METODY WNIOSKOWANIA STATYSTYCZNEGO W BADANIACH EKONOMICZNYCH

Studia Ekonomiczne

ZESZYTY NAUKOWE
WYDZIAŁOWE

UNIWERSYTETU EKONOMICZNEGO

W KATOWICACH

METODY WNIOSKOWANIA STATYSTYCZNEGO W BADANIACH EKONOMICZNYCH

Redaktor naukowy
Janusz L. Wywiał

Katowice 2013

Redaktor
Magdalena Pazura

Skład
Krzysztof Słaboń

ISSN 2083-8611

Wersją pierwotną Studiów Ekonomicznych jest wersja papierowa

WYDAWNICTWO UNIWERSYTETU EKONOMICZNEGO W KATOWICACH
ul. 1 Maja 50, 40-287 Katowice, tel. 32 257-76-30, fax 32 257-76-43
www.wydawnictwo.ue.katowice.pl, e-mail: wydawnictwo@ue.katowice.pl

SPIS TREŚCI

WSTĘP

Niniejszy zeszyt naukowy ma charakter specjalnego wydania Studiów Ekonomicznych z okazji przypadającego na 2012 r. jubileuszu siedemdziesiątych urodzin profesora Uniwersytetu Ekonomicznego doktora habilitowanego Józefa Kolonki. Autorzy dedykują Profesorowi swoje artykuły. Stanowi to rodzaj uszanowania dorobku Profesora oraz jego zasług akademickich. Wśród autorów są przyjaciele Profesora Józefa Kolonki lub Jego uczniowie, którzy tym sposobem wyrażają swoją wdzięczność za jego naukowe wskazówki.

Profesor Józef Kolonko jest znaną i szanowaną osobowością Uniwersytetu Ekonomicznego w Katowicach. Cenione są jego zasługi na polach naukowo-badawczym i dydaktycznym. Szeroko o zasługach Profesora pisano niedawno przy okazji wręczania Mu Medalu Uniwersytetu Ekonomicznego w Katowicach. Tutaj nakreślimy jedynie Jego drogę zawodową. Profesor Józef Kolonko ukończył studia z tytułem magistra ekonomii w 1965 r. w ówczesnej Wyższej Szkole Ekonomicznej (WSE) w Katowicach na Wydziale Przemysłu. W tym samym roku odbył staż asystencki w WSE w Katowicach, a następnie został przyjęty na stanowisko asystenta do Katedry Statystyki, wówczas kierowanej przez profesora doktora habilitowanego Zbigniewa Pawłowskiego, który jest niezwykle zasłużoną postacią dla Uniwersytetu Ekonomicznego w Katowicach. W 1967 r. awansował na stanowisko starszego asystenta. Rozprawę doktorską na temat: *Ekonometryczny model kształtowania się płac w wybranym przedsiębiorstwie przemysłu maszynowego* przygotował pod kierunkiem profesora Zbigniewa Pawłowskiego. W wyniku pomyślnej obrony tej rozprawy doktorskiej ówczesna Rada Wydziału Przemysłu w 1971 r. nadała Józefowi Kolonce tytuł doktora nauk ekonomicznych. W tym samym roku został on awansowany na stanowisko adiunkta. Pracę habilitacyjną przygotowywał m.in. przebywając na kilkumiesięcznym stażu naukowym w Netherlands Economic Institute (Holenderskim Instytucie Ekonomicznym) w Rotterdamie, gdzie studiował pod kierunkiem późniejszego doktora honoris causa Akademii Ekonomicznej w Katowicach – profesora Jeana H.P. Paelincka. W 1980 r. ukazała się praca Józefa Kolonki pt.: *Analiza dyskryminacyjna i jej zastosowanie w ekonomii*, wydana przez renomowane Państwowe Wydawnictwo Naukowe w Warszawie. Doprowadziło to do nadania Mu w 1981 r. tytułu naukowego doktora habilitowanego przez Radę Wydziału Przemysłu Akademii Ekonomicznej (AE) w Katowicach.

Finalizowanie przewodu habilitacyjnego zbiegło się z okresem wielkich przemian społeczno-politycznych w Polsce, czego m.in. wyrazem było powstanie Niezależnego Samorządnego Związku Zawodowego „Solidarność" w Akademii Ekonomicznej w Katowicach, do czego w znaczący sposób przyczynił się doktor Józef Kolonko.

Działalność związkowa ówcześnie już dr. hab. Józefa Kolonki bynajmniej nie przeszkodziła w Jego aktywności na polach naukowym i dydaktycznym. W 1981 r. został mianowany dyrektorem Instytutu Ekonometrii. W 1982 r. objął stanowisko docenta, natomiast dyrektorem Instytutu był do 1987 r. Dodajmy, że był to czas wielkich wyzwań dotyczących m.in. zmian merytorycznych oraz instytucjonalnych związanych z działalnością naukowo-dydaktyczną, które w konsekwencji po stanie wojennym doprowadziły do reformowania szkolnictwa wyższego. W 1991 r. dr. hab. Józefowi Kolonce nadano stanowisko profesora nadzwyczajnego Akademii Ekonomicznej w Katowicach w Instytucie Ekonometrii. W latach 1993-2000 profesor Józef Kolonko kierował nowo utworzonym Zakładem Ekonomicznych Zastosowań Statystyki, w Instytucie Ekonometrii. Następnie, począwszy od 2000 r. w ramach Katedry Statystyki kierował kolejno Zakładem Metod i Analiz Danych Statystycznych oraz Zakładem Statystycznej Analizy Jakości Danych. Profesor Józef Kolonko był promotorem prac magisterskich oraz doktorskich. Rozprawy doktorskie napisane pod jego kierunkiem charakteryzowały się bardzo wysokim poziomem naukowym. Miało to korzystny wpływ na dalszy rozwój naukowo-badawczy autorów tych prac.

Z merytorycznego punktu widzenia profesor Józef Kolonko intensywnie zajmował się szeroko rozumianymi metodami klasyfikacji oraz wykorzystaniem metod statystycznych i ekonometrycznych w analizach finansowych i przedsięwzięciach inwestycyjnych. Wiele osiągnął również w zakresie analizy jakości danych statystycznych, m.in. w kontekście budowy i wdrażania systemów utrzymania jakości kształcenia w szkołach wyższych. Profesor Józef Kolonko kierował wieloma zespołami osób, które zrealizowały około 20 problemów naukowo-badawczych. Efektem tych przedsięwzięć oraz innej działalności naukowo-badawczej było około 100 prac naukowo-badawczych, ekspertyz ekonomicznych, a w tym 30 opublikowanych artykułów i książek. O wysokiej jakości jego działalności świadczy m.in. przyznana mu w 1978 r. nagroda Ministra Nauki i Szkolnictwa Wyższego stopnia III. W latach 1998-2010 otrzymał siedemnaście nagród indywidualnych lub zespołowych rektora. Zasługi naukowo-badawcze Józefa Kolonki zostały docenione w gronie statystyków polskich, jako że był członkiem Komitetu Ekonometrii i Statystyki oraz członkiem Komisji Matematycznej Głównego Urzędu Statystycznego. Profesor Józef Kolonko cieszy się wielką popularnością w społeczności akademickiej Uniwersytetu Ekonomicznego w Katowicach; wybierano Go do różnych znaczących kolegialnych organów

Uczelni, w szczególności do Senatu Akademii Ekonomicznej w siedmiu kadencjach. Pełnił funkcję przewodniczącego Senackiej Komisji ds. Rozwoju w trzech kadencjach. Był również przewodniczącym Senackiej Komisji ds. Współpracy Międzynarodowej. Pełnił funkcję przewodniczącego Uczelnianej Komisji Wyborczej. Był także członkiem Senackiej Komisji ds. Nowelizacji Statutu, Komisji Oceny Ryzyka Realizacji Projektów oraz przewodniczącym Uczelnianej Komisji ds. Studiów Podyplomowych. Zasługi profesora Józefa Kolonki doceniono m.in. poprzez przyznanie Mu Krzyża Kawalerskiego Orderu Odrodzenia Polski, Brązowego i Złotego Krzyża Zasługi, Złotego Medalu za Długoletnią Służbę.

Profesor Józef Kolonko był inicjatorem założenia Śląskiej Międzynarodowej Szkoły Handlowej (ŚMSH). Kierował nią jako dyrektor w latach 1991-2004. Szkoła jest prowadzona przez Uniwersytet Ekonomiczny oraz Uniwersytet Śląski w Katowicach we współpracy z Ecole Superier du Comerse de Tuluse oraz University of Strathclyde in Glasgow. Zasługi profesora Kolonki związane z utworzeniem i rozwojem ŚMSH doceniły władze Uniwersytetu Śląskiego, przyznając Mu medal: Zasłużony dla Uniwersytetu Śląskiego. Profesor Józef Kolonko był również w gronie inicjatorów założenia Górnośląskiej Wyższej Szkoły Przedsiębiorczości (GWSH) w 2005 r. Profesor zajmował się działalnością inicjującą rozwój na Górnym Śląsku. Ma to odzwierciedlenie m.in. w fakcie, iż był jednym z założycieli i w ciągu 5 lat przewodniczącym Rady Nadzorczej Górnośląskiego Banku Gospodarczego, aktualnie Getin Banku. Te zasługi dostrzeżono, przyznając Profesorowi w 1978 r. Złoty Krzyż Zasługi dla województwa śląskiego. O Jego profesjonalizmie w tej dziedzinie świadczy fakt, że został powołany w 2008 r. na członka Rady Naukowej Narodowego Banku Polskiego.

Widzimy więc ogromną aktywność profesora Józefa Kolonki, począwszy od działalności naukowo-badawczej i dydaktycznej poprzez ekspercką po aktywność na polu przedsiębiorczości. Może więc stanowić on wzór dla młodszych pokoleń ekonomistów polskich.

Janusz L. Wywiał

Józef Biolik
Uniwersytet Ekonomiczny w Katowicach

DYLEMATY MODELOWANIA GOSPODARKI NARODOWEJ

Wprowadzenie

Jednym z obszarów aktywności naukowej Profesora Józefa Kolonki w latach 1985-1990 był najpierw udział, a następnie kierowanie zespołem badawczym zajmującym się budową ekonometryczno-symulacyjnego średniookresowego modelu gospodarki Polski, a w dalszym etapie konstrukcją modeli wspomagających planowanie gospodarcze[*].

Zmiana systemu gospodarczego w 1990 r. spowodowała, że prace związane z modelowaniem gospodarki narodowej przerwano, a wartościowych rezultatów tych badań z punktu widzenia metodologicznego nie opublikowano.

1. Koncepcja budowy modelu gospodarki narodowej

Prace związane z budową modelu gospodarki Polski rozpoczęły się w 1982 r. W początkowym etapie pracami kierowała dr Elżbieta Stolarska. Zbudowany model MES-1 zawierał następujące sfery:

- zatrudnienia,
- produkcji,
- dochodu narodowego,
- funduszu pieniężnego,
- spożycia
- wymiany zagranicznej.

Był to model rekurencyjny, dynamiczny. W pierwszym etapie został oszacowany na podstawie danych z lat 1961-1981. Opierając się na tych równaniach, obliczono prognozy zmiennych endogenicznych na 1982 r. i porównano ich wartości

[*] W ramach problemów badawczych MR oraz C.P.B.P, których koordynatorem był Instytut Badań Systemowych PAN w Warszawie.

z zaobserwowanymi w tym samym roku realizacjami. Niektóre spośród otrzymanych prognoz można było uznać za wystarczająco dokładne, lecz znaczna ich część charakteryzowała się dużymi błędami prognoz (równania bloku dochodu narodowego i wymiany z zagranicą oraz blok spożycia).

Przed przystąpieniem do dalszych badań dokonano reestymacji modelu, dołączając do próby dane za kolejne lata: 1982 oraz 1983. Przedmiotem wstępnej analizy były zmiany ocen parametrów oraz zmiany współczynnika determinacji R^2.

Będący przedmiotem analizy w 1985 r. model MES (Model ekonometryczno-symulacyjny) zawierał następujące sfery gospodarki:
- blok równań zatrudnienia – 14 równań,
- blok równań produkcji – 12 równań,
- blok równań tworzenia oraz podziału produktu globalnego i dochodu narodowego – 6 równań,
- blok równań spożycia i przyrostu oszczędności – 10 równań.

Model został oszacowany na podstawie danych z lat 1961-1983. Do oceny jego jakości zastosowano:
- ocenę adekwatności replikatywnej,
- ocenę adekwatności prognostycznej,
- ocenę adekwatności strukturalnej.

Model cechuje się wysoką adekwatnością replikatywną, jeśli w ramach próby pozwala z dużą dokładnością odtworzyć zaobserwowane wartości zmiennych endogenicznych. Klasycznym przykładem miary adekwatności replikatywnej jest współczynnik determinacji R^2. Należy jednak zauważyć, że wysoki stopień adekwatności replikatywnej nie gwarantuje ani dobrych prognoz, ani adekwatności strukturalnej.

