

ISSN 1898-6447

Uniwersytet
Ekonomiczny
w Krakowie

Zeszyty Naukowe

Cracow Review
of Economics
and Management

Nr 5 (965)

Kraków 2017



Uniwersytet
Ekonomiczny
w Krakowie

Zeszyty Naukowe

Cracow Review
of Economics
and Management

Nr 5(965)

Kraków 2017

Rada Naukowa

Andrzej Antoszewski (Polska), Slavko Arsovski (Serbia), Josef Arlt (Czechy), Daniel Baier (Niemcy), Hans-Hermann Bock (Niemcy), Ryszard Borowiecki (Polska), Giovanni Lagioia (Włochy), Tadeusz Markowski (Polska), Martin Mizla (Słowacja), David Ost (USA), Józef Pocięcha (Polska)

Komitet Redakcyjny

Joanna Dzwonczyk, Ryszard Kowalski (sekretarz), Barbara Pawełek, Aleksy Pocztowski (redaktor naczelny), Krystyna Przybylska, Tadeusz Sikora, Grzegorz Strupczewski (sekretarz), Wanda Sułkowska, Angelika Wodecka-Hyjek (sekretarz), Bernard Ziębicki

Redaktor statystyczny

Paweł Ulman

Redaktorzy Wydawnictwa

Agnieszka Penarska, Monika Rusin, Seth Stevens (teksty w j. angielskim)

Projekt okładki i układ graficzny tekstu

Marcin Sokołowski

Zdjęcie Profesora J. Steczkowskiego

Krzysztof Włodek

Czasopismo jest indeksowane w następujących bazach:

BazEkon (<https://bazybg.uek.krakow.pl/bazekon>), CEEOL (www.ceeol.com),

CEJSH (<http://cejsh.icm.edu.pl>) oraz ICI World of Journals (<https://journals.indexcopernicus.com>)



Ministerstwo Nauki
i Szkolnictwa Wyższego

Recenzowanie artykułów w języku angielskim oraz redakcja językowa artykułów w języku angielskim – zadania finansowane w ramach umowy 799/P-DUN/2017 ze środków Ministra Nauki i Szkolnictwa Wyższego przeznaczonych na działalność upowszechniającą naukę

© Copyright by Uniwersytet Ekonomiczny w Krakowie, Kraków 2017

ISSN 1898-6447

Wersja pierwotna: publikacja drukowana

Teksty artykułów są dostępne na stronie internetowej czasopisma:

www.zeszyty-naukowe.uek.krakow.pl oraz w bazie CEEOL

Wydawnictwo Uniwersytetu Ekonomicznego w Krakowie

31-510 Kraków, ul. Rakowicka 27, tel. 12 293 57 42, e-mail: wydaw@uek.krakow.pl

www.zeszyty-naukowe.uek.krakow.pl

Zakład Poligraficzny Uniwersytetu Ekonomicznego w Krakowie, 31-510 Kraków, ul. Rakowicka 27

Zam. 278/2017

Spis treści

Józef Pocięcha	
Prof. dr hab. inż. Jan Steczkowski (1923–2016) – życie i twórczość	5
Janusz L. Wywiłł	
On the Evaluation of Sample Size Required for a Good Approximation by the Normal Curve for Some Statistics	17
Jacek Osiewalski, Jerzy Marzec	
Dwuwymiarowe zmienne licznikowe – bayesowskie modelowanie selekcji próby	31
Andrzej Czyżewski, Aleksander Grzelak	
Wykorzystanie modelu przepływów międzygałęziowych do ocen makroekonomicznych gospodarki.....	51
Tadeusz Grabiński	
Uproszczona metoda delimitacji wektorowej	69
Michał Major	
Karty kontrolne sum skumulowanych z możliwością akceptacji procesu. Przegląd podstawowych narzędzi	87
Jan Tatar	
Zbieżność stochastyczna ciągów wektorów losowych	107



Jan Steczkowski

| Józef Pocięcha

| Prof. dr hab. inż. Jan Steczkowski (1923–2016) – życie i twórczość

Dzieciństwo i okres wojenny

Jan Kanty Steczkowski urodził się w Krakowie w dniu 11 października 1923 r. w znanej i szanowanej rodzinie. Dzieciństwo miał szczęśliwe, aczkolwiek, jak sam wspominał, był dzieckiem niesfornym i sprawiał kłopoty. Był niespełna szesnastoletnim chłopcem, gdy wraz z wybuchem wojny skończyło się jego szczęśliwe dzieciństwo, a rozpoczęły się trudne lata niemieckiej okupacji. Aby nie narażać się na aresztowanie przez okupanta, podjął naukę w legalnie działającej, znanej przed wojną, Szkole Rolniczej w Czernichowie. W szkole tej zdał konspiracyjną małą maturę.

W 1942 r. rozpoczął działalność konspiracyjną w szeregach Armii Krajowej [Między liczbą a treścią... 2000, s. 5–6]. Najpierw działał w placówce szkolnej o kryptonimie „Pardwa”, podporządkowanej IV batalionowi VI Dywizji Armii Krajowej Kraków. Z czasem placówka ta przekształciła się w konspiracyjną Szkołę Podchorążych AK. Młody Jan Steczkowski ukończył ją, otrzymując stopień wojskowy kaprała podchorążego. W związku z przygotowaniem planów operacji „Burza” po zaopatrzeniu w fałszywe dokumenty został przerzucony do Warszawy, a potem na Podlasie w okolice Radzyna, gdzie walczył w oddziale partyzanckim należącym do 35 pułku piechoty AK pod dowództwem majora „Muellera”, komendanta radzyńskiego obwodu AK. W oddziale tym jako młody, silny chłopak nosił CKM. Po wkroczeniu na te tereny Armii Czerwonej w lipcu 1944 r. jego oddział został rozproszony, a on sam został złapany, aresztowany

i osadzony w obozie przejściowym na Majdanku. Uniknął wywózki do sowieckich łagrów, wykorzystując okazję do ucieczki. Drugą połowę 1944 r. Profesor Jan Steczkowski wspominał jako jeden z najtrudniejszych okresów w swoim życiu; tułaczka po lasach, głód i niepewność jutra żołnierza AK bez przydziału były przygnębiające, jak sam niechętnie o tym mówił. Linia frontu oddzielająca Lubelszczyznę od Krakowa odcięła mu drogę powrotu do domu. W styczniu 1945 r. po przesunięciu linii frontu na zachód dotarł do domu rodzinnego w Krakowie. Tutaj spotkała go kolejna przykra niespodzianka. Jego dom rodzinny, jako jeden z nielicznych w Krakowie, został w czasie ofensywy zimowej trafiony bombą lotniczą; jego mieszkanie było zniszczone. Snuł plany podjęcia studiów – myślał o chemii na nowo tworzonej Politechnice Krakowskiej, lecz jego losy potoczyły się inaczej. Za przynależność do Armii Krajowej w sierpniu 1945 r. został aresztowany i skazany na dwa lata więzienia. W październiku 1946 r. został zwolniony na mocy amnestii.

Okres stalinowski

Po wyjściu z więzienia zaczął szukać swojego miejsca. Najpierw uzupełnił braki w średnim wykształceniu, a następnie w związku z zakazem wstępu na uczelnie państwowe¹ podjął studia na kierunku ogólnohandlowym w ówczesnej prywatnej Akademii Handlowej w Krakowie. Ukończył je w 1950 r., ale bez tytułu magistra, gdyż ideologiczny recenzent jego pracy magisterskiej odrzucił ją jako niemarksistowską [*Między liczbą a treścią...* 2000, s. 6].

W 1950 r. prywatna Akademia Handlowa została upaństwowiona i przekształcona w Wyższą Szkołę Ekonomiczną. Jej finansowanie jako uczelni państwowej zostało przejęte przez budżet państwa, co w ówczesnym systemie nakazowo-rozdzielczym wiązało się z przedstawianiem szczegółowych planów wydatków. W tym okresie „totalnego zamieszania” pojawił się na uczelni Jan Steczkowski we własnych studenckich sprawach.

„Spotkałem się z dr. Julianem Popielem, asystentem prof. Feliksa Młynarskiego i razem sekretarzem Szkoły, tak nazywał się bowiem dyrektor Uczelni, który popatrzył na mnie i – zapewne w ostatecznej rozterce – zapytał, czy nie podjąłbym się opracowania budżetu, zgodnego z wymaganiami resortu. Nie wiem, co mną wówczas powodowało – z pewnością chroniczny brak gotówki – bo wyraziłem zgodę. Otrzymałem z Uczelni stosowne upoważnienie, kupiłem ryżę papieru i pojechałem do Warszawy. W Ministerstwie namówiłem pewnego urzędnika, by mi udostępnił – jego zdaniem – najlepiej opracowany budżet uczelni ekonomicznej i zasiadłem do pracy. Wówczas wszystko musiało być zaplanowane: nawet papier toaletowy, ścierki i benzyna według z góry określonych norm zużycia. Przepisałem wszystkie pozycje użyczonego mi wzoru, co trwało dwa dni. Gdy wróciłem do Krakowa (...), kolejno zapraszałem przedstawicieli poszczególnych działów

¹ Dzieci tzw. bz-tów (byłych ziemian) oraz kułaków nie miały wstępu na uczelnie państwowe.

i zadawałem pytania w rodzaju: iloma samochodami dysponuje administracja, ile wynosi roczne zużycie benzyny, przeliczając według odpisanej przeze mnie normy itd. Tak pracowaliśmy trzy dni, zarywając noce, abym mógł opasły tom, zawierający szczegółowo rozpisany budżet dostarczyć do Ministerstwa... Najważniejsze jednak było to, że budżet został przyjęty. Ubocznym produktem przedstawionej działalności okazała się złożona mi oferta pracy kierowania Oddziałem Planowania i Inwestycji [*Rodowód Akademii Ekonomicznej w Krakowie...* 2006, s. 6].

Zarówno w dotychczasowej Akademii Handlowej, jak i powstałej z jej upaństwowienia Wyższej Szkole Ekonomicznej brakowało pomieszczeń, w których można byłoby prowadzić działalność dydaktyczną. Dotychczas wykorzystywane budynki przy ul. Sienkiewicza 4 i 5 były zdecydowanie za małe. Po perturbacjach dotyczących przyszłości uczelni ekonomicznej, związanych z planami jej ewentualnej likwidacji, decyzją Ministra Szkolnictwa Wyższego, z dniem 1 sierpnia 1952 r. WSE przejęła nieruchomość przy ul. Rakowickiej 27 wraz z terenem 81 829 m² [*Akademia Ekonomiczna w Krakowie...* 2002, s. 110]. Przeprowadzką kierował powołany na p.o. dyrektora administracyjnego Jan Steczkowski.

Na gospodarstwie zostałem sam, jako p.o. dyrektor administracyjny, dosłownie z kilkoma osobami. Inż. Władysława Skrzypka, mojego zastępcę, poprosiłem o to, aby każdy papierek, rachunek czy umowę poddawał wnikliwej kontroli, a ja będę pilnował realizacji robót i uzyskiwał zezwolenia na roboty i materiały. Sprawa była trudna, ponieważ obowiązywała gospodarka planowa, w której uzyskanie niezaplanowanych wcześniej materiałów i sił fachowych do robót w zasadzie można było załatwić na czarnym rynku albo dając łapówkę, a ja nie dość, że dysponowałem skromnym budżetem – na moje szczęście – nie umiałem tego robić... Architektonicznie cenny główny budynek był mocno zdewastowany, woda lała się po ścianach. Trzeba było wszystkie pomieszczenia ogrzewane piecami odmalować, sprawdzić liczne piece kaflowe i zaadaptować część piwnic, tak aby udało się w nich umieścić księgozbiór oraz zbudować dla niego odpowiednie półki i windę. Można by wymieniść całą litanię prac do wykonania. I udało się – nowy rok akademicki (1952/53) rozpoczął się bez szczególnych zakłóceń [*Rodowód Akademii Ekonomicznej w Krakowie...* 2006, s. 9–10].

Jak wiadomo, nad częścią środkową budynku głównego Uniwersytetu wznosi się piękna kopuła z krzyżem. Profesor Jan Steczkowski był tą osobą, która w okresie stalinowskim uratowała krzyż i kopułę przed rozbiórką. W swoich wspomnieniach Profesor tak to opisuje:

Potem chciano zlikwidować krzyż na kopule głównego gmachu. Komitet Uczelniany zażądał ode mnie nawet, abym usunął całą kopułę, a gdy odmówiłem, nazwano mnie papistą. Mimo pilnych robót musiałem się zająć sprawą. Udałem się do konserwatora miasta Krakowa dr. Józefa Lepiarczyka i poprosiłem z głupia frant o zezwolenie na zdjęcie kopuły. Ten dosłownie spąsował i krzyknął: „Chcecie likwidować jedną z najładniejszych kopuł w Krakowie? Cały Budapeszt ma wiele takich kopuł i nic im to nie przeszkadza!”. Poprosiłem, aby brak zgody przekazał mi na piśmie. Zrobił to bez wahania. Następnie

będąc w Warszawie poszedłem do Biura ds. Wyznań i ponowiłem prośbę. Odpowiedzieli: „Chyba macie w Krakowie za dużo zabytków. Popatrzcie, my na Nowym Świecie kopuły i herby restaurujemy, a wy chcecie je niszczyć?”. I tym razem poprosiłem o negatywną odpowiedź na piśmie. Następnie udałem się do Ministerstwa i ponowiłem prośbę. Zapytali, o ile to zwiększy kubaturę budynku i ile to będzie kosztowało. Gdy podałem, że koszty będą duże – odmówili finansowania takiej renowacji. Również i tu werdykt na piśmie dostałem. Po powrocie do Krakowa zjawiłem się w KU PZPR i stwierdziłem: „Popatrzcie, papistą jest konserwator miasta, zagnieździł się też w resorcie, a co gorsza i w Biurze ds. Wyznań”. Zdębieli, na pewien czas odczepili się ode mnie i mogłem zając się sensowną pracą [Rodowód Akademii Ekonomicznej w Krakowie... 2006, s. 10].

Pracując zawodowo, Jan Steczkowski podjął w 1953 r. studia na Wydziale Rolnym Wyższej Szkoły Rolniczej w Krakowie. 22 listopada 1957 r. uzyskał dyplom inżyniera rolnictwa. Równolegle w latach 1955–1957 ukończył w WSE studia II stopnia, otrzymując tytuł magistra ekonomii na podstawie pracy magisterskiej pt. *Założenia projektowe rozbudowy Wyższej Szkoły Ekonomicznej w Krakowie*, liczącej ok. 200 stron maszynopisu, obronionej 1 marca 1957 r. [Między liczbą a treścią... 2000, s. 6].

Podjęcie pracy naukowo-dydaktycznej

Przełom październikowy umożliwił Janowi Steczkowskiemu podjęcie pracy naukowo-dydaktycznej. Jeszcze jako student we wrześniu 1956 r. został przyjęty na asystenturę w Katedrze Statystyki WSE, kierowanej przez prof. Jerzego Fiericha. Po roku, 1 września 1957 r. awansował na stanowisko starszego asystenta. Inny wybitny statystyk krakowski, prof. Stefan Schmidt, również przyjął go od 1 marca 1957 r. na stanowisko starszego asystenta w Katedrze Statystyki Matematycznej WSR, gdzie Jan Steczkowski pracował do 30 września 1961 r. [Między liczbą a treścią... 2000, s. 7].

Z chwilą podjęcia pracy naukowo-dydaktycznej Jan Steczkowski przestał pracować w administracji Uczelni.

Uzyskiwanie stopni i tytułów naukowych

W grudniu 1963 r. uzyskał stopień doktora nauk ekonomicznych na podstawie pracy doktorskiej pt. *Zasady i metody rejonizacji produkcji rolniczej*; jej promotorem był prof. Jerzy Fierich. Praca ta, opublikowana w 1966 r. przez PWRiL [Steczkowski 1966], była w tych czasach jedną z najważniejszych prac z taksonomii i jej zastosowań w badaniach ekonomiczno-rolniczych. 1 marca 1964 r. Jan Steczkowski objął stanowisko adiunkta w Katedrze Statystyki WSE.

Habilitację uzyskał w 1971 r. w Wyższej Szkole Ekonomicznej we Wrocławiu, na podstawie pracy habilitacyjnej pt. *Statystyczna procedura określania struktury zbiorowości* [Steczkowski 1970]. Recenzentami pracy i dorobku naukowego byli

profesorowie: Zbigniew Czerwiński z WSE w Poznaniu, Zdzisław Hellwig z WSE we Wrocławiu oraz Kazimierz Zając z WSE w Krakowie. Decyzją ministra oświaty i szkolnictwa wyższego z dnia 26 lutego 1971 r. Jan Steczkowski został awansowany na stanowisko docenta.

Tytuł naukowy profesora nadzwyczajnego nauk ekonomicznych uzyskał w styczniu 1980 r., a tytuł profesora zwyczajnego w czerwcu 1989 r.

Funkcje akademickie w macierzystej uczelni

W 1972 r. w Wyższej Szkole Ekonomicznej utworzono Międzywydziałowe Studium Doktoranckie, a funkcję kierownika tego studium powierzono Profesorowi Janowi Steczkowskiemu. Świadczyło to o jego wysokiej pozycji naukowej w WSE w tych latach. Funkcję tę pełnił w latach 1972–1975.

W latach 1978–1981 był prorektorem ds. nauki Akademii Ekonomicznej w Krakowie. W tych trudnych latach starał się promować rozwój badań naukowych.

W 1978 r. stworzył Zakład Statystycznych Metod Kontroli Jakości w Instytucie Metod Rachunku Ekonomicznego [*Akademia Ekonomiczna w Krakowie...* 1997, s. 150]. Zakładem tym kierował (z trzyletnią przerwą) aż do przejścia na emeryturę w 1994 r.

Profesor Jan Steczkowski działał także na forum komisji rektorskich i senackich. Należy wymienić jego działalność w Komisji ds. Rozwoju Uczelni, dla której przedstawiał założenia rozwoju. Wniósł wkład w rozwój działalności publikacyjnej Uczelni, przez długie lata będąc członkiem Senackiej Komisji ds. Wydawnictw oraz Senackiej Komisji ds. Badań Naukowych.

Uczestnictwo w gremiach naukowych

Profesor Jan Steczkowski aktywnie uczestniczył w pracach Komitetów i Komisji Naukowych PAN. Był członkiem Komitetu Ekonomiki Rolnictwa Wydziału V PAN w Warszawie, a w jego ramach działał w Komisji Ekonomiki Obrotu Rolniczego oraz w sekcji II Organizacji Gospodarstw Rolniczych; był także członkiem Zespołu Zastosowań Metod Ekonometrycznych w Rolnictwie. Jako ceniony w tym czasie ekspert był członkiem Rady Naukowej Instytutu Rozwoju Wsi i Rolnictwa PAN.

Jako specjalista z zakresu zastosowań metod taksonomicznych był członkiem Komitetu Przestrzennego Zagospodarowania Kraju PAN, działając w Komisji Problematyki Przestrzennej Rolnictwa.

Był członkiem Komitetu Statystyki i Ekonometrii I Wydziału PAN w Warszawie i długoletnim członkiem Komisji Nauk Ekonomicznych PAN Oddział w Krakowie, a w jej ramach członkiem Komitetu Redakcyjnego „*Folia Oeconomica Cracoviensia*”.

Uczestniczył także w pracach Rady Naukowej Instytutu Rolnictwa i Leśnictwa Krajów Tropikalnych na Wydziale Leśnym Akademii Rolniczej w Krakowie.

Kontakty zagraniczne

Profesor Jan Steczkowski miał liczne kontakty zagraniczne, po części związane z jego przeszłością wojenną i studiami na Akademii Handlowej. Jego przyjaciółmi byli m.in. znany socjolog prof. Adam Podgórecki oraz dr Andrzej Ciechanowiecki, Kawaler Maltański, w Akademii Handlowej założyciel Klubu Logofagów, później znany marszałek londyński oraz fundator licznych dzieł dla muzeów Krakowa i Warszawy [*Rodowód Akademii Ekonomicznej w Krakowie...* 2006, s. 5]. Prof. Adam Podgórecki zapraszał go do uniwersytetów kanadyjskich. Jan Steczkowski przebywał także czterokrotnie na Uniwersytecie w Oksfordzie oraz odbył staż w London School of Economics and Political Science. Na zaproszenie Profesora Jana Steczkowskiego na Akademii Ekonomicznej w Krakowie przebywał prof. Alan Stuart, autor napisanego wraz z M.G. Kendalllem epokowego dzieła z zakresu statystyki pt. *The Advanced Theory of Statistics*. Wyjeżdżał także do środkowoeuropejskich ośrodków naukowych (Praga, Bratysława, Pecz, Belgrad i Sarajewo).

Działalność dydaktyczna i kształcenie kadr naukowych

Profesor Jan Steczkowski przywiązywał wielką wagę do dydaktyki. W swojej macierzystej uczelni wykładał rachunek prawdopodobieństwa i statystykę matematyczną, metodę reprezentacyjną, statystyczne metody sterowania jakością, a także statystyczne metody odbiorczych badań jakości. Przez wiele lat prowadził na Uniwersytecie Jagiellońskim wykłady ze statystyki dla studentów etnografii i socjologii. Wypromował ponad 330 magistrów, spośród których kilkunastu zostało pracownikami naukowymi, a kilku jest obecnie profesorami.

Wypromował sześciu doktorów i był recenzentem 34 prac doktorskich. Recenzował także 20 prac habilitacyjnych oraz dorobek naukowy dziewięciu osób ubiegających się o tytuł naukowy profesora [*Między liczbą a treścią...* 2000, s. 9].

Odnaczenia i nagrody

Za zasługi dla Uczelni oraz Polski Profesor Jan Steczkowski został odznaczony Krzyżem Kawalerskim Orderu Odrodzenia Polski, Złotym Krzyżem Zasługi, Krzyżem Armii Krajowej, Krzyżem Partyzanckim, Odznaką Akcji „Burza”. Otrzymał Medal Komisji Edukacji Narodowej oraz dziewięciokrotnie nagrodę Ministra Oświaty i Szkolnictwa Wyższego. Był wielokrotnie nagradzany nagrodą Rektora macierzystej uczelni.

Ważniejsze prace Profesora Jana Steczkowskiego

Zainteresowania naukowe Profesora Jana Steczkowskiego były różnorodne i zmieniały się w czasie. Każdy etap działalności naukowej Profesor zamykał publikacją monografii naukowej, dlatego też łatwo jest scharakteryzować etapy jego rozwoju naukowego, przypominając jego ważniejsze monografie. Ogólnie można wyróżnić pięć obszarów tematycznych zainteresowań naukowych Profesora, które scharakteryzowane zostaną przez przedstawienie pięciu najważniejszych monografii naukowych jego autorstwa.

Profesor Jan Steczkowski był uczniem prof. Jerzego Fiericha. Z tego też względu na początku swojej drogi naukowej zajmował się metodami taksonomicznymi i ich zastosowaniem w badaniach ekonomiczno-rolniczych. Na ten temat napisał swoją pracę doktorską, opublikowaną później pt. *Zasady i metody rejonizacji produkcji rolniczej* [Steczkowski 1966]. Była ona w tym czasie podstawową pracą podsumowującą pierwszy etap zastosowań metod taksonomicznych w badaniach ekonomiczno-rolniczych w Polsce. Istotnym wkładem Profesora Jana Steczkowskiego było zastosowanie metod taksonomicznych do wyboru cech, czyli tzw. dualne podejście w taksonomii [Pociecha i Zajac 1990].

Pracą, która prezentowała zainteresowania Profesora Jana Steczkowskiego problematyką statystycznych metod kontroli jakości, była książka pt. *Statystyczne metody kontroli jakości produktów* [Cyran, Steczkowski i Zajac 1973]. Wielkie znaczenie tej książki polegało przede wszystkim na tym, że podejmowała ona problematykę statystycznych metod kontroli jakości w polskim piśmiennictwie naukowym po niemal dwudziestoletniej przerwie. Stała się ona podstawową pracą przedstawiającą opis metod statystycznej kontroli jakości w procesie produkcji lub odbioru towarów, pokazującą możliwości ich praktycznych zastosowań. Napisane później przez osoby skupione wokół Profesora artykuły naukowe dały asumpt do utworzenia na Akademii Ekonomicznej w Krakowie Zakładu Statystycznych Metod Kontroli Jakości.

Rozprawa habilitacyjna Profesora Jana Steczkowskiego pt. *Statystyczna procedura określania struktury zbiorowości* [1970] mieści się w nurcie prac dotyczących fundamentów metod statystycznych. Przedmiotem rozważań w tej pracy było dochodzenie do empirycznych rozkładów liczebności mających postać szeregów rozdzielczych przez zastosowanie odmiennej od tradycyjnej procedury grupowania danych statystycznych oraz opisu uzyskanych rozkładów za pomocą miar koncentracji i asymetrii. Autor rozpatrywał problemy grupowania danych w nawiązaniu do grupowania taksonomicznego, a następnie ustalał podstawowe własności ich rozkładu statystycznego, głównie na podstawie charakterystyk jego ciągłości i nieciągłości i zaproponowanej miary skokowości. Jako podstawowe charakterystyki rozkładu wykorzystał miary koncentracji i asymetrii. W ostatnim rozdziale przedstawił empiryczny przykład zastosowań proponowanego podejścia.

W historii rozwoju metod statystycznych praca ta stanowi przyczynek do analizy struktury zbiorowości statystycznej.

Na początku lat 80. pojawił się nowy wątek badań statystycznych, dotyczący zagadnień analizy cech jakościowych, którego efektem jest, napisana wspólnie z prof. Aleksandrem Zeliąsiem, praca pt. *Statystyczne metody analizy cech jakościowych* [Steczkowski i Zeliś 1981]. Jej przedmiotem są metody służące opisowi zjawisk wyrażonych za pomocą cech mierzonych w skalach słabych: nominalnej i porządkowej. Książka odpowiadała ówczesnemu zapotrzebowaniu na metody służące statystycznej analizie cech jakościowych, które w przeciwieństwie do cech ilościowych w pracach z zakresu statystyki traktowane były marginalnie. Autorzy przedstawili zasady pomiaru cech statystycznych i skal pomiaru mające podstawowe znaczenie dla dalszej analizy statystycznej i interpretacji uzyskiwanych wyników. Omówili zasady i techniki wyodrębniania typów metodami taksonomicznymi. Przedstawili specyfikę liczenia oraz interpretacji podstawowych charakterystyk cech jakościowych, jakimi są miary położenia, zmienności asymetrii i koncentracji oraz wskaźniki statystyczne. W rachunku prawdopodobieństwa odpowiednikiem cechy jakościowej jest zmienna skokowa. Z tego względu autorzy omówili wybrane rozkłady zmiennej losowej skokowej jedno- i dwuwymiarowej oraz ich parametry. Dużo miejsca poświęcili metodom estymacji i testowania frakcji jako podstawowego parametru charakteryzującego rozkład cechy jakościowej oraz innym testom nieparametrycznym, a także specyficznym miarom współzależności cech jakościowych, jakimi są miary skojarzenia cech oraz miary korelacji rang. Praca ta w tamtych czasach wypełniała lukę w polskiej literaturze statystycznej, gdyż dotychczas badacze zajmujący się analizami cech jakościowych zmuszeni byli do korzystania w tym zakresie z prac anglojęzycznych lub prac z zakresu psychologii i socjologii matematycznej.

Kolejną pracą napisaną wraz z prof. Aleksandrem Zeliąsiem, podobnie jak poprzednia wypełniającą lukę w polskiej literaturze statystycznej tamtych czasów, była książka pt. *Analiza wariacyjna i kowariancyjna w badaniach ekonomicznych* [Steczkowski i Zeliś 1982]. Jednym z podstawowych źródeł rozwoju metod statystycznych było prowadzenie doświadczeń rolniczych. Na podstawie potrzeb dotyczących statystycznego opracowania wyników prowadzonych doświadczeń powstała gałąź wiedzy statystycznej nazywana planowaniem eksperymentów. W naukach społeczno-ekonomicznych nie ma możliwości nieograniczonego wykonywania eksperymentów, z tego też względu metody statystyczne wyrosłe na gruncie doświadczalnictwa rolniczego nie były szerzej stosowane do badań społeczno-ekonomicznych. Niektóre z nich mają jednak szersze zastosowanie niż tylko w klasycznym eksperymencie. Należy do nich metoda analizy wariacji, pozwalająca na testowanie równości przeciętnych w wielu populacjach i jednocześnie badania wpływu wyróżnionych czynników na zmiany tych przeciętnych.

W omawianej książce autorzy, wychodząc od problemu konstruowania schematów badań, przedstawili formalne podstawy analizy wariancji i kowariancji. Dużo miejsca poświęcili kwestiom weryfikacji założeń analizy wariancji, wyodrębnianiu jednorodnych zespołów średnich grupowych oraz schematom rozwiązań układów klasyfikacyjnych. Zwrócili też uwagę na nieparametryczne i przybliżone procedury analizy wariancji. Rozszerzeniem analizy wariancji jest analiza kowariancji, pozwalająca na badanie zależności pomiędzy zmienną objaśnianą a pewnymi zmiennymi objaśniającymi. W tym względzie stanowi ona w doświadczalnictwie alternatywę dla klasycznych metod analizy regresji. Prezentowana praca została wydana w serii Biblioteka Ekonometryczna PWN i uzupełniała dotychczasowe metody badań ekonometrycznych o metody wywodzące się z doświadczalnictwa rolniczego.

W pracach naukowych Profesora Jana Steczkowskiego już od początku lat 80. ubiegłego wieku pojawiają się zagadnienia dotyczące badań z zakresu teorii i zastosowań metody reprezentacyjnej w analizach społeczno-ekonomicznych. Ukoronowaniem prac z tego zakresu jest monografia naukowa pt. *Metoda reprezentacyjna w badaniach zjawisk ekonomiczno-społecznych* [1995]. Praca ta w polskiej literaturze przedmiotu stanowiła kolejny etap rozwoju metod próbkowania zastosowany do zagadnień społeczno-ekonomicznych po wydanej ponad 20 lat wcześniej pracy Ryszarda Zasepy [1972] i stanowiła bodziec do publikowania w latach późniejszych kolejnych prac z tego zakresu. W książce tej przedstawiono istotę i teoretyczne podstawy metody reprezentacyjnej oraz zakres jej praktycznych zastosowań. Dogłębnie omówiono etapy projektowania badań reprezentacyjnych. W kolejnych rozdziałach przedstawiono schematy losowania indywidualnego nieograniczonego ze zwracaniem i bez zwracania, losowania warstwowego, losowania systematycznego, losowania próby przy zróżnicowanych prawdopodobieństwach wyboru oraz losowania zespołowego. W zakresie estymacji parametrów skupiono uwagę na estymatorach ilorazowych i regresyjnych. W końcowych rozdziałach pracy zwrócono uwagę na wielofazowe losowanie próby, badania panelowe oraz na podejście bayesowskie i analizę sekwencyjną. Na tej monografii oraz na innych pracach Profesora Jana Steczkowskiego wykształciły się zastępy współczesnych specjalistów z zakresu, zyskujących coraz bardziej na popularności, niewyczerpujących badań statystycznych.

Profesor Jan Steczkowski we wspomnieniach uczniów i kolegów

Na koniec warto dodać kilka słów na temat tego, jak zapamiętali Profesora jego uczniowie i młodszy koledzy. Do jego wyróżniających cech charakteru należy zaliczyć stanowczość i niezmiennność poglądów. Był dzieckiem wojny i żołnierzem Armii Krajowej. Nigdy się tego nie wyrzekł i w okresie powojennym cierpiał za to, przebywając ponad rok w więzieniu, tracąc możliwość podjęcia studiów na uczelni

państwowej, a potem podjęcia pracy naukowo-dydaktycznej czy znosząc szykany PZPR-owskich aparatczyków w okresie pracy w administracji WSE [*Rodowód Akademii Ekonomicznej w Krakowie...* 2006, s. 7].

Solidarnościowy przełom 1980 r. zastał go na stanowisku prorektora ds. nauki. Od chwili powstania organizacji uczelnianej NSZZ Solidarność w październiku 1980 r. był jej członkiem jako jedyny z ówczesnego grona rektorskiego. Po 1989 r. włączył się w działalność Krakowskiego Towarzystwa Przemysłowego oraz był aktywnym uczestnikiem dyskusji nad przyszłym kształtem ustroju ekonomicznego wolnej Polski.

Profesor cechował się dociekliwością i wysokim poziomem inteligencji. Gdy zabierał głos w dyskusji, jego analizy były dogłębne, a riposty cięte. W dyskusjach trudno było mu dorównać, niezależnie od tego, czy toczyły się na tematy naukowe, czy społeczno-ekonomiczne. Swoim współpracownikom i uczniom nie skąpił rad. Jego uczniowie korzystali z uwag dotyczących metodologii badań naukowych.

Profesora wyróżniały także jego pogodny charakter i życzliwość dla ludzi. Był duszą towarzystwa, godzinami potrafił żartować i opowiadać dowcipy. Do dzisiaj jego młodszy koledzy, opowiadając dowcip, wspominają: „To był kawał Steczkowskiego”. Nie wypowiadał o innych negatywnych sądów, był dla wszystkich życzliwy. Wszyscy lubili przebywać w jego towarzystwie.

Prof. dr hab. inż. Jan Steczkowski zmarł w dniu 14 lutego 2016 r. (w dziesiątą rocznicę śmierci swojego młodszego kolegi – prof. Aleksandra Zeliasia) w 93. roku życia.

Literatura

- Akademia Ekonomiczna w Krakowie w latach 1925–2000* [2002], red. J.M. Małecki, Wydawnictwo Akademii Ekonomicznej w Krakowie, Kraków.
- Akademia Ekonomiczna w Krakowie w okresie organizacji instytucyjnej 1969–1992* [1997], red. J.M. Małecki, Akademia Ekonomiczna w Krakowie, Kraków.
- Cyran J., Steczkowski J., Zając K. [1973], *Statystyczne metody kontroli jakości produktów*, PWE, Warszawa.
- Między liczbą a treścią. Księga jubileuszowa dla uczczenia 75. urodzin Profesora Jana Steczkowskiego* [2000], Akademia Ekonomiczna w Krakowie, Kraków.
- Pocięcha J., Zając K. [1990], *Wkład ośrodka krakowskiego w rozwój teorii i zastosowań metod taksonomicznych* [w:] *Taksonomia – teoria i jej zastosowania*, Akademia Ekonomiczna w Krakowie, Kraków.
- Rodowód Akademii Ekonomicznej w Krakowie. Wspomnienia sprzed przeszło 50 lat prof. zw. dr. hab. inż. Jana Steczkowskiego* [2006], Kraków.
- Steczkowski J. [1966], *Zasady i metody rejonizacji produkcji rolniczej*, PWRiL, Warszawa.
- Steczkowski J. [1970], *Statystyczna procedura określania struktury zbiorowości*, Zeszyty Naukowe WSE w Krakowie, Seria Specjalna: Rozprawy Habilitacyjne, nr 21, Kraków.

Steczkowski J. [1995], *Metoda reprezentacyjna w badaniach zjawisk ekonomiczno-społecznych*, PWN, Warszawa–Kraków.

Steczkowski J., Zeliaś A. [1981], *Statystyczne metody analizy cech jakościowych*, PWE, Warszawa.

Steczkowski J., Zeliaś A. [1982], *Analiza wariancyjna i kowariancyjna w badaniach ekonomicznych*, PWN, Warszawa.

Zasępa R. [1972], *Metoda reprezentacyjna*, PWE, Warszawa.

| Janusz L. Wywiat

On the Evaluation of Sample Size Required for a Good Approximation by the Normal Curve for Some Statistics*

Abstract

Testing hypotheses or evaluation confidence intervals requires knowledge of some statistics' distributions. It is convenient if the probability distribution of the statistic converges to normal distribution when the sample size is sufficiently large. This paper examines the problem of how to evaluate sample size in order to determine that a statistic's distribution does not depart from normal distribution by more than an assumed amount. Two procedures are proposed to evaluate the necessary sample size. The first is based on Berry-Esseen inequality while the second is based on simulation procedure. In order to evaluate the necessary sample size, the distribution of the sample mean is generated by replicating samples of a fixed size. Next, the normal distribution of the evaluated sample means is tested. The size of the generated samples is gradually increased until the hypothesis on the normality of the sample mean distribution is not rejected. This procedure is applied in the cases of statistics other than sample mean.

Keywords: sample size, central theorem, sampling scheme, computer simulation, chi-square test of goodness of fit.

JEL Classification: C12, C15.

| Janusz L. Wywiat, University of Economics in Katowice, Department of Statistics, Econometrics and Mathematics, 1 Maja 50, 40-287 Katowice, e-mail: janusz.wywiat@ue.katowice.pl

| * This paper presents the results of a research project conducted with the financial support of the National Science Centre in Poland (grant number: DEC-2012/07/B/HS4/03073).

