POLSKA AKADEMIA NAUK
KOMITET STATYSTYKI I EKONOMETRII

PRZEGLĄD STATYSTYCZNY

STATISTICAL REVIEW

TOM 56

2

2009

DOM WYDAWNICZY ELIPSA

DOM WYDAWNICZY ELIPSA
WARSZAWA 2009

WYDAWCA

Komitet Statystyki i Ekonometrii Polskiej Akademii Nauk

Wydanie publikacji dofinansowane przez Ministra Nauki i Szkolnictwa Wyższego

Strona www „Przeglądu Statystycznego":
http://www.przegladstatystyczny.pan.pl

Nakład 300 egz.

DOM WYDAWNICZY ELIPSA

Realizacja wydawnicza:
Dom Wydawniczy ELIPSA,
ul. Inflancka 15/198, 00-189 Warszawa
tel./fax (0-22) 635 03 01, 635 17 85, e-mail: elipsa@elipsa.pl
www.elipsa.pl

PRZEGLĄD STATYSTYCZNY
R. LVI – ZESZYT 2 – 2009

BOGUSŁAW GUZIK

UWAGI NA TEMAT ZASTOSOWANIA METODY DEA DO USTALANIA ZDOLNOŚCI KREDYTOWEJ

1. WSTĘP

Artykuł nawiązuje do podejścia proponującego wykorzystanie metod *Data Envelopment Analysis* do ustalania zdolności kredytowej. Sprawie tej poświęcono wiele prac. Za autorów podejścia uchodzi m.in. Truott, Rai, Zhang [14]; Yeh [15]; Emel, Oral, Reisman, Yolalan [8]. W latach późniejszych pojawiły się ważne prace: Paradi, Asmild, Simak [13]; Chang, Chiang, Tang [7]; Min, Lee [12], a w literaturze krajowej: Gospodarowicz [10], Feruś [9]. Podejście DEA do analizy ryzyka kredytowego bardzo wyraźnie rozszerza krąg ekonomicznych zastosowań tej metody i czyni ją jeszcze bardziej atrakcyjną dla specjalistów. Wzbogaca również krąg metod *credit scoringu* o bardzo ciekawą, stosunkowo prostą i użyteczną technikę analityczną, co z kolei powinno być atrakcyjne dla ekonomistów.

Wydaje się jednak, że dotychczasowe „klasyczne" propozycje zastosowania DEA w ocenie ryzyka kredytowego mają pewne wady i dlatego należy wprowadzić do nich stosowne modyfikacje i uzupełnienia. Dyskusja „wad" klasycznego podejścia oraz propozycje jego uzupełnień są właśnie celem niniejszego artykułu.

2. KLASYCZNY SCHEMAT ZASTOSOWANIA DEA W *CREDIT SCORINGU*

Tradycyjne podejście DEA do analizy ryzyka kredytowego jasno i kompetentnie opisano w cytowanych pracach Min, Lee [12] oraz Emel, Oral, Reisman, Yolalan [8]. Na ich tle scharakteryzujemy omawianą problematykę. Etapy badania według [12], s. 1763-1764 są następujące:

1. Wybór zbioru obserwacji.
2. Identyfikacja wstępnych wskaźników finansowych.
3. Ustalenie końcowej listy wskaźników finansowych.
4. Kalkulacja wskaźnika zdolności kredytowej za pomocą DEA.
5. Konstrukcja funkcji dyskryminacyjnej. Walidacja za pomocą metod regresji, metod dyskryminacyjnych oraz testowanie aktualnych przypadków bankructwa.
6. Propozycja metody ratingu kredytowego wykorzystującej statystyczny rozkład wskaźnika zdolności kredytowej.

Cytowane uporządkowanie etapów badania i zamieszczone w artykule [12]oraz szerokie do nich komentarze są bardzo pomocne i wartościowe dla analityków bankowych. W niniejszym artykule odniesiemy się tylko do tych etapów, w których bezpośrednio wykorzystuje się metodę DEA lub jej wyniki.

3. KALKULACJA WSKAŹNIKÓW ZDOLNOŚCI KREDYTOWEJ ZA POMOCĄ DEA

3.1. UKIERUNKOWANY NA NAKŁADY STANDARDOWY MODEL CCR

W klasycznych propozycjach sugeruje się wykorzystanie modelu CCR[1]. Z formalnego punktu widzenia model ten jest pewnym (stosunkowo prostym) zadaniem programowania liniowego. Istnieje kilka jego wersji: (a) standardowa lub kanoniczna; (b) ukierunkowana na rezultaty lub na nakłady; (c) pierwotna (mnożnikowa) lub dualna (obwiedniowa).

Dla wyjaśnienia dodajmy, że model jest *standardowy*, gdy wszystkie ograniczenia mają charakter nierówności słabych, a kanoniczna, gdy ograniczenia są równaniami (postać kanoniczna, to wersja z luzami, „slack'ami"). Model jest *ukierunkowany na nakłady*, gdy postuluje się minimalizację nakładów przy dolnych ograniczeniach na wielkość rezultatów, a ukierunkowany na rezultaty, gdy postuluje się maksymalizację rezultatów przy górnych ograniczeniach na wielkość nakładów. W klasycznej pracy Charnesa, Coopera, Rhodesa [5] *modelem pierwotnym* nazwano wersję, w której wyznacza się jednostkowe wyceny nakładów oraz rezultatów, tzw. mnożniki. Modelem dualnym natomiast nazwano wersję, w której wyznacza się tzw. wagi intensywności służące do określenia optymalnej technologii zbioru obiektów. Ustalenie, co jest wersją pierwotną a co dualną, jest jednak sprawą konwencji, gdyż z matematycznego punktu widzenia model dualny do modelu dualnego to model pierwotny. Dlatego też pojawiają się propozycje, by zamiast nazw model „pierwotny" czy „dualny" używać terminów model mnożnikowy (*multiplier model*) – wtedy, gdy wyznacza się wyceny jednostkowe, czyli mnożniki oraz model obwiedniowy (*envelopment model*) – wtedy, gdy wyznacza się wagi intensywności) por. np. [6]. Model mnożnikowy nazwać można modelem wycen jednostkowych, a obwiedniowy – modelem struktury technologii lub modelem wag intensywności.

Ukierunkowany na nakłady standardowy (obwiedniowy) model CCR ma postać:

Dane

– Lista obiektów, $j = 1, \ldots, J$. W odniesieniu do problematyki badania zdolności kredytowej będzie to zbiór wniosków kredytowych, firm, projektów, wnioskodawców itp.

– Lista nakładów $n = 1, \ldots, N$ oraz lista rezultatów $r = 1, \ldots, R$. W problematyce *credit scoring* nakładami są te wielkości, które wnioskujący muszą ponosić, aby uzyskiwać rezultaty określające zdolność kredytową[2].

---

[1] Model CCR opracowany został przez Charnsa, Coopera, Rhodesa [5]

[2] Np. w [12], s. 1767 rozpatruje się trzy „nakłady" w sensie DEA: (1) wskaźnik wydatków finansowych do sprzedaży, (2) wskaźnik zobowiązań bieżących, (3) relacja pożyczek do aktywów ogółem oraz trzy rezultaty: (1) wskaźnik adekwatności kapitałowej, (2) wskaźnik płynności bieżącej, (3) wskaźnik zyskowności. Natomiast w [9], s. 53 rozpatruje się dwa nakłady: (1) wskaźnik rotacji aktywów w dniach, (2) wskaźnik ogólnego zadłużenia oraz cztery rezultaty: (1) stopa zysku netto, (2) ROA, (3) ROE, (4) wskaźnik płynności bieżącej.

– Wielkości rezultatów w poszczególnych obiektach, y_{rj} ($r = 1, \ldots R$; $j = 1, \ldots, J$).
– Wielkości nakładów w poszczególnych obiektach, x_{nj} ($n = 1, \ldots N$; $j = 1, \ldots, J$).
Wszystkie rezultaty są nieujemne, a nakłady są dodatnie.

Rezultaty mają charakter maksymant (stymulant): wyższa wartość rezultatu oznacza większą zdolność kredytową.

Nakłady to wielkości lub okoliczności, które są niezbędne dla uzyskania rezultatów. Wyższy poziom rezultatu wymaga – *ceteris paribus* – wyższego poziomu nakładu.

W celu wyznaczenia efektywności dla każdego obiektu konstruuje się jego dotyczące zadanie decyzyjne. Zadanie dla obiektu *o*-tego ma postać:

Zmienne decyzyjne

θ_o – *mnożnik poziomu nakładów obiektu o-tego,*

λ_{oj} – *waga intensywności technologii obiektu* j-ego ($j = 1, \ldots, J$).

Waga intensywności λ_{oj} określa krotność technologii empirycznej obiektu *j*-ego występującą w technologii wspólnej zbioru wszystkich obiektów. W modelu CCR obiekt *o*-ty też może tworzyć technologię wspólną. Technologia wspólna zorientowana jest na uzyskanie rezultatu nie mniejszego od rezultatów obiektu *o*-tego przy nakładach nie większych od poczynionych przez ten obiekt. Nakład *n*-ty w technologii wspólnej wynosi:

$$\bar{x}_{no} = \sum_{j=1}^{J} \lambda_{oj} x_{nj} \qquad (n = 1, \ldots, N) \tag{1}$$

a rezultat *r*-ty:

$$\bar{y}_{ro} = \sum_{j=1}^{J} \lambda_{oj} y_{rj} \qquad (r = 1, \ldots, R) \tag{2}$$

Mnożnik θ_o określa krotność nakładów obiektu *o*-tego, jaką muszą zastosować wszystkie obiekty w swej technologii wspólnej, aby uzyskać rezultaty obiektu *o*-tego.

Funkcja celu

$$\min \theta_o \tag{3}$$

Warunki ograniczające i znakowe

$$\bar{y}_{ro} \geq y_{ro} \qquad (r = 1, \ldots, R), \tag{4}$$

$$\bar{x}_{no} \leq \theta_o x_{no} \qquad (n = 1, \ldots, N), \tag{5}$$

$$\theta_o, \lambda_{oj} \geq 0 \qquad (j = 1, \ldots, J). \tag{6}$$

Niekiedy dodaje się warunek zmiennych korzyści skali, tzw. warunek BCC:

$$\sum_{j=1}^{J} \lambda_{oj} = 1 \tag{7}$$

wówczas mowa jest o zadaniu BCC[3]. Niekiedy wreszcie dodaje się też warunek:

$$\theta_o \leq 1. \tag{8}$$

który jednak w zadaniu CCR oraz BCC spełniony jest „automatycznie" i dlatego można go pominąć (w innych zadaniach może on być konieczny).

Wskaźnikiem efektywności obiektu *o*-tego jest uzyskany na podstawie modelu CCR *optymalny* mnożnik nakładów, $\tilde{\theta}_o$. Określa on, jaką przynajmniej krotność nakładów obiektu *o*-tego musiałyby wykorzystać wszystkie obiekty w swej optymalnej technologii wspólnej, aby otrzymać rezultaty nie gorsze od uzyskanych przez obiekt *o*-ty. Obiekt jest w pełni efektywny (w sensie Farrella), gdy $\tilde{\theta}_o = 1$, co pomijające rzadkie szczególne przypadki ma miejsce, gdy zorientowana na obiektu *o*-ty optymalna technologia wspólna redukuje się do technologii empirycznej obiektu *o*-tego. Jeśli $0 < \tilde{\theta}_o < 1$, to obiekt *o*-ty jest nieefektywny, bowiem wspólna technologia[4] dla uzyskania rezultatów obiektu *o*-tego potrzebuje mniej nakładów niż miało to miejsce w tym obiekcie.

W problematykę zdolności kredytowej „rezultatami" są zmienne charakteryzujące różne strony zdolności kredytowej, a „nakładami" – zmienne charakteryzujące wielkości niezbędne dla uzyskania zdolności kredytowej. W tym sensie analogia między badaniem zdolności kredytowej a badaniem efektywności na podstawie modelu DEA jest oczywista:

Wskaźnik efektywności $\tilde{\theta}_o$ można traktować jako syntetyczny wskaźnik zdolności kredytowej. Obiekt o wyższym poziomie wskaźnika efektywności to obiekt o wyższej zdolności kredytowej.

Przechodzimy do omówienia i analizy głównych wątpliwości dotyczących tradycyjnego nurtu zastosowań DEA do analizy zdolności kredytowej.

3.2. PROBLEM 1 – CZY POTRZEBNY JEST POSTULAT ZMIENNYCH KORZYŚCI SKALI?

Pierwsza wątpliwość dotyczy warunku BCC. Wyraża on postulat, by technologia wspólna była liniową wypukłą kombinacją liniową technologii empirycznych, czyli ich średnią ważoną. Wagami są współczynniki λ_{oj} ($j = 1, \ldots, J$).

Z matematycznego i ekonomicznego punktu widzenia jest to bardzo wygodne założenie, gdyż pozwala dowieść wielu ciekawych własności ekonomicznych i matematycznych modelu DEA. Z praktycznego jednak punktu widzenia jest to założenie bardzo silne i słuszne jedynie w odniesieniu do obiektów o mniej więcej tej samej skali. Jeśli natomiast obiekty są różnej skali założenie (7) prowadzi do deformacji rozwiązania.

Dla przykładu weźmy pod uwagę warunek (4) dotyczący rezultatu numer *r*. Zapiszemy go w nieco innej formie:

$$\sum_{j=1}^{J} \lambda_{oj} y_{rj} \geq y_{ro} \qquad (1 \leq r \leq R). \tag{4'}$$

[3] W odniesieniu do analizy *credit scoring* warunek taki sugeruje się np. w pracach [14] s. 407 oraz [13] s. 155. Warunek BCC pochodzi od Bankera, Charnesa, Coopera [3]. Cały model (3)-(7), będący rozszerzeniem modelu CCR o warunek (7), nazywany jest modelem BCC.

[4] Z uwagi na strukturę matematyczną zadania, nie ma w niej wtedy technologii obiektu *o*-tego.

Jeśli obiekt *o*-ty jest duży w tym sensie, że jego rezultat jest większy od rezultatu w każdym innym obiekcie, a więc jeśli $y_{ro} > y_{rj}$ ($j = 1, \ldots, J$; $j \neq o$), to dla obiektu *o*-tego nie istnieją takie λ_{oj} ($j \neq o$), które jednocześnie: są nieujemne, sumują się do 1, spełniają warunek (4)'. Jeśli bowiem wszystkie $y_{rj} < y_{ro}$ ($j \neq o$), to średnia ważona takich wartości y_{rj} musi być mniejsza od y_{ro}.

Jedyną wchodzącą w grę wartością jest wówczas $\lambda_{o,o} = 1$, a to oznacza, że jedynym rozwiązaniem modelu BCC jest to, że obiekt *o*-ty jest w pełni efektywny[5].

Podobnie jest, gdy obiekt *o*-ty jest mały w tym sensie, że jego *n*-ty nakład jest mniejszy od tego nakładu w każdym innym obiekcie, a więc gdy $x_{no} < x_{nj}$ ($j = 1, \ldots, J$; $j \neq o$). Wówczas nie istnieją takie λ_{oj} ($j \neq o$) nieujemne i sumujące się do 1, aby możliwe było spełnienie warunku:

$$\sum_{j \neq o} \lambda_{oj} x_{nj} \leq \theta_o x_{no} \qquad (1 \leq 1 \leq N). \tag{5)'}$$

gdy $\theta_o \leq 1$. Jedynym rozwiązaniem jest $\lambda_{o,o} = 1$, co oznacza, że efektywność $\theta_o = 1$.

Generalny wniosek jest następujący:

Wprowadzenie bardzo często rekomendowanego warunku BCC może doprowadzić do sytuacji, że wiele obiektów będzie w pełni efektywnych, co w odniesieniu do problematyki *credit scoring* oznaczałoby, że wiele obiektów jest „zdrowych" ekonomicznie i ma najwyższą zdolność kredytową. Niestety, ów wysoki wskaźnik efektywności (czyli „zdolności kredytowej") często będzie jednak wynikał tylko z formalnych konsekwencji dostosowania się zadania do postulatu BCC, a nie z rzeczywistych osiągnięć obiektu.

Propozycja rozwiązania problemu 1: *należy zrezygnować z postulatu zmiennych korzyści skali*

Warunek (7) można postawić tylko wtedy, gdy obiekty są mniej więcej tej samej skali i gdy rzeczywiście jesteśmy pewni, że mają miejsce tzw. zmienne korzyści skali.

3.3. PROBLEM 2 – REDUNDANCJA LICZBY OBIEKTÓW NAJLEPSZYCH, JAK PRZED NIĄ SIĘ UCHRONIĆ?

W modelu CCR oraz w wielu innych modelach DEA, w których w technologii wspólnej zbioru obiektów może uczestniczyć technologia empiryczna badanego obiektu dzieje się tak, że liczba obiektów o największej efektywności (równej 1) może być bardzo duża – niekiedy połowa a nawet więcej. Dochodzi więc do swego rodzaju redundancji (nadmiarowości) obiektów najlepszych. Trudno bowiem przyjąć, że połowa czy więcej obiektów, to obiekty, co do których nie można mieć żadnych zastrzeżeń i które są wzorcowe. Dodatkowo dochodzi jeszcze kłopot interpretacyjny, gdyż w sytuacji, gdy wiele obiektów ma wskaźnik efektywności równy 1, nie można utworzyć pełnego

[5] Dodajmy, że z tego właśnie powodu efektywność obiektu w sensie BCC jest nie mniejsza od efektywności w sensie CCR, a liczba obiektów w pełni efektywnych w modelu BCC jest nie mniejsza od występującej w modelu CCR.

rankingu obiektów. Obiekty z efektywnością 1 muszą być plasowane na tym samym, pierwszym, miejscu.

Propozycja rozwiązania problemu 2: *dla uzyskania pełnego rankingu wniosków należy zastosować model SE-CCR*

Model nadefektywności (*super-efficiency*) CCR, kodowany dalej jako SE-CCR[6], jest bardzo prostą modyfikacją modelu CCR polegającą na wykluczeniu obiektu *o*-tego ze zbioru obiektów tworzących technologię wspólną zorientowaną na obiekt *o*-ty. Formalnie można go zapisać następująco:

Dane – jak w CCR

Zmienne decyzyjne

ρ_o – *mnożnik poziomu nakładów obiektu o-tego,*

λ_{oj} ($j = 1, \ldots, J$ oprócz $j = o$) – *wagi intensywności.*

Funkcja celu

$$\min \rho_o \tag{9}$$

Warunki ograniczające i znakowe

$$\sum_{j \neq o} \lambda_{oj} y_{rj} \geq y_{ro} \qquad (r = 1, \ldots, R), \tag{10}$$

$$\sum_{j \neq o} \lambda_{oj} x_{nj} \leq \rho_o x_{no} \qquad (n = 1, \ldots, N), \tag{11}$$

$$\rho_o, \lambda_{oj} \geq 0 \qquad (j \neq o). \tag{12}$$

Uzyskany na podstawie modelu SE-CCR optymalny mnożnik nakładów $\tilde{\rho}_o$ jest *wskaźnikiem rankingowym* obiektu *o*-tego. Określa on, jaką przynajmniej krotność nakładów obiektu *o*-tego musiałyby wykorzystać pozostałe obiekty w swej optymalnej technologii wspólnej, aby otrzymać rezultaty nie gorsze od uzyskanych przez obiekt *o*-ty.

Obiekt jest w pełni efektywny w sensie Farrela (czyli w sensie CCR), gdy $\tilde{\rho}_o \geq 1$, a więc gdy inni – nawet w swej optymalnej technologii wspólnej – dla uzyskania rezultatów otrzymanych przez obiekt *o*-ty, muszą poczynić nakłady nie mniejsze od poniesionych przez obiekt *o*-ty. W przypadku $\tilde{\rho}_o < 1$ obiekt *o*-ty jest nieefektywny, bowiem inne obiekty dla uzyskania rezultatów obiektu *o*-tego potrzebują mniej nakładów niż ich obiekt ten zużył. Efektywność w sensie Farrella obiektu *o*-tego wynosi:

$$\tilde{\theta}_o = \begin{cases} 1 & \text{gdy} \quad \tilde{\rho}_o \geq 1 \\ \tilde{\rho}_o & \text{gdy} \quad \tilde{\rho}_o < 1 \end{cases}. \tag{13}$$

Korzystając ze wskaźników rankingowych można zaproponować nowy wskaźnik efektywności:

[6] Model ten lub jego idee proponowano w pracach Banker i Gilford [4] oraz Andersen-Petersen [2].

$$\tilde{\pi}_o = \frac{\tilde{\rho}_o}{M} \tag{14}$$

gdzie:

$$M = \max\{\tilde{\rho}_1, \ldots, \tilde{\rho}_J\}.$$

Wskaźnik $\tilde{\pi}_o$ jest unormowany na przedziale [0, 1]. Obiekt jest tym bardziej efektywny, im $\tilde{\pi}_o$ jest większe. Efektywność w sensie tego miernika nazywać będziemy π-efektywnością, a efektywność w sensie modelu CCR – θ-efektywnością.

Przykład

W tabeli 1 podano informacje opisujące kształtowanie się sześciu nakładów: $X_1, \ldots, X_6$ oraz trzech rezultatów: Y_1, Y_2, Y_3 w 19 obiektach. Przykład jest umowny[7], gdyż niestety, we wszystkich znanych autorowi pracach dotyczących zastosowania metod DEA do analizy *credit scoring*, brak jest danych oryginalnych (lub ich fragmentów), pozwalających na sprawdzenie obliczeń oraz kontynuację badań.

Tabela 1

Informacje statystyczne

Nakład/ Rezultat	O_1	O_2	O_3	O_4	O_5	O_6	O_7	O_8	O_9	O_{10}
X_1	31955	15647	9852	7287	3320	5591	7467	5089	5627	5387
X_2	2654	1498	1148	717	580	651	494	323	386	440
X_3	58938	31778	35594	12988	23044	10152	17607	14937	11550	9713
X_4	19836	18714	15935	19764	11077	13223	9245	3866	6879	3871
X_5	14917	4771	14260	5208	12620	10983	3597	5347	3492	2038
X_6	82900	51793	41438	38561	24669	18871	24169	16069	15550	14002
Y_1	13431	10513	8657	13517	2844	9364	3155	1123	2181	4145
Y_2	3785	2377	1955	936	724	1026	1034	651	780	459
Y_3	2254	2279	1358	824	855	1188	1331	603	572	228

Nakład/ Rezultat	O_{11}	O_{12}	O_{13}	O_{14}	O_{15}	O_{16}	O_{17}	O_{18}	O_{19}
X_1	2017	1339	1552	1715	982	1772	783	1164	856
X_2	102	60	80	117	56	87	93	67	68
X_3	6150	7038	2176	5794	643	1688	1654	1026	1029

[7] Choć dane pochodzą z autentycznych projektów empirycznych.

cd. tabeli 1

Nakład/ Rezultat	O_{11}	O_{12}	O_{13}	O_{14}	O_{15}	O_{16}	O_{17}	O_{18}	O_{19}
X_4	1646	839	4636	1386	5336	254	528	285	132
X_5	3960	4859	6806	652	1383	1037	193	123	289
X_6	5899	4249	1173	6717	4800	1442	2211	1461	989
Y_1	2777	2059	1490	578	667	771	206	381	40
Y_2	314	208	105	167	137	230	72	110	132
Y_3	225	200	68	144	112	166	44	44	11

Źródło: dane umowne.

Wskaźniki θ-efektywności oraz wskaźniki rankingowe z modelu SE-CCR i wskaźniki π-efektywności podano w tabeli 2. Przedstawiono również ranking według wskaźnika θ oraz według wskaźnika rankingowego ρ (lub, co na to samo wychodzi, według wskaźnika π).

Tabela 2

Wskaźniki efektywności i ranking obiektów na podstawie modeli CCR oraz SE-CCR

Obiekt	θ-efektywność	Ranking według θ-efektywności	Wskaźnik rankingowy ρ	π-efektywność	Ranking według π-efektywności
O_1	0,886	17	0,886	0,283	17
O_2	1	1	1,426	0,457	12
O_3	1	1	1,139	0,365	15
O_4	1	1	1,944	0,622	5
O_5	1	1	1,212	0,388	13
O_6	1	1	1,651	0,528	7
O_7	1	1	1,976	0,632	4
O_8	1	1	1,188	0,380	14
O_9	1	1	1,038	0,332	16
O_{10}	1	1	1,430	0,458	11
O_{11}	1	1	1,502	0,481	10
O_{12}	1	1	1,566	0,501	9
O_{13}	**1**	**1**	**2,376**	**0,760**	**2**
O_{14}	0,740	18	0,740	0,237	18

cd. tabeli 2

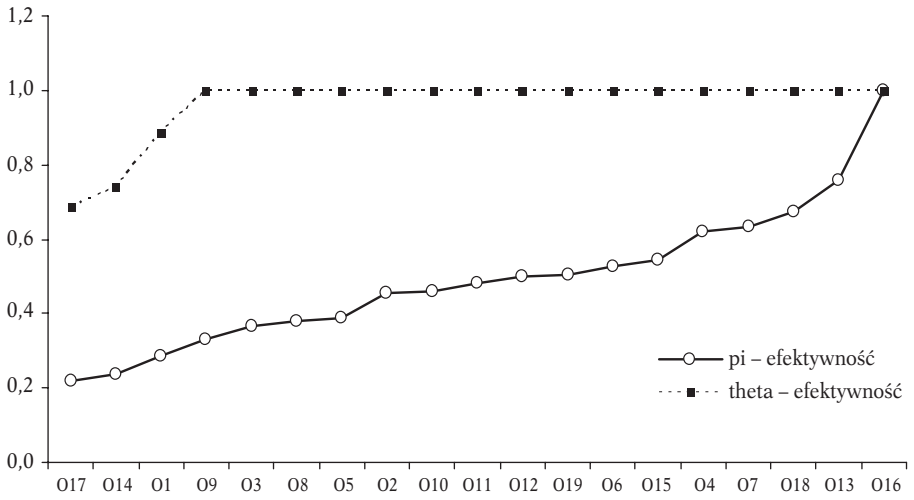
Obiekt	θ-efektywność	Ranking według θ-efektywności	Wskaźnik rankingowy ρ	π-efektywność	Ranking według π-efektywności
O_{15}	1	1	1,700	0,544	6
O_{16}	**1**	**1**	**3,124**	**1**	**1**
O_{17}	0,686	19	0,686	0,219	19
O_{18}	**1**	**1**	**2,103**	**0,673**	**3**
O_{19}	1	1	1,575	0,504	8

Źródło: obliczenia własne.

Tradycyjny model CCR sugeruje, że aż 16 na 19 obiektów jest efektywnych, co w odniesieniu do badania zdolności kredytowej oznaczałoby, że aż 16 na 19 wniosków odznacza się znakomitą zdolnością kredytową.

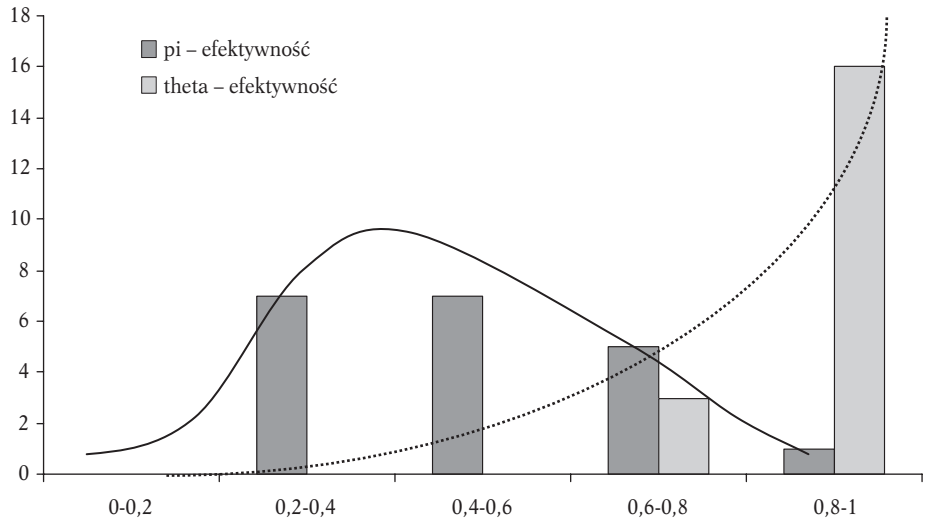
Inne wnioski wynikają natomiast z modelu SE-CCR. Według niego istnieje jeden, zdecydowanie najlepszy obiekt O_{16}. Efektywność kilku pozostałych (O_4, O_6, O_7, O_{12}, O_{13}, O_{15}, O_{18}, O_{19}) plasuje się w granicach 50-80% efektywności obiektu najlepszego, a pozostałych wynosi mniej, nawet tylko 20% (np. O_{17}).

Na rys. 1 oraz 2 zilustrowano kształtowanie się obu rodzajów efektywności.



Rysunek 1. Efektywność obiektów

Źródło: opracowanie własne.



Wykres 2. Histogram efektywności

Źródło: opracowanie własne.

Z rys. 2 widać, że rozkład θ-efektywności może być skrajnie asymetryczny. Ponadto z definicji jest to zmienna dwustronnie ucięta, gdyż wskaźnik θ plasuje się w przedziale (0, 1]. Natomiast przebieg wskaźnika π-efektywności i związanego z nim wskaźnika rankingowego ρ, jest o wiele bardziej symetryczny. Ponadto wskaźnik rankingowy jest ucięty tylko prawostronnie w zerze, co znaczy, że po jego zlogarytmowaniu można otrzymać rozkład $\ln\rho$, być może, podobny do rozkładu normalnego.

3.4. PROBLEM 3 – CZY EFEKTYWNOŚĆ WZGLĘDNA DOBRZE CHARAKTERYZUJE ZDOLNOŚĆ KREDYTOWĄ?

Tradycyjne metody *Data Envelopment Analysis* pozwalają określić efektywność obiektu względem danego zbioru obiektów, czyli efektywność *względną*. Nie pozwalają natomiast określić efektywności „bezwzględnej", czyli efektywności w zbiorze wszystkich możliwych obiektów. A ponieważ w danym zbiorze zawsze przynajmniej jeden obiekt ma efektywność względną równą 1, dlatego możliwe jest, że:

1. jeśli nawet wszystkie obiekty mają niskie wartości zmiennych charakteryzujących zdolność kredytową (czyli niskie rezultaty), to i tak niektóre z nich będą miały efektywność równą lub prawie równą 1[8];

[8] Np. wśród ocen: 2, 3(=), 3(–), 3 oceną najlepszą, w pełni efektywną, jest trzy.

2. jeśli natomiast we wszystkich obiektach poziom zmiennych charakteryzujących zdolność kredytową jest bardzo korzystny (rezultaty są bardzo wysokie), to i tak dla niektórych z tych obiektów wskaźniki efektywności będą małe[9].

W pierwszym wypadku, zgodnie z ogólną ideą ustalania efektywności za pomocą metod DEA należałoby uznać, że wiele obiektów odznacza się wysoką zdolnością kredytową, wbrew oczywistym faktom wynikającym z małej wartości zmiennych charakteryzujących zdolność kredytową. W drugim wypadku natomiast, wbrew oczywistym faktom oznaczającym wysoką zdolność kredytową, bo zmienne ją charakteryzujące osiągają bardzo korzystne wartości, trzeba by było przyjąć, że niektóre obiekty mają bardzo niezadowalającą zdolność kredytową.

Przykład

Można sprawdzić, że w trzech następujących sytuacjach:

sytuacja 1: nakłady są takie jak w tabeli 1 oraz rezultaty, takie jak w tabeli 1,

sytuacja 2: nakłady są takie jak w tabeli 1, ale rezultaty są 100 razy *mniejsze*,

sytuacja 3: nakłady są takie jak w tabeli 1, ale rezultaty są 100 razy *większe*,

wskaźniki efektywności będą takie same, mimo radykalnych zmian rezultatów przy tym samym poziomie nakładów[10]!

Oczywiście, analogiczna własność zachodzi też dla warunków dotyczących nakładów.

Pokazane przed chwilą, bardzo niedogodne, własności metody DEA można usunąć przez wprowadzenie do zadania DEA pewnego wirtualnego obiektu wzorcowego lub nawet kilku wirtualnych obiektów wzorcowych. Wtedy, oczywiście, efektywność nadal będzie względna, ale podstawą porównań będzie nie jakiś dowolny, wybrany przez procedurę, obiekt empiryczny, lecz z góry określony wirtualny wzorzec lub wzorce. Jeśli wzorzec jest „doskonały", to relatywna efektywność względem niego będzie zbliżona do efektywności „bezwzględnej".

Propozycja rozwiązania problemu 3: *dla wyeliminowania redundancji obiektów efektywnych należy wprowadzić „wirtualny" lub rzeczywisty obiekt wzorcowy i zastosować model DEA z dozwolonymi wzorcami.*

Konstrukcja wzorca

Jak obiekt wzorcowy skonstruować, jest sprawą pomysłowości badacza. Może to być jakiś, istniejący gdzie indziej, obiekt empiryczny o znacznie lepszych własnościach od obiektów przez nas badanych, np. znakomita firma uzyskująca najwyższe ratingi; mogą to być obiekty uznawane za najlepsze przez ekspertów. Można też próbować samodzielnie „statystycznie" skonstruować obiekt wzorcowy. Poniżej omawiamy trzy takie propozycje.

---

[9] Np. wśród ocen 5, 6 najgorszą oceną jest 5.

[10] Matematyczny dowód tej własności (ma ona zresztą miejsce dla wielu modeli DEA) jest bardzo prosty: Warunek dla rezultatów nie ulega zmianie, jeśli je wszystkie pomnożymy przez pewną dodatnią stałą a. Nierówność $\sum_j \lambda_{oj} y_{rj} \geq y_{ro}$ jest bowiem identyczny z nierównością $\sum_j \lambda_{oj} (a y_{rj}) \geq a y_{ro}$. Co więcej, będzie tak, gdy poszczególne rezultaty pomnożymy przez różne dodatnie stałe a_r ($r = 1, ..., R$).

Propozycja 1: Wzorzec ekstremalny.

Za wielkość *r*-tego rezultatu w wirtualnym obiekcie wzorcowym przyjmujemy maksymalną zarejestrowaną wielkość rezultatu empirycznego, a za wielkość *n*-tego nakładu – minimalną zarejestrowaną wielkość nakładu empirycznego:

$$\begin{aligned} y_{rw} &= \max(y_{r1}, \ldots, y_{rJ}), r = 1, \ldots, R, \\ x_{nw} &= \min(x_{n1}, \ldots, x_{nJ}), n = 1, \ldots, N. \end{aligned} \tag{15}$$

Indeks *w* oznacza „wzorcowy".

Propozycja 2. Wzorzec przeciętny.

Wzorzec ekstremalny na ogół jest sztuczny i praktycznie nieosiągalny, co w dużym stopniu może wynikać np. z różnej wielkości obiektów. Wiadomo, że duże rezultaty wystąpią raczej w dużych niż małych obiektach, a małe nakłady – w obiektach raczej małych niż dużych. Wzorzec (15) łączy więc dwie przeciwne strony: pod względem rezultatów odwołuje się do obiektów dużych, a pod względem rezultatów – do obiektów małych. Dlatego można sugerować, aby wzorzec wirtualny był podobny do obiektu „przeciętnego", np. żeby był określony na poziomie mediany czy średniej poszczególnych nakładów oraz rezultatów. W wypadku wzorca medianowego:

$$\begin{aligned} y_{rw} &= \textit{mediana}\ (y_{r1}, \ldots, y_{rJ}),\ r = 1, \ldots, R, \\ x_{nw} &= \textit{mediana}\ (x_{n1}, \ldots, x_{nJ}),\ n = 1, \ldots, N. \end{aligned} \tag{16}$$

Propozycja 3. Wzorzec kwartylowy. Wzorzec ekstremalny jest dość sztuczny, zaś wzorzec przeciętny może uplasować się w otoczeniu autentycznych obiektów. Pośrednim wyjściem jest przyjęcie wzorca kwartylowego:

$$\begin{aligned} y_{rw} &= \textit{kwartyl 3.rzędu w szeregu}\ (y_{r1}, \ldots, y_{rJ}),\ r = 1, \ldots, R, \\ x_{nw} &= \textit{kwartyl 1.rzędu}\ \text{w szeregu}\ (x_{n1}, \ldots, x_{nJ}),\ n = 1, \ldots, N. \end{aligned} \tag{17}$$

W tabeli 3 podano omawiane trzy wzorce statystyczne dotyczące danych z tabeli 1.

Tabela 3

Wzorce statystyczne

Nakład/Rezultat	Wzorzec ekstremalny	Wzorzec medianowy	Wzorzec kwartylowy
X_1	783	3320	1446
X_2	56	323	83
X_3	643	9713	1932
X_4	132	4636	1113
X_5	123	3960	1210
X_6	989	14002	3230

cd. tabeli 3

Nakład/Rezultat	Wzorzec ekstremalny	Wzorzec medianowy	Wzorzec kwartylowy
Y_1	13517	2181	6401
Y_2	3785	459	981
Y_3	2279	228	1022

Źródło: obliczenia własne.

Wersje modelu DEA

Generalnie możliwe są dwie wersje modeli DEA z góry ustalonymi wzorcami.

1. DEA rozszerzona o wzorce

Procedura jest prostym uogólnieniem tradycyjnej DEA: zbiór obiektów rozszerzany jest o obiekty wzorcowe, a obliczenia przeprowadza się tak, jak w tradycyjnej DEA, dopuszczając, że do rozwiązania może wejść każdy obiekt empiryczny oraz każdy obiekt wzorcowy.

Ta wersja badania może być przydatna dla sprawdzenia, czy obiekt słusznie został uznany za wzorzec. Jest oczywiste, że obiekt może być uznany za wzorzec, gdy jego efektywność na tle obiektów empirycznych jest nie mniejsza od pewnej dość dużej liczby większej od 1 (np. jego nadefektywność jest większa od 2).

2. DEA z dozwolonymi wzorcami

Obliczenia prowadzone są przy założeniu, że do rozwiązania mogą wejść tylko ustalone wcześniej obiekty wzorcowe. Mianowicie konstruuje się zadanie DEA z obiektami empirycznymi oraz obiektami wzorcowymi, z tym że dodatkowo postuluje się, by w modelu dla obiektu *o*-tego zerowe były wagi intensywności λ_{oj} dla obiektów niewzorcowych, $1 \leq j \leq J$.

Tę właśnie wersję modelu DEA zaproponujemy do ustalania zdolności kredytowej – zob. część 5 artykułu. Tam też znajduje się opis modelu oraz przykład. Na razie nie czynimy tego, gdyż nie chcemy przerywać dyskusji tradycyjnego podejścia DEA do mierzenia zdolności kredytowej.

Przykład

Korzystając z pierwszej wersji modelu SE-CCR *rozszerzonego o wzorce* sprawdzimy, czy podane w tabeli 3 wzorce: ekstremalny, medianowy oraz kwartylowy mogą być traktowane jako wzorcowe. Wyniki obliczeń zestawiono w tabeli 4.

Tabela 4

Wskaźniki efektywności i ranking obiektów

Obiekt	Wskaźniki rankingowe przy wzorcu:			
	ekstremalnym	medianowym	kwartylowym	bez wzorca (klasyczny)
O1	0,028	0,886	0,377	0,886
O2	0,050	**1,426**	0,607	**1,426**
O3	0,051	**1,139**	0,292	**1,139**
O4	0,107	**1,944**	0,496	**1,944**
O5	0,089	**1,212**	0,365	**1,212**
O6	0,097	**1,651**	0,378	**1,651**
O7	**2,007**	**2,042**	0,832	**1,976**
O8	1,012	**1,188**	0,660	**1,188**
O9	0,553	**1,015**	0,990	**1,038**
O10	**1,512**	**1,512**	0,623	**1,430**
O11	**1,448**	**1,462**	0,958	**1,502**
O12	**2,972**	**2,972**	**1,301**	**1,566**
O13	0,093	**2,376**	0,641	**2,376**
O14	0,030	0,740	0,311	0,740
O15	0,049	**1,700**	0,418	**1,700**
O16	0,050	**3,124**	0,989	**3,124**
O17	0,019	0,686	0,443	0,686
O18	0,029	**2,103**	**1,103**	**2,103**
O19	0,035	**1,575**	**1,125**	**1,575**
Efektywność wzorca	77,103	0,782	6,262	–

Źródło: obliczenia własne.

Wyniki sugerują, że wzorzec medianowy nie powinien być rozpatrywany, gdyż jest on nieefektywny.

Należy zwrócić uwagę, że przy różnych wzorcach sytuacja danego obiektu może być różna. Na przykład obiekt O_{16} był najlepszy przy wzorcu medianowym, był zupełnie bez znaczenia przy wzorcu ekstremalnym i osiągnął niezły wynik przy wzorcu kwartylowym. Jest to zrozumiałe, gdyż każdy wzorzec ma na ogół inną strukturę nakładów w stosunku do rezultatów, a więc za każdym razem dany obiekt oceniany jest w stosunku do obiektów o różnych na ogół strukturach.

4. FUNKCJA DYSKRYMINACYJNA

4.1. PROCEDURA KONSTRUKCJI FUNKCJI DYSKRYMINACYJNEJ

W klasycznej analizie ryzyka kredytowego dla stwierdzenia, czy dany obiekt jest „dobry" czy też „zły" konstruuje się funkcję dyskryminacyjną, np. słynną funkcję Z Altmana[11], a następnie, na podstawie wartości zmiennych niezależnych tej funkcji w analizowanym obiekcie, oblicza się jej wartość i ustala przydział obiektu do odpowiedniej grupy, np. grupy „nie ryzykownej: lub „ryzykownej"[12].

Podobne postępowanie proponuje się w zastosowaniach metod DEA do analizy ryzyka kredytowego. Za funkcję dyskryminacyjną przyjmuje się funkcję regresji, w którym zmienną zależną jest syntetyczny miernik zdolności kredytowej E, a zmiennymi niezależnymi są wykorzystane w modelu DEA zmienne charakteryzujące nakłady oraz rezultaty. Tak więc funkcja dyskryminacyjna ma formę:

$$\tilde{E} = b + \sum_n b_n X_n + \sum_r a_r Y_r \tag{18}$$

gdzie parametry wyznaczono jakąś metodą estymacji statystycznej, np. klasyczną mnk. W funkcji (18) wystąpić mogą wszystkie lub niektóre nakłady X_n oraz rezultaty Y_r. W przypadku modelu CCR za zmienną zależną E bierze się wskaźnik θ-efektywności, a w przypadku modelu SE-CCR – wskaźnik rankingowy ρ lub wskaźnik π-efektywności.

Niech γ będzie punktem odniesienia (= punktem „odcięcia") syntetycznego miernika zdolności kredytowej. W zależności od kontekstu syntetycznym miernikiem jest wskaźnik empiryczny E lub jego aproksymacja $\tilde{E}$ określona wzorem (18).

Obiekty, dla których syntetyczny miernik zdolności kredytowej $\tilde{E}$ lub E jest większy od γ kwalifikowane są jako „zdrowe", „dobre", „mało ryzykowne", „płynne" itp., a dla których syntetyczny miernik jest niewiększy od γ – kwalifikowane są do grupy przeciwnej „niezdrowych ekonomicznie", „złych", „wysoce ryzykowanych", „niepłynnych" itp.

Jest naturalne, że niekiedy trzeba wydzielić kilka klas zdolności kredytowej. Wówczas podstawa odniesienia ma charakter wielopunktowy; są to kolejne progi $\gamma_1 < \gamma_2 < \ldots < \gamma_L$. Wówczas, na przykład, jeśli $\tilde{E} < \gamma_1$ – to obiekt skrajnie niedobry, $\gamma_1 < \tilde{E} \leq \gamma_2$ – obiekt w dużym stopniu jest niedobry, …, $\tilde{E} > \gamma_L$ – obiekt jest „celujący". Klasyfikacja może też mieć charakter nierozłączny; oprócz obiektów „dobrych" oraz „złych" wyróżnia się obiekty „ani dobre, ani złe".

W literaturze dotyczącej zastosowania DEA do analizy wniosków kredytowych, dla ustalania punktów odcięcia γ proponuje się wykorzystać analizy ekspertów[13] lub niektóre wielkości statystycznych, np. za punkt odcięcia proponuje się przyjąć medianę[14]. Jeśli chodzi o wielopunktową podstawę odniesienia, to też można sugerować zastosowanie analizy eksperckiej lub statystycznej. W tym ostatnim przypadku – wykorzystujemy odpowiednie kwantyle w rozkładzie syntetycznego miernika zdolności kredytowej.

---

[11] Altman [1].

[12] Bardzo ładny przegląd różnych opracowanych w Polsce metod dyskryminacji w odniesieniu do klasyfikacji obiektów na „zdrowe" i „niezdrowe") zawiera artykuł Hamrol, Chodakowski [11].

[13] Np. [8], s. 115.

[14] Np. [12], s. 1767.

4.2. PROBLEM 4 – CZY NALEŻY WERYFIKOWAĆ FUNKCJĘ DYSKRYMINACYJNĄ?

W tradycyjnym podejściu funkcja dyskryminacyjna (18) określana jest na drodze ekonometrycznej i nie dziwią przeto zalecenia, aby w odniesieniu do niej stosować wszystkie podstawowe postulaty estymacji statystycznej, np. dobre dopasowanie oraz istotność zmiennej objaśniających.

Sądzimy, że to nie wystarcza. Jak bowiem wiadomo, modele regresyjne, choć powszechnie używane, mają „wady", które niekiedy poddają w wątpliwość sens ich stosowania. Np. z uwagi na wewnętrzne korelacje zmiennych niezależnych oraz przechodniość relacji skorelowania, znak współczynnika regresji może być zupełnie inny niż oczekiwany, a model wielu zmiennych na ogół nie jest „sumą" modeli dla pojedynczych zmiennych niezależnych[15]. Ponieważ na podstawie funkcji dyskryminacyjnej dokonywane są niekiedy bardzo ważne dla banku i dla wnioskodawcy analizy, powinna ona być wszechstronnie zweryfikowana, i to, przede wszystkim, pod względem merytorycznym. Ocena statystyczna jest tu tylko pomocnicza.

Propozycja rozwiązania problemu 4: *obok postulatów statystycznych istotności oraz dobrego dopasowania, należy dokonać merytorycznej oceny funkcji dyskryminacyjnej, badając przynajmniej poprawność znaków jej współczynników oraz ich skalę.*

W zastosowaniu DEA do *credit scoring* zmienną zależną (dyskryminatą) jest syntetyczny miernik zdolności kredytowej, który ma charakter maksymanty.

Dlatego też, jeśli coraz wyższy poziom zmiennej niezależnej (nakładu lub rezultatu) jest – *ceteris paribus* – oceniany coraz korzystniej z punktu widzenia zdolności kredytowej, to poprawny jest dodatni znak współczynnika regresji przy tej zmiennej. Jeśli zaś coraz wyższy poziom zmiennej niezależnej jest w warunkach *ceteris paribus* oceniany coraz bardziej niekorzystnie, to poprawny jest ujemny znak odpowiedniego współczynnika regresji[16].

Przykład

W jednej z prac podano funkcję dyskryminacyjną, w której dyskryminata zależy ujemnie od grającego rolę rezultatu wskaźnika płynności. Jest to sprzeczne z wiedzą ekonomiczną. Z kolei w innym artykule proponuje się funkcję dyskryminacyjną, w której miernik zdolności kredytowej zależy dodatnio od wskaźnika ogólnego zadłużenia i zależy ujemnie od ROA.

4.3. PROBLEM 5 – CZY SZACOWANIE FUNKCJI DYSKRYMINACYJNEJ JEST KONIECZNE DLA ZASTOSOWANIA METOD DEA DLA OCENY ZDOLNOŚCI KREDYTOWEJ?

Spojrzenie praktyczne

Z czysto praktycznego punktu widzenia funkcja dyskryminacyjna jest przede wszystkim potrzebna dla klasyfikowania nowych obiektów na „dobre" i „ryzykowne"

[15] Byłoby tak, gdyby w materiale statystycznym wszystkie zmienne niezależne były względem siebie ortogonalne.

[16] W rzeczywistości sprawa jest bardziej skomplikowana. W DEA za nakłady trzeba uznawać tylko takie wielkości, których wzrost jest niezbędny dla wzrostu rezultatów. Prowadzi to do tzw. postulatu koincydencji nakładów i rezultatów w modelach DEA.

pod względem zdolności kredytowej. Wówczas, podstawiając do odpowiedniego wzoru wartości zmiennych niezależnych charakteryzujących nowy obiekt, otrzymujemy wartość dyskryminaty dla nowego obiektu. Przyjmując ponadto pewien punkt odcięcia, można ustalić, czy obiekt ten zaliczyć do „dobrych" czy też nie.

Natomiast funkcja dyskryminacyjna dla obiektów już badanych w modelu DEA w zasadzie nie jest potrzebna, bo dysponujemy dla nich „prawdziwymi" (= empirycznymi) wartościami dyskryminaty. Aproksymacja ma jedynie sens, jeśli analityk chce bardziej elegancko zaprezentować wyniki obliczeń.

Spojrzenie formalne

Klasyczne podejście do estymacji funkcji dyskryminacyjnej jako funkcji regresji wydaje się intuicyjnie zrozumiałe: skoro w modelu DEA efektywność obiektu jest rozwiązaniem optymalnym odpowiedniego zadania decyzyjnego, czyli skoro jest ona przekształceniem empirycznych nakładów i rezultatów, to naturalne wydaje się, by funkcję dyskryminacyjną skonstruować jako model regresji, w którym zmiennymi niezależnymi są właśnie nakłady i rezultaty wykorzystywane przy określaniu efektywności. Idea ta jest jednak kontrowersyjna.

1. Po pierwsze, rodzi się pytanie, czy opis świata za pomocą modeli optymalizacyjnych jest tożsamy z opisem za pomocą modeli regresyjnych? Oczywiście, stosowanie jednego z nich nie wyklucza drugiego. Chodzi jednak o to, czy poprawne jest dyskryminowanie wartości uzyskanych z jednego podejścia za pomocą wartości uzyskanych w innym podejściu, jeśli w obu podejściach używa się *tych samych* zmiennych? Podobnie jest w wypadku liniowych równań przepływów bilansowych i liniowych modeli produkcji końcowej. Mimo zewnętrznego podobieństwa obu podejść układy równań liniowych wielu zmiennych nie są to przecież podejścia równoważne.

2. Po drugie, należy wyrazić wątpliwość, czy w ogóle konstrukcja „regresyjnej" funkcji dyskryminacyjnej jest poprawna. Jeśli bowiem przyjmuje się, że efektywność – jako rozwiązanie optymalne zadania decyzyjnego – zależy od nakładów i rezultatów, czyli przyjmuje się, że $\mathbf{E} = f(\mathbf{X}, \mathbf{Y})$, to dziwne na tym tle jest szacowanie modelu regresyjnego typu (18) w formie $\mathbf{E} = g(\mathbf{X}, \mathbf{Y})$. Bo to oznacza, jakby nie wierzono w wyniki modelu decyzyjnego. W przytoczonych wzorach symbol $\mathbf{X}$ to empiryczna macierz nakładów, $\mathbf{Y}$ – empiryczna macierz rezultatów, $\mathbf{E}$ – wektor wskaźników efektywności obiektów.

3. Po trzecie, jeśli funkcja dyskryminacyjna, g, jest liniowa – a tak jest najczęściej – to aproksymacja rozwiązania zadania decyzyjnego oznacza rozpatrywanie dość „dziwnej" zależności, że rozwiązanie optymalne $f(\mathbf{X}, \mathbf{Y})$, już będące funkcją $\mathbf{X}$, $\mathbf{Y}$, jeszcze dodatkowo zależy od tychże $\mathbf{X}$ oraz $\mathbf{Y}$.

Naturalnie można byłoby ratować to podejście i konstruować funkcję dyskryminacyjną względem nakładów i rezultatów nie ujętych w modelu DEA, ale to też będzie wątpliwe. Skoro bowiem w modelu DEA syntetyczną charakterystykę zdolności kredytowej uzależniono od $\mathbf{X}$ oraz $\mathbf{Y}$, to dlaczego teraz uzależniać ją od innych zmiennych, powiedzmy $\mathbf{Z}$ oraz $\mathbf{V}$?

Propozycja rozwiązania problemu 5: *proponujemy, aby stosując metody DEA, wręcz zrezygnować z konstrukcji funkcji dyskryminacyjnej, a klasyfikację obiektów przeprowadzać na podstawie modelu DEA z dozwolonymi wzorcami.*

Odpowiednią procedurą omówiono poniżej.

5. PROPOZYCJA PROCEDURY DYSKRYMINACJI WNIOSKÓW KREDYTOWYCH

Proponujemy, by zgodnie z duchem *Data Envelopment Analysis* dla określenia zdolności kredytowej nowego obiektu (lub ich grupy), a jeśli trzeba, to i „starych" obiektów rozwiązać odpowiednie zadania z *dozwolonymi* wzorcami, a więc zadania, w których wagi intensywności mogą być dodatnie tylko dla obiektów wzorcowych (przykładowy model z dozwolonymi wzorcami sformułowano niżej). Badamy wówczas zdolność kredytową obiektów albo „starych", albo „nowych", albo „i starych i nowych" na tle z góry ustalonych obiektów wzorcowych.

Następnie należy przeprowadzić zwykłe zadanie klasyfikacji obiektów, kierując się ustalonymi punktami odcięcia w odniesieniu do uzyskanych wskaźników efektywności (lub nadefektywności).

Nie ma natomiast potrzeby konstrukcji jakiejkolwiek funkcji dyskryminacyjnej (regresyjnej lub innej).

Przy obecnych środkach obliczeniowych rozwiązanie modelu DEA nawet dla dużych zbiorów danych nie stwarza żadnego kłopotu[17].

5.1. UKIERUNKOWANY NA NAKŁADY STANDARDOWY MODEL SE-CCR Z DOZWOLONYMI WZORCAMI

Niech B oznacza zbiór badanych obiektów (mogą to być tylko „stare" obiekty lub tylko „nowe", lub obie te grupy łącznie). Natomiast W niech oznacz zbiór obiektów-wzorców. Dysponujemy informacjami o wielkości nakładów oraz rezultatów dla wszystkich obiektów z obu zbiorów.

Zadanie dla obiektu $o \in B$ polega na znalezieniu takich wag intensywności λ_{oj}, gdzie $j \in W$ oraz takiego mnożnika nakładów ρ_o, że:

$$\min \rho_o \tag{19}$$

przy warunkach:

$$\sum_{j \in W} \lambda_{oj} y_{rj} \geq y_{ro} \qquad (r = 1, \ldots, R), \tag{20}$$

$$\sum_{j \in W} \lambda_{oj} x_{nj} \leq \rho_o x_{no} \qquad (n = 1, \ldots, N), \tag{21}$$

$$\rho_o, \lambda_{oj} \geq 0 \quad j \neq o, \quad j \in W, \quad o \in B. \tag{22}$$

Zadanie to można rozwiązać jako „zwykle" zadanie SE-CCR, narzucają jednak warunki $\lambda_{oj} = 0$ dla $j \notin W$.

[17] Jeśli oceniamy zdolność kredytową tylko nowych obiektów, to można rozpatrywać tylko zbiór tych obiektów. Oczywiście, z uwagi na ustalenie wzorców, nie będzie błędem rozpatrzenie zbioru wszystkich, starych i nowych obiektów (choć wiele rachunków wykonamy niepotrzebne).