Model ma wysoką adekwatność prognostyczną, jeśli na podstawie zadanych wartości zmiennych objaśniających pozwala on dostatecznie dokładnie przewidywać przyszłe wartości zmiennych objaśnianych przez poszczególne równania modelu. Do oceny poziomu adekwatności prognostycznej można wykorzystać mierniki dokładności prognoz ex ante oraz ex post.

Model adekwatny strukturalnie powinien nie tylko dawać wystarczająco dokładne prognozy, ale przede wszystkim dobrze odtwarzać rzeczywisty mechanizm tworzenia tych wartości. Należy zauważyć, że wysoki stopień adekwatności replikatywnej nie gwarantuje ani dobrych prognoz, ani adekwatności strukturalnej. W trakcie analiz pojawiła się idea, by w klasie modeli o „dopuszczalnej" adekwatności strukturalnej i w związku z tym o dobrych własnościach prognostycznych, poszukiwać modeli o stabilnych parametrach (niezmieniających się w obszarze obserwacji bez względu na dobór obserwacji w ramach dostępnej próby).

Wstępnym wymaganiem adekwatności strukturalnej mogą być tylko dobre prognozy. Spełnienie tego postulatu pozwala w drugiej kolejności podjąć badanie stabilności parametrów.

Wrażliwość modelu ekonometrycznego ujawnia się m.in. w jego reakcji na zmiany wartości zmiennych objaśniających poza obszarem próby statystycznej, na podstawie której został oszacowany. Pożądaną w tym sensie cechą modelu jest, aby prognozy obliczone na jego podstawie były trafne. Oznacza to, że istnieje wymóg, aby model cechował się wysokim stopniem adekwatności prognostycznej. Do oceny tej adekwatności posłużyła analiza błędów prognoz. W analizie przyjęto następujące określenia:

- prognozy wystarczająco dokładne to takie, których błąd względny nie przekracza 5%;
- prognozy średnio dokładne to takie, których błąd względny mieści się w granicach 5%-10%;
- prognozy mało dokładne to takie, których błąd względny mieści się w granicach 10%-25%;
- prognozy złe to takie, których względny błąd prognozy przekracza poziom 25%.

Tabela 1

Struktura błędów prognoz obliczonych na podstawie modelu MES

Rodzaj prognoz ze względu na błędy	Prognozy na 1982 r.		Prognozy na 1983 r.		Prognozy na 1984 r.	
	Ilość	%	Ilość	%	Ilość	%
Wystarczająco dokładne	29	50,0%	23	39,66%	25	43,10%
Średnio dokładne	7	12,07%	9	15,52%	7	12,07%
Mało dokładne	12	20,69%	14	24,14%	14	24,14%
Złe	10	17,24%	12	20,69%	12	20,69%

Źródło: Obliczenia własne.

Prognozy obarczone dużym względnym błędem dotyczyły produkcji globalnej, eksportu, importu, dochodu narodowego i inwestycji oraz spożycia z dochodów osobistych.

Okres prognozowany (lata 1982-1984) dotyczył lat kryzysu, gospodarki niedoboru, więc uzyskanie prognoz trafnych było mało prawdopodobne. W trakcie prac nad ekonometryczno-symulacyjnym modelem gospodarki Polski przeprowadzono analizę stabilności parametrów modelu, opartą na wielokrotnym szacowaniu parametrów poszczególnych równań modelu na podstawie „pełzająco" zmieniających się zbiorów informacji statystycznych. Na wstępie zbiory n informacji dotyczących każdej ze zmiennych modelu podzielono na $(n - m + 1)$ podzbiorów liczących po m kolejnych elementów.

Konkretnie wyróżniono po 10 podzbiorów zawierających 14 obserwacji. Następnie oszacowano parametry odpowiednich równań modelu, uzyskując po 10 ocen każdego z parametrów występujących w modelu. W wielu przypadkach zaobserwowany rozstęp wartości części parametrów wielokrotnie przewyższał wartości ich średnich błędów szacunku. W tych warunkach konieczne stało się zbadanie, czy zmiany wartości parametrów mają charakter regularny i jaki jest ich kierunek. Oszacowano więc dla każdego parametru współczynniki trendu wielomianowego stopnia pierwszego, drugiego, a także modelu autoregresyjnego. Najlepsze wyniki (ze względu na zgodność) uzyskano dla trendu wielomianowego stopnia drugiego, natomiast najgorsze dla modeli autoregresyjnych. Na 115 badanych parametrów w 38 przypadkach R^2_w przekraczał wartość 0,90; w 15 przypadkach $R^2 \in (0{,}80, 0{,}90)$, w 16 przypadkach $R^2 \in (0{,}70, 0{,}80)$ oraz w 12 przypadkach $R^2 \in (0{,}6, 0{,}7)$. W świetle uzyskanych wyników niezbędne wydało się wprowadzenie do modelu uzmiennionych parametrów.

Wyniki badania doprowadziły do wniosku, że wbrew dobrym własnościom replikatywnym w obrębie próby, niektóre równania nie dają dobrych prognoz, ponadto część równań nie spełniała wymogu stabilności parametrów. W związku z dezaktualizacją części równań stwierdzono, że model nie wykazuje zadowalającego stopnia adekwatności strukturalnej. Przyczyną dezaktualizacji może być:

- pojawienie się nowych czynników wpływających na wielkość objaśnianą, których wcześniejsza specyfikacja nie była możliwa albo nie wydawała się konieczna;
- nieuwzględnienie w modelu zmienności parametrów;
- przyjęcie niewłaściwej postaci analitycznej – takiej, która w obszarze próby niewiele różniła się od prawidłowej.

Należy zauważyć, że zmienność parametrów w czasie może się ujawnić właśnie wskutek niewyspecyfikowania nieznanych lub trudno mierzalnych czynników.

Po stwierdzeniu utraty aktualności modelu najczęściej:

- wprowadza się nowe zmienne objaśniające,
- uzmiennia się parametry strukturalne,
- dokonuje się zmiany postaci analitycznej,
- dokonuje się nowej specyfikacji całego modelu według nowej koncepcji.

Wprowadzenie nowych zmiennych w trakcie modyfikacji modelu uzasadnione jest wówczas, gdy:

- specyfikując model, przeoczono jakąś istotną zmienną, nie doceniano jej wagi, traktując ją jako czynnik mało znaczący;
- pojawił się wpływ takich czynników, które w okresie próby rzeczywiście nie odgrywały dużej roli, albo w ogóle nie istniały;
- uznano konieczność wprowadzenia nowej zmiennej wcześniej branej pod uwagę, ale trudno mierzalnej; w tej sytuacji konstruktorzy modelu często posługują się zmiennymi zerojedynkowymi, traktując ich użycie jako zastępczy, niedoskonały sposób wyrażania czynnika o charakterze jakościowym.

W odniesieniu do modelu MES-1 autorzy uznali, że przeprowadzone badania nie podważyły adekwatności modelu w początkowym okresie próby. Model ten miał duże walory poznawcze, w latach 1961-1981 z dużą dokładnością przybliżał rzeczywiste obserwacje wyjaśnianych zmiennych. Nie było też wystarczających podstaw do kwestionowania jego adekwatności strukturalnej w tym okresie. Uzyskane wyniki poddały w wątpliwość jego adekwatność w ostatnich latach próby oraz w latach 1982-1984.

Utrata aktualności modelu miała związek z faktem, że model MES-1 odwzorowywał badany mechanizm, zakładając, że jest on w przybliżeniu stabilny. Tymczasem w ostatnich latach okresu 1960-1984 nastąpiły znaczne zmiany mechanizmu funkcjonowania gospodarki. Autorzy zrezygnowali z prób modyfikowania modelu i zdecydowali się na budowę nowego modelu.

2. Koncepcja modelu MGS

Model MGS (Model Górny Śląsk) głębiej wnikał w sferę podziału dochodu narodowego, zaniedbując szczegółową analizę spożycia. Przesunięcie ciężaru modelu w kierunku podziału dochodu było uzasadnione ówczesną sytuacją gospodarki, a zwłaszcza oceną możliwości jej unowocześnienia przy jednoczesnym zachowaniu rosnącego ciągle zadłużenia. Model MGS opisywał gospodarkę narodową Polski w latach 1960-1983. Był to model nieliniowy, dynamiczny (z dwuokresowymi opóźnieniami) rekurencyjny. Wystąpiło w nim 47 zmiennych endogenicznych oraz 17 egzogenicznych. Model MGS był więc w porównaniu z modelem MES-1 bardziej przydatny do badania skutków decyzji makroekonomicznych, co było szczególnie ważne w świetle zapowiadanej restrukturyzacji gospodarki Polski i konieczności spłat długu zagranicznego.

Bardzo dobre formalne własności modelu MGS uzasadniały jego wykorzystanie dla celów prognostycznych i symulacyjnych. Zarówno prognozy obarczone małymi błędami, jak i wyniki symulacji w okresie 1984-1990 uzasadniały dobre walory modelu MGS.

Model MGS składał się z równań pogrupowanych w bloki:
- zatrudnienia,
- produkcji,
- handlu zagranicznego,
- dochodu narodowego,
- podziału dochodu narodowego i spożycia.

Model MGS zastosowano do symulacji w latach 1984-1990. Różnorodne scenariusze symulacyjne przedstawiały dopuszczalne decyzje przy wielu wariantach zmian parametrów, które interpretowano jako zmiany efektywności bądź udziałów.

Wyniki symulacji bazowej* potwierdziły znany obraz i własności ówczesnej gospodarki Polski – silna inercja oraz zdominowanie przez kompleks paliwowo-energetyczny. Analiza wyników symulacji bazowej na 1990 r. wskazywała, że nie pojawiły się eksplozyjne zmiany żadnych wielkości. Symulacyjne przebiegi zmiennych endogenicznych wykazywały wewnętrzną zgodność. Innym uzasadnieniem przydatności modelu MGS były jego własności prognostyczne, obliczone na podstawie modelu.

Rozkład względnych błędów prognoz na 1984 r. przedstawia tab. 2.

Tabela 2

Względne błędy prognoz oraz prognoz skorygowanych, dotyczące 1984 r.

Bloki zmiennych endogenicznych	Wielkość względnego błędu prognozy			
	0%-2%	2%-5%	5%-10%	Powyżej 10%
	Liczba przypadków			
Prognozy zatrudnienia w działach i sferach gospodarki	6	5	5	0
Skorygowane prognozy zatrudnienia	11	3	2	0
Prognozy produkcji globalnej w działach i gałęziach	2	5	3	4
Skorygowane prognozy produkcji globalnych	8	3	2	1
Prognozy ważniejszych makroekonomicznych charakterystyk gospodarki	2	1	3	2
Skorygowane prognozy makroekonomicznych charakterystyk gospodarki narodowej	2	2	2	2

Nota: Korekta polegała na zmianie parametrów opisujących efektywność pracy żywej zgodnie z zaobserwowanym w ostatnim okresie wzrostem wydajności.

Źródło: Obliczenia własne.

Na podstawie oszacowanego modelu MGS przeprowadzono analizę symulacyjną, zakładając cztery scenariusze symulacyjne, od skrajnie pesymistycznego do skrajnie optymistycznego.

Wyniki symulacji w zasadzie potwierdzały oczekiwania – ówczesna gospodarka Polski była sztywna, mało podatna na zmiany strukturalne, a przy tym miała skłonność do zachowania prymatu przemysłu ciężkiego. Ponadto kontynuacja dotychczasowej polityki w zakresie inwestycji prowadzi do gwałtownego, niemal katastroficznego pogorszenia zdolności gospodarki do odnowienia

* Przez symulację bazową rozumie się wyniki takiej symulacji zmiennych endogenicznych, przy oszacowanych w próbie wartościach parametrów i wartościach zmiennych objaśniających, jak w ostatnim roku próby.

majątku trwałego*. Nawet przy najbardziej optymistycznych scenariuszach rósł dług zagraniczny i fakt ten nie wpływał w zasadniczy sposób na udział konsumpcji w dochodzie narodowym. Wyniki symulacji wskazywały, że sfera spłaty zadłużenia i sfera spożycia były od siebie słabo uzależnione. Autorzy analizy symulacyjnej stwierdzili, że możliwości spłaty kredytów zależą w zasadniczym stopniu od zmian efektywności gospodarki.

Klasyczną formą prezentacji wyników prac związanych z modelowaniem ekonometrycznym i ich wykorzystania do celów planistycznych jest gotowy model wraz z ewentualnymi wynikami badań symulacyjnych dla różnych ustalonych przez autorów scenariuszy. Takie postępowanie powoduje, że charakterystyczną cechą przedstawionego odbiorcy wyniku jest jego zamkniętość oraz w zasadzie nierozszerzalność. Odbiorca może zapoznać się z wynikami i je w całości lub w pewnej części zakwestionować, bądź w pełni zaakceptować. Jeśli nawet model jest do przyjęcia dla jednego użytkownika, to często ze względu na zróżnicowane potrzeby może być niezadowalający dla innych. Tak właśnie zrodziła się koncepcja otwartych modeli ekonometrycznych. W założeniu otwarta forma modelu powinna pozwalać łatwo dostosowywać postać modelu, jego strukturę, stopień szczegółowości, próbę, na której się opiera do potrzeb użytkownika. Koncepcję tę zastosowano dla modelowania rolnictwa i sfery spożycia artykułów żywnościowych. Otwarty model składa się z przyjaznego użytkownikowi pakietu programów zawierającego procedury niezbędne przy konstrukcji modelu, procedury estymacyjne i weryfikujące jakość równań modelu, sprzężone z pakietem banki danych źródłowych.