1. Introduction

Statistical inference procedures such as testing hypotheses or evaluating confidence intervals depend on the distribution properties of test statistics or estimators which can be evaluated on the basis of complex samples. Usually, a statistic's exact distribution is not known except as some function of a normal simple random sample. In this situation, it is convenient if the probability distribution of the statistic converges to normal distribution when the sample size is sufficiently large. This leads to the problem of how to evaluate sample size so that the departure of the distribution of the statistic from normal distribution is not larger than the level assumed. This problem is frequently taken into account in statistics literature, but usually where simple random sample is involved. This paper discusses the problem under complex samples drawn from fixed and finite populations.

In some situations, it is possible to observe all values of an auxiliary variable in an entire population. Moreover, let us assume that the value of the correlation coefficient between the auxiliary variable and the variable under consideration here is close to one. In this case, we can expect that the degree of convergence to normal distribution of, e.g., the sample average of the auxiliary variable and the distribution of the sample mean of the variable under consideration will be similar. This allows us to assess the size of the sample, providing a sufficient degree of convergence of the sample mean distribution to normal distribution.

Two procedures are proposed to evaluate the necessary sample size. The first is based on the Berry-Esseen equality, while the second is based on a simulation procedure. The sample mean's distribution is generated by replicating samples of fixed size. Next, the normal distribution of the evaluated sample means is then tested. The size of the generated samples is gradually increased until the hypothesis on the normality of the sample mean distribution is not rejected. The normality of generated values of the sample mean is tested by means of the chi-square test of goodness of fit. The hypothesis on normal distribution is verified under the assumed significance level as well as the power of the test. The outlined procedure is used to assess the necessary sample size of statistics other than sample mean. Complex sampling schemes are also taken into account.

The properties of central limit theorems allow us to evaluate sample sizes in such a way that the probability distribution of, e.g., the standardised sample mean does not differ from standard normal distribution by more than an assumed level. The distribution of the simple random sample frequency is approximated by means of several methods reviewed e.g. by G. A. F. Seber (2013) and T. P. Ryan (2013). In the case of continuous or integer random variables, a bootstrap version of the statistics can be analysed. In this case, the statistical distribution can be

approximated by means of the well-known F. Y. Edgeworth (1907) expansion, which has been detailed by P. Hall (1992). In the case of sampling from a fixed population, the central theorems have been considered e.g. by Y. G. Berger (1998), W. A. Fuller (2009) and J. Hájek (1964, 1981).

Using appropriately prepared computer simulation experiments, it is possible to determine what sample size is necessary to assure sufficient convergence of the distribution of a statistic to the appropriate asymptotic distribution. This problem has been considered e.g. by M. R. Chernick and C. Y. Liu (2002) and T. P. Ryan (2013) in the context of sample frequency. F. Greselin and M. Zenga (2006) considered the simulation analysis for determining sample size, which assures sufficient convergence of Gini's statistic to normality. Some similar ideas are developed below.

Let $(x; y)$ be highly correlated variables observed in a population U of size N . These variables' values are denoted by $(x_i; y_i)$, $i = 1, \dots, N$. We assume that the values of the auxiliary variable x are observed in the whole population U but the values of the variable under study y are observed only in a sample s of size $n < N$ drawn from U . The random sample will be denoted by S and its observation by s treated as the set consisting of the population elements. The sample is drawn from the population according to sampling design denoted by $P(s) > 0$ for all $s \in \mathbf{S}$ and $\sum_{s \in \mathbf{S}} P(s) = 1$, where \mathbf{S} is the sample space, see e.g. C. M. Cassel *et al.* (1977) or Y. Tillé (2006).

2. Numerical Approximation of Sample Size

Let $z_{x,S}$ and $z_{y,S}$ be statistics evaluated based on data observed in sample S . Because values of x are observed in the whole population U , it is possible to observe values $z_{x,s}$ of $z_{x,S}$ in all samples which can be drawn from population U . In practice, values of variable y are observed only in one sample s . Let us assume that the two-dimensional normal distribution (with the marginal distributions equal to standard normal distribution and the correlation coefficient close to one) is the limit distribution of statistics $(z_{x,S}, z_{y,S})$. This convergence can be proved using H. Cramér's (1946) results. Hence, we can expect that when we evaluate sample size n_o , which assures the sufficient convergence of statistic $z_{x,S}$ to standard normal distribution, then the same sample size is also sufficient for the convergence of $z_{y,S}$ to standard normal distribution.

Let us assume that $x_i = y_i + d_i$, $i = 1, \dots, N$. x_i can be treated as a measure of y_i contaminated by error d_i . The following notation will be useful:

$$\bar{x} = \frac{1}{N} \sum_{i \in U} x_i, \quad c_{x,r} = \frac{1}{N} \sum_{i \in U} (x_i - \bar{x})^r, \quad v_x = c_{x,2}, \quad \tau_{x,r} = \frac{1}{N v_x^{r/2}} \sum_{i \in U} |x_i - \bar{x}|^r,$$

$$\eta_{x,r} = \frac{c_{x,r}}{v_x^{r/2}}, \quad \bar{x}_S = \frac{1}{n} \sum_{i \in S} x_i, \quad v_{x,S} = \frac{1}{n-1} \sum_{i \in S} (x_i - \bar{x})^2, \quad r = 2, 3, \dots$$

Definitions of the parameters \bar{y} , $\eta_{y,r}$, $\eta_{d,r}$, $\tau_{y,r}$, $\tau_{d,r}$, $\lambda_{y,r}$, $\lambda_{d,r}$, \bar{y}_S , v_y , v_d , $v_{y,S}$, $v_{d,S}$ are analogous with the above ones. Under the assumption that variables y and d are independent, the squared correlation coefficient between x and y is equal to:

$$\kappa = \frac{v_y}{v_y + v_d} = \frac{v_y}{v_x} = 1 - \frac{v_d}{v_x}, \quad 0 < \kappa \leq 1.$$

The following standardised sample means will be considered:

$$z_{y,S} = \frac{\bar{y}_S - \bar{y}}{\sqrt{v_{y,S}}} \sqrt{n}, \quad z_{x,S} = \frac{\bar{x}_S - \bar{x}}{\sqrt{v_{x,S}}} \sqrt{n}. \quad (1)$$

When S is the simple random sample drawn with replacement from U , the Berry-Esseen inequality, following M. Krzyśko (2000), becomes:

$$\sup_s |F_{y,s}(z) - \Phi(z)| \leq \zeta \frac{\tau_{y,3}}{\sqrt{n}},$$

where $F_{x,s}(z)$ is the sample distribution of $z_{x,S}$, $\Phi(z)$ is the distribution of standard normal random variable, and $\pi^{-1/2} \leq \zeta < 0,8$. J. L. Wywił (2016) showed that:

$$\sup_s |F_{y,s}(z) - \Phi(z)| \leq 0,8 \frac{\tau_{y,3}}{\sqrt{n}} \leq 0,8 \sqrt{\frac{\eta_{y,4}}{n}}.$$

Under the assumption that variables x and d are independent, after appropriate algebraic computations it can be shown that:

$$\begin{aligned} v_y^2 \eta_{y,4} &= c_{y,4} = c_{x,4} + 6v_x v_d + c_{d,4} = c_{x,4} + 6v_x^2 v_d^2 + v_x^2 (1 - \kappa)^2 \eta_{d,4} = \\ &= \frac{1}{\kappa^2} \left(\eta_{x,4} + 6(1 - \kappa) + \eta_{x,4} \frac{v_d^2}{v_x^2} \right). \end{aligned}$$

Finally:

$$\eta_{y,4} = \kappa^{-2} (\eta_{x,4} + 6(1 + \kappa) + \eta_{d,4} (1 - \kappa)^2) = f(\kappa),$$

$\eta_{y,4}$ is a strictly decreasing function of κ because we can show that $f'(\kappa) < 0$ for $\kappa \in (0; 1]$. Hence, inequality:

$$\sup_{s \in S} |F(z_{y,s}) - \Phi(z_{y,s})| \leq 0,8 \sqrt{\frac{f(\kappa)}{n}}.$$

Let us assess the necessary sample size when the approximate values of the parameters $\eta_{x,4}$, $\eta_{d,4}$, and κ are known. $f(0)$ takes an infinitely large value and

$f(1) = \eta_{x,4}$. When we assume that $0,8 \sqrt{\frac{f(\kappa)}{n}} \leq \Delta_0$, where Δ_0 is an admissible

difference between the sample distribution of the statistic and the standard normal distribution, the necessary sample size yields the following expression:

$$n \geq n_o = \left\lceil \frac{0.64(\eta_{x,4} + 6(1 + \kappa) + \eta_{d,4}(1 - \kappa)^2)}{\Delta_0^2 \kappa^2} \right\rceil.$$

Hence, necessary sample size n_o is a decreasing function of coefficient κ .

Example 1. For instance, when $\eta_{x,4} = 4$, $\eta_{d,4} = 4$, $\kappa = 0.99$ and $\Delta = 0.01$, then $n > n_o = 26676$. If $\eta_{x,4} = 4$, $\eta_{d,4} = 4$, $\kappa = 0.9$ and $\Delta = 0.01$ then $n > n_o = 42539$.

3. Simulation Evaluation of Sample Size

3.1. Chi-square Test of Goodness of Fit

Sample size n_o will be evaluated on the basis of the following simulation experiment. Under assumed sample size, the normality of $z_{x,S}$ is tested on the basis of its simulated values. In order to do that, a series of samples ($s_j, j = 1, \dots, r$) each of size n are drawn independently from population U according to assumed sampling design. Let $\mathbf{z}_{x,S}^{(n)} = (z_{x,s_1}^{(n)}, \dots, z_{x,s_r}^{(n)})$ be the sequence of the statistics evaluated based on the sequence of samples. Next the normal distribution of $z_{x,S}$ can be tested on the basis of data $\mathbf{z}_{x,S}^{(n)}$. When the hypothesis on normality is not rejected, we can expect that the distribution of statistic $z_{x,S}$ for the sample size $n_o = n$ is sufficiently close to standard normal distribution. If the hypothesis on normality is rejected, a new series of samples is drawn, but each of them is of size $n + d$, where $d \geq 1$. Using these samples, the sequence $\mathbf{z}_{x,S}^{(n+d)}$ is evaluated, allowing us to again test the normality of $z_{x,S}$ but for larger sample size $n + d$. The procedure is repeated until the hypothesis on the normality is not rejected. In order to verify the hypothesis that $z_{x,S}$ has standard normal distribution, several test statistics can be used, e.g. Kolmogorov or Shapiro-Wilk statistics. However, the powers of these tests cannot be easily controlled. That is why we use the chi-square test of goodness of fit.

Usually, $z_{x,S}$ is used to construct confidence intervals or test statistics on the expected value of variable x . In this case only the quantiles of high or small degrees of $z_{x,S}$ have to be close to the appropriate quantiles of standard normal distribution. J. L. Wywiał (2016) proposed the following procedure for evaluating the necessary sample size. When $Z \sim N(0;1)$, we expect that z , where $P(Z < z_\lambda) = \lambda$ is close to $z_{n,\lambda}$ where $P(z_{x;n} < z_{n,\lambda}) = \lambda$. Usually, $\lambda = 0.01; 0.05; 0.1; 0.9; 0.95; 0.99$. Let $\boldsymbol{\lambda} = [\lambda_1, \dots, \lambda_K]$, where $P(Z < z_{k,\lambda}) = \lambda_k, k = 1, \dots, K$. More formally, the following hypothesis must be tested:

$$H_0: \boldsymbol{\lambda} = \boldsymbol{\lambda}_0, \quad H_1: \boldsymbol{\lambda} = \boldsymbol{\lambda}_1 \neq \boldsymbol{\lambda}_0.$$

Let $\boldsymbol{\omega} = [\omega_1, \dots, \omega_{K+1}]$, where $\omega_k = \lambda_k - \lambda_{k-1}$, $k = 2, \dots, K$, $\omega_{1k} = \lambda_1$, $\omega_{K+1} = 1 - \lambda_K$. The above hypotheses can be rewritten as follows:

$$H_0: \boldsymbol{\omega} = \boldsymbol{\omega}_0, \quad H_1: \boldsymbol{\omega} = \boldsymbol{\omega}_1 \neq \boldsymbol{\omega}_0.$$

We can verify this using chi-square test of goodness of fit under fixed significance level α and power α . Several variants of the test have been considered – e.g. by F. C. Drost *et al.* (1989) and T. J. Santer and D. E. Duffy (1989). The test statistic is as follows:

$$Q_{n,r} = r \sum_{k=1}^{K+1} \frac{(W_k - \omega_k)^2}{\omega_k}, \quad (2)$$

where:

$$W_k = \frac{1}{r} \sum_{j=1}^r I_k(z_{x,s_j}^{(n)}), \quad (3)$$

if $z_{\lambda_{k-1}} < z_{x,s_j}^{(n)} \leq z_{\lambda_k}$, then $I_k(z_{x,s_j}^{(n)}) = 1$ otherwise $I_k(z_{x,s_j}^{(n)}) = 0$, $k = 1, \dots, K + 1$, $z_{\lambda_0} = -\infty$, $z_{\lambda_{K+1}} = \infty$. Under a sufficiently large number r , the statistic $Q_{n,r}$ has chi-square distribution with K degrees of freedom (denoted by $\chi_K^2(\kappa)$) and the following non-centrality parameter:

$$r\delta(\boldsymbol{\omega}_0, \boldsymbol{\omega}_1) = \sum_{k=1}^{K+1} \frac{(\omega_{0,k} - \omega_{1,k})^2}{\omega_{0,k}}. \quad (4)$$

The quantity $\delta(.,.)$ can be treated as distance between distributions specified by the hypotheses H_0 and H_1 . Particularly, we will consider the following vector of probabilities:

$$\boldsymbol{\omega}_0 = [0.01 \ 0.04 \ 0.05 \ 0.8 \ 0.05 \ 0.04 \ 0.01],$$

$$\boldsymbol{\omega}_1^{(1)} = [0.012 \ 0.048 \ 0.06 \ 0.76 \ 0.06 \ 0.048 \ 0.012],$$

$$\boldsymbol{\omega}_1^{(2)} = [0.011 \ 0.044 \ 0.055 \ 0.78 \ 0.055 \ 0.044 \ 0.011].$$

Let us note that we do not consider, e.g., the alternative: $\boldsymbol{\omega}_1^{(*1)} = [0.008 \ 0.036 \ 0.04 \ 0.832 \ 0.04 \ 0.036 \ 0.008]$ because the chi-square test does not select the difference between $\boldsymbol{\omega}_1^{(1)}$ alternatives and $\boldsymbol{\omega}_1^{(*1)}$. In this case the non-centrality coefficient takes the same value. Expression (4) allows us to calculate that $\delta(\boldsymbol{\omega}_0, \boldsymbol{\omega}_1^{(1)}) = 0.01$, $\delta(\boldsymbol{\omega}_0, \boldsymbol{\omega}_1^{(2)}) = 0.0025$.

When hypothesis H_0 is true, the test statistic $Q_{n,r}$ has the central chi-square distribution χ_K^2 with $K = 6$ degrees of freedom, provided that r is large. W. G. Cochran (1952) wrote that convergence to asymptotic distribution is sufficiently accurate when $r_0 = 5/\omega_0$ where $\omega_0 = \min_{k=1, \dots, K+1} \{\omega_k\}$. Hence, in our

case, for $\omega_0 = 0.01$, $r_0 = 500$. The algorithm for evaluating r is as follows. Firstly, based on χ_K^2 distribution, the critical value q_α of the test is determined under an assumed significance level α . Next, the power of the test is calculated for $r \geq r_0$ according to $\beta_r = P(\chi_K^2(r\delta(\omega_0, \omega_1)) \geq q_\alpha | H_1)$ and $\alpha = P(\chi_K^2 \geq q_\alpha | H_0)$. The r_0 is treated as the start number of sample replication. If for fixed $r \geq r_0$, β_r is not less than the assumed level β , then size r is sufficiently large and it will be denoted by $r_\#$. Otherwise, the power is calculated for $r + 10$ and so on.

Example 2. Consider these hypotheses:

$$H_0: \omega = \omega_0, \quad H_1: \omega = \omega_1^{(2)}. \quad (5)$$

The significance level is $\alpha = 0.05$ and power $\beta = 0.95$, $\delta(\omega_0, \omega_1^{(2)}) = 0.0025$. The above algorithm leads to the necessary number of the sample replication being $r_\# = 8350$. The next variants for calculating $r_\#$ are presented in the first three columns of Table 1 – also see (Wywił 2016).

3.2. Evaluation of Sample Size in Order to Assure the Normal Distribution of Some Statistics

Let us consider the determination of sample size n of a simple random sample in order to assure convergence of the standardised sample mean distribution to standard normal distribution when the sample is drawn from non-normal distribution. Let us suppose that a sample of size n is drawn with replacement from a population of size N , where values of variable x are observed. Next, the statistic $z_{x,S}$ is evaluated in the case when simple random sample is drawn with replacement. For sampling without replacement, the test statistic is as follows:

$$z_{1x,S} = \frac{\bar{x}_S - \bar{x}}{\sqrt{(N-n)v_{x,S}}} \sqrt{Nn}.$$

Our purpose is to evaluate the sample sizes so that $z_{x,S}$ and $z_{1x,S}$ converge sufficiently well to standard normal distribution. In order to do this, the sample sizes are replicated r -times. Values of the statistics are calculated on the basis of the replicated samples. Next, the value of the chi-square test statistic is calculated by means of expression (2). Bear in mind that under assumed significance level α and the number of sample replications $r_\#$, the chi-square test has power β . If the test rejects hypothesis H_0 , then an increase of d has to be added to sample size n and the described algorithm has to be repeated for $n + d$. When the test does not reject the hypothesis, we state that $n = n_{\alpha,\beta}$.

Table 1. The Necessary Sample Sizes for Testing Normal Distributions of Statistics $z_{x,S}$ and $z_{1x,S}$ under Assumed Significance Levels and Powers of the chi-square Test

α	β	$r_{\#}$	p	\underline{n}	\underline{n}_1
0.1	0.9	5870	1	920	930
			2	520	510
			4	300	320
0.05	0.95	8350	1	1060	1040
			2	590	560
			4	320	340
0.01	0.99	14 010	1	1440	1380
			2	780	790
			4	450	430
0.005	0.995	16 420	1	1530	1550
			2	860	780
			4	470	410

Source: the author's own calculations.

Example 3. Let us consider a population of $N = 100,000$ values generated according to gamma probability distribution with shape parameter p and a scale parameter of one. Using the algorithm, the necessary sample sizes are evaluated. A computer simulation implements the above algorithm under the hypothesis given by (5) and several combinations of the significance levels and powers. The obtained results lead to Table 1. The algorithm for evaluating necessary sample sizes is replicated 10-times, which lets us compute the mean sample sizes denoted by \underline{n} and \underline{n}_1 in the case of statistics $z_{x,S}$ and $z_{1x,S}$, respectively. In Table 1 we take into account only such α and β that $\alpha + \beta = 1$. Note, however, that this assumption is unnecessary.

Table 1 shows that the assessed mean sample sizes for both sampling without replacement and sampling with replacement are comparable. In general, when the significance level decreases and the power increases, the necessary sample size increases.

3.3. Evaluation of Complex Random Sample Size in Order to Assure the Normality of Some Statistics

Using the results of the previous sub-section, we can evaluate the necessary sample size for two complex sampling designs. The first is the well-known D. B. Lahiri (1951), H. Midzuno (1952) and A. R. Sen (1953) sampling design, which is defined by the following probability function:

$$P_2(s) = \binom{N}{n}^{-1} \frac{\bar{x}_s}{\bar{x}}, \quad s \in \mathbf{S},$$

where \mathbf{S} is sample space. The sampling design is defined for a positive valued variable observed in the whole population. The inclusion probabilities of the first and second order are as follows:

$$\pi_k = \frac{N-n}{(N-1)N} \frac{x_k - \bar{x}}{\bar{x}} + \frac{n}{N}, \quad \pi_{k,l} = \frac{(n-1)n}{(N-1)N} \frac{n-1}{N-2} \left(\pi_k + \pi_l - \frac{2n}{N} \right), \quad k \neq l = 1, \dots, N.$$

The sampling scheme implementing the sampling design is as follows. Let $p_k = x_k/x_U$, where $k = 1, \dots, N$ and $x_U = N\bar{x}$. The first element is drawn from the population into the sample with probability $p_k, k = 1, \dots, N$. The next $n - 1$ elements are drawn without replacement from the remaining $N - 1$ elements of the population as a simple sample of size $n - 1$.

The population mean \bar{y} is estimated by means of the following Horvitz-Thompson (1952) estimator:

$$\bar{y}_{HT,S} = \frac{1}{N} \sum_{k \in S} \frac{y_k}{\pi_k}.$$

This is an unbiased estimator of \bar{y} , when $\pi_k > 0$ for all $k = 1, \dots, N$. The unbiased estimator of variance proposed by A. R. Sen (1953), F. Yates and P. M. Grundy (1953) is as follows:

$$V_{1S}(\bar{y}_{HT,S}) = -\frac{1}{2N} \sum_{k \in S} \sum_{l \in S, k \neq l} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right) \frac{\Delta_{k,l}}{\pi_{k,l}},$$

where $\Delta_{k,l} = \pi_{k,l} - \pi_k \pi_l$ for $k \neq l$ and $\Delta_{k,k} = \pi_k(1 - \pi_k)$. This estimator is useful only when $\pi_{k,l} > 0$ for all $k, l = 1, \dots, N$ and $k \neq l$.

Next, the sampling design provides samples s of fixed size n drawn with replacement from population U with the above defined probabilities $p_k, k = 1, \dots, N$. This is a particular case of the multinomial sampling design (Tillé 2006, pp. 70–73). In this case the parameter \bar{y} is estimated by means of the following Hansen-Hurvitz (1943) estimator:

$$\bar{y}_{HH,S} = \frac{1}{n} \sum_{k \in S} \frac{y_k}{p_k}.$$

The unbiased estimator of its variance is:

$$V_{2S}(\bar{y}_{HH,S}) = \frac{1}{n(n-1)} \sum_{k \in S} \left(\frac{y_k}{p_k} - \bar{y}_{HH,S} \right)^2.$$

Let us consider the following statistic:

$$t_{HT,S} = \frac{\bar{y}_{HT,S} - \bar{y}}{\sqrt{V_{1S}(\bar{y}_{HT,S})}}, \quad t_{HH,S} = \frac{\bar{y}_{HH,S} - \bar{y}}{\sqrt{V_{2S}(\bar{y}_{HH,S})}}. \tag{6}$$

As we did in section 3.2, here we evaluate the necessary sample size in order to assure sufficient convergence of the distributions of statistics $t_{HT,S}$ and $t_{HH,S}$ to standard normal distribution.

Example 4. The simulation analysis is based on three appropriately generated sets of values $(y_k, x_{i,k})$, $k = 1, \dots, N$ of a two-dimensional random variable denoted by (Y, X_i) , where $X_i = Y + D_i$ and $i = 1, 2, 3$. Variable Y has gamma distribution with a shape parameter equal to 4 and a scale parameter of 1. Moreover, $D_1 \sim N(0; 0.425)$, $D_2 \sim N(0; 0.294)$ and $D_3 \sim N(0; 0.125)$. The correlation coefficients are: $\rho(Y, X_1) = 0.9$, $\rho(Y, X_2) = 0.95$ and $\rho(Y, X_3) = 0.99$. For instance, values of X_i can be treated as observations of Y contaminated by errors which are values of D_i , $i = 1, 2, 3$. According to the above sampling design, samples are replicated r -times. On the basis of such samples, the values of statistic $t_{HT,S}$ and $t_{HH,S}$ are calculated. As in Example 2, the hypothesis given by (5) is tested by means of the chi-square statistic. If the hypothesis is rejected, then sample size n is increased to $n + 10$ and the procedure is repeated. The algorithm is replicated until the hypothesis is not rejected.

Table 2. The Necessary Sample Sizes for Ensuring Normal Distributions of the Statistics $t_{HT,S}$ and $t_{HH,S}$ under Assumed Significance Levels and Powers of the Test

α	β	r	ρ	\underline{n}	
				$\bar{y}_{HT,S}$	$\bar{y}_{HH,S}$
0.1	0.9	5870	0.9	360	50
			0.95	440	50
			0.99	450	60
0.05	0.95	8350	0.9	530	40
			0.95	620	60
			0.99	760	60
0.01	0.99	14 010	0.9	570	60
			0.95	660	60
			0.99	1330	60

Source: the author's own calculations.

A sample size obtained in such a way is treated as sufficient for normal distribution of the statistic being considered under the assumed significance level as well as the power of the chi-square test statistic. The algorithm for evaluating the necessary sample sizes is replicated 10-times, allowing us to compute \underline{n} . The results of the simulation experiments can be found in Table 2.

Based on Table 2, we can say that the necessary sample size, in order to assure sufficient convergence of statistic $t_{HT,S}$ to normality under sampling design $P_2(s)$, is at least seven times larger than it is for $t_{HH,S}$ under a sample drawn with replacement with probabilities proportional to the auxiliary variable values. Under the considered variants of significance levels and powers of the test, the necessary sample size in order to ensure the normal distribution of $t_{HH,S}$ distribution oscillates around 60. For $t_{HT,S}$ distribution, the necessary sample size increases when the significance level decreases and the power increases.

4. Conclusions

Both of the methods considered for evaluating necessary sample sizes in order to ensure statistics are normally distributed were based on the assumption that an auxiliary variable is known from a whole population. The method requires the assumption that standard normal distribution is the asymptotic distribution of the statistics under analysis. The first method of determining the sample size is based on the Berry-Esseen inequality. The particular case of dependence between variables x and y considered in Example 2 lets us conclude that the necessary sample size decreases when the correlation coefficient between these variables increases.

In the case of the next method, the necessary sample size is evaluated by means of appropriate formulation of hypotheses on the normality coefficient. The tested and alternative hypotheses (see expression (5)) are constructed in such a way that the tails of the standard normal distribution are especially taken into account. The proposed simulation algorithm based on testing appropriate statistical hypothesis leads to the following conclusion. Under the evaluated sample size n_o , the hypothesis H_0 on normal distribution of considered statistic is not rejected. This decision is wrong (type II error) with a probability of $v = 1 - \beta$. Moreover, in the previous steps of the algorithm, when the sample size was shorter than the optimal level n_o , the alternative hypothesis H_1 (determining the non-admissible distribution) was accepted. This was the wrong decision (type I error) with probability α . Hence, in consequence, the probabilities α and v measure the risk of assessing the necessary sample size on the level n_o . When under the fixed distance measure $\delta(\omega_0, \omega_1)$ (see expression (4)) between the distributions specified by the hypotheses H_0 and H_1 , we decrease the level of α or the level of v , then size n_o increases. This usually causes the costs of data observations to rise. Hence, compromise levels for δ , α , v and n_o need to be found. This procedure can be applied to more complicated statistics or more complex sampling schemes than those considered in this paper. Moreover, it is possible to generalise the obtained results on distributions of bootstrap type statistics.

Bibliography

- Berger Y. G. (1998), *Rate of Convergence to Normal Distribution for Horvitz-Thompson Estimator*, "Journal of Statistical Planning and Inference", vol. 67, [https://doi.org/10.1016/s0378-3758\(97\)00107-9](https://doi.org/10.1016/s0378-3758(97)00107-9).
- Cassel C. M., Särndal C. E., Wretman J. H. (1977), *Foundation of Inference in Survey Sampling*, John Wiley & Sons, New York–London–Sydney–Toronto.
- Chernick M. R., Liu C. Y. (2002), *The Saw-toothed Behavior of the Power versus Sample and Software Solutions: Single Binomial Proportion Using Exact Methods*, "The American Statistician", vol. 56, <https://doi.org/10.1198/000313002317572835>.
- Cochran W. G. (1952), *The chi-squared Test of Goodness of Fit*, "Annals of Mathematical Statistics", vol. 23, <https://doi.org/10.1214/aoms/1177729380>.
- Cramér H. (1946), *Mathematical Methods of Statistics*, Princeton University Press, Princeton.
- Drost F. C., Kallenberg W. C. M., Moore D. S., Oosterhoff J. (1989), *Power Approximations to Multinomial Tests of Fit*, "Journal of the American Statistical Association", vol. 84, <https://doi.org/10.2307/2289856>.
- Edgeworth F. Y. (1907), *On the Representation of a Statistical Frequency by a Series*, "Journal of the Royal Statistical Society", vol. A 70.
- Fuller W. A. (2009), *Sampling Statistics*, John Wiley & Sons, Hoboken, New Jersey.
- Greselin F., Zenga M. (2006), *Convergence of the Sample Mean Difference to the Normal Distribution: Simulation Results*, "Statistica & Applicazioni", vol. 4, no 1.
- Hájek J. (1964), *Asymptotic Theory of Rejective Sampling with Varying Probabilities from a Finite Population*, "Annals of Mathematical Statistics", vol. 35, <https://doi.org/10.1214/aoms/1177700375>.
- Hájek J. (1981), *Sampling from a Finite Population*, ed. V. Dupač, Marcel Dekker, Inc., New York–Basel.
- Hall P. (1992), *The Bootstrap and Edgeworth Expansion*, Springer-Verlag, New York.
- Hansen M. H., Hurvitz W. N. (1943), *On the Theory of Sampling from Finite Population*, "Annals of Mathematical Statistics", vol. 14, <https://doi.org/10.1214/aoms/1177731356>.
- Horvitz D. G., Thompson D. J. (1952), *A Generalization of Sampling without Replacement from a Finite Universe*, "Journal of the American Statistical Association", vol. 47, <https://doi.org/10.1080/01621459.1952.10483446>.
- Krzyśko M. (2000), *Statystyka matematyczna*, Wydawnictwo Naukowe Uniwersytetu im. Adama Mickiewicza w Poznaniu, Poznań.
- Lahiri D. B. (1951), *A Method of Sample Selection Providing Unbiased Ratio Estimator*, "Bulletin of the International Statistical Institute", vol. 33.
- Midzuno H. (1952), *On the Sampling System with Probability Proportional to Sum of Size*, "Annals of the Institute of Statistical Mathematics", vol. 3, <https://doi.org/10.1007/bf02949779>.
- Ryan T. P. (2013), *Sample Size Determination and Power*, John Wiley & Sons, Hoboken, New Jersey.
- Santer T. J., Duffy D. E. (1989), *The Statistical Analysis of Discrete Data*, Springer-Verlag, New York.
- Seber G. A. F. (2013), *Statistical Models for Proportions and Probabilities*, Springer Briefs in Statistics, Heidelberg–New York–Dordrecht–London.

- Sen A. R. (1953), *On the Estimate of the Variance in Sampling with Varying Probabilities*, "Journal of the Indian Society of Agricultural Statistics", vol. 5.
- Tillé Y. (2006), *Sampling Algorithms*, Springer, New York.
- Wywił J. L. (2016), *Contributions to Testing Statistical Hypotheses in Auditing*, Wydawnictwo Naukowe PWN, Warszawa.
- Yates F., Grundy P. M. (1953), *Selection without Replacement from Within Strata with Probability Proportional to Size*, "Journal of the Royal Statistical Society", Series B, vol. 15.

Symulacyjne wyznaczanie niezbędnego rozmiaru próby zapewniającego wystarczającą zbieżność rozkładu pewnych statystyk do rozkładu normalnego (Streszczenie)

Podczas testowania hipotez lub wyznaczania przedziałów ufności rozkładu pewnych statystyk zwykle nie są znane. Wygodne jest, gdy rozkłady takich statystyk można przybliżyć rozkładem normalnym. Celem pracy jest wyznaczenie takiej liczebności próby, przy której rozkład statystyki jest dostatecznie dobrze aproksymowany rozkładem normalnym. Zaproponowano dwie procedury postępowania. Jedna z nich daje aproksymację liczebności próby na podstawie nierówności Berry-Esseena. Druga metoda polega na generowaniu serii prób o ustalonej liczebności, na podstawie których wyznacza się wartości statystyki. Opierając się na tych wartościach, testuje się normalność rozkładu statystyki. W razie odrzucenia hipotezy o normalności zwiększa się rozmiar generowanych prób. Procedurę tę powtarza się aż do ustalenia liczebności próby, przy której hipoteza o normalności nie jest odrzucona.

Słowa kluczowe: rozmiar próby, twierdzenia centralne, schemat losowania, symulacja komputerowa, test chi-kwadrat zgodności.

Jacek Osiewalski

Jerzy Marzec

Dwuwymiarowe zmienne licznikowe – bayesowskie modelowanie selekcji próby*

Streszczenie

W artykule przedstawiono propozycję łącznego modelu statystycznego dwóch zmiennych licznikowych, z których jedna może być zdegenerowana w zerze. Rozważane jest modelowanie oparte na przełączaniu między dwu- i jednowymiarowym modelem regresji poissonowskiej, przy czym przełączanie zależy od zaobserwowanej wartości trzeciej, dychotomicznej zmiennej. Zalecana jest analiza bayesowska; w dwóch szczególnych przypadkach proponowanego modelu bayesowskiego sformułowano konsekwencje ważne dla wnioskowania. W części empirycznej rozważane jest łączne modelowanie liczby płatności gotówką i kartą w Polsce, z wykorzystaniem danych zarówno dla posiadaczy kart, jak i osób ich nieposiadających.

Słowa kluczowe: dwuwymiarowe modele regresji Poissona, przełączanie między rozkładem niezdegenerowanym i zdegenerowanym, faktoryzacja funkcji wiarygodności, płatności kartą płatniczą i gotówką.

Klasyfikacja JEL: C25, C24, C51.

Jacek Osiewalski, Uniwersytet Ekonomiczny w Krakowie, Wydział Zarządzania, Katedra Ekonometrii i Badań Operacyjnych, ul. Rakowicka 27, 31-510 Kraków, e-mail: eosiewa@cyf-kr.edu.pl

Jerzy Marzec, Uniwersytet Ekonomiczny w Krakowie, Wydział Zarządzania, Katedra Ekonometrii i Badań Operacyjnych, ul. Rakowicka 27, 31-510 Kraków, e-mail: marzecj@uek.krakow.pl

* Artykuł stanowi wynik realizacji projektu sfinansowanego ze środków przyznanych Wydziałowi Zarządzania Uniwersytetu Ekonomicznego w Krakowie w ramach dotacji na utrzymanie potencjału badawczego.

1. Wprowadzenie

Przy łącznym modelowaniu zmiennych licznikowych można spotkać się z sytuacją, gdy jedna z nich jest z konieczności zerem dla wielu obserwowanych obiektów. Na przykład jeśli badamy determinanty i współzależność liczby przejazdów mieszkańców miasta transportem publicznym i własnymi samochodami, to dla osób bez samochodu liczba przejazdów tym środkiem jest stale równa zero. Powstaje pytanie, jakie są możliwości i konsekwencje wnioskowania o determinantach liczby przejazdów transportem publicznym oraz o zależności między oboma liczbami przejazdów na podstawie danych dotyczących wszystkich badanych mieszkańców miasta – wobec badania tych determinant i tej samej zależności na podstawie danych dotyczących tylko mieszkańców miasta posiadających samochód. Wykorzystanie tych ostatnich danych oznacza wstępną selekcję obserwacji i uniemożliwia przenoszenie wyników analizy na wszystkich mieszkańców. Aby wykorzystać cały zbiór obserwacji i umożliwić wyciąganie ogólniejszych wniosków, autorzy zaproponowali model statystyczny uwzględniający przełączanie między dwoma modelami zmiennych licznikowych: modelem dwuwymiarowym i jednowymiarowym; za przełączanie odpowiada dychotomiczny model stosownej zmiennej zero-jedynkowej (reprezentującej w przytaczanym przykładzie posiadanie samochodu). Takie podejście pozwala ująć różne sytuacje jako przypadki szczególne i sformułować kluczową testowalną hipotezę identyczności mechanizmu określającego generowanie (w dwóch grupach obiektów) wartości tej zmiennej licznikowej, która nigdy nie jest zdegenerowana.

Zasadniczą częścią składową omawianego w tej pracy modelu przełącznikowego jest dwuwymiarowy model zmiennych licznikowych, opisujący przypadek, w którym żadna ze zmiennych nie jest skoncentrowana w zerze. Regresja poissonowska jest znanym modelem analizy zmiennych licznikowych. Istnieją jej dwuwymiarowe uogólnienia, lecz większość z nich charakteryzuje się ograniczeniami dotyczącymi znaku współczynnika korelacji między zmiennymi, inne zaś prowadzą do komplikacji natury statystyczno-numerycznej (zob. m.in. [Kocherlakota i Kocherlakota 1992, Winkelman 2008]). Modele, które dopuszczają zarówno korelację dodatnią, jak i ujemną, można uzyskać wykorzystując np. kopule lub mieszkanki rozkładów. Innym podejściem jest warunkowy model Poissona, który zaproponowali P. Berkhout i E. Plug [2004]. Omówienie tych zagadnień, wraz z odwołaniami do literatury, można znaleźć m.in. w artykule [Marzec 2012]. Warto podkreślić, że w kontekście modeli dwuwymiarowych nie pojawia się w literaturze kwestia selekcji próby.