5.2. PRZYKŁAD EMPIRYCZNY

Należy zaklasyfikować do odpowiednich grup cztery obiekty scharakteryzowane nakładami i rezultatami podanymi w tabeli 5 (obiekty O_{20}, O_{21}, O_{22} i O_{23}). W tabeli podano też wzorce. Są to dwa wzorce „eksperckie", E_1 i E_2, z tym że wzorzec ekspercki E_2 jest identyczny z obiektem O_{19} oraz jeden wzorzec statystyczny – wzorzec kwartylowy.

Tabela 5

Nakłady i rezultaty czterech nowych obiektów oraz wzorce

Nakład/ Rezultat	Nowe obiekty				Wzorce		
	Obiekt O20	Obiekt O21	Obiekt O23	Obiekt O24	E1	O19	kwartylowy
X_1	1003	3000	1700	3700	5000	856	1446
X_2	1500	323	80	1800	1000	68	83
X_3	640	7131	1900	1400	8000	1029	1932
X_4	500	636	2113	1130	7000	132	1113
X_5	921	960	3210	674	3000	289	1210
X_6	1000	1402	2230	632	15000	989	3230
Y_1	400	2810	5600	6005	4000	40	6401
Y_2	700	1459	1980	1342	4000	132	981
Y_3	300	765	1002	400	3000	11	1022

Źródło: obliczenia własne.

W celu obliczenia syntetycznego wskaźnika zdolności kredytowej wykorzystamy ukierunkowany na nakłady standardowy model SE-CCR. Wskaźnikiem zdolności kredytowej będzie wskaźnik rankingowy ρ. Obiekty poklasyfikujemy na pięć grup. Przyjmiemy, że zdolność kredytowa jest:

(i) niewystarczająca mała, gdy $\rho < 0{,}5$
(ii) niewystarczająca umiarkowana, gdy $0{,}5 \leq \rho < 0{,}8$
(iii) wystarczająca umiarkowana, gdy $0{,}8 \leq \rho < 1{,}5$
(iv) wystarczająca dobra, gdy $1{,}5 \leq \rho < 2{,}5$
(v) wystarczająca bardzo dobra, gdy $\rho > 2{,}5$.

Wskaźniki rankingowe obiektów oraz przydział do grup przedstawia tabela 6.

Tabela 6

Wskaźniki rankingowe oraz klasyfikacja obiektów

Obiekt	ρ	Grupa	Obiekt	ρ	Grupa	Obiekt	ρ	Grupa
O_1	0,270	i	O_8	0,191	i	O_{16}	0,987	Iii
O_2	0,536	ii	O_9	0,250	i	O_{17}	0,341	i
O_3	0,276	i	O_{10}	0,384	i	O_{18}	0,864	iii
O_4	0,491	i	O_{11}	0,353	i	O_{20}	2,305	iv
O_5	0,364	i	O_{12}	0,445	i	O_{21}	3,426	v
O_6	0,378	i	O_{13}	0,641	ii	O_{23}	2,923	v
O_7	0,418	i	O_{14}	0,249	i	O_{24}	6,991	v
			O_{15}	0,422	i	O_{19}	1,135	iii

Źródło: obliczenia własne.

W sensie przyjętych kryteriów większość obiektów nie ma wystarczającej zdolności kredytowej. Bardzo dobrą zdolnością kredytową charakteryzują się nowe obiekty O20-024.

6. PODSUMOWANIE

1. W artykule wskazano, że tradycyjna procedura określania zdolności kredytowej poprzez:

– rozwiązanie zadania CCR lub BCC,

– oszacowanie funkcji dyskryminacyjnej jako funkcji regresji, w której zmienną zależną jest wskaźnik efektywności a zmiennymi niezależnym są użyte w zadaniu DEA zmienne charakteryzujące nakłady oraz rezultaty ma wiele wad, w szczególności dotyczących funkcji dyskryminacyjnej.

2. W związku z tym zaproponowano procedurę określania zdolności kredytowej w ogóle bez funkcji dyskryminacyjnej. Procedura cały czas opiera się na pojęciach i metodach DEA. Jej etapy są następujące:

(i) Ustala się wstępne obiekty wzorcowe, np. na podstawie analizy eksperckiej lub statystycznej.

(ii) Rozwiązuje się zadanie SE-DEA z obiektami empirycznymi oraz wzorcowymi.

(iii) W oparciu o te wyniki weryfikuje się wstępne wzorce. Np. wzorzec może być zaakceptowany, gdy na tle obiektów empirycznych jest on w pełni efektywny w sensie Farrella. Obok tego mogą być, oczywiście, uwzględnianie inne kryteria.

(iv) Ustala się syntetyczny miernik zdolności kredytowej poszczególnych obiektów („starych", „nowych") poprzez rozwiązanie odpowiedniego zadania SE-DEA *z dozwolonymi wzorcami*, np. (19)-(22).

(v) Przydział obiektów do grup odbywa się poprzez porównywanie obliczonych w etapie (iv) syntetycznych mierników zdolności kredytowej z punktem (punktami) odcięcia.

3. Obok badania zdolności kredytowej proponowana procedura może być stosowana do pokrewnych zagadnień – do oceny ryzyka, badania zagrożenia bankructwem, kwalifikacji do grup jakościowych itp.

Akademia Ekonomiczna w Poznaniu

LITERATURA

[1] Altman E.I., [1968], *Financial ratios, discriminant analysis and the prediction of corporate bankruptcy*, „The Journal of Finance", 23, 4.
[2] Andersen P., Petersen N.C., [1993], *A procedure for ranking efficient units in Data Envelopment Analysis*, „Management Science", 39, 10.
[3] Banker R.D., Charnes A., Cooper W.W., [1984], *Some models for estimating technical and scale inefficiencies in data envelopment analysis*, „Management Science", 30.
[4] Banker R.D., Gilford J.L., [1988], *A relative efficiency model for the evaluation of public health nurse productivity*, Mellon University Mimeo, Cornegie.
[5] Charnes A., Cooper W.W., Rhodes E., [1978], *Measuring the efficiency of decision making units*, „European Journal of Operational Research", 2.
[6] Cook W.D., Seiford L.M., [2009], *Data envelopment analysis (DEA) – Thirty years on*, „European Journal of Operational Research", 192, 1.
[7] Chang E.W.I., Chiang Y.H., Tang B.S., [2007], *Alternative approach to credit scoring by DEA: Evaluating borrowers with respect to PFI projects*, „Building and Environment", 42, 4.
[8] Emel A.B., Oral M., Reisman A., Yolalan R., [2003], *A credit scoring approach for the commercial banking sector*, „Socio-Economic Planning Sciences", 37, 2.
[9] Feruś A., [2006], *Zastosowanie metody DEA do określania poziomu ryzyka kredytowego przedsiębiorstw*, „Bank i Kredyt", 7.
[10] Gospodarowicz A., [2004], *Możliwości wykorzystania metody DEA do oceny ryzyka kredytowego w kontekście Nowej Umowy Kapitałowej, w: Przestrzenno-czasowe modelowanie i prognozowanie zjawisk gospodarczych*, (red. A. Zeliaś), Wyd. Akademii Ekonomicznej w Krakowie, Kraków.
[11] Hamrol M., Chodakowski J., [2008], *Prognozowanie zagrożenia finansowego przedsiębiorstwa. Wartość predykcyjna polskich modeli analizy dyskryminacyjnej*, Badania Operacyjne i Decyzje, 3.
[12] Min J.H., Lee Q-C., [2008], *A practical approach to credit scoring*, „Expert System with Applications", 35, 4.
[13] Paradi J.C., Asmild M., Simak P.C., [2004], *Using DEA and worst practice DEA in credit risk evaluation*, „Journal of Productivity Analysis", 21.
[14] Truott M.D., Rai A., Zhang A., [1996], *The pontential use of DEA for credit applicant acceptance system*, „Computers and Operations Research", 23, 4.
[15] Yeh Q.J., [1996], *The application of data envelopment analysis in conjunction with financial ratios for bank performance evaluation*, „Journal of Operational Research Society", 47, 8.

Praca wpłynęła do redakcji w marcu 2009 r.

UWAGI NA TEMAT ZASTOSOWANIA METODY DEA DO USTALANIA ZDOLNOŚCI KREDYTOWEJ

Streszczenie

W artykule wskazano, że tradycyjna dla DEA procedura określania zdolności kredytowej poprzez rozwiązanie zadania CCR lub BCC oraz oszacowanie funkcji dyskryminacyjnej jako funkcji regresji, w której zmienną zależną jest wskaźnik efektywności, a zmiennymi niezależnym są użyte w zadaniu DEA zmienne charakteryzujące nakłady oraz rezultaty ma wiele wad. Głównie dotyczą one funkcji dyskryminacyjnej. W związku z tym zaproponowano procedurę określania zdolności kredytowej w ogóle bez funkcji dyskryminacyjnej.

Procedura cały czas opiera się na pojęciach i metodach DEA. W szczególności proponuje się wykorzystanie modelu nadefektywności DEA, a zamiast funkcji dyskryminacyjne proponuje się użyć modeli SE-DEA *z dozwolonymi wzorcami*. Typowanie obiektów do grup odbywa się tradycyjnie, poprzez porównanie miernika zdolności kredytowej (tu: wskaźnika rankingowego) z punktami odcięcia.

Słowa kluczowe: DEA, SE-CCR, zdolność kredytowa, DEA z dozwolonymi wzorcami

ESTIMATION OF CREDIT CAPACITY USING DATA ENVELOPMENT ANALYSIS

Summary

The article points out some disadvantages of traditional (based on DEA methodology) procedure of estimating credit capacity which consists in solving CCR and BCC models and estimating a discriminant function where efficiency indicator is a dependent variable and inputs and outputs used in DEA models are independent variables.

Since the main problems with this procedure are connected with discriminant function, the author suggests a procedure of credit capacity estimation which uses no discriminant function. The new method is based on DEA methodology, particularly on super-efficiency DEA models (SE-DEA models) with *permitted benchmarks*. Comparing the credit capacity indicator (here: ranking indicator) with cut-off points enables objects classification.

Key words: DEA, SE-CCR, credit capacity, DEA with permitted benchmarks

PRZEGLĄD STATYSTYCZNY
R. LVI – ZESZYT 2 – 2009

BARBARA DAŃSKA-BORSIAK

ZASTOSOWANIA PANELOWYCH MODELI DYNAMICZNYCH W BADANIACH MIKROEKONOMICZNYCH I MAKROEKONOMICZNYCH[1]

1. WPROWADZENIE

Terminem „dane panelowe" określa się we współczesnej ekonometrii dane, powstające z połączenia szeregów czasowych obserwacji dla jednostek przekrojowych. Charakteryzują się one tym, że liczba obserwacji w czasie T jest znacznie mniejsza niż liczba obiektów przekrojowych N. Modele ekonometryczne szacowane na podstawie danych panelowych, w których zakłada się, że na kształtowanie się zmiennej objaśnianej wpływają, oprócz zmiennych objaśniających, niemierzalne, stałe w czasie i specyficzne dla danego obiektu czynniki, zwane efektami grupowymi, nazywane są modelami panelowymi (ang. *panel data models*). Obecność w modelach panelowych efektów grupowych spowodowała konieczność opracowania specyficznych metod estymacji takich modeli. Stałość w czasie efektów grupowych jest przyczyną dodatkowych komplikacji metodologicznych, które pojawiają się przy estymacji modeli dynamicznych. Głównym celem artykułu jest przedstawienie przykładów zastosowania panelowych modeli dynamicznych w analizach ekonomicznych w skali mikro i makro. Szczególny nacisk położony został przy tym na wskazanie czynników różnicujących te dwa przypadki i konsekwencje zastosowania różnych metod estymacji w zależności od rodzaju próby.

2. DANE PANELOWE W MODELOWANIU MIKROEKONOMICZNYM (MIKROPANELE) I MAKROEKONOMICZNYM (MAKROPANELE)

W związku z dynamicznie rosnącą liczbą praktycznych zastosowań modeli panelowych w różnego rodzaju analizach ekonomicznych i wynikającym z tej różnorodności zróżnicowaniem próby pod względem wymiaru czasowego i przekrojowego, w najnowszych opracowaniach (por. np. [6]) pojawiło się rozróżnienie danych panelowych na makropanele i mikropanele. Mikropanele to dane, charakteryzujące się bardzo dużą liczbą obserwacji przekrojowych, sięgającą setek lub nawet tysięcy obiektów; liczba obserwacji w czasie jest natomiast bardzo niewielka – rzadko przekracza dziesięć, zazwyczaj jest to kilka obserwacji. Typowym przykładem mikropanelu są dane pocho-

[1] Praca naukowa finansowana ze środków na naukę w latach 2007-2009 jako projekt badawczy nr N111 0938 33

dzące z gospodarstw domowych lub inne dane indywidualne. Makropanele charakteryzują się natomiast bardziej ograniczoną liczbą obserwacji przekrojowych (N waha się od kilkunastu do kilkudziesięciu) i dłuższym horyzontem czasowym próby, sięgającym nawet $T = 30$ okresów. Są to zazwyczaj dane wykorzystywane w badaniach makroekonomicznych. Przykładem mogą być dane dla krajów Unii Europejskiej, krajów OECD, dla których dostępne są stosunkowo długie szeregi czasowe, dotyczące wielu zmiennych makroekonomicznych. Modele panelowe estymowane na podstawie makro i mikropaneli wymagają stosowania różnych metod ekonometrycznych. Potrzeba taka wynika z różnych wymiarów tych danych i związanych z nimi pożądanych własności asymptotycznych estymatorów. W przypadku mikropaneli szczególnie ważne są własności asymptotyczne dla $N \to \infty$ i skończonego T, a w przypadku makropaneli – dla $N \to \infty$ i $T \to \infty$ lub własności estymatorów w małych próbach. Ponadto w przypadku makropaneli o szczególnie długim wymiarze czasowym mogą pojawiać się problemy związane z niestacjonarnością, nie występujące w mikropanelach, w których wymiar T jest niewielki. Problemem, który jest typowy dla makropaneli jest natomiast zależność między obiektami (ang. *cross-unit dependence*), powodująca zjawisko korelacji przekrojowej. Nie występuje on zazwyczaj w modelach szacowanych na podstawie mikropaneli, gdyż obiekty są elementami próby losowej, a zatem mało prawdopodobna jest ich wzajemna korelacja.

Dostęp do baz danych zarówno makro, jak i mikropanelowych jest obecnie stosunkowo łatwy, zwłaszcza do baz międzynarodowych, oraz mikropanelowych baz amerykańskich i zachodnioeuropejskich. Przykłady takich baz, wraz z adresami witryn internetowych, są podane i opisane np. w pracy [6], s. 1-5. Pogląd na temat ogromnych rozmiarów baz danych mikropanelowych dać może przykład danych z gospodarstw domowych, zgromadzonych na Uniwersytecie Michigan (USA), znany jako *Panel Study of Income Dynamics* (PSID)[2]. Dane zaczęto zbierać w 1968 roku i pochodziły one z 4800 rodzin. Do 2003 roku w PSID zgromadzono dane o 65 000 jednostkach, a długość szeregów czasowych dochodzi aż do 36 lat.

3. ZARYS NAJCZĘŚCIEJ STOSOWANYCH METOD ESTYMACJI PANELOWYCH MODELI DYNAMICZNYCH

Dynamiczny model panelowy ma postać:

$$y_{it} = \gamma y_{i,t-1} + \mathbf{x}_{it}^T \boldsymbol{\beta} + u_{it} = \gamma y_{i,t-1} + \mathbf{x}_{it}^T \boldsymbol{\beta} + \alpha_i + \varepsilon_{it}, \quad i = 1, \ldots, N, \quad t = 1, \ldots, T. \quad (1)$$

gdzie $\varepsilon_{it} \sim N(0, \sigma_\varepsilon^2)$ dla wszystkich i, t, α_i – efekt grupowy, losowy lub nielosowy; jeśli α_i są losowe, to $\alpha_i \sim N(0, \sigma_\alpha^2)$, $\mathbf{x}_{it} = [x_{kit}]_{Kx1}$ jest wektorem zmiennych objaśniających o K współrzędnych, β jest wektorem parametrów ($Kx1$), jednakowych dla wszystkich i oraz t.

Jak już wspomniano, do estymacji panelowych modeli dynamicznych należy stosować odrębne metody, inne niż stosuje się dla modeli statycznych. Proponowana w literaturze metodologia bazuje w zasadzie na jednej z trzech metod: Metodzie Zmiennych

[2] Baza dostępna jest na stronie: http://psidonline.isr.umich.edu.

Instrumentalnych (por. [2] i [3]]), Metodzie Największej Wiarygodności (por. [17]), oraz Uogólnionej Metodzie Momentów (GMM) (por. [5], [1], [4], [8]). Najczęściej stosowane w praktyce są metody oparte na GMM, a szczególnie tzw. GMM pierwszych różnic (ang. *first- differenced GMM*), FDGMM, zaproponowana przez Arellano i Bonda w pracy [5], oraz systemowa GMM Blundella i Bonda (ang. *system GMM*), SGMM, por. [8]. Metody bazujące na GMM są omówione np. w książkach [6], [17]. Uwagi na temat GMM dla modeli szeregów czasowych i FDGMM w języku polskim znaleźć też można w artykule [15].

GMM jest metodą umożliwiającą estymację parametrów modelu bezpośrednio z warunków momentów, które mogą być liniowe lub nieliniowe względem parametrów. Postać i liczba warunków momentów wykorzystywanych w procesie estymacji zależą od założeń dotyczących korelacji między zmiennymi $\mathbf{x}_{it}$ a składowymi α_i i ε_{it}. Przy założeniu, że ε_{it} nie wykazuje korelacji w czasie, oraz $\mathbf{x}_{it}$ są skorelowane z α_i, możliwe jest przyjęcie trzech alternatywnych założeń na temat korelacji $\mathbf{x}_{it}$ z ε_{it}. Zmienne $\mathbf{x}_{it}$ mogą być:

– endogeniczne, to znaczy $\mathbf{x}_{it}$ jest skorelowany z wartością bieżącą ε_{it} i wartościami opóźnionymi $\varepsilon_{i,t-s}$, ale nieskorelowana z wartościami przyszłymi $\varepsilon_{i,t+s}$,

– z góry ustalone (słabo egzogeniczne), to znaczy $\mathbf{x}_{it}$ jest nieskorelowany z wartością bieżącą ε_{it}, ale skorelowany z wartościami opóźnionymi $\varepsilon_{i,t-s}$,

– ściśle egzogeniczne, to znaczy $\mathbf{x}_{it}$ jest nieskorelowany z wartością bieżącą ε_{it}, wartościami opóźnionymi $\varepsilon_{i,t-s}$, i z wartościami przyszłymi $\varepsilon_{i,t+s}$.

Możliwe jest też przyjęcie dodatkowych silniejszych założeń, dotyczących korelacji między efektami grupowymi α_i a zmiennymi $\mathbf{x}_{it}$, braku korelacji jednoczesnej: $E(\mathbf{x}_{it}\,\mathbf{u}_{it}) = \mathbf{0}$ dla $t = 1, \ldots, T$, warunków początkowych lub słabej stacjonarności. Wybór między powyższymi alternatywnymi założeniami wydaje się arbitralny. Ponieważ jednak w większości przypadków dodatkowe warunki momentów są tzw. warunkami ponadidentyfikującymi, to ich właściwość można testować, na przykład za pomocą testu Sargana[3]. Poniżej przedstawiona zostanie w zarysie zasadnicza idea dwóch metod: GMM pierwszych różnic Arellano i Bonda (ang. *first differenced GMM*), FDGMM i systemowej GMM Blundella i Bonda (ang. *system GMM*), SGMM. Jak już wspomniano, są one jednymi z najczęściej stosowanych w praktyce metodami estymacji panelowych modeli dynamicznych, a ponadto prezentowane w kolejnych częściach artykułu przykłady empiryczne bazują właśnie na tych metodach.

Zasadniczą ideę FDGMM przedstawić można następująco: oblicza się pierwsze różnice modelu, w celu usunięcia stałych w czasie efektów grupowych α_i, a następnie zmienne objaśniające w modelu pierwszych różnic zastępuje instrumentami, którymi są poziomy zmiennych, opóźnione o dwa lub więcej okresów. Estymatory parametrów strukturalnych uzyskuje się stosując GMM do modelu pierwszych różnic.

Dokładniej, zastosowanie tej metody wymaga, aby model (1) spełniał założenia: $|\gamma| < 1$, $E(\alpha_i) = E(\varepsilon_{it}) = 0$, $E(\alpha_i\, \varepsilon_{it}) = 0$, $E(\varepsilon_{is}\, \varepsilon_{it}) = 0$ dla $i = 1, \ldots, N$, $t \neq s$ (brak autokorelacji składnika losowego ε_{it}), $E(y_{i1}\, \varepsilon_{it}) = 0$ dla $i = 1, \ldots, N$, $t = 2, \ldots, T$. Jeśli $N \to \infty$ a T jest skończone, to uwzględnione założenia implikują warunki momentów, których postać zależy od charakteru zmiennych $\mathbf{x}_{it}$. Znaczenie mają założenia, doty-

[3] Por. np. [5], [6] lub w języku polskim [15].

czące korelacji tych zmiennych z obiema składowymi składnika losowego modelu (1): z ε_{it} i z α_i.

Jeśli zmienne $\mathbf{x}_{it}$ są ściśle egzogeniczne i skorelowane z efektami grupowymi α_i, to warunki momentów mają postać:

$$E(\Delta x_{kit}\,\Delta \varepsilon_{it}) = 0 \text{ dla } t = 1, \ldots, T,\ k = 1, \ldots, K, \tag{2}$$

a macierz instrumentów:

$$\mathbf{Z}_i = \begin{bmatrix} \left[y_{i1}, \Delta \mathbf{x}_{i2}^T\right] & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \left[y_{i1}, y_{i2}, \Delta \mathbf{x}_{i3}^T\right] & & \mathbf{0} \\ \vdots & & \ddots & \mathbf{0} \\ \mathbf{0} & \ldots & \mathbf{0} & \left[y_{i1}, .., y_{i,T-2}, \Delta \mathbf{x}_{iT}^T\right] \end{bmatrix}. \tag{3}$$

Przypadek drugi zachodzi, gdy zmienne $\mathbf{x}_{it}$ są z góry ustalone, w sensie ich korelacji z właściwym składnikiem losowym ε_{it}, tzn. dla każdego k zachodzi $E(x_{kit}\,\varepsilon_{is}) \neq 0$ dla $t > s$, oraz $E(x_{it}\,\varepsilon_{is}) = 0$ dla $s \geq t$, i $\mathbf{x}_{it}$ są skorelowane z α_i. Oznacza to, że bieżące i opóźnione wartości zmiennych $\mathbf{x}_{it}$ są skorelowane z bieżącymi wartościami składnika losowego. Właściwymi instrumentami dla modelu pierwszych różnic uzyskanego z (1) w okresie s są zatem $x_{i,s-1}, \ldots, x_{i1}$. Warunki momentów mają wówczas postać:

$$E(x_{ki,t-s}\,\Delta \varepsilon_{it}) = 0 \text{ dla } t = 3, \ldots, T,\quad s \geq 2, \tag{4}$$

a macierz instrumentów:

$$\mathbf{Z}_i = \begin{bmatrix} \left[y_{i1}, \mathbf{x}_{i1}^T, \mathbf{x}_{i2}^T\right] & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \left[y_{i1}, y_{i2}, \mathbf{x}_{i1}^T, \mathbf{x}_{i2}^T, \mathbf{x}_{i3}^T\right] & & \mathbf{0} \\ \vdots & & \ddots & \mathbf{0} \\ \mathbf{0} & \ldots & \mathbf{0} & \left[y_{i1}, .., y_{i,T-2}, \mathbf{x}_{i1}^T, \ldots, \mathbf{x}_{i,T-1}^T\right] \end{bmatrix}. \tag{5}$$

W praktyce najczęściej zachodzi jednak przypadek mieszany, gdy elementami wektora $\mathbf{x}_{it}$ są zarówno zmienne ściśle egzogeniczne, jak i z góry ustalone w sensie ich korelacji z ε_{it}, oraz nie wszystkie zmienne $\mathbf{x}_{it}$ są skorelowane z efektami grupowymi α_i. W takim przypadku Arellano i Bond w pracy [5] proponują wykorzystanie idei Hausmana i Taylora zaproponowanej w [16], a więc dekompozycję wektora $\mathbf{x}_{it}$ na dwa podzbiory: zmiennych nieskorelowanych z α_i, i zmiennych skorelowanych z α_i, oraz odpowiednią modyfikację macierzy instrumentów. Dalsze postępowanie polega na zastosowaniu Uogólnionej Metody Momentów (GMM). Estymator wektora parametrów strukturalnych $[\gamma\ \beta^T]^T$ modelu (1) ma postać:

$$\begin{bmatrix} \hat{\gamma} \\ \hat{\boldsymbol{\beta}} \end{bmatrix}_{GMM} = \left[\left(\sum_{i=1}^{N} (\Delta \mathbf{y}_{i,-1}^T)^* \mathbf{Z}_i\right) \mathbf{W}_N \left(\sum_{i=1}^{N} \mathbf{Z}_i^T (\Delta \mathbf{y}_{i,-1})^*\right)\right]^{-1} \left[\left(\sum_{i=1}^{N} (\Delta \mathbf{y}_{i,-1}^T)^* \mathbf{Z}_i\right) \mathbf{W}_N \left(\sum_{i=1}^{N} \mathbf{Z}_i^T (\Delta \mathbf{y}_{i,-1})^*\right)\right], \tag{6}$$

gdzie $\mathbf{W}_N$ jest symetryczną, dodatnio określoną macierzą wag, a macierz $\mathbf{Z}_i$ jest określona w zależności od charakteru zmiennych $\mathbf{x}_{it}$, wzorem (3) lub (5). Nazywany jest on estymatorem GMM pierwszych różnic (FDGMM). Zastosowanie (6) w praktyce wymaga wcześniejszej estymacji macierzy wag. Estymator dwustopniowy wektora $\left[\hat{\gamma}\ \hat{\boldsymbol{\beta}}\right]^T_{GMM2}$, (FDGMM2) uzyskuje się, zastępując we wzorze (6) macierz $\mathbf{W}_N$ jej zgodnym i asymptotycznie efektywnym estymatorem postaci ogólnej:

$$\hat{\mathbf{W}}_N^{opt} = \left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{Z}_i^T\,\Delta\hat{\boldsymbol{\varepsilon}}_i\,\Delta\hat{\boldsymbol{\varepsilon}}_i^T\mathbf{Z}_i\right)^{-1}. \tag{7}$$

$\Delta\hat{\boldsymbol{\varepsilon}}_i$ oznacza wektor reszt pewnego początkowego estymatora $\left[\hat{\gamma}\ \hat{\boldsymbol{\beta}}\right]^T_{GMM1}$ wektora $[\gamma\ \boldsymbol{\beta}^T]^T$. Estymator (7) wyznacza się stosując procedurą iteracyjną, zaś estymator $\left[\hat{\gamma}\ \hat{\boldsymbol{\beta}}\right]^T_{GMM1}$ uzyskuje się zgodnie ze wzorem (6) dla początkowej macierzy wag $\mathbf{W}_{N1}$. Estymator $\left[\hat{\gamma}\ \hat{\boldsymbol{\beta}}\right]^T_{GMM1}$ nazywa się jednostopniowym estymatorem GMM pierwszych różnic (FDGMM1). Jako początkową macierz wag przyjąć można $\mathbf{W}_{N1} = \mathbf{I}$, lub zgodnie z propozycją Arellano i Bonda:

$$\mathbf{W}_{N1} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{Z}_i^T\mathbf{G}\mathbf{Z}_i, \text{ gdzie:}$$

$$\mathbf{G} = \begin{bmatrix} 2 & -1 & 0 & \dots \\ -1 & 2 & \ddots & 0 \\ 0 & \ddots & \ddots & -1 \\ \vdots & 0 & -1 & 2 \end{bmatrix}.$$

Macierz wariancji-kowariancji estymatora (6) jest w ogólnym przypadku dana wzorem:

$$p\lim_{N\to\infty}\left[\left(\frac{1}{N}\sum_{i=1}^{N}\Delta\mathbf{y}_{i,-1}^T\mathbf{Z}_i\right)\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{Z}_i^T\,\Delta\boldsymbol{\varepsilon}_i\,\Delta\boldsymbol{\varepsilon}_i^T\mathbf{Z}_i\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{Z}_i^T\,\Delta\mathbf{y}_{i,-1}\right)\right]^{-1}, \tag{8}$$

a przy dodatkowych założeniach dotyczących normalności rozkładu i sferyczności ε_{it} jej środkowy czynnik redukuje się do $\sigma_\varepsilon^2\tilde{\mathbf{W}}_N^{opt}$.

Możliwość zastosowania FDGMM jest problematyczna w przypadku, kiedy opóźnione poziomy zmiennych są słabymi instrumentami dla zmiennych zróżnicowanych (tzn. zmienne instrumentalne są zbyt słabo skorelowane ze zmienną objaśniającą). Sytuacja taka ma miejsce między innymi wtedy, gdy liczba obserwacji w czasie jest mała i wykorzystuje się szeregi czasowe o wysokim stopniu trwałości[4] (ang. *persistent*

[4] Wydaje się, że pojęcie to funkcjonuje w literaturze anglojęzycznej wyłącznie w kontekście panelowych modeli dynamicznych, gdzie definiowane jest jako proces, którego wartości przyszłe są silnie skorelowane z wartościami bieżącymi (*a highly persistent process is a time series process where outcomes in the distant future are highly correlated with current outcomes*). Jest ono ściśle związane z pojęciem niestacjonarności: szereg, który posiada pierwiastek jednostkowy jest definiowany jako szereg o wysokim stopniu trwałości, dla którego wartość bieżąca równa jest wartości z okresu poprzedniego plus składnik losowy (*a unit root*

time series). Estymatory FDGMM mogą być wówczas silnie obciążone. Szczegółowe rozważania na temat obciążenia estymatorów, spowodowanego zastosowaniem słabych instrumentów znaleźć można m.in. w pracy [20].

Systemowy estymator GMM (SGMM) Blundella i Bonda wykorzystuje dodatkowe warunki momentów, które są właściwe również w sytuacji, gdy instrumenty FDGMM są słabe, oraz założenia odnośnie warunków początkowych, umożliwiające uzyskanie warunków momentów, które są właściwe również dla szeregów o wysokim stopniu trwałości. Zasadnicza idea SGMM polega na oszacowaniu systemu równań zarówno na przyrostach, jak i na poziomach. Instrumentami dla zmiennych objaśniających w równaniach na poziomach są opóźnione pierwsze różnice zmiennych. Instrumenty te są właściwe, jeśli model (1), oprócz założeń przyjętych w [5], spełnia założenie dotyczące wartości początkowej, postaci: $E\left[\left(y_{i1} - \frac{\alpha_i}{1-\gamma}\right)\alpha_i\right] = 0$. Jest to założenie dotyczące stacjonarności średniej[5] i wynika z niego warunek:

$$E(\alpha_i \, \Delta y_{i2}) = 0 \text{ dla } i = 1, \ldots, N. \tag{9}$$

Estymator SGMM dla modelu autoregresyjnego (tzn. modelu postaci (1), w którym $\beta = \mathbf{0}$), na którym koncentrują się rozważania Blundella i Bonda zawarte w [8] wykorzystuje dwa układy warunków momentów, postaci[6]:

$$\begin{cases} E\left(\mathbf{Z}_i^T \, \Delta \boldsymbol{\varepsilon}_i\right) = \mathbf{0} \\ E\left(\mathbf{Z}_i^{+T} \mathbf{u}_i^+\right) = \mathbf{0} \end{cases}, \tag{10}$$

gdzie $\mathbf{u}_i^+ = \left[\Delta \varepsilon_{i3}, \ldots, \Delta \varepsilon_{iT}, u_{i3}, \ldots, u_{iT}\right]^T$, $\Delta \boldsymbol{\varepsilon}_i = \left[\Delta \varepsilon_{i2}, \ldots, \Delta \varepsilon_{iT}\right]$, a macierze instrumentów mają postać:

$$\mathbf{Z}_i = \begin{bmatrix} [y_{i1}] & \mathbf{0} & \ldots & \mathbf{0} \\ 0 & [y_{i1}, y_{i2}] & & \mathbf{0} \\ \vdots & & \ddots & \mathbf{0} \\ 0 & \ldots & \mathbf{0} & [y_{i1}, \ldots, y_{i,T-2}] \end{bmatrix}, \quad \mathbf{Z}_i^+ = \begin{bmatrix} \mathbf{Z}_i & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \Delta y_{i2} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \Delta y_{i3} & \ldots & \mathbf{0} \\ \ldots & \ldots & \ldots & \ddots & \ldots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \ldots & \Delta y_{i,T-1} \end{bmatrix}. \tag{11}$$

Jeśli $\beta \neq \mathbf{0}$, to postać macierzy zmiennych instrumentalnych zależy od charakteru zmiennych egzogenicznych, analogicznie jak w przypadku FDGMM. W przypadku, gdy zmienne $\mathbf{x}_{it}$ są ściśle egzogeniczne, to postępowanie jest takie samo jak w przypadku estymatora FDGMM. Jeśli zaś zmienne $\mathbf{x}_{it}$ nie są ściśle egzogeniczne, to przyjmując

process is a highly persistent time series process where the current value equals last period's value, plus a weakly dependent disturbance).

[5] Wyróżniamy dwa rodzaje niestacjonarności: w średniej (trend deterministyczny) i wariancji (trend stochastyczny) procesu.

[6] Pierwszy z układów określonych wzorem (10) to układ warunków momentów wykorzystywany w FDGMM. Drugi z nich tworzą dodatkowe warunki uzyskane dzięki spełnieniu warunku (9).

dla tych zmiennych założenie początkowe typu (9) można uzyskać dodatkowe warunki momentów postaci:

$$E(\Delta \mathbf{x}_{it-1}\mathbf{u}_{it}) = 0 \text{ dla } i = 1, \ldots, N \text{ oraz } t = 3, \ldots, T. \tag{12}$$

Zatem opóźnione pierwsze różnice $\Delta\mathbf{x}_{it-1}$ mogą być zmiennymi instrumentalnymi w równaniach na poziomach i zespół wszystkich możliwych warunków momentów tworzą warunki (4) i (12). Wykorzystując ten układ warunków do estymacji GMM systemu $(T-2)$ równań na przyrostach i $(T-2)$ równań na poziomach, uzyskuje się jednostopniowy i dwustopniowy estymator SGMM, analogicznie jak opisano to dla FDGMM.

W pracy [8] wykazano, że zastosowanie systemowego estymatora GMM przynosi znaczący wzrost efektywności w stosunku do FDGMM szczególnie, gdy wartość parametru autoregresyjnego $\gamma \to 1$ lub gdy stosunek wariancji $\sigma_\alpha^2/\sigma_\varepsilon^2$ rośnie.

Przy założeniu homoskedastyczności ε_{it} estymatory jednostopniowe FDGMM i SGMM są równoważne estymatorom dwustopniowym. Jeśli zaś ε_{it} jest heteroskedastyczny, to estymatory dwustopniowe są bardziej efektywne. Bond, Hoeffler i Temple w pracy [10] stwierdzają jednak, powołując się na eksperymenty Monte Carlo, że wzrost efektywności jest z reguły niewielki, oraz że estymator dwustopniowy zbiega do swojego rozkładu asymptotycznego dość powoli. W skończonych próbach błędy szacunku tego estymatora mogą ponadto być silnie obciążone w dół.

Przegląd metod estymacji pod kątem ich przydatności do estymacji jednorównaniowych modeli dynamicznych na podstawie mikropaneli zawarty jest w artykule [11]. Metody, które bazują na GMM są, według autora, szczególnie użyteczne w przypadku modeli zawierających endogeniczne lub z góry ustalone zmienne objaśniające. Endogeniczność zmiennych objaśniających jest zjawiskiem naturalnym: założenie o ścisłej ich egzogeniczności oznaczałoby, że przeszłe wartości zmiennych nie wpływają na wartości bieżące, co byłoby trudne do przyjęcia w przypadku zjawisk często modelowanych na podstawie mikropaneli, takich jak: konsumpcja, dochód, inwestycje, etc. Autor zwraca też uwagę na własności estymatorów w przypadku małych prób (jeśli dodatkowo wymiar przekrojowy próby jest niewielki), oraz warunków początkowych w kontekście mikropaneli. Wpływ obserwacji początkowej na każdą kolejną obserwację jest tu szczególnie istotny ze względu na krótkość szeregu czasowego. Ilustracją tych rozważań są dwa przykłady empiryczne, omówione w części 4.

Własności estymatorów bazujących na GMM w kontekście modeli dynamicznych szacowanych na podstawie makropaneli są rozważane w [10]. W przypadku tego typu danych, podobnie jak w przypadku mikropaneli, często pojawiającymi się problemami są endogeniczność i błędy pomiaru zmiennych, jak również konsekwencje wynikające z pominiętych zmiennych. Źródłem problemów w przypadku wykorzystania makropaneli może też być mała próba, oraz wysoki stopień trwałości szeregu. Rozważania są zilustrowane przykładem zastosowania estymatorów FDGMM i SGMM z alternatywnymi zbiorami instrumentów do szacowania modelu wzrostu PKB i konwergencji. Omówienie wyników estymacji tego modelu zawarte jest w części 5.

4. PRZYKŁADY ZASTOSOWANIA PANELOWYCH MODELI DYNAMICZNYCH W BADANIACH MIKROEKONOMICZNYCH

W artykule [11] Bond prezentuje przykłady zastosowania alternatywnych estymatorów do oszacowania parametrów dwóch modeli: autoregresyjnego modelu stopy inwestycji dla firm brytyjskich, oraz modelu produkcji dla firm z USA.

Model stopy inwestycji jest szacowany na podstawie 4966 obserwacji, pochodzących z 703 firm, dla których dostępne były dane roczne za co najmniej 4 kolejne lata w okresie 1988-2000. Model ma postać:

$$\left(\frac{I_{it}}{K_{it}}\right) = c_t + \gamma\left(\frac{I_{i,t-1}}{K_{i,t-1}}\right) + (\alpha_i + \varepsilon_{it}), \tag{13}$$

gdzie I_{it} oznacza wydatki inwestycyjne brutto, K_{it} – kapitał akcyjny netto i koszt odtworzenia. Stosunek I_{it}/K_{it} jest zatem przybliżoną miarą stopy wzrostu kapitału akcyjnego netto plus stopy amortyzacji. Zmienność w czasie wyrazu wolnego c_t odzwierciedlać ma jednakowe dla wszystkich firm składowe stopy inwestycji, wynikające z cykliczności lub tendencji rozwojowej.

Do oszacowania modelu zastosowane zostały cztery alternatywne estymatory: KMNK, wewnątrzgrupowy (WG), 2MNK dla modelu pierwszych różnic i FDGMM. Wyniki prezentowane są w [11] w tabeli 1, na s. 31. Oceny parametru γ wahają się od –0,01 dla estymatora WG do 0,27 dla estymatora KMNK. Jest to duża rozbieżność, która potwierdza regułę, że w przypadku panelowych modeli dynamicznych estymator KMNK jest obciążony w górę, a WG – w dół.

Ocena parametru autoregresyjnego uzyskana na podstawie 2MNK wynosi 0,1626 z błędem szacunku 0,036. Estymator 2MNK dla modelu pierwszych różnic jest wyznaczony przy zastosowaniu zmiennej instrumentalnej $(I/K)_{i,t-2}$. Ponieważ model jest jednoznacznie identyfikowalny, nie można zastosować testu Sargana, zatem oceny jego jakości dokonać można na podstawie testu autokorelacji Arellano-Bonda[7]. Przy standardowym założeniu, że składniki losowe ε_{it} nie wykazują korelacji w czasie, w prawidłowo wyspecyfikowanym modelu należy się spodziewać, że przyrosty $\Delta\varepsilon_{it}$ będą wykazywały istotną ujemną korelację pierwszego rzędu i brak istotnej korelacji drugiego rzędu. Wynik testu Arellano-Bonda (tabela 1, s. 31) potwierdza prawidłową specyfikację modelu.

Kolejną metodą zastosowaną do estymacji modelu (13) jest jednostopniowa FDGMM. Autor prezentuje wyniki uzyskane dla dwóch różnych zbiorów instrumentów. W pierwszej wersji zbiór zmiennych instrumentalnych jest ograniczony do $(I/K)_{i,t-2}$ i $(I/K)_{i,t-3}$. Mimo tego, że w procesie estymacji nie wykorzystuje się pełnej macierzy $\mathbf{Z}_i$, danej wzorem (5), zauważalna jest pewna poprawa precyzji estymacji w stosunku do 2MNK (błąd szacunku spada z 0,0362 do 0,0327, przy ocenie parametru γ równej 0,1593). Poprawa wynikać może z wykorzystania dodatkowych warunków momentów.

[7] Test Arellano-Bonda bada występowanie autokorelacji drugiego rzędu składnika losowego $\Delta\varepsilon$ w modelu pierwszych różnic. Występowanie w modelu pierwszych różnic autokorelacji rzędu wyższego niż 1 oznaczałoby, że warunki momentów są niespełnione, a więc instrumenty użyte podczas estymacji GMM nie są właściwe. Hipoteza H_0 testu Arellano-Bonda głosi, że autokorelacja taka nie występuje (por. np. [5], [6]).

Autor odnotowuje, że macierz $\mathbf{Z}_i$ wykorzystana do 2MNK składa się z jednej kolumny (postaci $[(I/K)_{i1}\ (I/K)_{i2}\ \ldots\ (I/K)_{i,T-2}]^T$, a FDGMM, nawet w ograniczonej wersji, wykorzystuje macierz liczącą 23 kolumny. Model jest obecnie identyfikowalny z nadmiarem, zatem do oceny jego jakości zastosować można test Sargana warunków ponadidentyfikujących. Wartość statystyki *s* jest równa 23,84, a wyznaczone *p-value* = 0,36 oznacza, że warunki ponadidentyfikujące są właściwe, a więc model jest poprawny. Wynik testu Sargana potwierdzony jest wynikiem testu Arellano-Bonda.

Następnie jednostopniowa FDGMM zastosowana jest ponownie, z pełnym zbiorem zmiennych instrumentalnych, wynikających z założeń przyjętych w [5]. Macierz $\mathbf{Z}_i$ wykorzystywana w tym przypadku liczy aż 78 kolumn. Wykorzystanie dodatkowych warunków momentów przynosi dalszą poprawę precyzji estymacji. Ocena parametru γ wynosi obecnie 0,156, a błąd jego szacunku 0,0318. Testy Sargana i Arellano-Bonda potwierdzają hipotezę o poprawności modelu. Powodem, dla którego (oprócz FDGMM z pełnym zbiorem instrumentów) stosowana jest FDGMM z ograniczonym zbiorem instrumentów jest to, że estymatory GMM wyznaczane na podstawie zbyt wielu warunków momentów mogą być obciążone w małych próbach, a przyczyną tego obciążenia (ang. *overfitting bias*) może być właśnie nadmiar wykorzystanych warunków momentów. Porównanie obu estymatorów FDGMM pokazuje jednak, że w omawianym modelu ten problem nie wystąpił, co więcej utrata efektywności spowodowana przez pominięcie zmiennych instrumentalnych, którymi są bardziej odległe w czasie opóźnienia jest bardzo niewielka.

Dodatkowo model (13) estymowany został za pomocą dwustopniowej FDGMM z pełnym zbiorem instrumentów. Wyznaczona za pomocą tej metody ocena parametru γ wynosi 0,1806. Błąd szacunku tego parametru szacowany jest dwoma metodami. Błąd asymptotyczny obliczony według wzoru (8) ma wartość 0,0227, a więc jest o około 30% mniejszy niż błąd estymatora jednostopniowego, przy zbliżonych wartościach ocen parametru. Autor stwierdza, że tak duży wzrost efektywności estymacji w wyniku zastosowania estymatora dwustopniowego jest nieuzasadniony i nigdy nie był odnotowany w symulacjach. Błąd obliczony przy zastosowaniu korekty Windmeijera[8], jest równy 0,0314, więc jest praktycznie taki sam, jak błąd estymatora jednostopniowego z pełnym zbiorem instrumentów. Rezultat ten potwierdza udowodnioną w literaturze tezę (por. np. [12]), że praktyczne wnioskowanie na asymptotycznej macierzy wariancji-kowariancji dwustopniowego estymatora GMM jest ryzykowne.

Innym przykładem zastosowania panelowego modelu dynamicznego w badaniach mikroekonomicznych jest dynamiczna specyfikacja funkcji produkcji, estymowana przez Blundella i Bonda w [9]. Przykład ten pokazuje dlaczego specyfikacja dynamiczna może być konieczna dla zapewnienia identyfikowalności pewnych parametrów, mimo że dynamika jako taka nie jest centralnym punktem analizy. Punktem wyjścia jest funkcja produkcji Cobba-Douglasa, postaci:

[8] Windmeijer w pracy [21] wykazał, że zasadniczą przyczyną obciążenia dwustopniowego estymatora GMM w małych próbach jest obecność początkowych estymatorów parametrów strukturalnych w macierzy wag. Udowodnił on też, że wielkość tego obciążenia można estymować, oraz podał wzór według którego można wyznaczyć korektę estymatorów wariancji parametrów, zwaną w późniejszych opracowaniach korektą Windmeijera.

$$\begin{aligned} y_{it} &= \beta_n n_{it} + \beta_k k_{it} + \lambda_t + (\alpha_i + \varepsilon_{it} + m_{it}) \\ \varepsilon_{it} &= \gamma \varepsilon_{i,t-1} + e_{it} \qquad\qquad |\gamma| < 1 \\ e_{it},\ m_{it} &\sim MA(0) \end{aligned} \tag{14}$$

gdzie y_{it}, n_{it}, k_{it} oznaczają kolejno logarytmy: sprzedaży, zatrudnienia i kapitału akcyjnego firmy i w roku t, λ_t jest specyficznym dla danego roku wyrazem wolnym (jednakowym dla wszystkich firm), odzwierciedlającym postęp technologiczny, β_n, β_k, γ są parametrami. Składowe składnika losowego to: efekt grupowy α_i, potencjalnie autoregresyjny „właściwy" składnik losowy ε_{it}, oraz składowa m_{it}, odzwierciedlająca błędy pomiaru, o której zakłada się, że nie wykazuje autokorelacji. Zakłada się, że zatrudnienie i kapitał są skorelowane z efektami grupowymi α_i, oraz z pozostałymi dwiema składowymi składnika losowego (e_{it} i m_{it}). Przy tych założeniach nie można wskazać właściwych warunków momentów dla modelu statycznego, postaci (14), o ile składnik losowy ε_{it} rzeczywiście jest autoregresyjny, to znaczy $\gamma \neq 0$. Model (14) zapisać można jednak równoważnie w postaci:

$$\begin{aligned} y_{it} = \beta_n n_{it} - \gamma\beta_n n_{i,t-1} + \beta_k k_{it} - \gamma\beta_k k_{i,t-1} + \gamma y_{i,t-1} + \\ + (\lambda_t - \gamma\lambda_{t-1}) + \left(\alpha_i(1-\gamma) + e_{it} + m_{it} - \gamma m_{i,t-1}\right) \end{aligned}, \tag{15}$$

lub w postaci:

$$y_{it} = \pi_1 n_{it} + \pi_2 n_{i,t-1} + \pi_3 k_{it} + \pi_4 k_{i,t-1} + \pi_5 y_{i,t-1} + \lambda_t^* + \left(\alpha_i^* + w_{it}\right), \tag{16}$$

z dwoma nieliniowymi restrykcjami wspólnego czynnika: $\pi_2 = -\pi_1\pi_5$ i $\pi_4 = -\pi_3\pi_5$. Ponieważ składnik losowy ($\varepsilon_{it} + m_{it}$) modelu (14) wykazuje autokorelację to możliwe są dwa przypadki: (a) nie występują błędy pomiaru, to znaczy $w_{it} = e_{it}$, oraz var $(m_{it}) = 0$, to składnik losowy w_{it} modelu (16) nie wykazuje autokorelacji, lub (b) jeśli w niektórych szeregach występują niesystematyczne (ang. *transient*) błędy pomiaru, to $w_{it} \sim MA(1)$. W każdym z tych dwóch przypadków możliwe jest uzyskanie zgodnych estymatorów wektora parametrów $\boldsymbol{\pi} = [\pi_1 \ \ldots \ \pi_5]$ za pomocą GMM. Po wyznaczeniu zgodnych estymatorów wektora $\boldsymbol{\pi}$ i wariancji var $(\boldsymbol{\pi})$, możliwe jest nałożenie i testowanie restrykcji wspólnego czynnika, postulowanych w modelu (16) i wyznaczenie zgodnych estymatorów parametrów β_n, β_k, γ.

W pracy [9] oszacowano parametry modelu (14) na podstawie zrównoważonego panelu danych z 509 firm prowadzących działalność badawczo-rozwojową w USA w latach 1982-1989. Kapitał akcyjny i zatrudnienie było mierzone na koniec roku rozrachunkowego w danej firmie.

Szeregi obserwacji na zmiennych y_{it}, n_{it}, k_{it} okazały się charakteryzować wysokim stopniem trwałości, co zostało wykazane poprzez konstrukcję i estymację modeli AR(1) dla tych zmiennych. Oceny MNK parametru autoregresyjnego prezentowane w [9] wynoszą dla wszystkich szeregów 0,99. Oceny wyznaczone na podstawie SGMM są niewiele niższe – od 0,92 do 0,96. Zaskakujące, zdaniem autorów jest, że ocena FDGMM parametru autoregresyjnego dla zmiennej n_{it} jest niemal identyczna (0,920) jak ocena SGMM (0,923). Dla pozostałych dwóch zmiennych potwierdzone zostało

przypuszczenie, że w związku z wysokim stopniem trwałości szeregów oceny FDGMM są obciążone w dół – wynoszą one 0,768 dla zmiennej k_{it} i 0,775 dla y_{it}.

Autorzy prezentują wyniki estymacji modeli (14) i (16) wyznaczone na podstawie KMNK, WG, oraz jednostopniowych FDGMM i SGMM. Zgodnie z oczekiwaniami oceny parametru γ KMNK są obciążone w górę, a oceny WG – obciążone w dół. FDGMM stosowana jest dwukrotnie: z opóźnionymi o dwa okresy poziomami jako zmiennymi instrumentalnymi, oraz powtórnie z instrumentami, którymi są poziomy opóźnione o trzy okresy. Właściwość instrumentów w pierwszym przypadku została odrzucona przez test Sargana. Odrzucenie to wynika z występowania błędów pomiaru m_{it}. Instrumenty w drugim przypadku są właściwe, a restrykcje wspólnego czynnika spełnione. Jednakże oceny wszystkich parametrów (w tym parametru γ przy opóźnionej zmiennej objaśnianej) są zbliżone do ocen estymatora wewnątrzgrupowego, co sugeruje możliwość wystąpienia obciążenia wynikającego z małej próby i problemu słabych instrumentów.

Właściwość poziomów opóźnionych o trzy okresy jako instrumentów dla równań na przyrostach i przyrostów opóźnionych o dwa okresy, jako instrumentów dla równań na poziomach jest na granicy istotności w teście Sargana. Jednakże wynik różnicowego testu Sargana, badającego poprawność dodatkowych warunków momentów, uzyskanych dzięki wprowadzeniu instrumentów dla równań na poziomach, potwierdza właściwość tych instrumentów na poziomie istotności 0,1. Ponadto o poprawności SGMM świadczy to, że ocena parametru γ uzyskana tą metodą jest wyższa niż ocena FDGMM i niższa niż ocena KMNK. Restrykcje wspólnego czynnika są również spełnione. Dlatego autorzy [9] uznają estymator SGMM za najwłaściwszy dla rozważanego modelu.

5. PRZYKŁADY ZASTOSOWANIA PANELOWYCH MODELI DYNAMICZNYCH W BADANIACH MAKROEKONOMICZNYCH (MODELOWANIU WZROSTU)

Wzmożone zainteresowanie konstrukcją i wynikami estymacji modeli wzrostu i konwergencji datuje się od początku lat 90. XX wieku, a jako przykłady znaczących prac z tego zakresu wymienić można prace [7] i [19]. Klasyczne modele wzrostu, oparte na pracy Solowa, zostały w późniejszych latach podważone przez modele bazujące na teorii wzrostu endogenicznego. Generalnie uważa się, że stwierdzenie występowania zjawiska konwergencji jest potwierdzeniem słuszności założeń Solowa, a stwierdzenie, że konwergencja nie występuje przemawia za poprawnością teorii wzrostu endogenicznego. Wspólną cechą badań empirycznych nad wzrostem gospodarczym, które bazują na danych przekrojowych jest założenie jednakowej zagregowanej funkcji produkcji dla wszystkich krajów. Do przyjęcia takiego założenia zmusza brak możliwości uwzględnienia w modelach danych przekrojowych niemierzalnych różnic między krajami. Jedną z pierwszych prac, które wykorzystują dane panelowe do estymacji dynamicznego modelu wzrostu jest praca [18]. Jak zauważa jej autor, wykorzystanie danych panelowych daje możliwość uwzględnienia wspomnianych różnic w procesie konstrukcji funkcji produkcji, poprzez wyodrębnienie efektów grupowych. Jednym z celów pracy [18] było porównanie wyników uzyskanych na podstawie danych panelowych z wynikami uzyskanymi na podstawie regresji przekrojowej, prezentowanych w [19]. Dlatego w obu

pracach wykorzystano analogiczne trzy zbiory krajów (98, 75 i 22 kraje). Ponieważ konstrukcja panelu na podstawie danych rocznych nie eliminowałaby wpływu zakłóceń krótkookresowych, wykorzystano dane z okresów 5-letnich z lat 1960-1985. Estymacji dokonano czterema metodami: KMNK na podstawie danych przekrojowych, KMNK na podstawie danych panelowych (ang. *pooled regression*), metodą minimalnej odległości[9] (ang. *minimum distance*), na podstawie modelu z efektami nielosowymi (estymator WG). Wyniki uzyskane w [18] na podstawie modeli panelowych wskazują, że tempo konwergencji jest znacznie szybsze niż wyznaczone na podstawie regresji przekrojowej, a oszacowane wartości elastyczności produkcji względem kapitału – niższe i bardziej akceptowalne. Z punktu widzenia teorii wzrostu, wykorzystanie modeli panelowych daje możliwość wyodrębnienia wpływu różnic technologicznych i instytucjonalnych między krajami na wzrost gospodarczy i na proces konwergencji.

W świetle późniejszej wiedzy, a szczególnie prac [5] oraz [8], stwierdzić trzeba, że wyniki prezentowane w [18] są obciążone w górę, co jest skutkiem zastosowania estymatora WG do estymacji modelu dynamicznego. Istotnie, tempo konwergencji wyznaczone dla najmniejszej grupy krajów na podstawie estymatora WG, wynoszące aż 9% wydaje się zbyt szybkie. Ogólnie stwierdzić trzeba, że estymacja dynamicznych modeli wzrostu wymaga zastosowania właściwych metod, ze względu na występowanie problemów związanych z endogenicznością zmiennych objaśniających, błędami pomiaru zmiennych lub pominiętymi zmiennymi.