Podsumowanie

Zmiany systemu gospodarczego, a tym samym brak porównywalnych danych statystycznych potrzebnych do konstrukcji i estymacji modelu, pojawienie się nowych obszarów badawczych, spowodowało zaniechanie dalszych prac nad modelowaniem gospodarki narodowej przez zespół Profesora Józefa Kolonki. Osiągnięte rezultaty metodologiczne tych badań są nadal aktualne i można je wykorzystać na różnych poziomach agregacji (gospodarka narodowa, gospodarka regionalna czy pojedyncze obiekty gospodarcze). Niektóre rezultaty analiz symulacyjnych, dla wybranych scenariuszy można znaleźć w pracy P. Chrzana et al. (1988). Jednym z powodów nie opublikowania rezultatów prac był charakter Profesora Józefa Kolonki, natura badacza nieustannie poszukującego prawdy.

* Zob. (Chrzan et al., 1988), scenariusze B,C,D.

Uważał bowiem, że każdy osiągnięty rezultat badań jest niezadowalający, należy nieustannie poszukiwać rozwiązań lepszych.

Proces transformacji do rozwiniętej gospodarki rynkowej jest procesem długotrwałym. W. Welfe (2000)wymienia tu kilka podokresów, w których pojawiają się specyficzne problemy mające wpływ na specyfikę modelowania:

- komercjalizacja i prywatyzacja prowadzą do zwiększenia wrażliwości podmiotów gospodarczych na zmiany relatywnych cen, co dotyczy przede wszystkim relacji cen dóbr pochodzenia krajowego i dóbr importowanych;
- stabilizacja gospodarcza pociąga za sobą wydłużenie czasu dostosowań, a także zmniejszenie ryzyka, co pobudza aktywność inwestycyjną i przyciąga zagraniczne inwestycje bezpośrednie;
- efekty wzrostu gospodarczego ujawniają się w formie nie tylko przyrostu realnych wynagrodzeń i dochodów, ale także w postaci tendencji do finansowania wydatków z kredytu konsumpcyjnego;
- podstawową rolę odgrywają instrumenty polityki pieniężnej oraz fiskalnej.

Do opisu tej fazy transformacji sugeruje się stosowanie przede wszystkim modeli kwartalnych o orientacji popytowej. W modelach tych podstawową rolę powinny odgrywać bloki równań, generujące w sferze realnej popyt krajowy i zagraniczny, popyt na produkcję krajową i import, zatrudnienie i bezrobocie, następnie ceny i wynagrodzenia przeciętne oraz przepływy finansowe.

Na początku lat 90. XX w. rozpoczęto regularną publikację prognoz gospodarczych opartych na modelach serii W. W następnych latach kolejne trzy ośrodki: Instytut Badań nad Gospodarką Rynkową, Niezależny Ośrodek Badań Ekonometrycznych oraz Centrum Danych Makroekonomicznych przedstawiały swoje prognozy rozwoju gospodarczego Polski. Były one obliczane na podstawie własnych modeli makroekonomicznych. Prognozy poszczególnych kategorii makroekonomicznych były obarczone błędami różnej wielkości.

Tabela 3

Przeciętne względne błędy prognoz dla czterech grup podmiotów prognozujących

Podmioty prognozujące	Przeciętne względne błędy prognoz		
	1995	1996	1997
Cztery ośrodki (LIFEA, IBNGR, NOBE, CDMiF)	1,16	1,05	0,81
Rząd	0,96	0,75	0,92
Źródła międzynarodowe	1,19	0,88	0,95
Inne prognozy	1,44	0,85	1,14

Nota: W celu porównywania błędów prognoz dla różnych podmiotów prognozujących oraz różnych lat skorygowano względne błędy prognoz o wielkości średnie dla danej kategorii (D" = D'/średnie D' dla danej kategorii); D' można uznać za teoretyczny procentowy błąd względny danej prognozy wykonanej w grudniu roku poprzedzającego rok prognozowany.

Źródło: (Maciejewski, 2000, s. 166).

Z danych zamieszczonych w tab. 3 wynika, że w poszczególnych latach wartość średniego błędu prognozy ulega znaczącym zmianom. W. Maciejewski sugeruje, że zmiany te mogą być wynikiem zmieniających się warunków gospodarczych, wewnętrznych i zewnętrznych, a także wynikiem ulepszania technik prognozowania przez poszczególne podmioty prognozujące. Warto zwrócić także uwagę na fakt, że bardzo krótkie okresy (lata 1995-1997) nie pozwoliły na wykorzystanie klasycznych metod porównań prognoz opartych na szeregach czasowych.

Bibliografia

Chrzan P., Fornal L., Kolonko J., Stolarska E., Zadora K. (1988): *Symulacja na podstawie modelu ekonometrycznego MGS (wyniki dla lat 1984-1990)*. W: *Makromodele gospodarki i ich zastosowania* (materiały pokonferencyjne) Red. E. Stolarska. Wydawnictwo Akademii Ekonomicznej, Katowice.

Welfe W. (2000): *Zasady makromodelowania gospodarki okresu* transformacji. W: *Gospodarka Polski w okresie transformacji*. Red. A. Welfe. Polskie Wydawnictwo Ekonomiczne, Warszawa.

Maciejewski W. (2000): *Dokładność makroekonomicznych prognoz gospodarki polskiej w latach 1995-1997*. W: *Współczesne problemy badań statystycznych i ekonometrycznych*. Red. A. Zeliaś. Wydawnictwo Akademii Ekonomicznej, Kraków, s. 158-171.

DILEMMAS OF MODELING THE NATIONAL ECONOMY

Summary

This paper presents problems of the construction of econometric models (MES-1 and MGS) in the Polish economy. These models were constructed in the 80's by a team led by Professor Kolonko. Model of the Polish economy – MES – 1 covered areas of: employment, production, national income, public money fund, consumption and foreign trade. The main object of this model was prediction. This model had a high compliance measured by determination ratio. Estimated prognosis had relatively high accuracy. To estimate the predictive accuracy of the model measures ex-ante and ex-post were used. MGS model was used to the simulation analysis. The transition to a market economy and changes in the economic system in the early 90's led to abandonment of further work on the use of constructed models.

Czesław Domański
Uniwersytet Łódzki

ROZKŁAD LAMBDA-TUKEY'A I PRÓBA JEGO ZASTOSOWANIA*

Wprowadzenie

W literaturze przedmiotu prezentowane są różnorodne rozkłady empiryczne, z których do najważniejszych należą: system krzywych K. Pearsona (1894), zawarty w pracy Pearsona (1948, por. także Domański, Pruska, 2000), system Johnsona przedstawiony w pracy Hahna i Shapiro (1967), rozkład Burra (1973) czy rozkład Tukey'a (1960).

W artykule przedstawiony będzie rozkład Lambda-Tukey'a z czterema parametrami, pozwalający na prezentację wielu różnorodnych kształtów krzywych. Zamieszczono także fragmenty tablic wartości parametrów opracowane dla tego rozkładu, które ułatwiają szacowanie jego parametrów.

Do innych ważnych zastosowań prezentowanego rozkładu należy generowanie liczb losowych dla badań symulacyjnych oraz analiz Monte Carlo sprawdzających odporność procedur statystycznych.

1. Uogólnienie rozkładu λ Tukey'a

Z reguły ciągły rozkład prawdopodobieństwa definiuje się za pomocą dystrybuanty lub funkcji gęstości. Alternatywnie można go określić przez funkcję kwantylową (percentylową). Ujmując to najprościej, funkcja kwantylowa jest funkcją odwrotną do dystrybuanty.

Badania nad uogólnionym rozkładem λ Tukey'a prowadzili m.in. Ramberg, Tadikamalla, Dudkiewicz, Mykytka (1979). Prezentowany przez tych autorów rozkład jest czteroparametrowy, uwzględniający parametry: położenia, skali, skośności i kurtozy.

* Praca napisana w ramach projektu sfinansowanego ze środków Narodowego Centrum Nauki przyznanych na podstawie decyzji numer DEC-2011/01/B/HS4/02746.

Szczególnym przypadkiem funkcji kwantylowej jest funkcja λ Tukey'a (1960):

$$R(p) = \frac{p^{\lambda} - (1-p)^{\lambda}}{\lambda}, \quad 0 \le p \le 1 \tag{1}$$

określona dla wartości $\lambda \neq 0$.

Jeżeli $\lambda \to 0$, to otrzymamy rozkład logistyczny.

Ramberg i Schmeiser (1974) przedstawili rozkład (1) z czterema parametrami danego funkcją kwantylową postaci:

$$R(p) = \lambda_1 + \frac{p^{\lambda_3} - (1-p)^{\lambda_4}}{\lambda_2}, \quad 0 \le p \le 1 \tag{2}$$

gdzie:

λ_1 – parametr położenia,

λ_2 – parametr skali,

λ_3 – parametr skośności,

λ_4 – parametr kurtozy.

Funkcja gęstości odpowiadająca (2) dana jest wzorem:

$$f(x) = f[R(p)] = \frac{\lambda_2}{\lambda_3 p^{\lambda_3 - 1} + \lambda_4 (1-p)^{\lambda_4 - 1}}, \quad 0 \le p \le 1 \tag{3}$$

Wyznaczenie funkcji gęstości dla ustalonych parametrów $\lambda_1, \lambda_2, \lambda_3 \, i \, \lambda_4$ wymaga znalezienia wartości (2) i (3) dla argumentu p z przedziału [0,1]. Następnie nanosi się wartości $f[R(p)]$ na osi Y względem wartości *R(p)* odłożonych na osi X.

Rozkład ten, którego szczególnym przypadkiem jest oryginalny rozkład λ pozwala uzyskać również skośne krzywe. Zauważmy, że dystrybuanta tego rozkładu nie występuje w postaci jawnej.

Wzory na wartość oczekiwaną, wariancję oraz współczynnik skośności i kurtozy uogólnionego rozkładu λ dane są wzorami:

$$\mu \equiv E(X) = aE(Y) + b = \lambda_1 + \frac{1}{\lambda_2}\left(\frac{1}{\lambda_3 + 1} - \frac{1}{\lambda_4 + 1}\right)$$

$$\sigma^2 \equiv E(X - E(X))^2 = a^2 E(Y - E(Y))^2 = \frac{1}{\lambda_2^2}(A_2 - A_1^2)$$

$$\beta_3 \equiv \frac{1}{\sigma^3} E(X - E(X))^3 = \frac{\lambda_2^3}{(A_2 - A_1^2)^{\frac{3}{2}}} \cdot \frac{1}{\lambda_2^3}(A_3 - 3A_1A_2 + 2A_1^3) \tag{4}$$

$$\beta_4 \equiv \frac{1}{\sigma^4}(E(X - E(X))^4) = \frac{1}{(A_2 - A_1^2)^2}(A_4 - 4A_1A_3 + 6A_1^2A_2 + 3A_1^4)$$

gdzie:

$$A_1 = \sum_{j=0}^{1}(-1)^j\binom{1}{j}\beta(\lambda_3(1-j)+1,\lambda_4 j+1) = \beta(\lambda_3+1,1)-\beta(1,\lambda_4+1) = \frac{1}{\lambda_3+1}-\frac{1}{\lambda_4+1}$$

$$A_2 = \sum_{j=0}^{2}(-1)^j\binom{2}{j}\beta(\lambda_3(2-j)+1,\lambda_4 j+1) = \beta(2\lambda_3+1,1)-2\beta(\lambda_3+1,\lambda_4+1)+\beta(1,2\lambda_4+1) = \tag{5}$$

$$= \frac{1}{2\lambda_3+1}+\frac{1}{2\lambda_4+1}-2\beta(\lambda_3+1,\lambda_4+1)$$

$$A_3 = \sum_{j=0}^{3}(-1)^j\binom{3}{j}\beta(\lambda_3(3-j)+1,\lambda_4 j+1) = \beta(3\lambda_3+1,1)-3\beta(2\lambda_3+1,\lambda_4+1)+3\beta(\lambda_3+1,2\lambda_4+1)-$$

$$-\beta(1,3\lambda_4+1) = \frac{1}{3\lambda_3+1}-\frac{1}{3\lambda_4+1}-3\beta(2\lambda_3+1,\lambda_4+1)+3\beta(\lambda_3+1,2\lambda_4+1)$$

$$A_4 = \sum_{j=0}^{4}(-1)^j\binom{4}{j}\beta(\lambda_3(4-j)+1,\lambda_4 j+1) = \beta(4\lambda_3+1,1)-4\beta(3\lambda_3+1,\lambda_4+1)+6\beta(2\lambda_3+1,2\lambda_4+1)-$$