Jako główną część składową proponowanego modelu statystycznego wykorzystano specyfikację ZIP-CP (*zero inflated Poisson – conditional Poisson*), którą zaproponowano w pracy [Marzec i Osiewalski 2012]. Jest to dwuwymia-

rowa regresja typu Poissona, ogólniejsza niż model P-CP (*Poisson – conditional Poisson*), który wprowadzili P. Berkhout i E. Plug [2004]. W modelu P-CP przyjmuje się brzegowy rozkład Poissona dla jednej zmiennej i warunkowy rozkład Poissona dla drugiej (przy ustalonej pierwszej); model ten jest łatwy w estymacji i dopuszcza korelację różnego znaku (dodatnią albo ujemną), ale znak ten zależy od znaku jednego parametru, a nie od zmiennych objaśniających. W modelu ZIP-CP dwuwymiarowej regresji typu Poissona zamiast brzegowego rozkładu Poissona pierwszej z dwóch zmiennych wprowadza się rozkład typu ZIP, w wersji „płatkowej” (*hurdle model*), co prowadzi do znaku kowariancji (między oboma zmiennymi licznikowymi) zależnego od wartości zmiennych objaśniających. Charakterystyki modelu ZIP-CP wynikają z własności dwuwymiarowego skokowego rozkładu ZIP-CP, który wprowadził i zbadał J. Osiewalski [2012]. Druga część proponowanego modelu przełącznikowego to jednowymiarowa regresja Poissona dla drugiej zmiennej – w przypadku gdy pierwsza jest zdegenerowana (skoncentrowana w zerze). Jak już wspomniano, trzecią częścią jest specyfikacja dychotomiczna, opisująca przełączanie między przypadkiem dwuwymiarowym (niezdegenerowanym) i jednowymiarowym (zdegenerowanym).

Następny punkt pracy poświęcony jest prezentacji probabilistycznych podstaw modelu, tj. rozkładów skokowych wykorzystywanych w budowie trzech części składowych tego modelu – w szczególności rozkładu ZIP-CP. W trzecim punkcie omówiono proponowany model statystyczny i postać funkcji wiarygodności oraz przedstawiono analizę bayesowską tego modelu, zwracając uwagę na jego dwa przypadki szczególne. W czwartym, empirycznym punkcie pracy zaprezentowano nowe wyniki, uzyskiwane na podstawie pełnego zbioru danych, w łącznym badaniu liczb transakcji dokonywanych kartą bankową i gotówką (zob. [Polasik, Marzec, Fiszedler i Górka 2012] oraz [Marzec i Osiewalski 2012]). Przykład ten ilustruje problemy modelowania i wnioskowania w sytuacji zmiennych licznikowych, z których jedna (liczba płatności kartą) jest zdegenerowana dla wielu badanych jednostek (osób nieposiadających kart). W piątym punkcie zawarto uwagi końcowe.

Proponowany w tej pracy przykład empiryczny wpisuje się w badania rozwoju obrotu bezgotówkowego w Polsce, które są prowadzone od kilku lat (zob. np. [Polasik i Maciejewski 2009, Fiszedler i Polasik 2009, Polasik 2015, Polasik, Wisniewski i Lightfoot 2012, Górka 2013, Goczek i Witkowski 2015, 2016]). Z punktu widzenia banku centralnego interesującą kwestią jest określenie czynników motywujących do korzystania z kart płatniczych i identyfikacja tych barier utrudniających działalność przedsiębiorstw handlowych, które są związane z dodatkowymi opłatami *interchange* za transakcje dokonane przy użyciu kart. Dla gospodarki i finansów państwa wymierne korzyści rodzi ograniczenie transakcji gotówkowych między klientem detalicznym a sprzedawcą na rzecz

transakcji dokonywanych kartą, co częściowo przyczyniłoby się do zmniejszenia szarej strefy. Ważnym elementem badań wzbogacających obecny stan wiedzy na temat płatności kartą i gotówką są propozycje nowych modeli ekonometrycznych, opisujących złożone decyzje podejmowane przez konsumentów.

2. Probabilistyczne podstawy nowego modelu statystycznego

Rozważamy łączny rozkład prawdopodobieństwa trzech zmiennych losowych (Y_1, Y_2, Y_3) , z których trzecia ma rozkład dwupunktowy (jest zmienną zero-jedynkową), druga może przyjąć dowolną wartość całkowitą nieujemną, a pierwsza ma rozkład jednopunktowy, gdy $Y_3 = 0$ ($\Pr\{Y_1 = 0 \mid Y_3 = 0\} = 1$), może zaś przyjąć dowolną wartość całkowitą nieujemną, gdy $Y_3 = 1$. Zatem przy $Y_3 = 0$ rozkład (warunkowy) pary (Y_1, Y_2) jest tożsamy z rozkładem pary $(0, Y_2)$, czyli odpowiada rozkładowi pojedynczej zmiennej Y_2 . Jedynie przy $Y_3 = 1$ rozkład pary (Y_1, Y_2) jest dwuwymiarowym rozkładem na zbiorze wszystkich par liczb całkowitych nieujemnych. Temu ostatniemu poświęcamy specjalną uwagę, rozważając przypadek prostszy: P-CP (zob. [Berkhout i Plug 2004]) i ogólniejszy: ZIP-CP (zob. [Osiewalski 2012]).

Przy $Y_3 = 1$ rozkład prawdopodobieństwa pary (Y_1, Y_2) jest następujący:

$$\Pr\{Y_1 = i, Y_2 = j \mid Y_3 = 1\} = \Pr\{Y_1 = i \mid Y_3 = 1\} \Pr\{Y_2 = j \mid Y_3 = 1, Y_1 = i\} = g(i)h(j, i), \quad (1)$$

przy czym $i, j \in N \cup \{0\}$. Jeśli rozkład zmiennej Y_1 jest rozkładem Poissona o wartości oczekiwanej i wariancji λ_1 , a rozkład warunkowy Y_2 przy ustalonej wartości zmiennej Y_1 jest rozkładem Poissona o wartości oczekiwanej i wariancji $\lambda_2 \exp(\alpha Y_1)$, czyli

$$g(i) = e^{-\lambda_1} (\lambda_1)^i / i!, \quad h(j, i) = \exp(-\lambda_2 e^{\alpha i}) (\lambda_2)^j e^{\alpha i j} / j!, \quad (2)$$

to mamy rozkład dwuwymiarowy P-CP o momentach postaci [Berkhout i Plug 2004]:

$$E(Y_2 \mid Y_3 = 1) = \lambda_2 \exp[\lambda_1 (e^\alpha - 1)], \quad (3)$$

$$\text{Var}(Y_2 \mid Y_3 = 1) = E(Y_2 \mid Y_3 = 1) + [E(Y_2 \mid Y_3 = 1)]^2 \{\exp[\lambda_1 (e^\alpha - 1)^2] - 1\}, \quad (4)$$

$$\text{Cov}(Y_1, Y_2 \mid Y_3 = 1) = \lambda_1 (e^\alpha - 1) E(Y_2 \mid Y_3 = 1). \quad (5)$$

Jeśli $\alpha \neq 0$, to wariancja (4) zmiennej Y_2 jest większa od wartości oczekiwanej (3). Zależność między obu zmiennymi sprawia, że rozkład zmiennej Y_2 odpowiada empirycznie częściej sytuacji zwiększonej wariancji danych licznikowych. Rozkład zmiennej Y_1 , czyli rozkład Poissona, nie ma tej właściwości. Jest to pierwszy powód uogólnienia dwuwymiarowego rozkładu P-CP przez wprowa-

dzenie rozkładu typu ZIP na miejsce brzegowego rozkładu Poissona. Modele regresji dla skokowej zmiennej objaśnianej z nadmierną liczbą zer spopularyzował głównie D. Lambert [1992], a A.C. Cameron i P.K. Trivedi [1998, 2005] oraz R. Winkelman [2008] przedstawiają stosowne modele danych licznikowych z przykładami ich zastosowań w ekonomii.

Należy zauważyć, że znak kowariancji między Y_1 i Y_2 , czyli znak wyrażenia (5), zależy jedynie od znaku stałej α , a nie od wielkości λ_1, λ_2 , parametryzowanych głębiej (uzależnianych od zmiennych objaśniających) w statystycznych zastosowaniach tego modelu probabilistycznego. Uogólnienie, które zaproponował J. Osiewalski [2012], dopuszcza związek znaku kowariancji i wielkości λ_1 . Ta ogólniejsza klasa rozkładów (oznaczana gwiazdką) jest określona przez ten sam warunkowy rozkład Y_2 przy ustalonym Y_1 :

$$\Pr^* \{Y_2 = j \mid Y_3 = 1, Y_1 = i\} = h(j, i) = \Pr \{Y_2 = j \mid Y_3 = 1, Y_1 = i\} \quad (6)$$

oraz przez rozkład zmiennej Y_1 , który odmiennie niż w (1) traktuje wartość 0:

$$\Pr^* \{Y_1 = i \mid Y_3 = 1\} = g^*(i) = \begin{cases} \gamma & \text{dla } i = 0, \\ \frac{1-\gamma}{1-g(0)} g(i) & \text{dla } i \in N, \end{cases} \quad (7)$$

gdzie γ jest ustaloną liczbą z przedziału $(0, 1)$, funkcje g i h są zaś takie same jak w (1). Jeśli $\gamma = g(0)$, to $\Pr^* \{Y_1 = i \mid Y_3 = 1\} = g^*(i) = g(i) = \Pr \{Y_1 = i \mid Y_3 = 1\}$ i mamy przypadek (1). Jeśli $\gamma \neq g(0)$, a funkcje g i h zadane są nadal wzorami (2), to rozkład zmiennej Y_1 jest typu ZIP, zaś warunkowy dla Y_2 przy ustalonym Y_1 pozostaje rozkładem Poissona. Rozkład łączny to ZIP-CP, a jego momenty mają ogólną postać:

$$E^*(Y_1^m Y_2^n \mid Y_3 = 1) = \frac{(1-\gamma)E(Y_1^m Y_2^n \mid Y_3 = 1) + (\gamma - g(0))0^m E(Y_2^n \mid Y_3 = 1, Y_1 = 0)}{1 - g(0)}, \quad (8)$$

gdzie wykorzystuje się znaną postać momentów rozkładu P-CP (dla $m = 0$ przyjmując $0^m = 1$). W szczególności otrzymujemy:

$$E^*(Y_1 \mid Y_3 = 1) = (1 - g(0))^{-1} (1 - \gamma) \lambda_1, \quad (9)$$

$$E^*(Y_2 \mid Y_3 = 1) = (1 - g(0))^{-1} \left[(1 - \gamma) E(Y_2 \mid Y_3 = 1) + (\gamma - g(0)) \lambda_2 \right], \quad (10)$$

$$\text{Var}^*(Y_1 \mid Y_3 = 1) = \frac{1-\gamma}{1-g(0)} \lambda_1 \left(1 + \frac{\gamma - g(0)}{1-g(0)} \lambda_1 \right), \quad (11)$$

$$\begin{aligned} & \text{Var}^*(Y_2 \mid Y_3 = 1) = \\ & = \frac{1-\gamma}{1-g(0)} \left\{ \text{Var}(Y_2 \mid Y_3 = 1) + \frac{\gamma - g(0)}{1-g(0)} \left[E(Y_2 \mid Y_3 = 1) - \lambda_2 \right]^2 + \frac{\gamma - g(0)}{1-\gamma} \lambda_2 \right\}, \end{aligned} \quad (12)$$

$$\begin{aligned} \text{Cov}^*(Y_1, Y_2 | Y_3 = 1) &= \\ &= \frac{\lambda_1 \lambda_2 (1 - \gamma)}{(1 - \exp(-\lambda_1))^2} \left\{ \left[(1 - e^{-\lambda_1}) e^\alpha - (1 - \gamma) \right] \exp(\lambda_1 (e^\alpha - 1)) - \gamma + e^{-\lambda_1} \right\}, \end{aligned} \quad (13)$$

gdzie $E(Y_2 | Y_3 = 1)$ i $\text{Var}(Y_2 | Y_3 = 1)$ są momentami rozkładu P-CP danymi w (3) i (4). Widzimy, że zmienne tworzące parę (Y_1, Y_2) o rozkładzie prawdopodobieństwa ZIP-CP:

- 1) są skorelowane ujemnie, jeśli $\left[(1 - e^{-\lambda_1}) e^\alpha - (1 - \gamma) \right] \exp(\lambda_1 (e^\alpha - 1)) < \gamma - e^{-\lambda_1}$,
- 2) są skorelowane dodatnio, jeśli $\left[(1 - e^{-\lambda_1}) e^\alpha - (1 - \gamma) \right] \exp(\lambda_1 (e^\alpha - 1)) > \gamma - e^{-\lambda_1}$,
- 3) są nieskorelowane, jeśli $\left[(1 - e^{-\lambda_1}) e^\alpha - (1 - \gamma) \right] \exp(\lambda_1 (e^\alpha - 1)) = \gamma - e^{-\lambda_1}$.

W przypadku $\gamma = g(0) = e^{-\lambda_1}$, tj. rozkładu Poissona dla Y_1 (przy $Y_3 = 1$), złożona formuła kowariancji (13) sprowadza się do znacznie prostszej postaci (5), gdzie znak kowariancji zależy jedynie od znaku stałej α . W pozostałych przypadkach, tj. gdy rozkład Y_1 jest typu ZIP, znak kowariancji (13) zależy od wartości przyjmowanych przez λ_1 i α (a nie tylko od znaku tej drugiej stałej). Oczywiście, konkretna wartość kowariancji w rozkładzie ZIP-CP (a nie sam jej znak) oraz – w konsekwencji – wartość współczynnika korelacji zależą od wszystkich stałych występujących w funkcji prawdopodobieństwa tego rozkładu, tj. od $\gamma, \lambda_1, \lambda_2$ i α .

Zauważmy też, że zwiększenie prawdopodobieństwa zerowej wartości Y_1 (w stosunku do rozkładu Poissona o wartości oczekiwanej i wariancji λ_1), czyli przyjęcie rozkładu ZIP z $\gamma > g(0)$, prowadzi do wariancji (11) większej niż wartość oczekiwana (9). Rozkład ZIP-CP umożliwia modelowanie zwiększonej wariancji obu obserwowanych zmiennych licznikowych, chociaż nie są one traktowane symetrycznie.

Powyższe rozważania dotyczyły jedynie rozkładu warunkowego pary (Y_1, Y_2) przy $Y_3 = 1$, czyli bardziej złożonej części specyfikacji trójwymiarowej. Rozkład zmiennej Y_2 przy $Y_3 = 0$ – i tym samym przy jedynej wartości Y_1 (równej 0) – przyjmujemy tak, aby można było badać identyczność rozkładu warunkowego zmiennej Y_2 przy $Y_1 = 0$ w obu sytuacjach: $Y_3 = 0$ i $Y_3 = 1$. Zakładamy zatem, że jest to rozkład Poissona o funkcji prawdopodobieństwa:

$$\Pr\{Y_2 = j | Y_3 = 0, Y_1 = 0\} = h_0(j) = \exp(-\lambda_{2,0}) (\lambda_{2,0})^j / j!, \quad (14)$$

z parametrem $\lambda_{2,0}$ niekoniecznie równym λ_2 .

Podsumowując dotychczas przyjęte założenia, wprowadzamy następujący łączny rozkład trzech zmiennych skokowych:

$$\Pr\{Y_1 = i, Y_2 = j, Y_3 = l\} = \begin{cases} pg^*(i)h(j, i), & i, j \in N \cup \{0\}, l = 1, \\ (1 - p)h_0(j), & i = 0, j \in N \cup \{0\}, l = 0, \\ 0, & i \in N, j \in N \cup \{0\}, l = 0, \end{cases} \quad (15)$$

gdzie $p = \Pr\{Y_3 = 1\}$. Brzegowy rozkład pary (Y_1, Y_2) jest swoistą mieszanką dwuwymiarowego rozkładu ZIP-CP i jednowymiarowego rozkładu Poissona:

$$\Pr\{Y_1 = i, Y_2 = j\} = pg^*(i)h(j, i) + (1-p)I_{(0)}(i)h_0(j), \quad i, j \in N \cup \{0\}, \quad (16)$$

gdzie $I_A(\cdot)$ oznacza funkcję charakterystyczną zbioru A ; jego momenty można zapisać jako:

$$E(Y_1^m Y_2^n) = pE^*(Y_1^m Y_2^n | Y_3 = 1) + (1-p)0^m E(Y_2^n | Y_3 = 0, Y_1 = 0), \quad (17)$$

przy czym $E^*(Y_1^m Y_2^n | Y_3 = 1)$ to moment zwykły lub mieszany rzędu (m, n) w rozkładzie ZIP-CP, dany ogólnym wzorem (8), zaś $E(Y_2^n | Y_3 = 0, Y_1 = 0)$ to moment zwykły rzędu n w rozkładzie Poissona z parametrem $\lambda_{2,0}$.

3. Model statystyczny

Rozważamy T trójwymiarowych zmiennych losowych $(Y_{1t}, Y_{2t}, Y_{3t}; t = 1, 2, \dots, T)$, gdzie Y_{3t} są zmiennymi zero-jedynkowymi. Przy $Y_{3t} = 1$, pary (Y_{1t}, Y_{2t}) mają różne rozkłady typu ZIP-CP:

$$\Pr^*\{Y_{1t} = i, Y_{2t} = j | Y_{3t} = 1\} = g_t^*(i)h_t(j, i), \quad i, j \in N \cup \{0\}, \quad (18)$$

gdzie

$$\Pr^*\{Y_{1t} = i | Y_{3t} = 1\} = g_t^*(i) = \begin{cases} \gamma_t & \text{dla } i = 0, \\ \frac{1 - \gamma_t}{1 - g_t(0)} g_t(i) & \text{dla } i \in N; \end{cases} \quad g_t(i) = e^{-\lambda_{1t}} (\lambda_{1t})^i / i!, \quad (19)$$

$$\Pr^*\{Y_{2t} = j | Y_{3t} = 1, Y_{1t} = i\} = h_t(j, i) = \exp[-\lambda_{2t} \exp(\alpha i)] (\lambda_{2t})^j \exp(\alpha ij) / j!, \quad (20)$$

$$\lambda_{1t} = \exp(x_t \beta_1), \quad \lambda_{2t} = \exp(w_t \beta_2), \quad \gamma_t = \exp(-e^\delta \lambda_{1t}) = \exp(-\exp(\delta + x_t \beta_1)); \quad (21)$$

x_t i w_t są wierszami wartości zmiennych objaśniających, które mogą się pokrywać (w części lub w całości). Zmienne te określają prawdopodobieństwa pojawienia się poszczególnych par wartości Y_{1t} i Y_{2t} ; wpływ x_t i w_t na te prawdopodobieństwa jest determinowany wielkością poszczególnych składowych kolumn β_1 i β_2 , wielkością parametru zależności α oraz wielkością parametru δ , który decyduje o odchyleniu prawdopodobieństwa, że $Y_{1t} = 0$, od wartości wynikającej z rozkładu Poissona. Zauważmy, że momenty rozkładu pary (Y_{1t}, Y_{2t}) , podane w poprzednim punkcie pracy, zależą teraz od zmiennych objaśniających.

W literaturze specyfikacja oparta na wzorze (19) jest nazywana modelem płótkowym – zob. [Cameron i Trivedi 2005, s. 680]. Porównanie tej specyfikacji z oryginalnym modelem ZIP przedstawia R. Winkelmann [2008]. Głównymi zaletami przedstawionej w niniejszym artykule propozycji są prostota parametryzacji

i względna łatwość estymacji oraz prostota testowania zasadności redukcji specyfikacji (19) do standardowego modelu Poissona.

Przy $Y_{3t} = 0$, pary $(Y_{1t}, Y_{2t}) = (0, Y_{2t})$ mają rozkłady zdegenerowane (bo zmienne Y_{1t} mają rozkład jednopunktowy), zaś jako warunkowe rozkłady Y_{2t} przyjmujemy różne rozkłady Poissona – przez analogię do (20):

$$\Pr\{Y_{2t} = j \mid Y_{3t} = 0, Y_{1t} = 0\} = h_{0,t}(j) = \exp[-\lambda_{2t,0}](\lambda_{2t,0})^j / j!, \quad \lambda_{2t,0} = \exp(w_t \beta_{2,0}). \quad (22)$$

Jeśli $\beta_2 = \beta_{2,0}$, to $\Pr^*\{Y_{2t} = j \mid Y_{3t} = 1, Y_{1t} = 0\} = \Pr\{Y_{2t} = j \mid Y_{3t} = 0, Y_{1t} = 0\}$, czyli sposób generowania wartości zmiennej Y_{2t} przy $Y_{1t} = 0$ jest identyczny bez względu na wartość zero-jedynkowej zmiennej Y_{3t} . Do weryfikacji hipotezy, że $\beta_2 = \beta_{2,0}$, potrzebny jest trójwymiarowy model statystyczny, tj. parametryczna klasa rozkładów postaci:

$$\Pr\{Y_{1t} = i, Y_{2t} = j, Y_{3t} = l; \theta\} = \begin{cases} p_t g_t^*(i) h_t(j, i), & i, j \in N \cup \{0\}, l = 1, \\ (1 - p_t) h_{0,t}(j), & i = 0, j \in N \cup \{0\}, l = 0, \\ 0, & i \in N, j \in N \cup \{0\}, l = 0, \end{cases} \quad (23)$$

gdzie $p_t = \Pr\{Y_{3t} = 1\} = 1 - F(-z_t \beta_3)$, z_t jest wektorem zmiennych objaśniających, zaś F jest dystrybuantą reprezentującą dychotomiczny model zmiennej Y_{3t} . W badaniu empirycznym przyjmujemy model logitowy, czyli zakładamy dystrybuantę rozkładu logistycznego. Warto byłoby w przyszłości rozważyć modele dychotomiczne oparte na dystrybuancie skośnego rozkładu t Studenta, które J. Osiewalski i J. Marzec [2004a, 2004b] wprowadzili jako alternatywę dla dwóch podstawowych specyfikacji: logitowej i probitowej. W modelu statystycznym trójki zmiennych (Y_{1t}, Y_{2t}, Y_{3t}) wektor parametrów θ jest kolumną grupującą $\delta, \alpha, \beta_1, \beta_2, \beta_3$ i $\beta_{2,0}$. Zakładamy, że przy dowolnie ustalonym θ trójwymiarowe obserwacje są niezależne.

Jeśli zaobserwowano $Y_{1t} = y_{1t}$, $Y_{2t} = y_{2t}$ i $Y_{3t} = y_{3t}$ ($t = 1, 2, \dots, T$), to odpowiadająca tym wartościom funkcja wiarygodności ma postać:

$$L(\theta; y) = \left[\prod_{t: y_{3t}=1, y_{1t}=0} \gamma_t h_t(y_{2t}, 0) \right] \left[\prod_{t: y_{3t}=1, y_{1t}>0} \frac{1 - \gamma_t}{1 - g_t(0)} g_t(y_{1t}) h_t(y_{2t}, y_{1t}) \right] \left[\prod_{t: y_{3t}=1, y_{1t}=0} h_{0,t}(y_{2t}) \right] \left[\prod_{t: y_{3t}=1} p_t \right] \left[\prod_{t: y_{3t}=0} (1 - p_t) \right] = L_1(\beta_1, \beta_2, \alpha, \delta) L_2(\beta_{2,0}) L_3(\beta_3), \quad (24)$$

gdzie y oznacza macierz $(3 \times T)$ zawierającą zaobserwowane wartości zmiennych Y_{1t} , Y_{2t} i Y_{3t} . Dwa pierwsze czynniki funkcji wiarygodności L odpowiadają dwuwymiarowej składowej mieszanki i tworzą funkcję L_1 parametrów $\delta, \alpha, \beta_1, \beta_2$; kolejny czynnik odpowiada składowej jednowymiarowej i stanowi funkcję L_2 parametru $\beta_{2,0}$; ostatnie dwa czynniki odpowiadają dychotomicznej zmiennej przełącznikowej i tworzą funkcję L_3 parametru β_3 . Jeśli brak jest związków między trzema wyróżnionymi grupami parametrów, czyli charakteryzują się one swobodą

zmienności (*variation freeness*), to wnioskowanie o każdej z nich prowadzi się odrębnie, na podstawie odpowiedniego czynnika L_r ($r = 1, 2, 3$), a nie na podstawie pełnej funkcji wiarygodności. Zauważmy, że faktoryzacja funkcji wiarygodności nie wynika z konkretnej postaci rozkładów, które przyjmujemy, lecz z samej struktury modelu statystycznego i ze swobody zmienności poszczególnych grup parametrów. Na potrzeby wnioskowania brak związków między parametrami (lub ich obecność) formalizuje się ściśle na gruncie bayesowskim, gdzie wprowadza się rozkład *a priori* (miarę probabilistyczną) na przestrzeni parametrów i można wtedy rozważać stochastyczną niezależność grup (wektorów) parametrów. W tej pracy skupiamy się na dwóch przypadkach: niezależności *a priori* trzech wyróżnionych grup parametrów i równości $\beta_2 = \beta_{2,0}$.

Przy separowalności funkcji wiarygodności niezależność *a priori* ($\delta, \alpha, \beta_1, \beta_2$), $\beta_{2,0}$ i β_3 prowadzi do niezależności *a posteriori* tych parametrów (czyli niezależności warunkowej przy znanych obserwacjach). Oznacza to całkowitą odrębność wnioskowania o każdej grupie parametrów oraz zasadność wykorzystania tylko danych z $y_{3t} = 1$ w estymacji ($\delta, \alpha, \beta_1, \beta_2$) i tylko danych z $y_{3t} = 0$ w estymacji $\beta_{2,0}$. Oczywiście, wnioskowanie o funkcjach wektora θ wszystkich parametrów – takich jak np. $\text{Corr}(Y_{1t}, Y_{2t} | \theta)$, tj. bezwarunkowy współczynnik korelacji między pierwszymi dwoma elementami trójki (Y_{1t}, Y_{2t}, Y_{3t}) – musi być oparte na rozkładzie *a posteriori* całego wektora θ , wykorzystującym pełną funkcję wiarygodności i kompletne dane. Empirycznie interesujące może być porównanie dwóch współczynników korelacji – bezwarunkowego $\text{Corr}(Y_{1t}, Y_{2t} | \theta)$ i warunkowego, określanego tylko na podstawie modelu ZIP-CP, tj. $\text{Corr}(Y_{1t}, Y_{2t} | Y_{3t} = 1, \theta) = \text{Corr}^*(Y_{1t}, Y_{2t} | Y_{3t} = 1, \delta, \alpha, \beta_1, \beta_2)$.

W przypadku gdy zakładamy $\beta_2 = \beta_{2,0}$, czyli niezależny od Y_{3t} mechanizm generowania Y_{2t} , czynników L_1 i L_2 funkcji wiarygodności nie da się rozważać odrębnie, gdyż oba zależą od tego samego wektora β_2 . Wnioskowanie o samych parametrach (a nie tylko o takich ich funkcjach, jak bezwarunkowy współczynnik korelacji) musi wykorzystywać pełną funkcję wiarygodności, opartą na wszystkich obserwacjach. Ograniczenie się we wnioskowaniu jedynie do danych z $y_{3t} = 1$ oznaczałoby „błąd selekcji próby”. Testowanie hipotezy $\beta_2 = \beta_{2,0}$ możliwe jest oczywiście tylko w przypadku nienarzucającym takiej restrykcyj.

Pełna specyfikacja bayesowskiego modelu statystycznego z rozkładem próbkowym postaci (23), prowadzącym do funkcji wiarygodności (24), wymaga przyjęcia konkretnego rozkładu *a priori* wektora θ . Proponujemy założyć niezależność *a priori* parametrów i dla każdego indywidualnie przyjąć standardowy rozkład normalny $N(0, 1)$. Zerowe wartości oczekiwane *a priori* oznaczają, że największą szansę dajemy wstępnie modelowi bez zmiennych objaśniających. Jednostkowe odchylenia standardowe *a priori* dają gwarancję, że specyfikacje odległe od tej

najprostszej mają bardzo istotne wstępne szanse. Wydaje się, że taki prosty łączny rozkład *a priori* niesie słabą wiedzę wstępną, gwarantując zarazem łatwość symulacji Monte Carlo z rozkładu *a posteriori*, ale jego konkretna rola informacyjna (w stosunku do funkcji wiarygodności) oraz wrażliwość rozkładu *a posteriori* są kwestiami empirycznymi, które należy badać odrębnie dla każdego analizowanego zestawu danych.

4. Łączne modelowanie liczby płatności kartą i gotówką

Aby zilustrować przydatność empiryczną zaproponowanego modelu statystycznego, w szczególności analizy konsekwencji selekcji próby, wykorzystamy dane, które zgromadzono w celu prowadzenia badania opisanego w pracach [Polasik, Marzec, Fiszeder i Górka 2012] oraz [Marzec, Polasik i Fiszeder 2013]¹. Dane te zawierają m.in. informacje o posiadaniu karty płatniczej (y_{3t}) oraz o liczbie płatności gotówką (y_{2t}) i kartą (y_{1t}) dokonanych (w miesiącu) przez $T = 2518$ osób, które były ankietowane w jednym spośród trzech miesięcy: w październiku, listopadzie 2010 r. albo w styczniu 2011 r. Osoby nieposiadające karty stanowiły 52,7% wszystkich badanych.

Z analizy dwuwymiarowego rozkładu empirycznego dla liczby płatności gotówką i kartą warunkowego względem $y_{3t} = 1$ (w tym jego rozkładów brzegowych) wynika, że w sytuacji posiadania karty płatniczej średnia liczba płatności gotówką wynosi 20,5 (odchylenie standardowe jest równe 17,3), a średnia liczba płatności kartą wynosi 5 (przy odchyleniu standardowym 6,7). Współczynnik warunkowej korelacji empirycznej między y_{1t} a y_{2t} (przy $y_{3t} = 1$) kształtuje się na poziomie 0,008, co można interpretować jako brak nawet przybliżonej zależności liniowej między liczbą płatności kartą i gotówką. Jednowymiarowe rozkłady empiryczne przy $y_{3t} = 1$ sugerują potrzebę zastosowania modelu, w którym obie zmienne skokowe charakteryzują się rozkładem z nadwyżką zer (zob. [Marzec i Osiewalski 2012]).

Szeregi rozdzielcze liczby płatności gotówką przy $y_{3t} = 1$ i $y_{3t} = 0$ podano w tabeli 1. W przypadku braku karty płatniczej średnia liczba transakcji gotówką wynosi 22,5 ($\pm 19,8$) i jest wyższa niż w przypadku posiadania karty. Także mediana empirycznego rozkładu y_{2t} przy $y_{3t} = 0$ jest przesunięta na prawo w stosunku do mediany rozkładu y_{2t} przy $y_{3t} = 1$. Wyniki zmodyfikowanego

¹ Badanie to, obejmujące m.in. zebranie materiału statystycznego przez TNS Pentor, zostało sfinansowane przez Narodowy Bank Polski.

testu W^2 Andersona i Darlinga [1954]² wskazują na silne niepodobieństwo tych rozkładów.

Tabela 1. Empiryczne rozkłady liczby płatności gotówką y_{2t} , warunkowe względem posiadania ($y_{3t} = 1$) lub nieposiadania ($y_{3t} = 0$) karty płatniczej

Liczba płatności	Częstość ($y_{3t} = 1$)	Struktura (w %)	Częstość ($y_{3t} = 0$)	Struktura (w %)
0	24	2	0	0
(0; 5]	126	11	60	5
(5; 10]	248	21	275	21
(10; 15]	196	16	224	17
(15; 20]	148	12	208	16
(20; 25]	108	9	151	11
(25; 30]	85	7	123	9
(30; 35]	66	6	73	5
(35; 40]	55	5	57	4
(40; 45]	32	3	55	4
(45; 50]	32	3	26	2
> 50	70	6	76	6
Łącznie	1190	100	1328	100
Średnia	20,5	–	22,5	–
Mediana	16	–	18	–

Źródło: opracowanie własne.

Wyniki uzyskane w modelu P-CP na podstawie danych obejmujących 1190 posiadaczy kart wskazywały na niewielką dodatnią korelację między liczbą płatności gotówką i kartą. J. Marzec i J. Osiewalski [2012] potwierdzili to, stosując model ZIP-CP, ale jednocześnie pokazali, że jego redukcja do P-CP nie jest zasadna (zob. też wyniki dla parametrów α i δ prezentowane w tabeli 4). Korzystając z formalnego bayesowskiego porównywania modeli (poprzez czynnik Bayesa) ustalono ponadto, że w modelu ZIP-CP zmienna Y_{1t} musi wyrażać liczbę transakcji kartą, a Y_{2t} liczbę transakcji gotówką (nie na odwrót). Konieczność

² Test w wersji dla zmiennych skokowych, dany formułą:

$$W^2 = N \sum_{i=1}^N \frac{(F^{em}(a_{1,i}) - F^{em}(a_{0,i}))^2}{F^{em}(a_{0,i})(1 - F^{em}(a_{0,i}))} p^{em}(a_{0,i}),$$

zastosowano dla dwóch szeregów składających się z $N = 363$ obserwacji, gdzie F^{em} to dystrybuanta empiryczna, p^{em} to częstość, $a_{0,i}$ i $a_{1,i}$ są realizacjami zmiennych $Y_2 | Y_3 = 0$ i $Y_2 | Y_3 = 1$. Wartość statystyki W^2 wyniosła 10,9, a kwantyl rzędu 0,9999 rozkładu statystyki przy H_0 wynosi nie więcej niż 6,17 (dla 100 obserwacji).

identyfikacji właściwej kolejności zmiennych wynika z niesymetrycznej struktury modelu dwuwymiarowego.

Obecnie przedstawione zostaną wyniki uzyskane na podstawie pełnego zbioru danych, uwzględniającego osoby bez kart płatniczych. Podobnie jak w pracy [Marzec i Osiewalski 2012] wykorzystano dane surowe, bez wag określających stopień reprezentatywności poszczególnych obserwacji³. Rozważamy model statystyczny omówiony w poprzedniej części pracy, na który składają się odrębne modele zmiennych licznikowych dla $T_1 = 1190$ par (Y_{1t}, Y_{2t}) przy $Y_{3t} = 1$ i dla $T_2 = 1328$ zmiennych Y_{2t} przy $Y_{3t} = 0$ oraz łączący je model zmiennej dychotomicznej Y_{3t} (dla $T = T_1 + T_2 = 2518$). Jako zmienne objaśniające wykorzystano główne cechy ankietowanych konsumentów. Przyjęto, że w każdym z trzech modeli składowych występuje ten sam zestaw (potencjalnych) zmiennych objaśniających.

W tabeli 2 przedstawiono zmienne objaśniające i ich typowe wartości, tj. średnie w przypadku zmiennych ciągłych i najczęstsze dla zmiennych dychotomicznych. Warto zwrócić uwagę, że – na prezentowanym poziomie agregacji informacji z badania ankietowego – głównymi czynnikami określającymi posiadanie karty wydają się: deklarowany dochód w rodzinie, wykształcenie, stan cywilny i dostęp do Internetu. Więcej wniosków uzyskamy, analizując wyniki podane w tabeli 3. W całym zbiorze badanych kobiety stanowiły 56% ankietowanych, odsetek osób będących w formalnym związku wyniósł 56%, a 61% ankietowanych posiadało dostęp do Internetu. Czynnikiem wyjaśniającym różnicowanie liczby transakcji kartą bądź gotówką może być miejsce zamieszkania. W miastach mieszkało 63% wszystkich badanych. Wśród posiadaczy kart płatniczych 71% było mieszkańcami miast. Odsetek osób bez kart, a mieszkających w miastach, był niższy i wyniósł 56%. Bezwarunkowa częstość posiadania karty była równa 47,3%, jednak częstość posiadania karty przez klienta pod warunkiem, że mieszka w mieście, wyniosła 53%. Natomiast udział posiadaczy karty wyniósł: 49% wśród mężczyzn, 55% wśród zamężnych albo żonatych oraz 58% wśród osób posiadających dostęp do Internetu.