Autorami, którzy pierwsi zastosowali FDGMM do estymacji modelu wzrostu na podstawie panelowego modelu dynamicznego byli Caselli, Esquivel i Lefort (por. [13]. Jak zauważają Bond, Hoeffler i Temple w [10], wykorzystanie do estymacji modeli wzrostu danych panelowych i zastosowanie GMM ma przewagę nad regresją przekrojową i innymi metodami estymacji. Po pierwsze, dzięki wykorzystaniu danych panelowych można oszacować model na przyrostach, co pozwala uniknąć obciążenia estymatorów, wynikającego z pominięcia zmiennych objaśniających stałych w czasie (stanowią one składową efektu grupowego i są eliminowane podczas obliczania pierwszych różnic). W szczególności, w warunkowych równaniach konwergencji pominiętą zamienną może być początkowy poziom efektywności, który jest nieobserwowalny, a dodatkowo skorelowany z początkowym poziomem dochodu, będącym zazwyczaj jedną ze zmiennych objaśniających. Po drugie, zastosowanie zmiennych instrumentalnych pozwala na wyznaczenie zgodnych estymatorów parametrów w modelach, zawierających endogeniczne zmienne objaśniające, takie jak np. stopa inwestycji, a także w przypadku obecności błędów pomiaru zmiennych.

Bond, Hoeffler i Temple (por. [10], w dalszym ciągu BHT) zwracają uwagę na to, że zastosowanie FDGMM do estymacji modeli wzrostu może być niepoprawne, co jest związane z problemem słabych instrumentów. Dane dotyczące produkcji mają często cechy procesu o wysokim stopniu trwałości, a ponadto panelowe modele wzrostu estymowane są często na podstawie kilku obserwacji w czasie, które dodatkowo mogą być

[9] Metoda minimalnej odległości została opracowana przez Chamberlaina w pracy [14]. Zamiast eliminowania efektów grupowych α_i poprzez wyznaczenie pierwszych różnic, specyfikuje się α_i jako funkcję tych zmiennych, z którymi może ona być skorelowana.

średnimi z kilku lat. Te cechy powodują, że opóźnione poziomy zmiennych są słabymi instrumentami dla równań na przyrostach, co skutkuje obciążeniem estymatorów.

Model wzrostu Solowa, zastosowany przez BHT ma postać:

$$\Delta y_{it} = \alpha_{0t} + (\gamma - 1)y_{i,t-1} + \mathbf{x}_{it}^T\boldsymbol{\beta} + \alpha_i + \varepsilon_{it}, \text{ dla } i = 1, \ldots, N, t = 2, \ldots, T \qquad (17)$$

gdzie: Δy_{it} jest przyrostem logarytmu PKB *per capita* w okresie 5 lat, $y_{i,t-1}$ – logarytmem PKB *per capita* na początku tego okresu, a $\mathbf{x}_{it}$ – wektorem zmiennych objaśniających, mierzonych na początku lub podczas tego okresu, którymi mogą być: logarytm stopy inwestycji (s_{it}), logarytm stopy wzrostu populacji (n_{it}) plus 0,05, gdzie 0,05 reprezentuje sumę wspólnej dla wszystkich obiektów egzogenicznej stopy postępu technicznego (g) i wspólnej stopy deprecjacji (δ). W rozszerzonym modelu Solowa regresorami mogą też być miary jakości kapitału ludzkiego, na przykład logarytm wskaźnika skolaryzacji (enr_{it}) dla szkół średnich (stosunku liczby uczniów tych szkół do liczby ludności w wieku odpowiadającym ustawowemu okresowi pobierania nauki na tym poziomie edukacji). Nieobserwowalny efekt grupowy α_i odzwierciedla między innymi różnice początkowego poziomu efektywności, zaś zmieniający się w czasie wyraz wolny (efekt czasowy α_{0t}) odzwierciedla jednakowe dla wszystkich obiektów zmiany produktywności.

BHT reestymowali za pomocą SGMM model, który oszacowany został oryginalnie w pracy [13] za pomocą FDGMM, na podstawie danych z okresów 5-letnich z lat 1960-1985. Model (17) można oczywiście zapisać równoważnie w postaci:

$$y_{it} = \alpha_{0t} + \gamma y_{i,t-1} + \mathbf{x}_{it}^T\boldsymbol{\beta} + \alpha_i + \varepsilon_{it}, \text{ dla } i = 1, \ldots, N, t = 2, \ldots, T \qquad (18)$$

Spełnienie dodatkowych warunków momentów, wymaganych dla zastosowania SGMM jest zagwarantowane, jeśli średnie po czasie wszystkich zmiennych (objaśnianej i objaśniających) są stałe dla każdego obiektu. Zostało to pokazane w pracy [9] dla funkcji produkcji. BHT zauważają, że o ile stałość średnich dla stopy inwestycji i stopy wzrostu populacji jest zgodna z postulatami modelu wzrostu Solowa, to stałość średniej wartości PKB *per capita* jest problematyczna. Jednakże włączenie do modelu efektów czasowych α_{0t} pozwala na założenie, że długookresowy wzrost PKB *per capita*, oraz postęp techniczny są jednakowe dla wszystkich krajów. Założenie dotyczące jednakowego postępu technicznego zostało przyjęte w pracy [19] i jest odtąd standardowo przyjmowane w modelach wzrostu.

BHT podkreślają, że włączenie efektów czasowych do modelu (17) jest równoważne przekształceniu zmiennych w ich odchylenia od średnich grupowych, czyli ich transformacji za pomocą operatora wewnątrzgrupowego. W estymowanym przez nich modelu, tak samo jak w pracy [13], zmienne zostały przekształcone w odchylenia od średnich, co umożliwia porównywalność wyników, oraz pominięcie efektów czasowych. Wyniki prezentowane przez autorów uzyskane są na podstawie estymatorów jednostopniowych FDGMM i SGMM ze średnimi błędami szacunku odpornymi na występowanie heteroskedastyczności.

Podstawowy model, postaci (17), w którym elementami wektora $\mathbf{x}_{it}$ są $\ln(s_{it})$ i $\ln(n_{it} + g + \delta)$, szacowany jest czterema metodami: MNK, WG, FDGMM, SGMM. Stopa inwestycji i stopa wzrostu populacji zostały uznane za zmienne endogeniczne, co

przy zastosowaniu dwóch ostatnich metod znalazło odzwierciedlenie w zastąpieniu ich zmiennymi instrumentalnymi. Szczegółowe wyniki zawarte są w pracy BHT w tabeli 1 na s. 31. Instrumentami w modelu pierwszych różnic, wykorzystanymi w FDGMM są dla odpowiednich zmiennych: $\ln(y_{i,t-2})$, $\ln(s_{i,t-2})$, $\ln(n_{i,t-2} + g + \delta)$, oraz wcześniejsze opóźnienia. Dodatkowymi instrumentami w równaniach na poziomach, wykorzystanymi w SGMM są dla odpowiednich zmiennych: $\Delta\ln(y_{i,t-1})$, $\Delta\ln(s_{i,t-1})$, $\Delta\ln(n_{i,t-1} + g + \delta)$. Wśród najważniejszych wniosków wymienić należy to, że ocena FDGMM parametru przy dochodzie początkowym $y_{i,t-1}$, równa –0,537 jest niższa niż ocena WG (–0,031), co wskazuje na jej obciążenie w dół, właściwe dla małej próby. Ponadto, odpowiadająca jej ocena tempa konwergencji $\hat{\beta} = 0{,}6$, jest znacznie wyższa niż uzyskana innymi metodami. SGMM daje ocenę parametru przy dochodzie początkowym równą –0,101, przy czym jest to wartość znajdująca się pomiędzy oceną WG i MNK. Wynikająca stąd ocena tempa konwergencji $\hat{\beta} = 0{,}021$ wydaje się znacznie sensowniejsza, oznacza bowiem, że stopa konwergencji wynosi 2%. Wyniki testu Sargana i różnicowego testu Sargana wskazują na właściwość użytych instrumentów i poprawność dodatkowych warunków momentów, wykorzystanych przez SGMM. Zauważalne jest też zmniejszenie błędów szacunku estymatorów SGMM w stosunku do FDGMM. Traktowanie stopy inwestycji i stopy wzrostu populacji jako potencjalnych zmiennych endogenicznych umożliwia uwzględnienie zmiennego w czasie, to znaczy niesystematycznego błędu pomiaru każdej z nich. Przy założeniu występowania niesystematycznego błędu pomiaru PKB *per capita* zarówno poziom tej zmiennej z okresu $t-2$ w równaniu na przyrostach, jak i przyrost z okresu $t-1$ w równaniu na poziomach nie są właściwymi instrumentami. Mimo że testy Sargana nie wykazały niewłaściwości instrumentów użytych w SGMM, BHT ponownie zastosowali SGMM z wyłączeniem wymienionych zmiennych instrumentalnych. Dokładniej, $\ln(y_{i,t-2})$ została pominięta w zbiorze instrumentów dla modelu pierwszych różnic, a $\Delta\ln(y_{i,t-1})$ – zastąpiona przez $\Delta\ln(y_{i,t-2})$ w równaniach na poziomach. Wyniki, również prezentowane przez autorów w tabeli 1, różnią się bardzo nieznacznie od uzyskanych w podstawowej wersji SGMM. Wydaje się to wskazywać na brak problemów z błędami pomiaru PKB *per capita*. Oprócz wspomnianego już określenia stopy konwergencji na poziomie 2%, oceny SGMM wskazują na istotny, pozytywny wpływ stopy inwestycji (ok. 0,19) na poziom PKB *per capita* właściwy dla stanu wzrostu zrównoważonego (ang. *steady state*) i to, co podkreślają autorzy, przy uwzględnieniu efektów grupowych i możliwej endogeniczności inwestycji.

BHT rozważają też model rozszerzony, w którym do zbioru zmiennych objaśniających dołączono wskaźnik skolaryzacji dla szkół średnich. Wyniki są prezentowane przez autorów w tabeli 2 na s. 32. Dodanie tej zmiennej poprawiło nieco jakość estymatora FDGMM – ocena parametru przy dochodzie początkowym jest porównywalna z oceną WG (–0,331 i –0,321). Wiadomo jednak, że estymator WG w małych próbach jest, w przypadku modeli dynamicznych, obciążony ponadto tempo konwergencji, obliczone na podstawie tych ocen wynosiłoby aż 0,08. Bardziej poprawne wyniki uzyskano na podstawie SGMM – stopa konwergencji wyznaczona z modelu uwzględniającego wpływ wskaźnika skolaryzacji wynosi 1,7%, jednakże błąd szacunku parametru przy tej zmiennej (–0,046) przekracza wartość oceny tego parametru (–0,018). BHT sugerują nieuwzględnianie wskaźnika skolaryzacji bezpośrednio w modelu, ale raczej dołączenie opóźnionych wartości tej zmiennej do zbioru instrumentów, wykorzysty-

wanych w FDGMM, w celu rozwiązania problemu słabych instrumentów. Rezultaty przedstawione są w tabeli 3 na s. 33. Dokładniej, zbiór instrumentów dla równań na przyrostach podstawowego modelu Solowa, został rozszerzony o $\ln(enr_{i,t-1})$ oraz wcześniejsze opóźnienia. W ten sposób uzyskane oceny FDGMM są bliskie ocenom SGMM. Opóźniony wskaźnik skolaryzacji pomaga zatem przewidywać wzrost PKB *per capita* w równaniach postaci zredukowanej, podczas gdy bieżące wartości tej zmiennej nie wpływają na właściwy dla stanu wzrostu zrównoważonego poziom PKB *per capita* w rozważanym modelu. BHT interpretują ten rezultat jako oddziaływanie wskaźnika skolaryzacji na wzrost gospodarczy za pośrednictwem stopy inwestycji.

6. ZAKOŃCZENIE

W prezentowanym artykule przedstawiono możliwości zastosowania panelowych modeli dynamicznych w badaniach ekonomicznych. Zaprezentowano też zarys dwóch alternatywnych metod estymacji takich modeli, mających, jak się wydaje, najszersze zastosowanie praktyczne. Analiza prezentowanych przykładów empirycznych konstrukcji i estymacji modeli mikro- i makroekonometrycznych, przy zastosowaniu alternatywnych metod, ilustruje znaczenie doboru odpowiedniej metody estymacji dla prawidłowości wnioskowania ekonomicznego.

Uniwersytet Łódzki

LITERATURA

[1] Ahn S.C., Schmidt P., [1995], *Efficient estimation of models for dynamic panel data*, „Journal of Econometrics", 68, s. 5-28.
[2] Anderson, T.W., Hsiao C., [1981], *Estimation of Dynamic models with Error Components*, „Journal of the American Statistical Association", 76, 598-606.
[3] Anderson T.W., Hsiao C., [1982], *Formulation and estimation of dynamic models using panel data*, „Journal of Econometrics", 18, 47-82.
[4] Arellano M., Bover O., [1995], *Another look at the instrumental variable estimation of error –components models*, „Journal of Econometrics", Vol. 68, s. 29-52.
[5] Arellano M., Bond S., [1991], *Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations*, „Review of Economic Studies", Vol. 58, s. 277-297.
[6] Baltagi B., [2008], *Econometric Analysis of Panel Data*, 4th edn. Wiley, Chichester.
[7] Barro R.J., Sala-i-Martin X., [1992], *Convergence*, „Journal of Political Economy", Vol. 100, s. 223-251.
[8] Blundell R., Bond S., [1998], *Initial conditions and moment restrictions i dynamic panel data models*, „Journal of Econometrics", Vol. 87(1), s. 115-143.
[9] Blundell R., Bond S., [2000], *GMM estimation with persistent panel data: an application to production functions*, „Econometric Reviews", Vol. 19, s. 321-340.
[10] Bond S., Hoeffler A., Temple J., [2001], *GMM estimation of empirical growth models*, CEPR discussion paper, 3048, Centre for Economic Policy Research, London, United Kingdom.
[11] Bond S., [2002], *Dynamic panel data models: a guide to micro data methods and practice*, Centre for microdata and practice (CEMMAP), working paper CWP09/02.
[12] Bond S., Windmeijer F. [2005], *Reliable inference for GMM estimators? Finite sample properties of alternative test procedures in linear panel data models*, „Econometric Reviews", Vol. 24(1), s. 1-37.

[13] Caselli F., Esquivel G., Lefort F. [1996], *Reopening the convergence debate: anew look at cross-country growth empirics*, „Journal of Economic Growth", Vol. 1, s. 363-389.
[14] Chamberlain G., [1982], *Multivariate Regression Models for Panel Data*, „Journal of Econometrics", Vol. 18, s. 5-46.
[15] Dańska-Borsiak B., [2008], *Wybrane zagadnienia stosowalności Uogólnionej Metody Momentów dla modeli klasycznych i panelowych*, „Przegląd Statystyczny", tom 55(3), s. 47-60.
[16] Hausman J.A., Taylor W.E., [1981], *Panel data and unobservable individual effects*, „Econometrica", Vol. 49, 1377-1398.
[17] Hsiao C., [2003], *Analysis of Panel Data*, 2nd edn. Cambridge, Cambridge University Press.
[18] Islam N., [1995], *Growth empirics: a panel data approach*, „Quarterly Journal of Economics", Vol. 110(4), s. 1127-1170.
[19] Mankiw N., Romer D., Weil D., [1992], *A contribution to the empirics of economic growth*, „Quarterly Journal of Economics", Vol. 107(2), s. 407-437.
[20] Staiger D., Stock J., [1997], *Instrumental variables regression with weak instruments*, „Econometrica", Vol. 65, 557-586.
[21] Windmeijer F., [2005], *A finite sample correction for the variance of linear efficient two-step GMM estimators*, „Journal of Econometrics", Vol. 126, s. 25-51.

Praca wpłynęła do redakcji w marcu 2009 r.

ZASTOSOWANIA PANELOWYCH MODELI DYNAMICZNYCH W BADANIACH MIKROEKONOMICZNYCH I MAKROEKONOMICZNYCH

Streszczenie

Modele ekonometryczne szacowane na podstawie danych panelowych, w których zakłada się, że na kształtowanie się zmiennej objaśnianej wpływają, oprócz zmiennych objaśniających, niemierzalne, stałe w czasie i specyficzne dla danego obiektu czynniki, zwane efektami grupowymi, nazywane są modelami panelowymi (ang. *panel data models*). Stałość w czasie efektów grupowych jest przyczyną komplikacji metodologicznych, pojawiających się przy estymacji modeli dynamicznych. W artykule prezentowana jest zasadnicza idea dwóch najczęściej stosowanych metod estymacji takich modeli, bazujących na Uogólnionej Metodzie Momentów (GMM). Głównym celem artykułu jest przedstawienie przykładów zastosowania panelowych modeli dynamicznych w analizach ekonomicznych w skali mikro i makro. Szczególny nacisk położony został przy tym na wskazanie czynników, różnicujących te dwa przypadki i konsekwencje zastosowania różnych metod estymacji w zależności od rodzaju próby.

Słowa kluczowe: model dynamiczny, dane panelowe, GMM Pierwszych Różnic, Systemowa GMM, metody estymacji, model wzrostu, model produkcji

DYNAMIC PANEL DATA MODELS IN MICROECONOMIC AND MACROECONOMIC RESEARCH

Summary

Econometric models based on panel data, in which the presence of unobservable, constant over time, group-specific effects is assumed are called panel data models. The constancy over time of the group effects causes some methodological complications in the case of dynamic models. In this paper the main ideas of the two methods, which are most often used for estimation of dynamic panel data models are presented. The methods are: first-differenced GMM and system GMM. The main goal of this paper is to present some examples of applications of dynamic panel data models in micro and macroeconomic analyses. Special

interest is in showing the consequences of using different methods according to the type of data – macro or micro.

Key words: dynamic model, panel data, first differenced GMM, system GMM, estimation methods, the growth model, production model.

PRZEGLĄD STATYSTYCZNY
R. LVI – ZESZYT 2 – 2009

JANUSZ WYWIAŁ

TESTING OF THE PREDICTION UNBIASEDNESS ON THE BASIS OF JANUS QUOTIENT

1. INTRODUCTION

The problem of choice the appropriate predictor is being considered. In this paper the analysis is generally focused on the problem of unbiasedness of the predictors. Several tests attempting to verify the unbiasedness of three predictors of the linear trend are proposed. They are based on some modifications of the well known Janus quotient. In some sense the construction of the test statistics is close to those considered [1] which were used to test the linearity of the trend. It is suggested that the proposed tests can be useful in more general problems like testing hypotheses referring to non-linear trends.

Let us consider the following model: $(Y_t) = (Y_1, Y_2, \ldots, Y_n, \ldots, Y_{n+m})$, $E(Y_t) = \mu_t$, $D^2(Y_t) = \sigma^2$, $Cov(Y_t, Y_k) = 0$, for $t \neq k$ and $t = 1, \ldots, n + m$, $k = 1, \ldots, n + m$. Moreover, let $\mathbf{Y}^T = \left[\mathbf{Y}_n^T \ \mathbf{Y}_m^T\right]$.

Let Y_{tp} be a predictor of Y_t, where $t = n + 1, \ldots, n + m$ and let Y_{tp} be a function of the variables $(Y_1 \ldots Y_n \ldots Y_{t-1})$. The vector of prediction errors is denoted by $\mathbf{U}^T = [U_1 \ldots U_m]$ where $U_t = Y_t - Y_{tp}$. The vector of prediction bias is $\delta^T = [\delta_1 \ldots \delta_m]$ where $\delta_t = E(U_t)$. A predictor is unbiased when $\delta^T = \mathbf{0}$. The prediction variance is as follows:

$$D^2(U_t) = D^2(Y_t) + 2Cov(Y_t, Y_{tp}) + D^2(Y_{tp}).$$

The mean square prediction error is as follows:

$$MSE(U_t) = E(U_t)^2 = D^2(U_t) + \delta_t^2.$$

The prediction ex-post variance is defined in the following way:

$$S_U^2 = \frac{1}{m} \sum_{t=n+1}^{n+m} U_t^2 = \frac{1}{m} \mathbf{U}^T \mathbf{U}. \tag{1}$$

Let $\hat{Y}_t$ be an estimator of the t-th value of the trend. The vector of residuals is denoted by $\mathbf{e}^T = [e_1 \ldots e_n]$ where $e_t = Y_t - \hat{Y}_t$. The residual variance is as follows.

$$S_e^2 = \frac{1}{m}\sum_{t=1}^{n} e_t^2 = \frac{1}{m}\mathbf{e}^T\mathbf{e}. \tag{2}$$

2. THE JANUS QUOTIENT

In [2] the following Janus coefficient is proposed.

$$J = \frac{S_U^2}{S_e^2} \tag{3}$$

Let us assume that $\mathbf{Y} \sim N(\mu,\ \sigma^2\ \mathbf{I}_{n+m})$ where $\mu^T = \left[\mu_n^T\ \ \mu_m^T\right]$, $E(\mathbf{Y}_n) = \mu_n$, $E(\mathbf{Y}_m) = \mu_m$,

Let $\mathbf{Z}^T = [\mathbf{e}^T\ \mathbf{U}^T]$ and

$$\mathbf{Z} = \mathbf{CY} \tag{4}$$

where

$$\mathbf{C} = \begin{bmatrix} \mathbf{A} & \mathbf{O}_{n\times m} \\ \mathbf{B}_{m\times n} & \mathbf{B}_{m\times m} \end{bmatrix} \tag{5}$$

Particularly we have

$$\mathbf{e} = \mathbf{AY}_n, \qquad \mathbf{U} = \mathbf{BY}. \tag{6}$$

where $\mathbf{B} = [\mathbf{B}_{m\times n}\ \mathbf{B}_{m\times m}]$.

Hence, $\mathbf{Z} \sim N(\mu_z,\ \Sigma_z)$ where $\mu_z^T = \left[\mu_e^T\ \ \mu_U^T\right]$, $\mu_e^T = \mathbf{A}\mu_n$, $\mu_U^T = \mathbf{B}\mu$,

$$\Sigma_z = \begin{bmatrix} \Sigma_{ee} & \Sigma_{eU} \\ \Sigma_{Ue} & \Sigma_{UU} \end{bmatrix} \tag{7}$$

where

$$\sum_{ee} = \sigma^2\mathbf{AA}^T,\ \sum_{UU} = \sigma^2\mathbf{BB}^T,\ \sum_{eU} = \sigma^2\mathbf{AB}_{m\times n}^T = \sum_{Ue}^T. \tag{8}$$

The Janus coefficient can be rewritten in the following way.

$$J = \frac{n}{m}\frac{\mathbf{Y}^T\mathbf{B}^T\mathbf{BY}}{\mathbf{Y}^T\mathbf{A}^T\mathbf{AY}}. \tag{9}$$

The Janus quotient is a function of two quadratic forms of the normal vector **Y**. The distributions of the quadratic forms are not necessarily chi-square ones and, moreover, their distributions may not be independent. That distribution function of the *J* statistic can be transformed in the following way:

$$F_J(J < x) = P(\mathbf{Y}^T(n\mathbf{B}^T\mathbf{B} - mx\mathbf{A}^T\mathbf{A})\mathbf{Y} < 0) = P(\mathbf{Y}^T\mathbf{G}\mathbf{Y} < 0) = F(0) \tag{10}$$

where

$$\mathbf{G} = n\mathbf{B}^T\mathbf{B} - \begin{bmatrix} xm\mathbf{A}^T\mathbf{A} & \mathbf{O}_{n\times m} \\ \mathbf{O}_{m\times n} & \mathbf{O}_{m\times m} \end{bmatrix}. \tag{11}$$

The probability $F(0)$ can be evaluated on the basis of the Appendix 6.1.

3. THE MODIFIED JANUS QUOTIENTS

In [6] there are some modifications of the Janus quotient proposed. The first of them is as follows:

$$J_* = \frac{n}{m}\frac{Q_U}{Q_e} \tag{12}$$

where

$$Q_U = \mathbf{U}^T\sum\nolimits_{UU}^{-}\mathbf{U}, \tag{13}$$

$$Q_e = \mathbf{e}^T\sum\nolimits_{ee}^{-}\mathbf{e}, \tag{14}$$

$\sum_{UU}^{-}$ and $\sum_{ee}^{-}$ are the pseudo-inverses of the matrices $\sum_{UU}$ and $\sum_{ee}$, respectively. Hence, $\sum_{UU}\sum_{UU}^{-}\sum_{UU} = \sum_{UU}$ and $\sum_{ee}\sum_{ee}^{-}\sum_{ee} = \sum_{ee}$.

On the basis of general results presented in [5] the random variable Q_U has noncentral chi-square distribution $\chi_k(\kappa_U)$ with $k = R(\Sigma_{UU})$ degree of freedom and non-centrality parameter $\kappa_U = \mu_U^T\Sigma_{UU}^{-}\mu_U$. Similarly, the quadratic form Q_e has a noncentral chi-square distribution $\chi_h(\kappa_e)$ with $h = R(\Sigma_{ee})$ degrees of freedom and non-centrality parameter $\kappa_e = \mu_e^T\Sigma_{ee}^{-}\mu_e$.

Let us assume that $\kappa_e = \mathbf{0} = E(\mathbf{e})$, which means that the trend function is well fitted to the data in the period $t = 1, \ldots, n$. In this case, the quadratic form Q_e has central chi-square distribution $\chi_k(0)$. When a predictor gives the unbiased forecasts in the period $t = n + 1, \ldots, n + m$, then $\kappa_U = \mathbf{0} = E(\mathbf{U})$ and Q_U has a central chi-square distribution $\chi_k(0)$. In the case, when at least one forecast is biased in the period $t = n + 1, \ldots, n + m$, the quadratic form has a noncentral distribution. Hence, a significantly large value of the J_* statistic leads to rejecting the hypothesis that $\kappa_U = 0$, which means that the considered predictor is unbiased in the period $t = n + 1, \ldots, n + m$. In order to evaluate the p-value of the test, we have to know the distribution of the J_* test statistic. It is an F distribution with k and h degrees of freedom under the additional assumption that quadratic forms Q_U and Q_e are independent. In a more general case, when they are not independent, the distribution function is approximated on the basis of the Appendix 6.1. and the expression (10), where the **G** matrix should be defined as follows.

$$\mathbf{G} = n\mathbf{B}^T\Sigma_{UU}^{-}\mathbf{B} - \begin{bmatrix} xm\mathbf{A}^T\Sigma_{ee}^{-}\mathbf{A} & \mathbf{O}_{n\times m} \\ \mathbf{O}_{m\times n} & \mathbf{O}_{m\times m} \end{bmatrix}. \tag{15}$$

The lack of the independence of the quadratic forms leads to numerous difficulties in evaluating the *p*-value of the test statistic. In order to omit this problem, we introduce a test statistic which can be treated as the second modification of the Janus quotient. Firstly, let us define the vector of the following conditional prediction errors:

$$\mathbf{U}_{U/e} = \mathbf{U} - \Sigma_{Ue}\Sigma_{ee}^{-}\mathbf{e}. \tag{16}$$

So, $\mathbf{U}_{U/e} \sim N(\mu_{U/e}, \Sigma_{UU/e})$ where

$$\mu_{U/e} = \mu_U - \Sigma_{Ue}\Sigma_{ee}^{-}\mu_e, \tag{17}$$

$$\Sigma_{UU/e} = \Sigma_{UU} - \Sigma_{Ue}\Sigma_{ee}^{-}\Sigma_{eU}. \tag{18}$$

The distribution of the vector $\mathbf{U}_{U/e}$ does not dependent on the vector $\mathbf{e}$. Moreover, if $\mu_e = E(\mathbf{e}) = \mathbf{0}$, then $\mu_{U/e} = E(\mathbf{U}_{U/e}) = E(\mathbf{U}) = \mu_U$. Hence, distribution of the square form

$$Q_{U/e} = \mathbf{U}_{U/e}^T\Sigma_{UU/e}^{-}\mathbf{U}_{U/e}, \tag{19}$$

has the noncentral chi-square distribution $\chi_k(\kappa_{U/e})$ with $k = R(\Sigma_{UU/e})$ degrees of freedom and the non-centrality parameter $\kappa_{U/e} = \mu_{U/e}^T\Sigma_{UU/e}^{-}\mu_{U/e}$ and it is not dependent on the quadratic form Q_e.

Under the assumption that $\mu_e = E(\mathbf{e}) = \mathbf{0}$, the following statistic

$$J_{U/e} = \frac{h}{r}\frac{Q_{U/e}}{Q_e}, \tag{20}$$

has the *F* noncentral distribution with *r* and *h* degrees of freedom where $r = R(\Sigma_{UU/e})$. The $J_{U/e}$ coefficient can be treated as a next modified Janus quotient. When the prediction is unbiased, then $\mu_{U/e} = E(\mathbf{U}_{U/e}) = E(\mathbf{U}) = \mu_U = \mathbf{0}$ and the test statistic $J_{U/e}$ has the central *F* distribution. So, a significantly large value of the test statistic leads to the rejection of the hypothesis that $E(\mathbf{U}) = \mu_U = \mathbf{0}$, which means that the prediction is unbiased in the considered time period $t = n + 1, \ldots, n + m$.

4. TESTING THE UNBIASEDNESS OF SOME PREDICTORS

Let us introduce the following assumptions (apart from those stated in the Introduction): We will consider the linear trend $\mu_t = E(Y_t) = at + b$ for $t = 1, \ldots, n$. It can change in the next period $t = n + 1, \ldots, n + m$, so it is possible that $E(Y_t) \neq at + b$ for at least $t = n + 1, \ldots, n + m$. We are going to consider the three predictors of the $\mathbf{Y}_m$ vector based on an extrapolation of the trend. They will be denoted by $\hat{\mathbf{Y}}_g$, $g = 1, 2, 3$. Let $\mathbf{U}_g = \mathbf{Y}_m - \hat{\mathbf{Y}}_g$ be the vector of prediction errors of the $g - th$ predictor,

$g = 1, 2, 3$. We will test the hypothesis that $H_0 : \mathbf{U}_g = \mathbf{0}$, against the alternative one that $H_0 : \mathbf{U}_g \neq \mathbf{0}$.

The vector of residuals of the well-known mean square error estimator of the linear trend is as follows:

$$\mathbf{e} = \mathbf{M}\mathbf{Y}_n \tag{21}$$

where

$$\mathbf{M} = \mathbf{I}_n - \mathbf{X}_n\left(\mathbf{X}_n^T\mathbf{X}_n\right)^{-1}\mathbf{X}_n^T, \qquad \mathbf{X}_n = \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ .. & .. \\ n & 1 \end{bmatrix}. \tag{22}$$

The first predictor is the following:

$$\hat{\mathbf{Y}}_1 = \mathbf{b}_1\mathbf{Y}_n \tag{23}$$

where

$$\mathbf{b}_1 = \mathbf{X}_m\left(\mathbf{X}_n^T\mathbf{X}_n\right)^{-1}\mathbf{X}_n^T \tag{24}$$

where

$$\mathbf{X}_n = \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ .. & .. \\ n & 1 \end{bmatrix}, \quad \mathbf{X}_m = \begin{bmatrix} n+1 & 1 \\ n+2 & 1 \\ .. & .. \\ n+m & 1 \end{bmatrix}.$$

The vector of the prediction error is $\mathbf{U}_1 = \mathbf{B}_1\mathbf{Y}$ where

$$\mathbf{B}_1 = [-\mathbf{b}_1 \ \mathbf{I}_m] \tag{25}$$

The expression (8) leads to the following well-known ones.

$$\Sigma_{UU} = \sigma^2\left(\mathbf{I}_m + \mathbf{X}_m\left(\mathbf{X}_n^T\mathbf{X}_n\right)^{-1}\mathbf{X}_m^T\right), \ \Sigma_{eU} = \mathbf{O}_{n\times m}. \tag{26}$$

On the basis of the expressions (12)-(14) we have

$$J_{*1} = \frac{n-2}{m}\frac{\mathbf{U}_1^T\left(\mathbf{B}_1\mathbf{B}_1^T\right)^{-1}\mathbf{U}_1}{\mathbf{e}^T\mathbf{M}\mathbf{e}}. \tag{27}$$

The residual vector $\mathbf{e}$ and the vector of the prediction errors $\mathbf{U}_1$ are independent. So, the J_{*1} statistic has the F distribution with m and $h = n - 2$ degrees of freedom when the hypothesis H_0 is true.

The next vector predictor is as follows:

$$\hat{\mathbf{Y}}_2 = \mathbf{b}_2\mathbf{Y} \tag{28}$$

where

$$\mathbf{b}_2 = \begin{bmatrix} \mathbf{b}_{2,n} & \mathbf{O}_{1\times m} \\ \mathbf{b}_{2,n+1} & \mathbf{O}_{1\times(m-1)} \\ \cdots & \cdots \\ \mathbf{b}_{2,n+m-2} & \mathbf{O}_{1\times 2} \\ \mathbf{b}_{2,n+m-1} & 0 \end{bmatrix}, \tag{29}$$

$$\mathbf{b}_{2,t} = \mathbf{x}_{t+1}\left(\mathrm{X}_t^T\mathbf{X}_t\right)^{-1}\mathbf{X}_t^T,\ t = n+1,\ \ldots,\ n+m-1$$
$$\mathbf{x}_{t+1} = \left[t+1\ \ 1\right],\ \mathbf{X}_t = \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ .. & .. \\ t & 1 \end{bmatrix}. \tag{30}$$

The vector of the prediction error is $\mathbf{U}_2 = \mathbf{B}_2\mathbf{Y}$ where

$$\mathbf{B}_2 = \mathbf{g} - \mathbf{b}_2,\ \mathbf{g} = \left[\mathbf{O}_{m\times n}\ \ \mathrm{I}_m\right] \tag{31}$$

The expression (8) leads to the following ones:

$$\Sigma_{UU} = \sigma^2\mathbf{B}_2^T\mathbf{B}_2. \tag{32}$$

$$\Sigma_{eU} = \sigma^2\left[\mathbf{M}\ \ \mathbf{O}_{n\times m}\right]\mathbf{B}_2^T = \mathbf{O}_{n\times m}. \tag{33}$$

On the basis of the expressions (12)-(14) we have

$$J_{*2} = \frac{n-2}{m}\frac{\mathbf{U}_2^T\left(\mathbf{B}_2\mathbf{B}_2^T\right)^{-1}\mathbf{U}_2}{\mathbf{e}^T\mathbf{M}\mathbf{e}}. \tag{34}$$

The residual vector $\mathbf{e}$ and the vector of the prediction errors $\mathbf{U}_1$ are independent. So, the J_{*2} statistic has the F distribution with m and $h = n - 2$ degrees of freedom when the hypothesis H_0 is true.

The third vector predictor is as follows:

$$\hat{\mathbf{Y}}_3 = \mathbf{b}_3\mathbf{Y} \tag{35}$$

where

$$\mathbf{b}_3 = \begin{bmatrix} \mathbf{b}_{3,n} & \mathbf{O}_{1\times m} & \\ 0 & \mathbf{b}_{3,n+1} & \mathbf{O}_{1\times(m-1)} \\ \mathbf{O}_{1\times 2} & \mathbf{b}_{3,n+2} & \mathbf{O}_{1\times(m-2)} \\ \cdots & \cdots & \cdots \\ \mathbf{O}_{1\times(m-2)} & \mathbf{b}_{3,n+m-2} & \mathbf{O}_{1\times 2} \\ \mathbf{O}_{1\times(m-1)} & \mathbf{b}_{3,n+m-1} & 0 \end{bmatrix}, \quad \mathbf{X}_m = \begin{bmatrix} n+1 & 1 \\ n+2 & 1 \\ .. & .. \\ n+m & 1 \end{bmatrix},$$

The $\mathbf{b}_{2,t}$ vector is defined by the right hand of the equation(30), but in this case

$$\mathbf{x}_{t+1} = [t+1 \;\; 1], \; \mathbf{X}_t = \begin{bmatrix} t-n+1 & 1 \\ t-n+2 & 1 \\ \cdots & \cdots \\ t & 1 \end{bmatrix}.$$

The vector of the prediction error is $\mathbf{U}_3 = \mathbf{B}_3\mathbf{Y}$ where

$$\mathbf{B}_3 = \mathbf{g} - \mathbf{b}_3, \; \mathbf{g} = \left[\mathbf{O}_{m\times n} \;\; \mathbf{I}_m\right]. \tag{36}$$

The expression (8) leads to the following ones.

$$\Sigma_{UU} = \sigma^2\mathbf{B}_3^T\mathbf{B}_3, \; \Sigma_{eU} = \sigma^2\left[\mathbf{M} \;\; \mathbf{O}_{n\times m}\right]\mathbf{B}_3^T. \tag{37}$$

On the basis of the expressions (12)-(14) we have

$$J_{*3} = \frac{n-2}{m}\frac{\mathbf{U}_3^T\left(\mathbf{B}_3\mathbf{B}_3^T\right)^{-1}\mathbf{U}_3}{\mathbf{e}^T\mathbf{M}\mathbf{e}} = \frac{n-2}{m}\frac{\mathbf{Y}^T\mathbf{B}_3^T\left(\mathbf{B}_3\mathbf{B}_3^T\right)^{-1}\mathbf{B}_3\mathbf{Y}}{\mathbf{Y}_n^T\mathbf{M}\mathbf{Y}_n}. \tag{38}$$

The J_{*3} statistic is proportionate to the ratio of two dependent quadratic forms of the normal vector. The value of its distribution can be evaluated on the basis of the appendix and the expression (10) where

$$\mathbf{G} = n\mathbf{B}_3^T\left(\mathbf{B}_3\mathbf{B}_3^T\right)^{-1}\mathbf{B}_3 - \begin{bmatrix} xm\mathbf{M} & \mathbf{O}_{n\times m} \\ \mathbf{O}_{m\times n} & \mathbf{O}_{m\times m} \end{bmatrix}. \tag{39}$$

According to the expressions (16)-(20) we define the following vector of the conditional prediction errors and its parameters.

$$\mathbf{U}_{U_3/e} = \mathbf{U}_3 - \mathbf{B}_3\begin{bmatrix} \mathbf{M} \\ \mathbf{O}_{m\times n} \end{bmatrix}\mathbf{e}. \tag{40}$$

When $E(\mathbf{e}) = \mathbf{0}$ then $\mathbf{U}_{U_3/e} \sim N(\mu_{U_3}, \Sigma_{U_3U_3/e})$ where $\mu_{U_3} = E(\mathbf{U}_3)$ and

$$\Sigma_{U_3U_3/e} = \sigma^2\mathbf{B}_3^T\mathbf{B}_3 - \sigma^2\begin{bmatrix} \mathbf{M} & \mathbf{O}_{n\times m} \\ \mathbf{O}_{m\times n} & \mathbf{O}_{m\times m} \end{bmatrix}. \tag{41}$$

The equation (20) leads to the following modified Janus coefficient.

$$J_{U_3/e} = \frac{h}{r}\frac{\mathbf{U}_{U_3/e}^T \Sigma_{U_3U_3/e}^{-} \mathbf{U}_{U_3/e}}{\mathbf{e}^T \mathbf{M} \mathbf{e}}. \tag{42}$$

The J_{U_3}/e statistic has the F distribution with $r = R(\Sigma_{U_3U_3/e}$ and $h = n - 2$ degrees of freedom when the hypothesis H_0 is true.

Example 1. Let us consider the following time series $\{y_t\} = \{10, 12, 13, 12, 16, 15, 17, 18, 20, 19, 21, 23, 26, 29, 29, 31, 34, 36, 39, 40\}$. On the basis of the first $n = 10$ observations, the following trend function was determined by means of the least square error: $\hat{y}_t = 1.07t + 9.33$. Similarly, in the next period $t = 11, \ldots, 20$, the trend was: $\hat{y}_t = 2.12t - 2.08$.

Values of the modified Janus quotients (with the p-values of the test in the brackets) were evaluated be means of the computer programmes are as follows: $J_{*1} = 256.70$ ($p = 0.000$), $J_{*2} = 164.73$ ($p = 0.000$), $J_{*3} = 4.39$ ($p = 0.103$) and $J_{U_3/e} = 5.21$ ($p = 0.014$). Each of the statistics J_{*1}, J_{*2} and $J_{U_3/e}$ has F distribution with 10 and 8 degrees of freedom.

The analysis of the p-values of the tests based on the statistics J_{*1} and J_{*2} leads to rejection of the hypotheses that the predictors $\hat{\mathbf{Y}}_1$ and $\hat{\mathbf{Y}}_2$ are unbiased. The hypothesis that the $\hat{\mathbf{Y}}_3$ predictor is unbiased is not rejected by any tests based on the statistics J_{*3} and $J_{U_3/e}$ under the significance level $\alpha = 0.01$.

5. CONCLUSIONS

Although the proposed tests are constructed under a very strong assumption connected with defining the model of time series, they can support the choice of the most accurate predictor. The analysis of the example 1 leads to conclusion that it is especially possible to assess how useful in practice the adaptive predictors are. It is easy to show that the tests can be useful in the cases of other trend functions as well as in the case of time series with seasonal factors. More complicated models with more explanatory variables can be considered, too. Moreover, it seems that testing a hypothesis referring to the parameters of exponential smoothing methods is possible, too.

The analysis of the properties of the tests should be developed in the direction of examining their power as well as their robustness of their assumptions. The author would like to consider those problems in his next papers.

6. APPENDIX

6.1. APPROXIMATION OF THE DISTRIBUTION FUNCTION OF THE QUADRATIC FORMS

In [4] the following procedure for the approximation of the distribution of the form $Q = \mathbf{Y}^T \mathbf{C} \mathbf{Y} = \sum_{i=1}^{r} d_i Z_i^2$ where $Z_i \sim N(0, 1)$, Z_i, Z_j are independent for $i \neq j$, $i = 1, \ldots, r$, $j = 1, \ldots, r$ is proposed. He showed that

$$Q \approx cU_k + v \tag{43}$$

where $Z_k \sim \chi_k$ and

$$c = \frac{\theta_3}{\theta_2},\ \upsilon = -\frac{\theta_2^2}{\theta_3} + \theta_1,\ k = \frac{\theta_2^3}{\theta_3^2},\ \theta_h = \sum_{i=1}^{k} = tr\mathbf{C}^h,\ h = 1,2,\ldots \tag{44}$$

Under this notation

$$P(Q < q) \approx P(U_k < u) \tag{45}$$

where

$$u = \frac{q - \theta_1}{\sqrt{\theta_2}}\sqrt{k} + k. \tag{46}$$

Let us note that in [3] there are other approximations of the distribution function of the quadratic forms considered.

6.2. PROOF OF INDEPENDENCE OF THE VECTORS **e** AND $\mathbf{U}_2$

Evaluation of the expression (33) is as follows. Let e_i be the $i-th$ residual ($i = 1, \ldots, n$) end U_{t+1} be the $(t + 1) - th$ prediction error ($t = n + 1, \ldots, n + m$).

$$\begin{gathered}
Cov(e_i, U_{t+1}) = E\left(Y_i - \mathbf{x}_i\left(\mathbf{X}_n\mathbf{X}_n^T\right)^{-1}\mathbf{X}_n^T\mathbf{Y}_n\right)\left(Y_{t+1} - \mathbf{x}_{t+1}\left(\mathbf{X}_n\mathbf{X}_n^T\right)^{-1}\mathbf{X}_t^T\mathbf{Y}_t\right),\\
Cov(e_i, U_{t+1}) = -\mathbf{x}_{t+1}\left(\mathbf{X}_n\mathbf{X}_n^T\right)^{-1}\mathbf{X}_t^T E(\mathbf{Y}_t y_i) +\\
+\mathbf{x}_i\left(\mathbf{X}_n\mathbf{X}_n^T\right)^{-1}\mathbf{X}_n^T E\left(\mathbf{Y}_n\mathbf{Y}_t^T\right)\mathbf{X}_t\left(\mathbf{X}_n\mathbf{X}_n^T\right)^{-1}\mathbf{x}_{t+1}^T,\\
Cov(e_i, U_{t+1}) = -\mathbf{x}_{t+1}\left(\mathbf{X}_n\mathbf{X}_n^T\right)^{-1}\mathbf{x}_i^T +\\
+\mathbf{x}_i\left(\mathbf{X}_n\mathbf{X}_n^T\right)^{-1}\mathbf{X}_n^T\left[\mathbf{I}_n\ \ \mathbf{O}_{n\times(t-n)}\right]\mathbf{X}_t\left(\mathbf{X}_n\mathbf{X}_n^T\right)^{-1}\mathbf{x}_{t+1}^T,\\
Cov(e_i, U_{t+1}) = -\mathbf{x}_{t+1}\left(\mathbf{X}_n\mathbf{X}_n^T\right)^{-1}\mathbf{x}_i^T +\\
+\mathbf{x}_i\left(\mathbf{X}_n\mathbf{X}_n^T\right)^{-1}\mathbf{X}_n^T\mathbf{X}_n\left(\mathbf{X}_n\mathbf{X}_n^T\right)^{-1}\mathbf{x}_{t+1}^T,\\
Cov(e_i, U_{t+1}) = -\mathbf{x}_{t+1}\left(\mathbf{X}_n\mathbf{X}_n^T\right)^{-1}\mathbf{x}_i^T + \mathbf{x}_i\left(\mathbf{X}_n\mathbf{X}_n^T\right)^{-1}\mathbf{x}_{t+1}^T = 0.
\end{gathered}$$

This leads straightforwardly to the expression (33).

Akademia Ekonomiczna w Katowicach

REFERENCES

[1] Azzalini A., Bowman A., [1993], *On nonparametric regression for checking linear relationships*, „The Journal of the Royal Statistical Society", B55(2), s. 549-557.

[2] Gadd A., Wold H., [1964], *The Janus quotient; a measure for the accuracy of prediction*, In Econometric Model Building, Amsterdam.

[3] Mathai A.M., Provost S.B., [1992], *Quadratic Forms in Random Variables (Theory and Applications)*, Marcel Decker, Inc., New York-Basel-Hong Kong.
[4] Pearson E.S., [1959], *Note on an approximation to the distribution of noncentral* χ^2, „Biometrika", 46, pp. 346-346.
[5] Rao C.R., Mitra S.K., [1971], *Generalized inverse of matrices and its applications*, John Wiley and Sons, New York-London-Sydney-Toronto.
[6] Wywiał J., [1995], *Weryfikacja hipotez o błędach predykcji adaptacyjnej*, Ossolineum, Wrocław-Warszawa-Kraków.

Praca wpłynęła do redakcji we wrześniu 2008 r.

O TESTOWANIU NIEOBCIĄŻONOŚCI PREDYKCJI NA PODSTAWIE WSPÓŁCZYNNIKA JANUSOWEGO

Streszczenie

W pracy jest rozważany problem wyboru odpowiedniego predykatora. Przyjęto, że postulowanym kryterium tego wyboru jest jego nieobciążoność. W pracy są proponowane testy na nieobciążoność predykcji trzech predyktorów trendu liniowego. Punktem wyjścia konstrukcji sprawdzianów testów jest znany współczynnik Janusowy używany do oceny dokładności ciągów wyznaczanych prognoz, który jest ilorazem wariancji predykcji *ex-post* i wariancji resztowej. Z formalnego punktu widzenia każdy z rozważanych sprawdzianów testów jest ilorazem dwóch form kwadratowych wektora zmiennych o rozkładzie normalnym. Te formy kwadratowe mogą być zależne. Dlatego w celu wyznaczenia rozkładu prawdopodobieństwa sprawdzianu testu jest adaptowana jedna ze znanych metod pozwalających na przybliżone wyliczanie wartości jego dystrybuanty. Rozważania zilustrowano przykładem. Otrzymane w pracy wyniki można uogólnić na przypadek predykcji na podstawie modelu regresji.

Słowa kluczowe: Test statystyczny, błąd predykcji, nieobciążoność, współczynnik Janusowy, forma kwadratowa wektora o rozkładzie normalnym, aproksymacja funkcji dystrybuanty, wariancja resztowa, wariancja predykcji.

TESTING OF THE PREDICTION UNBIASEDNESS ON THE BASIS OF JANUS QUOTIENT

Summary

The problem of choosing the appropriate predictor is being considered. Generally, in this paper the analysis is focused on the problem of unbiasedness of the predictors. Several tests attempting to verify the unbiasedness of three predictors of the linear trend are proposed. They are based on some modifications of the well-known Janus quotient being a ratio of the variance of prediction errors and the residual variance. In general each of the considered test statistic can be represented as the ratio of two quadratic forms of normal vectors. These two quadratic forms can be dependent, so its distribution function has to be approximated. An example of testing hypothesis on unbiasedness is presented. The obtained results can be generalized in the case of prediction on the basis of regression models.

Keyword: Test statistic, prediction error, unbiasednees, Janus quotient, quadratic form of normally distributed vector, approximation of distribution function, residual variance, prediction variance.

PRZEGLĄD STATYSTYCZNY
R. LVI – ZESZYT 2 – 2009

JAN PURCZYŃSKI

METODY PROGNOZOWANIA Z WYKORZYSTANIEM TRENDU POTĘGOWEGO

1. WSTĘP

Praca dotyczy estymacji parametrów nieliniowych modeli trendu i stanowi, w pewnym sensie, kontynuację zagadnień rozpatrzonych w pracy [6], gdzie omówiono metody estymacji parametrów trendu wykładniczego dla celów predykcji. Natomiast niniejsza praca jest poświęcona metodom estymacji parametrów trendu potęgowego na potrzeby procesu prognostycznego.

Celem badań zaprezentowanych w artykule jest wyznaczenie postaci wzorów przybliżonych umożliwiających estymację parametrów trendu potęgowego posiadających zbliżone właściwości, do oszacowania parametrów uzyskanych MNK w odniesieniu do modelu I (wzór (1)). W celu określenia przydatności tych wzorów wykonano symulacje komputerowe. Ponadto zostanie wyprowadzony wzór określający błąd *ex ante* prognozy uzyskanej MNK dla trendu potęgowego.

Ze względu na charakter składnika losowego, rozpatrzone zostaną trzy modele:

– Model addytywny (Model I)

$$y_t = A \cdot t^B + \varepsilon a_t \ \ gdzie \ t = 1, 2, \ldots, n \tag{1}$$

– Model multiplikatywny (Model II)

$$y_t = A \cdot t^B \cdot e^{\varepsilon m_t} \tag{2}$$

– Model mieszany (Model III)

$$y_t = A \cdot t^B \cdot e^{\varepsilon m_{tt}} + \varepsilon a_t \tag{3}$$

gdzie:

εa_t – składnik losowy o rozkładzie $N(0, \sigma a)$

εm_t – składnik losowy o rozkładzie $N(0, \sigma m)$

y_t – dane empiryczne.

W przypadku modelu II zalecana jest metoda transformacji liniowej [5, 7], polegająca na logarytmowaniu równania (2):

$$\ln(y_t) = \ln(A) + B \cdot \ln(t) + \varepsilon m_t. \tag{4}$$

Stosując MNK do równania (4), otrzymuje się następujące oszacowania parametrów B i A:

$$BL = \frac{\sum_{t=1}^{n} L_t(t' - \bar{t}')}{\sum_{t=1}^{n}(t' - \bar{t}')^2}; \; AL = \exp(\bar{L} - BL \cdot \bar{t}') \tag{5}$$

gdzie: $L_t = \ln(y_t)$; $\bar{L} = \frac{1}{n}\sum_{t=1}^{n} L_t$; $t' = \ln(t)$; $\bar{t}' = \frac{1}{n}\sum_{t=1}^{n} t'$;

$AL.$; BL – oszacowania parametrów A i B.

Dla modelu addytywnego (model I) zalecana jest MNK w odniesieniu do równania (1), w wyniku czego, uzyskuje się układ równań:

$$\sum_{t=1}^{n} y_t \cdot t^{BS} \ln(t) \cdot \sum_{t=1}^{n} t^{2BS} = \sum_{t=1}^{n} y_t \cdot t^{2BS} \cdot \sum_{t=1}^{n} t^{2BS} \ln(t) \tag{6a}$$

$$AS = \frac{\sum_{t=1}^{n} y_t \cdot t^{BS}}{\sum_{t=1}^{n} t^{2BS}}, \tag{6b}$$

gdzie: AS i BS oznaczają oszacowania parametrów A i B uzyskane MNK.

W celu wyznaczenia oszacowania BS należy rozwiązać równanie (6a) metodą numeryczną [2]. Następnie, z równania (6b) oblicza się oszacowanie AS.

Omawiając zagadnienie szacowania parametrów nieliniowych funkcji regresji, do których zalicza się funkcja potęgowa, autorzy podręczników i zbiorów zadań zwracają uwagę na ograniczenia stosowania transformacji logarytmicznej wynikające z założenia postaci modelu multiplikatywnego (wzór (2)) [4]. W pracy [3] autorzy dodatkowo podkreślają dwa inne problemy związane ze stosowaniem transformacji: brak przeniesienia własności estymatora $\ln(A)$ (wzór (4)) na estymator parametru A oraz nierównoważność kryteriów MNK stosowanej do funkcji krzywoliniowej (wzór (2)) i funkcji zlogarytmowanej (wzór (4)). W związku z czym, w pracy [3] proponuje się rozwiązanie układu równań (6) za pomocą arkusza kalkulacyjnego np. Microsoft Excel.

Jak już wspomniano powyżej, w niniejszej pracy zostaną zaproponowane wzory przybliżone umożliwiające estymację parametrów trendu potęgowego.

2. UPROSZCZONA METODA SZACOWANIA PARAMETRÓW TRENDU POTĘGOWEGO

Załóżmy, że dla modelu addytywnego (1) została wyznaczona postać teoretyczna $\hat{y}_t$:

$$\hat{y}_t = \hat{A} \cdot t^{\hat{B}} \tag{7}$$

gdzie: $\hat{A}$ i $\hat{B}$ oznaczają oszacowania parametrów trendu.

Punktem wyjścia jest różnica symetryczna

$$\hat{y}_{t+1} - \hat{y}_{t-1}. \tag{8}$$

Ze wzoru (7) i (8), otrzymuje się:

$$\hat{y}_{t+1} - \hat{y}_{t-1} = \hat{A} \cdot t^{\hat{B}} \cdot \left[\left(1 + \frac{1}{t}\right)^{\hat{B}} - \left(1 - \frac{1}{t}\right)^{\hat{B}}\right]. \tag{9}$$

Stosując rozwinięcie w szereg potęgowy [2]:

$$(1 \pm x)^m = 1 \pm mx + \frac{m(m-1)}{2!}x^2 \pm \frac{m(m-1)(m-2)}{3!}x^3 + \ldots \tag{10}$$

i uwzględniając trzy pierwsze wyrazy szeregu (10) oraz zależność (7), wzór (9) przyjmuje postać:

$$\hat{y}_{t+1} - \hat{y}_{t-1} = 2\hat{B} \cdot \frac{\hat{y}_t}{t}. \tag{11}$$

W celu zmniejszenia wpływu składnika losowego, wykonuje się obustronne sumowanie we wzorze (11), uzyskując:

$$\hat{B}_1 = \frac{\sum_{t=2}^{n-1}(y_{t+1} - y_{t-1})}{2\sum_{t=2}^{n-1}\frac{y_t}{t}}. \tag{12}$$

Mnożąc równanie (11) przez t, oraz sumując obustronnie, uzyskuje się:

$$\hat{B}_2 = \frac{\sum_{t=2}^{n-1}(y_{t+1} - y_{t-1}) \cdot t}{2\sum_{t=2}^{n-1} y_t}. \tag{13}$$

Poprawnym sposobem estymacji parametru B jest metoda najmniejszych kwadratów, która zastosowana do wzoru (11), prowadzi do zależności:

$$\hat{B}_3 = \frac{\sum_{t=2}^{n-1}(y_{t+1} - y_{t-1}) \cdot \frac{y_t}{t}}{2\sum_{t=2}^{n-1}\left(\frac{y_t}{t}\right)^2}. \tag{14}$$

Oszacowanie parametru A uzyskuje się ze wzoru (6b):

$$\hat{A}_j = \frac{\sum_t y_t \cdot t^{B_j}}{\sum_t t^{2B_j}}; \; j = 1,2,3, \tag{15}$$

gdzie: $\hat{B}_j$ określają wzory (12), (13), (14).

