-Rosse method to assess of the competition level in the cooperative banks sector.....	314
Marcin Pelka , Ensemble learning with the application of <i>boosting</i> in symbolic data analysis.....	322
Justyna Wilk , Comparative study of symbolic data classification software.....	332
Tomasz Bartłomowicz, Justyna Wilk , Application of symbolic data analysis methods for domain database searching.....	341
Kamila Migdał-Najman , A proposal of hybrid clustering method based on self-learning networks.....	351
Dorota Rozmus , Comparison of accuracy of spectral clustering and cluster ensembles stability based on bagging idea.....	360
Krzysztof Najman , A dynamic grouping based on self-learning GNG networks.....	369
Małgorzata Misztal , Influence of data imputation methods on the results of object classification using classification trees in the case of small data sets – simulation assessment.....	379
Mariusz Kubus , The application of pre-conditioning of explanatory variable for feature selection.....	386
Barbara Batóg, Jacek Batóg , Application of discriminant analysis to the identification of factors determining the rate of return on the capital market.....	395

Katarzyna Wójcik, Janusz Tuchowski , Comparative analysis of text documents similarity measures based on frequency matrix and based on domain knowledge.....	405
Iwona Staniec , Factor analysis in the identification of areas that determine the improvement of management systems in Polish organizations.....	415
Marek Lubicz, Maciej Zięba, Adam Rzechonek, Konrad Pawełczyk, Jerzy Kołodziej, Jerzy Błaszczyk , Comparative analysis of selected data mining approaches to the classification of medical data with missing values (covariates).....	425
Iwona Foryś , The log-linear analysis using to select the factors determining the attractiveness of the price of flats on the secondary market on the example of local housing market.....	435
Ewa Genge , Trimming approach to the mixtures of normal distributions.....	443
Jerzy Korzeniewski , Efficiency assessment of Ichino method and mean value method of selecting variables in cluster analysis.....	450
Andrzej Dudek , SMS – proposal of new clustering algorithm.....	459
Artur Mikulec , Evaluation methods for the grouping result in cluster analysis.....	468
Małgorzata Machowska-Szewczyk , Fuzzy clustering algorithm for objects described by symbolic or fuzzy variables.....	478
Artur Zaborski , PROFIT analysis and its using in the research of preferences.....	487
Karolina Bartos , Cluster analysis of selected countries due to the structure of their citizens' consumer expenditures – the use of Kohonen networks.....	495
Barbara Batóg, Magdalena Mojsiewicz, Katarzyna Wawrzyniak , Classification of households according to the impulses of concluding the insurance contract by means of qualitative variable models.....	504
Izabela Kurzawa , The application of LA/AIDS model to examine price elasticities of demand of households in the urban-rural relationship.....	512
Aleksandra Luczak, Feliks Wysocki , Linear ordering methods of objects described by a set of metric and ordinal characteristics.....	522
Agnieszka Sompolska-Rzechuła , The comparison of the classical and positional taxonomic analysis of the quality of life differentiation in Zachodniopomorskie voivodeship.....	531
Joanna Banaś, Małgorzata Machowska-Szewczyk , Evaluation of intensity of mailboxes using with the ordered probit model.....	540
Iwona Bąk , Segmentation of pensioners and annuitants households in terms of expenditures on recreation and culture.....	551
Aneta Becker , Application of ANP method to organize Polish voivodships in terms of dynamics of the use of ICT in 2008-2010.....	561
Katarzyna Dębowska , The classification of sectors' financial situation using the methods of multivariate statistical analysis.....	570

Anna Domagała , Proposal of a new method for variable selection in DEA models (combined forward stepwise selection method).....	579
Henryk Gierszal, Karina Pawlina, Maria Urbańska , Statistical analysis in demand research of ICT services in mobile networks.....	589
Hanna Gruchociak , Construction of regression estimator for two-level data	600
Tomasz Klimanek, Marcin Szymkowiak , Application of spatial models in indirect estimation of some labor market characteristics	609
Jarosław Lira , Forecasting of hog livestock production profitability in Poland	618
Christian Lis , The utilization of taxonomic methods in the appraisal of competitiveness of south Baltic ports	627
Beata Bieszk-Stolorz, Iwona Markowicz , The application of the multinomial logit model in evaluating employment odds for the unemployed job seekers	636
Lucyna Przezbórska-Skobiej, Jarosław Lira , Agritourism space of Poland and its valuation.....	645
Paweł Ulman , Model of expenses distribution and demand functions.....	654
Maria Urbańska, Tadeusz Mizera, Henryk Gierszal , Methods of statistical analysis in research of molluscs	663

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 pożądaných 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 statystycznych¹, 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

$$Y^T = [y_1, y_2, y_3, \dots, y_n]. \quad (1)$$

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

$$X^T = [x_1, x_2, x_3, \dots, x_m]. \quad (2)$$

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

$$\Omega: X \times \dots \times X \rightarrow Y. \quad (3)$$

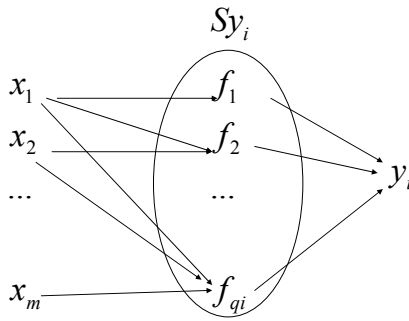
Transformacje mogą mieć różny charakter, a oszacowanie wartości określonej zmiennej wyjściowej y_i uzyskiwane jest drogą przekształcenia jednej lub kilku zmiennych wejściowych

$$y_i = f(X). \quad (4)$$

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

$$Sy_i = \{f_1^i(X), \dots, f_{q_i}^i(X)\}, \quad (5)$$

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 **badania**. Założono, że istnieje skończony zbiór r możliwych do przeprowadzenia różnych badań, które można przedstawić w postaci wektora