Uzyskany przy założeniu niezależności *a priori* rozkład *a posteriori* parametrów trójwymiarowego modelu statystycznego, danego wzorem (23), próbkowano stosując metody MCMC (Monte Carlo typu łańcuchów Markowa); zastosowano sekwencyjną wersję algorytmu Metropolisa i Hastingsa. Wykorzystując niezależność *a posteriori*, wynikającą z separowalności funkcji wiarygodności (24) i niezależności *a priori* poszczególnych wektorów parametrów, dokonano osobno estymacji parametrów każdego z trzech modeli składowych, tj. β_1, β_2, α i δ w modelu ZIP-CP (M_1), $\beta_{2,0}$ w modelu Poissona dla liczby transakcji gotówką

³ W badaniach opisanych w pracach [Polasik, Marzec, Fiszedler i Górka 2012] oraz [Marzec, Polasik i Fiszedler 2013] użyto danych ważonych.

w przypadku braku karty (M_2), oraz β_3 w modelu logitowym posiadania karty (M_3). Łączna liczba parametrów wyniosła 34.

Tabela 2. Przeciętne (średnie lub najczęstsze) wartości zmiennych objaśniających

Zmienna objaśniająca	$T = 2518$	$T_1 = 1190$	$T_2 = 1328$
	łącznie	posiadający kartę	bez karty
Płeć (1 – mężczyzna, 0 – kobieta)	0	0	0
Wiek (w latach)	41,2	40,1	42,2
Stan cywilny (1 – żonaty lub zamężna, 0 – nie)	1	1	0
Miejsce zamieszkania (1 – miasto, 0 – wieś)	1	1	1
Miesięczny dochód w rodzinie (w tys. zł)	2,9	3,3	2,5
Wykształcenie (lata nauki)	12,3	13,2	11,5
Dostęp do Internetu (1 – tak, 0 – nie)	1	1	0

Źródło: opracowanie własne.

Tabela 3. Udział wartości 1 w przypadku dychotomicznych zmiennych objaśniających (w %)

Zmienna objaśniająca	$T = 2518$	$T_1 = 1190$	$T_2 = 1328$
	łącznie	posiadający kartę	bez karty
Płeć (1 – mężczyzna, 0 – kobieta)	44	45	42
Stan cywilny (1 – żonaty lub zamężna, 0 – nie)	56	65	48
Miejsce zamieszkania (1 – miasto, 0 – wieś)	63	71	56
Dostęp do Internetu (1 – tak, 0 – nie)	61	76	49

Źródło: opracowanie własne.

W tabeli 4 zaprezentowano wartości oczekiwane i odchylenia standardowe *a posteriori* parametrów. Postulowany w modelu M_1 wpływ wszystkich zmiennych objaśniających na liczbę płatności gotówką, gdy konsument korzysta równocześnie z karty, został potwierdzony przez dane. Natomiast tylko posiadanie przez konsumenta dostępu do Internetu, jego wykształcenie i dochód powodują znaczące zróżnicowanie liczby płatności kartą. W czystym modelu Poissona (M_2) płeć konsumenta i jego dochód wydają się nie mieć wpływu na zróżnicowanie liczby transakcji gotówką w sytuacji braku karty. W modelu logitowym (M_3) determinantami posiadania karty płatniczej okazują się wszystkie zmienne objaśniające z wyjątkiem wieku, który wyraźnie nie ma znaczenia.

Tabela 4. Wartości oczekiwane i odchylenia standardowe *a posteriori* parametrów (β) modeli

Zmienna/parametr	Model	$E(\beta y)$	$D(\beta y)$	Model	$E(\beta y)$	$D(\beta y)$
„I”	M_1 : płatności kartą	0,911	0,098	M_3 : model logitowy posiadania karty	-5,455	0,330
Płeć		-0,044	0,025		0,181	0,092
Wiek		-0,002	0,001		0,001	0,003
Stan cywilny		-0,048	0,029		0,596	0,102
Miejsce zamieszkania		-0,007	0,028		0,483	0,096
Dochód		0,051	0,010		0,185	0,039
Wykształcenie		0,056	0,006		0,297	0,024
Internet		0,361	0,039		0,622	0,106
„I”		M_1 : płatności gotówką	2,825		0,050	M_2 : model Poissona – transakcje gotówką (gdy brak karty)
Płeć	-0,101		0,013	-0,014	0,013	
Wiek	0,008		0,001	0,002	0,001	
Stan cywilny	-0,158		0,015	0,082	0,015	
Miejsce zamieszkania	0,145		0,015	0,126	0,015	
Dochód	0,016		0,006	-0,009	0,006	
Wykształcenie	-0,008		0,003	0,062	0,003	
Internet	-0,085		0,016	0,152	0,016	
α	-	0,004	0,001	-	-	-
δ	-	-1,876	0,041	-	-	-

Źródło: opracowanie własne.

Warto zauważyć, że występują duże różnice w wartościach oczekiwanych *a posteriori* parametrów opisujących liczbę transakcji gotówką w modelach M_1 i M_2 . Aż dla czterech (z siedmiu) zmiennych objaśniających (stan cywilny, dochód, wykształcenie i Internet) znaki tych charakterystyk są przeciwne – co oznacza, że kierunek wpływu danej zmiennej na liczbę wykonanych transakcji gotówką jest inny w zależności od tego, czy konsument posiada dodatkowy instrument płatności w postaci karty. Odchylenia standardowe *a posteriori* większości parametrów są stosunkowo małe. Spostrzeżenia te sugerują, że równość $\beta_2 = \beta_{2,0}$ nie zachodzi. Oznaczałoby to, że we wnioskowaniu o samych parametrach selekcja próby nie powoduje negatywnych konsekwencji i można ograniczyć się do każdego modelu oddzielnie (dla odpowiednich podzbiorów danych).

W celu zweryfikowania hipotezy $\beta_2 = \beta_{2,0}$ zastosowano bayesowski odpowiednik testu chi-kwadrat. Niech $\kappa = \beta_2 - \beta_{2,0}$; opierając się na idei testu nie-

bayesowskiego, dla zmiennej wielowymiarowej κ rozważa się formę kwadratową postaci (por. [Marzec, Osiewalski 2008]):

$$\tau = \tau(\kappa; y) = (\kappa - E(\kappa | y))' (V(\kappa | y))^{-1} (\kappa - E(\kappa | y)), \quad (25)$$

gdzie $E(\kappa | y) = E(\beta_2 | y) - E(\beta_{2,0} | y)$ i $V(\kappa | y) = V(\beta_2 | y) + V(\beta_{2,0} | y)$, a sumowanie macierzy kowariancji *a posteriori* wynika z niezależności *a posteriori* obu porównywanych wektorów parametrów, zachodzącej w proponowanym modelu ogólnym (bez restrykcji). Jednowymiarowa zmienna τ jest losowa jako funkcja zarówno obserwacji, jak i parametrów modelu; we wnioskowaniu na podstawie danych interesuje nas jej rozkład *a posteriori*, czyli warunkowy względem danych, o gęstości $p(\tau | y)$. Testowanie hipotezy $\kappa = 0$ sprowadza się do zbadania, czy wartość $\tau(0; y)$ leży w obszarze największej gęstości $p(\tau | y)$, przy wysokim, ustalonym prawdopodobieństwie *a posteriori* $(1 - \alpha)$ tego obszaru. Jeśli tak, to nie odrzucamy hipotezy $\kappa = 0$ i przechodzimy do analizy modelu z tą restrykcją, uniemożliwiającą odrębne traktowanie dwóch podzbiorów obserwacji. Jeśli wartość $\tau(0; y)$ znajduje się poza obszarem wysokiej gęstości *a posteriori*, to równość $\kappa = 0$ jest nieuzasadniona w świetle dostępnych danych i ją odrzucamy, pozostając przy wnioskach z modelu ogólnego, umożliwiającego wstępny podział obserwacji na dwie grupy i osobne traktowanie każdej z nich.

Wyniki uzyskane za pomocą tego testu świadczą przeciwko hipotezie $\kappa = 0$. Rozkład *a posteriori* zmiennej losowej $\tau(\kappa; y)$ jest jednomodalny i prawostronnie asymetryczny, a jego modalna wynosi 5,7. Przedział $(0, 20)$ zawiera $\tau(\kappa; y)$ z prawdopodobieństwem *a posteriori* równym 0,99, zaś wartość $\tau(0; y)$ wynosi 973,85. Zatem wartość ta znajduje się bardzo daleko w prawym ogonie rozkładu *a posteriori* dla τ , czyli założenie równości wektorów β_2 i $\beta_{2,0}$ nie znajduje uzasadnienia. Wystarczająca jest estymacja parametrów z wykorzystaniem separowalności funkcji wiarygodności danej wzorem (24); taka estymacja parametrów nie jest obciążona „błędem selekcji próby”.

Na koniec prezentujemy wyniki dla współczynników korelacji między liczbą transakcji oboma instrumentami płatniczymi (Y_{1t}, Y_{2t}) . Przypomnijmy, że współczynnik ten jest funkcją wszystkich parametrów trzech podmodeli, więc bez względu na wynik wcześniej prezentowanego testu wyznaczenie jego charakterystyk *a posteriori* jest możliwe tylko w modelu łącznym. Syntetyczne wyniki estymacji zostały pokazane w tabeli 5. Dla wszystkich obserwacji ($T = 2518$) otrzymano rozkłady *a posteriori* dla $\text{Corr}(Y_{1t}, Y_{2t} | \theta)$, czyli bezwarunkowego współczynnika korelacji, skupione blisko zera – ale wyłącznie po stronie wartości dodatnich; charakteryzowały się one małym odchyleniem standardowym. Średnia wartość oczekiwana *a posteriori* wyniosła 0,072, przy czym najmniejsza 0,031, a największa 0,16; korelacja jest więc bardzo słaba, ale dodatnia.

Tabela 5. Uśrednione (po obserwacjach) wartości oczekiwane *a posteriori* współczynników korelacji pary (Y_{1t}, Y_{2t})

Współczynnik korelacji	Średnia ocena	
	gdy konsument posiada kartę ($Y_{3t} = 1$)	gdy brak karty ($Y_{3t} = 0$)
$Corr(Y_{1t}, Y_{2t} \theta)$	0,072 (dla $T = 2518$)	
$Corr(Y_{1t}, Y_{2t} \theta)$	0,065 (dla $T_1 = 1190$)	0,079 (dla $T_2 = 1328$)
$Corr(Y_{1t}, Y_{2t} Y_{3t} = 1, \theta)$	0,073 (dla $T_1 = 1190$)	–

Źródło: opracowanie własne.

W modelu ZIP-CP – tylko dla posiadaczy karty płatniczej – średnia wartość oczekiwana *a posteriori* warunkowego współczynnika korelacji $Corr(Y_{1t}, Y_{2t} | Y_{3t} = 1, \theta)$ wyniosła 0,073. Średnia ocena korelacji bezwarunkowej (między liczbą transakcji kartą i gotówką) jest praktycznie taka sama jak korelacji warunkowej przy $Y_{3t} = 1$, choć średnie cząstkowe korelacji bezwarunkowej (liczone dla posiadaczy karty i dla osób bez karty) są odmienne.

5. Podsumowanie

Omawiany trójwymiarowy rozkład skokowy i zbudowany na tej podstawie bayesowski model statystyczny zaproponowano w celu łącznego modelowania dwóch zmiennych licznikowych, z których pierwsza może być zdegenerowana w zerze. Zaproponowany przez autorów model statystyczny polega na zastosowaniu zmiennej dychotomicznej (zero-jedynkowej) do przełączania między dwoma modelami zmiennych licznikowych: dwu- i jednowymiarowym, przy czym model jednowymiarowy jest otrzymywany z dwuwymiarowego przez odpowiednie warunkowanie. O ile przedstawiony schemat modelowania ma walor ogólności, o tyle wybór konkretnych klas modeli składowych może podlegać zmianom. Wybierając specyfikację ZIP-CP dla dwuwymiarowej zmiennej licznikowej i logistyczną dla zmiennej dychotomicznej kierowano się prostotą obu, prowadzącą do prostego modelu trójwymiarowego, a także zbadanymi dobrymi własnościami modelu ZIP-CP. Zmiana specyfikacji logistycznej na inną, np. opartą na dystrybucie skośnego rozkładu Studenta i wykorzystującą interakcje między zmiennymi objaśniającymi (tzw. model II rzędu, zob. [Osiewalski i Marzec 2004a]), nie jest trudna i może podnieść jakość modelu – choć nie musi (jest to kwestia empiryczna). Trudne będzie zastąpienie specyfikacji ZIP-CP, głównej części modelu przełącznikowego, innym modelem. Użycie w tym celu

alternatywnych specyfikacji dla dwóch powiązanych zmiennych licznikowych będzie przedmiotem dalszych badań autorów.

Jeśli chodzi o specyficznie bayesowski element zaproponowanego modelu – rozkład *a priori*, to jego postać może oczywiście podlegać zmianom, ale należy zwrócić uwagę na dwa kluczowe elementy. Separowalność funkcji wiarygodności względem parametrów modeli składowych może być w pełni wykorzystana tylko przy niezależności *a priori* tych grup parametrów, więc nie należy z niezależności rezygnować. Z kolei szczególna postać rozkładu *a priori* (normalny o średniej 0 i wariancji 1), przyjęta przez nas dla każdego indywidualnego parametru, nie ma znaczenia, jeśli liczba obserwacji jest bardzo duża (jak w zaproponowanym przykładzie empirycznym). Oczywiście, przy małej liczbie obserwacji warto dokonać analizy wrażliwości wyników na rozkład *a priori* w ramach szerszej klasy (np. rozkładów Studenta).

W zaproponowanym modelu bayesowskim można łatwo zastosować test typu Lindleya, tj. bayesowski odpowiednik testu F bądź chi-kwadrat, by zbadać zasadniczą restrykcję identyczności parametrów opisujących w dwóch sytuacjach tę zmienną licznikową, która jest niezdegenerowana (jest nietrywialnie obserwowana) dla obu wartości zmiennej zero-jedynkowej. W dalszych badaniach warto zastosować bayesowskie porównywanie mocy wyjaśniającej konkurencyjnych modeli poprzez ich prawdopodobieństwa *a posteriori*, co wymaga odpowiedniej metody obliczenia brzegowej gęstości wektora obserwacji w każdym z modeli. W przypadku stosowania metod Monte Carlo łańcuchów Markowa (w celu próbkowania rozkładu *a posteriori* co najwyżej kilkudziesięciu parametrów modeli) właściwym narzędziem jest skorygowany estymator średniej arytmetycznej, który zaproponowała A. Pajor [2017].

Poza testowaniem ważnej restrykcji, proponowany model trójwymiarowy pozwala badać skutki wstępnej selekcji obserwacji, polegającej na usunięciu tych, dla których obserwowana jest tylko jedna zmienna licznikowa. W przykładzie empirycznym, dotyczącym modelowania liczby płatności dwoma instrumentami płatniczymi (kartą i gotówką), wykazano, że samo wnioskowanie o poszczególnych parametrach nie było zagrożone „błędem selekcji próby”, gdyż odrzucono restrykcję wiążącą parametry opisujące liczbę płatności gotówkowych w sytuacjach posiadania karty płatniczej i jej braku. Pokazano też, że głębsze wnioskowanie o korelacji między liczbami płatności kartą i gotówką, odróżniające korelację warunkową (względem posiadania karty) od bezwarunkowej, możliwe jest dopiero na gruncie modelu pełnego (trójwymiarowego). Empirycznie, oba rodzaje korelacji okazały się podobne co do wartości.

Literatura

- Anderson T.W., Darling D.A. [1954], *A Test of Goodness of Fit*, „Journal of the American Statistical Association”, vol. 49, nr 268.
- Berkhout P., Plug E. [2004], *A Bivariate Poisson Count Data Model Using Conditional Probabilities*, „Statistica Neerlandica”, vol. 58, nr 3, <https://doi.org/10.1111/j.1467-9574.2004.00126.x>.
- Cameron A.C., Trivedi P.K. [1998], *Regression Analysis of Count Data*, Cambridge University Press, New York.
- Cameron A.C., Trivedi P.K. [2005], *Microeconometrics: Methods and Application*, Cambridge University Press, New York.
- Fiszeder P., Polasik M. [2009], *Modelowanie liczby transakcji dokonywanych przy użyciu gotówki i kart płatniczych na rynku polskim*, „Acta Universitatis Nicolai Copernici – Ekonomia”, vol. 39, https://doi.org/10.12775/AUNC_ECON.2009.029.
- Goczek Ł., Witkowski B. [2015], *The Determinants of Cash-free Transactions*, „The National Bank of Poland Working Paper Series”, nr 146.
- Goczek Ł., Witkowski B. [2016], *Determinants of Card Payments*, „Applied Economics”, vol. 48, <https://doi.org/10.1080/00036846.2015.1102846>.
- Górka J. [2013], *Efektywność instrumentów płatniczych w Polsce*, Wydawnictwo Naukowe Wydziału Zarządzania Uniwersytetu Warszawskiego, Warszawa.
- Kocherlakota S., Kocherlakota K. [1992], *Bivariate Discrete Distributions*, Marcel Dekker, New York.
- Lambert D. [1992], *Zero-inflated Poisson Regression, with an Application to Defects in Manufacturing*, „Technometrics”, vol. 34, <https://doi.org/10.2307/1269547>.
- Marzec J. [2012], *Wybrane dwuwymiarowe modele dla zmiennych licznikowych w ekonomii*, „Zeszyty Naukowe Uniwersytetu Ekonomicznego w Krakowie – Metody Analizy Danych”, nr 884.
- Marzec J., Osiewalski J. [2008], *Bayesian Inference on Technology and Cost Efficiency of Bank Branches*, „Bank i Kredyt”, vol. 39, nr 9.
- Marzec J., Osiewalski J. [2012], *Dwuwymiarowy model typu ZIP-CP w łącznej analizie zmiennych licznikowych*, „Folia Oeconomica Cracoviensia”, nr 53.
- Marzec J., Polasik M., Fiszeder P. [2013], *Wykorzystanie gotówki i karty płatniczej w punktach handlowo-usługowych w Polsce: zastosowanie dwuwymiarowego modelu Poissona*, „Bank i Kredyt” vol. 44, nr 4.
- Osiewalski J. [2012], *Dwuwymiarowy rozkład ZIP-CP i jego momenty w analizie zależności między zmiennymi licznikowymi* [w:] *Spotkania z królową nauk. Księga jubileuszowa dedykowana Profesorowi Edwardowi Smadze*, Wydawnictwo Uniwersytetu Ekonomicznego w Krakowie, Kraków.
- Osiewalski J., Marzec J. [2004a], *Model dwumianowy II rzędu i skośny rozkład Studenta w analizie ryzyka kredytowego*, „Folia Oeconomica Cracoviensia”, nr 45.
- Osiewalski J., Marzec J. [2004b], *Uogólnienie dychotomicznego modelu probitowego z wykorzystaniem skośnego rozkładu Studenta*, „Przegląd Statystyczny”, t. 51.
- Pajor A. [2017], *Estimating the Marginal Likelihood Using the Arithmetic Mean Identity*, „Bayesian Analysis”, vol. 12, nr 1, <https://doi.org/10.1214/16-BA1001>.
- Polasik M. [2015], *Stan i potencjał rozwoju sieci akceptacji kart płatniczych w Polsce*, „Acta Universitatis Nicolai Copernici, Ekonomia”, vol. 46, https://doi.org/10.12775/AUNC_ECON.2015.002.

- Polasik M., Maciejewski K. [2009], *Innowacyjne usługi płatnicze w Polsce i na świecie*, „Materiały i Studia NBP”, nr 241.
- Polasik M., Marzec J., Fiszeder P., Górka J. [2012], *Modelowanie wykorzystania metod płatności detalicznych na rynku polskim*, „Materiały i Studia NBP”, nr 265.
- Polasik M., Wisniewski T.P., Lightfoot G. [2012], *Modelling Customers' Intentions to Use Contactless Cards*, „International Journal of Banking, Accounting and Finance”, vol. 4, nr 3, <https://doi.org/10.1504/IJBAAF.2012.051590>.
- Winkelmann R. [2008], *Econometric Analysis of Count Data*, Springer-Verlag, Berlin Heidelberg.

Bivariate Count Variables – Bayesian Modelling of Sample Selection

(Abstract)

The article presents a joint statistical model of two count variables, one of which can be degenerated at zero. We consider a modelling framework based on switching between a bivariate Poisson regression model and a univariate one, where the switching depends on the observed outcome of the third, dichotomous variable. Bayesian analysis is advocated; in two special cases of our Bayesian model, important consequences for inference are stated. In the empirical section we consider joint modelling of the number of cash and bank card transactions in Poland with the use of data for both cardholders and non-holders.

Keywords: bivariate Poisson regression models, switching between non-degenerate and degenerate distributions, likelihood factorisation, bank card and cash payments.

Andrzej Czyżewski
Aleksander Grzelak

Wykorzystanie modelu przepływów międzygałęziowych do ocen makroekonomicznych gospodarki

Streszczenie

Celem artykułu jest przedstawienie możliwości wykorzystania modelu przepływów międzygałęziowych do ocen makroekonomicznych gospodarki, z uwzględnieniem dotychczasowych doświadczeń i propozycji na przyszłość. W trakcie badań stwierdzono, że model przepływów międzygałęziowych stanowi użyteczny instrument oceny i interpretacji zjawisk gospodarczych. Przede wszystkim chodzi tu o ocenę związków, a także zdiagnozowanie źródeł zróżnicowanego rozwoju pomiędzy działami w obrębie gospodarki danego kraju. Istnieje potrzeba rozbudowy interpretacji modelu przepływów międzygałęziowych o czwartą część tablicy przepływów, co oznacza konieczność publikacji danych dotyczących podziału wytworzonych dochodów. W dalszej kolejności interesującą byłaby próba wykorzystania tego modelu do prezentacji skali nierównowagi podażowej w gospodarce oraz uwzględnienie kontekstu środowiskowego analiz makroekonomicznych.

Andrzej Czyżewski, Uniwersytet Ekonomiczny w Poznaniu, Wydział Ekonomii, Katedra Makroekonomii i Gospodarki Żywnościowej, al. Niepodległości 10, 61-875 Poznań, e-mail: Andrzej.Czyzewski@ue.poznan.pl

Aleksander Grzelak, Uniwersytet Ekonomiczny w Poznaniu, Wydział Ekonomii, Katedra Makroekonomii i Gospodarki Żywnościowej, al. Niepodległości 10, 61-875 Poznań, e-mail: agrzelak@interia.pl

Słowa kluczowe: model przepływów międzygałęziowych, makroekonomia, gospodarka, tworzenie i podział dochodów.

Klasyfikacja JEL: E01, E10, E66.

1. Wprowadzenie

Procesy zachodzące w gospodarce cechuje złożoność. Stwarza to potrzebę wykorzystania w makroekonomicznych ocenach wielu teorii i modeli. Jedną z takich możliwości są bilanse pieniężnych przepływów międzygałęziowych. Poprzez analizę strumieni przepływów wyrażających wartość transakcji pomiędzy dostawcami i odbiorcami (*input-output*) model konkretyzuje idee funkcjonowania gospodarki (mechanizmu rynkowego i budżetowego), jego wewnętrzne powiązania, struktury i zależności oraz efekty decydujące o procesach reprodukcji. Pomimo że ten nurt badań nie znalazł się na razie w Polsce w centrum zainteresowania, to może być użytecznym instrumentem oceny funkcjonowania gospodarki w ujęciu zagregowanym w perspektywie wieloletniej [Tomaszewicz 1994]. Jest to tym bardziej uzasadnione, że istnieją interesujące przesłanki dalszego rozwoju podejścia *input-output* w analizach ekonomicznych [Boratyński, Przybyliński i Świerczewska 2015]. Wykorzystując założenia teorii równowagi ogólnej, model ten umożliwia analizę tworzenia i podziału wytworzonych efektów makroekonomicznych oraz oddziaływania procesów globalnych na gospodarkę poprzez eksport i import. Walorem modelu jest możliwość wnioskowania dedukcyjnego, co pozwala na jego szerokie wykorzystanie, także w kształceniu na studiach ekonomicznych. Nie bez znaczenia jest również uniwersalność modelu w przypadku badania współzależności zachodzących w obrębie gospodarek danych krajów, w tym oceny efektów zmian aktywności gospodarczej [Bess i Ambargis 2011]. Świadczy też o tym analiza cech systemów gospodarczych niezależnych od sposobu regulacji produkcji i jej podziału [Leontief 1936, Stankiewicz 2007]. Z drugiej strony trzeba zauważyć, że znaczny zakres agregacji danych wymaga ostrożności w interpretacji tablicy przepływów międzygałęziowych [Lonc 1985].

Celem artykułu jest przedstawienie możliwości wykorzystania modelu przepływów międzygałęziowych w analizach makroekonomicznych gospodarki oraz sformułowanie konkluzji co do możliwości poszerzenia jego zastosowania w badaniach ekonomicznych. W związku z tym przedstawiono rekomendacje dotyczące modyfikacji dotychczasowego sposobu przedstawiania bilansów przepływów międzygałęziowych, stąd zaprezentowane rozważania mają charakter teoretyczny, z elementami postulatycznymi.

2. Model input-output – ujęcie retrospektywne i metodyka

Gospodarka danego kraju składa się z wielu różnych działów¹ wzajemnie ze sobą powiązanych. Związki te realizowane są poprzez transakcje kupna i sprzedaży dóbr oraz usług, które integrują wszystkie podmioty gospodarcze. Wymiernym efektem tych transakcji są strumienie, które mają charakter strukturotwórczy, a ponadto ich wartość decyduje o sile zależności pomiędzy działami². Istnienie przepływów produktów (usług) między działami w gospodarce stwarza zapotrzebowanie na analizę nakładów i wyników w skali poszczególnych działów (grup przedsiębiorstw) oraz całej gospodarki. Teoretyczne podstawy rozwoju modelu wynikają z uproszczonej wersji klasycznej teorii równowagi ogólnej nawiązującej do opisu zależności gospodarki zarówno w zakresie tworzenia PKB, jak i jego struktur. Ponadto model *input-output* jest na tyle elastycznym instrumentem, że umożliwia ocenę funkcjonowania gospodarek w różnych systemach gospodarczych zgodnie z metodyką liczenia agregatów makroekonomicznych zarówno według MPS (*material product system*), jak i SNA (*system of national account*). W pierwszym przypadku uwzględniane były jedynie działy produkcyjne, a więc wytwarzające dobra i usługi materialne, przy założeniu, że sfera niematerialna jest redystrybuowana poprzez budżet (spożycie zbiorowe). Z kolei w przypadku metody SNA uwzględniana jest pełna aktywność gospodarcza badanych podmiotów.

Pierwszym, który zauważył i wykorzystał sens analizy przepływów dóbr w gospodarce, był F. Quesnay, nadworny lekarz Ludwika XV, który w formie tablicy ekonomicznej przedstawił transfery dóbr między trzema działami gospodarki: rolnictwem (klasa produkcyjna), sferą pozarolniczą (klasa jałowa) i właścicielami (władza świecka i duchowna) [Quesnay 1928]. Na tej podstawie określił współzależności sfery wytwórczej gospodarki, podziału wytworzonego produktu społecznego oraz sfery dochodowej, odpowiadając na pytanie, kto i jakiego rodzaju otrzymuje dochód.

Kolejnym etapem rozwoju tego podejścia były marksowskie schematy reprodukcji (1) i (2):

$$P_1 = c_1 + v_1 + m_1, \quad (1)$$

$$P_2 = c_2 + v_2 + m_2,$$

$$c_2 = v_1 + m_1. \quad (2)$$

¹ Dział gospodarki rozumiany jest w artykule tak jak w opracowaniach GUS.

² W tym przypadku należy uwzględnić stopień agregacji działów. Przykładowo, za 2010 r. GUS w publikowanych danych przepływów międzygałęziowych dokonał podziału na 77 działów, a w 2005 r. było to tylko 55 działów. Wzrost liczby działów sprzyja większej szczegółowości analiz, ale z drugiej strony może utrudniać syntetyczne oceny.

Całkowity produkt globalny P niezbędny do wyprodukowania danego produktu równy jest sumie c (wartości zużytych środków produkcji) oraz kosztu siły roboczej V i wartości produktu dodatkowego m [Marks 1955]. W celu zachowania równowagi przepływów pomiędzy działem pierwszym (środki produkcji) i drugim (środki konsumpcji) musi być spełniony warunek równowagi (2). Koncepcję teorii równowagi ogólnej, przez przedstawienie powiązanych układów równań produkcji w warunkach cen wolnokonkurencyjnych, skonkretyzował L. Walras [1926]. Najbardziej przejrzystym modelem związków typu dostawca–odbiorca jest zapis szachownicowy przepływów rzeczowo-finansowych W. Leontiefa [1963], który przyjmujemy za podstawę dalszych rozważań. Jego istota sprowadza się do założenia, że gospodarka narodowa stanowi agregat zasobów i strumieni składających się z kilku sprzężonych ze sobą układów, które opisano metodą nakładów i wyników (*input-output*) w formie tabelarycznej (szachownicowej) – tabela 1.

Tabela 1. Wzorcowy układ tabelaryczny przepływów międzygałęziowych

$i \backslash j$	Produkcja globalna	Stadia pośrednie			Odbiorcy końcowi		
		1	2	j	razem	C	I
1	X_1	w_{11}	w_{12}	$\dots w_{1j}$	x_1	C_1	I_1
2	X_2	w_{21}	w_{22}	$\dots w_{2j}$	x_2	C_2	I_2
·	·	·	·	·	·	·	·
i	X_i	w_{i1}	w_{i2}	$\dots w_{ij}$	x_i	C_i	I_i
$i + 1$	V	V_1	V_2	$\dots V_j$			
$i + 2$	S	S_1	S_2	$\dots S_j$			
Ogółem	X	X_1	X_2	$\dots X_j$			

Oznaczenia: w_{ij} – wartość przepływów pomiędzy działami, V – płace, S – nadwyżka operacyjna brutto plus podatki, C – konsumpcja, I – inwestycje.

Źródło: [Leontief 1963].

Przyjęto założenie, że produkt każdego działu (*output*) może służyć jako środek produkcji (*input*) w każdym innym dziale bądź inwestycji lub jako środek konsumpcji. Model Leontiefa zawiera trzy części. Pierwsza przedstawia transakcje pomiędzy przedsiębiorstwami zgrupowanymi w gałęzi. Odzwierciedla więc ona popyt pośredni. Druga część obrazuje wielkość i strukturę popytu przeznaczonego dla odbiorców końcowych. Z kolei trzecia przedstawia wielkość i strukturę dochodów poszczególnych gałęzi. W dalszym etapie rozwoju tego modelu P. Samuelson [1951] i R. Solow [1959] dokonali jego rozszerzenia, wykorzystując metody optymalizacyjne. Z podejścia tego powstały także równania równowagi przepływów międzygałęziowych (3) [Lange 1961]. Lewa strona tego równania określa,

co pobiera (zakupuje) i wytwarza dany dział produkcji, a prawa, co sprzedaje pozostałym działom (3):

$$\sum_{i=1}^n w_{ij} + V_j + S_j = \sum_{j=1}^m w_{ij} + x_i \quad (i = 1, 2, \dots, n; j = 1, 2, \dots, m), \quad (3)$$

gdzie:

w_{ij} – nakłady na środki produkcji,

V_j – nakłady na płace,

S_j – wartość dodatkowa (nadwyżka operacyjna brutto + podatki – według SNA),

x_i – produkt (popyt) końcowy danego działu.

Wykorzystując podejście W. Leontiefa, stworzono wiele modyfikacji modelu przepływów międzygałęziowych. Idea przepływów międzygałęziowych ma swój wymiar zarówno teoretyczny, głęboko osadzony w historii myśli ekonomicznej, jak i aplikacyjny, związany z publikowaniem przez urzędy statystyczne stosownych bilansów. W tym miejscu przedstawimy przykładową interpretację tablicy, wykorzystywaną do objaśniania współzależności gospodarczych (tabela 2). Podejście to zawiera interpretację modelu dla jego czterech części. Uwagę zwraca występowanie w tablicy ćwiartki IV, która w praktyce rzadko jest wypełniana³. W ćwiartce I (lewa górna część tablicy) zawarto popyt pośredni zgłaszany przez działy gospodarki. Zostały tu zaprezentowane wzajemne transakcje między działami. W wierszach przedstawiony został strumień popytu pośredniego. Są to zakupy produktów (usług), które służą dalszemu przetworzeniu. Kolumny natomiast interpretujemy jako strukturę kosztów bezosobowych poszczególnych działów [Czyżewski 2011]. W części II tablicy przedstawiono popyt końcowy: konsumentów indywidualnych, budżetu, sektora bankowego, inwestorów oraz restytucyjny. Na podstawie zawartych tam danych można dokonać oceny strumieni odnoszących się do rozdysponowania produktów danych działów służących zaspokojeniu finalnego popytu przez podmioty gospodarcze. Jednocześnie ta ćwiartka tablicy może być też rozpatrywana przez pryzmat popytu potencjalnego, tj. aspiracji wyróżnionych podmiotów. Popyt, do którego aspirują, jest wyższy od efektywnego (zrealizowanego). Różnica między tymi kategoriami na gruncie ekonomii keynesowskiej sprowadza się do istnienia luki popytowej. W gospodarce rynkowej uwidocznione jest to poprzez występowanie nierównowagi podażowej.

³ W materiałach GUS w Polsce w odniesieniu do przepływów publikowane są tylko trzy części. Wynika to stąd, że na poziomie zagregowanym pomija się część czwartą dotyczącą podziału dochodów. Niemniej jednak procesy, które związane są z tym zjawiskiem, są istotne dla funkcjonowania gospodarki i dlatego w rozważaniach w niniejszym artykule podział dochodów został uwzględniony na poziomie teoretycznym [Czyżewski 2011].

Tabela 2. Bilans pieniężnych przepływów międzygałęziowych

i	j	Część I		Część II				Razem produkcja globalna i eksport		
		Popyt pośredni		Popyt końcowy podmiotów gospodarczych (wydatki)		Razem popyt końcowy	Razem produkcja globalna i eksport			
		działy	razem popyt pośredni	konsumenci indywidualni	budżet				banki	inwestorzy
	I	j	Σ	I	II	III	IV	V	Σ	$\Sigma\Sigma$
Koszty materiałowe i usługi	w_{11}	w_{1j}	Σ	C'_1	C''_1	K_1	I_1	p_1	Σ	$\Sigma\Sigma$
	\cdot	\cdot	\cdot	\cdot	\cdot	\cdot	\cdot	\cdot	\cdot	\cdot
	\cdot	\cdot	\cdot	\cdot	\cdot	\cdot	\cdot	\cdot	\cdot	\cdot
	\cdot	\cdot	\cdot	\cdot	\cdot	\cdot	\cdot	\cdot	\cdot	\cdot
	w_{i1}	w_{ij}	Σ	C'_i	C''_i	K_i	I_i	p_i	Σ	$\Sigma\Sigma$
Razem koszty materiałowe i usługi	Σ	Σ	Σ	Σ	Σ	Σ	Σ	Σ	Σ	$\Sigma\Sigma$
$i+1$	V_1	V_j	Σ	V'_1	V''_1	V_k	V_l	-	Σ	Σ
$i+2$	B_1	B_j	\cdot	B'_1	B''_1	B_k	B_l	-	\cdot	\cdot
$i+3$	A_1	A_j	\cdot	-	-	-	-	A_p	\cdot	\cdot
$i+4$	Z_1	Z_j	\cdot	Z'_1	Z''_1	Z_k	Z_l	-	\cdot	\cdot
$i+5$	m_1	m_j	\cdot	m'_1	m''_1	m_k	m_l	-	\cdot	\cdot
$i+6$	TA	TA_j	Σ	TA'_1	TA''_1	TA_k	TA_l	-	Σ	$\Sigma\Sigma$
Razem dochody brutto	Σ	Σ	Σ	Σ	Σ	Σ	Σ	Σ	Σ	$\Sigma\Sigma$
Razem produkcja globalna i import (koszty i dochody brutto)	$\Sigma\Sigma$	$\Sigma\Sigma$	$\Sigma\Sigma$	$\Sigma\Sigma$	$\Sigma\Sigma$	$\Sigma\Sigma$	$\Sigma\Sigma$	$\Sigma\Sigma$	$\Sigma\Sigma$	$\Sigma\Sigma$

Oznaczenia: w_{ij} – przepływ dóbr i usług i wytworzonych w danej gospodarce bądź pochodzących z importu uzupełniającego produkcję tej gospodarki (wraz z marżami handlowymi), zużytych przez dział j ; $V_j(i+1)$ – płace pracownikom (konsumentów); $B_j(i+2)$ – dochody poszczególnych działów z tytułu dotacji budżetowych; $A_j(i+3)$ – amortyzacja; $Z_j(i+4)$ – dochody poszczególnych działów z tytułu nadwyżki ekonomicznej; $m_j(i+5)$ – dochody poszczególnych działów z tytułu zmiany ilości pieniądza pozostającego do ich dyspozycji (głównie z tytułu kredytów, a także emisji); $TA_j(i+6)$ – podatki bezpośrednie i pośrednie; $C'_i(I)$ – struktura popytu końcowego (wydatków) konsumentów indywidualnych na dobra i ; $C''_i(II)$ – struktura popytu końcowego budżetu (wydatków) na dobra i ; $K_i(III)$ – zakupy bankowe; struktura popytu sektora bankowego na dobra i ; $I_i(IV)$ – struktura popytu końcowego sfery inwestycji na dobra i (zakupy inwestycyjne); $p_i(V)$ – struktura popytu restytucyjnego podmiotów gospodarczych (wydatki na rzecz amortyzacji).