3. WYNIKI SYMULACJI KOMPUTEROWYCH

W celu oceny przydatności proponowanych wzorów przybliżonych wykonano szereg symulacji komputerowych dla różnych wartości parametrów A i B oraz liczby obserwacji n. Poniżej zostaną zamieszczone wyniki dla $A = 20$, $n = 7$ oraz $B = 0{,}5$ i $B = 1{,}5$ oraz dla okresu prognozowania $T = 8,9,10$. Przy określonym poziomie składnika losowego wykonywano $M = 5000$ symulacji z użyciem generatora liczb losowych o rozkładzie normalnym, obliczając zgodnie ze wzorami (1)-(3) wartości obserwacji $y_{m,t}$. Dla każdej symulacji wyznaczano wartość prognozy $YP_{m,T}$:

$$YP_{m,T} = \hat{A}_m \cdot T^{\hat{B}_m} \tag{16}$$

gdzie:

$m = 1, 2 \ldots, M$ – numer kolejnej symulacji,
$\hat{A}_m$ i $\hat{B}_m$ – oszacowania parametrów trendu dla poszczególnych symulacji,
T – okres prognozy.

Analogicznie określano wartość realizacji zmiennej prognozowanej $YR_{m,T}$, podstawiając do wzorów (1)-(3) $t = T$. Do oceny jakości prognozy zastosowano błąd względny prognozy *ex post* b_T::

$$b_T = \frac{\sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(YP_{m,T} - YR_{m,T}\right)^2}}{\bar{YR}_T} \cdot 100 \tag{17}$$

gdzie: $\bar{YR}_T = \frac{1}{M}\sum_{m=1}^{M} YR_{m,T}$.

Zakładając zgodność wariancji błędu spowodowanego obecnością składnika losowego dla modelu I i modelu II, w pracy [5] podano zależność pomiędzy wartością odchylenia standardowego σa i σm. Dla małych wartości σm zachodzi wzór przybliżony:

$$\sigma m = \frac{\sigma a}{\sqrt{E(yd^2)}}, \tag{18}$$

gdzie: $E(yd^2) = \frac{A^2}{n}\sum_{t=1}^{n} t^{2B}$.

Wykonując symulacje, uruchamiano generator liczb losowych o rozkładzie $N(0, \sigma a)$, a następnie dla modelu I, ze wzoru (1) obliczano wartości y_t. Przy realizacji modelu II określano równoważną wartość odchylenia standardowego σm zgodnie ze wzorem (18). Model mieszany (III) był realizowany zgodnie ze wzorem:

$$y_t = A \cdot t^B \cdot e^{\frac{\varepsilon m_t}{\sqrt{2}}} + \frac{\varepsilon a_t}{\sqrt{2}}. \tag{19}$$

W celu rozwiązania równania (6a), stosowano metodę bisekcji (połowienia), która w odróżnieniu od metody siecznych i metody stycznych, daje gwarancję zbieżności procesu iteracyjnego [2].

Przy określonym poziomie składnika losowego (σa), wartość błędu prognozy *ex post* zależy od wartości parametrów trendu A i B. W celu ograniczenia wpływu wartości parametrów funkcji potęgowej dokonano normalizacji, wprowadzając względną wartość poziomu składnika losowego σv:

$$\sigma v = \frac{\sigma a}{\bar{y}}; \text{ gdzie: } \bar{y} = \frac{1}{n}\sum_{t=1}^{n} y_t. \tag{20}$$

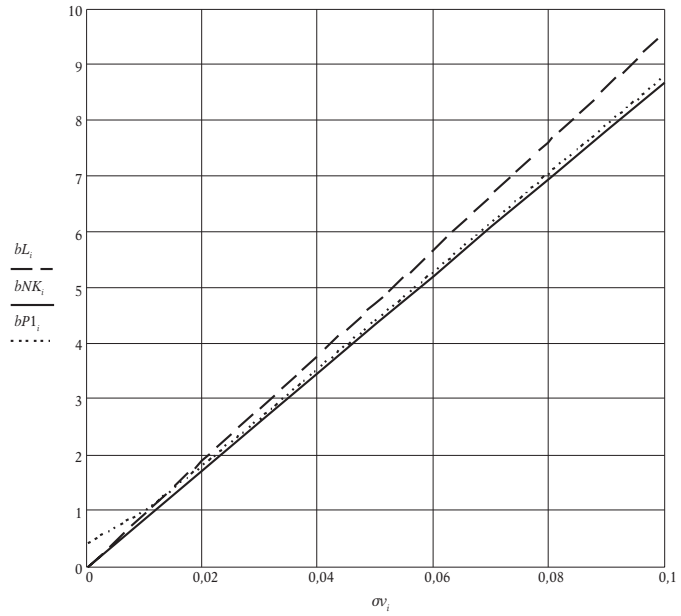
W wyniku normalizacji, dla wszystkich symulacji komputerowych, parametr σv należał do tego samego przedziału zmienności [0 – 0,1].

W pierwszej kolejności wykonano symulacje dla addytywnego modelu składnika losowego – wzór (1): $A = 20$, $B = 0{,}5$, $n = 7$, okres prognozy $T = 8$. Spośród trzech metod przybliżonych najmniejszym błędem *ex post* charakteryzuje się metoda I (wzory (12), (15), (17)). Metoda przybliżona III (wzory (14), (15), (17)) prowadzi do błędu *ex post* stanowiącego 1,015 wartości błędu metody I. Najgorzej wypada metoda II (wzory (13), (15), (17)), dla której błąd *ex post* wynosi 1,043 wartości błędu metody I.

Wszystkie rysunki zamieszczone w pracy przedstawiają wartości błędu względnego *ex post* (wyrażonego w procentach) w funkcji poziomu składnika losowego σv.

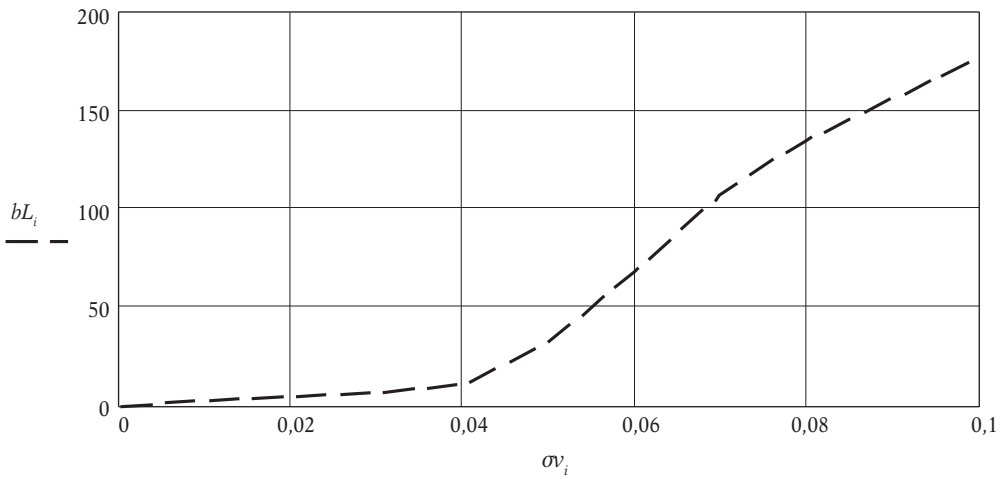
Na rysunku 1 porównano wartości błędu *ex post* uzyskane dla następujących metod: *bL* – metoda transformacji logarytmicznej, *bNK* – metoda najmniejszych kwadratów, *bP*1 – metoda przybliżona I. Do najmniejszych wartości błędu prognozy *ex post* prowadzi MNK, natomiast największym błędem obarczone są wyniki metody transformacji logarytmicznej.. Zbliżone wartości błędu wykazują prognozy uzyskane metodą najmniejszych kwadratów (*bNK*) oraz metodę przybliżoną (*bP*1).

W celu oceny przydatności metod przybliżonych, wykonano symulacje dla addytywnego modelu składnika losowego – parametr $B = 1{,}5$. Najmniejszym błędem prognozy *ex post* wyróżnia się metoda najmniejszych kwadratów, przy czym błąd prognozy rośnie liniowo w funkcji poziomu składnika losowego σv i dla $\sigma v = 0{,}1$ osiąga wartość 6,21%. Stosunek wartości błędów *ex post* poszczególnych metod przybliżonych do wartości błędu *ex post* MNK wynosi: 1,051 dla metody I, 1,012 dla metody II, 1,017 dla metody III.



Rysunek 1. Wartości błędu względnego *ex post* (w %) dla addytywnego modelu składnika losowego w funkcji poziomu składnika losowego σv_i; parametr $B = 0{,}5$, okres prognozy $T = 8$. Linią przerywaną bL_i oznaczono wyniki metody transformacji logarytmicznej, linia ciągła bNK_i odpowiada metodzie MNK, linia kropkowana $bP1_i$ odpowiada metodzie przybliżonej I.

Źródło: opracowanie własne.



Rysunek 2. Wartości błędu względnego *ex post* (w %) dla addytywnego modelu składnika losowego w funkcji poziomu składnika losowego σv_i – parametr $B = 1{,}5$. Wyniki odnoszą się do metody transformacji logarytmicznej.

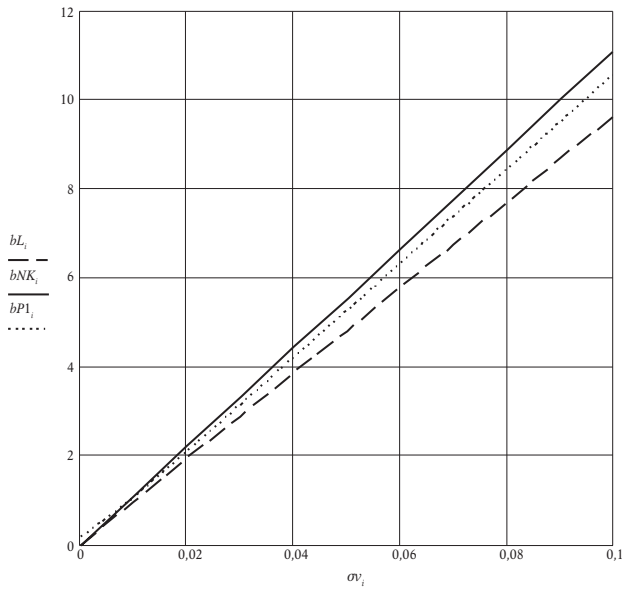
Źródło: opracowanie własne.

Na rysunku 2 przedstawiono wartości błędu względnego *ex post* bL_i odnoszące się do metody transformacji logarytmicznej, uzyskane dla addytywnego modelu składnika losowego – parametr $B = 1{,}5$. Z rysunku wynika odcinkami liniowa zależność wartości błędów od względnego poziomu składnika losowego σv, przy czym, powyżej wartości $\sigma v = 0{,}04$ następuje gwałtowny wzrost współczynnika kierunkowego prostej $bL\,(\sigma v)$. Podczas gdy dla $\sigma v = 0{,}04$ błąd wynosi $bL = 10{,}6\%$, to dla $\sigma v = 0{,}1$ osiąga wartość $bL = 177\%$. Przyczyny należy upatrywać w wysokim poziomie składnika losowego, który zgodnie ze wzorem (20), dla $\sigma v = 0{,}1$ i $\bar{y} = 175{,}49$ wynosi $\sigma a = 17{,}55$. Wykonując logarytmowanie wzoru (1) popełnia się błąd szczególnie duży dla małych wartości t, gdy wartości składnika losowego są porównywalne z wartością funkcji potęgowej.

Wykonując eksperyment numeryczny dla multiplikatywnego modelu składnika losowego z wartością parametru $B = 0{,}5$ stwierdzono, że najmniejszy błąd *ex post* zapewnia metoda transformacji logarytmicznej, który dla $\sigma v = 0{,}1$ osiąga wartość 11,23%. Dla pozostałych metod uzyskano: $\frac{bNK}{bL} = 1{,}020$, $\frac{bP1}{bL} = 1{,}035$, $\frac{bP2}{bL} = 1{,}106$, $\frac{bP3}{bL} = 1{,}019$.

Na uwagę zasługują zbliżone wartości błędów bL, BNK i bP3.

Symulacje komputerowe wykonane dla multiplikatywnego modelu składnika losowego z wartością parametru $B = 1{,}5$ wykazały, że spośród metod przybliżonych, najmniejszy błąd *ex post* zapewnia metoda przybliżona I, który dla $\sigma v = 0{,}1$ przyjmuje wartość 10,61%. Dla pozostałych metod uzyskano: $\frac{bNK}{bP1} = 1{,}048$, $\frac{bP2}{bP1} = 1{,}058$, $\frac{bP3}{bP1} = 1{,}027$.



Rysunek 3. Wartości błędu względnego *ex post* (w %) dla multiplikatywnego modelu składnika losowego w funkcji poziomu składnika losowego σv_i – parametr $B = 1{,}5$. Linią przerywaną bL_i oznaczono wyniki metody transformacji logarytmicznej. Linia ciągła odpowiada MNK (bNK_i). Błąd metody przybliżonej I ($bP1_i$) zaznaczono linią kropkowaną.

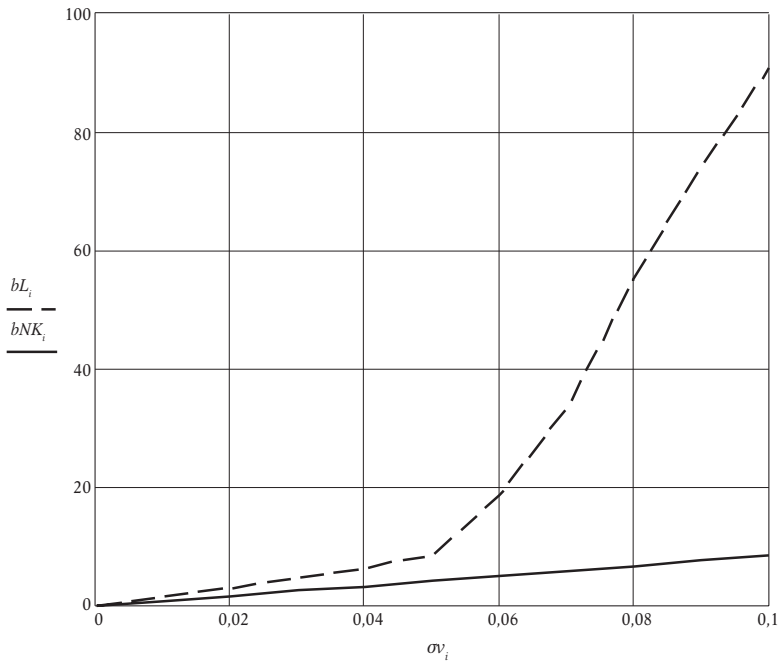
Źródło: opracowanie własne.

Na rysunku 3 porównano wartości błędu *ex post* uzyskane dla multiplikatywnego modelu składnika losowego (wartość parametru $B = 1{,}5$) następujących metod: *bL* – metoda transformacji logarytmicznej, *bNK* – metoda najmniejszych kwadratów, *bP*1 – metoda przybliżona I. Do najmniejszych wartości błędu prognozy *ex post* prowadzi metoda transformacji logarytmicznej. Metoda przybliżona I charakteryzuje się mniejszym błędem niż MNK.

Eksperyment numeryczny wykonany dla mieszanego modelu składnika losowego (wzór (3)) z wartością parametru $B = 0{,}5$ wykazał, że najmniejszy błąd *ex post* zapewnia MNK, który dla $\sigma v = 0{,}1$ przyjmuje wartość 8,51%. Dla pozostałych metod uzyskano: $\frac{bP1}{bNK} = 1{,}018$, $\frac{bP2}{bNK} = 1{,}097$, $\frac{bP3}{bNK} = 1{,}009$, $\frac{bL}{bNK} = 1{,}011$.

Z powyższych danych wynika, że aż cztery metody prowadzą do zbliżonych wartości błędu względnego *ex post*: MNK, metoda transformacji logarytmicznej oraz metody przybliżone I i III.

Jako ostatni przebadano model mieszany z wartością parametru $B = 1{,}5$ stwierdzając, że najmniejszy błąd *ex post* zapewnia metoda przybliżona I, który dla $\sigma v = 0{,}1$ osiąga wartość 8,15%. Dla pozostałych metod uzyskano: $\frac{bNK}{bP1} = 1{,}021$, $\frac{bP2}{bP1} = 1{,}026$, $\frac{bP3}{bP1} = 1{,}007$.



Rysunek 4. Wartości błędu względnego *ex post* (w %) dla modelu mieszanego (wzór (3)) w funkcji poziomu składnika losowego σv_i – parametr $B = 1{,}5$. Linią przerywaną bL_i oznaczono wyniki metody transformacji logarytmicznej. Linia ciągła odpowiada MNK (bNK_i).

Źródło: opracowanie własne.

Na rysunku 4 porównano wartości błędu *ex post* uzyskane dla modelu mieszanego (wartość parametru $B = 1{,}5$) w wyniku zastosowania metody transformacji logarytmicznej (bL) oraz MNK (bNK). Z rysunku wynika odcinkowo liniowa zależność wartości błędów od względnego poziomu składnika losowego σv, przy czym, powyżej wartości $\sigma v = 0{,}05$ następuje gwałtowny wzrost współczynnika kierunkowego prostej $bL\,(\sigma v)$. Podczas gdy, dla $\sigma v = 0{,}05$ błąd wynosi $bL = 8{,}59\%$, to dla $\sigma v = 0{,}1$ osiąga wartość $bL = 90{,}90\%$. Występuje tu, nieco załagodzona, sytuacja zaprezentowana na rysunku 2, odnoszącym się do addytywnego modelu składnika losowego – wartości błędu bL na rysunku 4 stanowią (w przybliżeniu) 50% wartości błędu bL przedstawionego na rysunku 2. Przyczyną tego jest jednakowy udział składnika addytywnego i składnika multiplikatywnego w modelu mieszanym (wzór (3)).

Podsumowując wyniki symulacji komputerowych pod kątem przydatności poszczególnych metod przybliżonych należy stwierdzić, że dla sześciu rozpatrzonych przypadków, metoda przybliżona I prowadzi do najmniejszego błędu prognozy *ex post* dla trzech wariantów: model addytywny ($B = 0{,}5$), model multiplikatywny ($B = 1{,}5$), model mieszany ($B = 1{,}5$). Metoda przybliżona III okazała się najlepsza dla dwóch przypadków: model multiplikatywny ($B = 0{,}5$) oraz model mieszany ($B = 0{,}5$). Metoda przybliżona II zapewniała najmniejszy błąd *ex post* dla modelu addytywnego z parametrem $B = 1{,}5$. Mając na uwadze, że w tym ostatnim przypadku metoda III minimalnie ustępuje metodzie II można zalecić stosowanie metod przybliżonych I i III. Metody przybliżone I i III zapewniają mniejszy błąd prognozy niż MNK dla modelu mieszanego i mutiplikatywnego z parametrem $B = 1{,}5$. Metoda transformacji logarytmicznej prowadzi do najmniejszego błędu, spośród rozpatrywanych metod, dla multiplikatywnego modelu składnika losowego. W przypadku modelu addytywnego i mieszanego metoda transformacji logarytmicznej może być stosowana dla małych wartości parametru B ($B < 0{,}8$). Ograniczenie wartości $B < 0{,}8$ wynika stąd, że dla addytywnego modelu składnika losowego ($\sigma v = 0{,}1$) wykonano symulacje komputerowe dla zmieniających się wartości parametru $B = 0{,}5;\ 0{,}6;\ 0{,}7;\ 0{,}8;\ 0{,}9;\ 1{,}0$, uzyskując wartości błędu $bL = 9{,}57;\ 9{,}69;\ 10{,}07;\ 10{,}85;\ 12{,}29;\ 15{,}68\%$. Oznacza to, że dla $B > 0{,}8$ następuje gwałtowny wzrost błędu. Ostatnia uwaga odnosi się do modelu rozpatrzonego w pracy ($A = 20$, $n = 7$, $T = 8$, $\sigma v = 0{,}1$). W celu jej uogólnienia należałoby rozpatrzyć modele z inną wartością obserwacji n oraz okresu prognozy T – w wyniku normalizacji poziomu składnika losowego (wzór (20) parametr A nie ma wpływu na wartość względnego błędu *ex post*.

4. BŁĄD *EX ANTE* PROGNOZY UZYSKANEJ MNK DLA TRENDU POTĘGOWEGO

Stosując metodę transformacji logarytmicznej, w pierwszej kolejności oblicza się odchylenie standardowe reszt S' modelu zlinearyzowanego yL_t:

$$S' = \sqrt{\frac{\sum_{t=1}^{n}(L_t - yL_t)^2}{n-2}} \tag{21}$$

gdzie: $yL_t = \ln AL + BL \cdot t'$

L_t, AL, BL, t' określa wzór (5).

a następnie wyznacza się błąd prognozy *ex ante* [5, 7]:

$$DAL_T = S' \sqrt{1 + \frac{1}{n} + \frac{(T' - \bar{t}')^2}{\sum_{t=1}^{n}(t' - \bar{t}')^2}} \cdot YP_T \tag{22}$$

gdzie: $T' = \ln(T)$; S' określa wzór (21); YP_T – wzór(16); $\bar{t}'$ – wzór (5).

Natomiast autor nie spotkał w literaturze wzoru określającego błąd prognozy *ex ante* dla oszacowania parametrów trendu potęgowego wyznaczonych MNK.

Punktem wyjścia wyprowadzenia tej zależności jest macierz informacji Fishera [1]:

$$I = \frac{1}{\hat{\sigma}a^2} \cdot \begin{bmatrix} \sum_t \left(\frac{\partial \hat{y}_t}{\partial \hat{A}}\right)^2 & \sum_t \frac{\partial \hat{y}_t}{\partial \hat{A}} \cdot \frac{\partial \hat{y}_t}{\partial \hat{B}} \\ \sum_t \frac{\partial \hat{y}_t}{\partial \hat{A}} \cdot \frac{\partial \hat{y}_t}{\partial \hat{A}} & \sum_t \left(\frac{\partial \hat{y}_t}{\partial \hat{B}}\right)^2 \end{bmatrix} \tag{23}$$

gdzie: $\hat{y}_t$ określa wzór (7)

$\hat{\sigma}a^2$ – oszacowanie wariancji składnika losowego.

Z zależności (7) i (23) otrzymuje się:

$$I = \frac{1}{\hat{\sigma}a^2} \cdot \begin{bmatrix} f_1(\hat{B}) & \hat{A}f_2(\hat{B}) \\ \hat{A}f_2(\hat{B}) & \hat{A}^2 f_3(\hat{B}) \end{bmatrix} \tag{24}$$

gdzie: $f_1(\hat{B}) = \sum_t t^{2 \cdot \hat{B}}$; $f_2(\hat{B}) = \sum_t \ln(t) \cdot t^{2 \cdot \hat{B}}$; $f_3(\hat{B}) = \sum_t \ln^2(t) \cdot t^{2 \cdot \hat{B}}$. (24a)

Następnie wyznacza się odwrotność macierzy informacji Fishera

$$I^{-1} = \frac{\hat{\sigma}a^2}{\det} \cdot \begin{bmatrix} \hat{A}^2 \cdot f_3(\hat{B}) & -\hat{A} \cdot f_2(\hat{B}) \\ -\hat{A} \cdot f_2(\hat{B}) & f_1(\hat{B}) \end{bmatrix} \tag{25}$$

gdzie: $\det = \hat{A}^2\left[f_1(\hat{B})f_3(\hat{B}) - f_2(\hat{B})^2\right]$

$f_1(\hat{B})$, $f_2(\hat{B})$, $f_3(\hat{B})$ opisuje wzór (24a).

Elementy macierzy I^{-1} leżące na głównej przekątnej określają, zgodnie z nierównością Rao-Cramera, dolną granicę wariancji estymatorów parametrów A i B [1]:

$$Var(\hat{A}) \geq V_A = \frac{\hat{\sigma}a^2 f_3(\hat{B})}{f_1(\hat{B})f_3(\hat{B}) - f_2(\hat{B})^2} \tag{26}$$

$$Var(\hat{B}) \geq V_B = \frac{\hat{\sigma}a^2 f_1(\hat{B})}{\hat{A}^2\left(f_1(\hat{B})f_3(\hat{B}) - f_2(\hat{B})^2\right)}. \tag{27}$$

W dalszej części przyjmuje się uproszczenie polegające na założeniu, że wariancja oszacowań parametrów uzyskanych MNK przyjmuje dolną granicę Rao-Cramera, oznaczoną we wzorach (26) i (27) odpowiednio V_A i V_B.

Błąd prognozy D_T wyraża się wzorem:

$$D_T = \hat{y}_T - yd_T + \varepsilon a_T \tag{28}$$

gdzie: $\hat{y}_T = \hat{A}T^{\hat{B}}$; $yd_T = AT^B$.

Dokonując rozwinięcia funkcji $\hat{y}_T$ w szereg Taylora dla funkcji dwóch zmiennych w otoczeniu punktu (A, B) i ograniczając się do wyrażeń zawierających pierwsze pochodne, uzyskuje się:

$$\hat{y}_T = AT^B + (\hat{A} - A)\frac{\partial\, yd_T}{\partial\, A} + (\hat{B} - B)\frac{\partial\, yd_T}{\partial\, B}. \tag{29}$$

Ze wzorów (28) i (29), otrzymuje się:

$$D_T = (\hat{A} - A)T^B + (\hat{B} - B)A \cdot \ln(T) \cdot T^B + \varepsilon a_T. \tag{30}$$

Wariancja błędu prognozy $V(D_T)$, wynosi:

$$V(D_T) = T^{2B}\left[V_A + V_B(A \cdot \ln(T))^2 + 2\,\mathrm{cov}(\hat{A},\hat{B})A\ln(T)\right] + \hat{\sigma}a^2 \tag{31}$$

gdzie: V_A i V_B określają wzory (26) i (27); $\mathrm{cov}(\hat{A},\hat{B})$ odpowiada elementom leżącym poza główną przekątną macierzy I^{-1} (wzór (25):

$$\mathrm{cov}(\hat{A},\hat{B}) = \frac{-\hat{\sigma}a^2 f_2(\hat{B})}{\hat{A}\left(f_1(\hat{B})f_3(\hat{B}) - f_2(\hat{B})^2\right)}. \tag{32}$$

Ze wzorów (26), (27), (31), (32) oraz przyjętego uproszczenia $A = \hat{A}$ i $B = \hat{B}$, otrzymuje się:

$$V(D_T) = \hat{\sigma}a^2\left[1 + \frac{T^{2\hat{B}}\left[\sum_t \ln^2\left(\frac{T}{t}\right)t^{2\hat{B}}\right]}{f_1(\hat{B})f_3(\hat{B}) - f_2(\hat{B})^2}\right]. \tag{33}$$

Uwzględniając, że błąd prognozy *ex ante* DA_T wyraża się zależnością:

$$DA_T = \sqrt{V(D_T)}, \tag{34}$$

uzyskuje się:

$$DA_T = S\sqrt{1 + \frac{T^{2\hat{B}}\sum_t \left(t^{\hat{B}}\ln\left(\frac{T}{t}\right)\right)^2}{\sum_t t^{2\hat{B}}\sum_t t^{2\hat{B}}\ln^2(t) - \left(\sum_t t^{2\hat{B}}\ln(t)\right)^2}} \tag{35}$$

gdzie: $t = 1, 2, \ldots, n$;

$$\hat{\sigma}a = S = \sqrt{\frac{\sum_t (y_t - \hat{y}_t)^2}{n-2}} \tag{35a}$$

y_t – obserwacje; $\hat{y}_t$ określa wzór (7); T – okres prognozy.

Zgodnie z wcześniejszymi założeniami upraszczającymi, zależność (35) określa błąd prognozy *ex ante* dla addytywnego modelu składnika losowego (wzór (1)).

W celu zweryfikowania wzorów (22) i (35) określających błąd *ex ante*, wykonano symulacje komputerowe dla trzech modeli składnika losowego (wzory (1), (2), (3)). Dla kolejnych wartości względnego poziomu składnika losowego $\sigma v = 0, 0.01, 0.02, \ldots, 0.1$ wykonano $M = 5000$ powtórzeń z wykorzystaniem generatora liczb losowych. Wartości błędu *ex post* wyznaczano zgodnie z poniższym wzorem opisującym średni błąd kwadratowy prognozy:

$$DP_T = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(YP_{m,T} - YR_{m,T}\right)^2}. \tag{36}$$

Następnie obliczano współczynnik W stanowiący stosunek błędu *ex ante* do błędu *ex post*:

$$W_T = \frac{DA_T}{DP_T}. \tag{37}$$

Współczynnik ten został oznaczony odpowiednio: WNK_T dla MNK; $WP1_T$, $WP2_T$, $WP3_T$ dla metod przybliżonych I, II, III; WL_T dla metody transformacji logarytmicznej. W przypadku współczynników WNK_T, $WP1_T$, $WP2_T$, $WP3_T$ błąd *ex ante* wyznaczano ze wzoru (35) a dla WL_T stosowano wzór (22). Zamieszczone w tabeli 1 wartości współczynników uzyskano w rezultacie uśrednienia wyników dla zmieniającego się poziomu składnika losowego σv.

Tabela 1

Stosunek wartości błędu *ex ante* do wartości błędu *ex post* (wzór (37))

Addytywny model składnika losowego (wzór (1))										
	WNK_T		$WP1_T$		$WP2_T$		$WP3_T$		WL_T	
	B = 0,5	B = 1,5	B = 0,5	B = 1,5	B = 0,5	B = 1,5	B = 0,5	B = 1,5	B = 0,5	B = 1.5
$T = 8$	1,023	1,034	1,020	1,011	1,015	1,025	1,016	1,025	1,368	---
$T = 9$	1,022	1,048	1,022	1,002	1,004	1,036	1,012	1,027	1,345	---
$T = 10$	0,995	1,038	0,981	0,970	0,961	1,019	0,969	1,008	1,283	---
Multiplikatywny model składnika losowego (wzór (2)										
$T = 8$	0,77	0,506	0,77	0,57	0,75	0,53	0,79	0,54	1,013	1,013
$T = 9$	0,75	0,517	0,76	0,56	0,73	0,52	0,77	0,54	1,000	1,000
$T = 10$	0,75	0,532	0,73	0,56	0,69	0,52	0,75	0,55	0,974	0,974
Model mieszany składnika losowego (wzór (3))										
$T = 8$	1,003	0,703	0,999	0,734	0,968	0,703	1,016	0,718	1,37	---
$T = 9$	0,939	0,690	0,940	0,72	0,902	0,688	0,951	0,704	1,27	---
$T = 10$	0,911	0,703	0,91	0,74	0,855	0,697	0,922	0,721	1,22	---

Źródło: opracowanie własne.

Analizując wyniki zawarte w tabeli 1 stwierdza się następujące fakty.

W przypadku addytywnego modelu składnika losowego wartości współczynnika WNK_T należą do przedziału $\langle 0{,}995,\ 1{,}048\rangle$; współczynniki $WP1_T$, $WP2_T$, $WP3_T$ zawierają się w przedziałach $\langle 0{,}97,\ 1{,}022\rangle$, $\langle 0{,}961,\ 1{,}036\rangle$, $\langle 0{,}969,\ 1{,}027\rangle$. Dla okresu prognozy $T = 10$ wartości współczynników są mniejsze od 1, co oznacza, że błąd prognozy *ex ante* jest mniejszy od rzeczywistego błędu prognozy wyznaczonego na drodze symulacji komputerowych. Należy stwierdzić stosunkowo dużą dokładność oszacowania błędu *ex ante* (wzór (35)) – wartości współczynników różnią się od liczby 1 o mniej niż 5%. Metoda transformacji logarytmicznej, dla parametru $B = 0{,}5$, prowadzi do wartości z przedziału $\langle 1{,}283,\ 1{,}368\rangle$, co oznacza zawyżone wartości błędu prognozy *ex ante*. W tabeli nie zamieszczono wartości współczynnika WL_T (dla wartości parametru $B = 1{,}5$), ponieważ wykazywał zbyt duże wahania (należy do przedziału $\langle 0{,}23,\ 1{,}54\rangle$), aby można było uśrednić jego wartości.

W przypadku multiplikatywnego modelu składnika losowego, zbyt małe wartości WNK_T i WP_T oznaczają nieprzydatność wzoru (35). Metoda transformacji logarytmicznej prowadzi do wartości WL_T z przedziału $\langle 0{,}974,\ 1{,}013\rangle$, co oznacza dużą dokładność wzoru (22) – zaniżone wartości błędu prognozy *ex ante* występują tylko dla okresu prognozy $T = 10$.

Dla modelu mieszanego z parametrem $B = 1{,}5$ zrezygnowano z zamieszczenia współczynnika WL_T, który należał do przedziału $\langle 0{,}26,\ 1{,}53\rangle$. Wartości współczynników

dla MNK i metod przybliżonych wskazują na możliwość stosowania wzoru (35) dla parametru $B = 0{,}5$. Natomiast, dla $B = 1{,}5$ wzór (35) nie powinien być stosowany. Odnośnie metody transformacji logarytmicznej można zauważyć, że wzór (22) jest nieprzydatny dla modelu mieszanego.

4. ZAKOŃCZENIE

W pracy rozpatrzono wybrane aspekty związane ze stosowaniem trendu potęgowego w procesie prognostycznym.

Symulacje komputerowe opisane w p. 3 wykazały dużą wrażliwość metody transformacji logarytmicznej na założony model składnika losowego. Metoda ta sprawdza się, jak powszechnie wiadomo, dla modelu multiplikatywnego a także dla modelu mieszanego i modelu addytywnego składnika losowego przy małych wartościach parametru B ($B < 0{,}8$). Natomiast metody tej nie powinno się stosować dla modeli I i III, gdy estymowany parametr B przyjmuje duże wartości ($B > 0{,}8$). Tym samym potwierdziły się zastrzeżenia dotyczące ograniczenia transformacji logarytmicznej wyrażane w pracach [3] i [4].

W odróżnieniu od pracy [3], gdzie proponuje się wykorzystanie arkusza kalkulacyjnego Microsoft Excel do rozwiązania układu równań (6), w pracy zaproponowano metody przybliżone, które prowadzą do wyników zbliżonych, jakie uzyskuje się MNK. Metody te można uporządkować w następującej kolejności: metoda przybliżona I (wzory (12), (15)), metoda przybliżona III (wzory (14), (15)) oraz metoda przybliżona II (wzory (13), (15)).

W przypadku metody I wzór (12) można zapisać w postaci dogodniejszej do obliczeń:

$$\hat{B}_1 = \frac{y_n + y_{n-1} - y_1 - y_2}{2\sum_{t=2}^{n-1} \frac{y_t}{t}}. \tag{38}$$

Stosując metody przybliżone, jak również MNK, można wyznaczyć błąd *ex ante* na podstawie wzoru (35). Przyjmując jako kryterium najmniejszą wartość błędu *ex ante* zaleca się wykonanie obliczeń trzema metodami przybliżonymi, a wartość prognozy wyznacza się tą metodą, która spełnia to kryterium.

Uniwersytet Szczeciński

LITERATURA

[1] Chow G.C., [1995], *Ekonometria*, Wydawnictwo Naukowe PWN, Warszawa.
[2] Dahlquist G., Bjorck A., [1983], *Metody numeryczne*, PWN, Warszawa.
[3] Jurkiewicz T., Plenikowska-Ślusarz T., [2001], *Algorytmy optymalizacyjne w estymacji nieliniowych funkcji regresji*, „Wiadomości Statystyczne" z. 10, s. 10-15.
[4] Kmenta J., [1990], *Elements of Econometrics* (second edition), Macmillan Publishing Co.

[5] *Prognozowanie gospodarcze. Metody i zastosowania*, [2001], pod redakcją Cieślak M. Wydawnictwo Naukowe PWN, Warszawa.
[6] Purczyński J., [2008], *Wybrane aspekty prognozowania z wykorzystaniem trendu wykładniczego*, „Przegląd Statystyczny" z. 1, s. 27-44.
[7] Zeliaś A., Pawełek B., Wanat S., [2004], *Prognozowanie ekonomiczne. Teoria, przykłady, zadania*, PWN, Warszawa.

Praca wpłynęła do redakcji w lutym 2009 r.?

METODY PROGNOZOWANIA Z WYKORZYSTANIEM TRENDU POTĘGOWEGO

Streszczenie

W pracy rozpatrzono wybrane aspekty związane ze stosowaniem trendu potęgowego w procesie prognostycznym. Zaproponowano przybliżone metody estymacji parametrów trendu (wzory (12)-(15)), które prowadzą do zbliżonych wyników, jakie daje metoda najmniejszych kwadratów (MNK). Wyprowadzono wzór (35) określający błąd *ex ante* prognozy wyznaczonej metodami przybliżonymi oraz MNK dla addytywnego modelu składnika losowego. Wykonano symulacje komputerowe uwzględniające trzy modele składnika losowego (addytywny, multiplikatywny, mieszany) mające na celu określenie zakresu przydatności metody transformacji logarytmicznej, MNK oraz metod przybliżonych.

Słowa kluczowe: prognozowanie, estymacja parametrów trendu potęgowego, błąd prognozy *ex ante*, symulacje komputerowe

METHODS OF FORECASTING WITH THE USE OF POWER TREND

Summary

In this paper selected aspects concerning the use of power trend in the forecasting process were considered. Approximate methods for estimating trend parameters (equations (12)-(15)) were proposed. The methods yielded results similar to results given by least squares method (LSM). Formula (35) determining *ex ante* error of the forecast determined by the approximate methods and LSM for random element additive model was defined. Computer simulations were done – including three models of random element (additive, multiplicative, mixed) – with the aim of determining the range of usefulness of logarithmic transformation method, LSM and the approximate methods.

Key words: forecasting, estimation of power trend parameters, forecast *ex ante* error, computer simulations

PRZEGLĄD STATYSTYCZNY
R. LVI – ZESZYT 2 – 2009

GRZEGORZ KRZYKOWSKI

BAYESIAN TECHNIQUES IN RANDOMIZED RESPONSE

1 INTRODUCTION

Opinion polls are regularly done for nearly 70 years. The first professional reports from an opinion poll were compiled by the American Institute of Public Opinion founded by George Gallup. There were earlier attempts to recognize views of particular groups of people tied by common profession, domicile or religion. However, they were usually opinions of whole groups, or sometimes representatives of these groups, but it was rare to randomely choose these representatives.

Preparing and realizing a contemporary opinion poll takes few major stages, which are carried out using advanced research techniques [9].

According to this procedure an opinion poll is finished with the moment of compiling a final report and releasing the results. At this stage the results and techniques of research are judged by the recipients of the report and a wide group of people interested in the results. This group is usually quite roughly acquainted with the methods of designing and realizing the research and of the statistical analysis of empirical data. Nevertheless, it is the orderer of the research who has a voice in making a decision about the shape and techniques of research. Often, but not always, they go by economical reasons and override content-related problems. As a consequence from the very beginning of a research project the researchers fight stereotypes that are hard to overpass and concern the nature of arising problems. In particular, the problems concern the interpretation of random issues. In one of the first polls concerning the analysis of voting preferences a sample of 10 million people[1] was taken arguing that accurate prognosis needs a huge set of data. It is only the spectacular mistakes in the results of the poll that allow researchers to argue for changing the research techniques. Another argument for new research methods is the expansion of the information technology, infrastructure and the rise of social awareness. In the times of a different view on the research results very important is the transformation of the informative media market. Society expects the media to provide accurate and reliable information substantiated by thorough research. Additionally, the research has become releatively cheap and this results with a huge group of individual research orderers. Banks, economic and social entities have researches done, the results of which are used on their own hook for substantiating argumentation concerning current economic or

[1] A 1936 poll concerning presidential election which was won by Franklin D. Roosevelt, whereas the sample favored his opponent, Alf Landon

political decisions. At the end of XXth century we observed an unimaginably dynamic growth of computer techniques. This growth is reflected by the change of approach to the analysis of statistical data. The main modification of the hitherto existing way of analyzing data is introducing the Bayesian approach in a big way. This approach not only changed the purely technical side of the form of analysis of the survey data, but the main change is the introduction of a new philosophy of giving answers to posed questions in the domain of the observation carried on.

The aim of the work is presenting and solving the issue of estimating the result in the randomized response technique. The techniques used so far of analysing this issue were based on the classical rules of statistical inferring. These methods were analytically correct and easy to compute, but as it sometimes happens, the results obtained in certain cases were unreasonable[2]. In the presented work we deeply analyse the issue of randomized response using of Bayesian methods. We present various choices of prior distributions and we show why some of them do not give the expected results. We present results and Bayesian procedures with the use of a modern V@R technology [7].

2. RANDOMIZED RESPONSE – RESEARCH EXPERIMENT DESCRIPTION

When constructing a survey research we are faced with a variety of partial problems which require detailed techniques. One of classic issues connected with formulating survey questions is balancing possible answers in the sense of positive and negative answer. This issue lies in the domain of psychological investigation and concerns aversion to or acceptance of a selected answer due to a too suggestive question. After the September 11th attack on World Trade Center it was justified to ask a question like the following one: Do you think that the United States should take a military action in retaliation for the terrorist attacks on New York and Washington? This question would suggest the answer too much and thus the actual question asked by The Gallup Organization was: Do you think that the United States should or should not take a military action in retaliation for the terrorist attacks on New York and Washington? Balance of the answers can also be achieved by proposing several answers with a growing strength of the answer. It is worse when we have to deal with sensitive questions.

The issue of truthfulness of the answers to a given questions appeared on the very beginning of the construction of the surveys. We usually assume, that the respondents answer the survey questions honestly and verdically. On the other hand we are aware, that if the question is socially or religiously or professionally sensitive, the truthfulness of the answers is up in the air [5]. The respondents try to avoid answering a sensitive question and most of them either chooses not to answer at all or chooses the answer which is socially accepted – they simply do not tell the truth. In such cases the result of the survey is erroneous and does not reflect the public opinion about the question

[2] Estimators obtained by the method of moments or plug-in method not always give an unambiguous result and not always fall in the allowable range of parameters (compare eg. [1, Chapter 2.1])

posed. The Randomized Response technique aims at discharging the respondent from publishing his standpoint about given issue. This kind of freeing from remorses was used from way back. The biblical penalty of stoning and later the execution by shooting by a firing squad had similar features. The firing squad chosen to shoot the convict was given rounds of ammunition, some of which were blank. They were distributed randomely and in fact no one knew who was the one who shot the convict.

In the classical approach [10] the research technique of randomized response is answering one of two connected questions randomely chosen by the respondent. Balance plays a significant role in the construction of the questions. Suppose the socially sensitive question (denoted Q_0) is as follows:

Q_0: *Did you ever drive in the state of significant alcohol intoxication?*

We add second, complementary question of opposite nature balancing the research issue (denoting Q_1):

Q_1: *Did you always qive up driving when in the state of significant alcohol intoxication?*

The questions may have quite simple form.

Q_0: *Will you vote for candidate x in the next presidential election?*

Q_1: *Will you not vote for candidate x in the next presidential election?*

Questions Q_0 and Q_1 are constructed in such a way that answering "Yes" to one of them is equivalent to answering "No" to the other. We propose the respondent to toss a die (in private) and answering question Q_0 when he/she gets the outcome of 1, 2, 3, 4 or 5, and answering Q_1 when he/she gets 6, but we expect the question to be answered honestly. The result is an answer of "Yes" or "No", but the researcher does not know which question was answered. As a result we don't know the opinion of the person questioned. The assessment of the respondent's standpoint is influenced by his/her opinion about the presented issue together with the frequency of choosing question Q_0. In the simpliest approach the respondent tosses a die and then gives the answer. This schema implies a computational process after the diagram in the figure (1).

From this, the probabilites of getting a positive and negative answer in a single sample are, accordingly:

$$\Pr(\text{Yes}) = \mu\theta + (1-\mu)(1-\theta) = \frac{2}{3}\theta + \frac{1}{6},\ \Pr(\text{No}) = \frac{5}{6} - \frac{2}{3}\theta,$$

where $1 - \mu$ is the probability of getting „6" and θ is the probability of the support for the candidate.

If the answers come from n respondents, we can try to assess the parameter θ, denoting „real" frequency of a positive answer to the sensitive question posed. Using the method of maximum likelihood (method of moments, plug-in principle) we get estimator for θ of the following form

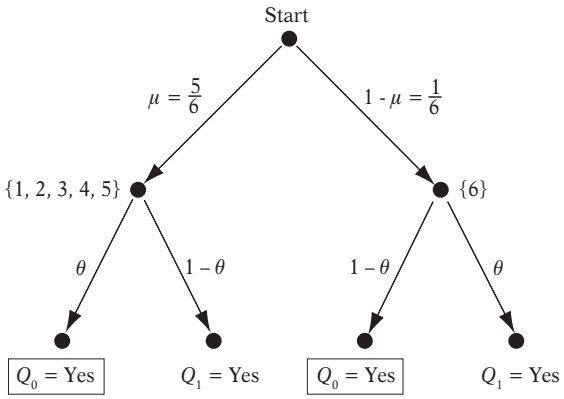


Figure 1. Engel's diagram

$$\hat{\theta}_{ML} = \frac{y_T + \mu - 1}{2\mu - 1} = \frac{3}{2} y_T - \frac{1}{4}, \tag{1}$$

where $y_T = \frac{n_T}{n}$ is the fraction of the answer „Yes" in the n element simple random sample from a binomial distribution with parameters $\left(n, \frac{2}{3}\theta + \frac{1}{6}\right)$.

Right at the first sight we notice that the estimator (1) has significant shortcomings. One of them is inappropriate range. For $n = 20$ respondents and $n_T = 3$ „Yes" answers from (1) we get a negative value of

$$\hat{\theta}_{ML} = -\frac{11}{80}.$$

We get similarly inappropriate (greater than 1) assesment of θ for $n_T = 17$. In such cases we assume values of 0 and 1 to be the estimator of greatest reliabilty, but this is not a favourable solution.

The literature of the subject gives examples of attempts to analyse the survey results using the randomized response technique. These attempts usually enrich the form of posing questions with the cases of not connected questions or systems of questions with multiple answers. The estimation optimalisation in those cases uses classical estimation techniques with a drift towards minimizing the variance [8]. We also find few suggestions of a different approach. One of them concerned using logit transformation [5], [2], another was based on the Bayesian approach [6] and was orientated towards considering a case where in place of the answer „Yes" there were several different answers considered. However, until now there were no studies that would transgress the standards found in monographic studies.

3. BAYESIAN APPROACH – MOTIVATION

In the Bayesian approach we assume that the assesed parameter is a random variable (that is, a random variable of a known distribution). At first we assume that

the distribution of the estimated parameter is known. We call it prior distribution. As a result of observation obtained from a survey we get a modification of prior distribution called posterior distribution. The posterior distribution is obtained by applying Bayes formula to a known conditional distribution of obtained empirical data an the prior distribution. The posterior distribution is the final result of the estimation.

Earlier works the Bayesian estimator was understood as the expected value of the posterior distribution. This value was derived from the conditions of minimum risk with a square loss function. This procedure shows some similarity to classical methods of estimation and justifies the name of this value. Nowadays authors consequently avoid the name „Bayesian estimator" using more often the „prior expected value" or, less accurately, „prior average". For example, such rule is used in monograph [4]. In another monograph [1] the term „Bayesian procedure" is used, and usage of „Bayesian estimator" is retired.

In practice the posterior distribution is presented by density or frequency graphs and by giving quantile characteristics, the average and the standard deviation of the distribution.

In theory the Bayesian approach can be used in almost every model. In practice we use it when we are aiming less at estimating the parameters of the distribution and more on modyfing the hitherto existing knowledge about the parameter. In the presented issue (the randomized response) we usually have very much initial information. The very approach to the question as a sensitive one is introducing an imprecise, but entrenched knowledge about the answers. It evinces in a supposition that the answers will be dishonest, so the classical empirical frequencies will be underrated (or overrated). The common usage of the maximum likelihood estimator according to formula (1) is supported by the prior information. We do assume, that in this approach it is unlikely for the values of the estimator to exceed the range and we treat it as the proper estimate of the frequency of an answer to a sensitive question. However, it turns out that the probability of a negative estimate of the frequency parameter θ may be quite big. We have equalities:

$$\Pr_\theta\left(\hat{\theta}_{ML} < 0\right) = \Pr_\theta\left(n_T < (1-\mu)n\right) = F_\theta\left((1-\mu)n\right)$$

where $F_\theta(\cdot)$ is the cumulative distribution function of a random variable of binomial distribution with parameters (n, θ). As for any $\theta \in (0, 1)$ and $\mu \in \left(\frac{1}{2}, 1\right)$

$$\lim_{n \to \infty} F_\theta\left((1-\mu)n\right) = 1,$$

then $\Pr_\theta\left(\hat{\theta}_{ML} < 0\right)$ may be arbitrarily close to one, so it does not have to be as small as it is commonly taken. Likewise for fixed n and $\mu \in \left(\frac{1}{2}, 1\right)$

$$\lim_{\theta \to 1} \Pr_\theta\left(\hat{\theta}_{ML} > 1\right) = 1 - \lim_{\theta \to 1} \Pr_\theta\left(n_T \leq n\mu\right) = 1,$$

so for every size of the sample and every $\mu \in \left(\frac{1}{2}, 1\right)$ we can select θ, such that $\Pr_\theta\left(\hat{\theta}_{ML} > 1\right)$ is arbitrarily close to one. From this exceeding value of one by the esti-

mator given by the formula (1) is quite a common effect when the value of estimated parameter θ is big.

The Bayesian approach is always connected with some limitations in the choice of parameters and interpretation of the results, but these are limitations we can influence and modify as needed. When using the maximum likelihood estimators we can at most interpret the results.

4. CONSTRUCTION OF A MODEL

Usually a survey of sensitive issues is preceded by partial or preparatory surveys. The results obtained from these preliminary information constitute the initial knowledge which is grouped to form an prior distribution. The issue of constructing an prior distribution is complicated and demands a great experience in analysing empirical data.

The researcher has to be very well acquainted to the content-related side of the analysed phenomenon. In the randomized response technique setting the conditional distribution from which the random sample is taken is quite natural and does not cause too much discrepancy between researchers. It is assumed that it is a binomial distribution with proper parameters, which usually reflects well the nature of the examined phenomenon. It is more difficult to set an prior distribution. We will discuss this issue in the next sections.

Let us assume the simpliest Warner model [10]. In this model we denote by μ a constant value corresponding to the frequency at which we draw question Q_0: We assume that $\mu \in \left(\frac{1}{2}, 1\right)$ which means that as a result of drawing a question by the respondent we obtain Q_0 more often. Then $1 - \mu$ is the probability of obtaining question Q_1. By $\theta \in (0, 1)$ we denote the value of a random variable Θ reflecting the frequency of a positive answer to a sensitive question Q_0. Let X be a random variable corresponding to the answer, which was given by the respondent. The random variable takes values „Yes" and „No" and its conditional distribution (on the condition Θ) has the following form:

$$\begin{aligned} \Pr(X = \text{Yes} \,|\, \Theta = \theta) &= \mu\theta + (1 - \mu)(1 - \theta), \\ \Pr(X = \text{No} \,|\, \Theta = \theta) &= 1 - (\mu\theta + (1 - \mu)(1 - \theta)). \end{aligned} \tag{2}$$

We will use values one and zero interchangeably instead of „Yes" and „No" accordingly, and for simplicity of notation we denote

$$\psi_\mu(\theta) = \mu\theta + (1 - \mu)(1 - \theta). \tag{3}$$

Let $Y_1, Y_2, \ldots, Y_n$ be a random n size simple sample from a distribution given by formula (2). By n_T we denote the number of „Yes" answers in this sample, and by n_N the number of „No" answers. With the above assumptions the combined conditional distribution of the random vector $Y = (Y_1, Y_2, \ldots, Y_n)$ has the following form:

$$\Pr(Y = y \,|\, \Theta = \theta) = \left(\psi_\mu(\theta)\right)^{n_T} \left(1 - \psi_\mu(\theta)\right)^{n_N} \tag{4}$$

where $\psi_\mu(\theta)$ is given by formula (3) and

$$\frac{1}{2} < \mu < 1,\ 0 < \theta < 1,\ y = (y_1,\ y_2,\ \ldots,\ y_n) \in \{Yes,\ No\}^n. \tag{5}$$

The obtained conditional distribution is a binomial distribution with parameters $(n,\ \psi_\mu(\theta))$.

4.1 THE CHOICE OF PRIOR DISTRIBUTION – NONINFORMATIVE PRIORS

In the prior distribution we intend to gather the hitherto existing information about the assessed parameter represented by the random variable Θ. If there are no clues concerning this parameter, we can assume that the prior distribution of variable Θ is one of classical noninformative distributions. For a binomial distribution noninformative distributions are well known [11]. For a randomized response procedure where the parameter of success is given by formula (3), that is as a linear function of an unknown conditioning parameter Θ, there are no such data. Below we bring up noninformative priors for the analysed model deriving from classical assumptions concerning the construction of these distributions and we will present the consequences of these choices. As the first noninformative distribution we assume the uniform distribution on the interval (0, 1), which means the density of variable Θ has the following form:

$$f_\Theta(\theta) = 1,\ \theta \in (0,1).$$

Then we get the posterior distribution

$$f_{(\Theta|Y)}(\theta \,|\, y) = \psi_\mu^{n_T}(\theta)\left(1 - \psi_\mu(\theta)\right)^{n_N},\ \theta \in (0,1)$$

where notation is given in formula (4), and the range of parameters is described by inequalities (5).

For $\mu = \frac{5}{6}$, $n = 100$ the conditional expected value of parameter Θ is given by the formula

$$\mathbb{E}\left(\Theta \,|\, Y = y\right) = C_\mu \int_0^1 \theta \left(\frac{2}{3}\theta + \frac{1}{6}\right)^{n_T} \left(\frac{5}{6} - \frac{2}{3}\theta\right)^{n_N} d\theta,$$

where $C_\mu = C_{\frac{5}{6}}$ is a normalisation constant, and n_T and n_N are the number of successes and failures in the sample, accordingly. The graph of this conditional expected value as a function of the number of „Yes" answers (that is as a function of n_T) if presented in figure 2. As it shows for the frequency of successes up to $\frac{1}{2}$ the expected value of posterior distribution is smaller than the empirical result, that is smaller than $\frac{n_T}{n}$, and for the frequency of successes above $\frac{1}{2}$ the expected value of osterior distribution is

above the empirical result. It is worthwile to consider this dependance for a changing parameter μ. For $\mu = 1$ we get a classical result, ie. the expected value of posterior distribution is a linear function of the number of successes (See Section (5.3)) and for $\frac{n_T}{n} < \frac{1}{2}$ is above the empirical result $\frac{n_T}{n}$, and for $\frac{n_T}{n} > \frac{1}{2}$ is below the empirical result (conversely to the figure 2). Together with the decrease of the value of parameter μ we observe the empirical results and the expected value of posterior distribution diverging to the form as in figure 2, which means the expected value of posterior distribution for empirical result $\frac{n_T}{n} < \frac{1}{2}$ falls below this result and for empirical data such that $\frac{n_T}{n} > \frac{1}{2}$ it is greater than the empirical value.