$$-4\beta(\lambda_3+1,3\lambda_4+1)+\beta(1,4\lambda_4+1) = \frac{1}{4\lambda_3+1}+\frac{1}{4\lambda_4+1}-4\beta(3\lambda_3+1,\lambda_4+1)+6\beta(2\lambda_3+1,2\lambda_4+1)-$$

$$-4\beta(\lambda_3+1,3\lambda_4+1)$$

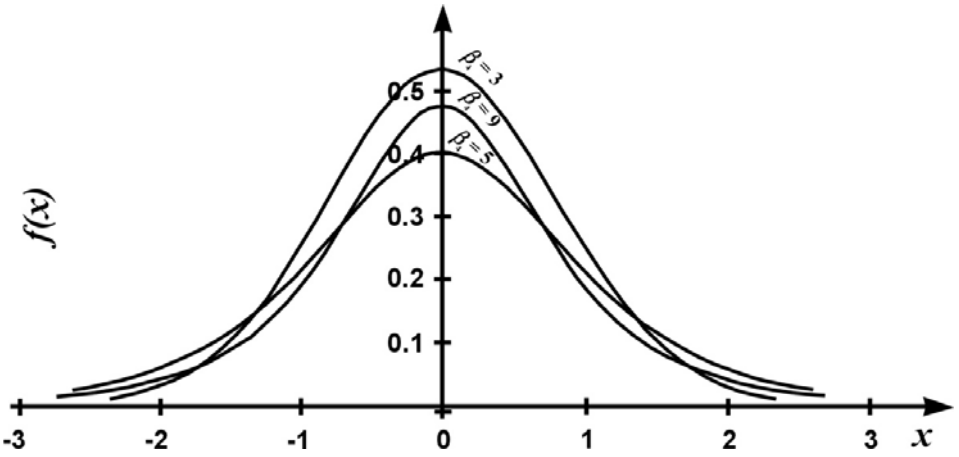
przy czym β oznacza funkcję beta.

Stąd *k*-ty moment można otrzymać, gdy $\min(\lambda_3,\lambda_4)\rangle\frac{1}{k}$.

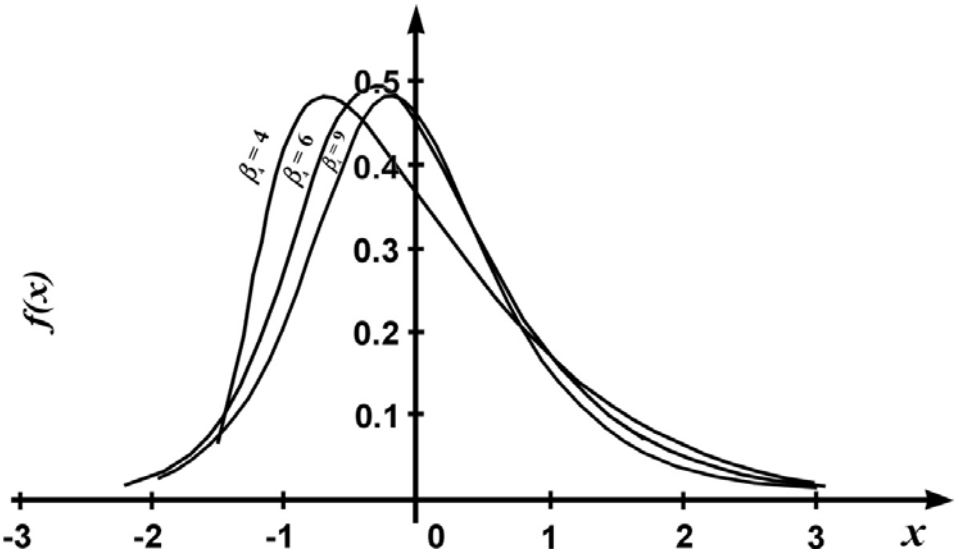
Wartość ta zależy tylko od parametrów $\lambda_3 i \lambda_4$, w konsekwencji współczynniki skośności i kurtozy również zależą tylko od tych parametrów.

Prezentowany rozkład z czteroma parametrami pozwala uzyskać wiele różnorodnych kształtów krzywych, co zostało pokazane na rys. 1-3. Na rys. 1 przedstawiona została funkcja gęstości dla parametrów β_3 = 0 oraz β_4 = 3,5,9, natomiast na rys. 2 – dla β_3 = 1 oraz β_4 = 1,6,9, a na rys. 3 dla parametrów β_3 = 0,0.5,1 oraz β_4 = 4.

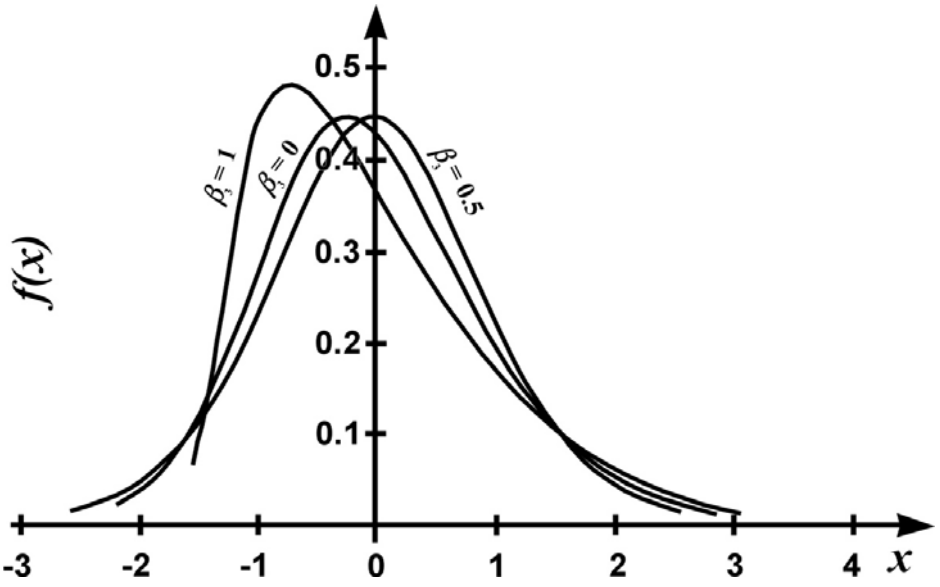
Ramberg, Dudewicz, Tadikamalla i Mykytka (1979) przedstawili tablice wartości $\lambda_1, \lambda_2, \lambda_3 i \lambda_4$ dla wybranych parametrów β_3 i β_4 oraz dla μ = 0 i σ = 1. Wielkości zamieszczone w tych tablicach zostały uwzględnione przy konstrukcji rozkładów prezentowanych na rys. 1-3.



Rys. 1. Funkcja gęstości dla $\beta_3 = 0$ oraz $\beta_4 = 3,5,9$



Rys. 2. Funkcja gęstości dla $\beta_3 = 1$ oraz $\beta_4 = 4,6,9$



Rys. 3. Funkcja gęstości dla $\beta_3 = 0,0.5,1$, $\beta_4 = 4$

Wartości A_k dla k = 1,2,3,4 (por. wzór 5) zależą tylko od parametrów $\lambda_3 \, i \, \lambda_4$, stąd współczynniki skośności i kuriozy zależą tylko od tych parametrów.

Parametry λ_i uogólnionego rozkładu λ Tukey'a obliczamy z równań:

$$\begin{aligned} &\mu = 0 \\ &\sigma^2 = 1 \\ &\beta_3 = \beta_3^* \\ &\beta_4 = \beta_4^* \end{aligned} \tag{6}$$

gdzie $\beta_3^* \, i \, \beta_4^*$ są obliczone na podstawie wyników z próby.

Uwzględniając wzory (4), otrzymujemy:

$$\begin{cases} \lambda_1 + \dfrac{1}{\lambda_2}\left(\dfrac{1}{\lambda_3 + 1} - \dfrac{1}{\lambda_4 + 1}\right) = 0 \\ \dfrac{1}{\lambda_2^2}(A_2 - A_1^2) = 1 \\ \dfrac{1}{(A_2 - A_1^2)^{\frac{3}{2}}}(A_3 - 3A_1A_2 + 2A_1^3) = \beta_3^* \\ \dfrac{1}{(A_2 - A_1^2)^2}(A_4 - 4A_1A_3 + 6A_1^2A_2 - 3A_1^4) = \beta_4^* \end{cases} \tag{7}$$

Do równań (7) podstawimy parametry λ_i (i = 1,2,3,4) z tablicy 4 artykułu Ramberga i in. (1979).

Na podstawie programu „Mathematica"* lewe strony równań (6) oznaczone są literami f, g, h i w.

Po podstawieniu parametrów λ powinno się otrzymać:

$$f = 0$$

$$g = 1$$

$$h = \beta_3$$

$$w = \beta_4$$

Obliczenia wykonane są dla 4 przypadków:

1. Podstawiamy $\lambda_1 = -1{,}245, \lambda_2 = 0{,}2445, \lambda_3 = 0{,}0178, \lambda_4 = 0{,}4748$

W wyniku otrzymujemy:

$$\begin{cases} f = 0{,}0002007596 \\ g = 0{,}999776 \\ h = 0{,}500145 \\ w = 2{,}40007 \end{cases}$$

W cytowanych tablicach dla wybranych parametrów λ $\beta_3 = 0{,}5$ $\beta_4 = 2{,}4$.

2. $\lambda_1 = -0{,}045, \lambda_2 = -0{,}1198, \lambda_3 = -0{,}0569, \lambda_4 = -0{,}0617$

Wyniki:

$$\begin{cases} f = 0{,}000277763 \\ g = 0{,}999848 \\ h = -0{,}148876 \\ w = 5{,}39963 \end{cases}$$

W tablicach $\beta_3 = 0{,}15, \beta_4 = 5{,}4$.

3. $\lambda_1 = -0{,}134, \lambda_2 = -0{,}2501, \lambda_3 = -0{,}0977, \lambda_4 = -0{,}1242$

Wyniki:

$$\begin{cases} f = 0{,}0000837921 \\ g = 0{,}99988 \\ h = -0{,}651818 \\ w = 8{,}20468 \end{cases}$$

* Obliczenia zostały wykonane przez dr Katarzynę Bolonek-Lasoń.

W tablicach $\beta_3 = 0{,}65, \beta_4 = 8{,}2.$

4. $\lambda_1 = -0{,}499, \lambda_2 = 0{,}1497, \lambda_3 = 0{,}0538, \lambda_4 = 0{,}1438$

Wyniki:

$$\begin{cases} f = -0{,}000216105 \\ g = 1{,}00005 \\ h = 0{,}550225 \\ w = 3{,}40036 \end{cases}$$

W tablicach $\beta_3 = 0{,}55, \beta_4 = 3{,}4.$

W przedstawionych przypadkach otrzymujemy wyniki zgodne z wartościami z tablicy 4 artykułu Ramberga i in. (1979).

2. Przykłady zastosowań dla indeksów giełdowych

Dane empiryczne dotyczą tygodniowych notowań indeksu DAX z okresu 03.01.1997-27.07.2012 (813 obserwacji, por. rys. 4). Na podstawie tych danych wyznaczamy parametry rozkładu:

$\mu = 5391{,}972$

$\sigma = 1338{,}23$

$$\beta_3 = \frac{\mu_3}{\sigma^3} = -0{,}05$$

$$\beta_4 = \frac{\mu_4}{\sigma^4} = 2{,}14$$

Z tab. 1 dla $\beta_3 = 0{,}05$ i $\beta_4 = 2{,}2$ odczytujemy $\lambda_1 = -0{,}802$, $\lambda_2 = 0{,}3314$, $\lambda_3 = 0{,}1128$, $\lambda_4 = 0{,}5802$. Przekształcamy wielkości parametrów λ_1 i λ_2 według wzorów (uwzględniamy wartość bezwzględną ze względu na to, że wartości λ_1 i λ_2 w tab. 1 podane są dla zmiennej o wartości oczekiwanej zero i wariancji jeden):

$$\lambda_1(\mu,\sigma) = \lambda_1(0{,}1)\sigma + \mu = -0{,}802 \cdot 1338{,}23 + 5391{,}972 = 4318{,}7$$

$$\lambda_2(\mu,\sigma) = \lambda_2(0{,}1)/\sigma = 0{,}3314/1338{,}23 = 0{,}00025$$

gdzie μ i σ to średnia i odchylenie standardowe obliczone na podstawie danych empirycznych.