Źródło: [Czyżewski 2011].

Produkcja, która w uproszczeniu może być odzwierciedleniem łącznego popytu, do którego podmioty aspirują, nie znajduje realnego (efektywnego) popytu w warunkach rynkowych. Rośnie więc presja na konsumenta poprzez działania promocyjne.

Ćwiartka III tablicy przepływów (lewa dolna część tablicy) pokazuje dochody i ich strukturę, które zostały wytworzone w poszczególnych działach gospodarki. W wierszach wyróżnione są poszczególne elementy wartości dodanej, m.in. płace, nadwyżka ekonomiczna. W oddzielnych wierszach zapisano wpływ budżetu państwa i banków na dochody przedsiębiorstw, poprzez m.in. dochody z tytułu zmiany ilości pieniądza emisyjnego i kredytowego oraz podatki. Na podstawie informacji zawartych w tej ćwiartce tablicy możemy określić makroekonomiczne efekty działalności poszczególnych działów, w tym zwłaszcza w postaci wartości dodanej. Część IV tablicy dotyczy wtórnego podziału wytworzonych dochodów, które z kolei służą realizacji popytu finalnego⁴. W tej ćwiartce tablicy dochód zostaje rozdzielony pomiędzy podmioty występujące w gospodarce: konsumentów indywidualnych, budżet, bank oraz inwestorów. Po stronie dochodów budżetu uwzględnia się głównie wpływy z podatków i opłat, ale także z innych tytułów, np. zysków przedsiębiorstw państwowych. Z kolei wydatki obejmują środki przeznaczone na konsumpcję zbiorową, inwestycje, dotacje i subwencje, a także ewentualny podział wpływów z innych tytułów. W odniesieniu do banków zawiera ona informacje o wpływach z racji otwieranych rachunków i emisji pieniądza oraz wydatków z tytułu kredytów obrotowych i inwestycyjnych. Taka konstrukcja tablicy gwarantuje możliwość obserwacji zarówno strony rzeczowej, jak i finansowej przepływów [Czyżewski 2011], np. procesy globalizacji oddziałują na tę część bilansu przepływów międzygałęziowych poprzez presję wywieraną na zmniejszenie roli państwa w procesach gospodarczych [Czyżewski i Grzelak 2009].

Warto w tym miejscu podkreślić, że model przepływów międzygałęziowych stosowany współcześnie w praktyce gospodarczej zakłada, iż II i III ćwiartka tablicy wystarczą do interpretacji procesów równowagi gospodarczej (tabela 3). Tym samym przyjmuje się, że produkt krajowy brutto wytworzony (PKB) równy jest podzielonemu, tak jakby nie występowała nierównowaga gospodarcza (tu podaźowa), a popyt efektywny równy był potencjalnemu, do którego aspirują uczestnicy rynku, co jest oczywiście założeniem idealizującym.

⁴ W praktyce gospodarczej w Polsce wykorzystywano ją jedynie sporadycznie w celu koordynacji strony finansowej planów gospodarczych. Ostatnie publikowane dane z tego zakresu dotyczą bilansu gospodarki polskiej za 1957 r. [*Opracowanie Biura Ekonomicznego NBP...* 1958].

Tabela 3. Bilans pieniężnych przepływów międzygałęziowych stosowanych współcześnie w praktyce gospodarczej w Polsce

$\begin{matrix} j \\ \backslash \\ i \end{matrix}$		Część I				Część II						Ogółem zużycie pośrednie i popyt końcowy
		Zużycie pośrednie (PKD)		Popyt końcowy		spożycie		akumulacja brutto		eksport	razem	
		działy	razem	przez instytucje komercyjnie	przez instytucje rządowe i samorządowe	razem	nakłady brutto na środki trwałe	przystrost rzeczowych środków obrotowych	razem			
Produkty (PKWiU)	1	w_{11}	w_{1j}	C_1	C_1'	C_1''	Σ	N_1	O_1	E_{x_1}	3Σ	4Σ

	i	w_{i1}	w_{ij}	C_i	C_i'	C_i''	Σ	N_i	O_i	E_{x_i}	3Σ	4Σ
Razem produkty		Σ	Σ	Σ	Σ	Σ	Σ	Σ	Σ	Σ	3Σ	4Σ
a		a_1	A_j	Ca	Ca'	Ca''	Σ	Na	Oa	E_{xa}	3Σ	4Σ
b		b_1	B_j	Cb	Cb'	Cb''	Σ	Nb	Ob	E_{xb}	3Σ	4Σ
Razem zużycie pośrednie		Σ	Σ	Σ	Σ	Σ	Σ	Σ	Σ	Σ	6Σ	8Σ
Część III	$i+1$	V_1	V_j									
	$i+2$	TA_1	TA_j									
	$i+3$	A_1	A_j									
	$i+4$	Z_1	Z_j									
Wartość dodana brutto		Σ	Σ	Σ	Σ	Σ	Σ	Σ	Σ	Σ	6Σ	8Σ
Produkcja globalna		Σ	Σ	Σ	Σ	Σ	Σ	Σ	Σ	Σ	3Σ	

Oznaczenia: a – wykorzystanie produktów importowanych, cif ; b – podatki od produktów pomniejszone o dotacje do produktów; $V_j (i+1)$ – koszty związane z zatrudnieniem (płace); $TA_j (i+2)$ – podatki od producentów; $A_j (i+3)$ – amortyzacja środków trwałych; $Z_j (i+4)$ – nadwyżka operacyjna netto.

Źródło: [Bilans przepływów... 2014].

3. Zastosowania makroekonomiczne modelu input-output – ujęcie teoretyczne

Istnieje wiele możliwości wykorzystania przepływów międzygałęziowych do oceny funkcjonowania gospodarki. Chodzi tu o identyfikację związków pomiędzy działami w gospodarce zachodzących poprzez przepływy międzygałęziowe, jak również ich znaczenie w tworzeniu produktu globalnego czy wartości dodanej. Wykorzystać tu można n -równań opisujących równowagę nakładów i efektów (4) w całej gospodarce:

$$X_i - \sum_{j=1}^n w_{ij} = x_i \quad (i = 1, 2, \dots, n; j = 1, 2, \dots, m), \quad (4)$$

gdzie:

X_i – produkcja globalna,

w_{ij} – wartość przepływów pomiędzy działami,

x_i – produkcja (popyt) finalna w i -tym dziale.

Bilanse przepływów międzygałęziowych mogą służyć także do oceny struktury strumieni produktów „zasilających” dany dział (odpowiednie kolumny w I części tablicy przepływów międzygałęziowych). Pozwala to na określenie zakresu samozaopatrzenia czy wzajemnych powiązań pomiędzy działami w układzie przedmiotowym i dynamicznym (dla ujęć wieloletnich). Oceniając z kolei struktury rozdysponowania produktów danych działów, a w szczególności elementy popytu końcowego (spożycie, akumulację), można dokonać oceny ich pozycji w gospodarce. W literaturze dotyczącej tego tematu można znaleźć wiele przykładów takiego rodzaju zastosowań przepływów międzygałęziowych, szczególnie w przypadku gospodarki żywnościowej [Mrówczyńska-Kamińska 2015, Grzelak 2006]. Identyfikacja wzajemnych powiązań jest szczególnie istotna w przypadku programów restrukturyzacji określonych działów gospodarki. W ujęciu wieloletnim można ocenić dynamikę rozwoju poszczególnych działów, jak też charakter zmian w zakresie oceny pozycji w procesach gospodarczych.

Rozwijając ten wątek, warto zbadać strukturę nakładów bieżących, a przez odwrócenie tzw. współczynników „chłonności” określić efektywność poszczególnych rodzajów nakładów. Celowi temu służą m.in. współczynniki pracochłonności i produktochłonności (materiałochłonności), inaczej określane jako współczynniki techniczne produkcji (5):

$$a_{ij} = \frac{w_{ij}}{X_j} \quad (i = 1, 2, \dots, n; j = 1, 2, \dots, m), \quad (5)$$

gdzie: X_j – produkcja globalna.

Współczynniki te można zapisać w postaci macierzy (6):

$$A = [a_{ij}] = \begin{pmatrix} a_{11} & \dots & a_{1j} \\ \dots & \dots & \dots \\ a_{i1} & \dots & a_{ij} \end{pmatrix}. \quad (6)$$

Techniczny współczynnik produkcji może charakteryzować technologię procesu produkcyjnego każdego z działów. Natomiast pod względem interpretacyjnym wskazuje, jaka wartość i -tego produktu danego działu jest niezbędna do wyprodukowania jednostki produkcji globalnej j -tego działu. W ten sposób model przepływów może służyć do krótkookresowego prognozowania przyszłej wartości produkcji zarówno globalnej, jak i finalnej przy założeniu niezmienniej technologii produkcji⁵. Mechanizm rynkowy wywołuje presję na obniżenie tych współczynników i tym samym relatywne zmniejszenie znaczenia popytu pośredniego w gospodarce. Związane jest to z wprowadzaniem innowacji, a tym samym postępowaniem technicznym. Różne jest jednak tempo tych zmian, co przesądza o przeobrażeniach strukturalnych w gospodarce. Stąd współczynniki materiałochłonności służyć mogą także oszacowaniu międzygałęziowych dyfuzji innowacji [Wolff 2011, Świeczewska i Tomaszewicz 2012] czy identyfikacji, które w największym stopniu wpływają na wzrost potencjału innowacyjnego gospodarki [Świeczewska 2014]. Interesujący sposób wykorzystania modelu przepływów międzygałęziowych przedstawił B. Czyżewski [2013]. Na podstawie danych empirycznych dokonał szacunku międzygałęziowych przepływów rent ekonomicznych w Polsce z rolnictwa do jego otoczenia.

Podobnie jak w przypadku produktochłonności (materiałochłonności) można wyznaczyć pracochłonność poszczególnych działów (7):

$$P_j = \frac{V_j}{X_j} \quad (j = 1, 2, \dots, m), \quad (7)$$

gdzie:

P_j – pracochłonność,

V_j – koszty wynagrodzeń,

X_j – produkcja globalna.

Określa się w tym przypadku relację kosztów wynagrodzeń na jednostkę produkcji globalnej danego działu. Zmniejszenie tej relacji wskazuje na wzrost wydajności pracy i tym samym bardziej efektywne wykorzystanie zasobów pracy.

⁵ W literaturze dotyczącej tego tematu istnieje wiele przykładów takich zastosowań, np. modele CGE (*computable general equilibrium*) [Handbook... 2013] czy modele zintegrowane (*input-output econometric models*).

Z drugiej jednak strony należy mieć świadomość, że wiązać się to może ze zwiększeniem bezrobocia i wzrostem obciążeń socjalnych budżetu państwa.

Analiza tablicy przepływów międzygałęziowych pozwala także na ocenę efektywności makroekonomicznej poszczególnych działów gospodarki. Rozumiana może być ona jako udział wartości dodanej brutto w łącznej wartości dodanej brutto wytworzonej w gospodarce lub w produkcji globalnej danego działu (8):

$$E_j = \frac{WD_j}{X_j} \quad (j = 1, 2, \dots, m), \quad (8)$$

gdzie:

E_j – efektywność makroekonomiczna działu j ,

WD_j – wartość dodana działu j ,

X_j – produkcja globalna.

Relacja ta pozwala wstępnie określić dysparytety rozwojowe w gospodarce i tym samym dostosować instrumenty polityki gospodarczej z uwzględnieniem paradygmatu rozwoju zrównoważonego.

Kolejnym sposobem określania efektywności może być ocena relacji popytu końcowego do wartości strumieni zasilających dany dział (innymi słowy, jest to wskaźnik efektywności powiązań międzygałęziowych) [Czyżewski i Grzelak 2007] (9):

$$E_{pm} = \frac{x_i}{\sum_{i=1}^n w_{ij}}, \quad (9)$$

gdzie:

E_{pm} – efektywność powiązań międzygałęziowych,

x_i – produkcja (popyt) finalna w i -tym dziale,

w_{ij} – wartość przepływów pomiędzy działami.

Wskaźniki efektywności służą do określenia pozycji konkurencyjnej danego działu względem pozostałych działów. Pośrednio mogą także wskazywać na transfer wypracowanych już efektów i rent do otoczenia (głównie poprzez system cen). Określają także poprawę bądź pogorszenie rozwoju danego działu przez pryzmat kreacji efektów dochodowych. Ponadto związki te mogą stanowić podstawę do konturowania symulacji makromodeli. Istnieje wiele opracowań dotyczących wykorzystania modelu *input-output* do oceny efektywności działów, np. działu rolnego [Woś 1973, Czyżewski i Smędzik-Ambroży 2015]. Należy w tym miejscu podkreślić, że omawiane współczynniki efektywności powiązań międzygałęziowych obarczone są uproszczeniem wynikającym z techniczno-bilansowego ujęcia danych w tablicy przepływów międzygałęziowych. Chodzi tu o założenie jednoczesnej transformacji nakładów w efekty, tj. brak odroczenia

w czasie [Czyżewski i Helak 1991]. Nie przekreśla to jednak wnioskowania o zaistniałych tendencjach w tym zakresie.

Powiązanie działań gospodarki z zagranicą można analizować przez pryzmat zmian udziału eksportu w łącznym bądź finalnym popycie na produkty (usługi), jak również z perspektywy importochłonności danego działu. Pierwsza z wymienionych możliwości pozwala na ocenę zmian konkurencyjności zewnętrznej. W przypadku natomiast współczynnika importochłonności (bezpośredniej), definiowanej jako wartość produktów zużytych bezpośrednio przez dział, a pochodzących z importu, ocenić możemy znaczenie „zasilen” przez strumienie produktów z importu. Tablica przepływów międzygałęziowych może być również wykorzystana do oceny dynamiki poszczególnych składników przy porównaniach wieloletnich. Chodzi w szczególności o te elementy, które związane są ze zmianami popytu finalnego na produkty danego działu, wartości dodanej czy strumieni dóbr i usług zasilających dany dział. Istotne znaczenie w tym przypadku ma wycena poszczególnych składników, a zwłaszcza zastosowanie właściwych deflatorów przy porównaniach w cenach stałych.

4. Możliwości uzupełnienia statystyki bilansów przepływów międzygałęziowych*

Dotychczasowe doświadczenia związane z wykorzystaniem bilansów przepływów międzygałęziowych do ocen makroekonomicznych pozwalają na sformułowanie określonych rekomendacji. Najważniejszy z postulatów dotyczy potrzeby wypełnienia IV ćwiartki tablicy przepływów. Brak tego istotnie ogranicza możliwości analizy w zakresie podziału PKB i związanych z tym procesów redystrybucji budżetowej, a także efektywności polityki fiskalnej. Na poziomie zagregowanych ocen umożliwiłoby to pełniejsze określenie roli budżetu państwa w procesach rozwojowych gospodarki.

Innym zagadnieniem jest kwestia nierównowagi podażowej ocenianej z perspektywy bilansów w modelu przepływów międzygałęziowych. Zarówno popyt efektywny (zrealizowany), wytworzone dochody, jak i podzielone efekty teoretycznie powinny być sobie równe. Wynika to z tożsamości w rachunku makroekonomicznym. Strumienie wydatków podmiotów gospodarczych powinny pokrywać się z łącznymi ich dochodami za pośrednictwem mechanizmu rynkowego. Jak wobec tego wyjaśnić realnie występującą nierównowagę rynkową, tzn. podażową? Gospodarka w tych warunkach nie wykorzystuje zazwyczaj wszystkich swoich możliwości w sensie pełnego wykorzystania zdolności produkcyj-

* Opracowano na podstawie [Czyżewski i Grzelak 2012].

nych. W tym przypadku można mówić, używając terminologii keynesowskiej, o równowadze, ale w warunkach niepełnego wykorzystania mocy wytwórczych w gospodarce (luka popytowa). Sytuacja ta przejawia się nadwyżką produkcji nad popytem, wzrostem poziomu zapasów czy niewykorzystanymi mocami wytwórczymi. W modelu przepływów międzygałęziowych mogłoby to zostać uwidocznione w niezrealizowanych aspiracjach konsumentów i producentów w II ćwiartce tablicy przepływów. Chodzi w szczególności o uwzględnienie zarówno zapasów, oszczędności podmiotów, jak i szacunku alternatywnej wielkości popytu potencjalnego ponad poziom naturalnej stopy bezrobocia. Wówczas w ćwiartce IV przedstawiony mógłby być popyt efektywny, tj. zrealizowany, natomiast w ćwiartce II popyt potencjalny, który byłby wyższy (co do aspiracji popytowych) od łącznej wartości ćwiartki III przedstawiającej dochody będące do dyspozycji podmiotów gospodarczych. Skutkiem nierównowagi są nadwyżkowe zapasy i niewykorzystane zasoby produkcyjne (zwłaszcza pracy) i związane z tym bezrobocie. Można to także interpretować jako nienadążanie opłaty pracy realnej w odniesieniu do wzrostu wydajności i (lub) jako niższą wydajność podatkową podmiotów gospodarczych, co wiąże się z problem ukrywania dochodów czy cenami transferowymi [Czyżewski i Grzelak 2009]. W związku z powyższym interesujące byłoby oszacowanie przybliżonej⁶ nierównowagi poprzez przedstawienie (alternatywnie) w ćwiartce II popytu potencjalnego, tj. efektywnego, oraz niezrealizowanych aspiracji. Ponadto rozwinięcie tego podejścia umożliwiłoby oszacowanie nierównowagi podażowej na poziomie poszczególnych działów. Stanowiłoby to przesłankę precyzyjnie zaadresowanej polityki gospodarczej w zakresie stosowania nieautomatycznych stabilizatorów koniunktury. Wypełnienie IV ćwiartki tablicy przepływów międzygałęziowych pozwoli także na pogłębienie interpretacji odnoszącej się do oceny relacji pieniężnych w gospodarce. Chodzi m.in. o szacunkowe zaprezentowanie (w ćwiartce III) wielkości depozytów i kredytów (lub nowo otwieranych/udzielanych) dla poszczególnych działów gospodarki i na konkretyzację skutków oddziaływania polityki pieniężnej, działalności banków, w tym kreacji pieniądza dla procesów gospodarczych.

Istnieje także możliwość pełnego wykorzystania modelu *input-output* do badań w skali regionalnej. Pozwolą one na odniesienie sytuacji ekonomicznej danego działu w województwie do sytuacji ogólnokrajowej, co umożliwi identyfikację stymulant i destymulant rozwojowych w poszczególnych regionach czy ocenę ekonomiczną konkretnych przedsięwzięć inwestycyjnych. Zastosowanie modelu przepływów międzygałęziowych w badaniach regionalnych pozwoliłoby na pełniejsze dopasowanie lokalnej aktywności gospodarczej do istniejących potrzeb, uwzględniając zewnętrzne (względem regionu) rynki zbytu. Z reguły jednak

⁶ Przybliżonej w tym sensie, że nie sposób jest zaprezentować wszystkich aspektów nierównowagi podażowej, takich chociażby jak niewykorzystane zasoby produkcyjne.

urzędy statystyczne nie publikują tablic *input-output* w układzie regionalnym. Istnieje jednak możliwość samodzielnej dekompozycji danych zawartych w tablicy przepływów na podstawie programowania matematycznego czy współczynników lokalizacji w celu stworzenia stosownych macierzy regionalnych. W literaturze dotyczącej tego zagadnienia jest niewiele przykładów takiego kierunku wykorzystania modelu przepływów międzygałęziowych [Malaga 1992, Tomaszewicz i Trębska 2005, Zawalińska 2009]. Przepływy regionalne konstruowane są przez urzędy statystyczne np. w Finlandii i Hiszpanii. Warto też nadmienić, że istnieje możliwość wykorzystania podejścia *input-output* w modelu CGE (*computable general equilibrium* – model równowagi ogólnej). F. Perali, L. Pironi i G. Standardi [2012], wykorzystując ten model, badali wpływ zniesienia ceł w światowym handlu produktami rolnymi na gospodarki krajów Unii Europejskiej (UE-15). W ten sposób możliwe stało się wskazanie krajów (działów), które relatywnie skorzystają lub stracą na liberalizacji handlu.

Wobec narastającej potrzeby wykorzystania zmiennych środowiskowych do oceny rozwoju gospodarczego pojawiły się też zmodyfikowane koncepcje tablic przepływów międzygałęziowych [Duchin i Steenge 2007, Plich 2002]. Chodzi w tym przypadku o odpowiednie uwzględnienie w macierzach po stronie efektów (*output*) emisji dwutlenku węgla i innych gazów cieplarnianych. W ten sposób ocena funkcjonowania badanych podmiotów staje się pełniejsza, gdyż uwzględnia ujemne efekty zewnętrzne. Modele tego typu z jednej strony wciąż są wyzwaniem, z drugiej jednak są koniecznością w świetle współcześnie obserwowanych tendencji rozwoju gospodarki światowej.

Publikowanie pełnych bilansów międzygałęziowych w Polsce co 5 lat i to z dużym przesunięciem czasowym (ok. 4-letnim) znacznie ogranicza ich praktyczne wykorzystanie, jeśli chodzi o bieżące monitorowanie zjawisk gospodarczych. Uzasadniony jest zatem postulat, aby publikacje te ukazywały się częściej (np. co 3 lata) i szybciej, maksymalnie z 2-letnim opóźnieniem. Ważne jest też, by dawały podstawę do pełnej oceny podziału wytworzonych dochodów i były komplementarne z ujednoliconymi tablicami publikowanymi przez OECD dla 40 krajów [The World... 2012]. Umożliwiłoby to prowadzenie międzynarodowych porównań funkcjonowania gospodarek i ich poszczególnych działów.

5. Podsumowanie

Rozważania zawarte w artykule skłaniają do kilku konkluzji:

1. Wykorzystanie w badaniach podejścia *input-output* stanowi komplementarny instrument oceny i interpretacji zjawisk gospodarczych, umożliwiający pogłębienie analiz makroekonomicznych. Chodzi tu przede wszystkim o inter-

pretację związków, jak również zdiagnozowanie źródeł zróżnicowanego rozwoju pomiędzy działami w obrębie gospodarki danego kraju, czy też zmian pozycji rynkowej badanych działów gospodarki. Nie bez znaczenia jest także wykorzystanie modelu w symulacjach jako narzędzia do oceny wpływu polityki gospodarczej czy szoków egzogenicznych (nowe technologie, odkrycia złóż) na wielkości makroekonomiczne i gałęziowe.

2. Ocena funkcjonowania gospodarki z wykorzystaniem modelu przepływów międzygałęziowych w ujęciu dynamicznym pozwala na ocenę przeobrażeń strukturalnych w ramach danej gospodarki. Zrozumienie ich charakteru i tendencji jest szczególnie przydatne w kreowaniu polityki gospodarczej państwa w zakresie stosowania nieautomatycznych stabilizatorów koniunktury.

3. Istnieje potrzeba rozbudowy modelu przepływów międzygałęziowych o czwartą część i tym samym publikacji danych dotyczących podziału wytworzonych dochodów oraz szerszego uwzględnienia w praktyce gospodarczej cech opisujących sferę pieniężną. Interesująca byłaby próba wykorzystania modelu przepływów do przedstawienia skali nierównowagi podażowej zarówno pomiędzy poszczególnymi częściami tego modelu, tj. dochodu do dyspozycji (III), popytu efektywnego, potencjalnego (II) oraz realnego podziału wytworzonych dochodów (IV), jak i w ramach poszczególnych działów. Istnieją także możliwości konstruowania regionalnych modeli przepływów międzygałęziowych, co dałoby szansę na zwiększenie skuteczności działań w ramach różnych rodzajów polityki regionalnej. Wyzwaniem jest również uwzględnienie zmiennych środowiskowych w analizach *input-output*.

4. Można sądzić, że będziemy mieć do czynienia z dalszym rozwojem modeli wykorzystujących metodę *input-output* w miarę doskonalenia aparatu ekonometrycznego, budowy bardziej złożonych modeli prognostycznych z uwzględnieniem powiązań globalnych, kontekstu środowiskowego czy dezagregacji danych do poziomu regionalnego [Dietzenbacher i in. 2013].

Literatura

- Bess R., Ambargis Z. [2011], *Input-Output Models for Impact Analysis: Suggestions for Practitioners Using RIMS II Multipliers*, Presented at the 50th Southern Regional Science Association Conference, Louisiana.
- Bilans przepływów międzygałęziowych w bieżących cenach bazowych w 2010 r.* [2014], GUS, Warszawa.
- Boratyński J., Przybyliński M., Świeczewska I. [2015], *Metody input-output: wybrane kierunki rozwoju* [w:] *Nauczyciel akademicki wobec nowych wyzwań edukacyjnych*, red. P. Wdowiński, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.
- Czyżewski A. [2011], *Przepływy międzygałęziowe jako makroekonomiczny model gospodarki*, Wydawnictwo Uniwersytetu Ekonomicznego w Poznaniu, Poznań.

- Czyżewski A., Grzelak A. [2007], *The Use of Input-Output to Evaluate the Agriculture Situation in Poland after 1990*, „Management”, vol. 11, nr 2.
- Czyżewski A., Grzelak A. [2009], *Możliwości oceny rozwoju rolnictwa w warunkach globalnych z zastosowaniem tabeli przepływów międzygałęziowych*, „Roczniki Naukowe SERiA”, t. XI, z. 2.
- Czyżewski A., Grzelak A. [2012], *Możliwości wykorzystania statystyki bilansów przepływów międzygałęziowych*, „Przegląd Statystyczny”, nr 1.
- Czyżewski A., Helak K. [1991], *Przekształcenia w kompleksie gospodarki żywnościowej w Polsce*, „Wież i Rolnictwo”, nr 3.
- Czyżewski A., Smędzik-Ambroży K. [2015], *The Efficiency of the Agricultural Sector in Poland in the Light Output-Input Model*, „Management”, vol. 19, nr 1, <https://doi.org/10.1515/manment-2015-0009>.
- Czyżewski B. [2013], *Renty ekonomiczne w gospodarce żywnościowej w Polsce*, PWE, Warszawa.
- Dietzenbacher E., Lenzen M., Los B., Guan D., Lahr M.L., Sancho F., Suh S., Yang C. [2013], *Input-Output Analysis: The Next 25 Years*, „Economic Systems Research”, vol. 25, nr 4, <https://doi.org/10.1080/09535314.2013.846902>.
- Duchin F., Steenge A. [2007], *Mathematical Models in Input-Output Economics*, Rensselaer Polytechnic Institute, Working Papers in Economics, <http://www.economics.rpi.edu/workingpapers/rpi0703.pdf> (data dostępu: grudzień 2016).
- Grzelak A. [2006], *Powiązania rolnictwa z otoczeniem z perspektywy bilansów przepływów międzygałęziowych w warunkach transformacji systemowej*, „Polityka Gospodarcza”, nr 13.
- Handbook of Computable General Equilibrium Modeling* [2013], eds P.B. Dixon, D. Jorgenson, vol. 1A and 1B, North-Holland.
- Lange O. [1961], *Teoria reprodukcji i akumulacji*, PWN, Warszawa.
- Leontief W. [1936], *Quantitative Input and Output Relations in the Economic System of the United States*, „The Review of Economics and Statistics”, vol. 18, nr 3 (August), <https://doi.org/10.2307/1927837>.
- Leontief W. [1963], *Studia nad strukturą gospodarki amerykańskiej*, PWN, Warszawa.
- Lonc T. [1985], *Związki rolnictwa z gospodarką narodową na początku lat 1980-tych*, „Zagadnienia Ekonomiki Rolnej”, nr 6.
- Malaga K. [1992], *Struktura produkcyjna gospodarki żywnościowej Wielkopolski w świetle przepływów międzygałęziowych* [w:] *Gospodarka żywnościowa w Polsce i regionie*, red. A. Czyżewski, PWE, Warszawa.
- Marks K. [1955], *Kapitał*, t. 2, Książka i Wiedza, Warszawa.
- Mrówczyńska-Kamińska A. [2015], *Gospodarka żywnościowa w krajach Unii Europejskiej. Kierunki rozwoju, przepływy i współzależności*, Wydawnictwo Uniwersytetu Przyrodniczego w Poznaniu, Poznań.
- Opracowanie Biura Ekonomicznego NBP* [1958], NBP, Warszawa.
- Perali F., Pieroni L., Standardi G. [2012], *World Tariff Liberalization in Agriculture: An Assessment Using a Global CGE Trade Model for EU15 Regions*, „Journal of Policy Modeling”, vol. 34, nr 2, <https://doi.org/10.1016/j.jpolmod.2012.01.002>.
- Plich M. [2002], *Budowa i zastosowanie wielosektorowych modeli ekonomiczno-ekologicznych*, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.
- Quesnay F. [1928], *Pisma wybrane*, tłum. B.J. Pietkiewiczówna, Gebethner i Wolf, Warszawa.

- Samuelson P. [1951], *Abstract of a Theorem Concerning Substitutability in Open Leontief Model* [w:] *Activity Analysis in Production and Allocation*, ed. T. Koopmans, John Wiley and Sons, New York.
- Solow R. [1951], *Competitive Valuation in a Dynamic Input-Output System*, „Econometrica”, vol. 27, nr 1.
- Stankiewicz W. [2007], *Historia myśli ekonomicznej*, PWE, Warszawa.
- Świczewska I. [2014], *The Externalities of Enterprises' Innovative Activity – An Input-Output Approach*, „Folia Oeconomica Stetinensia”, vol. 13, nr 3, <https://doi.org/10.2478/fofi-2013-0023>.
- Świczewska I. Tomaszewicz Ł. [2012], *Rola innowacji w procesie wzrostu efektywności polskiej gospodarki. Ujęcie gałęziowe* [w:] *Modele i prognozy w ekonomii i finansach*, red. A. Jakimowicz, Wydawnictwo Naukowe PWN, Warszawa.
- Tomaszewicz Ł. [1994], *Metody analizy input-output*, PWE, Warszawa.
- Tomaszewicz Ł., Trębska J. [2005], *Regional and Interregional Input-Output Tables for Poland* [w:] *Modeling Economies in Transition*, eds W. Welfe, P. Wdowiński, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.
- Walras L. [1926], *Elements d'economie politique pure*, R. Pichon et R. Durnad-Auzias Editeurs, Paris.
- Wolff E. [2011], *Spillovers, Linkages and Productivity Growth in the US Economy, 1958 to 2007*, NBER Working Paper, No 16864.
- The World Input-Output Database (WIOD): Contents, Sources and Methods* [2012], ed. M. Timmer, WIOD Working Paper No. 10, <http://www.wiod.org/publications/papers/wiod10.pdf> (data dostępu: grudzień 2016).
- Woś A. [1973], *Rolnictwo w bilansie przepływów międzygałęziowych*, „Zagadnienia Ekonomiki Rolnictwa”, nr 1.
- Zawalińska K. [2009], *Regionalne efekty wsparcia Unii Europejskiej dla rozwoju obszarów wiejskich*, IRWiR, Warszawa.

The Use of an Input-output Model for the Macroeconomic Evaluation of the Economy

(Abstract)

The objective of the article is to present the possibilities of using an input-output model in macroeconomic evaluations of the economy taking into account past experience and suggestions for the future. During the study we found that the input-output model is a useful tool for evaluating and interpreting economic phenomena. It can be used primarily to evaluate relationships and to diagnose the sources of diversified development between branches within a given economy. There is a need to expand the interpretation of the input-output model by adding a fourth part. This would entail publishing data on the distribution of generated income. Further, it would be interesting to attempt to use this model to present supply inequality while taking into account the environmental context of the macroeconomic analysis.

Keywords: the input-output model, macroeconomics, economy, the creation and distribution of national income.

| Tadeusz Grabiński

Uproszczona metoda delimitacji wektorowej

Streszczenie

W artykule przedstawiono propozycję graficznej procedury porządkowania i grupowania obiektów opisanych pojedynczą cechą ujętą w formie szeregu przekrojowego, czasowego lub przekrojowo-czasowego. Metoda ta nawiązuje do klasycznej metody porządkowania opracowanej ponad 100 lat temu przez Jana Czekanowskiego. Zaletami proponowanej procedury są mała pracochłonność, intuicyjność i możliwość realizacji za pomocą arkusza kalkulacyjnego z zastosowaniem kilku elementarnych funkcji. Metodę wykorzystano do analizy taksonomicznej zbioru 28 krajów Unii Europejskiej opisanych trzema cechami: stopa bezrobocia, zagrożenie ubóstwem oraz długość życia, odnoszącymi się do dwóch lat – 2010 oraz 2015 r.

Słowa kluczowe: taksonometria, porządkowanie, grupowanie, delimitacja, szeregi przekrojowo-czasowe.

Klasyfikacja JEL: C38.

1. Wprowadzenie

W artykule przedstawiono propozycję procedury delimitacji (porządkowania i grupowania) obiektów wymagającej znajomości kilku podstawowych funkcji arkusza kalkulacyjnego oraz niewielkiego nakładu czasu pracy. Jej dodatkową zaletą jest intuicyjność i możliwość śledzenia rezultatów we wszystkich etapach.

Podstawą analizy jest wektor danych źródłowych wyrażonych w tych samych jednostkach pomiarowych (strumieniowych lub wskaźnikowych). Może to być

szereg przekrojowy (zbiór obiektów opisanych jedną cechą w jednym okresie), szereg czasowy (pojedynczy obiekt opisany jedną cechą w kilku okresach) lub szereg przekrojowo-czasowy (zbiór obiektów opisanych jedną cechą w kilku okresach).

Nie jest to więc metoda *stricte* z obszaru taksonometrii wielowymiarowej, gdzie analizie podlega zbiór cech. Tym niemniej podstawą delimitacji w prezentowanej metodzie może być także cecha syntetyczna uzyskana z agregacji kilku cech cząstkowych metodami wielowymiarowej analizy porównawczej. Dlatego też delimitacja oparta na wektorze danych również może mieć charakter wielowymiarowy. Innym sposobem uzyskania wielowymiarowości delimitacji wektorowej jest przeprowadzenie analizy dla kilku cech i porównanie uzyskanych wyników.

2. Charakterystyka danych źródłowych

Omawianą metodę delimitacji wektorowej wykorzystano do analizy taksonometrycznej zbioru 28 krajów Unii Europejskiej opisanych trzema cechami: stopa bezrobocia (B), zagrożenie ubóstwem (U) oraz długość życia (D), odnoszącymi się do dwóch lat (2010 i 2015).

Dane zaczerpnięto z *Rocznika statystyki międzynarodowej 2016* (<http://stat.gov.pl/obszary-tematyczne/roczniki-statystyczne/roczniki-statystyczne/rocznik-statystyki-miedzynarodowej-2016,10,5.html>). W tabeli 1 podano wykaz krajów podlegających analizie oraz przyjęte dla nich skróty wykorzystywane w diagramach.

Tabela 1. Wykaz analizowanych krajów

N	Kraj	Skrót	N	Kraj	Skrót	N	Kraj	Skrót	N	Kraj	Skrót
1	Austria	AT	8	Estonia	EE	15	Irlandia	IE	22	Polska	PL
2	Belgia	BE	9	Grecja	EL	16	Włochy	IT	23	Portugalia	PT
3	Bułgaria	BG	10	Hiszpania	ES	17	Litwa	LT	24	Rumunia	RO
4	Cypr	CY	11	Finlandia	FI	18	Luksemburg	LU	25	Szwecja	SE
5	Czechy	CZ	12	Francja	FR	19	Łotwa	LV	26	Słowenia	SI
6	Niemcy	DE	13	Chorwacja	HR	20	Malta	MT	27	Słowacja	SK
7	Dania	DK	14	Węgry	HU	21	Holandia	NL	28	Wielka Brytania	UK

Źródło: opracowanie własne.