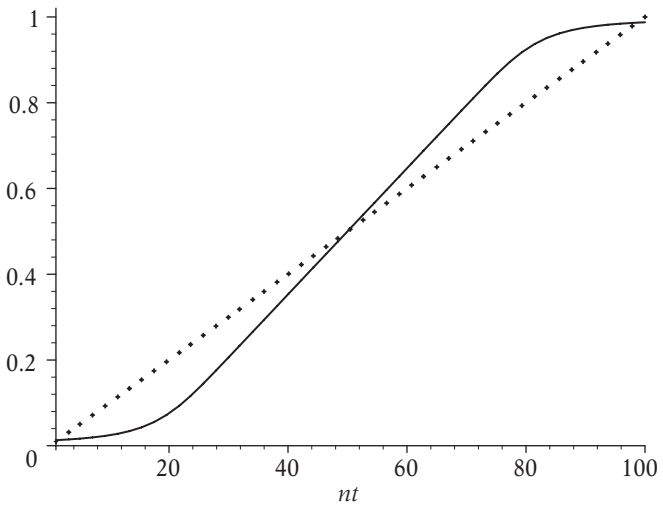


Figure 2. Expected value of the posterior distribution as a function of the number of successes. Dotted line denotes fraction $n_T/100$; solid line denotes the expected value

Another classical noninformative distribution in the described randomized response procedure which we consider is the Jaffreys distribution basing on the Fisher amount of information. In this case the density of the parameter Θ has the form (see Section 5.1)

$$f_{\Theta}(\theta) = \frac{2\mu - 1}{2\,arcsin(2\mu - 1)} \frac{1}{\sqrt{\psi_{\mu}(\theta)\big(1 - \psi_{\mu}(\theta)\big)}},\ \theta \in (0,1),\ \mu \in \Big(\frac{1}{2}, 1\Big),$$

and therefore the density of posterior distribution fulfills the proportion

$$f_{(\Theta|Y)}(\theta \,|\, y) \overset{\theta}{\propto} \psi_{\mu}(\theta)^{n_T - \frac{1}{2}} \big(1 - \psi_{\mu}(\theta)\big)^{n_N - \frac{1}{2}}.$$

This distribution has the same feature as the uniform distribution. For the frequency of successes up to $\frac{1}{2}$ the conditional expected value is below the empirical frequency $\frac{n_T}{n}$, and for the frequency of successes above $\frac{1}{2}$, it is above. The graph of conditional expected value as a function of the number of successes is similar to the graph in figure 2.

The third classical noninformative distribution, prior MDIP distribution, also has similar features. We obtain this distribution by maximizing entropy (see section 5.2). In this case the density of prior distribution is given by formula

$$f_\Theta(\theta) = C_\mu \psi_\mu(\theta)^{\psi_\mu(\theta)} \left(1 - \psi_\mu(\theta)\right)^{\left(1 - \psi_\mu(\theta)\right)},\ \theta \in (0,1)$$

where $C_\mu = C_{\frac{5}{6}}$ is the normalization constant, and $\psi_\mu(\theta)$ is given by the formula (4) with limitations to the range of parameters (5).

4.2 CHOOSING BETA DISTRIBUTION AS THE PRIOR DISTRIBUTION

Another group of prior distributions are the distributions that bear some information obtained from hitherto research. In case of a sample taken from a binomial distribution we take as a standard prior distribution Beta distribution with parameters α, β with different way of selecting the α and β parameters. The main reason after choosing Beta distribution is the fact, that Beta distribution is conjugate for binomial distribution [3, §9.4 Tw.2]. In the case of randomized response model Beta distribution is not conjugate (except the case of $\mu = 1$). However the group of Beta distributions is fairly rich and we can expect that taking an prior distribution from this group may be successful used. Let's assume that the „knowledge" about the nature of frequency expressed by random parameter Θ resolves to the fact that we can set the expected value and variance

$$\mathbb{E}\Theta = m_0,\ Var(\Theta) = \sigma_0^2,$$

where m_0 and σ_0^2 are known values. We then assume that the random variable Θ has a Beta distribution with parameters α and β connected to values m_0 and σ_0^2. We can compute the parameters of Beta distribution starting from a system of equations (eg. [4, Sec. A.2])

$$\frac{\alpha}{\alpha + \beta} = m_0,\ \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} = \sigma_0^2. \tag{6}$$

This system has the following resolution

$$\alpha = m_0(1 - m_0)\left(\frac{m_0}{\sigma_0^2} - \frac{1}{1 - m_0}\right),\ \beta = (1 - m_0)^2\left(\frac{m_0}{\sigma_0^2} - \frac{1}{1 - m_0}\right). \tag{7}$$

We notice that if some random variable ξ has Beta distribution, and parameters are such that $\alpha \neq 1$ or $\beta \neq 1$ and $m = \mathbb{E}\xi$, $\sigma^2 = \mathrm{Var}(\xi)$, then from inequality $\mathbb{E}\xi > \mathbb{E}\xi^2$ we get the inequality

$$\frac{m}{\sigma^2} - \frac{1}{1-m} = \frac{\mathbb{E}\xi - \mathbb{E}\xi^2}{(1-m)\sigma^2} > 0.$$

Then the parameters α and β in the formula (7) are well described (are non-negative). In the same time we have to notice that the expected value m_0 and variance σ_0^2 in the Beta distribution fulfill the conditions $0 < m_0 < 1$ and $\sigma_0^2 \leq \frac{1}{4}$.

Additionally, a condition resulting from the relationship between variance and expected value has to be fulfilled, which implies the inequality

$$\sigma_0^2 < m_0(1-m_0),\ m_0 \in (0,1). \tag{8}$$

Solving this inequality we have that for a given $\sigma_0^2 \in \left(0, \frac{1}{4}\right)$ value m_0 fulfills the inequality

$$\frac{1}{2} - \sqrt{\frac{1}{4} - \sigma_0^2} < m <_0 \frac{1}{2} + \sqrt{\frac{1}{4} - \sigma_0^2} \tag{9}$$

Let us denote for simplicity

$$m_1 = \frac{1}{2} - \sqrt{\frac{1}{4} - \sigma_0^2},\ m_2 = \frac{1}{2} + \sqrt{\frac{1}{4} - \sigma_0^2} \tag{10}$$

Then $m_0 \in (m_1, m_2)$, $\sigma_0^2 \in \left(0, \frac{1}{4}\right)$.

The Beta distribution taken as the prior distribution has density

$$f_\Theta(\theta) = \frac{\Gamma(a+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1},\ \theta \in (0,1) \tag{11}$$

where α and β are described by the formulas (7) with the condition (8).

The subtle analysis of the choice of α and β parameters in a classical case has been presented in monograph [4, Chapter 5]. The analysis taken out by Gelman and coauthors concerns parameters α and β of Beta distribution, when the conditional distribution of a simple random sample is a binomial distribution. The authors assume as well the conditions resulting from setting prior expected value and variance in the Beta distribution. With the randomized response techniqe, however, we have quite a different model and completely dissimilar expectations about the results of the Bayesian estimation.

The conditional distribution of the random simple sample is given by the formula (4). From this, from the Bayes formula the posterior distribution fulfills the condition

$$f_{(\Theta|Y)}(\theta \mid y) \overset{\theta}{\propto} \psi_\mu(\theta)^{n_T}\left(1 - \psi_\mu(\theta)\right)^{n_N}\theta^{\alpha-1}(1-\theta)^{\beta-1} \tag{12}$$

where $\psi_\mu(\theta)$ is given by formula (3), $\theta \in (0, 1)$ and n_T is equal to the number of „Yes" answers and n_N is equal to the number of „No" answers for the result of the random sample $(y_1, y_2, \ldots, y_n)$. The randomized response research technique assumes in its essence underrating (or overrating) the frequency of classical answers. The respondent of a survey answers „No" (or „Yes") more frequently than it is owed to his attitude. Let us assume that the question is phrased in such a way, that respondents in a classical survey underrate the frequency of positive answers, thus hiding their preferences.

As a consequence, if a classically driven preliminary or partial survey suggests, that the expected value of the answer „Yes" is about $m_0 = 15\%$, then the survey using randomized response technique should give more „Yes" answers than 15%.

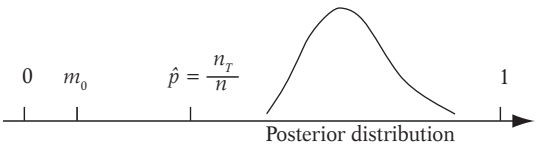


Figure 3. Location of the expected value of the prior distribution (m_0), the empirical value $\hat{p}$ and the posterior distribution

This number of „Yes" answers should imply higher estimate of the frequency parameter θ. The location of these values is schematically presented in figure 3. The schema given in figure 3 does not have to be fulfilled in every survey, but it reflects a natural situation basing on prior premises. This location is usually fulfilled when the success parameter has a value below $\frac{1}{2}$.

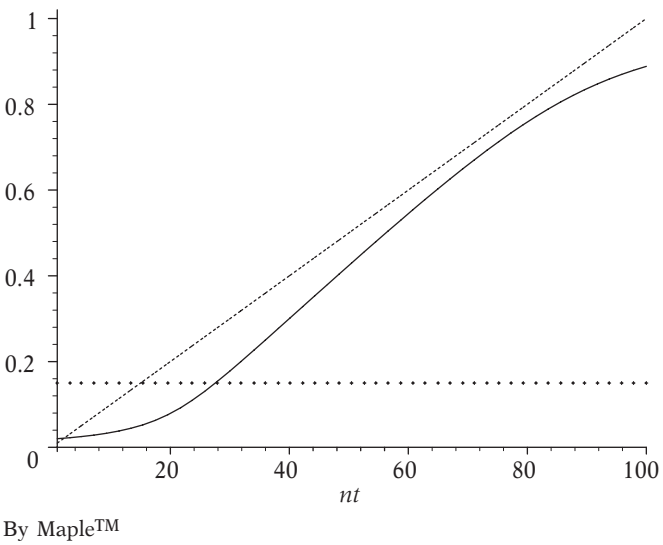


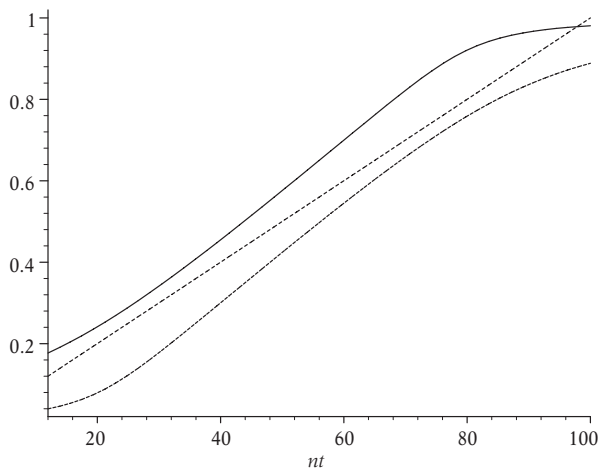
Figure 4. Expected value of posterior distribution as a function of the number of successes. Dotted line denotes fracton $\frac{n_T}{100}$, solid line denotes the expected value, and the crossed line denotes the limit of 15%

The approaches we presented, where the priors is noninformative or is a Beta distribution with parameters α and β derived from assumptions about expected value and variance of Beta distribution given prior, do not thoroughly reflect the nature of the answer to a sensitive question. They do not take into consideration the main characteristics of these answers, that is the underration of the frequency of positive answers to the sensitive question in the classical technique. We solve this problem by using some sort of symetrization by changing the meaning of parameters α and β in the prior assumed Beta distribution.

4.3 COMPUTATIONAL TECHNIQUES AND EXAMPLES

We start illustrating the application of Beta distribution as prior distribution from assuming an prior Beta distribution of expected value $m_0 = 0.15$ and variance $\sigma_0^2 = 0.01$. Then from the formula (7) we get that $\alpha = 1.7625$ and $\beta = 9.9875$. The graph of conditional expected value $\mathbb{E}(\Theta|Y = y)$ as a function of n_T that is the number of "Yes" answers (for $n = 100$ tries) is presented in figure (4).

The solid line, denoting the expected value in posterior distribution, is located below the dotted line, which proves that the conditional expected value for every number of successes $n_T > 2$ underrates the empirical value which comes from the number of obtained successes. The underration is quite substantial. As it can be seen in figure 4 only at $n_T = 27$ to 100 successes in the sample the expected value of posterior distribution is 15%. A similar phonomenon of underration of the frequency of successes by the expected value of posterior distribution is observed at different values of m_0 and σ_0^2. This phenomenon should not occur in the analysed issue, we should rather expect the location of expected value of posterior distribution such as shown in figure 3.



By Maple™

Figure 5. The expected value of posterior distribution as a function of the number of successes. Dotted line denotes the fraction $\frac{n_T}{100}$, solid line below the dotted one denotes the expected value of posterior distribution without modification, and the solid line above the dotted one – with modification.

This underration is caused by a characteristic of Beta distribution signaled in section (5.3), namely the fact that the expected value of posterior distribution is, as a result of assumed prior Beta distribution and binomial likelihood, a linear combination of the initial expected value and the empirical value. This fact is even deepened in case of introducing that parameter of constant frequency μ. The solution to this problem is reversing analogical relationship above the empirical value, in some sense, a symmetrical reflection. We achieve this solution by assuming as the prior distribution the Beta distribution, but with exchanging parameters α and β (compare section (5.3)).

After such symmetrization we get the expected result. The effect of this symmetrization taking place is presented in figure (5). The choice of parameters α and β of the prior distribution is conditioned by the intial information. We may assume, that from earlier analysis basing on classical techniques we have some approximation of the expected value m_0 and variance σ_0^2 of the prior distribution. Then from the formula (7) we find parameters α and β, and then we assume the prior distribution to be $\beta(\beta, \alpha)$. With this choice of parameters the posterior distribution is given by the formula

$$f_{(\Theta|\underline{X})}(\theta \mid \underline{x}) \overset{\theta}{\propto} \psi_\mu(\theta)^{n_T}\left(1 - \psi_\mu(\theta)\right)^{n_N}\theta^{\beta-1}(1-\theta)^{\alpha-1}, \tag{13}$$

where $\psi_\mu(\theta)$ is given by formula (3), $\theta \in (0, 1)$ and n_T is equal to the number of „Yes" answers, and n_N to the number of „No" answers for a random sample $\underline{x} = (x_1, x_2, \ldots, x_n)$. As m_0 and σ_0^2 unambiguously determine the Beta distribution

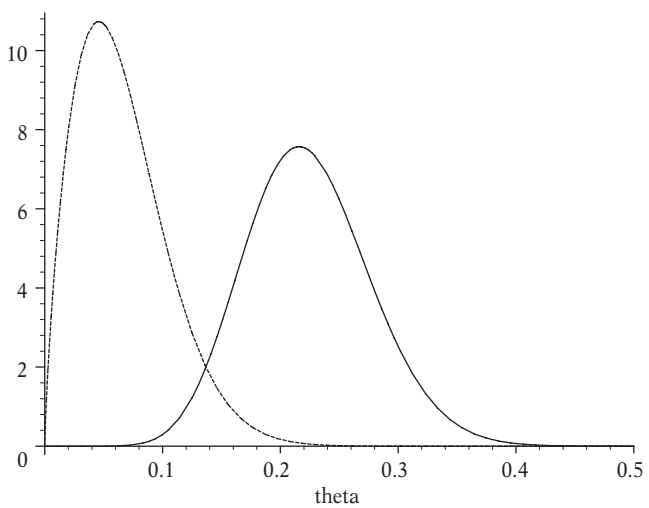


Figure 6. Density of posterior distribution before symmetrization (dotted line) and after symmetrization (solid line) for $m_0 = 0.15$, $\sigma_0^2 = 0.01$, $n = 100$, $n_T = 18$.

as the prior distribution, let $\alpha = k\ n\ m_0$ and $\beta = k\ n - \alpha$. Then the new parameter k fulfills the equation

$$k = \frac{1}{n}\left(\frac{m_0(1 - m_0)}{\sigma_0^2} - 1\right). \tag{14}$$

Value k is well chosen (non-negative) if conditions described by inequalities (9) are fulfilled. For example for $n = 100$, $m_0 = 0.15$, $\sigma_0^2 = 0.01$ value of k is 0.1175. Several more values of parameter k are given in table 1.

Table 1

Some values of the proportion parameter k in Beta distribution

μ_0	0.05	0.10	0.15	0.20
$\sigma_0^2 = 0.01$	0.0375	0.0800	0.1175	0.0150
$\sigma_0^2 = 0.04$	0.0019	0.0125	0.0219	0.0300

Eventually we assume as the prior distribution the Beta distribution with parameters $\alpha = k\ n\ (1 - m_0)$ and $\beta = k\ n\ m_0$, where k is given by formula (14).

Let us consider some examples. Let $m_0 = 0.15$ and $\sigma_0^2 = 0.01$ and as a result of simple random sample we obtained $n_T = 18$ successes. Then the posterior distribution has density presented in figure 6. The expected value of the posterior distribution is $\mathbb{E}\ (\Theta|\underline{X}) = 0.2219$, and values of quantiles are given in table 2. This table contains also the quantiles for the greater number of successes $n_T = 20$. Let us note that in this case the estimators of maximum likelihood given by formula (1) for $n = 100$, $n_T = 18$ and $n_T = 20$ are $\Theta_{ML} = 0.02$ and $\Theta_{ML} = 0.05$ accordingly and are of no sense whatsoever in this issue. In the second example we assume $m_0 = 0.10$, $\sigma_0^2 = 0.005$, $n = 200$, $n_T = 30$.

Table 2

Quantiles of posterior distribution ($n = 100$, $m_0 = 0.15$, $\sigma_0^2 = 0.01$)

n_T	$Q_{0.05}$	$Q_{0.10}$	$Q_{0.25}$	$Q_{0.50}$	$Q_{0.75}$	$Q_{0.90}$	$Q_{0.95}$
$n_T = 18$	0.1417	0.1579	0.1867	0.2212	0.2580	0.2930	0.3146
$n_T = 20$	0.1558	0.1729	0.2030	0.2389	0.2770	0.3129	0.3351

The graph of the posterior density is presented in figure 7. The expected value

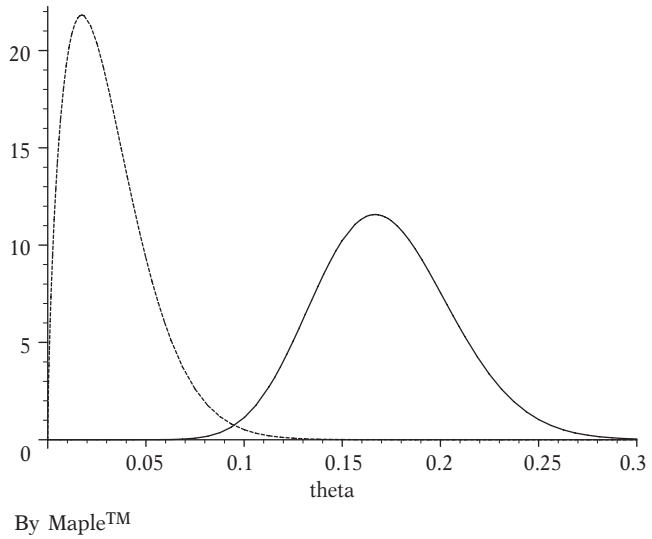


Figure 7. Density of posterior distribution before symmetrization (dotted line) and after symmetrization (solid line) for $m_0 = 0.10$, $\sigma_0^2 = 0.005$, $n = 200$, $n_t = 30$

of posterior distribution is $\mathbb{E}(\Theta|\underline{X}) = 0.1715$, and the values of quantiles are given in table 3. The estimator of greates reliability given by formula (1) for $n = 200$, $n_T = 30$ is negative and we have to assume $\Theta_{ML} = 0$, which does not conform to the conditions of the issue. Lastly, let us note that the Bayesian

Table 3

Quantiles of posterior distribution ($n = 200$, $m_0 = 0.10$, $\sigma_0^2 = 0.005$, $n_T = 30$)

$Q_{0.05}$	$Q_{0.10}$	$Q_{0.25}$	$Q_{0.50}$	$Q_{0.75}$	$Q_{0.90}$	$Q_{0.95}$
0.1175	0.1283	0.1474	0.1700	0.1940	0.2169	0.2310

estimatian is as always very sensitive to the size of the sample. For a greater size we have to increase the precision of prior distribution relatively decreasing the value of variance σ_0^2 in the Beta distribution.

5. APPENDIX

5.1 NONINFORMATIVE JEFFREYS' PRIOR

Lemma 1 *Let X be a random variable of a distribution belonging to a family of distributions fulfilling the condition*

$$\Pr(X = x) = \psi_\mu^x(\theta)\big(1 - \psi_\mu(\theta)\big)^{1-x},$$
$$\psi_\mu(\theta) = \theta\mu + (1-\theta)(1-\mu),\ x \in 0,1,\ \theta \in (0,1),\ \mu \in \Big(\frac{1}{2},1\Big).$$

Then the Fisher's amount of information of the parameter θ is given by the formula

$$I(\theta) = \frac{(2\mu-1)^2}{\psi_\mu(\theta)\big(1-\psi_\mu(\theta)\big)}.$$

Proof. For $x \in \{0,\ 1\}$, $\theta \in (0,\ 1)$, $\mu \in \Big(\frac{1}{2},1\Big)$ we denote

$$f(\theta;x) = \psi_\mu^x(\theta)\big(1-\psi_\mu(\theta)\big)^{1-x}.$$

Then from the definition of the Fisher's amount of information for a family of distributions indexed by parameter θ we have

$$I(\theta) = \mathbb{E}\Big(\frac{\partial \ln f}{\partial\,\theta}\Big)^2 = \mathbb{E}\Big[\frac{\partial}{\partial\,\theta}\big(X\ln\psi_\mu(\theta) - (1-X)\ln\big(1-\psi_\mu(\theta)\big)\big)\Big]^2$$
$$= E\Big[X\frac{\psi_\mu'(\theta)}{\psi_\mu(\theta)} + \frac{1-X}{1-\psi_\mu(\theta)}\psi_\mu'(\theta)\Big]^2 = \big[\psi_\mu'(\theta)\big]^2\frac{\mathbb{E}\big[X-\psi_\mu(\theta)\big]^2}{\psi_\mu(\theta)\big(1-\psi_\mu(\theta)\big)}$$
$$= \frac{2\mu-1}{\psi_\mu(\theta)\big(1-\psi_\mu(\theta)\big)}Var(X) = \frac{(2\mu-1)^2}{\psi_\mu(\theta)\big(1-\psi_\mu(\theta)\big)}.$$

The noninformative Jeffreys prior distribution is given by the relationship

$$f_\Theta(\theta) \overset{\theta}{\propto} \frac{1}{\sqrt{I(\theta)}},\ \theta \in (0,1).$$

As for $\mu \in \Big(\frac{1}{2},1\Big)$

$$\int_0^1 \frac{2\mu-1}{\sqrt{\psi_\mu(\theta)\big(1-\psi_\mu(\theta)\big)}}d\theta = \frac{2\,arcsin(2\mu-1)}{2\mu-1}$$

then

$$f_\Theta(\theta) = \frac{2\mu-1}{2\,arcsin(2\mu-1)}\frac{1}{\sqrt{\psi_\mu(\theta)\big(1-\psi_\mu(\theta)\big)}},\ \theta \in (0,1)$$

is a noninformative prior Jeffreys' distribution.

For $\mu = 1$ we get the noninformative Jeffreys' distribution of a binomial distribution with parameter θ [11].

5.2 NONINFORMATIVE MDPI PRIOR

Let X be a random variable taking a finite number of values m with positive probabilities $p_1, p_2, \ldots, p_m$. Then by the entropy of the distribution $p_1, p_2, \ldots, p_m$ we mean the value

$$H = \sum_{i=1}^{m} p_i \ln p_i.$$

It is easy to show that the entropy is in this case a value limited by $\ln m$ and the maximum value is reached for a uniform distribution on the set $1, 2, \ldots, m$. We define the prior MDIP (Maximum Data Information Prior) distribution by the releationship

$$f_\Theta(\theta) = C_\mu \exp\{H(\theta)\}.$$

If a random variable X has a distribution given by the formula

$$\Pr(X = x) = \psi_\mu^x(\theta)\big(1 - \psi_\mu(\theta)\big)^{1-x},\ x \in (0,1),$$

then the entropy of the distribution of the random variable X fulfills the condition

$$H(\theta) = -\big(\psi_\mu(\theta)\ln\psi_\mu(\theta) + \big(1 - \psi_\mu(\theta)\big)\ln\big(1 - \psi_\mu(\theta)\big)\big).$$

From this

$$f_\Theta(\theta) = C_\mu \psi_\mu(\theta)^{\psi_\mu(\theta)}\big(1 - \psi_\mu(\theta)\big)^{(1-\psi_\mu(\theta))},\ \theta \in (0,1)$$

is the prior MDIP distribution. Because entropy is limited, this distribution is a proper noninformative distribution. The constant C_μ can be computed for a given value of μ. In table 4 we computed come values of the normalization constant

Table 4

Normalization constant in the MDIP distribution

μ	$\mu = 1$ (binomial distribution)	$\mu - \frac{5}{6}$	$\mu - \frac{4}{5}$	$\mu - \frac{3}{4}$	$\mu - \frac{2}{3}$
C_μ	1.6184	1.8454	1.8759	1.9184	1.9626

5.3 BETA PRIOR

Lemma 2 *Let the conditional distribution of a random simple sample* $\underline{X} = (X_1, \ldots, X_n)$ *on the condition of parameter* Θ *be given by the formula*

$$\Pr\big(\underline{X} = \underline{x} \,|\, \Theta = \theta\big) = \theta^{\left(\sum_{i=1}^{n} x_i\right)}(1-\theta)^{\left(n - \sum_{i=1}^{n} x_i\right)}$$

where $\theta \in (0, 1)$, $\underline{x} \in \{0, 1\}^n$ *(binomial distribution). Let the prior distribution of parameter* Θ *be Beta distribution with parameters* (α, β). *Then the expected value of posterior distribution is given by the formula*

$$\mathbb{E}\big(\Theta \,\big|\, \underline{X} = \underline{x}\big) = tm + (1 - t)\hat{p},$$

where $m = \mathbb{E}\,\Theta$, $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $t = \frac{\alpha + \beta}{n + \alpha + \beta}$.

Proof. As parameter Θ has a distribution of $\beta(\alpha, \beta)$, then the prior density is given by formula

$$f_\Theta(\theta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \cdot \Gamma(\beta)} \theta^{\alpha - 1}(1 - \theta)^{\beta - 1},\ \theta \in (0,1).$$

From this, the density of posterior distribution fulfills the proportionality condition

$$f_{(\Theta \,|\, \underline{X})}(\theta \,|\, \underline{x}) \overset{\theta}{\propto} \theta^{x_T + \alpha - 1}(1 - \theta)^{x_N + \beta - 1},\ \theta \in (0,1),\ \underline{x} \in \{0,1\}^n,$$

where $x_T = \sum_{i=1}^{n} x_i$, $x_N = n - x_T$. Then the posterior distribution is $\beta(\alpha + n_T, \beta + n_N)$.

From this its expected value may be denoted as follows

$$\mathbb{E}\big(\Theta \,\big|\, \underline{X} = \underline{x}\big) = \frac{\alpha + n_T}{\alpha + \beta + n} = \frac{\alpha}{\alpha + \beta}\left(\frac{\alpha + \beta}{\alpha + \beta + n}\right) + \frac{n_T}{n}\left(\frac{n}{n + \alpha + \beta}\right),$$

which is the equality from the problem's thesis. Let us note, that if we set $\frac{\alpha}{\alpha + \beta} = m$, then

$$\lim_{\alpha + \beta \to \infty} \mathbb{E}\big(\Theta \,\big|\, \underline{X} = \underline{x}\big) = m,\ \lim_{n \to \infty}\Big(\mathbb{E}\big(\Theta \,\big|\, \underline{X} = \underline{x}\big) - \hat{p}\Big) = 0.$$

From this, when choosing parameters α and β of the prior distribution, we decide about the location of the expected value in such a way, that the greater $\alpha + \beta$ the smaller is the influence of the direct data in the posterior distribution. Howsoever, with the increase of the size of the sample, the influence of the prior expected value on the posterior expected value is smaller.

Lemma 3 *If a random variable* ξ *has a distribution* $\beta(\alpha, \beta)$, *then the random variable* $1 - \xi$ *has distribution* $\beta(\beta, \alpha)$:

Proof. Let us denote $\eta = 1 - \xi$. Then

$$f_\eta(\theta) = f_{\xi}(1 - \theta),\ \theta \in (0,1).$$

From this

$$f_\eta(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cdot\Gamma(\beta)}(1-\theta)^{\alpha-1}(\theta)^{\beta-1},\ \theta \in (0,1),$$

which gives the lemma's thesis.

Instytut Informatyki Uniwersytetu Gdańskiego

REFERENCES

[1] Bickel P.J., Doksum K.A., [2001], *Mathematical Statistics. Basic Ideas and Selected Topics. Vol. I*, Prentice Hall, Upper Saddle River, New Jersey, 07458, ii edition.
[2] Corstange D., [Sep. 2-5 2004], Sensitive Questions, Truthful Responses? Randomized Response and Hidden Logit as a Procedure to Estimate It. Annual Meeting of the American Political Science Association.
[3] DeGroot M.H., [1981], *Optymalne decyzje statystyczne*, Państwowe Wydawnictwo Naukowe, Warszawa.
[4] Gelman A., Carlin J.B., Stern H.S., Rubin D.B., **[2000],** *Bayesyan Data Analysis*, CHAPMAN & HALL=CRC, Boca Raton, Lndyn, New York, Washington D.C., II edition.
[5] van der Heijden P., van Gils G., Bouts J., Hox J., [2000], A Comparison of Randomized Response, Computer-Assisted Self-Interview, and Face-to-Face Direct Questioning, Sociological Methods and Research, 28(4):505-537.
[6] Kim J.-M., Heo T.-Y., [2005], A Bayesian Analysis of the Multinomial Randomized Response Model using Dirichlet Prior Distribution. Proceedings for the Spring Conference 2005 Korean Statistical Society, pages 239-244.
[7] RMG (RiskMetrics Group), [1999], *Risk Management: A Practical Guide*, Risk-Metrics Group, First edition.
[8] Sing S., Horn S., Singh R., Mangat N.S., [2003], *On the use of modified randomization device for estimating the prevalence of a sensitive attribute*, Statistics in Transition, 6(4):515-522.
[9] Szreder M., [2002], *Badania opinii*, Wydawnictwo Wyższej Szkoły Zarządzania, Gdańsk.
[10] Warner S., **[1965],** *Randomized response: a survey technique for eliminating evasive answer bias*, „Journal of American Statistical Association", 60:63-69.
[11] Yang R., Berger R., [1997], *A Catalog of Noninformative Priors*, Plann. Infer., 79:223-235.

Praca wpłynęła do redakcji w kwietniu 2009 r.

TECHNIKI BAYESOWSKIE W PROCEDURACH LOSOWANIA ODPOWIEDZI

Streszczenie

Zagadnienie prawdziwości odpowiedzi na drażliwe społecznie lub osobiście pytania ankieterów pojawiło się wraz z początkiem badań ankietowych. Często respondenci próbują uniknąć odpowiedzi lub odpowiadają niezgodnie z prawdą. Powstało wiele technik badawczych i analitycznych korygujących wyniki badań ankietowych pozwalających na ocenę rzeczywistej opinii respondentów. W pracy przedstawiono bayesowską analizę techniki Randomize Response. Zaprezentowano argumentację wyboru różnych rozkładów *a priori* zarówno nieinformujących, jak i rozkładów kompensujących informacje wstępne opisane rozkładami parametrycznymi.

Słowa kluczowe: Techniki bayesowskie, randomizowane odpowiedzi, model Warnera

BAYESIAN TECHNIQUES IN RANDOMIZED RESPONSE

Summary

The issue of truthfulness of the answers given to socially or personally sensitive questions appeared at the very beginning of the construction of the surveys. The respondents often try to avoid answering a sensitive question or they do not tell the truth. Various research and analysis techniques have been developed to adjust the surveys results in order to assess the real opinion of the respondents. The work introduces a bayesian analysis of the Randomized Response technique. It presents arguments for choosing various a priori distributions, both non-informative and compensating for the preliminary information described by parametrized distributions.

Key words: Bayesian techniques, randomized response, Warner model

PRZEGLĄD STATYSTYCZNY
R. LVI – ZESZYT 2 – 2009

BOGUSŁAW GUZIK

POSTULAT KOINCYDENCJI NAKŁADÓW I REZULTATÓW W BADANIACH EMPIRYCZNYCH DEA

1. WSTĘP

Przystępując do ustalania efektywności obiektów gospodarczych czy społecznych za pomocą metod *Data Envelopment Analysis* (DEA) trzeba, między innymi, określić listę rezultatów oraz listę nakładów, w sensie których oceniana będzie efektywność obiektów. Poza pojawiającymi się niekiedy oczywistymi sformułowaniami typu, że w charakterze nakładów należy brać te wielkości, które wykazują wyraźny związek merytoryczny z rezultatami, w literaturze na ogół brak jest jakichś sprecyzowanych sugestii w tym zakresie. Zupełnie odmienna sytuacja ma miejsce w ekonometrii, gdzie problem doboru zmiennych objaśniających jest znany i poświęcono mu wiele miejsca. Jak wiadomo bardzo wartościowe wyniki uzyskali tu uczeni polscy, np. Hellwig oraz Czerwiński[1].

W literaturze dotyczącej DEA na ogół twierdzi się też, że rezultaty mają charakter stymulant, co znaczy, że ich wzrost oceniany jest korzystnie, a nakłady mają charakter destymulant – co znaczy, że ich wzrost oceniany jest niekorzystnie. W artykule dyskutujemy z tą, naszym zdaniem nazbyt ogólnikową, tezą.

W niniejszej pracy formułujemy postulaty dotyczące ogólnych relacji między nakładami a rezultatami w modelach DEA. Wprowadzamy też kryteria zgodności (= *koincydencji*) nakładów z rezultatami. Postulat koincydencja w modelach DEA rozpatrujemy w innym sensie niż zaproponowany przez Z. Hellwiga[2] postulat dotyczący modeli ekonometrycznych.

2. POSTULAT KOINCYDENCJI NAKŁADÓW I REZULTATÓW

Niech X_n będą zmiennymi reprezentującymi wielkości nakładów $n = 1, \ldots, N$; natomiast Y_r niech będą zmiennymi reprezentującymi wielkości rezultatów $r = 1, \ldots, R$. W *Data Envelopment Analysis* na ogół przyjmuje się, co też zakładamy w całym artykule, że zmienne charakteryzujące rezultaty są *stymulantami*. Oznacza to, że większa wartość zmiennej Y_r oznacza większy poziom rezultatu r-tego rodzaju i jest tym korzystniej, im zmienna Y_r ma większą wartość. Natomiast w odniesieniu do nakładów, jak już wspomniano we wstępie, przyjmuje się, że mają one charakter *destymulant*, czyli

[1] Hellwig [7], Czerwiński [4].
[2] Hellwig [8].

takich zmiennych, że coraz ich większa wartość oceniana jest niekorzystnie, a coraz mniejsza – korzystnie.

Ostatnie określenie budzi jednak wątpliwości. Jest zrozumiałe, że chcielibyśmy minimalizować nakłady, bo to jest korzystne. Niemniej nie można rozumieć tego bezwarunkowo, jako sugestię, że każdy wzrost nakładów jest niekorzystny, a każdy spadek jest korzystny, i że dla uznania pewnej wielkości za nakład trzeba, aby jej spadek oceniany był korzystnie. Nie trzeba bowiem przekonywać, że spadek ów może być związany z jeszcze większym spadkiem rezultatów, co właśnie będzie niekorzystne. Z drugiej strony wzrost nakładu nie może być *a priori* uznany za niekorzystny, gdyż warunkiem wzrostu rezultatów na ogół jest wzrost nakładów. Dlatego „literaturowe" określenie nakładu jako destymulanty jest niewystarczające.

Za podstawową cechę nakładu należy uznać, że – oprócz oczywistego merytorycznego lub co najmniej symptomatycznego związku z rezultatami – jest to wielkość, która zmienia się *w tym samym kierunku* co rezultaty.

Postulat 1

– Jeśli rezultaty mając charakter stymulant, to zmienną X_i możemy uznać za *nakład*, gdy – w warunkach *ceteris paribus* – dla wzrostu rezultatów konieczny jest wzrost zmiennej X_i.

– Warunki *ceteris paribus* oznaczają, że wszystkie pozostałe nakłady X_n ($n = 1, \ldots, N$; $n \neq i$) są ustalone. Ustalona jest też technologia przekształcania nakładów w rezultaty.

Postulat 1 stwierdza, że jeśli nic się nie zmienia, to wzrost rezultatu może być osiągnięty jedynie poprzez wzrost nakładu. Przez symetrię do niego można byłoby sugerować, że zmienna X_i traktowana jako nakład, musi też mieć własność, że – w warunkach *ceteris paribus* – spadkowi rezultatów odpowiada spadek zmiennej X_i. Byłoby to jednak żądanie zbyt daleko idące, bowiem przyczyną spadku rezultatów niekoniecznie musi być zmniejszenie nakładów. Może nią być np. *rozrzutność* (*marnotrawstwo*) nakładów, czyli stabilizacja lub wzrost nakładów przy malejących rezultatach. Ponadto postulat spadku nakładów, gdy rezultaty maleją oraz wzrostu nakładów, gdy rosną rezultaty przesądzałby od razu, że rozpatrujemy tylko sytuacje, gdy każdy nakład jest efektywny w sensie Pareto, czego – oczywiście – tu nie czynimy[3].

Nie będziemy przyjmowali tego rozszerzonego postulatu. Ograniczymy się tylko do postulatu 1. Żąda on zgodności dodatniego kierunku zmian nakładów i rezultatów i dlatego nazwiemy go postulatem *dodatniej koincydencji kierunków zmian* nakładów i rezultatów, krócej: postulatem *koincydencji nakładów i rezultatów*.

Postulat 2

– W modelu DEA wszystkie zmienne reprezentujące nakłady muszą być koincydentne z rezultatami.

[3] Naturalnie, z matematycznego punktu widzenia obie te sytuacje (wzrost X – wzrost Y; spadek X – spadek Y) są równoprawne. Niemniej z ekonomicznego punktu widzenia są one różne. O ile w warunkach *ceteris paribus* wzrost rezultatów wymaga wzrostu nakładów, o tyle spadek rezultatów może mieć miejsce bez spadku nakładów. Oznacza to swego rodzaju „półefektywność" Pareto.

3. ILUSTRACJA EMPIRYCZNA SKUTKÓW BRAKU KOINCYDENCJI

Zilustrujemy na przykładzie liczbowym niektóre niekorzystne konsekwencje braku koincydencji nakładów i rezultatów w modelu DEA. Wykorzystamy ukierunkowany na nakłady model nadefektywności (*super-efficiency*) CCR, który kodujemy jako SE-CCR. Jak wiadomo daje on to samo co model CCR, a dodatkowo pozwala na rangowanie obiektów oraz przeprowadzanie wielu wartościowych analiz ekonomicznych[4].

3.1. DANE EMPIRYCZNE

W tabeli 1 podano dane empiryczne dotyczące dwóch rezultatów Y_1, Y_2 oraz czterech zmiennych wstępnie uznanych za nakłady: X_1, X_2, X_3, X_4 w 10 obiektach. Dane są uporządkowane w kolejności rosnącej obu rezultatów.

Tabela 1

Dane empiryczne

Wyszczególnienie	O_1	O_2	O_3	O_4	O_5	O_6	O_7	O_8	O_9	O_{10}
Y_1	100	103	141	181	220	223	250	298	299	300
Y_2	50	80	100	105	118	119	130	140	170	180
X_1	100	145	170	180	204	206	218	230	241	400
X_2	300	311	317	318	325	326	330	340	346	500
X_3	20	22	24	28	31	35	38	42	48	60
X_4	100	88	85	70	65	51	50	18	17	10

Źródło: dane umowne.

Korelacje zmiennych reprezentujący nakłady wynoszą:

Zmienne	X_1	X_2	X_3	X_4
X_1	1	0,943	0,944	-0,846
X_2	0,943	1	0,844	-0,681
X_3	0,944	0,844	1	-0,956
X_4	-0,846	-0,681	-0,956	1

[4] Model SE-CCR jest stosunkowo prostym uogólnieniem podstawowego dla DEA modelu CCR Charnesa, Coopera i Rhodesa [3]. Model SE-CCR zaproponowano w pracach Bankera i Gilforda [2] oraz Andersena-Petersena [1]. Opis modelu SE-CCR i jego własności – zob. np. Guzik [5]. Samo sformułowanie modelu SE-CCR podamy za chwilę.

Korelacje zmiennych X nie są równe 1, co jest ważne, gdyż nakłady skorelowane w stopniu 1 można reprezentować jednym nakładem, a wyniki modelu SE-CCR nie ulegną zmianie. Korelacje nie są też bliskie zeru, co odpowiada typowym sytuacjom badań empirycznych.

Jak można zauważyć, w ślad za wzrostem rezultatów Y_1, Y_2, rosną również nakłady X_1, X_2, X_3. Zmienne te są więc koincydentne z rezultatami. Natomiast wartości zmiennej X_4 – wbrew postulatowi koincydencji – maleją. Zmienna ta jest niekoincydentna względem rezultatów.

3.1. UKIERUNKOWANE NA NAKŁADY STANDARDOWE ZADANIE SE-CCR

Dla każdego obiektu mamy osobne zadanie SE-CCR. Zadanie dla obiektu o ($1 \leq o \leq J$) ma następującą postać:

Dane

– Lista obiektów $j = 1, \ldots, J$; lista nakładów $n = 1, \ldots, N$ oraz lista rezultatów $r = 1, \ldots, R$.

– Wielkości rezultatów w poszczególnych obiektach, y_{rj} ($r = 1, \ldots, R$; $j = 1, \ldots, J$).

– Wielkości nakładów w poszczególnych obiektach, x_{nj} ($n = 1, \ldots, N$; $j = 1, \ldots, J$).

Wszystkie rezultaty są nieujemne, a nakłady są dodatnie. Przez K_o oznaczamy zbiór wszystkich obiektów oprócz obiektu o-tego, czyli $K_o = \{1,\ldots, J\}\setminus\{o\}$.

Zmienne decyzyjne

ρ_o – *mnożnik poziomu nakładów obiektu o-tego,*

λ_{oj} ($j \in K_o$) – *wagi intensywności.*

Mnożnik ρ_o określa krotność nakładów obiektu o-tego, jaką muszą zastosować pozostałe obiekty (= *konkurenci technologiczni* obiektu o-tego), aby uzyskać przynajmniej jego rezultaty. Waga intensywności λ_{oj} określa krotność technologii empirycznej obiektu j-ego występującą w technologii wspólnej konkurentów obiektu o-tego.

Funkcja celu

$$\min \rho_o \tag{1}$$

Warunki ograniczające i znakowe

$$\sum_{j \in K_o} \lambda_{oj} y_{rj} \geq y_{ro} \quad (r = 1, \ldots, R); \tag{2}$$

$$\sum_{j \in K_o} \lambda_{oj} x_{nj} \leq \rho_o x_{no} \quad (r = 1, \ldots, N); \tag{3}$$

$$\rho_o, \lambda_{oj} \geq 0 \quad (j \in K_o). \tag{4}$$

Uzyskany na podstawie modelu SE-CCR optymalny mnożnik nakładów ρ_o jest *wskaźnikiem rankingowym* obiektu *o*-tego. Określa on, jaką przynajmniej krotność nakładów obiektu *o*-tego musiałyby wykorzystać pozostałe obiekty w swej optymalnej technologii wspólnej, aby otrzymać rezultaty nie gorsze od uzyskanych przez obiekt *o*-ty. Obiekt jest w pełni efektywny w sensie Farrela, gdy $\rho_o \geq 1$, a więc gdy inni – nawet w swej optymalnej technologii wspólnej – dla uzyskania rezultatów otrzymanych przez obiekt *o*-ty, muszą poczynić nakłady nie mniejsze od poniesionych przez obiekt *o*-ty. W przypadku $\rho_o < 1$ obiekt *o*-ty jest nieefektywny, bowiem inne obiekty dla uzyskania rezultatów obiektu *o*-tego potrzebują mniej nakładów niż ich zużył obiekt *o*-ty.

3.3. WYNIKI EMPIRYCZNE

Niech **X** oznacza ustaloną wcześniej listę zmiennych koincydentnych. Ponadto niech X_d oznacza pewną zmienną koincydentną nie występującą w liście **X**, zaś X^{nk} niech oznacza zmienną niekoincydentną.

Dla zbadania skutków braku koincydencji porównamy rozwiązanie modelu SE-CCR otrzymywane przy liście nakładów $\mathbf{X} \cup X_d$ zawierającej tylko zmienne koincydentne, z rozwiązaniem otrzymywanym przy „mieszanej" liście nakładów $\mathbf{X} \cup X^{nk}$, zawierającej zmienne koincydentne oraz zmienną niekoincydentną X^{nk}. W naszym przykładzie zmienną niekoincydentną jest X_4. Obie listy, $\mathbf{X} \cup X_d$ oraz $\mathbf{X} \cup X^{nk}$ są rozszerzeniem listy **X** o jeden element.

W tabeli 2 przedstawiono, wynikające z optymalnego rozwiązania modelu SE-CCR, optymalne mnożniki rankingowe dla poszczególnych wariantów listy zmiennych reprezentujących nakłady. Obok kolumny dla listy **X** podano kolumnę dla listy $\mathbf{X} \cup X_4$. Lista rezultatów jest stała i obejmuje Y_1 oraz Y_2.

Tabela 2

Optymalne mnożniki rankingowe dla różnych list zmiennych reprezentujących nakłady

Obiekt	Kombinacje zmiennych reprezentujących nakłady													
	1	1,4	2	2,4	3	3,4	1,2	1,2,4	1,3	1,3,4	1,2,3	1,2,3,4	1,2,3	1,2,3,4
O_1	0,782	0,782	0,382	0,382	0,705	0,705	0,782	0,782	0,782	0,782	0,782	0,782	0,782	0,782
O_2	0,782	0,782	0,524	0,524	0,873	0,873	0,782	0,782	0,905	0,905	0,905	0,905	0,905	0,905
O_3	0,834	0,834	0,642	0,642	1,095	1,095	0,834	0,834	1,095	1,095	1,095	1,095	1,095	1,095
O_4	0,827	0,827	0,672	0,672	0,959	0,959	0,827	0,827	0,974	0,974	0,974	0,974	0,974	0,974
O_5	0,857	0,857	0,780	0,780	1,058	1,058	0,857	0,857	1,058	1,058	1,058	1,058	1,058	1,058
O_6	0,859	0,859	0,788	0,788	0,898	0,924	0,859	0,859	0,935	0,935	0,935	0,935	0,935	0,935
O_7	0,905	0,905	0,871	0,871	0,927	0,945	0,905	0,905	0,961	0,961	0,961	0,961	0,961	0,961

cd. tabeli 2

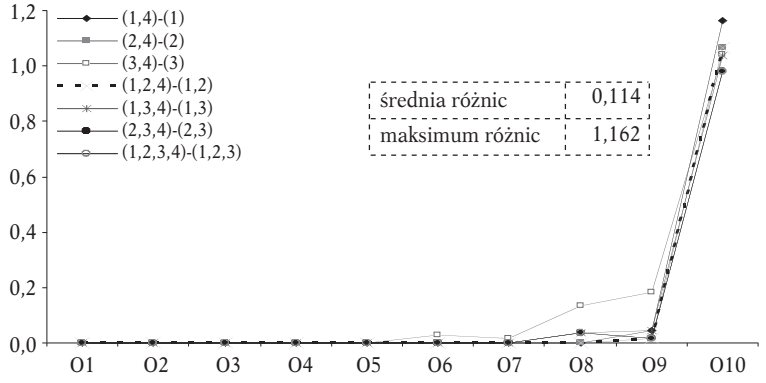
Obiekt	Kombinacje zmiennych reprezentujących nakłady													
	1	1,4	2	2,4	3	3,4	1,2	1,2,4	1,3	1,3,4	1,2,3	1,2,3,4	1,2,3	1,2,3,4
O_8	1,044	1,044	1,014	1,014	1,000	1,132	1,044	1,044	1,095	1,132	1,098	1,132	1,098	1,132
O_9	1,159	1,204	1,193	1,211	0,912	1,094	1,193	1,211	1,159	1,204	1,193	1,211	1,193	1,211
O_{10}	0,638	1,800	0,733	1,800	0,758	1,800	0,733	1,800	0,763	1,800	0,820	1,800	0,820	1,800

Źródło: obliczenia własne zmienna nr 4 – zmienna niekoincydentna.

Uwaga

Porównując kolumny tabeli 2 trzeba pamiętać o fakcie matematycznym, że każde rozszerzenie listy nakładów lub rezultatów skutkuje tym, że współczynniki rankingowe nie maleją, czyli rosną lub pozostają na poprzednim poziomie. Wynika to z własności zadania programowania liniowego, którego szczególnym przypadkiem jest model SE-CCR[5]. Dlatego też współczynniki rankingowe dla listy **X** są nie większe od współczynników dla listy $\mathbf{X} \cup \mathbf{D}$, gdzie **D** – lista zmiennych „dodanych". W szczególności więc, jeśli do listy zmiennych koincydentnych (nawet bardzo dużej) dołączymy tylko jedną zmienną niekoincydentną, czyli o przeciwnym charakterze niż dotychczasowe, to niektóre współczynniki rankingowe i tak wzrosną, a pozostałe nie ulegną zmianie.

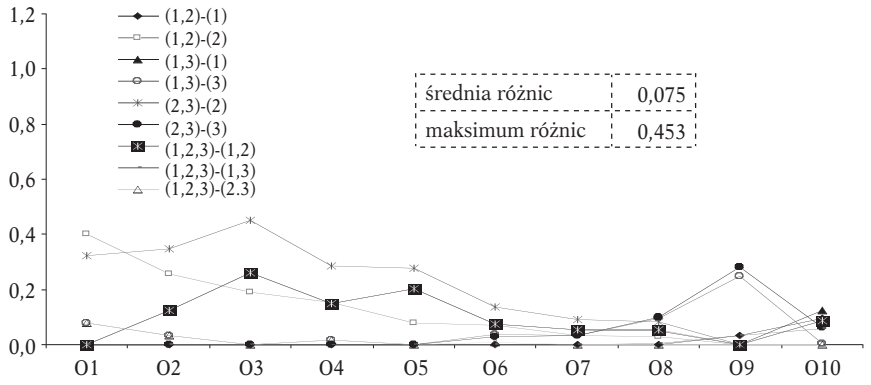
Na rys. 1 zilustrowano kształtowanie się różnic wskaźników rankingowych dla kombinacji niekoincydentnej $\mathbf{X} \cup X_4$ oraz kombinacji koincydentnej **X**.



Rysunek 1. Różnice między wskaźnikami rankingowymi dla kombinacji $\mathbf{X} \cup X_4$ oraz **X**

Natomiast na rys. 2 zilustrowano kształtowanie się różnic wskaźników rankingowych kombinacji koincydentnej $\mathbf{X} \cup X_d$ oraz kombinacji **X**. Oczywiście $X_d \notin \mathbf{X}$. Oba rysunki wykonano w tej samej skali.

[5] Zob. np. Guzik [5].



Rysunek 2. Różnice między wskaźnikami rankingowymi dla kombinacji $\mathbf{X} \cup X_d$ oraz $\mathbf{X}$

Przykładowe wnioski

1. Nie ma jakiejś jednej matematycznej reguły dotyczącej zmian wskaźnika rankingowego w przypadku dołączenia zmiennej niekoincydentnej do zbioru zmiennych koincydentnych[6]. Przyrosty wskaźnika rankingowego na skutek dołączenia do listy zmiennych koincydentnych pewnej innej zmiennej niekoincydentnej mogą być mniejsze lub większe od przyrostu wskaźnika rankingowego na skutek dołączenia zmiennej koincydentnej do listy zmiennych koincydentnych.

2. Ogólnie jednak – jak pokazuje ta i wiele innych analiz empirycznych – dołączenie zmiennej niekoincydentnej wprowadza większe niestabilności kształtowania się współczynników rankingowych oraz większe ich zróżnicowanie co do skali i – na ogół – większe zróżnicowanie co do średniej (por. skalę zmian na rys. 1 oraz rys. 2 i podane tam średnie zmian).

3. Przykład potwierdza to, co można było przypuszczać. Dla obiektów, w których wartość zmiennej niekoincydentnej jest bardzo korzystna (relatywnie duże rezultaty przy małych nakładach) współczynnik rankingowy dla listy uwzględniającej zmienną niekoincydentną jest wyraźnie większy niż w przypadku, gdy zamiast niej występuje zmienna koincydentna (zob. wyniki dotyczące O_8, O_9, O_{10}). Natomiast dla obiektów, w których wartość zmiennej niekoincydentnej jest mało korzystna (wysoka przy niskich nakładach), wskaźnik rankingowy w przypadku listy zawierającej tę zmienną niekoincydentną jest zauważalnie mniejszy niż gdyby zamiast niej występowała zmienna koincydentna (zob. wyniki dla O_1 – O_6).

4. Dołączenie pojedynczej zmiennej koincydentnej nawet do listy kilku zmiennych koincydentnych może spowodować, że obiekty, które do tej pory były niezbyt efektywne, nagle stają się wysoce efektywne. Zachodzi więc zjawisko podobne do zjawiska *katalizy* wykrytego i zbadanego w modelach ekonometrycznych przez Z. Hellwiga[7]. Natomiast inne obiekty, które przy różnych konfiguracjach zmiennych koincydentnych były efektywne, stają się nieefektywne.

[6] Oprócz tej oczywiście, że współczynnik rankingowy nie maleje.
[7] Hellwig [9].

Dla szerszego porównania zmian wskaźników rankingowych wskutek przyłączenia zmiennej niekoincydentnej zbadano jeszcze kształtowanie się wskaźników przy różnych poziomach zmiennej niekoincydentnej X_4 oraz różnych poziomach rezultatu Y_1 w obiekcie O_{10}. Do tej pory w tym obiekcie nakład $X_4 = 10$, rezultat $Y_1 = 300$. Obliczenia dotyczą listy rezultatów $\{Y_1, Y_2\}$ oraz listy zmiennych charakteryzujących nakłady: X_1, X_2, X_4 (dwie pierwsze są koincydentne, trzecia – nie jest). Relatywny zakres zmian rezultatu Y_1 oraz nakładu X_4 jest taki sam, mianowicie jest to 3-krotny spadek rezultatu Y_1 oraz 3-krotny wzrost zmiennej X_4. Wyniki obliczeń zawiera tabela 3.