Zmienna X oraz odpowiadająca jej funkcja gęstości przyjmuje postać:

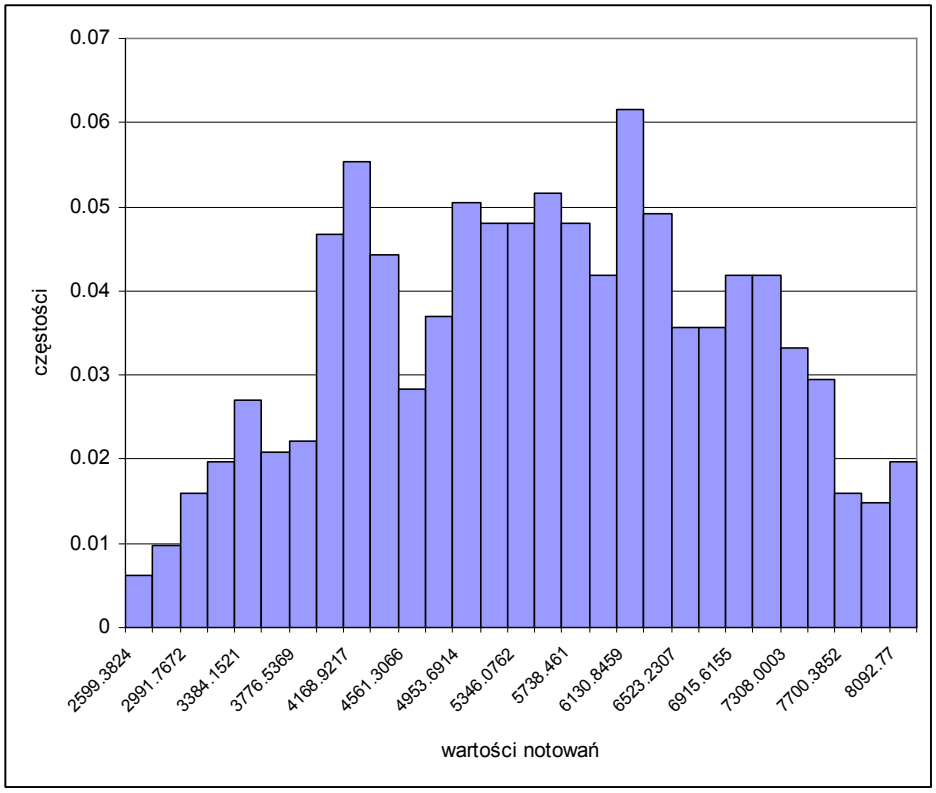
$$X = \frac{\lambda_1 + p^{\lambda_3} + (1-p)^{\lambda_4}}{\lambda_2}$$
$$f(x) = \frac{\lambda_2}{\lambda_3 p^{\lambda_3 - 1} + \lambda_4 (1-p)^{\lambda_4 - 1}} \tag{8}$$

Tabela 1

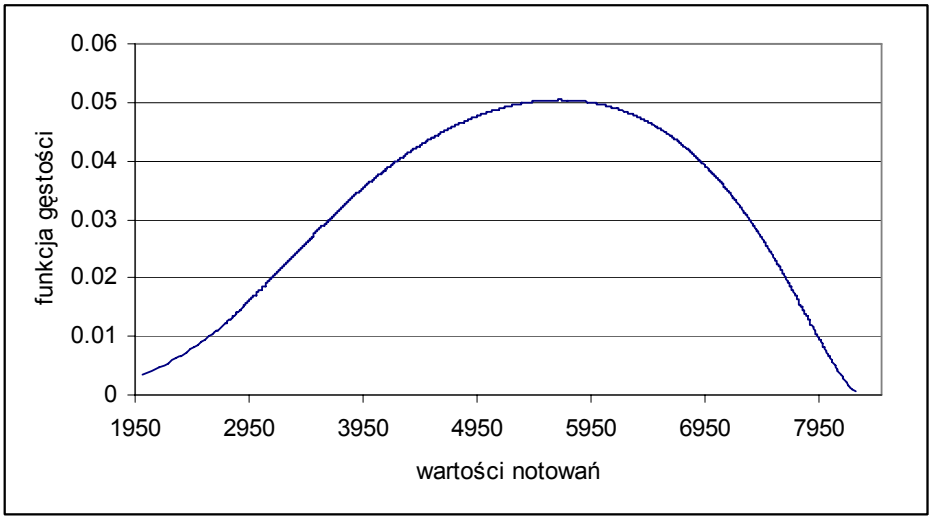
Wybrane wartości parametrów $\lambda_1, \lambda_2, \lambda_3 \, i \, \lambda_4$ dla współczynników skośności β_3 = 0.0,0.05,1 kurtozy β_4 = 1.0, …,9.0 *gdy* μ = 0 i σ = 1

$\beta_3 = 0.0$				
β_4	λ_1	λ_2	λ_3	λ_4
1.8	0.0	.5774	1.0000	1.0000
2.0	0.0	.4952	.5843	.5843
2.2	0.0	.4197	.4092	.4092
2.4	0.0	.3533	.3032	.3032
2.6	0.0	.2949	.2303	.2303
2.8	0.0	.2433	.1765	.1765
3.0	0.0	.1974	.1349	.1349
4.0	0.0	.0262	.0148	.0148
5.0	0.0	-.0676	-.0443	-.0443
6.0	0.0	-.1686	-.0802	-.0802
7.0	0.0	-.2306	-.1045	-.1045
8.0	0.0	-.2800	-.1223	-.1223
9.0	0.0	-.3203	-.1359	-.1359
$\beta_3 = 0.05$				
β_4	λ_1	λ_2	λ_3	λ_4
1.8	-1.703	.2861	.0000	.9502*
2.0	-1.229	.3122	.0505	.7603
2.2	-.802	.3314	.1128	.5802
2.4	-.375	.3328	.1876	.3941
2.6	-.143	.2924	.1973	.2605
2.8	-.083	.2429	.1625	.1903
3.0	-.059	.1975	.1276	.1425
4.0	-.026	.0264	.0146	.0153
5.0	-.016	-.0867	-.0435	-.0448
6.0	-.013	-.1682	-.0791	-.0810
7.0	-.011	-.1034	-.1034	-.1054
8.0	-.928+	-.2797	-.1212	-.1232
9.0	-.837+	-.3201	-.1348	-.1368
$\beta_3 = 1.00$				
β_4	λ_1	λ_2	λ_3	λ_4
3.4	-1.253	.1772	.0000*	.2854*
4.0	-.886	.1333	.0193	.1588
5.0	-.533	.0340	.9695+	.0285
6.0	-.379	-.0562	-.0187	-.0388
7.0	-.297	-.1291	-.0453	-.0790
8.0	-.248	-.1878	-.0670	-.1058
9.0	-.215	-.2356	-.0844	-.1249

Źródło: Na podstawie (Ramberg i in., 1979).



Rys. 4. Histogram wartości notowań indeksu DAX w latach 1997-2012



Rys. 5. Funkcja gęstości wyznaczona na podstawie równań (6) odpowiadająca notowaniom indeksu DAX

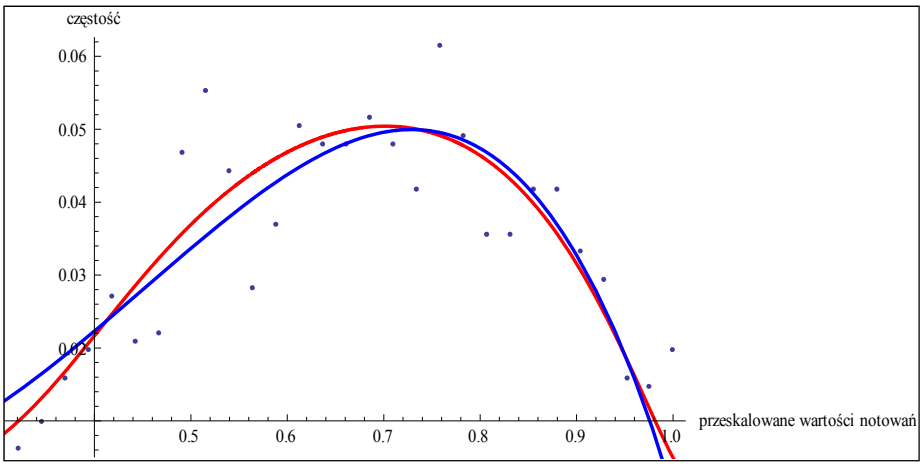
Wartości trzeciego i czwartego momentu danych empirycznych znajdują się w obszarze rozkładu beta, zatem w programie „Mathematica" dopasowujemy ten rozkład metodą najmniejszych kwadratów do danych empirycznych (rys. 6), otrzymując parametry rozkładu beta:

```
{BestFitParameters→ {a→3.64419, b→1.98305, c→2.50873},

                        |  Estimate  Asymptotic SE  CI
ParameterCITable→       a | 3.64419   0.452338      {2.7144, 4.57399}
                        b | 1.98305   0.190936      {1.59058, 2.37552}
                        c | 2.50873   0.132889      {2.23557, 2.78189}

EstimatedVariance→ 0.745372,

                                  DF     SumOfSq     MeanSq
                    Model         3      388.158     129.386
ANOVATable→         Error         26     19.3797     0.745372,
                    Uncorrected Total    29     407.538
                    Corrected Total      28     62.7104
```

$$f(x)=\frac{2{,}50873}{\beta(3{,}64419;\ 1{,}98305)}x^{2{,}64419}(1-x)^{0{,}98305}$$

Punkty na rys. 6 oznaczają dane empiryczne, czyli częstości występowania zmiennej w każdym przedziale, niebieska krzywa prezentuje funkcję gęstości rozkładu beta, natomiast czerwona krzywa przedstawia funkcję gęstości wyznaczoną na podstawie funkcji kwantylowej.

Stopień dopasowania rozkładu beta do danych empirycznych mierzony współczynnikiem determinacji wynosi około 74%, dla rozkładu wyznaczonego na podstawie funkcji kwantylowej współczynnik ten wynosi 76%.



Rys. 6. Funkcja gęstości rozkładu beta oraz rozkładu wyznaczonego na podstawie funkcji kwantylowej

Nie ma podstaw do odrzucenia hipotezy zerowej o zgodności rozkładu empirycznego z rozkładem wyznaczonym na podstawie funkcji kwantylowej (wartość statystyki testowej $\chi^2 = 30{,}53$, wartość krytyczna dla poziomu istotności α = 0,05 wynosi $\chi^2_{24} = 36{,}415$). Dla rozkładu beta test zgodności χ^2 odrzuca hipotezę zerową o zgodności tego rozkładu z rozkładem empirycznym (wartość statystyki testowej $\chi^2 = 38{,}85$).

Podsumowanie

Omawiany rozkład pozwala uzyskać szeroką gamę kształtów krzywych, które jako najprostsze przykłady pokazane są na rys. 1-3. Ze względu na wysoką elastyczność tego rozkładu znajduje on wiele różnorodnych zastosowań w przypadku, gdy rzeczywisty rozkład nie jest znany.

Wielu autorów zajmowało się badaniami własności rozważanego rozkładu (por. np. Chalabi, Scott, Wuertz, 2012). Literatura z tego zakresu jest stosunkowo bogata, co świadczy o dużych możliwościach zastosowań uogólnionego rozkładu λ. Tristano (2010) prezentuje np. uogólniony rozkład λ z pięcioma parametrami, który w dalszych badaniach autora będzie rozważany.

Bibliografia

Burr I.W. (1973): *Parameters for a General System of Distributions to Match a Grid of* α_3 *and* α_4. Comm. Statist., 2,1-21.

Chalabi Y., Scott D.J., Wuertz D. (2012): *An Asymmetry-Steepness Parameterization of the Generalized Lambda Distribution*, http://mpra.ub.uni-muenchen.de/37814.

D'Addaro R. (1949): *Ricerche sulla* curva dei redditi. „Giornale degli Economisti e Annali di Economia", 8, s. 91-114.

Domański Cz., Pruska K. (2000): *Nieklasyczne metody statystyczne*. Polskie Wydawnictwo Ekonomiczne, Warszawa.

Edgeworth F.Y. (1898): *On the Representation of Statistics by Mathematical Formule*. „Journal of the Royal Statistical Society", 1, s. 670-700.

Hahn G.J., Shapiro S.S. (1967): *Statistical Models in Engineering*. John Wiley & Sons, New York.

Johnson N.L. (1949): *Systems of Frequency Curves Generated by Methods of Translation*, „Biometrika" 44, s. 147-176.

Pearson K. (1894): *Contributions* to the *Mathematical Theory* of Evolution. Transactions of the Royal Society, 184. W: K. Pearson (1948), s. 1-40.

Pearson K. (1895): *Contributions* to the *Mathematical Theory* of Evolution. II Skew Variation in Homogeneous Material Philosophical, 186. W: K. Pearson (1948): s. 41-112.

Pearson K. (1948): *Karl Pearson's Early Statistical Papers*. Cambridge University Press.

Ramberg J.S., Schmeister B.W. (1974): *An Approximate Method for Generating Asymmetric Random Variables*. Canon. ACM, 17, s. 78-82.

Ramberg J.S., Tadikamalla P.R., Dudewicz E.J., Mykytka E.F., (1979): *A Probability Distribution and its Uses in Fitting Data*. „Technometries 21", No. 2, s. 201-214.

Tarsitano A. (2010): *Comparing Estimation Methods for the FPLD*. „Journal Probability and Statistics", Vol. 1, No. 1, s. 1-16.