Na podstawie tych informacji wyznaczono następujące cechy wynikowe:

1) statyczne wartości standaryzowane ustalane odrębnie dla każdego roku i każdej cechy z wykorzystaniem funkcji Excela *procent.pozycja*, zgodnie z którą

maksymalnej wartości w zbiorze podlegającym standaryzacji przypisuje się 1 (100%), natomiast wartości najmniejszej – 0 (0%)¹. W przypadku długości życia standaryzacji dokonywano według zmienionej formuły [1-procent.pozycja], dzięki czemu wszystkie 3 cechy uzyskały jednakowy status, wyrażający się w ujednoliconej ocenie ich wielkości, zgodnie z regułą: im niższe są wartości każdej z tych cech, tym korzystniej należy ocenić dany kraj;

2) dynamiczne wartości standaryzowane wyznaczone według analogicznych zasad jak wartości statyczne, z tą różnicą, że standaryzacji nie dokonuje się odrębnie dla każdego roku, lecz dla połączonych wartości danej cechy z lat 2010 i 2015. Oznacza to, że wartości krańcowe (0–1) przypisywane są krajom z najbardziej i najmniej korzystną wartością danej cechy spośród jej wartości z tych lat;

3) wartości zagregowane zdefiniowane jako średnie arytmetyczne z wartości standaryzowanych z obydwu lat w ujęciu statycznym [%B], [%U], [%D] oraz dynamicznym [%%B], [%%U], [%%D];

4) statyczne wskaźniki syntetyczne jako średnie arytmetyczne ze statycznych wartości standaryzowanych dla wszystkich cech B-U-D w trzech wariantach: [%X10] – dla 2010 r., [%X15] – dla 2015 r. oraz [%X] – dla obydwu lat;

5) dynamiczne wskaźniki syntetyczne jako średnie arytmetyczne z dynamicznych wartości standaryzowanych dla wszystkich cech B-U-D, także w trzech wariantach: [%%X10] – dla 2010 r., [%%X15] – dla 2015 r. oraz [%%X] – dla obydwu lat.

Tabela 2. Podstawowe parametry opisowe zmiennych wyjściowych

Parametr	Stopa bezrobocia			Zagrożenie ubóstwem			Długość życia		
	B10	B15	B	U10	U15	U	D10	D15	D
Min	4,4	4,9	5,2	14,4	14,0	14,2	70,7	74,5	72,7
Q1	6,3	6,8	6,6	18,8	18,4	18,6	75,1	77,7	76,3
Q2	7,7	8,9	8,8	22,5	23,5	23,2	79,6	81,3	80,5
Q3	10,1	11,5	11,2	27,7	28,8	27,9	80,0	81,9	81,0
Max	16,4	26,3	20,3	49,2	41,3	45,3	81,6	83,3	82,4
Średnie arytmetyczne	8,5	10,4	9,5	24,6	24,3	24,4	77,7	79,9	78,8
Odchylenia standardowe	3,05	5,22	3,67	8,35	6,96	7,50	3,36	2,91	3,11
V1 = odch./śred. (w %)	36	50	39	34	29	31	4	4	4
V2 = (Q3–Q1)/Q2 (w %)	49	53	53	40	44	40	6	5	6

Źródło: opracowanie własne.

¹ W nowszych wersjach programu Excel (2013, 2016) zamiast funkcji *procent.pozycja* należy skorzystać z funkcji *proc.poz.przedz.otw* lub *proc.poz.przedz.zamk*.

Liczbę zmiennych można jeszcze powiększyć, np. definiując zmienne syntetyczne o zróżnicowanych wagach lub sięgając po inne metody standaryzacji bądź poszerzając liczbę lat. W tabeli 2 przytoczono podstawowe parametry opisowe wyjściowych zmiennych (kwartyle, średnie arytmetyczne, odchylenia standardowe, współczynniki zmienności). Warto tu zwrócić uwagę na niewielkie (4–6%) zróżnicowanie krajów pod względem długości życia w odróżnieniu od pozostałych dwóch zmiennych, dla których współczynniki zmienności zawierają się w przedziale 30–50%.

3. Algorytm delimitacji wektorowej

Proponowana procedura delimitacji nawiązuje do najstarszej, liczącej ponad 100 lat, taksonometrycznej metody Czekanowskiego (por. [Czekanowski 1913, Piasecki 1971]). Wspólną cechą obydwu metod jest wykorzystanie diagraficznego macierzy odległości, na podstawie której można ustalić, czy istnieją grupy podobnych obiektów, a jeżeli tak, to czy możliwa jest identyfikacja tych grup. Prezentowana metoda różni się od metody Czekanowskiego procedurą porządkowania macierzy odległości. Zamiast dokonywać sekwencyjnych przestawień wierszy i kolumn macierzy odległości, szukając optymalnej wartości funkcji kryterium określającej stopień uporządkowania wartości w macierzy², przyjmuje się, że porządek wierszy i kolumn w macierzy odległości wynika z wybranej zmiennej delimitacyjnej.

Grupy obiektów ustala się, zaokrąglając wartości zmiennej delimitacyjnej do coraz to większych wartości. Drugim możliwym rozwiązaniem jest zaokrąglanie elementów macierzy odległości wyznaczonej na podstawie zmiennej delimitacyjnej. W utworzonej aplikacji przewidziano wykorzystanie obydwu możliwości zaokrągleń, zarówno wyjściowych wartości zmiennej delimitacyjnej, jak i wyznaczonych na jej podstawie elementów macierzy odległości.

Poniżej przedstawiono kolejne kroki w proponowanej metodzie.

1. Punktem wyjścia analizy jest wektor wartości dowolnej zmiennej delimitacyjnej. Może to być w szczególności zmienna syntetyczna reprezentująca zbiór innych zmiennych. Może to być zmienna w postaci szeregu czasowego. W tym przypadku uzyskuje się delimitację okresów, co określane jest jako periodyzacja. Wyjściowa zmienna nie może mieć zerowej wariancji. Ponadto powinna być wyrażona w jednostkach pozwalających na wykonywanie podstawowych operacji algebraicznych, np. wyznaczanie różnic lub ilorazów wartości.

2. Wartości zmiennej delimitacyjnej standaryzuje się za pomocą formuły oddającej relacje strukturalne wartości oryginalnych. Zakładając, że obliczenia wykonywane są w arkuszu kalkulacyjnym, proponuje się w tym celu wykorzy-

² Przykłady takich algorytmów można znaleźć w pracach: [Kozłowski 1972, Sołtysiak 1977a, b].

stać funkcję *procent.pozycja*, a w nowszych wersjach Excela funkcję *proc.poz.przedz.zamk*.

3. Wartości standaryzowanej zmiennej delimitacyjnej porządkuje się w kolejności rosnącej lub malejącej. Występuje tu problem identycznych wartości zmiennej, których pojawianie się nie pozwala wykonać niektórych operacji numerycznych przeprowadzanych w proponowanej metodzie. Dlatego też dokonuje się modyfikacji wszystkich wartości zmiennej zgodnie ze wzorem:

$$x'_i = x_i + x_i \cdot \text{los}()/1000 \quad (i = 1, 2, \dots, n), \quad (1)$$

gdzie *los()* to generator liczb losowych z przedziału (0, 1). Dzięki tej operacji w zbiorze wartości zmiennej delimitacyjnej nie ma identycznych realizacji.

4. Generuje się kwadratową macierz odległości o wymiarach $n \times n$, w której umieszcza się moduły różnic pomiędzy wszystkimi wartościami uporządkowanego wektora realizacji zmiennej delimitacyjnej:

$$d_{ij} = |x'_i - x'_j| \quad (i, j = 1, \dots, n). \quad (2)$$

5. Elementy macierzy modułów różnic formatuje się za pomocą funkcji Excela, zastępując wartości (2) symbolami graficznymi identyfikującymi rząd ich wielkości (formatowane warunkowe). W wyniku tej operacji na głównej przekątnej macierzy odległości pojawiają się symbole graficzne przypisane najniższym wartościom różnic (na głównej przekątnej będą to wartości zerowe). Im bardziej oddalone od głównej przekątnej są komórki macierzy (2), tym większe są tam wartości, co jest uwidocznione za pomocą ikon graficznych w różnych kolorach lub kształtach.

6. Grupy podobnych obiektów tworzą kwadraty składające się z ikon przypisanych najmniejszym wartościom modułów różnic. Jeżeli wynikowa macierz nie pozwala zidentyfikować wyraźnych ugrupowań (kwadratów identycznych ikon), to przeprowadza się procedurę stopniowego zaokrąglania wartości zmiennej delimitacyjnej (1). Do tego celu służą funkcje arkusza kalkulacyjnego Excel: *zaokr.w.góre*, *zaokr.w.dół*.

7. W analizach wykorzystano funkcje *zaokr.w.dół* z parametrem przyjmującym kolejno wartości: 0,1%, 1%, 2%, 3%, 4% itd. Pierwszy parametr [0,1%] generuje macierz podobieństwa na podstawie oryginalnych wartości zmiennej delimitacyjnej zaokrąglonych do 1 promila. Drugi parametr [1%] zaokrągla wartości zmiennej do 1%, kolejny parametr [2%] – do wielokrotności 2% itd. Na przykład przy zastosowaniu funkcji *zaokr.w.dół* z parametrem [2%] wektor wartości [0,13; 0,14; 0,17; 0,20] przyjmuje postać [0,12; 0,14; 0,16; 0,20], natomiast z parametrem [3%] wektor wynikowy ma składowe [0,12; 0,12; 0,15; 0,18].

8. Drugim narzędziem poszukiwania rozłącznych ugrupowań obiektów jest zaokrąglanie elementów macierzy odległości (2) za pomocą tej samej funkcji

zaokr.w.dół. W praktyce dobrze jest rozpocząć analizę od wartości oryginalnych, przyjmując obydwa parametry zaokrągleń (dla wartości zmiennej delimitacyjnej oraz dla elementów macierzy odległości) na poziomie 0,1%. Następnie należy stopniowo zmieniać ten parametr od 1% do 5–7%, najpierw dla elementów macierzy odległości, a następnie dla zmiennej delimitacyjnej.

9. Wprowadzanie do arkusza kolejnych parametrów określających rząd zaokrągleń generuje macierze o innej konfiguracji ikon. Wybór finalnej konfiguracji, a tym samym rezultat delimitacji proponuje się ustalać z zastosowaniem następującej procedury:

– na podstawie uporządkowanych rosnąco zaokrąglonych wartości zmiennej delimitacyjnej wyznacza się ciąg różnic pomiędzy sąsiednimi jej wartościami:

$$g_i = x'_i - x'_{i-1} \quad (i = 2, 3, \dots, n). \quad (3)$$

W odrębnej kolumnie zapisuje się niezerowe wartości różnic g_i , eliminując ich wartości zerowe;

– ustala się minimalną wielkość różnic g_i (po wyeliminowaniu wartości zerowych), a następnie w kolejnej kolumnie zapisuje wartości tych różnic większe od minimalnej;

– operacje z poprzedniego punktu powtarza się aż do momentu, w którym wszystkie różnice g_i zostaną wyeliminowane;

– wiersze, w których pojawi się najwięcej wartości różnic, wyznaczają delimitację zbioru obiektów. Są to miejsca, w których obserwuje się największe różnice pomiędzy sąsiednimi elementami uporządkowanych wartości zmiennej delimitacyjnej.

10. Druga metoda ustalania liczby ugrupowań opiera się na sumach komórek zawierających wartości zaliczone do kategorii wskazującej na maksymalne podobieństwo obiektów, wyznaczonych dla każdego wiersza. Miejsca podziału obiektów na grupy są zlokalizowane w wierszach, w których liczba maksymalnych podobieństw jest najmniejsza. Podobnie jak w poprzedniej metodzie można tu sekwencyjnie eliminować maksymalne liczby maksymalnych podobieństw, dochodząc do momentu, w którym pozostaną tylko wielkości minimalne.

4. Wyniki delimitacji krajów Europy

Przedstawioną procedurę delimitacji wykorzystano do grupowania krajów Europy, wykorzystując omówione dane źródłowe. Wyniki analizy zebrano w tabelach 3–10, zawierających diagramy dla:

– wyjściowych wartości zmiennej delimitacyjnej bez zaokrągleń, ze standaryzacją dynamiczną (tabela 3) oraz statyczną (tabela 4),

– zaokrąglonych wartości zmiennej delimitacyjnej na poziomie 5%, ze standaryzacją dynamiczną (tabela 5) oraz statyczną (tabela 6),

- zaokrąglonych wartości zmiennej delimitacyjnej na poziomie 20%, ze standaryzacją dynamiczną (tabela 7) oraz statyczną (tabela 8),
- zaokrąglonych wartości zmiennej delimitacyjnej na poziomie 30%, ze standaryzacją dynamiczną (tabela 9) oraz statyczną (tabela 10).

W każdym diagramie pogrubionymi krawędziami oznaczono grupy krajów o podobnych wartościach zmiennej delimitacyjnej. Podział krajów na grupy przeprowadzany jest tak, aby w kwadratach wzdłuż głównej przekątnej znajdowało się jak najwięcej komórek z identyfikatorami wskazującymi na maksymalne podobieństwo krajów. Elementom macierzy odległości przypisano różne kolory. Wartościom mierników wskazujących na najmniejszy poziom podobieństw klasyfikowanych obiektów przyporządkowany jest kolor zielony, kolor pomarańczowy wskazuje na przeciętny poziom podobieństwa obiektów, a kolor czerwony na poziom wysoki. Mierniki odległości pomiędzy obiektami o największym stopniu podobieństwa oznaczone są kolorem czarnym. Z definicji są to elementy głównej przekątnej macierzy odległości wskazujące na podobieństwo obiektu do samego siebie. Konfiguracja geometryczna elementów macierzy oznaczonych czarnym kolorem wskazuje na stopień separowalności analizowanego zbioru obiektów. Konfiguracje w formie dużych kwadratów pozwalają podzielić zbiór obiektów na odpowiednią liczbę rozłącznych, wyraźnie separowalnych grup. Natomiast konfiguracja w postaci wstęgi wzdłuż głównej przekątnej macierzy odległości oznacza sytuację, w której trudno jest wydzielić grupy obiektów podobnych do siebie i jednocześnie niepodobnych do obiektów zaliczonych do innych grup.

Przy ustalaniu grup krajów pomocne są zamieszczone po prawej stronie diagramów:

- maksymalne wartości różnic zmiennej delimitacyjnej,
- liczby komórek zawierających identyfikatory maksymalnego podobieństwa.

Ponadto dla celów porównawczych dokonano delimitacji krajów Europy klasyczną metodą Czekanowskiego. Wykorzystano w tym celu program do grupowania danych MaCzek, wersja 3.3.44, udostępniony na stronie <http://www.antropologia.uw.edu.pl/MaCzek>. Podstawą analizy w tym przypadku były wektory zawierające statyczne (%X) oraz dynamiczne (%%X) wskaźniki syntetyczne wyznaczone na podstawie danych z obydwu lat. Wyniki delimitacji zebrano w tabelach 11 i 12. Znajdują się tu macierze Czekanowskiego dla danych:

- nieuporządkowanych (według alfabetycznej kolejności krajów) – tabela 11, cecha [%%X],
- uporządkowanych za pomocą algorytmu uproszczonego – tabela 11, cecha [%%X],
- uporządkowanych za pomocą algorytmu genetycznego – tabela 12, cecha [%X],
- uporządkowanych za pomocą algorytmu UMCzek – tabela 12, cecha [%X].

Jak można zauważyć, w obydwu wariantach wyniki grupowania są bardzo zbliżone, pomimo różnego uporządkowania krajów. Grupę krajów o najkorzystniejszych wartościach zmiennych B-U-D tworzą: Holandia, Austria, Luksemburg, Szwecja i Czechy. Druga grupa krajów o nieco mniej korzystnych wartościach zmiennych diagnostycznych obejmuje 8 krajów: Francję, Niemcy, Wielką Brytanię, Danię, Finlandię, Belgię, Malte i Słowenię. W trzeciej grupie, obejmującej 7 krajów: Włochy, Irlandia, Hiszpania, Portugalia, Polska, Estonia, Cypr – zmienne diagnostyczne kształtują się na poziomie poniżej przeciętnego. W ostatniej grupie znajdują się kraje o najmniej korzystnych wartościach zmiennych. W odróżnieniu od poprzednich grup, kolejność krajów w tej grupie jest identyczna dla obydwu zmiennych: Słowacja, Grecja, Rumunia, Węgry, Chorwacja, Bułgaria, Litwa i Łotwa.

Wyniki grupowania krajów metodą uproszczoną zebrano w tabeli 13, w której także zamieszczono wyniki uzyskane klasyczną metodą Czekanowskiego (ostatnie dwie kolumny). Kolejne segmenty tej tabeli zawierają wyniki otrzymane w uproszczonej metodzie delimitacji dla zmieniającego się poziomu zaokrągleń zmiennej delimitacyjnej począwszy od 30%, gdzie wartości zmiennej zostały zaokrąglone do [0; 0,3; 0,6; 0,9]. Następne poziomy zaokrągleń to 20% z wartościami [0; 0,2; 0,4; 0,6; 0,8] oraz 5% (wartości od 0,10 do 0,90 z odstępem co 0,05). W ostatnim segmencie tej tabeli przytoczono wyniki grupowania na podstawie oryginalnych (niezaokrąglonych) wartości zmiennej diagnostycznej. Wyróżnione grupy krajów o podobnych wartościach zmiennej syntetycznej oznaczono w tabeli różnymi kolorami. Kolorem zielonym oznaczono najmniejsze wartości zmiennej syntetycznej, natomiast kolorem czerwonym – wartości największe. Elementy macierzy zawierające pośrednie wartości zmiennej syntetycznej oznaczono pozostałymi kolorami – od żółtego do niebieskiego. W poszczególnych segmentach tabeli wynikowa liczba grup krajów podobnych (a tym samym i kolorów) jest różna – od 3 grup w pierwszym segmencie (30-procentowy poziom zaokrągleń) do 6 grup w segmencie zawierającym wyniki analizy bez zaokrągleń zmiennej diagnostycznej.

Analizując uzyskane wyniki, należy zwrócić uwagę na wysoką zgodność delimitacji w wariantach opartych na zmiennych standaryzowanych dynamicznie ($%%X$) i statycznie ($%X$). Podobnie jak w przypadku klasycznej metody Czekanowskiego w delimitacji opartej na danych wyjściowych (bez zaokrągleń) kolejność krajów w obydwu wariantach (standaryzacja statyczna i dynamiczna) nie jest identyczna, ale w ramach grup o zbliżonych wartościach zmiennej syntetycznej uzyskuje się pełną zgodność przypisanych do nich krajów.

W przypadku delimitacji opartej na danych zaokrąglonych (do 5%, 20% i 30%) obserwuje się kilka rozbieżnych przyporządkowań. Dotyczy to:

- Finlandii w wariancie z danymi zaokrąglonymi do 5%,
- Słowenii, Wielkiej Brytanii i Hiszpanii w wariancie z danymi zaokrąglonymi do 20%,
- Danii, Malty i Łotwy w wariancie z danymi zaokrąglonymi do 30%.

Tabela 13. Porównanie wyników delimitacji krajów Europy

Lp.	%X	0.3	%%X	0.3	%X	0.2	%%X	0.2	%X	0.05	%%X	0.05	%X	0.0	%%X	0.0	CZ%X	CZ%%X
1	AT	0.0	AT	0.0	NL	0.0	NL	0.0	NL	0.10	NL	0.15	NL	0.14	NL	0.16	NL	NL
2	CZ	0.0	CZ	0.0	AT	0.0	AT	0.0	AT	0.15	AT	0.15	AT	0.17	AT	0.20	AT	AT
3	LU	0.0	DK	0.0	CZ	0.2	CZ	0.2	SE	0.20	LU	0.20	LU	0.20	LU	0.21	SE	LU
4	MT	0.0	LU	0.0	DE	0.2	DE	0.2	LU	0.20	SE	0.20	LU	0.22	SE	0.24	LU	SE
5	NL	0.0	NL	0.0	DK	0.2	DK	0.2	CZ	0.25	CZ	0.25	CZ	0.28	CZ	0.28	CZ	CZ
6	SE	0.0	SE	0.0	FI	0.2	FI	0.2	MT	0.25	DK	0.25	DK	0.30	DK	0.29	MT	DK
7	BE	0.3	BE	0.3	FR	0.2	FR	0.2	FR	0.30	DE	0.30	FR	0.30	DE	0.31	FR	MT
8	CY	0.3	CY	0.3	LU	0.2	LU	0.2	DE	0.30	MT	0.30	DE	0.31	MT	0.31	DE	DE
9	DE	0.3	DE	0.3	MT	0.2	MT	0.2	DK	0.30	FR	0.30	DK	0.32	FR	0.34	DK	FI
10	DK	0.3	FI	0.3	SE	0.2	SE	0.2	FI	0.35	FI	0.30	FI	0.36	FI	0.34	FI	FR
11	FI	0.3	FR	0.3	UK	0.4	SI	0.2	UK	0.40	SI	0.35	UK	0.40	SI	0.37	UK	SI
12	FR	0.3	IT	0.3	SI	0.4	UK	0.2	SI	0.40	UK	0.35	SI	0.41	UK	0.38	BE	UK
13	IT	0.3	MT	0.3	BE	0.4	BE	0.4	BE	0.40	BE	0.40	BE	0.41	BE	0.41	SI	BE
14	SI	0.3	SI	0.3	IT	0.4	IT	0.4	IT	0.40	IT	0.45	IT	0.42	IT	0.46	IT	IT
15	UK	0.3	UK	0.3	CY	0.4	CY	0.4	CY	0.45	CY	0.45	CY	0.46	CY	0.48	CY	CY
16	BG	0.6	BG	0.6	ES	0.4	EE	0.6	ES	0.55	EE	0.60	ES	0.57	IE	0.60	ES	IE
17	EE	0.6	EE	0.6	EE	0.6	EL	0.6	IE	0.60	ES	0.60	IE	0.63	ES	0.61	IE	ES
18	EL	0.6	EL	0.6	EL	0.6	ES	0.6	EE	0.60	IE	0.60	PL	0.64	PT	0.62	PL	PT
19	ES	0.6	ES	0.6	HR	0.6	HR	0.6	PL	0.60	PL	0.60	EE	0.64	PL	0.63	EE	EE
20	HR	0.6	HR	0.6	HU	0.6	HU	0.6	PT	0.65	PT	0.60	SK	0.65	EE	0.63	PT	PL
21	HU	0.6	HU	0.6	IE	0.6	IE	0.6	SK	0.65	SK	0.65	SK	0.65	SK	0.65	SK	SK
22	IE	0.6	IE	0.6	PL	0.6	PL	0.6	EL	0.65	EL	0.65	EL	0.70	EL	0.66	EL	EL
23	LT	0.6	LT	0.6	PT	0.6	PT	0.6	RO	0.70	RO	0.65	RO	0.70	RO	0.70	RO	RO
24	PL	0.6	PL	0.6	RO	0.6	RO	0.6	HU	0.70	HU	0.70	HU	0.70	HU	0.72	HU	HU
25	PT	0.6	PT	0.6	SK	0.6	SK	0.6	HR	0.75	HR	0.75	HR	0.78	HR	0.78	HR	HR
26	RO	0.6	RO	0.6	BG	0.8	BG	0.8	BG	0.80	BG	0.80	BG	0.84	BG	0.82	BG	BG
27	SK	0.6	SK	0.6	LT	0.8	LT	0.8	LT	0.85	LT	0.80	LT	0.87	LT	0.85	LT	LT
28	LV	0.9	LV	0.6	LV	0.8	LV	0.8	LV	0.90	LV	0.85	LV	0.90	LV	0.88	LV	LV

Źródło: opracowanie własne.

Tabela 14. Zgodność uporządkowań krajów wynikających ze zmiennych syntetycznych %X oraz %%X

Kraj	Miejsce krajów według różnych zmiennych				Moduły różnic miejsc				
	U1	U2	U3	U4	U5	U6	U7	U8	U9
NL	1	1	1	1					
AT	2	2	2	2					
SE	3	4	3	4	1	1	1		
LU	4	3	4	3	1	1	1		
CZ	5	5	5	5					
MT	6	8	6	7	2	2	1		1
FR	7	9	7	10	3	2	3		1
DE	8	7	8	8	1	1			1
DK	9	6	9	6	3	3	3		
FI	10	10	10	9	1		1		1
UK	11	12	11	12	1	1	1		
SI	12	11	13	11	2	1	2	1	
BE	13	13	12	13	1		1	1	
IT	14	14	14	14					
CY	15	15	15	15					
ES	16	17	16	17	1	1	1		
IE	17	16	17	16	1	1	1		
PL	18	19	18	20	2	1	2		1
EE	19	20	19	19	1	1			1
SK	20	21	21	21	1	1		1	
PT	21	18	20	18	3	3	2	1	
EL	22	22	22	22					
HU	23	24	24	24	1	1		1	
RO	24	23	23	23	1	1		1	
HR	25	25	25	25					
BG	26	26	26	26					
LT	27	27	27	27					
LV	28	28	28	28					
Liczba krajów:					U5	U6	U7	U8	U9
zajmujących identyczne miejsca					10	12	15	22	22
różniących się zajmowanymi miejscami					18	16	13	6	6
Liczba krajów różniących się:									
1 miejscem					12	12	8	6	6
2 miejscami					3	2	3		
3 miejscami					3	2	2		
Suma modułów różnic miejsc					27	22	20	6	6

Oznaczenia parametrów:

U1 – %X; U2 – %%X; U3 – CZ%X; U4 – CZ%%X; U5 – max(U1:U4) – min(U1:U4);

U6 – moduł (U1 – U2); U7 – moduł (U3 – U4); U8 – moduł (U1 – U3); U9 – moduł (U2 – U4).

Źródło: opracowanie własne.

Identyfikatory nazw krajów o rozbieżnych przyporządkowaniach oznaczono w tabeli 13 pogrubioną czcionką.

Nieco inaczej jest, jeżeli chodzi o zgodność grup krajów w przekroju wariantów analizy różniących się stopniem zaokrąglenia danych. W tym przypadku pomimo generalnej zgodności uporządkowań krajów (por. tabela 14) uzyskuje się różne konfiguracje i różne liczby grup krajów. W wariacie zaokrąglenia 30% otrzymano 3 ugrupowania, w wariacie z danymi zaokrąglonymi do 20% – 5 grup, natomiast w pozostałych dwóch wariantach (zaokrąglenia 5% oraz bez zaokrąglenia) pojawia się podział na 6 grup. Z klasycznej metody Czekanowskiego wynika jeszcze inna konfiguracja z 4 grupami.

Z informacji podanych w tabeli 14 wynika, że największa zgodność uporządkowań ma miejsce w przypadku kolejności wynikającej z klasycznej metody Czekanowskiego oraz zmiennych %X i %%X (dwie ostatnie kolumny tej tabeli). Jak można zauważyć, dla 22 krajów ich kolejność w tych metodach delimitacji się pokrywa. Nieco większe rozbieżności wystąpiły w uporządkowaniach krajów według zmiennych standaryzowanych (statycznie i dynamicznie). Liczba krajów o zgodnych pozycjach spada tu z 22 do 12–15. Zmiany w uporządkowaniach miały miejsce m.in. dla Francji, Danii, Portugalii, Malty, Słowenii i Polski.

Ogólnie biorąc, różnice te są niewielkie, a ponadto w trakcie grupowania większość z nich zanika, gdyż do danej grupy są zaliczane kraje o zbliżonych pozycjach wynikających ze zmiennych delimitacyjnych.

5. Wnioski

Wyniki zastosowania uproszczonej metody delimitacji wskazują na wysokie podobieństwo ugrupowań krajów w przekroju cech standaryzowanych dynamicznie i statycznie. W dalszych badaniach warto podjąć próbę ustalenia, czy jest to związane z małymi zmianami cech w analizowanych latach, czy też decyduje o tym algorytm standaryzacji.

Uzyskiwanie różnych wyników grupowania w zależności od stopnia zaokrąglenia danych wyjściowych może świadczyć o przydatności tego narzędzia w analizach. Jeżeli rezultaty analizy są identyczne, to stosowanie tej operacji jest bezzasadne.

Operacja zaokrąglania danych w analizach danych właściwie nie jest stosowana. Jak się wydaje, w niektórych sytuacjach może się ona okazać przydatna. Pozwala przyglądać się dużym zbiorom liczb z pewnym (stopniowanym) dystansem, łatwiej jest wówczas dostrzec istniejące w nich regularności. Inaczej mówiąc, zaokrąglanie

liczb to narzędzie numerycznej indukcji umożliwiające analizy danych od szczegółu do ogółu.

Do zalet uproszczonych metod delimitacji należy też zaliczyć ich prostotę, intuicyjność, możliwość śledzenia efektów analizy w każdym jej etapie, a także brak ograniczeń związanych z narzędziami analitycznymi.

Prezentowana metoda delimitacji opiera się na pojedynczych zmiennych, co może być uznane za jej wadę. Tym niemniej jeżeli analizowana cecha jest zmienną agregatową, to można przyjąć, że pozwala ona także analizować struktury wielowymiarowe.

Istotnym problemem jest kwestia wykorzystania w analizach danych wyraźnych w różnych skalach pomiarowych, a zwłaszcza zasadności wykonywania na zmiennych w skalach słabszych: porządkowej (rangi) lub nominalnej (kategorii) operacji i przekształceń stosowanych w przypadku zmiennych w skalach silnych (ilorazowa lub przedziałowa). Warto podjąć badania mające na celu ustalenie, jak dalece zmieniają się wyniki analiz (np. parametry współzależności) w przypadku zastąpienia zmiennych w skalach silnych zmiennymi w skalach słabszych.

Literatura

- Czekanowski J. [1913], *Zarys metod statystycznych w zastosowaniu do antropologii*, Prace Towarzystwa Naukowego Warszawskiego, nr 5.
- Kozłowski S. [1972], *O sposobie weryfikacji podziałów przestrzennych w odniesieniu do taksonomicznej metody różnic J. Czekanowskiego*, „Wiadomości Statystyczne”, nr 1.
- Piasecki Z. [1971], *Nowa metod taksonomiczna*, „Listy Biometryczne”, nr 30–31.
- Sołtysiak A. [1977a], *UMCzek 1.00. Ewolucyjny algorytm porządkowania diagramu Czekanowskiego*, „Biuletyn Antropologiczny”, t. 1.
- Sołtysiak A. [1977b], *MaCzek 1.00. Język programowania diagramu Czekanowskiego*, „Biuletyn Antropologiczny”, t. 1.

The Simplified Method of Vector Delimitation

(Abstract)

The paper presents a graphical procedure for sorting and grouping objects described by a single variable taken in the form of cross-sectional, time series or panel data. This method draws on the classical taxonomic method developed over 100 years ago by Jan Czekanowski. The advantages of simplified vector delimitation are that it is not labour intensive but is intuitive and can be handled with simple spreadsheet functions. The method was used for the taxonomic analysis of a set of 28 EU countries described by three characteristics: the unemployment rate, poverty risk and life expectancy for the two years 2010 and 2015.

Keywords: taxonomy, clustering, grouping, delimitation, cross-time data.

Michał Major

Karty kontrolne sum skumulowanych z możliwością akceptacji procesu. Przegląd podstawowych narzędzi

Streszczenie

Celem artykułu jest przegląd oraz analiza podstawowych kart kontrolnych sum skumulowanych, umożliwiających akceptację lub dyskwalifikację badanego procesu produkcyjnego. W literaturze przedmiotu można spotkać dwa rodzaje kart kontrolnych umożliwiających akceptację procesu – karty kontrolne Shewharta oraz karty kontrolne sum skumulowanych. Obydwie mogą być wykorzystywane przez menedżerów jakości lub menedżerów finansowych w trakcie monitorowania lub audytu procesu. W artykule omówiono wybrane procedury kontrolne oraz przedstawiono przykłady zastosowania opisanych rozwiązań.

Słowa kluczowe: statystyczna kontrola jakości, statystyczne sterowanie procesami, karty kontrolne akceptacji, karty kontrolne sum skumulowanych.

Klasyfikacja JEL: C12, C19, C44, L15.

1. Wprowadzenie

W gospodarce wolnorynkowej bardzo ważna jest możliwość monitorowania procesów ekonomicznych z punktu widzenia poziomu ich stabilności. W XX w. w dobie

dynamicznie rozwijającego się przemysłu i produkcji seryjnej podstawowym zadaniem była możliwość niezwłocznej identyfikacji momentu rozregulowania procesu produkcyjnego. Niezwykle pomocne w tym zakresie były stworzone w 1924 r. przez Waltera Shewharta (1891–1967) karty kontrolne. Umożliwiły one łatwą i szybką identyfikację momentu rozregulowania procesu i podjęcie koniecznej regulacji agregatu produkcyjnego. Procedury kontrolne zaproponowane przez W. Shewharta opierały się na klasycznej teorii weryfikacji hipotez statystycznych¹. Podobną rolę odegrały również karty kontrolne sum skumulowanych, oparte na sekwencyjnej weryfikacji hipotez statystycznych, której podstawy sformułowane zostały w pracach Abrahama Walda (1902–1950). Rozwój technik sum skumulowanych zawdzięczamy E.S. Page'owi z Uniwersytetu Cambridge, autorowi artykułu opublikowanego w czasopiśmie „Biomertika” w 1954 r. (zob. [Page 1954, s. 100–115]). Publikacja ta ukazała się 9 lat po publikacji A. Walda [1945, s. 117–186]. Podejście prezentowane przez A. Walda, a następnie E.S. Page'a stało w opozycji do klasycznego podejścia do weryfikacji hipotez J. Neymana i E. Pearsona [Neyman i Pearson 1933, s. 289–337]. Wspólną cechą kart kontrolnych Shewharta i kart kontrolnych sum skumulowanych – w ich pierwotnej postaci – jest to, że umożliwiają jedynie rejestrację sygnałów świadczących o destabilizacji (rozregulowaniu) procesu. Nie pozwalają natomiast na akceptację badanego procesu, czy inaczej, stwierdzenie, że jest on uregulowany. Wiąże się to z tym, że procedury te w pierwotnej formie zakładały tylko ryzyko zbędnej regulacji procesu (α) – określane w statystyce prawdopodobieństwem błędu pierwszego rodzaju. Pomijane lub też zerowane jest ryzyko błędu drugiego rodzaju (β). Efektem tego jest możliwość stwierdzenia rozregulowania procesu lub ogłoszenia braku podstaw do odrzucenia hipotezy o jego uregulowaniu. Obecnie ważna staje się nie tylko tolerancja przebiegu procesu, ale również możliwość jego akceptacji. Stąd też powstały opracowania przedstawiające modyfikacje klasycznych narzędzi i metod weryfikacji procesów. Dotyczyły one zarówno kart kontrolnych Shewharta (zob. [Iwasiewicz 1985, s. 57–86, 159–163; 1999, s. 239–242; 2001, s. 35–38; Major 2015b, s. 223–238], a także [PN-ISO 7966: 2001]), jak i kart kontrolnych sum skumulowanych (zob. [Major 1997, s. 47–54; Iwasiewicz 2008–2009, s. 71–90; 2011, s. 213–245; Major 2015a, s. 25–43; 2015b, s. 223–238]). O ile zmodyfikowane karty kontrolne Shewharta zostały już obszernie opracowane, o tyle ciągle za mało jest opracowań poświęconych zmodyfikowanym procedurom sum skumulowanym. Może to wynikać z tego, że wielu badaczy i użytkowników błędnie postrzega procedury sum skumulowanych jako narzędzia trudniejsze do zrozumienia, a tym samym do stosowania. Nie zmienia to jednak faktu, że pomimo pozornej komplikacji, dają one w wielu sytuacjach lepsze rezultaty niż

¹ Pod pojęciem klasycznych procedur weryfikacji hipotez statystycznych należy rozumieć procedury zaproponowane przez Jerzego Sławę Neymana oraz Egonę Pearsona.

karty kontrolne Shewharta. Ponadto, procedury sekwencyjne w wielu przypadkach generują znacznie mniejsze koszty kontroli niż ich odpowiedniki oparte na kartach kontrolnych Shewharta, a kwestia obniżki kosztów jest jedną z determinant konkurencyjności nowoczesnego przedsiębiorstwa.

Wymienione powyżej, a także inne merytoryczne kwestie stały się przesłanką do sformułowania celu niniejszego opracowania, jakim jest potrzeba rzetelnej i szczegółowej ilustracji tego typu procedur. Praca ta jest kontynuacją opracowań autora poświęconych procedurom kontrolnym z możliwością akceptacji procesu (zob. [Major 1997, s. 47–54; 2015a, s. 25–43; 2015b, s. 223–238]) oraz nawiązuje do wcześniejszych opracowań środowiska naukowego Katedry Statystyki Uniwersytetu Ekonomicznego w Krakowie.

W artykule przedstawiono kilka najważniejszych kart kontrolnych sum skumulowanych. Opisane narzędzia reprezentują dwa warianty oceny przebiegu procesu produkcyjnego. Pierwszy z nich dotyczy sytuacji, gdy jakość produktu (wyrobu, usługi) lub procesu jest oceniana ze względu na zero-jedynkową zmienną diagnostyczną (tzw. alternatywna ocena właściwości). Proces oceniany w ten sposób nazywa się procesem binarnym (zob. [Iwasiewicz 2008–2009, s. 71–90; 2005–2006, s. 103–116; 2011, s. 213–245]). Z drugim wariantem oceny mamy do czynienia wówczas, gdy jakość produktu (wyrobu, usługi) lub procesu oceniana jest ze względu na ciągłą (quasi-ciągłą) zmienną diagnostyczną (tzw. liczbowa ocena właściwości). Ten typ procesu i jego oceny nazwano roboczo procesem ciągłym lub quasi-ciągłym.