Tabela 3

Wskaźniki rankingowe dla różnych poziomów zmiennej X_4 oraz Y_1 w obiekcie O_{10}

Zmienna X_4	**10**	10	10	10	10	10	**20**	20	20	20	20	20	**30**	30	30	30	30	30
Zmienna Y_1	300	270	250	200	150	100	300	270	250	200	150	100	300	270	250	200	150	100
O_1	0,78	0,86	0,87	0,87	0,87	0,87	0,78	0,86	0,87	0,87	0,87	0,87	0,78	0,86	0,87	0,87	0,87	0,87
O_2	0,78	0,78	0,78	0,80	0,84	0,86	0,78	0,78	0,78	0,80	0,84	0,86	0,78	0,78	0,78	0,80	0,84	0,86
O_3	0,83	0,83	0,83	0,88	0,91	0,93	0,83	0,83	0,83	0,88	0,91	0,93	0,83	0,83	0,83	0,88	0,91	0,93
O_4	0,83	0,88	0,91	0,94	0,95	0,96	0,83	0,88	0,91	0,94	0,95	0,96	0,83	0,88	0,91	0,94	0,95	0,96
O_5	0,86	0,93	0,95	0,96	0,96	0,96	0,86	0,93	0,95	0,96	0,96	0,96	0,86	0,93	0,95	0,96	0,96	0,96
O_6	0,86	0,93	0,95	0,96	0,96	0,96	0,86	0,93	0,95	0,96	0,96	0,96	0,86	0,93	0,95	0,96	0,96	0,96
O_7	0,91	0,98	1,06	1,12	1,13	1,14	0,91	0,98	1,06	1,12	1,13	1,14	0,91	0,98	1,06	1,12	1,13	1,14
O_8	1,04	1,04	1,04	1,02	1,00	0,94	1,04	1,04	1,04	1,02	1,00	0,95	1,04	1,04	1,04	1,02	1,00	0,95
O_9	1,21	1,21	1,21	1,21	1,21	1,21	1,25	1,25	1,25	1,25	1,25	1,25	1,29	1,29	1,29	1,29	1,29	1,29
O_{10}	1,80	1,80	1,80	1,80	1,80	1,80	0,90	0,90	0,90	0,90	0,90	0,90	0,73	0,73	0,73	0,73	0,73	0,73

Źródło: obliczenia własne.

Przykładowe wnioski

1. Zmiany poziomu zmiennej niekoincydentnej X_4 wywoływały znacznie większe zmiany współczynników efektywności niż tej samej krotności zmiany rezultatu Y_1.

2. Dotyczy to szczególnie obiektów, w których występuje bardzo niski (czyli bardzo korzystny) poziom zmiennej niekoincydentnej (tu są to obiekty O_9 oraz O_{10}). Np. dwukrotny wzrost zmiennej X_4 z poziomu 10 do poziomu 20 spowodował dwukrotne zmniejszenie wskaźnika rankingowego obiektu O_{10}, przy $X_4 = 20$ obiekt O_{10} stał się wręcz nieefektywny. Natomiast dwukrotny spadek rezultatu Y_1 (np. z 300 do 150) przy ustalonym poziomie zmiennej X_4 nie wywoływał żadnej zmiany wskaźnika rankingowego obiektu O_{10}.

3. Odwrotnie jest w wypadku obiektów, dla których zmienna niekoincydentna ma wartość niekorzystną (dużą) – por. np. wyniki dla O_1 – O_6. W tym wypadku ich

efektywność poprawiała się tylko w miarę spadku rezultatu Y_1 w obiekcie O10 oraz – przy danym poziomie Y_1 – wręcz nie zmieniała się wraz ze zmianami poziomu zmiennej X_4 w obiekcie O10[8].

Niekoincydentność zmiennej reprezentującej nakłady jest z jednej strony anormalnym zjawiskiem ekonomicznym, gdyż uzyskiwanie coraz większych rezultatów kosztem coraz mniejszych nakładów dystansuje nawet znane z ekonomii matematycznej pojęcie „rogu obfitości"[9]. Z drugiej strony, na podstawie przeprowadzonych powyżej symulacji widać, że zmienne niekoincydentne deformują obraz efektywności obiektów. Trzeba więc takie zmienne usuwać z zadań DEA. W następnej części artykułu sformułowano niektóre metody rozpoznawania niekoincydentych zmiennych reprezentujących nakłady.

4. MIERNIKI KOINCYDENCJI

Podamy propozycje dwóch klas mierników koincydencji (= *zgodności*). Pierwsze oparte są na badaniu monotoniczności wzrostu i dlatego mówią one o koincydencji (zgodności) *uporządkowań*. Drugi rodzaj mierników to mierniki oparte na badaniu korelacji, mówią one o koincydencji (zgodności) *korelacyjnej*.

4.1. MIERNIK „SILNEJ" KOINCYDENCJI UPORZĄDKOWAŃ

Procedura jest prosta:

1. Dla danej zmiennej X_n ustalamy, jak często w ślad za wzrostem lub stabilizacją rezultatu Y_r, wzrastają lub nie ulegają zmianie wartości zmiennej X_n.

2. Miernik silnej koincydencji uporządkowań zmiennej X_n z rezultatem Y_r wynosi:

$$S_{n,r} = \frac{1}{J-1} \sum_{j=2}^{J} s_{n,r}^{j} \tag{5}$$

gdzie:

$$s_{nr}^{j} = \begin{cases} 1 \text{ gdy znak różnicy } (x_{n,j} - x_{n,j-1}) \text{ jest taki sam jak znak } (y_{r,j} - y_{r,j-1}) \\ 0 \text{ w przeciwnym przypadku} \end{cases} \tag{6}$$

$$j = 2, \ldots, J.$$

Jeśli jedna z wymienionych różnic jest zerowa, przyjmujemy, że ma miejsce zgodność, a więc przyjmujemy $s_{nr}^{j} = 1$.

---

[8] Dla uniknięcia nieporozumień dodać trzeba, że cały czas mówimy o efektywności *względnej*, czyli efektywności *na tle* badanego zbioru obiektów. Dlatego spadek rezultatu Y_1 w obiekcie O_{10} może powodować względnie lepszą sytuację w innych obiektach.

[9] *Róg obfitości* to otrzymanie choćby jakiegoś dodatniego rezultatu przy zerowych nakładach, zob. np. Panek [10], s. 71. Tu rezultat może być dowolnie duży, i w dodatku tym większy, im nakład jest mniejszy.

Definicja 1

Zmienna X_n jest koincydentna z rezultatem r-tym w sensie silnej zgodności uporządkowań, gdy częstość $S_{n,r}$ jest dostatecznie wysoka, a więc gdy:

$$S_{n,r} \geq \gamma \text{ gdzie } \gamma \leq 1 \text{ jest liczbą dostatecznie bliską 1, np. 0,90.} \quad (7)$$

Definicja 2

Zmienna X_n jest, w sensie silnej zgodności uporządkowań, koincydentna z układem rezultatów, Y_r, $r \in U$, gdy jest ona koincydenta w sensie silnej zgodności uporządkowań z wszystkimi rezultatami tego układu, czyli gdy:

$$S_{n,r} \geq \gamma \text{ dla wszystkich } r \in U. \quad (8)$$

Naturalnie, można przyjąć, że $\gamma = 1$. Jest to jednak postulat dość ostry. W praktyce bowiem zdarzają się nietypowe obserwacje, dla których incydentalnie postulat koincydencji nie jest spełniony. Dlatego dopuszczamy, że dla pewnej małej liczby obiektów, np. 5 czy 10%, postulat koincydencji może nie być spełniony.

Przykład

W tabeli 4 podano informacje o kształtowaniu się trzech rezultatów i czterech nakładów.

Tabela 4

Dane empiryczne

Wyszczególnienie	O_1	O_2	O_3	O_4	O_5	O_6	O_7	O_8	O_9	O_{10}
Y_1	4736	3262	33882	5000	146	95,4	500	500	25000	30000
Y_2	239	133	406	7,78	18,8	7,37	54,6	0,09	20,8	1,76
Y_3	55,4	54,6	50,9	53,8	57,0	55,9	52,8	53,1	53,6	59,0
X_1	19836	18714	11077	15935	8172	6879	19764	13223	9245	3866
X_2	626	466	1367	98,7	13,0	12,1	242	57,2	5,5	4,6
X_3	13292	8762	51182	5881	1143	708	5449	346	1547	1283
X_4	4276	2003	35473	2500	249	131	1784	367	355	372

Źródło: dane umowne.

Należy ustalić, czy zmienne proponowane jako nakłady są koincydentne w sensie silnej koincydencji uporządkowań przy $\gamma = 0{,}90$.

Tabela 5

Znaki przyrostów

Wyszczególnienie	O_1	O_2	O_3	O_4	O_5	O_6	O_7	O_8	O_9	O_{10}
Y_1		–1	1	–1	–1	–1	1	0	1	1
Y_2		–1	1	–1	1	–1	1	–1	1	–1
Y_3		–1	–1	1	1	–1	–1	1	1	1
X_1		–1	–1	1	–1	–1	1	–1	–1	–1
X_2		–1	1	–1	–1	–1	1	–1	–1	–1
X_3		–1	1	–1	–1	–1	1	–1	1	–1
X_4		–1	1	–1	–1	–1	1	–1	–1	1

Źródło: obliczenia własne.

Znaki przyrostów zmiennych pomiędzy sąsiednimi obiektami zawiera tabela 5. Występujący we wzorze (5) sygnał koincydencji silnej $s_{nr}^{j} = 1$, gdy iloczyn odpowiednich podanych w tabeli 5 liczb wynosi 1 lub zero.

Wartości miernika silnej koincydencji uporządkowań podano w tabeli 6.

Tabela 6

Mierniki silnej koincydencji uporządkowań

Zmienne	Y_1	Y_2	Y_3
X_1	5/9	5/9	4/9
X_2	7/9	7/9	2/9
X_3	8/9	8/9	3/9
X_4	8/9	6/9	3/9

Wnioski

W sensie miernika silnej koincydencji uporządkowań (przy $\gamma = 0{,}90$) z rezultatem Y_1 koincydentne są zmienne X_3 oraz X_4, a z rezultatem Y_2 – koincydenta jest zmienna X_3.

Nie ma takiej listy zmiennych X (choćby jednoelementowej), która byłaby koincydentna ze wszystkimi rezultatami. Tak więc żadna ze zmiennych $X_1, \ldots, X_4$ nie powinna być tak traktowana jako nakład przy badaniu efektywności uzyskiwania całego układu rezultatów Y_1, Y_2, Y_3 przez obiekty $O_1 - O_{10}$.

Kierując się postulatem koincydencji można badać efektywność co najwyżej układu zwierającego dwa rezultaty: Y_1 i Y_2 oraz jeden nakład – X_3.

4.2. MIERNIK „SŁABEJ" KOINCYDENCJI UPORZĄDKOWAŃ

Konstruując miernik silnej koincydencji uporządkowań przyjmowano, że przyrost nakładu jest koincydentny z przyrostem rezultatu, gdy ich znaki są identyczne. W sferze zjawisk społeczno-gospodarczych często jednak należy liczyć się z różnego rodzaju błędami lub nieokreślonością. Wobec tego moglibyśmy dopuścić, że niewielkie odchylenia jeszcze kategorycznie nie świadczą o braku koincydencji[10].

Procedura polegałaby na następującym

1. Ustalamy relatywną (δ_n) lub bezwzględną (Δ_n) wielkość nieodróżnialnego od zera (czyli nieistotnego) przyrostu zmiennej X_n.

2. Dalsze postępowanie przebiega jak w 4.1. z tym, że zamiast zdefiniowanego wzorem (6) wskaźnika monotoniczności bierzemy pod uwagę wskaźnik „przedziałowy":

$$w_{nr}^{j} = \begin{cases} 1 \text{ gdy } \left(x_{n,j} - x_{n,j-1}\right) + \Delta_n \geq 0, \text{ o ile przyrost } \left(y_{r,j} - y_{r,j-1}\right) \geq 0 \\ 1 \text{ gdy } \left(x_{n,j} - x_{n,j-1}\right) - \Delta_n \leq 0, \text{ o ile przyrost } \left(y_{r,j} - y_{r,j-1}\right) \leq 0 \\ 0 \text{ w przeciwnym przypadki; } j = 2, \ldots, J \end{cases} \tag{9}$$

3. Miernik słabej koincydencji uporządkowań określony jest jako:

$$W_{n,r} = \frac{1}{J-1} \sum_{j=2}^{J} w_{n,r}^{j}. \tag{10}$$

Definicja 3

Zmienna X_n jest koincydenta z rezultatem Y_r w sensie słabej zgodności uporządkowań, gdy

$$W_{n,r} \leq \gamma. \tag{11}$$

Definicja 4

Zmienna X_n jest, w sensie słabej zgodności uporządkowań, koincydentna z układem rezultatów $\{Y_r, r \in U\}$, gdy jest koincydenta z wszystkimi rezultatami układu w sensie słabej zgodności uporządkowań:

$$W_{n,r} \leq \gamma \text{ dla wszystkich } r \in U. \tag{12}$$

Wartość Δ_n może być ustalona bezpośrednio lub może wynikać z przyjętej wielkości nieodróżnialnego od zera relatywnego przyrostu δ_n, np. $\Delta_n = \delta_n x_{n,j}$.

Przykład

Przyjmiemy, że przyrost zmiennej X_n niewiększy od 5% wartości tej zmiennej w obiekcie $j-1$ może być traktowany jako przyrost nieodróżnialny od zerowego. W tabeli 7 podano przyrosty zmiennych oraz wielkości $\Delta_n = 0{,}05x_{nj}$

[10] Ogólniejsze ujęcie, naturalnie, prowadzi do zbiorów rozmytych.

Tabela 7

Przyrosty rezultatów i nakładów oraz wielkości Δ

Wyszczególnienie	O_1	O_2	O_3	O_4	O_5	O_6	O_7	O_8	O_9	O_{10}
Y_1		–1474	30620	–28882	–4854	–50,6	404,6	0	24500	5000
Y_2		–106,0	273,0	–398,2	11,0	–11,4	47,2	–54,5	20,7	–19,0
Y_3		–0,8	–3,7	2,9	3,2	–1,1	–3,1	0,3	0,5	5,4
X_1		–1122	–7637	4858	–7763	–1293	12885	–6541	–3978	–5379
Δ		935	553	796	408	343	988	661	462	193
X_2		–160	901	–1268,3	–85,7	–0,9	229,9	–184,8	–51,7	–0,9
Δ		23,3	68,3	4,9	0,65	0,61	12,1	2,86	0,28	0,23
X_3		–4530	42420	–45301	–4738	–435	4741	–5103	1201	–264
Δ		438	2559	294	57,1	35,4	272	17,3	77,3	64,1
X_4		–2273	33470	–32973	–2251	–118	1653	–1417	–12	17
Δ		100	1773	125	12,5	6,5	89,2	18,3	17,7	18,6

Źródło: obliczenia własne.

Z obliczeniowego punktu widzenia wygodnie będzie wprowadzić dwie wielkości. Nazwijmy „górnym" przyrostem zmiennej X_n wielkość *przyrost zmiennej* $+\Delta$, a „dolnym" przyrostem wielkość *przyrost zmiennej* $-\Delta$[11].

Zmienna X_n jest koincydentna ze zmienną Y_n w sensie słabego miernika zgodności uporządkowań, gdy znak przyrostu rezultatu Y_r jest zgodny ze znakiem bądź „dolnego", bądź „górnego" przyrostu zmiennej X_n.

Znaki „dolnego" oraz „górnego" przyrostu zmiennych reprezentujących nakłady zawiera tabela 8.

Tabela 8

Znaki przyrostów zmiennych

Wyszczególnienie	O_1	O_2	O_3	O_4	O_5	O_6	O_7	O_8	O_9	O_{10}
Y_1		–1	1	–1	–1	–1	1	0	1	1
Y_2		–1	1	–1	1	–1	1	–1	1	–1
Y_3		–1	–1	1	1	–1	–1	1	1	1
$X_{1(\text{dolny})}$		–1	–1	1	–1	–1	1	–1	–1	–1
$X_{1(\text{górny})}$		1	–1	1	–1	–1	1	–1	–1	–1

[11] Górny przyrost odpowiada wartości $(x_{nj} - x_{n,j-1}) + \Delta$, gdy przyrost nakładu jest ujemny, a dolny przyrost odpowiada wartości $(x_{nj} - x_{n,j-1}) - \Delta$, gdy przyrost nakładu jest dodatni.

cd. tabeli 8

Wyszczególnienie	O_1	O_2	O_3	O_4	O_5	O_6	O_7	O_8	O_9	O_{10}
$X_{2(\text{dolny})}$		–1	1	–1	–1	–1	1	–1	–1	–1
$X_{2(\text{górny})}$		–1	1	–1	–1	1	1	–1	–1	–1
$X_{3(\text{dolny})}$		–1	1	–1	–1	–1	1	–1	1	–1
$X_{3(\text{górny})}$		–1	1	–1	–1	–1	1	–1	1	–1
$X_{4(\text{dolny})}$		–1	1	–1	–1	–1	1	–1	–1	–1
$X_{4(\text{górny})}$		–1	1	–1	–1	–1	1	–1	1	1

Źródło: obliczenia własne.

Wartości miernika słabej zgodności uporządkowań podano w tabeli 9.

Tabela 9

Mierniki słabej koincydencji uporządkowań

Zmienne	Y_1	Y_2	Y_3
X_1	5/9	5/9	4/9
X_2	7/9	7/9	2/9
X_3	8/9	8/9	3/9
X_4	9/9	8/9	4/9

Źródło: obliczenia własne.

Wnioski

Słaba koincydencja uporządkowań (przy $\delta = 0{,}1$ oraz $\gamma \geq 0{,}9$) ma miejsce tylko dla rezultatu Y_1 oraz Y_2 względem nakładu X_3 oraz X_4. Ten układ dwóch nakładów i dwóch rezultatów zmiennych można uznać za rozwiązanie problemu ustalenia zbioru nakładów koincydentnych względem rezultatów.

Poprawa w stosunku do miernika silnej koincydencji uporządkowań nastąpiła w wyniku tego, że przy w obiekcie O_9 górna granica, a w obiekcie O_{10} dolna granica ma taki sam znak jak przyrost rezultatu Y_2.

4.3. MIERNIK SŁABEJ KOINCYDENCJI KORELACYJNEJ

Definicja 5

Zmienna X_n reprezentująca nakład n-ty jest koincydentna z rezultatem Y_r w sensie słabej zgodności korelacyjnej, gdy współczynnik korelacji prostej, k_{nr}, między wektorami obserwacji tych wielkości: $\mathbf{x}_n = [x_{nj}]_{j = 1,\ldots, J}$; $\mathbf{y}_r = [y_{rj}]_{j = 1,\ldots, J}$ jest dodatni:

$$k_{nr} > 0. \tag{13}$$

Definicja 6

Zmienna X_n jest koincydentna w sensie słabej zgodności korelacyjnej z układem rezultatów $\{Y_r; r \in U\}$, gdy w tym właśnie sensie jest koincydenta względem wszystkich rezultatów układu, czyli gdy:

$$k_{nr} > 0 \text{ dla wszystkich } r \in U. \tag{14}$$

Przykład

Współczynniki korelacji między scharakteryzowanymi w tabeli 4 rezultatami a zmiennymi reprezentującymi nakłady są następujące (tabela 10):

Tabela 10

Współczynniki korelacji prostej

Zmienne	Y_1	Y_2	Y_3
X_1	-0,472	0,297	-0,488
X_2	0,410	0,987	-0,564
X_3	0,535	0,936	-0,581
X_4	0,580	0,876	-0,582

Źródło: obliczenia własne.

Wnioski

W sensie miernika słabej koincydencji korelacyjnej wszystkie zmienne $X_1, \ldots, X_4$ są koincydentne tylko z rezultatem Y_2. Układ X_2, X_3, X_4 jest natomiast koincydentny z rezultatem Y_1. Żadna ze zmiennych reprezentujących nakłady nie jest koincydentna z rezultatem Y_3.

Ani jedna ze zmiennych $X_n (1 \leq n \leq 4)$ nie jest koincydentna z całym badanym układem rezultatów Y_1, Y_2, Y_3.

4.4. MIERNIK SILNEJ KOINCYDENCJI KORELACYJNEJ

Badanie współczynników korelacji prostej, choć często stosowane przez niestatystyków i nieekonometryków, jest jednak niekompletne, gdy dany rezultat zależy od wielu nakładów. Dlatego bardziej poprawne postępowanie polegałoby na badaniu zależności danego rezultatu od *wszystkich* nakładów, co oznacza badanie cząstkowego wpływu na rezultat Y_r poszczególnych zmiennych X_n na tle całego układu tych zmiennych. Prowadzi to, oczywiście, do koncepcji współczynników *korelacji cząstkowej*.

Niech więc c_{nr} oznacza współczynnik korelacji cząstkowej między zmienną X_n a rezultatem Y_r w sytuacji, gdy rezultat ten zależy od wszystkich zmiennych $X_1, \ldots, X_N$ według równania regresji liniowej:

$$Y_r = \sum_{n=1}^{N} a_{nr} X_n + a_{0r}, \tag{15}$$

gdzie parametry wyznaczono klasyczną mnk.

Jak wiadomo, znak współczynnika korelacji cząstkowej c_{nr} jest taki sam jak znak współczynnika regresji a_{nr}.

1. Postępowanie może polegać na oszacowaniu, dla każdego rezultatu Y_r, klasyczną metodą najmniejszych kwadratów liniowego równania regresji (15).

2. Następnie sprawdzamy, czy znaki współczynników kierunkowych są dodatnie.

Definicja 7

Zmienna X_n ($1 \leq n \leq N$) jest koincydentna w silnym sensie korelacyjnym z nakładem Y_r, gdy

$$a_{nr} > 0 \text{ (lub współczynnik korelacji cząstkowej } c_{nr} > 0). \tag{16}$$

Definicja 8

Zmienna X_n ($1 \leq n \leq N$) jest w sensie silnej zgodności korelacyjnej koincydentna względem układu rezultatów, U, gdy jest w tymże sensie koincydentna ze względu na każdy rezultat układu, a więc, gdy:

$$a_{nr} > 0 \text{ (lub } c_{nr} > 0) \text{ dla wszystkich } r \in U. \tag{17}$$

Uwagi

Jak wiadomo postulat koincydencji w modelach ekonometrycznych, a szerzej – w modelach regresyjnych – głosi, że znak współczynnika korelacji prostej k_{nr} powinien być identyczny ze znakiem współczynnika korelacji cząstkowej c_{nr}. Tu, w odniesieniu do koincydencji korelacyjnej w modelu DEA, takiego postulatu nie stawiamy. Postulat koincydencji DEA głosi mianowicie, że – z osobna – odpowiedni współczynnik korelacji cząstkowej lub korelacji prostej jest *dodatni*.

Jest zrozumiałe, że można proponować silniejszy postulat, zgodny z podejściem spotykanym w ekonometrii, żeby oba współczynniki korelacji (cząstkowej i prostej) jednocześnie były dodatnie[12]. Tego jednak, mając na uwadze zupełnie odmienną interpretację obu współczynników, nie proponujemy.

Dodajmy przy okazji, że ekonometryczny postulat koincydencji jest zachowany, gdy współczynniki korelacji prostej oraz korelacji cząstkowej są dodatnie lub ujemne. W przypadku współczynnika ujemnego, koincydencja korelacyjna w modelu DEA nie ma jednak miejsca.

[12] Co byłoby zgodne z ideą koincydencji w modelach ekonometrycznych.

Warto zaznaczyć, że po odpowiednich modyfikacjach sformułowane tu cztery metody badania koincydencji w modelach DEA, mogłyby być zaadaptowane dla potrzeb modelowania ekonometrycznego. Byłoby to jednak inne rozumienie koincydencji niż dotychczasowe.

Przykład

Na podstawie danych z tabeli 4, oszacowano klasyczną mnk liniowe równania regresji dla rezultatów:

$$Y_r = \sum_{n=1}^{4} a_{nr} X_n + a_{0r}. \tag{18}$$

Wyniki przedstawiono w tabeli 11.

Tabela 11

Parametry liniowych równań regresji (18)

Zmienne	Y_1	Y_2	Y_3
X_1	–1,591	–0,0035	–0,00045
X_2	–46,2	0,35	0,00056
X_3	4,71	0,012	0,00072
X_4	–4,17	–0,013	–0,00116

Źródło: obliczenia własne.

Wnioski

Zmienna X_1 oraz X_4 nie są koincydentne z żadnym rezultatem w sensie silnej zgodności korelacyjnej. Zmienna X_2 jest natomiast koincydentna z drugim i trzecim rezultatem, a zmienna X_3 – z pierwszym, drugim i trzecim.

Kierując się silnym kryterium korelacyjnym należałoby badać efektywność obiektów na podstawie modelu DEA zawierającego rozpatrywane trzy rezultaty oraz jeden nakład – X_3.

5. PODSUMOWANIE

W tabeli 12 zestawiono wyniki omówionych tu czterech kryteriów koincydencji rezultatów i nakładów w modelach DEA. Symbol T oznacza, że ma miejsce koincydencja w odpowiednim sensie.

Tabela 12

Zestawienie wyników badania koincydencji nakładów z rezultatami

Naklad	Y_1				Y_2				Y_3			
	Słabe porządkowe	Silne porządkowe	Słabe korelacyjne	Silne korelacyjne	Słabe porządkowe	Silne porządkowe	Słabe korelacyjne	Silne korelacyjne	Słabe porządkowe	Silne porządkowe	Słabe korelacyjne	Silne korelacyjne
X_1							T					
X_2			T				T	T				T
X_3	T	T	T	T	T	T	T	T				T
X_4	T	T	T		T		T					

Źródło: obliczenia własne.

Wnioski

1. To, że różne mierniki dają różne wskazania nie powinno nas dziwić.

2. Spośród wykorzystanych mierników, w przykładzie najbardziej „liberalny" jest miernik słabej koincydencji korelacyjnej, natomiast najbardziej „radykalny" jest miernik silnej koincydencji uporządkowań. Jest to intuicyjnie zrozumiałe i wydaje się, że na ogół tak będzie w większości problemów empirycznych.

3. Wyniki przykładu sugerują, że zapewnienie koincydencji w modelach DEA może być poważnym problemem empirycznym (podobnie jak uzyskanie dobrego modelu ekonometrycznego).

4. W przykładzie tylko jeden miernik (silny miernik korelacyjny) sugeruje, że możliwe jest określenie wspólnej listy nakładów dla wszystkich rezultatów. Jest to jednak lista bardzo wąska, jednoelementowa (X_3). Ze słabego kryterium korelacyjnego płynie natomiast sugestia, że sensowne jest zbudowanie model DEA tylko dla dwóch rezultatów: Y_1 oraz Y_2, ale natomiast z trzema nakładami: X_2, X_3, X_4. Z kolei według słabego kryterium porządkowego możliwe jest utworzenia zadania DEA dotyczącego dwóch nakładów (Y_1 oraz Y_2) z dwoma nakładami – X_3, X_4.

W każdym z tych wypadków dochodzi do redukcji wymiarów zadania DEA – albo liczby nakładów, albo liczby rezultatów.

5. Postulat koincydencji ma podłoże metodologiczne, bowiem dziwna byłaby ekonomia, w której w celu zwiększenia rezultatów należałoby zmniejszać nakłady. Ma też podłoże empiryczne. Jeśli bowiem brak jest koincydencji nakładów i rezultatów, wyniki są niestabilne i często sugerują zupełnie „przypadkowy" bardzo wysoki poziom efektywności niektórych obiektów.

6. Wydaje się przeto, że badacze posługujący się metodami DEA, powinni – na wzór ekonometrii – mieć większą świadomość metodologiczną. Nawet jeśli układ nakładów i rezultatów sformułowano w najlepszej wierze, należy za każdym razem sprawdzić, czy spełnione są warunki umożliwiające uzyskanie rzetelnych wyników. Jednym z tych warunków jest omawiany w artykule postulat koincydencji nakładów z rezultatami.

7. Dla jego uzyskania może być potrzebna redukcja albo modyfikacja listy nakładów lub listy rezultatów, lub nawet redukcja czy modyfikacja listy obiektów.

Uniwersytet Ekonomiczny w Poznaniu

LITERATURA

[1] Andersen P., Petersen N.C., [1993], *A procedure for ranking efficient units in Data Envelopment Analysis*, „Management Science", 39 (10).
[2] Banker R.D., Gilford J.L., [1988], *A relative efficiency model for the evaluation of public health nurse productivity*, Mellon University Mimeo, Cornegie.
[3] Charnes A., Cooper W.W., Rhodes E., [1978], *Measuring the efficiency of decision making units*, „European Journal of Operational Research".
[4] Czerwiński Z., [1976], *Przyczynek do dyskusji nad problemem „dobrego" modelu ekonometrycznego*, „Przegląd Statystyczny", 4.
[5] Guzik B., [2008], *Model nadefektywności DEA na tle modelu CCR*, „Wiadomości Statystyczne", 1.
[6] Guzik B., *Prosta metoda doboru zestawu nakładów w modelu DEA*, „Przegląd Statystyczny", złożone do druku.
[7] Hellwig Z., [1960], *Problem optymalnego wyboru predykant*, „Przegląd Statystyczny" 3.
[8] Hellwig Z., [1976], *Przechodniość relacji skorelowania zmiennych losowych i płynące stąd wnioski ekonometryczne*, „Przegląd Statystyczny" 1/1985.
[9] Hellwig Z., [1987], *Model ekonometryczny z kompensatorem różnicowym*, „Przegląd Statystyczny", 1.
[10] Panek E., [2003], *Ekonomia matematyczna*, Wyd. AE w Poznaniu, Poznań.

Praca wpłynęła do redakcji w marcu 2009 r.

POSTULAT KOINCYDENCJI NAKŁADÓW I REZULTATÓW W BADANIACH EMPIRYCZNYCH DEA

Streszczenie

W artykule wskazano na konieczność uwzględniania postulatów metodologicznych przy formułowaniu listy nakładów oraz rezultatów w zadaniach DEA. W szczególności sformułowano i uzasadniono postulat koincydencji nakładów i rezultatów, który głosi, że kierunek zmiennej uznanej za nakład musi być zgodny z kierunkiem zmian rezultatów. Zaproponowano cztery mierniki koincydencji dotyczące: silnej i słabej koincydencji uporządkowań oraz silnej i słabej koincydencji korelacyjnej.

Słowa kluczowe: DEA, dobór nakładów, koincydencja nakładów

INPUT AND OUTPUT COINCIDENCE POSTULATE IN DEA EMPIRICAL RESEARCH

Summary

The article shows that the selection of input and output variables to DEA models should be based on methodological postulates. Particularly, the author formulates and justifies the postulate of intput and output coincidence, which means that direction of change in input variable shoud be consistent with direction of change in output variable. Four measures of coincendence are provided: the measure of strong and weak order coincidence and strong and weak correlation coincidence.

Key words: DEA, selection of inputs, coincidence of inputs

PRZEGLĄD STATYSTYCZNY
R. LVI – ZESZYT 2 – 2009

IWONA MARKOWICZ, BEATA STOLORZ

MODEL PROPORCJONALNEGO HAZARDU COXA PRZY RÓŻNYCH SPOSOBACH KODOWANIA ZMIENNYCH

1. WSTĘP

Metody analizy przeżycia są coraz częściej stosowane w badaniach zjawisk społeczno-ekonomicznych[1]. Ze względu na brak konieczności znajomości rozkładu badanej zmiennej losowej szczególną wagę przywiązuje się do modeli nieparametrycznych bądź semiparametrycznych. Coraz powszechniej wykorzystywane są one do badania zjawisk innych niż czas trwania życia ludzkiego. Przeglądu metodologii analizy historii zdarzeń i ich aplikacji do badania czasu funkcjonowania firm autorki artykułu dokonały w ramach realizacji grantu MNiSW (N 111 011 31/1109). Warunkiem stosowania modeli analizy przeżycia jest odpowiednia baza danych umożliwiająca wyznaczenie czasu trwania zdefiniowanego stanu dla poszczególnych jednostek badanej zbiorowości. Zazwyczaj są to badania retrospektywne z wykorzystaniem sporządzanych rejestrów. Przykładem takiej bazy danych jest rejestr bezrobotnych.

Celem artykułu jest wskazanie wpływu sposobu kodowania zmiennych na oszacowania parametrów modelu regresji Coxa i ich interpretację. Autorki przedstawiły również związek między parametrami modelu szacowanymi dla danych zakodowanych w dwojaki sposób. Badaną kohortę stanowią osoby bezrobotne wyrejestrowane w określonym okresie czasu. Podziału na podgrupy dokonano ze względu na wiek, który jest determinantą czasu poszukiwania pracy, co autorki wykazały we wcześniejszych badaniach [7].

2. DANE STATYSTYCZNE WYKORZYSTANE W ANALIZIE

Analiza czasu oczekiwania na pracę została przeprowadzona w oparciu o indywidualne dane o bezrobotnych wyrejestrowanych z Powiatowego Urzędu Pracy w Szczecinie w I kwartale 2007 roku. Uzyskane informacje dotyczyły wieku i powodu wyrejestrowania. Powody wyrejestrowania były różne, natomiast powód, który został uznany za zdarzenie kończące obserwację to podjęcie przez dotychczasowego bezrobotnego pracy. Osoby wyrejestrowane z innych przyczyn, takich jak podjecie nauki, wyjazd za granicę, odmowa przyjęcia propozycji zatrudnienie, niestawienie się w PUP w wyznaczonym terminie, czy osiągnięcie wieku emerytalnego, stanowią obserwacje cenzurowane. Dla

[1] Por. [1], [6], [7].

tej grupy nie można ustalić okresu oczekiwania na pracę. Analizie poddano ogółem 4237 osób. Kategorie wieku zostały pogrupowane według klasyfikacji stosowanej przez PUP. Spośród wszystkich wyrejestrowanych dla 46% osób powodem było znalezienie pracy. Strukturę badanej zbiorowości przedstawiono w tabeli 1.

Tabela 1

Charakterystyka ilościowa badanych bezrobotnych

Cecha			Obserwacje		Razem
			pełne	cenzurowane	
Wiek	⟨18, 25)	(1)	431	569	1000
	⟨25, 35)	(2)	779	719	1498
	⟨35, 45)	(3)	310	369	679
	⟨45, 55)	(4)	361	474	835
	⟨55, 60)	(5)	61	112	173
	⟨60, 65)	(6)	4	48	52
Ogółem			1946	2291	4237

Źródło: opracowanie własne.

3. MODEL PROPORCJONALNEGO HAZARDU COXA[2]

Do zbadania wpływu potencjalnej zmiennej na czas pozostawania w rejestrze bezrobotnych nie można zastosować modeli regresji wielorakiej ze względu na nieznajomość rozkładu zmiennej zależnej oraz występowanie obserwacji cenzurowanych. Model proporcjonalnego hazardu Coxa zakłada, że funkcja hazardu [4] jest funkcją zmiennych niezależnych, którą można zapisać następująco [1]:

$$h(t{:}x_1, x_2, \ldots, x_n) = h_0(t)\exp(\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n) \tag{1}$$

gdzie:

$h(t{:}\ x_1, x_2, \ldots, x_n)$ – wynikowy hazard (szansa) przy danych n zmiennych niezależnych $x_1, x_2, \ldots, x_n$ i odpowiednim czasie przeżycia (oczekiwania),

$h_0(t)$ – hazard (szansa) odniesienia lub zerowa linia hazardu,

$\beta_1, \beta_2, \ldots, \beta_n$ – współczynniki modelu,

t – czas obserwacji.

Bazowa wartość $h_0(t)$ hazardu jest tą wartością hazardu, dla której wszystkie zmienne niezależne są równe zero.

[2] Por. [3].

Wieloczynnikowy model proporcjonalnego hazardu Coxa umożliwia ocenę jednoczesnego wpływu wielu zmiennych na czas do wystąpienia określonego zdarzenia.

4. SPOSOBY KODOWANIA ZMIENNYCH

Ze względu na różne sposoby kodowania zmiennych można obliczyć różne rodzaje ryzyka względnego. W artykule zostaną przedstawione dwa z nich, zgodnie z procedurami przedstawionymi przez Hosmer i Lemeshow [5].

Pierwszy rodzaj kodowania umożliwia wyznaczenie szansy opuszczenia przez bezrobotnego rejestru PUP w stosunku do wybranej kategorii danej zmiennej. Sposób kodowania w przypadku, gdy do porównania wybrano pierwszą grupę wieku przedstawia tabela 2. Kodowanie cech umożliwiło zastąpienie cechy ilościowej (wiek w latach) cechą kategoryzowaną (kodowanie 0-1). Poszczególne przedziały wieku ponumerowano od 1 do 6. Ponieważ jako odniesienie przyjęto pierwszy przedział wieku ($\langle 18, 25)$), to cechy w modelu oznaczono jako $Wiek\,(i,\ 1)$ dla $i \in \{2, 3, 4, 5, 6\}$.

Tabela 2

I rodzaj kodowania

Wiek bezrobotnych	Cecha *Wiek* (*i*, 1)				
	Wiek (2,1)	*Wiek* (3,1)	*Wiek* (4,1)	*Wiek* (5,1)	*Wiek* (6,1)
$\langle 18, 25)$	0	0	0	0	0
$\langle 25, 35)$	1	0	0	0	0
$\langle 35, 45)$	0	1	0	0	0
$\langle 45, 55)$	0	0	1	0	0
$\langle 55, 60)$	0	0	0	1	0
$\langle 60, 65)$	0	0	0	0	1

Źródło: opracowanie własne.

Wyniki estymacji parametrów modelu Coxa przedstawiono w tabeli 3. Jest to model ze zmiennymi kategoryzowanymi $Wiek\,(i,\ 1)$ dla $i \in \{2, 3, 4, 5, 6\}$.

Tabela 3

Wyniki estymacji parametrów modelu Coxa przy zastosowaniu kodowania I

	Cecha	β_i	Błąd parametru	Wartość statystyki t	$\exp(\beta_i)$	Statystyka Walda	p
Kodowanie I	*Wiek* (2,1)	–0,05824	0,060162	–0,96807	0,943423	0,93715	0,333018
	Wiek (3,1)	–0,54685	0,075151	–7,27666	0,578770	52,94979	0,000000
	Wiek (4,1)	–0,66539	0,072208	–9,21489	0,514073	84,91415	0,000000
	Wiek (5,1)	–0,91716	0,137533	–6,66870	0,399651	44,47152	0,000000
	Wiek (6,1)	–2,67375	0,502618	–5,31965	0,068993	28,29865	0,000000

Źródło: obliczenia własne z wykorzystaniem programu *Statistica*.

Oszacowany, metodą częściowej wiarygodności[3], model można przedstawić w następującej postaci:

$$h(t,x_2,x_3,x_4,x_5,x_6) = \\ = h_0(t)\exp(-0{,}05824x_2 - 0{,}54685x_3 - 0{,}66539x_4 - 0{,}91716x_5 - 2{,}67375x_6), \tag{2}$$

gdzie:

$$x_i = Wiek(i,\ 1),\ \text{dla}\ i = 2,\ \ldots,\ 6.$$

Wyrażenie $\exp\beta_i$ wyraża w tym przypadku stosunek szansy na znalezienie pracy przez bezrobotnego z i-tej grupy wieku w porównaniu z grupą pierwszą. Przyjmuje się więc, że $\beta_1 = 0$.

Istnieje również możliwość obliczenia szansy względnej między dowolnymi dwiema grupami wieku. Wartość odpowiedniego parametru beta wyznacza się jako stosunek funkcji proporcjonalnego hazardu dla porównywanych kategorii danej zmiennej, przy założeniu stałości pozostałych zmiennych objaśniających[4]. Zmiany ryzyka w zależności od grupy wieku wyznaczono na podstawie wzoru:

$$Wiek(i,j) = \frac{Wiek(i,1)}{Wiek(j,1)} = \frac{\exp\beta_i}{\exp\beta_j} = \exp(\beta_i - \beta_j),\ \text{dla}\ i,j = 2,\ldots,6 \tag{3}$$

a wyniki zaprezentowano w tabeli 4.

[3] Por. [8], s. 29-30, [5], s. 11-14.
[4] Por. [5], s. 123-124.

Tabela 4

Szansa względna uzyskania zatrudnienia wyznaczona na podstawie wzoru (3)

Szansa względna uzyskania zatrudnienia przez bezrobotnych	w stosunku do grupy wieku				
z grupy wieku	⟨18, 25)	⟨25, 35)	⟨35, 45)	⟨45, 55)	⟨55, 60)
⟨25, 35)	0,943423				
⟨35, 45)	0,57877	0,613479			
⟨45, 55)	0,514073	0,544902	0,888216		
⟨55, 60)	0,399651	0,423618	0,690518	0,777421	
⟨60, 65)	0,068993	0,073131	0,119206	0,134209	0,172633

Źródło: obliczenia własne z wykorzystaniem programu *Statistica*.

Drugi rodzaj kodowania umożliwia wyznaczenie szansy opuszczenia rejestru PUP przez bezrobotnego z danej grupy wieku względem średniej całej badanej kohorty (tabela 5). Ponieważ jako odniesienie przyjęto średnią dla kohorty oznaczoną jako s, to cechy w modelu oznaczono jako $Wiek(i, s)$ dla $i \in \{1, 2, 3, 4, 5, 6\}$.

Tabela 5

II rodzaj kodowania

Wiek bezrobotnych	Cecha *Wiek* (i, s)				
	Wiek (2, s)	*Wiek* (3, s)	*Wiek* (3, s)	*Wiek* (4, s)	*Wiek* (5, s)
⟨18, 25)	–1	–1	–1	–1	–1
⟨25, 35)	1	0	0	0	0
⟨35, 45)	0	1	0	0	0
⟨45, 55)	0	0	1	0	0
⟨55, 60)	0	0	0	1	0
⟨60, 65)	0	0	0	0	1

Źródło: opracowanie własne.

W przypadku podziału kohorty na n grup otrzymujemy $n - 1$ estymatorów parametrów $\beta_2^*, \beta_3^*, \ldots, \beta_n^*$, przy czym zachodzi warunek:

$$\sum_{i=1}^{n} \beta_i^* = 0, \tag{4}$$

czyli:

$$\beta_1^* = -\sum_{i=2}^{n} \beta_i^*. \tag{5}$$

Wyniki estymacji parametrów modelu regresji Coxa przedstawiono w tabeli 6.

Tabela 6

Wyniki estymacji parametrów modelu Coxa przy zastosowaniu kodowania II

	Cecha	β_i^*	Błąd parametru	Wartość statystyki t	$\exp(\beta_i^*)$	Statystyka Walda	p
Kodowanie 2	*Wiek* (1, *s*)	0,810257	0,096426	8,402876	2,248487	70,60831	0,000000
	Wiek (2, *s*)	0,75202	0,092535	8,12681	2,121274	66,04506	0,000000
	Wiek (3, *s*)	0,26340	0,099121	2,65740	1,301352	7,06177	0,007879
	Wiek (4, *s*)	0,14486	0,097587	1,48445	1,155882	2,20358	0,137700
	Wiek (5, *s*)	–0,10691	0,136522	–0,78311	0,898605	0,61326	0,433568
	Wiek (6, *s*)	–1,86363	0,417675	–4,46191	0,155109	19,90862	0,000008

Źródło: obliczenia własne z wykorzystaniem programu *Statistica*.

W tym przypadku model ze zmiennymi kategoryzowanymi *Wiek* (i, s) dla $i \in \{1, 2, 3, 4, 5, 6\}$ ma postać:

$$\begin{aligned} &h(t, x_2, x_3, x_4, x_5, x_6) = \\ &= h_0(t)\exp(0{,}810257x_1 + 0{,}75202x_2 + 0{,}2634x_3 + 0{,}14486x_4 - 0{,}10697x_5 - 1{,}86363x_6), \end{aligned} \tag{6}$$

gdzie:

$$x_i = \textit{Wiek}\,(i, s), \text{ dla } i = 1, \ldots, 6.$$

Wyrażenie $\exp(\beta_i^*)$ wyraża w tym przypadku stosunek szansy na znalezienie pracy przez bezrobotnego z i-tej grupy wieku w porównaniu ze średnią całej kohorty.

Również w przypadku drugiego kodowania można obliczyć szansę względną między i-tą grupą wieku, a grupą pierwszą. Można ją wyznaczyć korzystając z zależności [2]:

$$\textit{Wiek}(i,1) = \frac{\exp\beta_i^*}{\exp\beta_1^*} = \exp(\beta_2^* + \ldots 2\beta_i^* + \ldots + \beta_n^*). \tag{7}$$

Analogicznie wyznacza się szansę względną między dowolnymi dwiema grupami wieku:

$$\textit{Wiek}(i,j) = \frac{\textit{Wiek}(i,s)}{\textit{Wiek}(j,s)} = \frac{\exp\beta_i^*}{\exp\beta_j^*} = \frac{\exp\beta_i}{\exp\beta_j}. \tag{8}$$

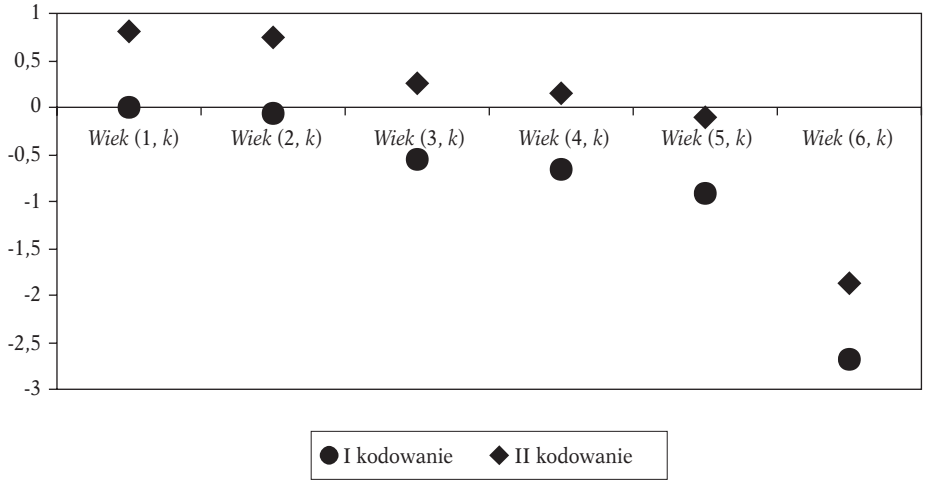
Otrzymane wartości szansy względnej są takie same, jak w przypadku kodowania pierwszego (tabela 4).

Stopień dopasowania modelu przy zastosowaniu obu sposobów kodowania jest oczywiście taki sam, wartość statystyki χ^2 wynosi 229,844 przy poziomie istotności $p = 0{,}0000$.

5. ZWIĄZKI MIĘDZY PARAMETRAMI MODELU COXA PRZY ZASTOSOWANIU OBU SPOSOBÓW KODOWANIA

Zastosowanie poszczególnych sposobów kodowania daje w wyniku różne oszacowania parametrów modelu Coxa i inna jest też ich interpretacja. Kodowanie I pozwala na wyznaczenie szansy na znalezienie pracy przez bezrobotnego z danej grupy wieku (i) względem możliwości zdobycia zatrudnienia osób z grupy pierwszej. Przykładowo bezrobotni w wieku od 45 do 55 lat mają prawie o połowę mniejszą szansę podjęcia pracy niż osoby w wieku od 18 do 25 lat. Przy interpretacji parametrów, w przypadku zastosowania kodowania II, punktem odniesienia jest średnia szansa znalezienia pracy całej badanej zbiorowości. Bezrobotni w wieku od 45 do 55 lat w tym przypadku o 15% szybciej znajdowali zatrudnienie niż przeciętnie w całej kohorcie.

Jak już wskazano oszacowania parametrów, jak też ich interpretacja, przy zastosowaniu obu sposobów kodowania są różne, ale przedstawiając na wykresie (rysunek 1) wyznaczone wartości parametrów modelu regresji Coxa można zauważyć istnienie pewnej zależności.



Rysunek 1. Wartości oszacowanych parametrów β i β^* (stała różnica między parametrami)

Źródło: opracowanie własne.

Wartości szansy względnej wyznaczone ze wzoru (3) przy zastosowaniu kodowania I i ze wzoru (8) przy zastosowaniu kodowania II, które zaprezentowano w tabeli 4, są

jednakowe. W związku z tym powinna istnieć zależność między parametrami oznaczonymi w artykule jako β i β^*.

Korzystając z zależności:

$$\exp\left(\beta_2^* + \ldots + 2\beta_i^* + \ldots + \beta_n^*\right) = \exp\beta_i, \quad \text{dla } i = 2, 3, \ldots, n \tag{9}$$

oraz

$$\exp\left(\beta_1^* + \beta_2^* + \ldots + \beta_i^* + \ldots + \beta_n^*\right) = \exp\beta_1 \tag{10}$$

otrzymujemy wzory przejścia między dwoma omówionymi sposobami kodowania:

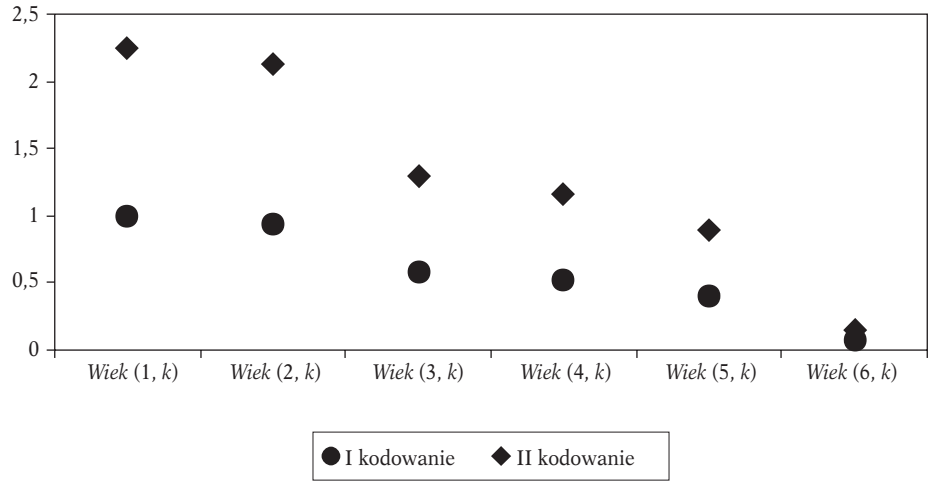
$$\beta_i^* = \beta_i - \frac{1}{n}\sum_{k=1}^{n}\beta_k, \text{ dla } k = 1, 2, \ldots, n. \tag{11}$$

Różnica między parametrami β i β^* równa się:

$$\beta_i - \beta_i^* = \frac{1}{n}\sum_{k=1}^{n}\beta_k, \text{ dla } k = 1, 2, \ldots, n, \tag{12}$$

czyli jest stała i równa się średniej arytmetycznej parametrów β_i, uzyskanych w wyniku pierwszego kodowania.

Ponieważ potwierdzono istnienie zależności między parametrami β i β^* można przypuszczać, że również zachodzi związek między szansami względnymi, wyznaczonymi odpowiednio w stosunku do pierwszej grupy wieku oraz średniego czasu pozostawania bez pracy. Oszacowane wartości szansy względnej w przypadku kodowania I i II przedstawiono na rysunku 2.



Rysunek 2. Oszacowane wartości szansy względnej w przypadku kodowania I i II (stały stosunek szans względnych)

Źródło: opracowanie własne.

Korzystając ze wzoru (12) można znaleźć związek między $\exp(\beta_i)$ i $\exp(\beta_i^*)$:

$$\frac{\exp(\beta_i)}{\exp(\beta_i^*)} = \exp\left(\frac{1}{n}\sum_{k=1}^{n}\beta_k\right), \text{ dla } k = 1, 2, \ldots, n. \tag{13}$$

Ze wzoru (13) można odczytać, że stosunki szansy względnej w przypadku kodowania pierwszego i drugiego są stałe i równe $\exp\left(\frac{1}{n}\sum_{k=1}^{n}\beta_k\right)$.

6. PODSUMOWANIE

Z przedstawionych badań wynikają następujące wnioski:

– parametry modelu Coxa można wyznaczyć stosując dwa sposoby kodowania zmiennych wpływających na czas poszukiwania pracy,

– kodowanie 0-1 oznaczone w artykule jako kodowanie I wymusza określenie podgrupy odniesienia; w analizowanym przykładzie wybrano podgrupę pierwszą, czyli bezrobotnych w wieku od 18 do 25 lat,

– jako grupę odniesienia można przyjąć dowolną podgrupę, którą badacz chce wyróżnić; może to być na przykład grupa najliczniejsza, najstarsza itp.,

– stosując kodowanie –1-0-1, oznaczone w artykule jako kodowanie II, punktem odniesienia jest średnia całej kohorty; w tym przypadku nie ma znaczenia wybór podgrupy, która zostanie oznaczona przez –1,

– między parametrami modelu proporcjonalnego hazardu Coxa w przypadku obu sposobów kodowania zmiennych istnieje związek; różnica między odpowiadającymi sobie parametrami jest stała i równa średniej arytmetycznej z parametrów otrzymanych dla kodowania 0-1;

– wyznaczenie szansy względnej na podjęcie pracy dla dowolnych dwóch podgrup jest możliwe przy zastosowaniu dowolnego rodzaju kodowania.

Uniwersytet Szczeciński

LITERATURA

[1] Bednarski T., [2005], *Ocena przydatności danych Bael dla charakterystyki rozkładu czasu poszukiwania pracy na przykładzie danych z lat 2001-2002*, Studia Ekonomiczne, nr 4, Instytut Nauk Ekonomicznych PAN, Warszawa.

[2] Colett D., [2003], *Modelling Survival Data in Medical Research*, Chapman & Hall/CRC, Boca Raton, Floryda.

[3] Cox D.R., Oakes D., [1984], *Analysis of Survival Data*, Chapman and Hall, London.

[4] Frątczak E., Gach-Ciepiela U., Babiker H., [2005], *Analiza historii zdarzeń. Elementy teorii, wybrane przykłady zastosowań*, SGH, Warszawa.

[5] Hosmer D.W., Lemeshow S., [1999], *Applied Survival Analysis. Regression Modeling of Time to Event Data*, John Wiley & Sons, INC, New York.

[6] Hozer J. (red.), [2002], *Badania statystyczne w ubezpieczeniach*, Wydawnictwo Naukowe Uniwersytetu Szczecińskiego, Szczecin.

[7] Markowicz I., Stolorz B., [2007], *Determinants of Labour Seeking Time Resulting From Labour Demand on Szczecin Labour Market*, The labour demand in the modern economy, Economics & Competition Policy, No. 10, Katedra Mikroekonomii Uniwersytetu Szczecińskiego, Szczecin.
[8] Rossa A., [2005], *Metody estymacji rozkładu czasu trwania zjawisk dla danych cenzurowanych oraz ich zastosowania*, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

Praca wpłynęła do redakcji w listopadzie 2008 r.

MODEL PROPORCJONALNEGO HAZARDU COXA PRZY RÓŻNYCH SPOSOBACH KODOWANIA ZMIENNYCH

Streszczenie

Metody analizy przeżycia są coraz częściej stosowane w badaniach zjawisk społeczno-ekonomicznych. Ze względu na brak konieczności znajomości rozkładu badanej zmiennej losowej szczególną wagę przywiązuje się do modeli nieparametrycznych bądź semiparametrycznych. Coraz powszechniej wykorzystywane są one do badania zjawisk innych niż czas trwania życia ludzkiego. Warunkiem stosowania modeli analizy przeżycia jest odpowiednia baza danych umożliwiająca wyznaczenie czasu trwania zdefiniowanego stanu dla poszczególnych jednostek badanej zbiorowości. Zazwyczaj są to badania retrospektywne z wykorzystaniem sporządzanych rejestrów. Przykładem takiej bazy danych jest rejestr bezrobotnych.