Tukey J.W. (1960): *The Practical Relationship Between the Common Transformations of Percentages of Counts and of Amounts*. Technical Report 36, Statistical Techniques Research Group, Princeton University.

LAMBDA-TUKEY DISTRIBUTION AND APPLICATION ATTEMPT

Summary

In the article the generalized Lambda-Tukey distribution was presented with the following four parameters of: location, scale, skewness and kurtosis.

The distribution presented, due to its high flexibility is widely applied, especially when empirical distributions are sophisticated and do not show desired accordance with known classical theoretical distributions.

The examples presented on the fitting of the DAX index distribution to the four parameter Tukey distribution turn out to be better than the ones for the beta distribution.

Wojciech Gamrot
Uniwersytet Ekonomiczny w Katowicach

ON KERNEL SMOOTHING AND HORVITZ-THOMPSON ESTIMATION

Introduction

Design-based estimation of population parameters usually relies on the knowledge of inclusion probabilities characterizing the sampling scheme. These are needed to construct sampling weights that form the well-known Horvitz-Thompson estimator of the population total and estimates for other parameters of interest. Sometimes, the complexity of sampling scheme prevents the exact calculation of inclusion probabilities. Such a situation arises for example for spatial sampling schemes (Fattorini and Ridolfi, 1997) some order sampling schemes (Rosen, 1997; Aires, 2000) as well as in fixed-cost sequential sampling where the composition of the sample depends on individual costs of sampled units (Pathak, 1976; Kremers, 1985).

The lack of exact inclusion probabilities does not necessarily render the Horvitz-Thompson statistic useless, since the statistician still possesses the knowledge of the sampling procedure used to draw the sample. When all the information needed to carry out sampling is readily available (such as: auxiliary variable values, unit sampling costs, adjacency matrix in spatial sampling), Fattorini (2006) proposes to conduct a simulation study and to estimate unknown inclusion probabilities, by drawing large numbers of sample replications and then counting appearances of individual units. By replacing unknown inclusion probabilities with estimates an alternative statistic known as empirical Horvitz--Thompson estimator is obtained.

Estimation of inclusion probabilities by simple sample proportions (or some statistics functionally dependent on it) usually requires large numbers of sample replications to achieve desired accuracy of Horvitz-Thompson estimates. Hence it appears reasonable to employ some form of strength-borrowing to capitalize on available auxiliary information and to improve accuracy of the simulation-based Horvitz-Thompson statistic. In this paper a nonparametric strength-borrowing technique is proposed for sampling schemes where first order inclusion probabilities satisfy simple ordering constraints. The fixed-cost sequential sampling scheme of Pathak (1976) is used as an example.

1. Estimators

Let the finite population be represented as a set of unit indices U={1, ... N}. Also, let y_1 , ... , y_N represent fixed values of some characteristic of interest and let $t = \sum_{i \in U} y_i$ be the population total to be estimated. An unordered sample s is drawn from U using some sampling scheme characterized by a set of first-order inclusion probabilities π_1, ... ,π_N where $\pi_i = P(i \in s)$ for $i \in U$. If inclusion probabilities were known, a design-unbiased Horvitz-Thompson estimator for t would be easily calculated from s according to the formula:

$$\hat{t} = \sum_{i \in s} \frac{y_i}{\pi_i} \tag{1}$$

When inclusion probabilities are impossible to calculate exactly, one may use the known sampling scheme to generate M independent sample replications s_1, ... ,$s_M \subseteq U$. For $i \in U$ let

$$k_i = \#\{r \in 1, \ \ldots \ , M : i \in s_r\} \tag{2}$$

be the number of replications containing the i-th unit. A very simple estimate of π_i is the sample proportion:

$$\hat{\pi}_i = \frac{k_i}{M} \text{ for } i \in s \tag{3}$$

However, when plugged into the formula (1) in place of π_i it could lead to division by zero if $k_i = 0$ for some $i \in s$. Such an event would require the i-th unit not to be drawn at all to any replication and is extremely unlikely for large M, but formally it prevents moments of the Horvitz-Thompson statistic from being computed. Hence, Fattorini (2006) proposes to estimate the inclusion probability π_i by the statistic:

$$\hat{\pi}_{iF} = \frac{k_i + 1}{M + 1} \text{ for } i \in s \tag{4}$$

and to estimate the population total t through the estimator:

$$\hat{t}_F = \sum_{i \in s} \frac{y_i}{\hat{\pi}_{iF}} \tag{5}$$

He derives an exact formula for its bias and a tight upper bound for the mean square error. However, as noted by the same author, the number of replications needed to guarantee high accuracy of this statistic may still be very large. This justifies efforts aimed at finding an alternative method of estimating π_i. Let

us notice, that during the simulation experiment involving generation of M replications, one may calculate estimates of inclusion probabilities not only for units in the sample s, but in fact for all N population units at negligible additional cost. Hence, any known relationships between individual inclusion probabilities corresponding to units included in s and units not included in s may be utilized to improve accuracy of estimates. In particular, such relationships may take the form of multiple inequality:

$$\pi_1 \leq \pi_2 \leq \ ... \ \leq \pi_N \tag{6}$$

As a simple example one may consider the well-known Pareto sampling scheme of Rosén (1997). By arranging population units in non-decreasing order with respect to known auxiliary variable on which the Pareto sampling is based one may easily guarantee that first-order inclusion probabilities characterizing this scheme satisfy the multiple inequality above. Gamrot (2012) proposed to incorporate the ordering constraint into empirical Horvitz-Thompson framework by calculating restricted estimates of inclusion probabilities satisfying (6) using isotonic regression algorithms such as Pool-Adjacent-Violators Algorithm (PAVA) or active set methods (see: Ayer et al., 1955; Robertson et al., 1988; Best and Chakravarti, 1990) and then by replacing unknown probabilities in (1) with these restricted estimates. However, isotonic regression only corrects for the breaches of ordering constraint (6) but it produces estimates equivalent to respective sample proportion when ordering is not violated. Hence properties of PAVA-based estimates should differ only slightly from sample proportions for larger replication numbers where such violations are rare. We will now propose another method that may be less prone to this unwelcome effect.

Let us start by noting that by definition we have $\pi_i \in [0,1]$ for $i \in U$. When N is large the ordering constraint (6) implies that either for all pairs (π_i, π_{i+1}) the difference $\pi_{i+1}-\pi_i$ is relatively small, or at least that the number of pairs where this difference is relatively large is itself not large. This leads to the intuition that for large N a particular inclusion probability π_i corresponding to the i-th population unit is unlikely to differ much from inclusion probabilities for its closest neighbors. Hence, combining probability estimates for inclusion probabilities of neighboring units may lead to better precision than using simple sample proportion.

A kernel estimator originally proposed by Rosenblatt (1956) appears to be a convenient way of forming a combined estimate of any individual inclusion probability in the population. For our purposes it is constructed as a weighted mean of simple proportions using the formula (see: Kulczycki, 2005; Härdle, 1992):

$$\hat{\pi}_{iK} = \frac{\sum_{j=1...N} \hat{\pi}_j w_{ij}}{\sum_{j=1...N} w_{ij}} \tag{7}$$

with

$$w_{ij} = \frac{1}{h} K\left(\frac{x_i - x_j}{h} \right) \tag{8}$$

where K(.) represents a certain non-negative symmetric real function having weak global maximum at 0 (so that K(x)=K(-x) and K(0)≥K(x) for x∈R) which is usually called a *kernel function* while h is a positive real constant known as *smoothing factor* or *bandwidth*. The symbol x_i represents for i∈U the value of some auxiliary characteristic of the i-th population unit. It is natural to intuitively assume it to be the unit index so that $x_i = i$ for i∈U. Another more interesting possibility of choosing x_i is discussed in the next section. Ultimately, the non-parametric empirical Horvitz-Thompson estimator of the population total is calculated according to the formula:

$$\hat{t}_K = \sum_{i \in s} \frac{y_i}{\hat{\pi}_{iK}} \tag{9}$$

Kulczycki (2005) argues, that the choice of a particular kernel function influences the accuracy of the kernel estimator (7) much less than the choice of bandwidth. In applications associated with sample surveys the normal kernel given by the formula:

$$K(x) = \exp\left(-\frac{x^2}{2} \right) \tag{10}$$

seems to be particularly popular (Giommi, 1987). From our perspective it is important that (10) always takes strictly positive values. As a result, all the terms w_{ij} in the linear combination (7) are strictly positive. Meanwhile, if the sampling scheme never produces empty samples (which may be safely assumed to be true), then at least one population unit belongs to some replication and consequently at least one of simple proportions $\hat{\pi}_1, \ldots, \hat{\pi}_N$ is strictly positive. This means that all kernel estimators $\hat{\pi}_{1K}, \ldots, \hat{\pi}_{NK}$ always take strictly positive (although possibly very small) values. Such an effect guarantees the finiteness of the Horvitz-Thompson statistic itself, and hence may be considered an advantage. In the following discussion it will be assumed that the normal formula (10) is used as a kernel.

As a general side note, it should also be stated that the proposed nonparametric estimator does not guarantee the constraint (6) to be satisfied. Although the likelihood of violating this restriction by individual estimates is apparently lower than for simple proportions computed through (3), such violations may still happen relatively often. Having said that one should keep in mind that the constraint (6) was discussed here only in order to motivate and justify the use of kernel smoothing, and was not meant to be strictly imposed.

In the following sections the proposed estimator (9) will be compared to other alternatives for a specific sampling design.

2. Application to fixed-cost sampling

Let us consider the fixed-cost sequential sampling scheme of Pathak (1976). It is characterized by varying inclusion probabilities which are generally difficult to calculate for larger sample sizes due to the combinatorial explosion (Schuster, 2000). Despite the existence of some sufficiency-based design-unbiased estimators which do not utilize inclusion probabilities, the empirical Horvitz-Thompson estimators may be of interest when nonresponse corrections need to be incorporated or when some modifications are made to the original scheme. In this paper the Pathak's scheme in its original form illustrates the use of nonparametric empirical Horvitz-Thompson approach. The sampling procedure is carried out as follows. Let $c_1, \ldots, c_N$ denote known per-unit costs of observing the characteristic under study for individual population units. Population units are drawn to the sample one-by-one without replacement and with equal probabilities until the total cumulative cost of the sample is greater or equal to some budget constraint C fixed in advance. The element for which this happens is not appended to the sample. The sample size is random in general, but instead the variability of random sample cost is largely limited.

Meanwhile, it may be shown that inclusion probabilities of the first order – although hard to compute – constitute a non-increasing function of the per-unit cost, so that:

$$\underset{i,j\in U}{\forall} c_i < c_j \Rightarrow \pi_i \geq \pi_j \tag{11}$$

and

$$\underset{i,j\in U}{\forall} c_i = c_j \Rightarrow \pi_i = \pi_j \tag{12}$$

Consequently, by arranging population units in a non-increasing order with respect to individual unit cost one may easily guarantee that inclusion probabilities satisfy the ordering constraint (6). This suggests that for most population units their inclusion probabilities should not differ dramatically from those having similar cost. This in turn justifies the use of nonparametric empirical Horvitz-Thompson estimator (9) for the population total, with costs $c_1, \ldots, c_N$ treated as auxiliary variables $x_1, \ldots, x_N$ in (8).

3. A simulation study

A simulation study was carried out in order to compare performance of the proposed non-parametric empirical Horvitz-Thompson estimator (9), the PAVA-based estimator proposed by Gamrot (2012) and the classic Fattorini's (2006) statistic (5). In experiments, the finite population was represented by the data set describing 695 farms in the Gręboszów municipality of the Dąbrowa Tarnowska district obtained during the agricultural census conducted by Polish Central Statistical Office in 1996. It was assumed that the cost of sampling individual units is strictly proportional to the farm area, which featured high positive skew and that the budget constraint C is equal to five percent of the total cost of exhaustively enumerating the whole population.

The simulation experiment accounted for two sources of randomness, namely the randomness of the actual sample s, and the randomness of inclusion probability estimates. It was carried out by drawing 20000 samples and executing an independent simulation study involving 300 sample replications for each such sample to arrive at population total estimates. Figure 1 shows the observed relative bias (RBIAS) of kernel-based estimates for h = 0.2, 0.4, ... , 30. Figure 2 shows the observed relative root mean square error (RRMSE) of kernel-based population total estimates for h = 0.2, 0.4, ... , 30. The corresponding levels of RRMSE's for PAVA-based Horvitz-Thompson estimator and for Fattorini's statistic are also shown in the Figure 2.

The relative bias of the proposed estimator exhibits rather complex behavior. For very small h it takes values very close to zero, but quite unstably fluctuating between positive and negative values. With growing h at first it also quickly grows, reaching 0.00537 for h = 4.2 but then it steadily decreases to reach 0.00010 for h=17.6 to finally slowly increase again for h>17.6. The biases of PAVA-based estimator and Fattorini's statistic do not depend on h and they are respectively equal to 0.00801 and –0.06470 with the absolute value of the latter obviously the greatest of all for any h. Hence one may conclude that for any h = 0.2, 0.4, ... , 30 the proposed estimator clearly dominated the other two by a wide margin in terms of bias.