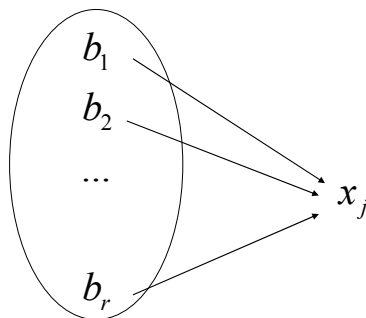
$$B^T = [b_1, b_2, b_3, \dots, b_r]. \quad (6)$$

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

$$\Phi : B \rightarrow X. \quad (7)$$

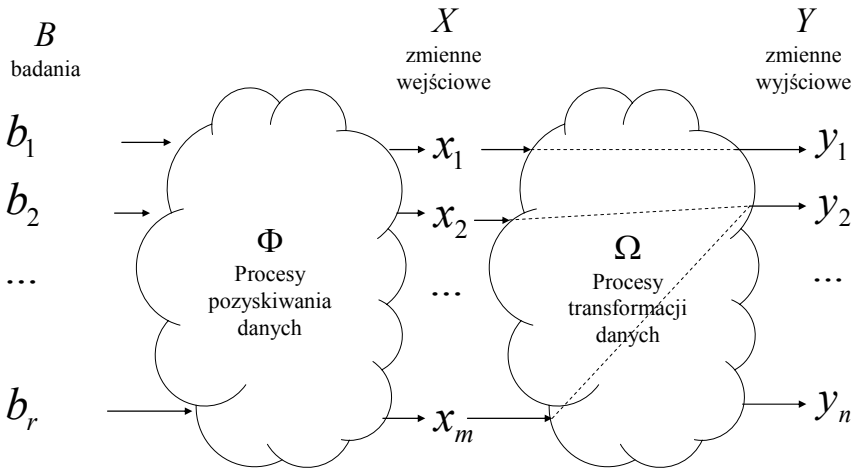
Przyjęto, że oszacowania wartości zmiennej x_j dokonuje się bezpośrednio w badaniu (bez dalszych transformacji). Założono także możliwość szacowania konkretnej, tej samej zmiennej x_j w różnych badaniach (rys. 2).



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 wyjściowych (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

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

Rozwiązując zadanie, należy szczegółowo ustalić

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

$$BX = \begin{bmatrix} bx_1^l & \dots & bx_1^l & \dots & bx_1^r \\ \dots & & bx_j^l & & \dots \\ & & \dots & & \\ bx_m^l & \dots & bx_m^l & \dots & bx_m^r \end{bmatrix}, \quad (9)$$

gdzie: bx_j^l – binarna zmienna określająca, 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

$$Sy^* = \{Sy_1^*, \dots, Sy_i^*, \dots, Sy_n^*\} \quad (10)$$

$$Sy_i^* = \{f_k^i(X) \in Sy_i\},$$

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

$$Z = \{Y, BX, Sy^*\}. \quad (11)$$

4. Funkcje kryterium

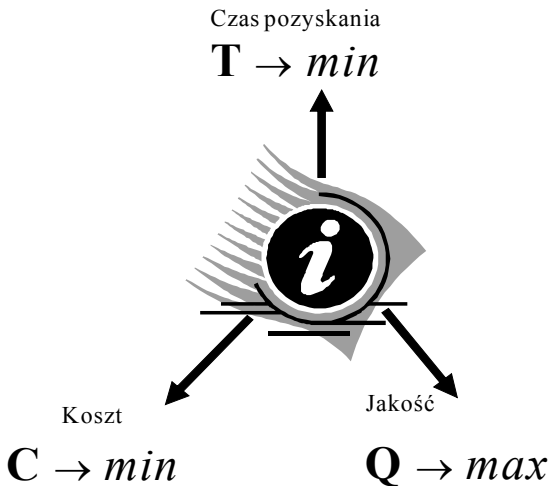
Znalezienie optymalnego sposobu przeprowadzania badań statystycznych wymaga wprowadzenia stosownych funkcji celu. Zakładając, że celem całego procesu badań statystycznych jest uzyskanie odpowiednich zmiennych wyjściowych Y , które można traktować jako informacje będące efektem działania systemu, niezbędne jest ustalenie podstawowych cech, które powinny charakteryzować dobrą informację². 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 informacji³ ($T(y)$) oraz wysoką jakość informacji⁴ ($Q(y)$), co można zapisać jako:

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

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]. Próbuąc 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

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

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

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

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

$$Q(y_i) = f(f_k^i(X), XQ), \quad (15)$$

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 pożądaných zmienných wyjściowych z wektora Y , którą można przedstawić jako całkowity koszt przeprowadzenia wszystkich badań (CY):

$$CY = \sum_{l=1}^r CB(b_l) \cdot \text{sgn}(n_l) \quad (16)$$

$$n_l = \sum_{j=1}^m bx_j^l, \forall l = \overline{1, \dots, r}$$

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

bx_j^l – binarna zmienna określająca, 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 zmienných wejściowych,

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

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

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

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 poszczególných zmienných wyjściowych⁵ ($Q(y_i)$)

$$QY = f(Q(y_1), \dots, Q(y_n)). \quad (18)$$

⁵ Ze wstępnego rozeznania wynika, że przy posiadaniu miary jakości przypisanej do poszczególných zmienných 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 pożądaných 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.