Celem artykułu jest wskazanie wpływu sposobu kodowania zmiennych na oszacowania parametrów modelu regresji Coxa i ich interpretację. Autorki przedstawiły również związek między parametrami modelu szacowanymi dla danych zakodowanych w dwojaki sposób. Badaną kohortę stanowią osoby bezrobotne wyrejestrowane w określonym okresie czasu. Podziału na podgrupy dokonano ze względu na wiek, który jest determinantą czasu poszukiwania pracy.

Słowa kluczowe: analiza przeżycia, modele semiparametryczne, model regresji Coxa, kodowanie.

THE COX PROPORTIONAL HAZARD MODEL FOR DIFFERENT METHODS OF ENCRYPTION OF VARIABLES

Summary

Methods of survival analysis are more and more often used in analysis of social and economic occurrences. Due to lack of distributional information regarding the random variable, much attention is put on non-parametric or semi-parametric models. They are more and more commonly used for analysis of occurrences different than life expectancy. The condition of use of models of survival analysis is appropriate database that makes possible estimation of duration time of defined state for particular elements of analysed population. They are usually retrospective analyses with use of records. The example of such database is unemployment records.

The article presents results of analysis of influence of encryption of variables on estimation of parameters of the Cox proportional hazard model and their interpretation. The authors also presented correlation between parameters of the model estimated for the data encrypted in two ways. The cohort consisted of the unemployed persons unregistered in specific period. Sub-clusters were allocated with respect to age that is a determinant of period of waiting for a job.

Key words: survival analysis, semi-parametric models, Cox regression model, encryption.

PRZEGLĄD STATYSTYCZNY
R. LVI – ZESZYT 2 – 2009

PIOTR KULCZYCKI, KARINA DANIEL

METODA WSPOMAGANIA STRATEGII MARKETINGOWEJ OPERATORA TELEFONII KOMÓRKOWEJ

1. WSTĘP

Rynek telefonii komórkowej jest nieustannie poddawany działaniom mającym na celu dostosowanie się do ulegających ciągłym zmianom wymagań klientów. Zarówno dynamiczny rozwój tego rynku, jak i poszerzający się zakres oferowanych usług, implikują coraz bardziej ekspansywną postawę operatorów wobec poszczególnych grup abonenckich. Ze względu na nasilającą się szczególnie w tej dziedzinie konkurencję ważne jest poszukiwanie narzędzi, które stanowić będą skuteczną pomoc w powiększaniu efektywnej aktywności firm operatorskich. Precyzyjne zdefiniowane potrzeb każdego klienta skutkuje zwiększeniem się szansy utrzymania go u danego operatora i uzyskania z tego tytułu możliwie jak największych zysków. Z kolei uniknięcie nadmiernego rozdrobnienia w charakteryzowaniu poszczególnych grup abonentów, a także utraty spójności w strategii postępowania wobec klientów, wymaga wprowadzenia w tym zakresie nowych rozwiązań o charakterze systemowym.

Powyższe szczególnie dotyczy rynku klientów korporacyjnych – grupy specyficznej i bardzo wymagającej. Kreowanie oferty wymaga w tym przypadku znacznej elastyczności w zakresie proponowanych usług oraz otwartości na indywidualne rozwiązania, które mogą stać się elementem budującym i utrwalającym współpracę, a także gwarantującym ją w dłuższym horyzoncie czasowym.

Przedmiotem prezentowanych tu badań jest zagadnienie wspomagania decyzji dotyczącej określenia najkorzystniejszej strategii postępowania wobec konkretnego klienta korporacyjnego – poprzez odpowiednie zastosowanie zniżek – w celu maksymalizacji zysku operatora sieci telefonii komórkowej, przy jednoczesnym zapewnieniu satysfakcjonującej abonenta jakości usług. Procedura ta została stworzona z wykorzystaniem metodyki statystycznych estymatorów jądrowych, a jej postać oraz wyniki otrzymane na podstawie prowadzonych badań zweryfikowane zostały z użyciem danych rzeczywistych uzyskanych od jednego z polskich operatorów sieci GSM, w oparciu o wiedzę teoretyczną i doświadczenie wynikające z praktyki zawodowej związanej z rynkiem biznesowym.

Statystyczne estymatory jądrowe wykorzystano w prezentowanej pracy do zagadnień wykrywania elementów nietypowych (odosobnionych), analizy skupień (klasteryzacji) oraz klasyfikacji. W opracowanym algorytmie zastosowano także teorię zbiorów rozmytych, umożliwiającą przede wszystkim wykorzystanie wiedzy eksperta – znawcy specyfiki funkcjonowania rynku biznesowego abonentów sieci telefonii komórkowej.

Niniejszy artykuł jest podzielony na trzy zasadnicze części. Sekcję drugą stanowią preliminaria matematyczne – przedstawiono tu wybrane zagadnienia związane z estymatorami jądrowymi, a także ich zastosowaniami do wykrywania elementów odosobnionych, jak również w algorytmach analizy skupień oraz klasyfikacji. Sekcja trzecia stanowi główną część pracy – został tu przedstawiony i szczegółowo opisany algorytm wspomagania decyzji dotyczącej strategii marketingowej operatora sieci telefonii komórkowej. W poszczególnych podsekcjach kolejno zawarto opis struktury, analizę i przetwarzanie danych, sposób określania tablic ryzyka z wykorzystaniem liczb rozmytych typu L-R, a także wyznaczania funkcji celu oraz proces klasyfikacji klienta, co kompletuje prezentowaną tu formułę wspomagania decyzji. W sekcji czwartej przedstawiono wyniki otrzymane na podstawie badań przeprowadzonych na zbiorze rzeczywistych danych. W pełni potwierdziły one poprawność i użyteczność prezentowanej tu procedury, dostarczając także wniosków i komentarzy oraz ilustracyjnych interpretacji.

2. PRELIMINARIA MATEMATYCZNE

2.1. STATYSTYCZNE ESTYMATORY JĄDROWE

Estymatory jądrowe należą do grupy metod nieparametrycznych, w których nie formułuje się arbitralnie założeń dotyczących przynależności rozkładu badanej zmiennej do określonej klasy. Niech zatem dana będzie n-wymiarowa zmienna losowa, której rozkład posiada funkcję gęstości f. Jej estymator jądrowy $\hat{f}:\mathbb{R}^n \to [0,\infty)$ wyznaczany jest na podstawie wartości m-elementowej próby losowej $x_1, x_2. \ldots, x_m$ i w swej podstawowej postaci dany jest wzorem

$$\hat{f}(x) = \frac{1}{mh^n}\sum_{i=1}^{m} K\left(\frac{x - x_i}{h}\right), \tag{1}$$

przy czym dodatni współczynnik h określa się mianem parametru wygładzania, natomiast mierzalna funkcja $K: \mathbb{R}^n \to [0, \infty)$, która jest symetryczna względem zera (tj. $K(x) = K(-x)$ dla każdego $x \in \mathbb{R}^n$), posiada w tym punkcie słabe maksimum globalne (czyli $K(0) \geq K(x)$ dla każdego $x \in \mathbb{R}^n$) oraz spełnia warunek jednostkowej całki (tzn. $\int_{\mathbb{R}^n} K(x)dx = 1$), nazywana jest jądrem. Wyboru postaci jądra K i wartości parametru wygładzania h dokonuje się najczęściej w oparciu o kryterium minimalizacji scałkowanego błędu średniokwadratowego. Okazuje się, iż postać jądra K nie wpływa w znaczącym stopniu na jakość statystyczną estymacji – stanowi to istotną praktyczną zaletę prezentowanej metody, gdyż pozwala zagwarantować cechy wymagane od estymatora (1) w konkretnym zastosowaniu, na przykład odpowiednią klasę regularności, przyjmowanie dodatnich wartości lub dogodność obliczeniową. W przeciwieństwie, istotne znaczenie ma właściwe wyznaczenie wartości parametru wygładzania h – istnieją wszakże dogodne algorytmy pozwalające obliczyć powyższą wartość na podstawie posiadanej próby losowej. W praktycznych zadaniach stosuje się ponadto dodatkowe

procedury polepszające własności estymatora (np. modyfikację parametru wygładzania lub transformację liniową) oraz dopasowujące jego własności do specyfiki rozważanego zagadnienia (m.in. ograniczenie nośnika). Estymatory jądrowe pozwalają również oszacować inne charakterystyki rozkładów zmiennych losowych, np. dystrybuantę lub wartość kwantyla. Szczegółowy opis powyższej metodyki można znaleźć w książce [11] oraz klasycznych monografiach [14, 15].

Statystyczne estymatory jądrowe stanowią dogodne narzędzie matematyczne przy rozwiązywaniu różnorodnych zadań z zakresu analizy i przetwarzania danych. W niniejszej pracy zostały one zastosowane do zagadnień wykrywania elementów odosobnionych, analizy skupień oraz klasyfikacji – zagadnienia te będą przedstawione poniżej w kolejnych podsekcjach.

2.2. WYKRYWANIE ELEMENTÓW ODOSOBNIONYCH

W wielu zagadnieniach analizy danych występuje zadanie wykrywania elementów nietypowych (odosobnionych), czyli takich, które w istotny sposób różnią się od ogółu populacji. Często umożliwia to eliminację tego typu elementów z posiadanego zbioru danych, co zwiększa jego homogeniczność (jednorodność), ułatwiając wnioskowanie, zwłaszcza w złożonych i nietypowych przypadkach. W praktyce, proces wykrywania elementów odosobnionych najczęściej jest dokonywany z wykorzystaniem procedur testowania hipotez statystycznych [1]. Poniżej zostanie przedstawiony test istotności, którego konstrukcja została oparta na metodyce estymatorów jądrowych.

Niech zatem dana będzie próba losowa $x_1, x_2, \ldots, x_m$ uznawana za wzorcową, czyli zawierająca możliwie jak najbardziej reprezentatywny zbiór elementów typowych. Niech ponadto $\alpha \in (0,1)$ oznacza założony poziom istotności. Testowana jest hipoteza stanowiąca, że ustalony element $\tilde{x} \in \mathbb{R}^n$ jest typowy, przeciwko hipotezie iż takowym nie jest, czyli iż należy go uznać za odosobniony. Stosowaną tu statystykę $S : \mathbb{R}^n \to [0, \infty)$ definiuje się wzorem

$$S(\tilde{x}) = \hat{f}(\tilde{x}), \tag{2}$$

przy czym $\hat{f}$ oznacza jądrowy estymator gęstości otrzymany dla wzmiankowanej powyżej wzorcowej próby losowej $x_1, x_2, \ldots, x_m$, podczas gdy zbiór odrzucenia (krytyczny) przyjmuje postać lewostronną $(-\infty, a]$, gdzie a stanowi jądrowy estymator kwantyla stopnia α wyznaczony na podstawie próby $\hat{f}(x_1), \hat{f}(x_2), \ldots, \hat{f}(x_m)$, przy założeniu iż nośnik zmiennej losowej jest ograniczony do zbioru liczb nieujemnych.

2.3. ANALIZA SKUPIEŃ

Celem analizy skupień jest dokonanie podziału zbioru danych – na przykład określonych w postaci próby losowej $x_1, x_2, \ldots, x_m$ na podgrupy (skupienia, klastry), z których każda zawiera „podobne" do siebie elementy, ale istotnie różniące się między poszczególnymi podgrupami [10]. W praktyce często pozwala to zdekomponować

duże zbiory danych o zróżnicowanych charakterystykach poszczególnych elementów, na podzbiory zawierające elementy o podobnych własnościach, co znacząco ułatwia lub wręcz umożliwia dalsze, szczegółowe wnioskowanie. W niniejszej podsekcji zostanie przedstawiona procedura analizy skupień oparta na statystycznych estymatorach jądrowych, wykorzystująca ideę metod gradientowych [5].

Przyjmuje się tu naturalne założenie, iż poszczególne skupienia odpowiadają modom (lokalnym maksimom) jądrowego estymatora gęstości $\hat{f}$ określonego na podstawie rozważanej próby losowej $x_1, x_2, \ldots, x_m$. W ramach wspomnianej procedury, poszczególne elementy są przemieszczane w kierunku zdefiniowanym przez gradient, według następującego algorytmu iteracyjnego:

$$x_j^0 = x_j \text{ dla } j = 1, 2, \ldots, m \tag{4}$$

$$x_j^{k+1} = x_j^k + b\frac{\nabla \hat{f}\left(x_j^k\right)}{\hat{f}\left(x_j^k\right)} \text{ dla } j = 1, 2, \ldots, m \text{ oraz } k = 1, 1, \ldots, \tag{5}$$

przy czym $b > 0$ (w praktyce rekomenduje się $b = h^2/(n + 2)$), natomiast ∇ oznacza operator gradientu. Dzięki odpowiedniemu doborowi postaci jądra K, możliwe jest uzyskanie dogodnej, analitycznej formuły gradientu $\nabla \hat{f}$.

W wyniku wykonywania kolejnych kroków iteracyjnych, elementy próby losowej przemieszczają się sukcesywnie, coraz wyraźniej grupując w pewną liczbę skupień. Ich ostateczną postać określa się po wykonaniu ustalonego k^*-tego kroku[1]. W tym celu obliczany jest estymator jądrowy wzajemnych odległości elementów $x_1^{k^*}, x_2^{k^*}, \ldots, x_m^{k^*}$ (przy założeniu nieujemności nośnika zmiennej), po czym zostaje wyznaczona najmniejsza wartość, w której estymator ten osiąga pierwsze lokalne minimum, z pominięciem ewentualnego minimum w zerze. Do poszczególnych skupień zalicza się te elementy, których odległość jest nie większa niż tak wyznaczona wartość. Dzięki możliwościom zmiany wartości parametru wygładzania h, możliwa staje się ingerencja w rząd wielkości liczby otrzymywanych skupień, aczkolwiek bez arbitralnych założeń dotyczących wartości tej liczby, co umożliwia dopasowanie jej do rzeczywistej struktury danych. Co więcej, ewentualne zmiany intensywności procedury modyfikacji parametru wygładzania, umożliwiają wpływ na proporcję liczby skupień wyznaczonych w obszarach zagęszczenia elementów posiadanej próby losowej do liczby skupień położonych w rejonach, gdzie są one rzadkie. Szczegółowy opis powyższego algorytmu analizy skupień zawarto w artykule [12].

2.4. KLASYFIKACJA

Niech dana będzie liczba $J \in \mathbb{N}\backslash\{0,1\}$. Załóżmy także, iż pozyskana z n-wymiarowej zmiennej losowej próba $x_1, x_2, \ldots, x_m$ została podzielona na J rozłącznych podzbiorów:

[1] Dla rozważanego w tej pracy zbioru danych, po przeprowadzeniu szczegółowych wstępnych badań, przyjęto $k^* = 50$. W pracy [11] znaleźć można ogólną formułę kryterium wyznaczania wartości k^*.

$$x_1^1, x_2^1, \ldots, x_{m_1}^1 \tag{6}$$

$$x_1^2, x_2^2, \ldots, x_{m_2}^2 \tag{7}$$

$$\vdots$$

$$x_1^J, x_2^J, \ldots, x_{m_J}^J, \tag{8}$$

gdzie $m_1, m_2, \ldots, m_j \in \mathbb{N}\backslash\{0\}$ oraz $\sum_{j=1}^{J} m_j = m$, reprezentujących wzorce wyróżnionych klas, przy czym wprowadzony dodatkowo górny indeks stanowi o przynależności elementu do wzorca danej klasy. Zadanie klasyfikacji polega na wskazaniu, do którego z nich należy przypisać ustalony element $\tilde{x} \in \mathbb{R}^n$ [10].

Metodyka estymatorów jądrowych dostarcza naturalnego aparatu matematycznego do rozwiązania powyższego zagadnienia w optymalnym – w sensie minimum wartości oczekiwanej strat – ujęciu bayesowskim. Otóż niech $\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_j$ oznaczają jądrowe estymatory gęstości określone dla prób odpowiednio (6)-(8). Jeżeli liczności $m_1, m_2, \ldots, m_J$ są proporcjonalne do „częstości" pojawiania się elementów z poszczególnych klas, to rozważany element $\tilde{x}$ należy zaliczyć do tej klasy, dla której wartość $m_1\hat{f}_1(\tilde{x}), m_2\hat{f}_2(\tilde{x}), \ldots, m_J\hat{f}_J(\tilde{x})$ jest największa.

3. ALGORYTM WSPOMAGANIA DECYZJI DOTYCZĄCEJ STRATEGII MARKETINGOWEJ

W poniższej części zostanie przedstawiony opis prezentowanej tu metody wspomagania strategii marketingowej operatora telefonii komórkowej wobec klienta korporacyjnego. Sama procedura jest wieloetapowa i z tego powodu jej opis został podzielony na poszczególne podsekcje, w których zawarte będą kolejne fazy procesu analizy i przetwarzania danych oraz wspomagania procesu decyzyjnego.

3.1. STRUKTURA DANYCH

W praktyce istnieje rozległe spektrum wielkości charakteryzujących poszczególnych abonentów. Dzieje się tak przede wszystkim ze względu na ciągły rozwój rynku telefonii komórkowej, związany między innymi z wprowadzaniem nowych rozwiązań mobilnych w zakresie szeroko rozumianych usług głosowych i transmisji danych, mających na celu zaspokojenie zwiększających się potrzeb klientów. Przystępując do analizy należy spośród różnorodnych wielkości wytypować te, które możliwie najlepiej będą opisywać rozważane zagadnienie przy jednoczesnej gwarancji efektywnej ich aktualizacji. Warto również zaznaczyć, że nie wszystkie z badanych charakterystyk mogą być w praktyce identyfikowalne w równych odstępach czasu, część z nich ma charakter losowy, a niektóre sezonowy. Co więcej, niektóre z nich są tworzone i obserwowane w nieregularnych odstępach czasu, gdy przykładowo służą one weryfikacji skuteczności zastosowania jednorazowych działań operatora mających na celu polepszenie relacji z klientami lub zweryfikowanie słuszności wprowadzenia konkretnego produktu.

Ostatecznie, po szczegółowej analizie aspektów ekonomicznych rozważanego w niniejszej pracy zagadnienia przyjęte zostało, iż podstawową charakterystykę poszczególnych klientów stanowić będą trzy wielkości:

– średni miesięczny przychód przypadający na kartę SIM (ARPU),
– staż w sieci,
– liczba aktywnych kart SIM.

Przeprowadzone badania potwierdziły, iż poszerzanie powyższej listy o dalsze wielkości praktycznie nie polepsza otrzymywanych wyników, natomiast zwiększa czas i złożoność obliczeń.

W dalszej części pomocniczo rozważana będzie również wartość średniego abonamentu danego klienta (średnia arytmetyczna wszystkich abonamentów jego konta). Iloczyn liczby kart SIM oraz tak wyznaczonego średniego abonamentu stanowi główną część stałych przychodów, jakie dany klient generuje dla operatora. Pozostały przychód uzależniony jest od czasu połączeń wykonywanych poza pakietem uwzględnionym w kwocie abonamentu. Jest on związany z rodzajem usług, z których korzysta klient, dlatego też coraz częściej uwzględnia się nie tylko opłaty za połączenia głosowe, ale także za transmisję danych oraz inne usługi bezpośrednio z nią związane.

Uzupełnienie dla prowadzonych analiz stanowić będą także pomocnicze wielkości o charakterze ilościowym i jakościowym, które w podlegającym badaniom zbiorze klientów wyróżniają jednostki prestiżowe, a także firmy transportowe oraz typu ochrona-monitoring, jak również przedsiębiorstwa małe, średnie, duże oraz bardzo duże. Zostaną one wykorzystane w procesie eliminacji nietypowych elementów i mało licznych skupień.

Niech zatem charakterystyka poszczególnych klientów dana będzie w postaci 3-wymiarowego wektora:

$$x_i = \begin{bmatrix} x_{i,1} \\ x_{i,2} \\ x_{i,3} \end{bmatrix} \text{ dla } i = 1, 2, \ldots, m, \tag{9}$$

gdzie $x_{i,1}$ stanowi średni miesięczny przychód na kartę SIM i-tego klienta, $x_{i,2}$ – jego staż w sieci, $x_{i,3}$ – liczbę kart SIM tegoż klienta, podczas gdy m stanowi liczność posiadanej bazy danych. Dodatkowo przez $y_{i,1}$ oznaczona zostanie wielkość jego średniego abonamentu, natomiast zmienne $y_{i,2}$, $y_{i,3}$, $y_{i,4}$ określono następująco:

$y_{i,2}$ – przynależność do jednostek prestiżowych charakteryzuje się za pomocą zmiennej skategoryzowanej o następujących wartościach: 0 – klient nieprestiżowy, 1 – klient istotny dla operatora z punktu widzenia strategii firmy, zwany dalej prestiżowym, 2 – klient wyjątkowo prestiżowy (kluczowy);

$y_{i,3}$ – wielkość przedsiębiorstwa charakteryzowana jest przez zmienną skategoryzowaną o wartościach: 1 – małe przedsiębiorstwo, 2 – średnie przedsiębiorstwo, 3 – duże przedsiębiorstwo, 4 – bardzo duże przedsiębiorstwo;

$y_{i,4}$ – dla firm transportowych oraz typu ochrona-monitoring jako jej wartość przypisuje się odpowiednio 1 oraz –1, natomiast 0 dla jednostek o innym profilu działalności.

Powyższe wielkości zostaną użyte do scharakteryzowania poszczególnych klientów korporacyjnych w procesie analizy i przetwarzania danych przedstawionym w dalszej części artykułu.

3.2. ANALIZA I PRZETWARZANIE DANYCH

We wstępnej fazie analizy, ze zbioru danych eliminowane są jednostki nietypowe (elementy odosobnione), zgodnie z procedurą przedstawioną w podsekcji 2.2. Zastrzeżone zostaje jednak, iż nie można usunąć tu jednostek kluczowych oraz wszystkich jednostek z grupy transportowych lub typu ochrona-monitoring. Niniejszym zwiększona zostaje jednolitość struktury danych, a efekt ten otrzymuje się poprzez pominięcie jedyne tych elementów, które miałyby pomniejsze znaczenie w dalszej procedurze.

Po selektywnym wyeliminowaniu elementów nietypowych, dokonuje się analizy skupień zbioru danych, z użyciem algorytmu opisanego w podsekcji 2.3. W rezultacie uzyskuje się podział zbioru danych, reprezentujących poszczególnych klientów korporacyjnych, na podgrupy elementów o zbliżonej charakterystyce. Następnie dokonuje się ponownej, nieco odmiennej eliminacji elementów nietypowych, realizowanej poprzez usunięcie skupień o małej liczności, aczkolwiek z ponownym zastrzeżeniem pozostawienia skupień zawierających jednostki kluczowe, a ponadto skupień składających się co najmniej w połowie z klientów prestiżowych, jak również pozostawienia co najmniej jednego skupienia z firmą transportową oraz jednego skupienia z jednostką typu ochrona-monitoring.

W konsekwencji przeprowadzonego powyżej procesu analizy i przetwarzania danych, ich zbiór pomniejszony jest o elementy wnoszące pomniejszą informację przedmiotową, a pozostałą jego część grupuje się w skupienia zawierające elementy o zbliżonej – z punktu widzenia rozważanego zadania – charakterystyce.

3.3. OKREŚLENIE TABLIC RYZYKA

Podstawę przedstawionej w niniejszej podsekcji procedury wspomagania decyzji podejmowanej wobec firm o charakterystyce właściwej dla poszczególnych, wyznaczonych uprzednio skupień, stanowić będzie wiedza eksperta skutkująca zdefiniowaniem wprowadzonych dla potrzeb dalszej części procedury tabel nazywanych poniżej tablicami ocen rezygnacji.

Na podstawie praktyki tworzenia ofert dla abonentów przyjęto dwie zmienne decyzyjne, których wartości stanowią kolejno o wielkości udzielonego upustu na abonament (oznaczana jako d_a) oraz wielkości rabatu za minutę połączenia (oznaczana przez d_m). Zmienne te zostały wybrane ze względu na fakt, iż stanowią one w każdym procesie negocjacyjnym prowadzonym między operatorem a klientem, podstawowe narzędzie służące tworzeniu nowej i modyfikowaniu obowiązującej oferty, która ma regulować współpracę w przyszłym okresie. Zakłada się, że wynegocjowany poziom upustów i rabatów obowiązywać będzie przez cały okres umowy między firmą i operatorem, czyli przez 24 miesiące.

W dalszej części pracy przyjęto uzasadnioną praktyką dyskretyzację wartości tych zmiennych do następujących „poziomów":

– możliwy upust na abonament $d_a \in \{0\%, 5\%,10\%, 15\%, 50\%, 99\%\}$,

– rabat za minutę połączenia $d_m \in \{0\%, 3\%, 6\%, 12\%, 25\%, 50\%\}$.

Dla każdej kombinacji powyższych „poziomów", ekspert powinien dokonać oceny ryzyka rezygnacji klienta z usług danego operatora. Biorąc pod uwagę przyjęte wartości zniżek, zadaniem eksperta jest przygotowanie dla każdego z otrzymanych skupień, 36 ocen ryzyka. W przypadku eksperta znającego specyfikę badanego rynku, a także zależności w nim występujące, możliwa jest specyfikacja występujących prawidłowości, co istotnie ułatwia określenie takich tablic. Co więcej, dla nieodległych skupień powyższe tablice będą relatywnie podobne. W celu usprawnienia powyższego procesu ekspert może wyznaczyć pewną ilość tablic wzorcowych, które będą stanowić element wyjściowy do dalszych prac, polegających jedynie na ich odpowiednich modyfikacjach. Podstawę do tworzenia tablic wzorcowych powinny stanowić przede wszystkim jednostki kluczowe i prestiżowe oraz firmy transportowe i typu ochrona-monitorig. Dodatkowo ekspert może wykonać modele tablic przygotowane dla klasycznych przedsiębiorstw małych, średnich, dużych i bardzo dużych. Dzięki takiemu ujęciu każda tworzona tablica będzie stanowić połączenie informacji zawartej w odpowiednich tablicach wzorcowych, co korzystnie wpłynie na jakość i czas jego pracy.

Ponieważ praktyka wskazuje, iż tego typu oceny formułuje się w sposób werbalny na podstawie nieprecyzyjnych przesłanek intuicyjnych, to wynik powinien zostać określony w postaci liczby rozmytej. Do powyższego celu użyte będą liczby typu L-R z uwagi na ich prostotę oraz wyrazistą interpretację. Liczba rozmyta $\mathcal{A}$ jest typu L-R w przypadku, jeśli jej funkcja przynależności dana jest w postaci:

$$\mu_{(w,\alpha,\beta)}(x) = \begin{cases} L\left(\dfrac{x-w}{\alpha}\right), & \text{gdy } x \le w \\ R\left(\dfrac{x-w}{\beta}\right), & \text{gdy } x > w \end{cases}, \tag{10}$$

gdzie w $\in \mathbb{R}$, $\alpha, \beta > 0$, a ponadto $L : (-\infty, 0] \to [0,1]$ jest funkcją niemalejącą oraz spełniająca warunek $L(0) = 1$, i analogicznie, $R : [0, \infty) \to [0,1]$ jest funkcją nierosnącą oraz $R(0) = 1$. Parametry α oraz β decydują zatem o lewo- i prawostronnym „rozmyciu" powyższej liczby względem wartości w. Ponadto przyjmuje się dla wartości indeksów $\alpha = 0$ albo $\beta = 0$:

$$\mu_{(w,0,\beta)}(x) = \begin{cases} 0, & \text{gdy } x < w \\ R\left(\dfrac{x-w}{\beta}\right), & \text{gdy } x \ge w \end{cases} \quad \text{dla } \beta > 0 \tag{11}$$

$$\mu_{(w,\alpha,0)}(x) = \begin{cases} L\left(\dfrac{x-w}{\alpha}\right), & \text{gdy } x \le w \\ 0 & \text{gdy } x > w \end{cases} \quad \text{dla } \alpha > 0. \tag{12}$$

Po dokonywanym na wstępie arbitralnym ustaleniu postaci funkcji L oraz R, liczbę rozmytą typu L-R definiują trzy parametry w, α oraz β, a zatem można ją zapisać

w postaci $\mathcal{A} = (w, \alpha, \beta)$ i w konsekwencji proces jej identyfikacji wymaga określenia jedynie tych trzech wartości.

Uwzględniając sposoby określania lojalności klienta korporacyjnego, wynikające ze stosowanej do takich ocen praktyki, przyjęto istnienie siedmiu stanów:

0 – praktycznie nie istnieje możliwość rezygnacji;
1 – ryzyko odejścia jest nikłe, uzależnione jedynie od czynników losowych;
2 – istnieje małe ryzyko rezygnacji;
3 – istnieje średnie ryzyko odejścia;
4 – istnieje duże ryzyko odejścia;
5 – jest omal pewne, że klient zrezygnuje z usług;
6 – praktycznie jest pewne, że klient zrezygnuje z usług.

W tej sytuacji można uściślić, iż $w, \alpha, \beta \in \{0, 1, \ldots, 6\}$, przy czym powinno być spełnione $w - \alpha \geq 0$ oraz $w - \beta \leq 6$.

Na przykładowej tablicy ocen rezygnacji, otrzymanej w trakcie numerycznej weryfikacji prezentowanej tu metody (zamieszczonej w podsekcji 4.3 jako tab. 1), można zauważyć, iż stosunkowo małym rozmyciem charakteryzują się te oceny eksperta, które dotyczą skrajnie małych lub skrajnie dużych zniżek, co świadczy w takich przypadkach o większej precyzyjności oceny pozytywnej albo negatywnej reakcji klienta. Jednakże im większe jest zróżnicowanie w proponowanych poziomach upustów i rabatów, czyli w prawym-górnym i lewym-dolnym obszarze tablicy, tym oceny eksperta stają się nieco mniej precyzyjne. Z powyższych obserwacji wynikają zauważalne prawidłowości przydatne przy określaniu tablic ocen rezygnacji.

Schemat postępowania w stosunku do każdego z typów klientów jest inny, wyszczególnione jednostki swoje decyzje o zmianie operatora podejmują na podstawie odmiennych przesłanek. Niektóre z nich – przede wszystkim jednostki kluczowe i prestiżowe lub bardzo duże przedsiębiorstwa – w sposób bardziej zdecydowany są w stanie podjąć decyzję o rezygnacji z usług firmy świadczącej usługi mobilne w momencie, kiedy ich często zbyt wysokie oczekiwania nie zostają w pełni spełnione przez operatora. Prestiż i wielkość klienta stanowią bowiem dla niego główny argument w procesie prowadzenia negocjacji z operatorem, tym samym formułowane oczekiwania są bardziej rygorystyczne, niż ma to miejsce w przypadku innych klientów. Wobec mniejszych klientów korporacyjnych ekspertowi jest trudniej precyzyjnie określić reakcje na zaproponowane im kombinacje upustów i rabatów. W tym przypadku pozytywna decyzja o pozostaniu w sieci zarówno przy zaproponowaniu niskiej, jak i wysokiej kombinacji zniżek, jest o wiele trudniejsza do ustalenia.

Opracowane tablice stanowią podstawę przy wyznaczaniu funkcji celu, której zadaniem jest wskazanie kombinacji upustu i rabatu przynoszących najmniejsze potencjalne straty operatorowi, uwzględniając iż klient może pozostać w sieci i z uwagi na otrzymane zniżki stać się mniej zyskownym lub – z drugiej strony – zrezygnować z usług danego operatora.

3.4. WYZNACZANIE FUNKCJI CELU

Na podstawie ekonomicznej analizy zagadnienia, w oparciu o elementy teorii preferencji rozmytych [4], przyjęta została następująca postać funkcji celu $\mathcal{P}$:

$$\mathscr{P}(d_a, d_m; x_i, y_{i,1}, w, \alpha, \beta) = \frac{\int_{-\infty}^{\infty} x\mu_{(w,\alpha,\beta)}(x)dx}{\int_{-\infty}^{\infty} \mu_{(w,\alpha,\beta)}(x)dx} \sum_{i \in \{i_1, i_2, \ldots, i_I\}} x_{i,3}(y_{i,1} + b) +$$
$$+ \left(6 - \frac{\int_{-\infty}^{\infty} x\mu_{(w,\alpha,\beta)}(x)dx}{\int_{-\infty}^{\infty} \mu_{(w,\alpha,\beta)}(x)dx}\right) \sum_{i \in (i_1, i_2, \ldots, i_I\}} x_{i,3}(d_a \cdot y_{i,1} + d_m \cdot b), \tag{13}$$

gdzie i_1, i_2, ..., i_I oznaczają indeksy tych elementów zbioru danych, które zostały zaliczone do rozważanego skupienia, I oznacza jego liczność, b jest iloczynem średniej stawki za minutę połączenia oraz średniego czasu połączeń generowanego poza pakietem objętym opłatą abonamentową. W przypadku funkcji przynależności danej dla liczb rozmytych typu L-R wzorami (10)-(12), i typowych postaci funkcji L oraz R, jej wartość będzie można wyrazić za pomocą dogodnego wzoru analitycznego.

Równanie (13) określające funkcję celu złożone jest z dwóch składników. Pierwszy z nich przedstawia iloczyn średniej rozmytej oceny odejścia oraz straty wynikłej z rezygnacji z usług, natomiast drugi stanowi iloczyn dopełnienia do średniej rozmytej oceny odejścia i straty powstałej w wyniku przyznania zniżek. Wielkości $y_{i,1} + b$ oraz $d_a \cdot y_{i,1} + d_m \cdot b$ reprezentują wartości potencjalnych strat w przypadkach – odpowiednio – nieprzyznania zniżek i rezygnacji z usług, a także zmniejszenia przychodów operatora wynikłych z przyznania rabatów i upustów.

Parametr b wprowadzony we wzorze (13) jest stały w relatywnie długich okresach czasu i obecnie jego wartość wynosi 24 PLN, jednakże ulega ona okresowej aktualizacji. Wartość tego parametru stanowi iloczyn średniej stawki za minutę połączenia przyjętej jako 0,40 PLN oraz 60 minut czasu połączeń. Ze względu na dynamiczny rozwój rynku telefonii mobilnej, jak również panującą na nim silną konkurencję, skutkującą sukcesywnym obniżaniem stawek za minutę połączenia zarówno u operatorów mobilnych, jak i stacjonarnych, powyższa stawka powinna być uaktualniana co 3-6 miesięcy, aczkolwiek w przypadku utrzymywania się stabilnej sytuacji, akceptowalne może być wydłużenie tego czasokresu do 12 miesięcy.

Tablice ocen rezygnacji określane są odrębnie dla kolejnych skupień. Jak podkreślano wcześniej, wyznaczenie dla każdego z nich 36 ocen dla poszczególnych wariantów rabatów i upustów, w praktyce nie stanowi dla eksperta nadmiernej trudności, biorąc pod uwagę dostępny czas przygotowania takiej informacji oraz istnienie zauważalnej systematyki wartości tych ocen. Następnie dla każdego ze skupień wyznaczane jest minimum funkcji celu, co pozwala wskazać, który wariant wartości upustu na abonament oraz rabatu za minutę połączenia wydaje się najkorzystniejszy z punktu widzenia spodziewanego zysku operatora. Ze względu na możliwość uelastycznienia indywidualnych negocjacji możliwe jest także ewentualne wskazanie tych zniżek, dla których wartość funkcji celu jest niewiele większa od minimalnej. W procesie negocjacji niezmiernie bowiem ważne jest wykorzystanie alternatywnych upustów i rabatów, które

z punktu widzenia klienta mogą wydawać się znacznie korzystniejsze, a operatorowi przynoszą zysk niewiele mniejszy.

W kolejnej podsekcji zostanie przedstawiona procedura klasyfikacji konkretnego klienta do jednego z otrzymanych skupień, co w połączeniu z powyższymi wynikami wyczerpuje prezentowaną tu procedurę.

3.5. KLASYFIKACJA KLIENTA

W wyniku przeprowadzonej dotychczas realizacji kolejnych etapów analizy danych dokonano dekompozycji elementów bazy klientów korporacyjnych na rozłączne grupy i ustalono dla każdej z nich odpowiednią strategię wspomagającą proces decyzyjny. Aby skorzystać z powyższych wyników w przypadku negocjacji prowadzonych z konkretnym abonentem, konieczne było przypisanie go do jednej z wyróżnionych w ten sposób grup. Celowi temu służy klasyfikacja klienta przedstawiona w poniższym podrozdziale.

Rozważany klient korporacyjny, z którym prowadzone są negocjacje dotyczące udzielenia mu zniżek, scharakteryzowany jest za pomocą trzech wielkości, w nawiązaniu do wzoru (9) zapisanych jako:

$$\tilde{x} = \begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \tilde{x}_3 \end{bmatrix}, \tag{14}$$

gdzie $\tilde{x}_1$ oznacza jego średni miesięczny przychód na kartę SIM, $\tilde{x}_2$ – staż w sieci, natomiast $\tilde{x}_3$ – liczbę kart SIM tegoż klienta.

W tym przypadku – jeśli prowadzone są renegocjacje warunków umowy – dane te mogą dotyczyć dotychczasowej historii abonenta w konkretnej sieci. Mogą one także stanowić informacje historyczne pochodzące od konkurencji, gdy została podjęta próba przejęcia klienta. Istnieje również możliwość nawiązania współpracy z całkowicie nowym użytkownikiem usług telefonii mobilnej – dane potrzebne do przeprowadzenia jego klasyfikacji oparte zostały na oszacowaniach wynikających ze znajomości rynku oraz na planowanej strukturze usług świadczonych temu klientowi. Wielkości charakteryzujące każdego klienta przyjęte do opracowanej analizy są dostępne wprost, na przykład na podstawie danych zawartych w miesięcznych fakturach za korzystanie z usług telekomunikacyjnych. Nie wymagają one dodatkowych przekształceń, jak również specjalnych opracowań. W przypadku wszystkich operatorów potrzebne do klasyfikacji informacje przyjmują formę właściwie identycznych raportów. Dodatkowe, nie zawarte w fakturach dane, na przykład informacja o stażu w danej sieci mogą być łatwo uzupełnione z dokumentacji operatora lub – w przypadku próby przejęcia klienta – na podstawie przedstawionych dokumentów.

Podstawą klasyfikacji jest procedura przedstawiona w podsekcji 2.4. W jej wyniku, podlegający klasyfikacji element (14) zostaje przypisany do konkretnej klasy (skupienia). Zniżka jaka powinna być wówczas zaproponowana została już wyznaczona w poprzedniej podsekcji. Należy podkreślić, iż umożliwia to stosowanie tej części algorytmu w czasie rzeczywistym w trakcie pertraktacji z klientem.

Tym samym zostaje ostatecznie skompletowana formuła procedury będącej przedmiotem badań prezentowanych w niniejszym artykule.

4. WERYFIKACJA NUMERYCZNA

Tematykę poniższej sekcji stanowi weryfikacja poprawności działania prezentowanej w niniejszej pracy procedury wspomagania strategii marketingowej operatora telefonii komórkowej w odniesieniu do klienta korporacyjnego. Została ona przeprowadzona na podstawie rzeczywistych danych uzyskanych od jednego z operatorów telefonii komórkowej w Polsce. W warunkach użyczenia zawarta została klauzula nieujawniania nazwy owego operatora, a także przedstawiania otrzymywanych wyników w formie gwarantującej uniemożliwienie skojarzenia poszczególnych elementów bazy danych z konkretnymi abonentami.

Statystyczne estymatory jądrowe zostały wykorzystane do zagadnień eliminacji elementów odosobnionych, analizy skupień i klasyfikacji. Zastosowanie jednej metodyki do wszystkich wymienionych zadań znacząco ułatwiło jej użycie, interpretacje uzyskanych wyników częściowych, a także umożliwiło wielokrotne wykorzystanie w programie poszczególnych bloków funkcjonalnych o uniwersalnym znaczeniu. Wieloetapowość opracowanej metody umożliwiła dodatkowo naturalny podział programu na niezależne moduły, znacznie ułatwiające zarówno sam proces pisania kodu, jak i weryfikację poszczególnych części.

Użyte zostało jądro normalne, z uwagi na jego dużą efektywność, a jednocześnie różniczkowalność i przyjmowanie dodatnich wartości w całej dziedzinie.

4.1. WSTĘPNA ANALIZA DANYCH

Przystępując do badań dysponowano pełnym zbiorem 2264 klientów rynku biznesowego, na koncie których liczba aktywnych kart SIM była większa lub równa 20. Każdy klient charakteryzowany był przez średni miesięczny przychód na kartę SIM, jego staż w sieci, liczbę kart SIM oraz dodatkowo wielkości pomocnicze określające jego średni abonament, przynależność do jednostek prestiżowych, wielkość przedsiębiorstwa, a także wyróżniające firmy transportowe i typu ochrona-monitoring. Najpierw dokonano standaryzacji pierwszych trzech wielkości.

Po wstępnej analizie struktury danych dokonano eliminacji części elementów bazy danych, w wyniku której zostały usunięte firmy liczące mniej niż 30 kart SIM, co w efekcie zmniejszyło ich liczbę do 1825. Ograniczenie to wynikło z faktu, iż firmy o liczbie kart SIM mieszczącej się w granicach 20-29 nie posiadały typowych cech klienta korporacyjnego. Dzięki temu zwiększyła się jednorodność bazy danych i w konsekwencji osiągnięto lepszą reprezentatywność charakterystyki klientów korporacyjnych.

4.2. ELIMINACJA ELEMENTÓW ODOSOBNIONYCH

W wyniku zastosowania procedury eliminacji elementów odosobnionych, z badanego zbioru usuniętych zostało 186 klientów. Wśród zidentyfikowanych elementów

odosobnionych nie znalazły się jednostki o znaczeniu kluczowym ani wszystkie firmy transportowe lub typu ochrona-monitoring. Zbiór 186 wyeliminowanych abonentów zawierał przedstawicieli wszystkich typów działalności.

4.3. ANALIZA SKUPIEŃ

W kolejnym etapie, na powstałej po wyeliminowaniu elementów odosobnionych próbie, zastosowano algorytm analizy skupień. Wyniki uzyskane dla typowej intensywności modyfikacji parametru wygładzania (ustanawianej przez wartość $c = 0{,}5$, por. [11 – podrozdział 3.1.6, 12]), wskazywały na zbytnią ilość skupień o niedużej liczności, ulokowanych w obszarach małego zagęszczenia elementów próby, a także zbyt liczne główne skupienie zawierające ponad połowę elementów. Zgodnie z własnościami stosowanego algorytmu [12], wartość ta została powiększona do $c = 1$. Uzyskano dzięki temu żądany efekt: liczba „peryferyjnych" skupień istotnie zmniejszyła się, a główne skupienie uległo rozbiciu. Otrzymana liczba skupień była zadowalająca, co spowodowało iż ewentualna zmiana wartości parametru wygładzania okazała się zbędna.

Ostatecznie uzyskano podział rozważanej teraz 1639-elemntowej próby na 26 skupień o następujących licznościach: 488, 413, 247, 128, 54, 41, 34, 34, 33, 28, 26, 21, 20, 14, 13, 12, 10, dwa skupienia 4-elementowe, trzy skupienia 3-elementowe, dwa skupienia 2-elementowe i dwa skupienia 1-elementowe. Warto zwrócić uwagę na wyraźnie zarysowane cztery grupy**:** pierwsza z nich to dwa duże skupienia o licznościach 488 i 413, następnie dwa średnie 247- i 128-elementowe, po czym małe – dziewięć liczących od 20 do 54 oraz wreszcie 13 skupień zawierających mniej niż 20 elementów.

Następnie przystąpiono do eliminacji skupień o licznościach mniejszych niż 20, ale z wyłączeniem tych, które zawierają klientów kluczowych (skupienia 14-, 13- i 10-elementowe) i składających się co najmniej w połowie z klientów prestiżowych (skupienie 12-elementowe).

Pozostawione skupienia zawierały firmy transportowe oraz typu ochrona-monitoring, co spowodowało, że warunek zapewnienia reprezentacji tego typu jednostek został spełniony bez dodatkowej ingerencji.

Ostatecznie do dalszej analizy pozostało 17 skupień.

4.4. BUDOWANIE TABLIC OCEN REZYGNACJI

Kolejnym etapem było określenie tablic ocen rezygnacji dla każdego z wyznaczonych powyżej skupień. Podstawą zawartych tam ocen była ekspercka wiedza, oparta na dostępnych charakterystykach jednostek wchodzących w skład poszczególnych skupień. Przykładowo przedstawiona zostanie poniżej tablica (tab. 1) przypisana do skupienia grupującego 54 klientów.

Tabela 1

Tablica ocen rezygnacji dla skupienia 54-elemntowego

(w, α, β)		d_m					
		0%	3%	6%	12%	25%	50%
d_a	0%	(4,0,2)	(4,0,1)	(4,0,1)	(4,2,1)	(4,2,1)	(4,2,0)
	5%	(4,0,1)	(4,0,1)	(4,1,1)	(3,1,2)	(3,1,2)	(3,2,1)
	10%	(4,0,1)	(4,1,1)	(4,2,1)	(3,2,1)	(3,2,1)	(3,2,0)
	15%	(3,1,2)	(3,2,2)	(3,2,1)	(3,2,1)	(3,2,0)	(3,2,0)
	50%	(3,1,2)	(3,2,1)	(3,2,0)	(2,1,1)	(2,2,0)	(2,2,0)
	99%	(2,1,2)	(2,1,1)	(2,2,1)	(2,2,0)	(2,2,0)	(1,1,0)

Dokonana analiza wchodzących w jego skład elementów wskazała, że skupienie to składa się wyłącznie z klientów dużych, a nawet bardzo dużych. Ponad połowę elementów stanowią tu firmy transportowe. Średnia liczba kart SIM dla tego skupienia wynosi prawie 300, przeciętna wysokość abonamentu kształtuje się na średnim poziomie 54 PLN, aczkolwiek wskaźnik ARPU jest relatywnie duży (około 70 PLN).

Zdaniem eksperta, ryzyko rezygnacji z usług zależy tu głównie od wysokości udzielonego rabatu za minutę połączenia. Dotyczyć on powinien nie tylko połączeń głosowych krajowych, ale w równej mierze połączeń roamingowych. Warte podkreślenia jest to, iż w ramach nawet wyższej opłaty abonamentowej, klient – w tym przypadku przeważnie przedsiębiorstwo transportowe – oczekiwać będzie stałych, możliwie najniższych stawek za usługi związane z intensywnością rozmów, a ogólniej transmisji danych, które generują wysokie koszty jednostkom prowadzącym tego typu działalność.

W rozpatrywanym przypadku oceny rezygnacji z usług uwzględniają – poprzez średnie wartości dla małych zniżek – niedużą możliwość utraty tego typu klienta, głównie wynikłą stąd, iż zmiana operatora dla jednostek prowadzących działalność transportową jest wyjątkowo kłopotliwa, ze względu na dużą mobilność użytkowników poszczególnych numerów.

4.5. WYZNACZANIE FUNKCJI CELU DLA POSZCZEGÓLNYCH SKUPIEŃ

Kolejnym etapem prezentowanej procedury było wyznaczenie dla każdego elementu tablic ocen rezygnacji wartości funkcji celu (13), a następnie wskazanie tych, dla których osiąga ona wartości najmniejsze. Dzięki szczególnej postaci liczb L-R danej wzorami (10)-(12), w przypadku gdy występujące tam funkcje przynależności mają postać liniową, ściślej opisaną następującą formułą

$$L(x) = \begin{cases} 0 & \text{gdy } x \in (-\infty, -1) \\ 1 + x & \text{gdy } x \in [-1, 0] \end{cases} \tag{15}$$

$$R(x) = \begin{cases} 1 - x & \text{gdy } x \in [0,1] \\ 0 & \text{gdy } x \in (1,\infty) \end{cases}, \tag{16}$$

to funkcję celu można wyrazić dogodnym do obliczeń wzorem analitycznym. Należy zaznaczyć iż postać liniowa (15)-(16) została również wskazana przez eksperta.

Niech zatem, dla dowolnie ustalonego skupienia, $i_1, i_2, \ldots, i_I$ oznaczają indeksy tych elementów zbioru danych, które zostały do niego zaliczone, natomiast I oznacza jego liczność. Uwzględniając zależności (10)-(12) oraz (15)-(16) funkcja celu przyjęła następującą postać analityczną:

$$\begin{aligned} \mathscr{P}\left(d_a, d_m; x_i, y_{i,1}, w, \alpha, \beta\right) = {} & \frac{\alpha(3w - \alpha) + \beta(\beta + 3w)}{3(\alpha + \beta)} \sum_{i \in (i_1, i_2, \ldots, i_I\}} x_{i,3}\left(y_{i,1} + b\right) + \\ & + \left(6 - \frac{\alpha(3w - \alpha) + \beta(\beta + 3w)}{3(\alpha + \beta)}\right) \sum_{i \in (i_1, i_2, \ldots, i_I\}} x_{i,3}\left(d_a \cdot y_{i,1} + d_m \cdot b\right), \end{aligned} \tag{17}$$

gdy $\alpha, \beta > 0$, a także dla wartości indeksów $\alpha = 0$ albo $\beta = 0$ odpowiednio

$$\begin{aligned} \mathscr{P}\left(d_a, d_m; x_i, y_{i,1}, w, 0, \beta\right) = {} & \frac{\beta(\beta + 3w)}{3\beta} \sum_{i \in (i_1, i_2, \ldots, i_I\}} x_{i,3}\left(y_{i,1} + b\right) + \\ & + \left(6 - \frac{\beta(\beta + 3w)}{3\beta}\right) \sum_{i \in (i_1, i_2, \ldots, i_I\}} x_{i,3}\left(d_a \cdot y_{i,1} + d_m \cdot b\right) \quad \text{dla } \beta > 0, \end{aligned} \tag{18}$$

$$\begin{aligned} \mathscr{P}\left(d_a, d_m; x_i, y_{i,1}, w, \alpha, 0\right) = {} & \frac{\alpha(3w - \alpha)}{3\alpha} \sum_{i \in (i_1, i_2, \ldots, i_I\}} x_{i,3}\left(y_{i,1} + b\right) + \\ & + \left(6 - \frac{3w - \alpha}{\alpha}\right) \sum_{i \in (i_1, i_2, \ldots, i_I\}} x_{i,3}\left(d_a \cdot y_{i,1} + d_m \cdot b\right) \quad \text{dla } \alpha > 0. \end{aligned} \tag{19}$$

Zgodnie z powyższym wobec każdego z 17 skupień obliczone zostały wartości wszystkich 36 pól związanej z nimi tablicy ocen rezygnacji. Następnie dla każdego skupienia wyznaczana została wartość minimalna oraz dwie wartości jej najbliższe. Odpowiadające im pary rabatów i upustów wskazały na wartość najkorzystniejszą z punktu widzenia strategii operatora, a także na dwie niewiele „gorsze", których celem było uelastycznienie procesu negocjacji. W przypadku tych dwóch ocen podany także został relatywny wzrost wskaźnika jakości wobec wartości najmniejszej. Wyraźnie zauważalna była zgodność przedstawionych tam wartości z dokonaną wcześniej analizą ekspercką, dotyczącą poszczególnych skupień.

Na tym etapie została zakończona faza dotycząca wstępnej analizy i przetwarzania danych, w rezultacie której uzyskano wyniki umożliwiające zakwalifikowanie rozważanego klienta do odpowiedniego skupienia oraz wskazanie najkorzystniejszych w jego przypadku wariantów upustów i rabatów.

4.6. KLASYFIKACJA KLIENTA

W celu ilustracji uzyskiwanych w ostatnim etapie wyników, poniżej rozważono rzeczywistego klienta, którego poddano procedurze klasyfikacji, zgodnie z metodyką przedstawioną w podsekcji 2.4. Klient ten został zaliczony do przykładowo rozważanego w podsekcji 4.4, 54-elementowego skupienia. Było to przedsiębiorstwo transportowe, działające na rynku krajowym oraz w Europie Wschodniej. Klient ten oczekiwał, w ramach renegocjacji warunków umowy, obniżenia kosztów miesięcznych. Duża wartość jego ARPU jest wynikiem przede wszystkim rozmów prowadzonych za granicą. Abonent został scharakteryzowany przez następujący wektor wielkości:

$$\tilde{x} = \begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \tilde{x}_3 \end{bmatrix} = \begin{bmatrix} 35 \\ 80 \\ 250 \end{bmatrix}. \tag{20}$$

Jak wspomniano, po zastosowaniu procedury klasyfikacji klienta tego przypisano do 54-elementowego skupienia. W przypadku tego skupienia wskazane zostały następujące zniżki:

– wariant najkorzystniejszy – upust na abonament $d_a = 15\%$, rabat za minutę połączenia $d_m = 25\%$;

– I wariant alternatywny – upust na abonament $d_a = 10\%$, rabat za minutę połączenia $d_m = 12\%$, wartość wskaźnika jakości większa o 0,75%;

– II wariant alternatywny – upust na abonament $d_a = 15\%$, rabat za minutę połączenia $d_m = 6\%$, wartość wskaźnika jakości większa o 2,46%.

Klient ten był zaklasyfikowany do skupienia, do którego należała również duża część innych jednostek transportowych, dzięki czemu oceny eksperta, na podstawie których wyznaczane były wartości funkcji celu, uwzględniły preferencje właściwe ich specyfice. Zaproponowane zniżki dla abonamentu na poziomie 10 i 15% oraz dla stawki za minutę 6%, 12% i 25% mogą być dla niego zadowalające przy założeniu, że korekcie podlegają także kwoty generowane przez klienta w roamingu.

5. KOŃCOWE KOMENTARZE I PODSUMOWANIE

W ostatniej dekadzie ubiegłego wieku rozpoczął się gwałtowny rozwój rynku telefonii komórkowej w Polsce. Obecnie nie widać przesłanek do oczekiwania spowolnienia tych procesów, natomiast można spodziewać się wzrostu konkurencji, poprzez pojawianie się nowych operatorów i zaostrzanie rywalizacji cenowej. Bardzo istotne w związku z tym staje się poszukiwanie nowoczesnych rozwiązań o charakterze systemowym, które pozwalają na tworzenie narzędzi skutecznych w procesie wspomagania decyzji dotyczącej określenia optymalnej strategii postępowania wobec konkretnego klienta, w celu maksymalizacji zysku operatora przy jednoczesnym możliwie dużym zadowoleniu abonentów.

Przedmiotem prezentowanego artykułu były opracowanie i weryfikacja procedury realizującej powyższy cel wobec klienta korporacyjnego, możliwej do zastosowania

w czasie rzeczywistym podczas negocjacji. Metodykę oparto na współczesnych metodach technik informacyjnych w zakresie analizy i eksploracji danych. W szczególności zastosowano koncepcję statystycznych estymatorów jądrowych oraz elementy logiki rozmytej. Estymatory jądrowe zostały wykorzystane do zagadnień eliminacji elementów odosobnionych, analizy skupień i klasyfikacji. Zastosowanie do wszystkich tych zadań jednolitej metodyki znacząco ułatwiło jej użycie, interpretacje uzyskiwanych wyników częściowych, a także umożliwiło wielokrotne wykorzystanie w programie poszczególnych bloków programowych[2].