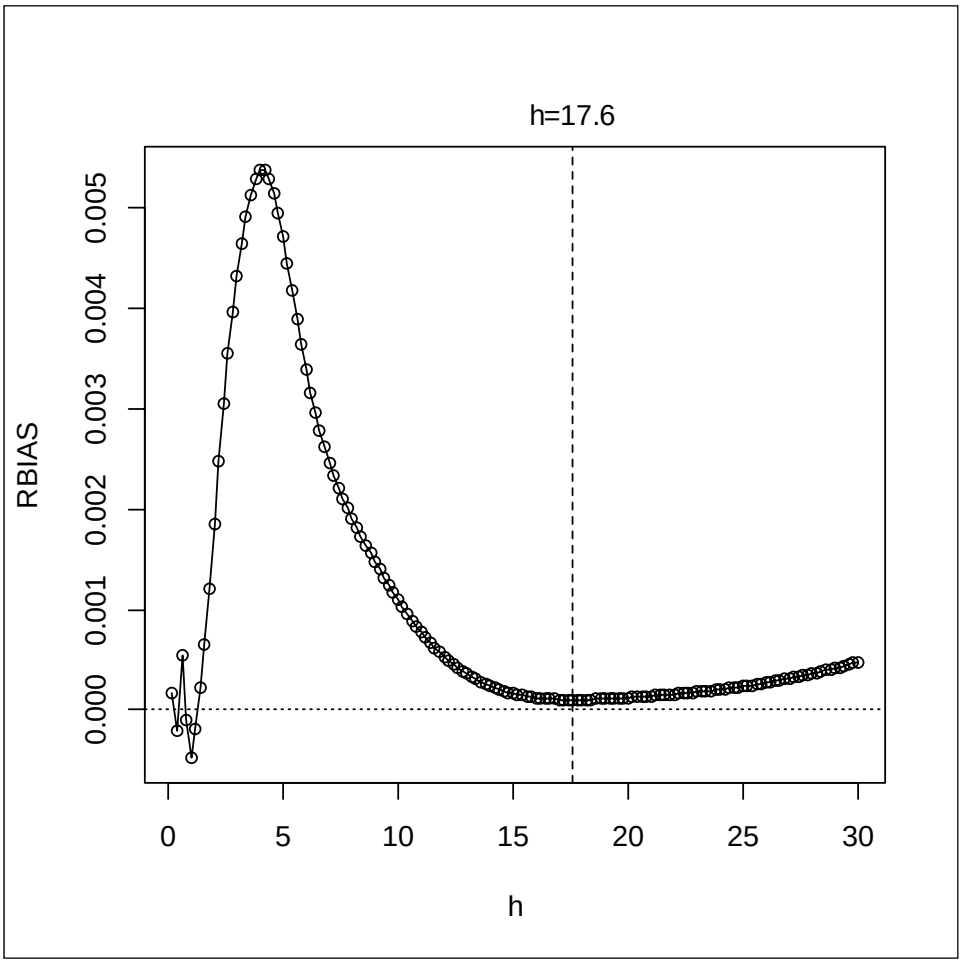


Figure 1. Relative bias of the proposed estimator for h = 0.2, 0.4, ... , 30.

The relative root mean square error of the proposed estimator also exhibited rather complicated behavior, reflecting to some extent the tendencies in the bias. It took the maximum value of 0.13877 for h = 0.1, but also featured two local minima around h = 1.2 and h = 15.8. For h = 15.8 it was equal to 0.12896 which is respectively about 12% and 3% lower than RRMSE's of PAVA-based estimator and Fattorini's statistic.

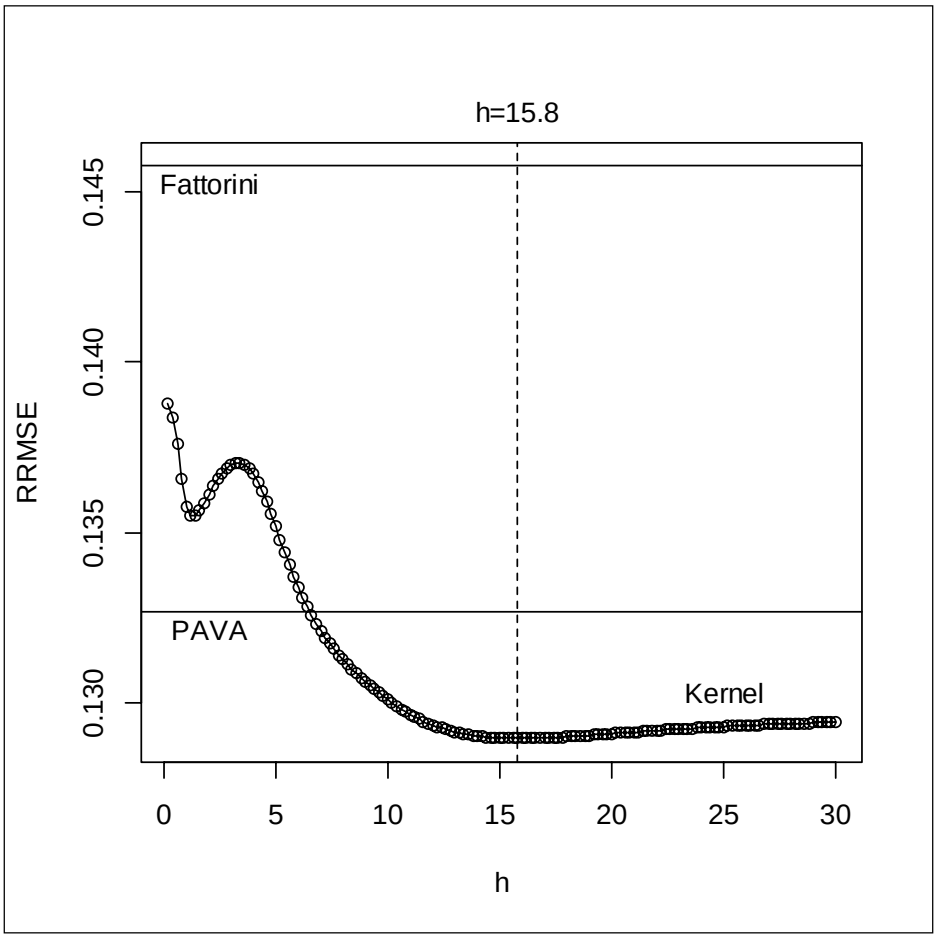


Figure 2. Relative root mean square errors of three population total estimators for h = 0.2, 0.4, ... , 30

Conclusion

Presented simulation results suggest that proposed nonparametric empirical Horvitz-Thompson estimator of the population total constitutes an attractive alternative to its two counterparts, especially in terms of bias reduction. The main challenge for it to gain a wider popularity most likely lies in choosing an optimal value for the smoothing factor h. In our study it could easily be chosen through simulation on the basis of known values of the characteristic under study in the whole population. In practice of the field work the statistician does not possess such information and would have to resort to using cross-validation or the plug-in method of Sheather and Jones (1991). Nevertheless the wide range of h-values for which the proposed estimator dominates its counterparts in terms of bias and mean square error seems to justify such approach.

References

Aires N. (2000): *Techniques to Calculate Exact Inclusion Probabilities for Conditional Poisson Sampling and Pareto πps Sampling Designs*, Phd thesis, Chalmers, Göteborg University, Göteborg.

Ayer M., Brunk H.D., Ewing G.M., Reid W.T., Silverman E. (1955): *An Empirical Distribution Function for Sampling with Incomplete Information*. *The Annals of Mathematical Statistics* 6(4), s. 641-647.

Best M.J., Chakravarti N. (1990): *Active Set Algorithms for Isotonic Regression*. A Unifying Framework. *Mathematical Programming* 47, s. 425-439.

Fattorini L., Ridolfi G. (1997): *A Sampling Design for Areal Units Based on Spatial Variability*. *Metron* 55, s. 59-72.

Fattorini L. (2006): *Appling the Horvitz-Thompson Criterion in Complex Designs: A Computer-Intensive Perspective for Estimating Inclusion Probabilities*. „Biometrica", 93(2), s. 269-278.

Gamrot W. (2012) Simulation-Assisted Horvitz-Thompson Statistic and Isotonic Regression. *Proceedings of the 30th International Conference on Mathematical Methods in Economics 2012* (accepted).

Giommi A. (1987): *Nonparametric Methods for Estimating Individual Response Probabilities*. „Survey Methodology", Vol. 13, No. 2, s. 127-134.

Härdle W. (1992): *Applied Nonparametric Regression*. Cambridge University Press.

Kulczycki P. (2005): *Estymatory jądrowe w analizie systemowej*. WNT, Warszawa.

Kremers W.K. (1985): *The Statistical Analysis of Sum-Quota Sampling*. Unpublished PHD thesis. Cornell University.

Pathak K. (1976): *Unbiased Estimation in Fixed-Cost Sequential Sampling Schemes*. „Annals of Statistics", 4 (5), s. 1012-1017.

Robertson T., Wright F.T., Dykstra R.L. (1988): *Order Restricted Statistical Inference*. Wiley, New York.

Rosén B. (1997): *On Sampling with Probability Proportional to Size*. „Journal of Statistical Planning and Inference", 62, s. 159-191.

Rosenblatt M. (1956): *Remarks on Some Nonparametric Estimates for the Density Function*. „Annals of Mathematical Statistics", No. 27, s. 832-837.

Schuster P. (2000): *Taming Combinatorial Explosion*. Proceedings of the National Academy of Sciences of the United States of America, 97 (14), s. 7678-7680.

Sheather S.J., Jones M.C. (1991): *A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation*. „Journal of the Royal Statistical Society", B, 53(3), s. 683-690.

ON KERNEL SMOOTHING AND HORVITZ-THOMPSON ESTIMATION

Summary

Estimation of the total value of fixed characteristic of interest in a finite population is considered for a complex sampling scheme featuring unknown inclusion probabilities. The general empirical Horvitz-Thompson statistic is adopted as an estimator for the unknown total. In the presence of additional knowledge on inclusion probabilities taking form of inequality constraints it is proposed to use the well-known kernel estimator for individual inclusion probabilities. For a fixed-cost sequential sampling scheme this leads to a new nonparametric empirical Horvitz-Thompson estimator of a total. Its properties are compared to known alternatives in a simulation study.

Eugeniusz Gatnar
Uniwersytet Ekonomiczny w Katowicach

ANALIZA DYSKRYMINACYJNA – STAN AKTUALNY I KIERUNKI ROZWOJU

Wprowadzenie

W literaturze statystycznej pojęcie analizy dyskryminacyjnej jako zadania polegającego na znalezieniu charakterystyki klas pojawiło się już w latach 30. XX w., gdy Fisher (1936) próbował w tym celu zastosować liniowe modele regresji. Prowadził on eksperymenty na najbardziej znanym w wielowymiarowej analizie statystycznej zbiorze danych, zawierającym charakterystykę trzech gatunków kosaćca (IRYS) za pomocą czterech zmiennych.

W Polsce jednym z pierwszych badaczy zajmującym się metodami analizy dyskryminacyjnej był prof. Józef Kolonko, który w swojej pracy (Kolonko, 1980) stosował pojęcie analizy dyskryminacyjnej w nieco innym, znacznie szerszym znaczeniu. W przedstawionym przez niego ujęciu, związanym z podejściem cybernetycznym obecnym w latach 70. XX w. w naukowej literaturze rosyjskiej, dyskryminacja zbioru obserwacji to jego podział („rozbicie"), który optymalizuje wartość pewnego funkcjonału stanowiącego kryterium. Analiza dyskryminacyjna dzieli się wobec tego na klasyfikację z wzorcem oraz klasyfikację bez wzorca (taksonomię).

Obecnie w polskiej literaturze statystycznej panuje zgodny pogląd, że analizą dyskryminacyjną jest zbiór metod prowadzących do znalezienia reguły klasyfikacyjnej, charakterystyki klas lub funkcji rozdzielających klasy, na podstawie zbioru uczącego, tj. zawierającego obiekty o znanej przynależności do klas. Jest to więc jedynie klasyfikacja z nauczycielem.

W niniejszym artykule, który ma charakter przeglądowy, zagadnienie dyskryminacji zostało zdefiniowane jako problem znalezienia charakterystyki klas, poprzez identyfikację sposobów ich odseparowania. W pierwszym punkcie zostanie sformułowane ogólne zadanie dyskryminacji, natomiast w drugim – klasyczne podejście zaproponowane przez Fishera, które prowadzi do powstania liniowych funkcji dyskryminacyjnych. Część trzecia zawiera omówienie założeń, których spełnienie pozwala na zastosowanie modeli liniowych. W przeciwnym przypadku można stosować kwadratowe funkcje dyskryminacyjne. W czwartym punkcie artykułu

przedstawiono najbardziej popularną nieparametryczną metodę dyskryminacji wykorzystującą drzewa klasyfikacyjne. Na końcu pracy znajduje się omówienie aktualnych kierunków rozwoju metod dyskryminacji.