2. Klasyczne procedury sekwencyjne

Teoria analizy sekwencyjnej sprowadza się do losowego pobierania pojedynczych lub małych zespołów zbiorowości generalnej oraz każdorazowego rozstrzygnięcia, czy zebrany zasób informacji pozwala na podjęcie określonej decyzji. Przy sformułowanej hipotezie zerowej H_0 i alternatywnej H_1 decyzje takie mogą dotyczyć:

- przyjęcia hipotezy H_0 ,
- odrzucenia hipotezy H_0 i przyjęcia hipotezy H_1 ,
- odłożenia decyzji do czasu pobrania następnego próbeki (n) do próby m .

Postać hipotezy zerowej i alternatywnej zależy od charakteru badanej zmiennej losowej (stymulanta, destymulanta, nominanta) i ograniczenia przedziału tolerancji. Przyjmując, że weryfikowany parametr oznaczono symbolem Q , w tabeli 1 ujęto trzy ogólne zapisy weryfikowanych hipotez (wzory (1a) i (1b), (2a) i (2b), (3a), (3b) i (3c)).

Tabela 1. Weryfikowane hipotezy

Ograniczenie przedziału tolerancji / rodzaj zmiennej	Weryfikowane hipotezy
Prawostronne / destymulanta	$H_0: Q = Q_0$ (1a) $H_1: Q = Q_1$ (1b) $(Q_0 < Q_1)$
Lewostronne / stymulanta	$H_0: Q = Q_0$ (2a) $H_{-1}: Q = Q_{-1}$ (2b) $(Q_{-1} < Q_0)$
Dwustronne / nominanta	$H_0: Q = Q_0$ (3a) $H_1: Q = Q_1$ (3b) $H_{-1}: Q = Q_{-1}$ (3c) $(Q_{-1} < Q_0 < Q_1)$

Źródło: opracowanie własne na podstawie [Wald 1945, Iwasiewicz 2008–2009].

Próba n jest w tym ujęciu sumą wszystkich próbek pobieranych w kolejnych k krokach, czyli:

$$n = n_1 + n_2 + \dots + n_k. \quad (4)$$

Jeżeli w każdym k -tym przedziale próbkowania liczebność próbki wynosi jeden, to wówczas liczność całej skumulowanej próby wyniesie $k(n = k)$.

Podczas każdego etapu badań obliczana jest wartość sekwencyjnego testu ilorazowego ogólnej postaci:

$$Q_k = \frac{\prod_{t=1}^n p_1(x_t | H_1)}{\prod_{t=1}^n p_0(x_t | H_0)}, \quad (5)$$

gdzie:

$p_1(x_t | H_1)$ – prawdopodobieństwo zdarzenia losowego $X = x_t$, gdy założona jest prawdziwość hipotezy H_1 ,

$p_0(x_t | H_0)$ – prawdopodobieństwo zdarzenia losowego $X = x_t$, gdy założona jest prawdziwość hipotezy H_0 ,

i sprawdzana jest prawdziwość nierówności:

$$A < Q_k < B, \quad (6)$$

gdzie:

$$A = \frac{\beta}{1 - \alpha}, \text{ natomiast } B = \frac{1 - \beta}{\alpha}.$$

Jeżeli $Q_k \leq A$, to przyjmuje się hipotezę H_0 z prawdopodobieństwem błędu nie większym niż β .

Gdy $Q_k \geq B$, wówczas należy przyjąć hipotezę H_1 z prawdopodobieństwem błędu nie większym niż α .

Natomiast gdy $A < Q_k < B$, to brak jest podstaw do podjęcia jednej z dwóch wymienionych powyżej decyzji i należy kontynuować badanie.

Postać sekwencyjnego testu ilorazowego (5) zależy od rodzaju zmiennej diagnostycznej X i charakteryzującego ją rozkładu prawdopodobieństwa. Zapisu wzoru (5) dokonano przy założeniu, że zmienna losowa X jest zmienną dyskretną. Wówczas wyrażenia stojące w tym wzorze należy interpretować jako funkcje rozkładu prawdopodobieństwa. W przypadku ciągłych zmiennych losowych wyrażenia we wzorze (5) są odpowiednimi funkcjami gęstości prawdopodobieństwa. Jest to więc stosunek dwóch funkcji wiarygodności otrzymanych na podstawie próby, które zostały wyznaczone przy założeniu prawdziwości hipotezy H_1 (licznik) i prawdziwości hipotezy H_0 (mianownik). W tabeli 2 zostały zestawione wzory ilorazu (5) dla najczęściej wykorzystywanych rozkładów zmiennej diagnostycznej.

Tabela 2. Sekwencyjny test ilorazowy Q_k

Typ rozkładu zmiennej diagnostycznej	Sekwencyjny test ilorazowy Q_k
Dwumianowy $X \sim B(p; 1)$	$Q_k = \frac{\prod_{i=1}^n p_1^{x_i} q_1^{1-x_i}}{\prod_{i=1}^n p_0^{x_i} q_0^{1-x_i}} \quad (q_1 = 1 - p_1; q_0 = 1 - p_0) \quad (7)$
Poissona $X \sim P(\lambda)$	$Q_k = \frac{\prod_{i=1}^n \frac{\lambda_1^{x_i}}{x_i!} e^{-\lambda_1}}{\prod_{i=1}^n \frac{\lambda_0^{x_i}}{x_i!} e^{-\lambda_0}} \quad (8)$
Normalny $X \sim N(\mu; \sigma)$	$Q_k = \frac{\prod_{i=1}^n \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x_i - \mu_1}{\sigma})^2}}{\prod_{i=1}^n \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x_i - \mu_0}{\sigma})^2}} \quad (9)$

Źródło: opracowanie własne na podstawie [Iwasiewicz, Paszek i Steczkowski 1988].

W dalszej części artykułu przyjęto założenie, że zmienna diagnostyczna X jest zmienną dyskretną lub ciągłą o rozkładzie normalnym z przedziałem tolerancji ograniczonym prawostronnie². Jeżeli obserwowaną charakterystyką z próby jest suma realizacji zmiennej X , to w kolejnych krokach postępowania oblicza się wartość statystyki:

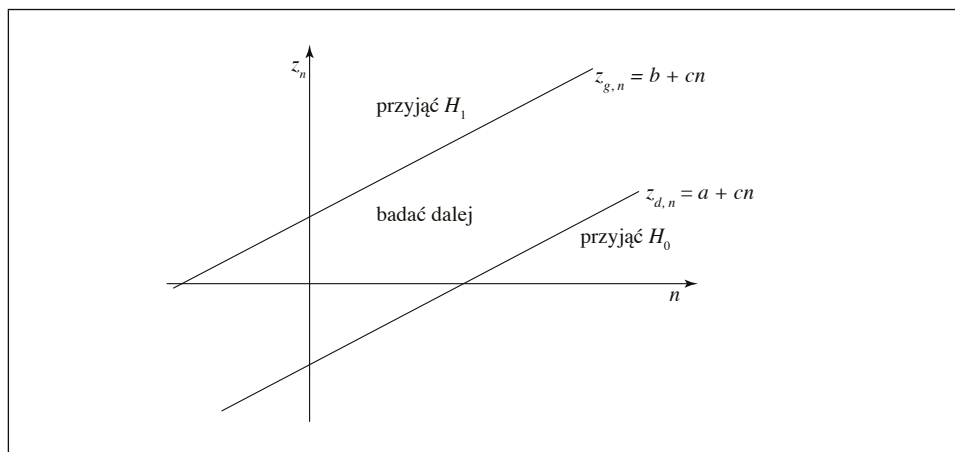
² W analogiczny sposób można zbudować procedury kontrolne dla zmiennych diagnostycznych ciągłych z przedziałem tolerancji ograniczonym lewostronnie lub dwustronnie. Szerzej zob. np. [Iwasiewicz, Paszek i Steczkowski 1988].

$$z_t = \sum_{i=1}^t x_i \quad (n = 1, 2, 3, \dots; t = 1, 2, 3, \dots, n). \quad (10)$$

Jeżeli w procesie weryfikacji hipotez stosowana jest graficzna wersja procedury kontrolnej, to wówczas obliczone wartości statystyki (10) według schematu:

$$t = 1: z_1 = x_1; t = 2: z_2 = z_1 + x_2; \dots; t = n: z_n = z_{n-1} + x_n, \quad (11)$$

nanosi się na diagram przeglądowy (wykres), przedstawiony schematycznie na rys. 1.



Rys. 1. Diagram przeglądowy klasycznej procedury sekwencyjnej przy prawostronnie ograniczonym przedziale tolerancji

Źródło: opracowanie własne.

Przedstawione na rys. 1 równania linii kontrolnych wynikają z nierówności (6) i powstają w wyniku jej przekształcenia³ do postaci:

$$z_{d,n} = a + cn < z_n < z_{g,n} = b + cn. \quad (12)$$

Wartości parametrów a , b i c zależą z kolei od przedstawionego w tabeli 2 rodzaju sekwencyjnego testu ilorazowego Q_k . Wzory opisujące te parametry zestawiono w tabeli 3.

Jeżeli w dowolnym kroku badania spełniona jest nierówność: $z_n \leq z_{d,n}$, wówczas przyjmuje się hipotezę H_0 (badany proces jest uregulowany), a prawdopodobieństwo tego, że prawdziwa jest hipoteza H_1 , nie przekracza β .

³ Dokładny proces przekształcenia nierówności (6) do postaci (12) czytelnik może znaleźć np. w pracy [Iwasiewicz, Paszek i Steczkowski 1988, s. 36 i dalsze].

Jeżeli zachodzi nierówność: $z_n \geq z_{g,n}$, wówczas przyjmuje się hipotezę H_1 (badany proces jest rozregulowany), a prawdopodobieństwo tego, że prawdziwa jest hipoteza H_0 , nie przekracza α .

Tabela 3. Parametry linii kontrolnych dla klasycznej procedury sekwencyjnej

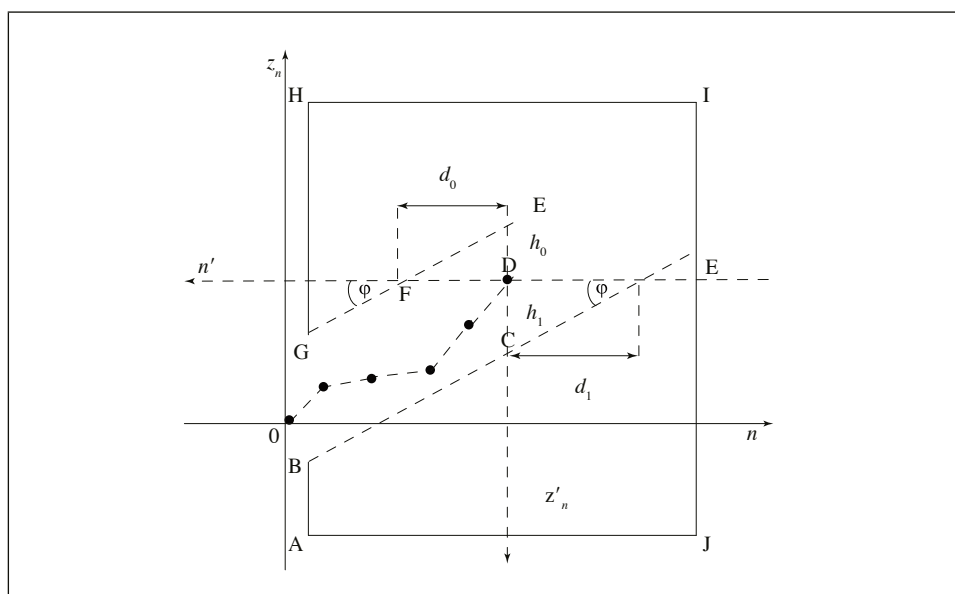
Typ rozkładu zmiennej diagnostycznej	Parametry linii kontrolnych
Dwumianowy	$a = \frac{\ln \frac{\beta}{1-\alpha}}{\ln \frac{p_1 q_0}{p_0 q_1}} \quad (13)$
	$b = \frac{\ln \frac{1-\beta}{\alpha}}{\ln \frac{p_1 q_0}{p_0 q_1}} \quad (14)$
	$c = \frac{\ln \frac{q_0}{q_1}}{\ln \frac{p_1 q_0}{p_0 q_1}} \quad (15)$
Poissona	$a = \frac{\ln \frac{\beta}{1-\alpha}}{\ln \frac{\lambda_1}{\lambda_0}} \quad (16)$
	$b = \frac{\ln \frac{1-\beta}{\alpha}}{\ln \frac{\lambda_1}{\lambda_0}} \quad (17)$
	$c = \frac{\lambda_1 - \lambda_0}{\ln \frac{\lambda_1}{\lambda_0}} \quad (18)$
Normalny (prawostronne ograniczenie przedziału tolerancji)	$a = \frac{\sigma^2}{\mu_1 - \mu_0} \ln \frac{\beta}{1-\alpha} \quad (19)$
	$b = \frac{\sigma^2}{\mu_1 - \mu_0} \ln \frac{1-\beta}{\alpha} \quad (20)$
	$c = \frac{\mu_1 + \mu_0}{2} \quad (21)$

Źródło: opracowanie własne na podstawie [Iwasiewicz, Paszek i Steczkowski 1988].

Jeżeli natomiast: $z_{d,n} < z_n < z_{g,n}$, to nie ma podstaw do przyjęcia żadnej z hipotez i należy badać dalej proces, zwiększając liczebność próby o jeden, a następnie powtórnie obliczyć parametry $z_{d,n}$, $z_{g,n}$ i z_n .

3. Procedura sum skumulowanych z możliwością akceptacji procesu

Z uwagi na fakt, że zasady funkcjonowania klasycznej procedury sum skumulowanych można znaleźć w wielu pozycjach literatury z zakresu zarządzania jakością (zob. np. [Iwasiewicz, Paszek i Steczkowski 1988, Montgomery 2009, s. 399–418]), w opracowaniu skupiono się tylko na zmodyfikowanej procedurze sum skumulowanych, umożliwiającej akceptację lub dyskwalifikację analizowanego procesu. Z formalnego punktu widzenia oznacza to, że zostanie uchylone założenie klasycznych kart kontrolnych sum skumulowanych głoszące, że ryzyko popełnienia błędu drugiego rodzaju $\beta = 0$, na rzecz założenia: $\beta > 0$. Niezmienne pozostaje założenie, że procedura sum skumulowanych jest klasyczną procedurą sekwencyjną realizowaną wstecznie. Oznacza to, że w każdym punkcie kończącym badaną sekwencję obserwacji wykreśla się pomocniczy układ współrzędnych obrócony w stosunku do pierwotnego o 180° . W każdym n -tym kroku postępowania bada się, czy zaobserwowany dotychczas ciąg obserwacji $z_n = \sum_{t=1}^n x_t$ wystarcza do przyjęcia jednej z dwóch weryfikowanych hipotez H_0 lub H_1 . W celu ułatwienia korzystania z tej metody konstruuje się tzw. maskownicę, którą przesuwa się na wykresie wraz ze wzrostem długości badanej sekwencji. Przykład takiej maskownicy przedstawiono na rys. 2.



Rys. 2. Zmodyfikowana maskownica

Źródło: opracowanie własne.

Zmodyfikowana maskownica ma dwie czynne krawędzie (linie kontrolne), odcinki BC i GF. Podczas posługiwania się maskownicą ważna pozostaje zasada, żeby punkt D pokrywał się z ostatnim z punktów obserwowanej sekwencji oraz odcinek DE był równoległy do osi odciętych n .

Przekroczenie – przez choć jeden z punktów skumulowanej sekwencji – in minus dolnej linii maskownicy sugeruje konieczność przyjęcia hipotezy H_1 (proces jest rozregulowany) z ryzykiem błędu nie większym niż α . Gdy przekroczona zostaje in plus górna linia maskownicy, wówczas przyjmuje się hipotezę H_0 (proces jest uregulowany) z ryzykiem błędu nieprzekraczającym β . Natomiast jeżeli punkty skumulowanej sekwencji układają się w korytarzu pomiędzy czynnymi krawędziami maskownicy, to wówczas kontynuuje się badania, powiększając próbkę o kolejną wartość x_i .

Aby zbudować maskownicę, niezbędna jest znajomość (zob. rys. 2) parametrów d_0, d_1, h_0, h_1 oraz $c = \text{tg}\varphi$. Parametr d_0 uzyskuje się w wyniku przekształcenia równania dolnej linii kontrolnej $z_{d,n}$ klasycznej procedury sekwencyjnej (zob. rys. 1). Przekształcenie polega na wyznaczeniu miejsca zerowego n_0 funkcji $z_{d,n}$. Pamiętając, że maskownica funkcjonuje w układzie współrzędnym obróconym o 180° , można zapisać, że:

$$d_0 = -n_0. \quad (22)$$

W analogiczny sposób wyznacza się równanie parametru d_1 , przekształcając tym razem równanie górnej linii kontrolnej $z_{g,n}$. Równania parametrów d_0 i d_1 dla rozważanych przykładowych trzech typów rozkładów zmiennej losowej zestawiono w tabeli 4.

Tabela 4. Równania parametrów maskownicy d_0 i d_1

Typ rozkładu	Parametry maskownicy
Dwumianowy	$d_0 = \frac{\ln \frac{\beta}{1-\alpha}}{\ln \frac{q_0}{q_1}} \quad (23)$
	$d_1 = \frac{\ln \frac{1-\beta}{\alpha}}{\ln \frac{q_0}{q_1}} \quad (24)$
Poissona	$d_0 = \frac{\ln \frac{\beta}{1-\alpha}}{\lambda_1 - \lambda_0} \quad (25)$
	$d_1 = \frac{\ln \frac{1-\beta}{\alpha}}{\lambda_1 - \lambda_0} \quad (26)$

cd. tabeli 4

Typ rozkładu	Parametry maskownicy
Normalny (prawostronne ograniczenie przedziału tolerancji)	$d_0 = \frac{2\sigma^2 \ln \frac{\beta}{1-\alpha}}{\mu_1^2 - \mu_0^2} \quad (27)$
	$d_1 = \frac{2\sigma^2 \ln \frac{1-\beta}{\alpha}}{\mu_1^2 - \mu_0^2} \quad (28)$

Źródło: opracowanie własne na podstawie [Iwasiewicz, Paszek i Steczkowski 1988].

Wartość parametru φ jest wyznaczana ze wzoru:

$$\varphi = \arctg c = \arctg \frac{\mu_0 + \mu_1}{2}, \quad (29)$$

natomiast parametry h_0 i h_1 wynikają ze wzorów:

$$h_0 = d_0 \operatorname{tg} \varphi = d_0 c, \quad (30)$$

$$h_1 = d_1 \operatorname{tg} \varphi = d_1 c. \quad (31)$$

Wartości parametrów h_0 , h_1 oraz c są niezbędne podczas konstrukcji numerycznego algorytmu weryfikacji stanu procesu. W trakcie jego przebiegu można wyróżnić dwie sekwencje przebiegu procesu – jedna prowadzi do akceptacji procesu (przyjęcia hipotezy H_0), natomiast druga do jego dyskwalifikacji (przyjęcia hipotezy H_1). Założono dalej, że sekwencja punktów prowadząca do akceptacji procesu będzie nazywana „sekwencją A”, natomiast sekwencja prowadząca do dyskwalifikacji procesu – „sekwencją B”. W obydwu przypadkach weryfikowana jest wartość skumulowana statystyki z_n^* postaci:

$$z_n^* = \sum_{i=1}^s (x_{it} - c), \quad (32)$$

gdzie:

$c = \operatorname{tg} \varphi$,

t – indeks bieżący,

i – indeks operacyjny,

s – największa w danym momencie wartość indeksu operacyjnego ($s = 1, 2, 3, \dots$).

Indeks t we wzorze (32) funkcjonuje przez cały okres badań, natomiast moment uruchomienia licznika indeksu i zależy od rodzaju sekwencji punktów.

Analizę przebiegu procesu rozpoczyna się od śledzenia znaku różnicy $x_i - c$. W chwili gdy $x_i - c < 0$, zostaje uruchomiony licznik indeksu operacyjnego i ($i = 1$) oraz rozpoczyna się „sekwencja A” skutkująca wyznaczeniem wartości statystyki (32). Obliczanie tej statystyki jest kontynuowane tak długo, aż zostanie spełniony jeden z poniższych warunków:

$$1^A \quad z_n^* \leq z_d^*, \text{ gdzie } z_d^* = d_0 c = h_0, \quad (33)$$

$$2^A \quad z_n^* \geq 0. \quad (34)$$

W przypadku gdy spełniony jest warunek 1^A (33), badanie kończy się przyjęciem hipotezy zerowej H_0 z ryzykiem błędu nie większym niż β , natomiast gdy spełniony jest warunek 2^A (34), przerywa się obliczanie wartości charakterystyki z_n^* i powraca do śledzenia znaku różnicy $x_t - c$. Spełnienie warunku (33) lub (34) pociąga za sobą również konieczność wyzerowania indeksu i ($i = 0$).

Uruchomienie licznika indeksu operacyjnego i może również nastąpić, gdy $x_t - c > 0$. Jest to wówczas początek „sekwencji B”, a zarazem moment rozpoczęcia kumulacji zgodnie ze wzorem (32). Proces takiej kumulacji trwa do momentu, gdy spełniony jest jeden z poniższych warunków:

$$1^B \quad z_n^* \geq z_g^*, \text{ gdzie } z_g^* = d_1 c = h_1, \quad (35)$$

$$2^B \quad z_n^* \leq 0. \quad (36)$$

Spełnienie warunku 1^B (35) skutkuje przyjęciem hipotezy alternatywnej H_1 z ryzykiem błędu nie większym niż α . Spełnienie drugiego warunku 2^B (36) oznacza natomiast konieczność przerwania kumulacji i powrót do śledzenia znaku różnicy $x_t - c$. Podobnie jak w poprzednim przypadku spełnienie warunku 1^B lub 2^B skutkuje wyzerowaniem indeksu i ($i = 0$).

4. Przykłady zastosowania kart kontrolnych

Przykład 1

W trakcie produkcji materiałów bawełnianych monitorowano ich kurczliwość podczas prania. W tym celu wyprodukowane próbki materiału poddawano okresowo badaniom na kurczliwość. Podczas badania założono, że kurczliwość bawełny nie powinna przekraczać 7%⁴. Co można powiedzieć o przebiegu procesu wytwarzania materiałów, jeżeli dla 10 losowo wybranych próbek odnotowano wartości procentowej kurczliwości podane w tabeli 5?

Tabela 5. Procentowa kurczliwość materiałów bawełnianych

t	1	2	3	4	5	6	7	8	9	10
n	1	2	3	4	5	6	7	8	9	10
x_t (%)	5	6	6	4,5	8	6,25	5	4	6	4,5

Źródło: dane umowne.

⁴ Według polskiej normy [PN-P-84004:2003] naturalna kurczliwość bawełny wynosi 6%.

Podczas analizy powyższego procesu założono, że: $\alpha = \beta = 0,05$ oraz że kurczliwość bawełny w praniu ma rozkład normalny ze stałą wariancją wynoszącą 1,5%.

Ponieważ zmienna diagnostyczna będąca procentową kurczliwością bawełny w praniu jest traktowana jako destymulanta jakości⁵ analizowanego procesu, hipoteza zerowa oraz hipoteza alternatywna będą miały postać:

$$H_0: \mu = \mu_0 = 5\%, \quad H_1: \mu = \mu_1 = 7\%.$$

Korzystając ze wzorów (19), (20) i (21), wyznaczono wartości parametrów a , b i c niezbędne do zastosowania klasycznej procedury sekwencyjnej.

$$a = \frac{1,5^2}{7-5} \ln \frac{0,05}{0,95} = -3,31; \quad b = \frac{1,5^2}{7-5} \ln \frac{0,95}{0,05} = 3,31; \quad c = \frac{7+5}{2} = 6.$$

Równania linii kontrolnych przedstawiały się następująco:

$$z_{d,n} = -3,31 + 6n; \quad z_{g,n} = 3,31 + 6n.$$

W tabeli 6 zestawiono obliczenia robocze dla danych wejściowych podanych w tabeli 5. Obserwując wyniki zawarte w tabeli 6 można zauważyć, że w dziesiątym kroku postępowania sekwencyjnego, przy liczności skumulowanej próbki $n = 10$, należy przyjąć hipotezę zerową H_0 i proces produkcji materiału bawełnianego uznać za uregulowany. Prawdopodobieństwo tego, że ocena ta jest fałszywa, nie przekracza $\beta = 0,05$.

Tabela 6. Analiza procesu kurczenia materiałów bawełnianych za pomocą klasycznej procedury sekwencyjnej

t	n	x_t (%)	\bar{z}_n	$\bar{z}_{d,n}$	$\bar{z}_{g,n}$	Decyzja
1	1	5	5	2,69	9,31	badać dalej
2	2	6	11	8,69	15,31	badać dalej
3	3	6	17	14,69	21,31	badać dalej
4	4	4,5	21,5	20,69	27,31	badać dalej
5	5	8	29,5	26,69	33,31	badać dalej
6	6	6,25	35,75	32,69	39,31	badać dalej
7	7	5	40,75	38,69	45,31	badać dalej
8	8	4	44,75	44,69	51,31	badać dalej
9	9	6	50,75	50,69	57,31	badać dalej
10	10	4,5	55,25	56,69	63,31	przyjąć H_0

Źródło: obliczenia własne.

⁵ Im kurczliwość 100-procentowej bawełny jest mniejsza, tym jakość bawełny jest wyższa. Jednak w pewnych sytuacjach niska lub wręcz zerowa kurczliwość materiału może świadczyć o tym, że dzianiny mają domieszkę innego materiału, np. elanobawełny.

Podobną analizę można przeprowadzić stosując zmodyfikowaną procedurę sum skumulowanych. Należy wtedy obliczyć dodatkowo wartości parametrów d_0, d_1 oraz h_0, h_1 (zob. wzory (27), (28), (30) i (31)). Mamy wówczas:

$$d_0 = \frac{2 \cdot 1,5^2 \ln \frac{0,05}{0,95}}{7^2 - 5^2} = -0,55; \quad d_1 = \frac{2 \cdot 1,5^2 \ln \frac{0,95}{0,05}}{7^2 - 5^2} = 0,55;$$

$$h_0 = d_0 c = z_d^* = -3,31; \quad h_1 = d_1 c = z_g^* = 3,31.$$

Mając te parametry, można przystąpić do powtórnej analizy danych przy użyciu karty kontrolnej sum skumulowanych z możliwością akceptacji procesu. Jej przebieg przedstawiono w tabeli 7.

Tabela 7. Analiza procesu kurczenia materiałów bawełnianych za pomocą zmodyfikowanej procedury sum skumulowanych

t	i	x_{ii}	n	$z_i^* = x_{ii} - c$	z_n^*	Uwagi
1	1	5	1	-1	-1	$x_i < c$ (zacząć kumulację, $i = 1$)
2	2	6	2	0	-1	kumulować ($i = i + 1$)
3	3	6	3	0	-1	kumulować ($i = i + 1$)
4	4	4,5	4	-1,5	-2,5	kumulować ($i = i + 1$)
5	5	8	5	2	-0,5	kumulować ($i = i + 1$)
6	6	6,25	6	0,25	-0,25	kumulować ($i = i + 1$)
7	7	5	7	-1	-1,25	kumulować ($i = i + 1$)
8	8	4	8	-2	-3,25	kumulować ($i = i + 1$)
9	9	6	9	0	-3,25	kumulować ($i = i + 1$)
10	10	4,5	10	-1,5	-4,75	$z_n^* < z_d^*$ (przyjąć H_0)

Źródło: obliczenia własne.

Jak widać, zmodyfikowana procedura sum skumulowanych doprowadziła również do przyjęcia hipotez zerowej. Proces kumulacji rozpoczął się w okresie $t = 1$ i trwał nieprzerwanie do okresu $t = 10$.

Przykład 2

W trakcie badania jakości linek holowniczych sprawdzanych jest wiele parametrów pod względem ich zgodności z wytycznymi zawartymi w Prawie o ruchu drogowym (długość holu, umiejscowienie chorągiewki ostrzegawczej i kolor linki) oraz zgodności z warunkami technicznymi określonymi przez Przemysłowy Instytut Motoryzacji (PIMOT) WT/008/PIMOT/93 – „Linki holownicze. Wymagania i badania w zakresie bezpieczeństwa użytkownika”. Pod uwagę bierze się wówczas instrukcję użytkownika oraz wytrzymałość holu. W trakcie weryfikacji

wytrzymałości linki zakłada się, że (<http://federacja-konsumentow.org.pl/download/LINKI%202.doc>, data dostępu: 22.09.2016):

– linki holownicze przeznaczone do holowania pojazdów o masie do 1500 kg powinny bez uszkodzeń przenieść obciążenie nie mniejsze niż 1200 daN w czasie 20 s,

– linki holownicze przeznaczone do holowania pojazdów o masie większej niż 1500 kg powinny przenieść bez uszkodzeń obciążenie nie mniejsze niż masa holowanego pojazdu. Na przykład jeżeli masa pojazdu wynosi 2000 kg, to obciążenie badawcze linki holowniczej powinno wynieść nie mniej niż 2000 daN w czasie 20 s. Bardzo ważną kwestią jest także pomiar wytrzymałości granicznej do zerwania, podczas którego określa się poziom przyłożonej siły, przy której linka ulega zerwaniu. Linka holownicza z zaczepem powinna wytrzymać obciążenie większe niż deklarowane w przepisach WT/008/PIMOT/93 oraz jej wytrzymałość powinna przekraczać wartość deklarowaną w ulotce przez producenta lub dystrybutora.

W dalszym ciągu rozważań przyjęto założenie, że przedmiotem badań jest partia linek, która przeszła pozytywnie weryfikację ze względu na wszystkie właściwości, z wyjątkiem wytrzymałości na rozciąganie mierzonej w jednostce czasu (20 s). Przyjęto, że badaniu poddano próbki linek przeznaczonych do holowania pojazdów o masie do 1500 kg. Założono, że obserwowana zmienna losowa X przybierała wartość 0, gdy linka przeszła pozytywnie próbę obciążenia nie mniejszą niż 1200 daN, i wartość 1, gdy linka w czasie próby wytrzymałości została zerwana lub też jej mocowania zostały na tyle zdeformowane, że nie kwalifikowała się ona do dalszego użytku. Założono dalej, że badanie przeprowadzono w ciągu 10 kolejnych momentów czasowych, powiększając za każdym razem liczebność próbki o 1 sztukę. Wyniki badania podano w tabeli 8.

Tabela 8. Wytrzymałość linek holowniczych

t	1	2	3	4	5	6	7	8	9	10
n	1	2	3	4	5	6	7	8	9	10
x_t	0	0	0	1	0	0	1	1	1	1

Źródło: dane umowne.

Należy sprawdzić, czy na podstawie tych danych można przyjąć jedną z następujących hipotez:

$$H_0: p = p_0 = 0,01,$$

$$H_1: p = p_1 = 0,03.$$

Symbole p_0 i p_1 oznaczają odpowiednio poziomy wadliwości: maksymalnej dopuszczalnej i minimalnej dyskwalifikującej badanego procesu. W trakcie weryfikacji hipotez przyjęto dodatkowe założenie $\alpha = \beta = 0,01$. Sprawdzenie zdefiniowanych hipotez odbyło się w pierwszej kolejności z wykorzystaniem klasycznej procedury sekwencyjnej, a następnie zmodyfikowanej procedury sum skumulowanych.

W trakcie wykorzystania klasycznej procedury sekwencyjnej wyznaczono wartości statystyki z_n postaci (10) i (11) i porównano je z liniami kontrolnymi $z_{d,n}$ i $z_{g,n}$ obliczonymi na podstawie wzorów (13)–(15). Wynikało to z założenia, że otrzymana w trakcie analizy charakterystyka będąca sumą zmiennych zero-jedynkowych ma rozkład dwumianowy. Wartości parametrów a , b i c oraz powstałe na ich podstawie równania linii kontrolnych przedstawiały się następująco:

$$a = \frac{\ln \frac{0,01}{0,99}}{\ln \frac{0,03 \cdot 0,99}{0,01 \cdot 0,97}} = -4,11; \quad b = \frac{\ln \frac{0,99}{0,01}}{\ln \frac{0,03 \cdot 0,99}{0,01 \cdot 0,97}} = 4,11; \quad c = \frac{\ln \frac{0,99}{0,97}}{\ln \frac{0,03 \cdot 0,99}{0,01 \cdot 0,97}} = 0,02;$$

$$z_{d,n} = -4,11 + 0,02n; \quad z_{g,n} = 4,11 + 0,02n.$$

Wyniki analizy danych empirycznych zestawiono w tabeli 9.

Tabela 9. Analiza wytrzymałości linek za pomocą klasycznej procedury sekwencyjnej

t	n	x_t	z_n	$z_{d,n}$	$z_{g,n}$	Decyzja
1	1	0	0	-4,09	4,12	badać dalej
2	2	0	0	-4,07	4,14	badać dalej
3	3	0	0	-4,05	4,16	badać dalej
4	4	1	1	-4,03	4,18	badać dalej
5	5	0	1	-4,02	4,20	badać dalej
6	6	0	1	-4,00	4,22	badać dalej
7	7	1	2	-3,98	4,23	badać dalej
8	8	1	3	-3,96	4,25	badać dalej
9	9	1	4	-3,94	4,27	badać dalej
10	10	1	5	-3,92	4,29	przyjąć H_1

Źródło: obliczenia własne.

Jak wynika z tabeli 9, przy liczebności próbki $n = 10$ została podjęta decyzja o przyjęciu hipotezy H_1 przy ryzyku błędu $\alpha = 0,01$.

W kolejnym kroku badania dane wejściowe zestawione w tabeli 8 poddano weryfikacji przy użyciu zmodyfikowanej procedury sum skumulowanych. W tym

celu, korzystając ze wzorów (23) i (24) oraz (30) i (31), wyznaczono kolejno parametry d_0, d_1 oraz h_0, h_1 .

$$d_0 = \frac{\ln \frac{0,01}{0,99}}{\ln \frac{0,97}{0,99}} = -225,15; \quad d_1 = \frac{\ln \frac{0,99}{0,01}}{\ln \frac{0,97}{0,99}} = 225,15;$$

$$h_0 = d_0 c = z_d^* = -4,11; \quad h_1 = d_1 c = z_g^* = 4,11.$$

Obliczenia robocze oraz wynikające z nich wnioski ujęto w tabeli 10.

Tabela 10. Analiza wytrzymałości linek za pomocą zmodyfikowanej procedury sum skumulowanych

t	i	x_{it}	n	$z_i^* = x_{it} - c$	z_n^*	Uwagi
1	1	0	1	-0,02	-0,02	$x_i < c$ (start sekw. A, $i = 1$)
2	2	0	2	-0,02	-0,04	kumulować ($i = i + 1$)
3	3	0	3	-0,02	-0,05	kumulować ($i = i + 1$)
4	1	1	4	0,98	0,98	$z_n^* > 0$ (start sekw. B, kumulować ($i = 1$))
5	2	0	5	-0,02	0,96	kumulować ($i = i + 1$)
6	3	0	6	-0,02	0,95	kumulować ($i = i + 1$)
7	4	1	7	0,98	1,93	kumulować ($i = i + 1$)
8	5	1	8	0,98	2,91	kumulować ($i = i + 1$)
9	6	1	9	0,98	3,89	kumulować ($i = i + 1$)
10	7	1	10	0,98	4,87	$z_n^* > z_g^*$ (przyjąć H_1)

Źródło: obliczenia własne.

W trakcie przebiegu badanego procesu dostrzec można dwie sekwencje. Pierwsza z nich (sekwencja A) prowadziła do przyjęcia hipotezy H_0 , ale została przerwana po czterech krokach, ponieważ wartość sumy skumulowanej z_n^* była większa od zera. Moment $t = 4$ był równocześnie punktem startu drugiej sekwencji typu B, która doprowadziła po siedmiu kolejnych krokach kumulacji do przyjęcia hipotezy H_1 . A zatem wnioski, jakie wynikają z zastosowania zmodyfikowanej procedury sum skumulowanych, są identyczne jak w przypadku użycia klasycznej procedury sekwencyjnej. Na koniec warto zauważyć, że przyjęcie hipotezy H_1 może nastąpić znacznie szybciej niż przyjęcie hipotezy H_0 . W analizowanym przykładzie potrzebnych było 5 przypadków zerwania linki holowniczej w dziesięcioelementowej próbie. Jednocześnie, aby potwierdzić prawdziwość hipotezy H_0 , należałoby przebadać przynajmniej 226 linek (miejsce zerowe linii $z_{d,n}$ lub parametr d_1) i otrzymać we wszystkich przypadkach potwierdzenie zgodności wytrzymałości linek z wymaganiami w tym zakresie.