Powstała metoda pozwala operatorowi sieci komórkowej aktywnie reagować na potrzeby zgłaszane przez klienta w procesie prowadzenia z nim negocjacji. Opracowany algorytm, przede wszystkim dzięki wzbogaceniu wiedzą eksperta, nie stanowi jedynie matematycznego aparatu służącego do kalkulacji powszechnie analizowanych w takich sytuacjach wskaźników lub weryfikowania wielkości abonamentów dla dostępnych planów taryfikacyjnych, na podstawie czego budowane są oferty utrzymaniowe i przejęciowe, jak to ma powszechnie miejsce w dotychczasowej praktyce branżowej. Koncepcja przedstawionego rozwiązania nie opiera się także na popularnej obecnie analizie *churnu*, czyli wskaźnika określającego poziom rezygnacji z usług. Podstawą zaproponowanej w niniejszej pracy metody są informacje zawarte w bazie danych klientów, a także wiedza eksperta o rynku, których połączenie z nowoczesnymi technikami informacyjnymi, stworzyło możliwość opracowania metody o charakterze systemowym pozwalającym na efektywne utrzymywanie dotychczasowych i pozyskiwanie nowych klientów z zapewnieniem maksymalnego zysku operatora.

Poprawność opracowanej tu metody została zweryfikowana z użyciem danych rzeczywistych. Analiza uzyskanych rezultatów potwierdziła ich poprawność zarówno w zakresie interpretacji poszczególnych skupień, jak i wyników otrzymanych na przykładach nowych klientów. Wyraźnie można było zauważyć, iż w skład poszczególnych skupień zaliczone zostały jednostki podobne z punktu widzenia operatora, a niekoniecznie rodzaju reprezentowanej przez nie branży. Przykładowo, firmy typu ochrona-monitoring ulokowały się pojedynczo w wielu różnych skupieniach, a w przeciwieństwie, jednostki transportowe praktycznie skupiły się w nielicznych skupieniach, zwłaszcza jednym, w którym stanowiły połowę jego liczności. Jedno ze skupień zawierało swoiste „zbiorowisko" różnorodnych jednostek, natomiast inny obejmował firmy „zaniedbane", dla których nie wynegocjowano dotychczas korzystnych warunków umów. Najważniejszą wielkością stanowiącą o skupianiu się jednostek była wartość średniego przychodu na kartę SIM, a w dalszej kolejności liczba kart SIM, i wreszcie staż w sieci. Także rozkład liczności skupień był zbieżny z oczekiwaniami. Dwa duże zawierały 27% i 23%

[2] Warto w tym miejscu wspomnieć, iż w literaturze czasem rekomenduje się dla potrzeb procedury klasyfikacji (podsekcja 2.4) metody wyznaczania wartości parametru wygładzania estymatora jądrowego oparte na kryterium minimalizacji błędu klasyfikacji (por. [10] – podrozdział 2.4), odmienne od ogólnie stosowanych, wykorzystujących kryterium scałkowanego błędu średniokwadratowego (por. [11] – podrozdział 3.1.5). Tych pierwszych nie da się jednak użyć do algorytmu analizy skupień (podsekcja 2.3). W prezentowanych badaniach konsekwentnie stosowane były zatem metody wykorzystujące kryterium scałkowanego błędu średniokwadratowego. W przeciwnym bowiem przypadku, użycie różnych kryteriów (a zatem różnych wartości parametru wygładzania, czyli w konsekwencji także różnych estymatorów jądrowych $\hat{f}$) do wyznaczania wzorców poszczególnych klas za pomocą algorytmu analizy skupień, po czym do dokonywanej na podstawie tych wzorców klasyfikacji, mogłoby w praktyce skutkować trudnymi do przewidzenia rezultatami.

i kolejne dwa o średniej wielkości 14% i 7% elementów początkowej bazy – były to elementy o typowych charakterystykach. Reszta skupień zawierała mniej niż 3%, najczęściej nietypowych, specyficznych firm. Także kończące proces weryfikacji przykłady wyników uzyskiwanych dla nowych klientów dawały wskazania właściwych – z punktu widzenia ich specyfiki i uwarunkowań operatora – propozycji rabatów i upustów.

Instytut Badań Społecznych, PAN

LITERATURA

[1] Barnett V., Lewis T., [1994], *Outliers in Statistical Data*, Wiley, Chichester.
[2] Burnett K., [2002], *Relacje z kluczowymi klientami. Analiza i zarządzanie*, Oficyna Ekonomiczna, Warszawa.
[3] Dyche J., [2002], *CRM. Relacje z klientami*, Helion, Warszawa.
[4] Fodor J., Roubens M., [1994], *Fuzzy Preference Modelling and Multicriteria Decision Support*, Kluwer, Dordrecht.
[5] Fukunaga K., Hostetler L.D., [1975], The estimation of the gradient of a density function, with applications in pattern recognition, IEEE Transactions on Information Theory, 21(1975)32-40.
[6] Hołubowicz W., Płóciennik P., [1994], *Cyfrowe systemy telefonii komórkowej GSM 900, GSM 1800, UMTS*, Holkom, Poznań.
[7] Hryniewicz O., [2004], *Wykłady ze statystyki dla studentów informatycznych technik zarządzania*, Wydawnictwo WSISiZ, Warszawa.
[8] Jagoda A., de Villepin M., [1993], *Mobile communications*, Wiley, Chichester.
[9] Kacprzyk J., [1986], *Zbiory rozmyte w analizie systemowej*, PWN, Warszawa.
[10] Krzyśko M., Wołyński W., Górecki T., Skorzybut M., [2008], *Systemy uczące się. Rozpoznawanie wzorców, analiza skupień i redukcja wymiarowości*, WNT, Warszawa.
[11] Kulczycki P., [2005], *Estymatory jądrowe w analizie systemowej*, WNT, Warszawa.
[12] Kulczycki P., Charytanowicz M., A Complete Gradient Clustering Algorithm Formed with Kernel Estimators, Applied Mathematics and Computer Science, 19(2009), w druku.
[13] Lake N., [2005], *Planowanie strategiczne w firmie*, Onepress, Gliwice.
[14] Silverman B.W., [1986], *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.
[15] Wand M.P., Jones M.C., [1995], *Kernel Smoothing*, Chapman and Hall, London.
[16] Wesołowski K., [1999], *Systemy radiokomunikacji ruchomej*, WKiŁ, Warszawa.

Praca wpłynęła do redakcji w marcu 2009 r.

METODA WSPOMAGANIA STRATEGII MARKETINGOWEJ OPERATORA TELEFONII KOMÓRKOWEJ

Streszczenie

Przedmiotem prezentowanych badań jest zadanie wspomagania decyzji w zakresie wyznaczania strategii postępowania wobec klienta korporacyjnego – abonenta sieci telefonii komórkowej. Badania przeprowadzono na rzeczywistej bazie danych, uzyskanej od jednego z polskich operatorów sieci GSM. Wzmiankowana strategia została określona z zastosowaniem metodyki statystycznych estymatorów jądrowych, wykorzystywanej tu do zagadnień wykrywania elementów nietypowych (odosobnionych), analizy skupień (klasteryzacji) i klasyfikacji. Z kolei elementów logiki rozmytej użyto do oceny możliwych do zaproponowania kombinacji

zniżek, natomiast teoria preferencji rozmytych została zastosowana przy wyznaczaniu wariantu postępowania jak najkorzystniejszego dla operatora w sensie maksymalizacji spodziewanego zysku.

Słowa kluczowe: analiza i eksploracja danych, techniki informacyjne, analiza systemowa, strategia marketingowa, rynek telefonii komórkowej, statystyczne estymatory jądrowe, logika rozmyta.

A METHOD FOR SUPPORTING THE MARKETING STRATEGY OF A MOBILE PHONE NETWORK PROVIDER

Summary

The subject of the research presented in this paper is the task of decision support concerning strategy of behavior towards a corporate client – a mobile phone network subscriber. Investigations were carried out on a real database made available by one of the Polish GSM network operators. This strategy was defined with the statistical kernel estimators methodology, applied to the problem of discovering atypical elements (outliers), clustering and classification. In turn, elements of fuzzy logic were used to assess different possible combinations of discounts to offer, while the theory of fuzzy preferences was applied in selecting the most advantageous – in the sense of maximizing expected profit – variant of behavior for the operator.

Key words: data analysis and mining, information technology, systems analysis, marketing strategy, mobile phone market, statistical kernel estimators, fuzzy logic.

PRZEGLĄD STATYSTYCZNY
R. LVI – ZESZYT 2 – 2009

ANNA WITASZCZYK

ESTYMACJA UOGÓLNIONEJ WARIANCJI WYBRANYCH ROZKŁADÓW WIELOWYMIAROWYCH

1. WPROWADZENIE

Potrzeba stosowania metod statystyki wielowymiarowej występuje w wielu dziedzinach nauki. Podstawowym założeniem większości metod wielowymiarowej analizy statystycznej jest to, że próba pochodzi z populacji o wielowymiarowym rozkładzie normalnym. Celem pracy jest przedstawienie problemu estymacji uogólnionej wariancji, czyli wyznacznika macierzy kowariancji, w przypadku obserwacji pochodzących z populacji o rozkładzie wielowymiarowym niekoniecznie normalnym. Przedstawiono tzw. G-estymator uogólnionej wariancji otrzymany przez Girko w oparciu o twierdzenia graniczne dla wyznaczników losowych przy bardzo ogólnych założeniach dotyczących rozkładu wektora losowego [np. 4, 5, 6]. Własności G-estymatora zostały zbadane za pomocą eksperymentów symulacyjnych.

2. UOGÓLNIONA WARIANCJA

Koncepcja wariancji, jako miary koncentracji rozkładów jednowymiarowych, może być rozszerzona na przypadek rozkładów wielowymiarowych na dwa sposoby: jako obiekt geometryczny zwany elipsoidą koncentracji oraz jako miara liczbowa, która nazywa się uogólnioną wariancją.

Niech $\vec{\xi}$ będzie m wymiarowym wektorem losowym o wektorze wartości oczekiwanych $\mathbf{a} = E\vec{\xi}$ i macierzy kowariancji $\mathbf{R}_m = E\left(\vec{\xi} - \mathbf{a}\right)\left(\vec{\xi} - \mathbf{a}\right)^T$. Wyznacznik macierzy kowariancji nazywa się uogólnioną wariancją wektora $\vec{\xi}$.

Niech $\mathbf{x}_1, \ldots, \mathbf{x}_n$ będą niezależnymi obserwacjami wektora losowego $\vec{\xi}$ i niech $\mathbf{A} = \sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$, gdzie $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$. Macierz $\hat{\mathbf{R}}_m = \frac{1}{n-1}\mathbf{A}$ nazywa się próbkową macierzą kowariancji. Jeżeli próba pochodzi z rozkładu normalnego $N_m(\mathbf{a}, \mathbf{R}_m)$, to $\hat{\mathbf{R}}_m$ jest nieobciążonym estymatorem parametru $\mathbf{R}_m$. Próbkowa macierz kowariancji ma rozkład Wisharta z $n-1$ stopniami swobody i macierzą kowariancji $\frac{1}{n-1}\mathbf{R}_m$ [np. 2, 8, 9].

Jeżeli wykładnik funkcji gęstości wielowymiarowego rozkładu normalnego $(\mathbf{x} - \mathbf{a})^T \mathbf{R}_m^{-1}(\mathbf{x} - \mathbf{a})$ przyrównamy do pewnej stałej c, to tak otrzymane równanie jest równaniem elipsoidy w m wymiarowej przestrzeni euklidesowej. Zmieniając c

otrzymuje się rodzinę elipsoid o wspólnym środku w punkcie **a**, które nazywa się elipsoidami koncentracji. Pierwszą osią główną każdej z elipsoid nazywa się prostą przechodzącą przez środek elipsoidy oraz przez punkt najbardziej od tego środka oddalony. Długość pierwszej osi głównej jest równa $2\sqrt{\lambda_1 c}$, gdzie λ_1 oznacza największą wartość własną macierzy $\mathbf{R}_m$, natomiast jej kierunek określony jest przez kosinusy kierunkowe unormowanego wektora własnego α_1 odpowiadającego wartości własnej λ_1. Druga oś określona jest przez wektor własny odpowiadający drugiej co do wielkości wartości własnej macierzy $\mathbf{R}_m$ itd.

W analizie wielowymiarowej wiele statystyk można interpretować za pomocą objętości, są one odpowiednikami odległości, z którymi mamy do czynienia w przypadku jednowymiarowym. Objętość dowolnej elipsoidy jest równa $\dfrac{\pi^{\frac{m}{2}} c^{\frac{m}{2}} \left|\mathbf{R}_m\right|^{\frac{1}{2}}}{\Gamma\left(\frac{1}{2}m + 1\right)}$ gdzie $c > 0$. Jeżeli więc porównuje się tylko objętości elipsoid koncentracji dwóch rozkładów, to jest to równoważne z porównaniem ich uogólnionych wariancji.

Analogicznie do uogólnionej wariancji, uogólnioną wariancją z próby nazywa się wyznacznik macierzy $\hat{\mathbf{R}}_m$.

Podamy geometryczną interpretację próbkowej uogólnionej wariancji. Niech $\mathbf{t}_i = \mathbf{x}_i - \bar{\mathbf{x}}$, $i = 1, \ldots, n$ i

$$\mathbf{Y}^T = \begin{bmatrix} \mathbf{y}_1^T \\ \vdots \\ \mathbf{y}_m^T \end{bmatrix} = \left[\mathbf{t}_1, \ldots, \mathbf{t}_n\right], \quad \mathbf{y}_i \in R^n.$$

Współrzędnymi wektora $\mathbf{y}_i$ są i-te składowe wektorów $\mathbf{t}_1, \ldots, \mathbf{t}_n$. Wtedy $\mathbf{A} = \mathbf{Y}^T\mathbf{Y}$ i $a_{ii} = \mathbf{y}_i^T \mathbf{y}_i$ jest kwadratem długości wektora $\mathbf{y}_i$, a $a_{ij} = \mathbf{y}_i^T \mathbf{y}_j$ jest iloczynem długości wektorów $\mathbf{y}_i$ i $\mathbf{y}_j$ przez kosinus kąta między nimi. Rozważmy równoległościan zbudowany na tych m wektorach. Dla $m = 2$ będzie to równoległobok zbudowany na wektorach $\mathbf{y}_1$ i $\mathbf{y}_2$, dla $m = 3$ otrzymuje się równoległościan zbudowany na wektorach $\mathbf{y}_1$, $\mathbf{y}_2$, $\mathbf{y}_3$ itd. Dla dowolnego m jest to figura ograniczona parami równoległych $(m - 1)$ wymiarowych hiperpłaszczyzn, z których jedna zawiera $m - 1$ wektorów spośród $\mathbf{y}_1, \ldots, \mathbf{y}_m$, a druga, równoległa do niej, przechodzi przez koniec pozostałego wektora. Wyznacznik macierzy $\mathbf{A}$, tj. $|\mathbf{A}| = \left|(n-1)\hat{\mathbf{R}}_m\right| = (n-1)^m \left|\hat{\mathbf{R}}_m\right|$ jest równy kwadratowi objętości tak utworzonego równoległościanu [1].

Przedstawimy teraz interpretację geometryczną $|\mathbf{A}|$ wykorzystując wektory $\mathbf{t}_1, \ldots, \mathbf{t}_n$, które są punktami przestrzeni m wymiarowej. Gdy $m = 1$, to $|\mathbf{A}| = \sum_{i=1}^{n} t_{1i}^2$, czyli wyznacznik macierzy A jest równy sumie kwadratów odległości tych punktów od początku układu współrzędnych. W ogólnym przypadku wyznacznik macierzy $\mathbf{A}$ jest równy sumie kwadratów objętości wszystkich równoległościanów zbudowanych na m wektorach wybranych spośród $\mathbf{t}_1, \ldots, \mathbf{t}_n$. Biorąc pod uwagę, że $\mathbf{t}_i = \mathbf{x}_i - \bar{\mathbf{x}}$ otrzymujemy, że próbkowa uogólniona wariancja $\left|\hat{\mathbf{R}}_m\right|$ jest proporcjonalna do sumy kwadratów objętości wszystkich równoległościanów utworzonych w następujący sposób: wektorami tworzącymi każdy równoległościan jest m wektorów, których wspólnym początkiem

jest punkt $\bar{\mathbf{x}}$, a końce znajdują się w m punktach spośród $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Współczynnik proporcjonalności wynosi $\frac{1}{(N-1)^p}$.

Uogólniona wariancja z próby pochodzącej z populacji o rozkładzie normalnym ma taki rozkład jak zmienna $\frac{|\mathbf{R}_m|}{(n-1)^m}$ pomnożona przez iloczyn m niezależnych zmiennych, z których i-ta ma rozkład χ^2 z $(n-i)$ stopniami swobody ($i = 1, \ldots, m$) (por. [8]).

Dla $m = 1$ lub $m = 2$ można otrzymać dokładny rozkład $|\hat{\mathbf{R}}_m|$, jednak dla większych wartości m jest on coraz bardziej skomplikowany i trudniej powiedzieć coś o jego własnościach, dlatego wygodniej posługiwać się rozkładem asymptotycznym. Jeżeli $\hat{\mathbf{R}}_m$ jest próbkową macierzą kowariancji o wymiarach $m \times m$ i $(n-1)$ stopniach swobody otrzymaną z próby z rozkładu normalnego $N_m(\mathbf{a}, \mathbf{R}_m)$, to zmienna losowa $\sqrt{n-1}\left(\frac{|\hat{\mathbf{R}}_m|}{|\mathbf{R}_m|} - 1\right)$ ma rozkład asymptotycznie normalny z wartością oczekiwaną zero i wariancją $2m$ [1].

Nieobciążony estymator wyrażenia $\ln|\mathbf{R}_m|$ ma postać:

$\ln|\hat{\mathbf{R}}_m| - \sum_{i=1}^{m} \psi\left(\frac{n-i}{2}\right) + m \ln\frac{n-1}{2}$, gdzie $\psi(x) = \frac{d \ln \Gamma(x)}{dx}$ (por. [8]).

3. G-ESTYMATOR UOGÓLNIONEJ WARIANCJI

W większości przypadków dokładne rozkłady wyznaczników macierzy losowych są nieznane albo mają skomplikowaną postać, dlatego poszukuje się ich rozkładów granicznych. W tradycyjnym podejściu przyjmuje się, że wymiar wektora losowego jest stały (m = const) i bada się własności estymatorów, gdy n dąży do nieskończoności. Girko w swoich pracach [np. 3, 4, 5, 6] rozważa ogólniejszy przypadek, kiedy wraz ze wzrostem liczby obserwacji również liczba składowych wektora losowego dąży do nieskończoności, przy czym faktycznie m zależy od n, co zaznaczymy stosując oznaczenie m_n. Liczby m_n i n spełniają następujący tzw. G-warunek:

$$\lim_{n \to \infty} \frac{m_n}{n} = const < \infty.$$

Przy powyższym założeniu rozważany jest następujący problem: przy odpowiednim doborze stałych normujących a_n, b_n znaleźć warunki zbieżności, a także ogólną postać rozkładów granicznych dla ciągu zmiennych losowych

$$\frac{\ln|\Xi_n| - a_n}{b_n}, \quad n = 1, 2, \ldots, \tag{1}$$

gdzie $\Xi_n = \left[\xi_{ij}^{(n)}\right]$ jest kwadratową macierzą losową stopnia n.

W GSA (General Statistical Analysis) zakłada się, że elementy macierzy Ξ_n są niezależne i mają niektóre jednakowe własności. W dowodach twierdzeń granicznych wykorzystuje się między innymi metodę przekształceń ortogonalnych. Polega ona na

tym, że wyznacznik macierzy Ξ_n przedstawia się w postaci iloczynu zmiennych losowych, co pozwala na uogólnienie teorii sumowania zmiennych losowych dla logarytmu wyznacznika macierzy losowej.

Poniżej przedstawione są dwa twierdzenia, które są podstawą konstrukcji estymatora uogólnionej wariancji [5].

Przy pewnych założeniach dotyczących elementów macierzy losowej Ξ_n zachodzi następujące centralne twierdzenie graniczne dla wyznaczników losowych:

Twierdzenie 1. Niech dla każdej wartości n elementy losowe $\xi_{ij}^{(n)}$, $i, j = 1, \ldots n$ macierzy Ξ_n będą niezależne, $E\xi_{ij}^{(n)} = 0$, $D\xi_{ij}^{(n)} = 1$, $E\left(\xi_{ij}^{(n)}\right)^4 = 3$, dla pewnego $\delta > 0$

$$\sup_n \max_{i,j=1,\ldots,n} E\left|\xi_{ij}^{(n)}\right|^{4+\delta} < \infty.$$

Wtedy

$$\lim_{n\to\infty} P\left\{\frac{\ln\left|\Xi_n^2\right| - \ln(n-1)!}{\sqrt{2\ln n}} < x\right\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{y^2}{2}}\, dy.$$

Dla zmiennej $\frac{1}{c_n}\ln\left|\hat{\mathbf{R}}_m\right|$, gdzie $\{c_n\}$ jest pewnym ciągiem zachodzi słabe prawo wielkich liczb i centralne twierdzenie graniczne.

Twierdzenie 2. Niech m_n wymiarowe wektory $\mathbf{x}_1, \ldots, \mathbf{x}_n$ dla każdego $n > m_n$ będą niezależne, mają jednakowe rozkłady o wektorze wartości oczekiwanych $\mathbf{a}$ i nieosobliwej macierzy kowariancji $\mathbf{R}_{m_n}$, dla pewnego $\delta > 0$

$$\sup_n \sup_{\substack{i=1,\ldots,n \\ j=1,\ldots,m_n}} E\left|\tilde{x}_{ij}\right|^{4+\delta} < \infty \tag{2}$$

gdzie $\tilde{x}_{ij}$ są składowymi wektora $\tilde{\mathbf{x}}_i = \mathbf{R}_{m_n}^{-\frac{1}{2}}(\mathbf{x}_i - \mathbf{a})$,

$$\lim_{n\to\infty}(n - m_n) = \infty, \quad \lim_{n\to\infty}\frac{m_n}{n} = 1 \tag{3}$$

1i dla każdej wartości $n > m_n$ wektory losowe $\tilde{\mathbf{x}}_i$ będą niezależne.

Wtedy

$$p\lim_{n\to\infty}\frac{1}{c_n}\left\{\ln\left|\hat{\mathbf{R}}_{m_n}\right| + \ln\frac{(n-1)^{m_n} n}{A_{n-1}^{m_n}(n-m_n)} - \ln\left|\mathbf{R}_{m_n}\right|\right\} = 0, \tag{4}$$

gdzie $A_n^m = n(n-1)\ldots(n-m+1)$ a $\{c_n\}$ jest pewnym ciągiem takim, że $\lim\limits_{n\to\infty}\frac{1}{c_n^2}\ln\frac{n}{n-m_n} = 0$.

Jeżeli ponadto

$$E\left(\tilde{x}_{ij}\right)^4 = 3 \quad \text{dla } i = 1, \ldots, n,\ j = 1, \ldots, m_n, \tag{5}$$

to

$$\lim_{n\to\infty} P\left\{\frac{\ln\left|\hat{\mathbf{R}}_{m_n}\right| + \ln\frac{(n-1)^{m_n}n}{A_{n-1}^{m_n}(n-m_n)} - \ln\left|\mathbf{R}_{m_n}\right|}{\sqrt{-2\ln\left(1-\frac{m_n}{n}\right)}} < x\right\} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{y^2}{2}}\,dy. \tag{6}$$

Z twierdzenia 2 wynika, że jeśli spełnione są warunki (2) i (3), to za estymator wielkości $\ln|\mathbf{R}_{m_n}|$ można przyjąć

$$G\left(\hat{\mathbf{R}}_{m_n}\right) = \ln\left(\left|\hat{\mathbf{R}}_{m_n}\right|\frac{(n-1)^{m_n}n}{A_{n-1}^{m_n}(n-m_n)}\right), \tag{7}$$

Przy czym, jeżeli spełniony jest warunek (5), to zachodzi centralne twierdzenie graniczne (6). $G\left(\hat{\mathbf{R}}_{m_n}\right)$ nazywa się G-estymatorem uogólnionej wariancji.

Twierdzenie 2 można uogólnić na przypadek, gdy $n > m_n$, $\lim\limits_{n\to\infty}(n-m_n) < \infty$, $\overline{\lim\limits_{n\to\infty}}\frac{m_n}{n} \le 1$.

4. PRZYKŁADY SYMULACYJNE

W celu zbadania własności G-estymatora przeprowadzone zostały eksperymenty symulacyjne.

Niech wektor losowy $(X_1, \ldots, X_p)$ będzie p wymiarową zmienną losową o rozkładzie normalnym $N_p(\mathbf{a}, \mathbf{R})$. Według tego rozkładu wygenerowano n wektorów i na ich podstawie wyznaczono wartość estymatora $G(\hat{\mathbf{R}})$ oraz klasycznego estymatora największej wiarogodności $\ln|\hat{\mathbf{R}}|$. Eksperymenty realizowano dla $k = 10000$ powtórzeń. Obliczone zostały średnie wartości estymatorów oraz błędy średniokwadratowe. Obliczenia wykonano dla różnych wartości p, n, $\ln|\mathbf{R}|$, $\mathbf{a} = \mathbf{0}$.

Następne eksperymenty, analogiczne, zostały przeprowadzone dla innych niż normalny rozkładów wielowymiarowych: rozkładu t-Studenta, rozkładów Pearsona, wykładniczego, Laplace'a, Weibulla, Beta.

Wyniki przedstawione są w tablicach 1-7. Na ich podstawie widać, że estymator $\ln|\hat{\mathbf{R}}|$ ze wzrostem liczebności próby dąży do wartości szacowanego parametru monotonicznie, ale o wiele wolniej niż G-estymator. $G(\hat{\mathbf{R}})$ stosunkowo szybko osiąga wartości bliskie $\ln|\mathbf{R}|$ i, w przypadku rozkładu normalnego, oscyluje w otoczeniu rzeczywistej wartości parametru przyjmując wartość większą bądź mniejszą niż $\ln|\mathbf{R}|$. Dla

ustalonego p wartości błędów średniokwadratowych obu estymatorów ze wzrostem n przyjmują zbliżone wartości. Dla małych n błąd średniokwadratowy G-estymatora jest zdecydowanie mniejszy.

Dla rozkładów innych niż wielowymiarowy rozkład normalny estymator $G(\hat{\mathbf{R}})$ ma podobne własności jak w przypadku N_p.

Eksperymenty pokazały, że estymator uogólnionej wariancji, otrzymany za pomocą metodologii GSA daje bardzo dobre oszacowanie i można go stosować nawet wtedy, gdy próba losowa nie jest zbyt liczna.

Ze względu na swoje własności G-estymatory znajdują z powodzeniem zastosowanie w zagadnieniach praktycznych [7].

Tabela 1

Własności estymatorów $G(\hat{\mathbf{R}})$ i $\ln|\hat{\mathbf{R}}|$ dla p wymiarowego rozkładu normalnego

Wartość parametru $\ln\|\mathbf{R}\|$	Wymiar wektora p	Liczebność próby n	Średnia wartość estymatora		Błąd średniokwadratowy	
			$G(\hat{\mathbf{R}})$	$\ln\|\hat{\mathbf{R}}\|$	$BSK(G(\hat{\mathbf{R}}))$	$BSK(\ln\|\hat{\mathbf{R}}\|)$
1,609	2	5	1,476	0,678	1,594	2,445
		10	1,583	1,242	0,534	0,668
		20	1,605	1,445	0,227	0,253
		30	1,613	1,509	0,147	0,157
		40	1,606	1,529	0,110	0,116
		50	1,610	1,548	0,085	0,089
		70	1,608	1,564	0,059	0,061
		100	1,611	1,581	0,041	0,042
1,946	3	5	1,643	–0,255	3,342	8,093
		10	1,902	1,176	0,871	1,462
		20	1,937	1,609	0,356	0,469
		30	1,948	1,736	0,223	0,267
		40	1,944	1,787	0,162	0,188
		50	1,944	1,820	0,129	0,145
		70	1,948	1,860	0,090	0,097
		100	1,938	1,877	0,062	0,067
2,398	5	8	2,182	–0,697	2,810	12,339
		10	2,295	0,239	1,750	6,398
		20	2,392	1,531	0,628	1,380
		30	2,391	1,845	0,398	0,704
		40	2,385	1,984	0,271	0,442
		50	2,386	2,070	0,220	0,327
		70	2,402	2,180	0,152	0,199
		100	2,394	2,240	0,104	0,129

cd. tabeli 1

Wartość parametru $\ln \lvert\mathbf{R}\rvert$	Wymiar wektora p	Liczebność próby n	Średnia wartość estymatora		Błąd średniokwadratowy	
			$G(\hat{\mathbf{R}})$	$\ln \lvert\hat{\mathbf{R}}\rvert$	$BSK(G(\hat{\mathbf{R}}))$	$BSK(\ln \lvert\hat{\mathbf{R}}\rvert)$
3,045	10	12	2,621	–5,647	5,499	80,867
		20	3,013	–0,587	1,556	14,740
		30	3,020	0,858	0,856	5,635
		40	3,054	1,506	0,591	2,959
		50	3,039	1,831	0,466	1,938
		70	3,034	2,196	0,316	1,036
		100	3,045	2,460	0,216	0,546
3,434	15	18	3,165	–8,313	4,655	142,579
		20	3,289	–6,102	3,335	94,257
		30	3,393	–1,744	1,462	28,269
		40	3,419	–0,157	0,992	13,891
		50	3,418	0,669	0,733	8,375
		70	3,427	1,545	0,505	4,073
		100	3,437	2,158	0,336	1,965
3,714	20	23	3,418	–12,662	5,113	273,175
		30	3,658	–6,331	2,420	103,311
		40	3,704	–2,969	1,484	46,137
		50	3,708	–1,331	1,058	26,505
		70	3,704	0,310	0,699	12,282
		100	3,703	1,420	0,447	5,706
3,932	25	28	3,611	–17,154	5,690	450,197
		30	3,761	–14,134	4,262	330,611
		40	3,879	–7,250	2,070	127,103
		50	3,892	–4,316	1,449	69,476
		70	3,925	–1,496	0,918	30,380
		100	3,924	0,320	0,586	13,635

Źródło: obliczenia własne.

Tabela 2

Własności estymatorów $G(\hat{\mathbf{R}})$ i $\ln|\hat{\mathbf{R}}|$ dla p wymiarowego rozkładu Studenta

Liczba stopni swobody v	Wartość parametru $\ln\|\mathbf{R}\|$	Wymiar wektora p	Liczebność próby n	Średnia wartość estymatora		Błąd średniokwadratowy	
				$G(\hat{\mathbf{R}})$	$\ln\|\hat{\mathbf{R}}\|$	$BSK(G(\hat{\mathbf{R}}))$	$BSK(\ln\|\hat{\mathbf{R}}\|)$
6	0,811	2	5	0,376	–0,422	2,283	3,315
			10	0,585	0,244	0,904	1,175
			20	0,691	0,531	0,453	0,517
			30	0,718	0,613	0,320	0,350
			40	0,738	0,660	0,242	0,259
			50	0,758	0,697	0,190	0,201
			70	0,771	0,727	0,139	0,144
			100	0,772	0,742	0,102	0,105
6	2,027	5	8	0,955	–1,924	5,535	19,995
			10	1,201	–0,855	3,739	11,362
			20	1,540	0,679	1,713	3,294
			30	1,674	1,127	1,153	1,838
			40	1,743	1,343	0,873	1,261
			50	1,795	1,479	0,717	0,964
			70	1,848	1,625	0,519	0,649
			100	1,888	1,734	0,379	0,446
6	4,055	10	12	1,592	–6,676	15,235	124,320
			20	2,586	–1,013	6,237	29,766
			30	2,987	0,826	3,974	13,258
			40	3,170	1,621	2,973	8,111
			50	3,327	2,120	2,324	5,538
			70	3,491	2,652	1,661	3,309
			100	3,643	3,067	1,198	2,003
15	0,286	2	5	0,082	–0,717	1,822	2,787
			10	0,209	–0,132	0,653	0,822
			20	0,247	0,087	0,291	0,329
			30	0,259	0,154	0,190	0,206
			40	0,273	0,196	0,137	0,145
			50	0,270	0,209	0,113	0,119
			70	0,278	0,234	0,078	0,081
			100	0,280	0,250	0,055	0,056

cd. tabeli 2

Liczba stopni swobody v	Wartość parametru $\ln \lvert\mathbf{R}\rvert$	Wymiar wektora p	Liczebność próby n	Średnia wartość estymatora		Błąd średniokwadratowy	
				$G(\hat{\mathbf{R}})$	$\ln \lvert\hat{\mathbf{R}}\rvert$	$BSK(G(\hat{\mathbf{R}}))$	$BSK(\ln \lvert\hat{\mathbf{R}}\rvert)$
15	0,716	5	8	0,270	–2,608	3,543	14,392
			10	0,400	–1,655	2,271	7,793
			20	0,569	–0,293	0,930	1,925
			30	0,611	0,064	0,567	0,980
			40	0,638	0,238	0,426	0,648
			50	0,664	0,348	0,341	0,473
			70	0,676	0,454	0,237	0,303
			100	0,689	0,535	0,164	0,196
15	1,431	10	12	0,365	–7,903	7,838	93,831
			20	0,965	–2,634	2,438	16,818
			30	1,118	–1,044	1,567	7,592
			40	1,180	–0,368	1,125	4,300
			50	1,221	0,014	0,861	2,825
			70	1,275	0,436	0,627	1,592
			100	1,320	0,745	0,419	0,877
30	0,138	2	5	–0,022	–0,820	1,718	2,611
			10	0,088	–0,253	0,583	0,733
			20	0,120	–0,039	0,253	0,284
			30	0,122	0,018	0,165	0,180
			40	0,128	0,051	0,122	0,130
			50	0,130	0,069	0,095	0,100
			70	0,132	0,088	0,067	0,070
			100	0,133	0,103	0,047	0,048
30	0,345	5	8	–0,008	–2,886	3,133	13,448
			10	0,146	–1,909	1,985	7,028
			20	0,271	–0,590	0,745	1,613
			30	0,313	–0,234	0,475	0,809
			40	0,310	–0,090	0,350	0,538
			50	0,333	0,017	0,263	0,371
			70	0,321	0,099	0,186	0,247
			100	0,335	0,181	0,133	0,160

cd. tabeli 2

Liczba stopni swobody v	Wartość parametru $\ln \lvert\mathbf{R}\rvert$	Wymiar wektora p	Liczebność próby n	Średnia wartość estymatora		Błąd średniokwadratowy	
				$G(\hat{\mathbf{R}})$	$\ln \lvert\hat{\mathbf{R}}\rvert$	$BSK(G(\hat{\mathbf{R}}))$	$BSK(\ln \lvert\hat{\mathbf{R}}\rvert)$
30	0,690	10	12	0,010	–8,258	6,474	86,086
			20	0,438	–3,161	1,993	16,762
			30	0,531	–1,630	1,145	6,502
			40	0,567	–0,981	0,822	3,600
			50	0,599	–0,609	0,628	2,306
			70	0,627	–0,211	0,442	1,251
			100	0,641	0,066	0,304	0,691

Źródło: obliczenia własne.

Tabela 3

Własności estymatorów $G(\hat{\mathbf{R}})$ i $\ln \lvert\hat{\mathbf{R}}\rvert$ dla p wymiarowych rozkładów Pearsona

Wartość parametru β_1	Wartość parametru $\ln \lvert\mathbf{R}\rvert$	Wymiar wektora p	Liczebność próby n	Średnia wartość estymatora		Błąd średniokwadratowy	
				$G(\hat{\mathbf{R}})$	$\ln \lvert\hat{\mathbf{R}}\rvert$	$BSK(G(\hat{\mathbf{R}}))$	$BSK(\ln \lvert\hat{\mathbf{R}}\rvert)$
0,2	0	2	5	–0,104	–0,902	1,582	2,385
			10	–0,023	–0,364	0,530	0,662
			20	–0,006	–0,165	0,232	0,259
			30	–0,007	–0,112	0,148	0,160
			40	–0,001	–0,078	0,108	0,114
			50	0,001	–0,060	0,085	0,088
			70	–0,002	–0,046	0,059	0,061
			100	–0,005	–0,035	0,041	0,042
0,2	0	3	5	–0,290	–2,188	3,378	8,079
			10	–0,039	–0,764	0,874	1,456
			20	–0,003	–0,324	0,352	0,457
			30	–0,004	–0,216	0,219	0,266
			40	–0,003	–0,159	0,159	0,184
			50	0,000	–0,124	0,128	0,143
			70	0,001	–0,086	0,092	0,099
			100	–0,001	–0,062	0,062	0,066

cd. tabeli 3

Wartość parametru β_1	Wartość parametru $\ln \lvert\mathbf{R}\rvert$	Wymiar wektora p	Liczebność próby n	Średnia wartość estymatora		Błąd średniokwadratowy	
				$G(\hat{\mathbf{R}})$	$\ln \lvert\hat{\mathbf{R}}\rvert$	$BSK(G(\hat{\mathbf{R}}))$	$BSK(\ln \lvert\hat{\mathbf{R}}\rvert)$
0,2	0	5	10	–0,080	–2,136	1,735	6,290
			20	–0,006	–0,868	0,637	1,390
			30	–0,002	–0,549	0,389	0,690
			40	–0,008	–0,408	0,275	0,441
			50	–0,001	–0,317	0,215	0,316
			70	–0,002	–0,224	0,150	0,200
			100	–0,001	–0,155	0,104	0,128
0,4	0	2	5	–0,120	–0,918	1,629	2,457
			10	–0,028	–0,369	0,543	0,679
			20	–0,010	–0,169	0,232	0,261
			30	–0,002	–0,106	0,147	0,156
			40	–0,005	–0,083	0,107	0,114
			50	0,003	–0,059	0,085	0,089
			70	–0,002	–0,046	0,060	0,062
			100	–0,004	–0,035	0,040	0,042
0,4	0	3	5	–0,328	–2,226	3,409	8,255
			10	–0,052	–0,778	0,877	1,481
			20	–0,011	–0,339	0,358	0,472
			30	–0,002	–0,214	0,226	0,272
			40	0,000	–0,157	0,163	0,187
			50	0,002	–0,122	0,128	0,143
			70	0,006	–0,082	0,089	0,096
			100	0,000	–0,061	0,062	0,066
0,4	0	5	10	–0,101	–2,156	1,763	6,402
			20	–0,023	–0,884	0,648	1,429
			30	–0,003	–0,550	0,395	0,697
			40	0,003	–0,397	0,280	0,438
			50	–0,009	–0,325	0,219	0,325
			70	–0,002	–0,224	0,155	0,205
			100	–0,007	–0,161	0,104	0,130

cd. tabeli 3

Wartość parametru β_1	Wartość parametru $\ln \lvert \mathbf{R} \rvert$	Wymiar wektora p	Liczebność próby n	Średnia wartość estymatora		Błąd średniokwadratowy	
				$G(\hat{\mathbf{R}})$	$\ln \lvert \hat{\mathbf{R}} \rvert$	$BSK(G(\hat{\mathbf{R}}))$	$BSK(\ln \lvert \hat{\mathbf{R}} \rvert)$
1	0	2	5	–0,315	–1,113	2,514	3,654
			10	–0,067	–0,408	0,730	0,892
			20	–0,011	–0,170	0,268	0,297
			30	0,003	–0,101	0,163	0,173
			40	0,009	–0,069	0,114	0,118
			50	0,007	–0,055	0,091	0,094
			70	0,012	–0,032	0,063	0,064
			100	0,011	–0,020	0,043	0,043
1	0	3	5	–0,601	–2,498	4,901	10,781
			10	–0,108	–0,834	1,163	1,846
			20	0,001	–0,327	0,407	0,514
			30	0,003	–0,209	0,245	0,289
			40	0,006	–0,151	0,178	0,200
			50	0,012	–0,112	0,134	0,146
			70	0,016	–0,072	0,095	0,099
			100	0,012	–0,049	0,064	0,067
1	0	5	10	–0,198	–2,254	2,273	7,312
			20	–0,011	–0,872	0,729	1,490
			30	–0,002	–0,549	0,434	0,736
			40	0,016	–0,384	0,304	0,451
			50	0,016	–0,300	0,234	0,324
			70	0,023	–0,200	0,163	0,203
			100	0,021	–0,133	0,107	0,124

Źródło: obliczenia własne.

Tabela 4

Własności estymatorów $G(\hat{\mathbf{R}})$ i $\ln|\hat{\mathbf{R}}|$ dla p wymiarowego rozkładu wykładniczego z parametrem $\lambda = 1$

Wartość parametru $\ln\|\mathbf{R}\|$	Wymiar wektora p	Liczebność próby n	Średnia wartość estymatora		Błąd średniokwadratowy	
			$G(\hat{\mathbf{R}})$	$\ln\|\hat{\mathbf{R}}\|$	$BSK(G(\hat{\mathbf{R}}))$	$BSK(\ln\|\hat{\mathbf{R}}\|)$
0	2	5	–0,795	–1,594	3,618	5,525
		10	–0,437	–0,778	1,604	2,018
		20	–0,241	–0,400	0,761	0,863
		30	–0,157	–0,261	0,497	0,540
		40	–0,123	–0,200	0,381	0,406
		50	–0,105	–0,166	0,325	0,341
		70	–0,075	–0,119	0,221	0,230
		100	–0,052	–0,083	0,153	0,157
0	3	5	–1,335	–3,232	7,213	15,878
		10	–0,648	–1,374	2,559	4,026
		20	–0,354	–0,682	1,198	1,538
		30	–0,263	–0,475	0,794	0,951
		40	–0,197	–0,353	0,592	0,678
		50	–0,169	–0,293	0,471	0,529
		70	–0,114	–0,202	0,342	0,370
		100	–0,085	–0,146	0,236	0,250
0	5	8	–1,488	–4,366	7,475	24,327
		10	–1,137	–3,192	5,208	14,106
		20	–0,591	–1,452	2,193	3,952
		30	–0,423	–0,970	1,434	2,195
		40	–0,334	–0,734	1,049	1,477
		50	–0,261	–0,577	0,832	1,097
		70	–0,195	–0,417	0,580	0,716
		100	–0,140	–0,294	0,414	0,481

Źródło: obliczenia własne

Tabela 5

Własności estymatorów $G(\hat{\mathbf{R}})$ i $\ln|\hat{\mathbf{R}}|$ dla p wymiarowego rozkładu Laplace'a z parametrem $\lambda = 1$

Wartość parametru $\ln\|\mathbf{R}\|$	Wymiar wektora p	Liczebność próby n	Średnia wartość estymatora		Błąd średniokwadratowy	
			$G(\hat{\mathbf{R}})$	$\ln\|\hat{\mathbf{R}}\|$	$BSK(G(\hat{\mathbf{R}}))$	$BSK(\ln\|\hat{\mathbf{R}}\|)$
1,386	2	5	0,845	0,046	2,750	4,252
		10	1,127	0,786	1,082	1,375
		20	1,250	1,091	0,512	0,580
		30	1,300	1,204	0,339	0,368
		40	1,321	1,244	0,245	0,261
		50	1,328	1,266	0,194	0,205
		70	1,351	1,307	0,142	0,147
		100	1,355	1,325	0,098	0,101
2,079	3	5	1,160	–0,737	5,399	12,485
		10	1,701	0,975	1,751	2,827
		20	1,887	1,559	0,760	0,994
		30	1,944	1,732	0,518	0,620
		40	1,982	1,826	0,378	0,433
		50	1,996	1,872	0,304	0,340
		70	2,011	1,923	0,218	0,238
		100	2,039	1,978	0,152	0,160
3,466	5	8	2,574	–0,304	5,095	18,513
		10	2,784	0,728	3,460	10,488
		20	3,141	2,280	1,399	2,699
		30	3,215	2,669	0,924	1,496
		40	3,279	2,879	0,657	0,966
		50	3,317	3,001	0,523	0,717
		70	3,362	3,140	0,364	0,460
		100	3,396	3,242	0,251	0,296

Źródło: obliczenia własne.

Tabela 6

Własności estymatorów $G(\hat{\mathbf{R}})$ i $\ln|\hat{\mathbf{R}}|$ dla p wymiarowego rozkładu Weibulla

Wartość parametru λ	Wartość parametru $\ln\|\mathbf{R}\|$	Wymiar wektora p	Liczebność próby n	Średnia wartość estymatora		Błąd średniokwadratowy	
				$G(\hat{\mathbf{R}})$	$\ln\|\hat{\mathbf{R}}\|$	$BSK(G(\hat{\mathbf{R}}))$	$BSK(\ln\|\hat{\mathbf{R}}\|)$
0,5	5,991	2	5	2,432	1,633	22,208	28,530
			10	3,697	3,356	10,587	12,267
			20	4,530	4,371	5,167	5,658
			30	4,900	4,796	3,384	3,622
			40	5,091	5,013	2,633	2,778
			50	5,219	5,158	2,221	2,319
			70	5,376	5,332	1,560	1,615
			100	5,529	5,499	1,122	1,151
0,5	8,987	3	5	3,381	1,483	46,849	71,721
			10	5,517	4,791	19,962	25,526
			20	6,795	6,469	9,400	10,945
			30	7,287	7,075	6,273	7,038
			40	7,605	7,448	4,639	5,097
			50	7,792	7,668	3,694	4,007
			70	8,063	7,975	2,665	2,835
			100	8,274	8,213	1,890	1,981
0,5	14,979	5	8	7,898	5,020	66,758	115,805
			10	9,021	6,965	49,150	77,869
			20	11,298	10,437	21,323	28,404
			30	12,165	11,618	13,719	17,093
			40	12,660	12,260	9,951	11,968
			50	13,020	12,704	7,767	9,105
			70	13,434	13,211	5,366	6,102
			100	13,790	13,636	3,745	4,135
2	–3,078	2	5	–3,221	–4,019	1,603	2,469
			10	–3,118	03,459	0,572	0,715
			20	–3,092	–3,251	0,246	0,276
			30	–3,078	–3,182	0,157	0,168
			40	–3,079	–3,157	0,116	0,122
			50	–3,082	–3,144	0,098	0,103
			70	–3,078	–3,122	0,067	0,068
			100	–3,079	–3110	0,045	0,046

cd. tabeli 6

Wartość parametru λ	Wartość parametru $\ln\|\mathbf{R}\|$	Wymiar wektora p	Liczebność próby n	Średnia wartość estymatora		Błąd średniokwadratowy	
				$G(\hat{\mathbf{R}})$	$\ln\|\hat{\mathbf{R}}\|$	$BSK(G(\hat{\mathbf{R}}))$	$BSK(\ln\|\hat{\mathbf{R}}\|)$
2	–4,617	3	5	–4.945	–6,842	3,364	8,207
			10	–4.658	–5,384	0,904	1,491
			20	–4,631	–4,959	0,380	0,497
			30	–4,637	–4,849	0,245	0,298
			40	–4,626	–4,783	0,180	0,208
			50	–4,628	–4,753	0,142	0,160
			70	–4,618	–4,706	0,101	0,109
			100	–4,622	–4,683	0,068	0,072
2	–7,695	5	8	–7,947	–10,825	2,888	12,624
			10	–7,809	–9,865	1,857	6,552
			20	–7,714	–8,575	0,684	1,458
			30	–7,713	–8,259	0,423	0,742
			40	–7,712	–8,112	0,311	0,485
			50	–7,702	–8,018	0,242	0,347
			70	–7,704	–7,926	0,170	0,223
			100	–7,702	–7,855	0,119	0,145

Źródło: obliczenia własne.

Tabela 7

Własności estymatorów $G(\hat{\mathbf{R}})$ i $\ln|\hat{\mathbf{R}}|$ dla p wymiarowego rozkładu Beta

Wartości parametrów (α, β)	Wartość parametru $\ln\|\mathbf{R}\|$	Wymiar wektora p	Liczebność próby n	Średnia wartość estymatora		Błąd średniokwadratowy	
				$G(\hat{\mathbf{R}})$	$\ln\|\hat{\mathbf{R}}\|$	$BSK(G(\hat{\mathbf{R}}))$	$BSK(\ln\|\hat{\mathbf{R}}\|)$
(0.5, 0.5)	–4,159	2	5	–4,063	–4,862	1,217	1,702
			10	–4,040	–4,381	0,239	0,274
			20	-4,090	–4,250	0,078	0,081
			30	–4,108	–4,212	0,046	0,046
			40	–4,123	–4,196	0,031	0,030
			50	–4,125	–4,187	0,024	0,024
			70	–4,135	–4,178	0,017	0,016
			100	–4,143	–4,173	0,011	0,011

cd. tabeli 7

Wartości parametrów (α, β)	Wartość parametru $\ln \lvert \mathbf{R} \rvert$	Wymiar wektora p	Liczebność próby n	Średnia wartość estymatora		Błąd średniokwadratowy	
				$G(\hat{\mathbf{R}})$	$\ln \lvert \hat{\mathbf{R}} \rvert$	$BSK(G(\hat{\mathbf{R}}))$	$BSK(\ln \lvert \hat{\mathbf{R}} \rvert)$
(0.5, 0.5)	–6,238	3	5	–6,221	–8,118	2,704	6,238
			10	–6,057	–6,783	0,441	0,705
			20	–6,125	–6,453	0,134	0,167
			30	–6,160	–6,372	0,074	0,086
			40	–6,183	–6,340	0,051	0,059
			50	–6,195	–6,320	0,039	0,044
			70	–6,202	–6,290	0,026	0,027
			100	–6,214	–6,275	0,017	0,018
(0.5, 0.5)	–10,397	5	8	–10,178	–13,056	1,977	8,997
			10	–10,123	–12,178	1,064	4,161
			20	–10,221	–11,082	0,273	0,711
			30	–10,273	–10,820	0,145	0,308
			40	–10,301	–10,701	0,096	0,179
			50	–10,316	–10,631	0,070	0,119
			70	–10,339	–10,562	0,046	0,069
			100	–10,354	–10,508	0,031	0,041
(2, 0.5)	–6,171	2	5	–6,698	–7,496	3,216	4,696
			10	–6,351	–6,692	1,006	1,245
			20	–6,236	–6,395	0,376	0,422
			30	–6,217	–6,321	0,233	0,254
			40	–6,190	–6,267	0,163	0,172
			50	–6,193	–6,255	0,128	0,135
			70	–6,179	–6,222	0,087	0,090
			100	–6,176	–6,206	0,060	0,061
(2, 0.5)	–9,256	3	5	–10,204	–12,101	6,350	13,545
			10	–9,503	–10,229	1,543	2,429
			20	–9,348	–9,676	0,579	0,747
			30	–9,308	–9,520	0,344	0,411
			40	–9,281	–9,438	0,247	0,280
			50	–9,286	–9,410	0,190	0,213
			70	–9,271	–9,359	0,130	0,140
			100	–9,267	–9,328	0,087	0,092

cd. tabeli 7

Wartości parametrów (α, β)	Wartość parametru $\ln \lvert \mathbf{R} \rvert$	Wymiar wektora p	Liczebność próby n	Średnia wartość estymatora		Błąd średniokwadratowy	
				$G(\hat{\mathbf{R}})$	$\ln \lvert \hat{\mathbf{R}} \rvert$	$BSK(G(\hat{\mathbf{R}}))$	$BSK(\ln \lvert \hat{\mathbf{R}} \rvert)$
(2, 0.5)	–15,427	5	8	–16,156	–19,034	5,067	17,549
			10	–15,891	–17,947	2,989	9,124
			20	–15,613	–16,474	1,049	2,111
			30	–15,523	–16,070	0,587	0,991
			40	–15,493	–15,894	0,435	0,648
			50	–15,480	–15,796	0,324	0,457
			70	–15,454	–15,676	0,218	0,279
			100	–15,444	–15,598	0,150	0,180
(1, 2)	–5.781	2	5	–5,818	–6,617	1,494	2,191
			10	–5,749	–6,090	0,456	0,551
			20	–5,757	–5,917	0,177	0,195
			30	–5,760	–5,864	0,107	0,113
			40	–5,773	–5,850	0,081	0,085
			50	–5,768	–5,829	0,063	0,065
			70	–5,769	–5,812	0,043	0,044
			100	–5,776	–5,806	0,030	0,031
(1, 2)	–8,671	3	5	–8,833	–10,730	3,189	7,402
			10	–8,642	–9,368	0,732	1,216
			20	–8,635	–8,963	0,282	0,366
			30	–8,645	–8,857	0,168	0,201
			40	–8,647	–8,804	0,119	0,135
			50	–8,650	–8,775	0,094	0,105
			70	–8,660	–8,748	0,065	0,070
			100	–8,663	–8,724	0,044	0,047
(1, 2)	–14,452	5	8	–14,525	–17,403	2,589	11,294
			10	–14,419	–16,475	1,551	5,642
			20	–14,398	–15,259	0,501	1,149
			30	–14,410	–14,957	0,290	0,543
			40	–14,424	–14,825	0,206	0,344
			50	–14,425	–14,742	0,157	0,240
			70	–14,434	–14,656	0,112	0,153
			100	–14,435	–14,588	0,075	0,093

Źródło: obliczenia własne.

Uniwersytet Łódzki, Katedra Metod Statystycznych

LITERATURA

[1] Anderson T.W., [1958], *An introduction to multivariate statistical analysis*, John Wiley and Sons, NY, London.
[2] Barra J.R., [1982], *Matematyczne podstawy statystyki*, PWN, Warszawa.
[3] Girko V.L., [1975], *Random Matrices* (in Russian), Kiev University, Kiev.
[4] Girko V.L., [1988], *Spectral Theory of Random Matrices* (in Russian), Nauka, Moscow.
[5] Girko V.L., [1995], *Statistical Analysis of Observations of Increasing Dimension*, Kluwer Academic Publishers.
[6] Girko V.L., [1990], *Theory of Random Determinants*, Kluwer Academic Publisher.
[7] Johnson B.A., Abramovich Y.I., Mestre X., [Aug. 2008], *MUSIC, G-MUSIC,and Maximum- Likelihood Performance Breakdown*, IEEE Transactions on Signal Processing, Vol. 56, No. 2.
[8] Krzyśko M., [2000], *Wielowymiarowa analiza statystyczna*, Rektor UAM, Poznań.
[9] Rao C.R., [1982], *Modele liniowe statystyki matematycznej*, PWN, Warszawa.

Praca wpłynęła do redakcji w maju 2009 r.

ESTYMACJA UOGÓLNIONEJ WARIANCJI WYBRANYCH ROZKŁADÓW WIELOWYMIAROWYCH

Streszczenie

Uogólniona wariancja, czyli wyznacznik macierzy kowariancji jest skalarną miarą rozrzutu rozkładów wielowymiarowych. Dokładny rozkład uogólnionej wariancji znany jest tylko dla wektorów losowych o wielowymiarowym rozkładzie normalnym. Dla wektorów losowych o dużych wymiarach przyjmuje on skomplikowaną postać, co stanowi utrudnienie w zastosowaniach praktycznych.

W pracy przedstawiono tzw. *G*-estymator logarytmu uogólnionej wariancji otrzymany na podstawie twierdzeń granicznych dla wyznaczników losowych i jego własności na przykładach symulacyjnych dla kilku wybranych rozkładów wielowymiarowych.

Słowa kluczowe: uogólniona wariancja, wektor losowy, macierz kowariancji, wyznacznik losowy.

ESTIMATION OF GENERALIZED VARIANCE OF CHOSEN MULTIVARIATE DISTRIBUTION

Summary

Generalized variance i.e. the determinant of the covariance matrix is a scalar measure of multivariate distribution dispersion. The exact distribution of the generalized variance is known only for multivariate normal vectors. For random vectors in high dimensional spaces it has a complicated formula very troublesome to apply.

An estimator of the logarithm of generalized variance derived with the help of limit theorems for random determinants was presented as well as its properties in examples of chosen simulation multivariate distributions.

Key words: generalized variance, random vector, covariance matrix, random determinant.

SPIS TREŚCI

CONTENTS