5. Podsumowanie

Przedstawione w opracowaniu procedury i przykłady nie wyczerpują zagadnienia zastosowania kart kontrolnych sum skumulowanych z możliwością akceptacji procesu. W omawianych przypadkach ograniczono się do sytuacji, gdy analizowana zmienna jest destymulantą jakości. Skupiono się na najczęściej występujących podczas kontroli jakości rozkładach zmiennych diagnostycznych, takich jak: rozkład dwumianowy, rozkład normalny i rozkład Poissona. Rozkład dwumianowy jest łączony często z przypadkami, gdy miarą jakości wykonania jest wadliwość, a badany proces ma charakter binarny. Rozkład Poissona ma natomiast zastosowanie, gdy miarą jakości produktu jest liczba wad lub niezgodności w jednostce produktu⁶. Z kolei rozkład normalny wykorzystywany jest najczęściej w sytuacji, gdy obserwowana zmienna diagnostyczna ma charakter ciągły (tzw. procesy quasi-ciągłe). Na uwagę zasługują również (nieujęte w tym opracowaniu) procedury kontrolne stosowane dla zmiennych diagnostycznych o rozkładzie wykładniczym. Rozkład wykładniczy jest najczęściej stosowany do opisu niezawodności nienaprawialnych elementów systemów technicznych, a używana zmienna diagnostyczna jest definiowana jako czas poprawnej pracy tychże elementów. Tego typu zmienna losowa będzie traktowana jako stymulanta jakości, co przełoży się na konieczność modyfikacji omawianych procedur.

Jak zaznaczono na wstępie artykułu, opisane w nim karty kontrole sum skumulowanych nie są jedynymi procedurami pozwalającymi na dyskwalifikację lub akceptację monitorowanych procesów. Należy wymienić również inne karty kontrolne, w tym procedury kontrolne oparte na kartach kontrolnych Shewharta. Opis ich funkcjonowania można znaleźć np. w cytowanych wcześniej pracach A. Iwasiewicza [1985, 1999, 2001], a także w normie PN-ISO 7966 z 2001 r. Wymagają one jednak wstępnego ustalenia niezbędnej liczebności próby potrzebnej do przyjęcia jednej z weryfikowanych hipotez. W procedurach sekwencyjnych oraz powstałych na ich podstawie procedurach sum skumulowanych liczebność próby jest parametrem określanym dopiero w momencie zaakceptowania lub dyskwalifikacji procesu. Jest to istotna zaleta tych narzędzi, pozwalająca na minimalizację ryzyka przeszacowania liczebności próby, a co się z tym wiąże, generowania zbytecznych kosztów badania i oceny. Jest to szczególnie ważne, gdy badania kontrolne mają charakter niszczący. Z tego względu zaprezentowane w niniejszym artykule metody sekwencyjne oceny procesu produkcyjnego wydają się znacznie skuteczniejsze i kosztowo efektywniejsze niż analogiczne procedury oparte na kartach kontrolnych Shewharta.

⁶ Procedury kontrolne sum skumulowanych dla zmiennych o rozkładzie Poissona zostały omówione np. w pracy [Major 2015b, s. 223–238].

Literatura

- Iwasiewicz A. [1985], *Statystyczna kontrola jakości w toku produkcji, systemy i procedury*, PWN, Warszawa.
- Iwasiewicz A. [1999], *Zarządzanie jakością – podstawowe problemy i metody*, Wydawnictwo Naukowe PWN, Warszawa–Kraków.
- Iwasiewicz A. [2001], *Karty kontrolne Shewharta z możliwością akceptacji procesu*, „Zeszyty Naukowe Uniwersytetu Szczecińskiego”, nr 320.
- Iwasiewicz A. [2005–2006], *Monitorowanie procesów binarnych*, „Folia Oeconomica Cracoviensia”, vol. XLVI–XLVII.
- Iwasiewicz A. [2008–2009], *Monitorowanie procesów binarnych za pomocą kart kontrolnych sum skumulowanych*, „Folia Oeconomica Cracoviensia”, vol. XLIX–L.
- Iwasiewicz A. [2011], *Analiza wielowymiarowych procesów binarnych jako metoda wspomaganie decyzji menedżerskich w zarządzaniu jakością* [w:] *Przedsiębiorcze aspekty rozwoju organizacji i biznesu*, red. A. Chodyński, Oficyna Wydawnicza AFM, Kraków.
- Iwasiewicz A., Paszek Z., Steczkowski J. [1988], *Sekwencyjne metody kontroli jakości*, Akademia Ekonomiczna w Krakowie, Kraków.
- Major M. [1997], *Sterowanie procesem za pomocą kart kontrolnych sum skumulowanych*, Materiały konferencyjne z I Krajowej Konferencji Naukowej Materiałoznawstwo–Odlewnictwo–Jakość, vol. 3: Jakość, Kraków.
- Major M. [2015a], *Acceptance Control Charts*, „Zeszyty Naukowe Uniwersytetu Ekonomicznego w Krakowie”, nr 7(943), <https://doi.org/10.5678/ZNUEK.2015.0943.0702>.
- Major M. [2015b], *Karty kontrolne sum skumulowanych z możliwością akceptacji procesu dla zmiennych diagnostycznych o rozkładzie Poissona* [w:] *Wielowymiarowość systemów zarządzania*, red. M. Giemza, T. Sikora, Wydawnictwo Naukowe PTTŻ, Kraków.
- Montgomery D.C. [2009], *Introduction to Statistical Quality Control*, Sixth ed., John Wiley & Sons, Inc., Jefferson City, USA.
- Neyman J., Pearson E.S. [1933], *On the Problem of the Most Efficient Tests of Statistical Hypotheses. Philosophical Transactions of the Royal Society of London. Series A: „Mathematical, Physical and Engineering Sciences”*, vol. 231, The Royal Society Publishing, <http://www.stats.org.uk/statistical-inference/NeymanPearson1933.pdf> (data dostępu: 23.12.2016).
- Page E.S. [1954], *Continuous Inspection Scheme*, „Biometrika”, vol. 41(1/2).
- PN-ISO 7966: 2001. Karty akceptacji procesu, Polski Komitet Normalizacyjny, Warszawa.
- PN-P-84004:2003. Tekstylnia – Działiny i wyroby dziane – Wartości wskaźnika zmiany wymiarów po praniu, Polski Komitet Normalizacyjny, Warszawa.
- Wald A. [1945], *Sequential Tests of Statistical Hypotheses*, „Annals of Mathematical Statistics”, vol. 16(2).

Acceptance Cumulative Sum Control Charts: Overview of Basic Tools

(Abstract)

The main goal of this article was to review and analyse basic cumulative sum control charts, which can be used to accept or disqualify an analysed production process. In the

literature, there are two kinds of acceptance control charts: Shewhart Control Charts and Cumulative Sum Control Charts (CUSUM Charts). Both can be used by quality managers or financial managers in monitoring or auditing processes. The article presents only selected process control procedures. The conclusion of the article provides examples of the applications of the solutions described.

Keywords: statistical quality control, statistical process control, acceptance control charts, cumulative sum control charts.

| Jan Tatar

Zbieżność stochastyczna ciągów wektorów losowych*

Streszczenie

W artykule zaproponowano uogólnienie na przypadek wielowymiarowy dwóch twierdzeń, znanych dla zmiennych losowych jednowymiarowych, dotyczących zbieżności stochastycznej, czyli zbieżności według prawdopodobieństwa. Uogólnianymi twierdzeniami są słabe prawa wielkich liczb Markowa i Chinczyna. Wynika z nich, że przy odpowiednich założeniach ciąg średnich arytmetycznych wektorów losowych jest stochastycznie zbieżny do średniej arytmetycznej ich wartości oczekiwanych. W przeprowadzonych dowodach wykorzystano „łączne momenty rozkładów prawdopodobieństwa wektorów losowych” zaproponowane we wcześniejszych pracach autora. Opierają się one na definicji potęgi wektora w przestrzeni z iloczynem skalarnym.

Słowa kluczowe: potęga wektora, moment rozkładu prawdopodobieństwa, wektor losowy, zbieżność stochastyczna.

Klasyfikacja JEL: C02, C10, C18, C32.

1. Wprowadzenie

W pracach [Tatar 1993, 1996b] zaproponowano odmienne od wcześniej stosowanego podejście do opisu wielkości losowych o charakterze wektorowym, tj. podejście, w którym charakterystyki rozkładów prawdopodobieństwa tych wiel-

Jan Tatar, Uniwersytet Ekonomiczny w Krakowie, Wydział Finansów i Prawa, Katedra Matematyki, ul. Rakowicka 27, 31-510 Kraków, e-mail: tatarj@uek.krakow.pl

* Artykuł stanowi wynik realizacji projektu badawczego sfinansowanego ze środków przyznanych Wydziałowi Finansów i Prawa Uniwersytetu Ekonomicznego w Krakowie w ramach dotacji na utrzymanie potencjału badawczego.

kości (w szczególności ich momenty) są rzeczywiście charakterystykami wektorów losowych nie zaś – jak to ma miejsce w ujęciu klasycznym – jednowymiarowych zmiennych losowych będących funkcjami współrzędnych badanych wektorów. Punktem wyjścia prowadzonych w przywołanych pracach rozważań było zdefiniowanie nowego pojęcia, jakim jest potęga wektora w przestrzeni euklidesowej (lub ogólniej: w przestrzeni Hilberta). Umożliwiło ono sformułowanie definicji tzw. momentów łącznych (zarówno zwyczajnych, jak i centralnych) rozkładów prawdopodobieństwa wektorów losowych.

W wielu kolejnych pracach uogólniono na przypadek wielowymiarowy m.in. takie pojęcia, jak współczynnik korelacji [Tatar 2008a], momenty absolutne [Tatar 2001], miary zależności [Tatar 2008b], miary asymetrii [Tatar 2000], funkcje charakterystyczne oraz półniezmienniki [Tatar 2004, 2006], rozkłady warunkowe [Tatar 2009], regresja liniowa [Najman i Tatar 2010, Budny i Tatar 2012], czy wreszcie kurtoza oraz eksces [Budny 2009, Budny i Tatar 2009]. Sformułowano i udowodniono także wielowymiarowe wersje niektórych znanych w literaturze probabilistycznej twierdzeń, np. nierówność Czebyszewa [Tatar 1996a, Osiewalski i Tatar 1997, 1999, Budny 2014a, b], nierówność Lapunowa [Tatar 2002] czy wybrane słabe prawa wielkich liczb [Tatar 2003]. Zaproponowane łączne charakterystyki wielowymiarowych rozkładów prawdopodobieństwa wykorzystano także w opisie i badaniu wielkości ekonomicznych i finansowych [Tatar 2013, Budny i Tatar 2014, Budny, Szklarska i Tatar 2014].

W niniejszej pracy zaproponowano i udowodniono uogólnienie na przypadek wektorów losowych kolejnych dwóch twierdzeń z grupy słabych praw wielkich liczb, czyli tych, które mówią o stochastycznej (tj. według prawdopodobieństwa) zbieżności ciągów wektorów losowych. Będą to uogólnienia twierdzeń Markowa i Chinczyna. Ich postaci dla zmiennych losowych jednowymiarowych można znaleźć np. w pracach [Feller 1969, Fisz 1969, Plucińska i Pluciński 2000].

2. Podstawowe pojęcia

Niech dana będzie przestrzeń wektorowa $(R^n, R, +, \cdot)$, w której określono klasyczny (euklidesowy) iloczyn skalarny postaci $(\cdot | \cdot): R^n \times R^n \rightarrow R$.

Definicja 1 [Tatar 1993, 1996b]. Potęgą stopnia k , gdzie $k \in N_0 = N \cup \{0\}$, wektora $v \in R^n$ nazywamy wielkość v^k określoną następująco:

$$v^0 = 1$$

oraz

$$v^k = \begin{cases} v^{k-1} \cdot v, & \text{dla } k \text{ nieparzystej} \\ (v^{k-1} | v), & \text{dla } k \neq 0 \wedge k \text{ parzystej} \end{cases}$$

Z powyższej definicji wynikają w szczególności następujące własności:

$$(w.1) \quad \forall v \in R^n, k \in N_0: k \text{ parzysta} \Rightarrow v^k \in R,$$

$$(w.2) \quad \forall v \in R^n, k \in N: k \text{ nieparzysta} \Rightarrow v^k \in R^n.$$

Jak wspomniano we wprowadzeniu, pojęcie potęgi wektora pozwoliło zdefiniować momenty wielowymiarowego rozkładu prawdopodobieństwa.

Niech zatem (Ω, S, P) będzie przestrzenią probabilistyczną oraz niech $\xi = (\xi_1, \xi_2, \dots, \xi_n): \Omega \rightarrow R^n$ będzie n -wymiarowym wektorem losowym o wartościach w R^n .

Definicja 2 [Tatar 1993, 1996b]. Dla dowolnej liczby $k \in N_0$ łącznym momentem zwykłym rzędu k wektora losowego ξ nazywamy wartość oczekiwaną $m_k = E(\xi^k)$, jeżeli $E(\|\xi^k\|) < +\infty$.

$$\text{W szczególności: } m_0 = 1, m_1 = (m_{1(1)}, m_{1(2)}, \dots, m_{1(n)}), m_2 = \sum_{i=1}^n m_{2(i)}.$$

Symbolem $m_{r(i)}$ ($r \in N, i = 1, \dots, n$) oznaczamy w powyższych formułach moment zwykły rzędu r w rozkładzie brzegowym jednowymiarowej zmiennej losowej ξ_i .

Prawdziwe są implikacje:

$$(w.3) \quad \text{jeżeli } k \text{ jest liczbą parzystą, to } m_k \in R,$$

$$(w.4) \quad \text{jeżeli } k \text{ jest liczbą nieparzystą, to } m_k \in R^n.$$

Definicja 3 [Tatar 1993, 1996b]. Momentem centralnym łącznym rzędu k , ($k \in N_0$) wektora losowego ξ nazywamy wartość oczekiwaną $\mu_k = [E(\xi - m_1)^k]$, jeżeli spełniona jest nierówność $E(\|\xi - m_1\|^k) < +\infty$.

$$\text{W szczególności otrzymujemy: } \mu_1 = 0 \in R^n, \mu_2 = \sum_{i=1}^n \mu_{2(i)}.$$

Symbolem $\mu_{2(i)}$ oznaczyliśmy moment centralny rzędu drugiego w rozkładzie jednowymiarowej zmiennej losowej ξ_i .

Centralny moment łączny rzędu drugiego nazywamy wariancją łączną wektora losowego ξ i oznaczamy przez $\sigma^2(\xi)$ lub $Var \xi$. Jest zatem: $\sigma^2(\xi) = \mu_2 = \sum_{i=1}^n \mu_{2(i)} = \sum_{i=1}^n \sigma_i^2$, gdzie σ_i^2 ($i = 1, \dots, n$) oznacza wariancję rozkładu zmiennej brzegowej ξ_i .

Warto podkreślić, że przy przyjętych definicjach zachodzi równość: $\sigma^2(\xi) = m_2 - m_1^2$.

Pierwiastek kwadratowy z wariancji łącznej nazywamy łącznym odchyleniem standardowym rozkładu wektora losowego ξ i oznaczamy przez $\sigma(\xi)$. Jest więc

$$\sigma(\xi) = \sqrt{\sigma^2(\xi)} = \sqrt{Var \xi} = \left(\sum_{i=1}^n \sigma_i^2 \right)^{1/2}.$$

Inne wprowadzone wcześniej pojęcia, konieczne do zrozumienia prowadzonych w niniejszej pracy rozważań, zostaną przypomniane bezpośrednio przed prezentacją głównych rezultatów.

3. Uogólnione wersje słabych praw wielkich liczb

3.1. Uwagi ogólne

Na początku rozważań przypomnimy pojęcie stochastycznej zbieżności ciągu wektorów losowych (inaczej: zbieżności według prawdopodobieństwa), dla którego rezerwujemy symbol „ \xrightarrow{p} ”.

Niech zatem $\xi_n, \xi: \Omega \rightarrow R^s, n=1, 2, 3, \dots$, będą wektorami losowymi oraz niech $c_n \in R^s, n=1, 2, 3, \dots$, oraz $c \in R^s$.

Definicja 4 (stochastyczna zbieżność ciągu wektorów losowych):

- $\xi_n \xrightarrow{p} 0 \Leftrightarrow \forall \varepsilon > 0: \lim_{n \rightarrow \infty} P(\|\xi_n\| > \varepsilon) = 0$,
- $\xi_n \xrightarrow{p} \xi \Leftrightarrow \xi_n - \xi \xrightarrow{p} 0$,
- $\xi_n \xrightarrow{p} c_n \Leftrightarrow \xi_n - c_n \xrightarrow{p} 0$,
- $\xi_n \xrightarrow{p} c \Leftrightarrow \xi_n - c \xrightarrow{p} 0$.

3.2. Uogólnione prawo Markowa

W dowodzie uogólnianego twierdzenia wykorzystamy wielowymiarową wersję nierówności Czebyszewa.

Twierdzenie 1 [Osiewalski i Tatar 1999]. Dla wektora losowego $\xi: \Omega \rightarrow R^s$ o skończonym drugim momencie zwyczajnym oraz dla dowolnej liczby rzeczywistej $r > 0$ zachodzi nierówność $P[\|\xi - m_1(\xi)\| \geq r \cdot \sigma(\xi)] \leq \frac{1}{r^2}$.

Tezę twierdzenia 1 można także zapisać inaczej: $\forall \varepsilon > 0: P[\|\xi - m_1(\xi)\| \geq \varepsilon] \leq \frac{\sigma^2(\xi)}{\varepsilon^2}$.

Twierdzenie 2 (uogólnione słabe prawo wielkich liczb Markowa). Niech $\xi_k: \Omega \rightarrow R^s (k=1, 2, 3, \dots)$ będzie ciągiem s -wymiarowych wektorów losowych o wartościach oczekiwanych $m(\xi_k)$ oraz wariancjach $Var \xi_k$ spełniających warunek $Var\left(\sum_{k=1}^n \xi_k\right) = o(n^2)$. Wówczas: $\frac{1}{n} \sum_{k=1}^n (\xi_k - m(\xi_k)) \xrightarrow{p} 0$.

Uwagi:

- założenie $Var\left(\sum_{k=1}^n \xi_k\right) = o(n^2)$ oznacza, że $\lim_{n \rightarrow \infty} \frac{1}{n^2} \sum_{k=1}^n Var \xi_k = 0$,
- tezę twierdzenia można także zapisać następująco:

$$\frac{1}{n} \sum_{k=1}^n \xi_k \xrightarrow{p} \frac{1}{n} \sum_{k=1}^n m(\xi_k)$$

lub

$$\forall \varepsilon > 0: \lim_{n \rightarrow \infty} P \left[\left\| \frac{1}{n} \sum_{k=1}^n (\xi_k - m(\xi_k)) \right\| < \varepsilon \right] = 1.$$

Dowód. Określmy nowy ciąg $\{\eta_n\}$ wektorów losowych postaci $\eta_n = \frac{1}{n} \sum_{k=1}^n \xi_k$ dla $n = 1, 2, 3, \dots$. Wartość oczekiwana każdego wektora η_n jest postaci:

$$m(\eta_n) = E(\eta_n) = \frac{1}{n} \sum_{k=1}^n E(\xi_k) = \frac{1}{n} \sum_{k=1}^n m(\xi_k).$$

Wykorzystując powyższe ustalenia oraz uogólnioną nierówność Czebyszewa (por. twierdzenie 1), otrzymujemy – dla dowolnego $\varepsilon > 0$ – następujący ciąg zależności:

$$\begin{aligned} P \left[\left\| \frac{1}{n} \sum_{k=1}^n (\xi_k - m(\xi_k)) \right\| \geq \varepsilon \right] &= P \left[\left\| \frac{1}{n} \sum_{k=1}^n \xi_k - \frac{1}{n} \sum_{k=1}^n m(\xi_k) \right\| \geq \varepsilon \right] = \\ &= P \left[\left\| (\eta_n - m(\eta_n)) \right\| \geq \varepsilon \right] \leq \frac{\frac{1}{n^2} \text{Var} \left[\sum_{k=1}^n (\xi_k - m(\xi_k)) \right]}{\varepsilon^2} = \frac{\text{Var} \left[\sum_{k=1}^n \xi_k - \sum_{k=1}^n m(\xi_k) \right]}{\varepsilon^2 \cdot n^2} = \\ &= \frac{\text{Var} \left[\sum_{k=1}^n \xi_k \right]}{\varepsilon^2 \cdot n^2} = \frac{1}{\varepsilon^2} \cdot \frac{1}{n^2} \cdot \text{Var} \left[\sum_{k=1}^n \xi_k \right]. \end{aligned}$$

Na mocy założenia oraz twierdzenia o trzech ciągach mamy zatem równość:

$$\lim_{n \rightarrow \infty} P \left[\left\| \frac{1}{n} \sum_{k=1}^n (\xi_k - m(\xi_k)) \right\| \geq \varepsilon \right] = 0,$$

czyli także

$$\lim_{n \rightarrow \infty} P \left[\left\| \frac{1}{n} \sum_{k=1}^n (\xi_k - m(\xi_k)) \right\| < \varepsilon \right] = 1,$$

a więc żądana tezę.

Innymi słowy, wykazaliśmy stochastyczną zbieżność ciągu $\left\{ \frac{1}{n} \sum_{k=1}^n \xi_k \right\}$ do wektora $\frac{1}{n} \sum_{k=1}^n m(\xi_k)$.

3.3. Uogólnione prawo Chinczyna

W tej części pracy wykorzystamy uogólnioną (wielowymiarową) postać funkcji charakterystycznej rozkładu prawdopodobieństwa zaproponowaną w pracy [Tatar 2004].

Definicja 5 [Tatar 2004]. Łączną funkcją charakterystyczną wektora losowego $\xi: \Omega \rightarrow R^s$ nazywamy funkcję $\varphi: R \rightarrow C$ postaci $\forall t \in R: \varphi(t) = E(e^{i \cdot t \cdot \xi})$.

Występującą w powyższej definicji wektorową potęgę dodatniej liczby rzeczywistej a rozumiemy (por. także [Tatar 2004]) jako każde odwzorowanie postaci:

$$d: V \times R_+ \rightarrow R_+$$

spełniające warunki:

- (i) $d(e_+, a) = 1$,
- (ii) $\forall v, w \in V: d(v + w, a) = d(v, a) \cdot d(w, a)$,
- (iii) $\forall v \in V, k \in R: d(k \cdot v, a) = [d(v, a)]^k$,
- (iv) $\forall v, w \in V: d(w, d(v, a)) = a^{(v/w)}$.

W dowodzie głównego w tej części pracy twierdzenia użyteczne także będą następujące dwa lematy.

Lemat 1. Niech dany będzie wektor losowy $\xi: \Omega \rightarrow R^s$ o funkcji charakterystycznej φ_ξ . Niech ponadto $a \in R$ oraz $v_o \in R^s$. Wówczas funkcja charakterystyczna wektora losowego $\gamma = a \cdot \xi + v_o$ jest postaci $\varphi_\gamma(t) = e^{i \cdot t \cdot v_o} \cdot \varphi_\xi(a \cdot t)$.

Dowód. Wykorzystując przypomnianą powyżej definicję 5 oraz własności wartości oczekiwanej, otrzymujemy:

$$\varphi_\gamma(t) = E(e^{i \cdot t \cdot \eta}) = E(e^{i \cdot t \cdot (a \cdot \xi + v_o)}) = E(e^{i \cdot t \cdot a \cdot \xi} \cdot e^{i \cdot t \cdot v_o}) = e^{i \cdot t \cdot v_o} \cdot E(e^{i \cdot (a \cdot t) \cdot \xi}) = e^{i \cdot t \cdot v_o} \cdot \varphi_\xi(a \cdot t).$$

Lemat 2. Niech dane będą niezależne wektory losowe $\xi_1, \xi_2, \dots, \xi_n: \Omega \rightarrow R^s$ o funkcjach charakterystycznych $\varphi_{\xi_1}, \varphi_{\xi_2}, \dots, \varphi_{\xi_n}$. Niech ponadto $a_1, a_2, \dots, a_n \in R$. Wówczas funkcja charakterystyczna wektora losowego $\gamma = \sum_{i=1}^n a_i \cdot \xi_i$ jest postaci $\varphi_\gamma(t) = \prod_{i=1}^n \varphi_{\xi_i}(a_i \cdot t)$.

Dowód. Z własności działania określonego w definicji 5 mamy:

$$\varphi_\gamma(t) = E(e^{i \cdot t \cdot (a_1 \cdot \xi_1 + a_2 \cdot \xi_2 + \dots + a_n \cdot \xi_n)}) = E(e^{i \cdot t \cdot a_1 \cdot \xi_1} \cdot e^{i \cdot t \cdot a_2 \cdot \xi_2} \cdot \dots \cdot e^{i \cdot t \cdot a_n \cdot \xi_n}).$$

Korzystając następnie z niezależności wektorów ξ_1, ξ_2 otrzymujemy:

$$\varphi_\gamma(t) = E(e^{i \cdot t \cdot a_1 \cdot \xi_1}) \cdot E(e^{i \cdot t \cdot a_2 \cdot \xi_2}) \cdot \dots \cdot E(e^{i \cdot t \cdot a_n \cdot \xi_n}) = \varphi_{\xi_1}(a_1 \cdot t) \cdot \varphi_{\xi_2}(a_2 \cdot t) \cdot \dots \cdot \varphi_{\xi_n}(a_n \cdot t),$$

czyli żądaną tezę.

Przechodzimy do sformułowania oraz dowodu twierdzenia będącego głównym wynikiem tej części pracy.

Twierdzenie 3 (uogólnione słabe prawo wielkich liczb Markowa). Niech $\xi_k: \Omega \rightarrow R^s$ ($k = 1, 2, 3, \dots$) będzie ciągiem niezależnych s -wymiarowych wektorów losowych o jednakowych rozkładach z wartością oczekiwaną $m = E(\xi_k) \equiv const$. Wówczas:

$$\frac{1}{n} \sum_{k=1}^n \xi_k \xrightarrow{p} m.$$

Dowód. Niech φ oznacza funkcję charakterystyczną wektora losowego $\xi_k - m$ (taką samą dla wszystkich $k = 1, 2, 3, \dots$).

W tej sytuacji – wobec lematów 1 i 2 – funkcja charakterystyczna wektora losowego $\eta_n = \frac{1}{n} \sum_{k=1}^n (\xi_k - m)$ jest postaci:

$$\varphi_{\eta_n}(t) = \left[\varphi\left(\frac{1}{n} \cdot t\right) \right]^n.$$

Korzystając z rozwinięcia funkcji φ w punkcie $\frac{1}{n} \cdot t$ w szereg Maclaurina, otrzymujemy:

$$\varphi_{\eta_n}(t) = \left[\varphi(0) + \varphi'(0) \cdot \frac{t}{n} + o\left(\frac{t}{n}\right) \right]^n.$$

Z kolei, wobec własności funkcji charakterystycznej φ oraz jej pierwszej pochodnej φ' mamy:

$$\varphi_{\eta_n}(t) = \left[1 + E(i \cdot (\xi_k - m)) \cdot \frac{t}{n} + o\left(\frac{t}{n}\right) \right]^n,$$

czyli

$$\varphi_{\eta_n}(t) = \left[1 + o\left(\frac{t}{n}\right) \right]^n,$$

a więc także

$$\ln \varphi_{\eta_n}(t) = n \cdot \ln \left[1 + o\left(\frac{t}{n}\right) \right].$$

Wykorzystując powtórnie rozwinięcie funkcji w szereg Maclaurina (tym razem funkcji $f(z) = \ln(1+z)$, czyli $\ln(1+z) = z + o(z)$), otrzymujemy:

$$\ln \varphi_{\eta_n}(t) = n \cdot \left[o\left(\frac{t}{n}\right) \right].$$

Stąd

$$\lim_{n \rightarrow \infty} \ln \varphi_{\eta_n}(t) = \lim_{n \rightarrow \infty} n \cdot \left[o\left(\frac{t}{n}\right) \right] = 0,$$

czyli także

$$\lim_{n \rightarrow \infty} \varphi_{\eta_n}(t) = 1.$$

Skoro ciąg funkcji charakterystycznych $\{\varphi_{\eta_n}(t)\}$ zmierza do funkcji charakterystycznej rozkładu jednopunktowego skoncentrowanego w zerze, więc ciąg dystrybuant wektorów losowych $\{\eta_n\}$ jest zbieżny do dystrybuanty tego rozkładu (jednopunktowego skoncentrowanego w zerze). Wynika stąd, że ciąg wektorów losowych $\{\xi_k - m\}$ jest stochastycznie (tzn. według prawdopodobieństwa) zbieżny do zera. Oznacza to, że $\frac{1}{n} \sum_{k=1}^n (\xi_k - m) \xrightarrow{P} 0$, czyli $\frac{1}{n} \sum_{k=1}^n \xi_k \xrightarrow{P} m$. Otrzymaliśmy więc tezę dowodzonego twierdzenia.

4. Podsumowanie

W pracy zaproponowano uogólnienie na przypadek wielowymiarowy dwóch twierdzeń (znanych dla zmiennych losowych jednowymiarowych) dotyczących zbieżności stochastycznej, czyli zbieżności według prawdopodobieństwa, tj. słabych praw wielkich liczb Markowa i Chinczyna. Twierdzenia te orzekają, że jeżeli ciąg wektorów losowych spełnia stosowne założenia, to ciąg ich średnich arytmetycznych jest zbieżny według prawdopodobieństwa do średniej arytmetycznej ich wartości oczekiwanych. W przeprowadzonych dowodach wykorzystano „łączne momenty rozkładów prawdopodobieństwa wektorów losowych”, które – dzięki nowemu podejściu – są rzeczywiście charakterystykami wektorów losowych nie zaś jednowymiarowych zmiennych losowych będących funkcjami współrzędnych badanych wektorów.

Literatura

- Budny K. [2009], *Kurtoza wektora losowego*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, nr 78, seria: Ekonometria, nr 26.
- Budny K., Tatar J. [2009], *Kurtosis of a Random Vector – Special Types of Distributions*, „Statistics in Transition – New Series”, vol. 10, nr 3.
- Budny K., Tatar J. [2012], *Regresja liniowa z wykorzystaniem nowej definicji momentów wektorów losowych*, „Zeszyty Naukowe Uniwersytetu Ekonomicznego w Krakowie”, nr 892.
- Budny K., Szklarska M., Tatar J. [2014], *Wielowymiarowa analiza sytuacji społeczno-demograficznej Polski [w:] 50 lat kształcenia ekonomistów w Kielcach*, red. E. Molendowski i A. Szplit, seria: Studia i Materiały. Miscellanea Oeconomicae, R. 18, nr 1, Kielce.
- Budny K., Tatar J. [2014], *Charakterystyki wielowymiarowych wielkości finansowych oparte na definicji potęgi wektora [w:] Metody wnioskowania statystycznego w badaniach ekonomicznych*, red. J. Kolonko, Uniwersytet Ekonomiczny w Katowicach, Katowice.
- Budny K. [2014a], *A Generalization of Chebyshev's Inequality for Hilbert-space-valued Random Elements*, „Statistics and Probability Letters”, vol. 88, <https://doi.org/10.1016/j.spl.2014.01.021>.
- Budny K. [2014b], *An Extension of the Multivariate Chebyshev's Inequality to a Random Vector with a Singular Covariance Matrix*, „Communication in Statistics – Theory and Methods”, vol. 45, nr 17, <https://doi.org/10.1080/03610926.2014.941499>.
- Feller W. [1969], *Wstęp do rachunku prawdopodobieństwa*, t. 1 i 2, PWN, Warszawa.
- Fisz M. [1969], *Rachunek prawdopodobieństwa i statystyka matematyczna*, PWN, Warszawa.
- Najman P., Tatar J. [2010], *Regresja wektorów losowych dla wielowymiarowego rozkładu normalnego [w:] Badania ekonometryczne w teorii i praktyce*, red. A.S. Barczak, Wydawnictwo Uniwersytetu Ekonomicznego w Katowicach, Katowice.

- Osiewalski J., Tatar J. [1997], *Silna wersja uogólnionej nierówności Czebyszewa*, Materiały z XV Seminarium Naukowego im. Prof. Zbigniewa Pawłowskiego, Akademia Ekonomiczna we Wrocławiu, Wrocław.
- Osiewalski J., Tatar J. [1999], *Multivariate Chebyshev Inequality Based on a New Definition of Moments of a Random Vector*, „Przegląd Statystyczny”, nr 2.
- Plucińska A., Pluciński E. [2000], *Rachunek prawdopodobieństwa. Statystyka matematyczna. Procesy stochastyczne*, WNT, Warszawa.
- Tatar J. [1993], *Moments of a Random Variable in a Hilbert Space*, Discussion Paper, No. 1, Cracow Academy of Economics.
- Tatar J. [1996a], *Nierówność Czebyszewa dla wielowymiarowych zmiennych losowych*, „Badania Operacyjne i Decyzje”, nr 2.
- Tatar J. [1996b], *O niektórych miarach rozproszenia rozkładów prawdopodobieństwa*, „Przegląd Statystyczny”, nr 3–4.
- Tatar J. [2000], *Asymetria wielowymiarowych rozkładów prawdopodobieństwa*, Materiały z XVIII Seminarium Naukowego im. Prof. Zbigniewa Pawłowskiego, Akademia Ekonomiczna w Krakowie, Kraków.
- Tatar J. [2001], *Momenty absolutne wielowymiarowych rozkładów prawdopodobieństwa*, Sprawozdania z Posiedzeń Komisji Naukowych, PAN, Oddział w Krakowie, T. 43/2, Kraków (streszczenie).
- Tatar J. [2002], *Nierówność Lapunowa dla wielowymiarowych rozkładów prawdopodobieństwa*, „Zeszyty Naukowe Akademii Ekonomicznej w Krakowie”, nr 549.
- Tatar J. [2003], *Prawa wielkich liczb dla wielowymiarowych wektorów losowych [w:] Zastosowania statystyki i matematyki w ekonomii*, red. W. Ostasiewicz, Wydawnictwo Akademii Ekonomicznej we Wrocławiu, Wrocław.
- Tatar J. [2004], *Funkcje charakterystyczne wielowymiarowych wektorów losowych*, Materiały z XXXVIII Konferencji Statystyków, Ekonometryków i Matematyków Akademii Ekonomicznych Polski Południowej, Akademia Ekonomiczna w Krakowie, Kraków.
- Tatar J. [2006], *Późnizmienniki i momenty w charakteryzacji wielowymiarowych rozkładów prawdopodobieństwa [w:] Matematyka – język uniwersalny. Księga jubileuszowa dla uczczenia 70. urodzin Profesora Tadeusza Stanisza*, Wydawnictwo Akademii Ekonomicznej w Krakowie, Kraków.
- Tatar J. [2008a], *Korelacja wektorów losowych o dowolnych wymiarach [w:] Postępy statystyki, ekonometrii i matematyki stosowanej w Polsce Południowej*, red. A. Żeliaś, J. Pocięcha, Uniwersytet Ekonomiczny w Krakowie, Kraków.
- Tatar J. [2008b], *Miary zależności wektorów losowych o różnych wymiarach*, „Zeszyty Naukowe Uniwersytetu Ekonomicznego w Krakowie”, nr 780.
- Tatar J. [2009], *Nowe charakterystyki warunkowych rozkładów wielowymiarowych*, „Studia i Prace Uniwersytetu Ekonomicznego w Krakowie”, nr 3.
- Tatar J. [2013], *Modele wskaźnikowe rynku kapitałowego wykorzystujące funkcję regresji wektorów losowych*, „Zeszyty Naukowe Uniwersytetu Ekonomicznego w Krakowie”, nr 923, <https://doi.org/10.15678/ZNUEK.2013.0923.03>.

Stochastic Convergence of Sequences of Random Vectors

(Abstract)

The paper presents a multidimensional generalisation (known for one-dimensional random variables) of two theorems regarding stochastic convergence – that is, convergence by probability. The generalised theorems are Markov's and Chinchyn's weak laws of great numbers. Both lead to the theory that, with the appropriate assumptions, a sequence of arithmetic averages of the random vectors converges their expected values to the arithmetic average. The proof for this thesis uses „whole moments of the multidimensional probability distribution”, which the author has proposed elsewhere. Their basis is a definition of the power of a vector in a space with a scalar product.

Keywords: power of vector, moment of probability distribution, random vector, stochastic convergence.

ISSN 1898-6447

UNIwersYTET EKONOMICZNY W KRAKOWIE
31-510 Kraków, ul. Rakowicka 27
WYDAWNICTWO
www.uek.krakow.pl

Zamówienia na wydane prace przyjmuje
Księgarnia Uniwersytetu Ekonomicznego w Krakowie
tel. 12 293-57-40, fax 12 293-50-11
e-mail: ksiegarnia@uek.krakow.pl