1. Zadanie dyskryminacji

Dyskryminacja to decyzja o przydzieleniu obiektu do klasy, która jest dokonywana na podstawie znajomości rozkładów zmiennych w klasach oraz prawdopodobieństw *a priori*. Jeżeli zbiór $\{C_1, \ldots, C_J\}$ zawiera wartości zmiennej objaśnianej Y (nazwy klas), to prawdopodobieństwo *a priori* dla klasy C_j ($j = 1, \ldots, J$) to $p(C_j)$.

Ważne jest także założenie o losowości wektora zmiennych $\mathbf{X} = (X_1, \ldots, X_L)$ i o tym, że jego warunkowe rozkłady prawdopodobieństw w klasach $g(\mathbf{X} \mid C_j)$ są znane. Jeżeli oba założenia są spełnione, to można wyznaczyć prawdopodobieństwo, że klasyfikowana obserwacja $[\mathbf{x}_i, y_i]$ należy do klasy C_j, na podstawie wzoru Bayesa:

$$p(C_j \mid \mathbf{x}_i) = \frac{g(\mathbf{x}_i \mid C_j)\, p(C_j)}{\sum_{k=1}^{J} g(\mathbf{x}_i \mid C_k)\, p(C_k)}. \tag{1}$$

Prawdopodobieństwo $p(C_j \mid \mathbf{x}_i)$ jest nazywane prawdopodobieństwem *a posteriori*, gdyż decyzja o przydzieleniu obserwacji $[\mathbf{x}_i, y_i]$ do klasy C_j podejmowana jest już po jej zaobserwowaniu. Mianownik (1) nie wpływa na wartość $p(C_j \mid \mathbf{x}_i)$, zatem zależy ona jedynie od rozkładu warunkowego $g(\mathbf{x}_i \mid C_j)$ ważonego prawdopodobieństwami *a priori*.

Model dyskryminacyjny (1) może błędnie zaklasyfikować obserwację $[\mathbf{x}_i, y_i]$, z czym związana jest pewna strata, którą wyraża przyjęta arbitralnie funkcja straty (*loss function*) $L(Y, f(\mathbf{X}))$, najczęściej funkcja zero-jedynkowa:

$$L\big(Y, f(\mathbf{X})\big) = \begin{cases} 1 & gdy \quad Y \neq f(\mathbf{X}) \\ 0 & gdy \quad Y = f(\mathbf{X}) \end{cases} \tag{2}$$

lub entropia krzyżowa (*cross entropy*):

$$L\big(Y, f(\mathbf{X})\big) = -2\sum_{j=1}^{J} (Y = C_j) \log p(C_j \mid \mathbf{X}). \tag{3}$$

Błąd klasyfikacji to spodziewana wielkość funkcji straty $e(f) = E\left[L(Y, f(\mathbf{X})\right]$.

Jeżeli obserwacja $[\mathbf{x}_i, y_i]$ została błędnie zaklasyfikowana, to wartość błędu związanego z taką decyzją wynosi:

$$e(\mathbf{x}_i) = \sum_{j=1}^{J} L(y_i, f(\mathbf{x}_i)) p(C_j|\mathbf{x}_i). \tag{4}$$

Wyrażenie (4) dla funkcji straty (2) osiąga minimum, gdy:

$$\hat{f}(\mathbf{x}_i) = \arg\min_{j=1,\ldots,J} \left\{1 - p(C_j|\mathbf{x}_i)\right\}. \tag{5}$$

Oznacza to, innymi słowy, że błąd predykcji (4) jest najmniejszy, gdy obserwacja $[\mathbf{x}_i, y_i]$ zostaje przydzielona do klasy, dla której prawdopodobieństwo *a posteriori* jest największe:

$$\hat{f}(\mathbf{x}_i) = C_j \;\; gdy \;\; p(C_j|\mathbf{x}_i) = \max_{k=1,\ldots,J} \left\{p(C_k|\mathbf{x}_i)\right\}. \tag{6}$$

Reguła (6) nazywana jest bayesowską regułą klasyfikacji (*Bayes rule*), a model (5) to optymalny klasyfikator bayesowski (*Bayes classifier*). Z kolei jego błąd klasyfikacji $e^{Bayes} = 1 - \max_j \left\{p(C_j|\mathbf{x}_i) \cdot p(C_j)\right\}$ nazywany jest błędem bayesowskim (*Bayes error*).

Bayesowska reguła klasyfikacji jest prosta, lecz wymaga znajomości rozkładów warunkowych w klasach $g(\mathbf{X}|C_j)$ oraz prawdopodobieństw *a priori*. W praktyce rozkłady te są estymowane na podstawie próby uczącej. Na przykład, najczęściej stosuje się frakcje obiektów należących do klasy C_j, tj. $p(C_j) = n_j / N$, gdzie $\sum_{j=1}^{J} p(C_j) = 1$. Można też przyjąć, że są równe, tj. $p(C_1) = \ldots = p(C_J)$, oszacować na podstawie zbioru testowego lub ustalić subiektywnie.

Zgodnie z regułą Bayesa, aby poprawnie sklasyfikować obserwację $[\mathbf{x}_i, y_i]$, należy znaleźć maksimum wyrażenia $g(\mathbf{x}_i|C_j)p(C_j)$. Wyniki klasyfikacji jednak nie zmienią się, jeśli na funkcję dyskryminacyjną nałożymy odwzorowanie monotonicznie rosnące, otrzymując np. funkcje: $f_j(\mathbf{X}) = g(\mathbf{X}|C_j)$, $f_j(\mathbf{X}) = g(\mathbf{X}|C_j)p(C_j)$ lub $f_j(\mathbf{X}) = \ln(g(\mathbf{X}|C_j)) + \ln(p(C_j))$. W tym ostatnim przypadku szukamy klasy C_j, dla której funkcja:

$$f_j(\mathbf{X}) = -\frac{1}{2}(\mathbf{X} - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu}_j) + \ln(p(C_j)) \tag{7}$$

osiąga maksimum, gdzie $\boldsymbol{\mu}_j$ jest środkiem ciężkości klasy C_j, zaś $\boldsymbol{\Sigma}$ macierzą wariancji i kowariancji. Warto zauważyć, że funkcja dyskryminacyjna (7) jest funkcją liniową.

W podobny sposób można skonstruować funkcje dyskryminacyjne, które separują poszczególne klasy. Biorąc pod uwagę np. klasy C_j oraz C_k, można zbudować funkcję:

$$f_{jk}(\mathbf{X}) = \ln\left(\frac{p(C_j \mid \mathbf{X})}{p(C_k \mid \mathbf{X})}\right) + \ln\left(\frac{p(C_j)}{p(C_k)}\right), \tag{8}$$

która wyznacza równanie hiperpłaszczyzny:

$$\ln\left(\frac{p(C_j)}{p(C_k)}\right) - \frac{1}{2}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_j + \boldsymbol{\mu}_k) + (\boldsymbol{\mu}_j - \boldsymbol{\mu}_k)\boldsymbol{\Sigma}^{-1}\mathbf{X} = 0\,. \tag{9}$$

Widać tutaj związek z liniowymi funkcjami dyskryminacyjnymi, zaproponowanymi przez Fishera, które zostaną omówione w następnej części artykułu.

W literaturze mowa jest także o naiwnym klasyfikatorze bayesowskim (*naive Bayes*), który opiera się na założeniu, że w klasie C_j zmienne $X_1,...,X_L$ są niezależne:

$$f_j(\mathbf{X}) = \prod_{l=1}^{L} f_{jl}(X_l)\,, \tag{10}$$

gdzie funkcje gęstości f_{lj} są przybliżane za pomocą rozkładu normalnego. Uzyskane za jego pomocą wyniki klasyfikacji są często bardziej dokładne niż w przypadku innych modeli, ponieważ obciążenie estymatorów funkcji gęstości w klasach nie przekłada się na obciążenie prawdopodobieństw *a posteriori*.

2. Liniowe modele dyskryminacyjne Fishera

Jak już wspomniano, zagadnienie dyskryminacji zostało po raz pierwszy sformułowane przez Fishera (1936), a zaproponowana przez niego metoda opiera się na redukcji wymiaru przestrzeni cech. Pierwotna przestrzeń $\mathbf{X}^L$ zostaje zredukowana do przestrzeni zmiennych kanonicznych (*canonical variables*) $\mathbf{Z}^{J-1} = Z_1 \times ... \times Z_{J-1}$.

Należy podkreślić, że wywodzące się od Fishera pojęcie liniowej analizy dyskryminacyjnej (Linear Discriminant Analysis – LDA) używane jest często w węższym znaczeniu i określa jedynie jego podejście. Fisher rozwiązał pro-

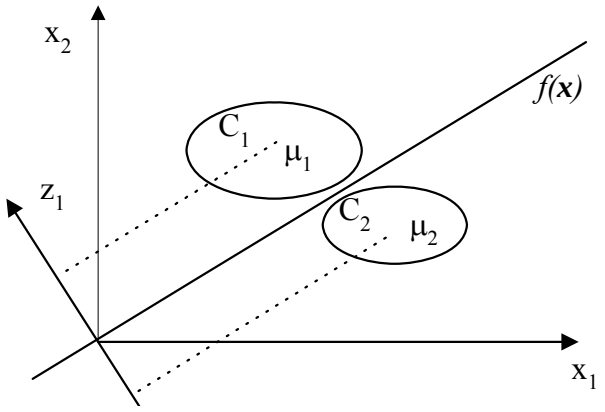
blem dla dwóch klas ($J = 2$), zaś uogólnienie dla $J > 2$ klas podali Rao (1948) i Bryan (1951).

Zmiennymi kanonicznymi są:

$$Z_j = \sum_{l=1}^{L} a_{jl} X_l , \tag{11}$$

które powodują, że w tej nowej przestrzeni klasy są optymalnie odseparowane, tj. ich środki ciężkości leżą jak najdalej od siebie.

Na rys. 1 pokazano dwie klasy C_1, C_2 w dwuwymiarowej przestrzeni cech X_1, X_2 oraz zmienną kanoniczną Z_1. Prosta $f(\mathbf{X})$, która je oddziela, jest ortogonalna względem Z_1.



Rys. 1. Separacja klas w przestrzeni dwóch zmiennych

Źródło: Gatnar (2008).

Poszukiwane jest więc maksimum funkcji:

$$Q = \frac{\left|\mathbf{a}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T\right|^2}{\mathbf{a}\mathbf{W}\mathbf{a}^T}, \tag{12}$$

gdzie licznik jest kwadratem odległości środków ciężkości klas wzdłuż kierunku $\boldsymbol{a}$, zaś $\mathbf{W}$ jest macierzą wariancji i kowariancji wewnątrzklasowej:

$$\mathbf{W} = \frac{1}{N-J} \sum_{j=1}^{J} \sum_{i=1}^{N_j} (\mathbf{x}_i - \boldsymbol{\mu}_j)^2 . \tag{13}$$

Maksimum (12) jest osiągane dla wektora $\hat{\mathbf{a}}$ o kierunku $\mathbf{W}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$. Hiperpłaszczyzna separująca obie klasy jest prostopadła do kierunku $\hat{\mathbf{a}}$ i przechodzi przez środek odcinka łączącego środki ciężkości klas. Ma ona postać:

$$(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)\mathbf{W}^{-1}\left(\mathbf{X} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\right) = 0\,. \tag{14}$$

Można pokazać, że licznik funkcji (12) jest równy $\mathbf{aBa}^T$, a zatem kryterium to sprowadza się do znalezienia takiego wektora $\hat{\mathbf{a}}$, aby iloraz wariancji:

$$F = \frac{\mathbf{aBa}^T}{\mathbf{aWa}^T} \tag{15}$$

był jak największy, gdzie $\mathbf{B}$ jest macierzą wariancji i kowariancji międzyklasowej, wyznaczoną na podstawie próby:

$$\mathbf{B} = \frac{1}{J-1}\sum_{j=1}^{J} N_j(\boldsymbol{\mu}_j - \boldsymbol{\mu})(\boldsymbol{\mu}_j - \boldsymbol{\mu})^T\,. \tag{16}$$

Wektor $\hat{\mathbf{a}}$ maksymalizujący kryterium (14) jest wektorem własnym macierzy $\mathbf{W}^{-1}\mathbf{B}$ odpowiadającym największej wartości własnej tej macierzy.

Zmienne kanoniczne są konstruowane w sposób sekwencyjny. Wektor $\hat{\mathbf{a}}$ wyznacza kierunek, wzdłuż którego klasy są najlepiej odseparowane, a kombinacja liniowa $\mathbf{Z}_1 = \hat{\mathbf{a}}_1\mathbf{X}^T$ jest pierwszą zmienną kanoniczną. Jeśli jakość klasyfikacji nie jest zadowalająca, można wyznaczyć kolejne wektory $\hat{\mathbf{a}}_2, \ldots, \hat{\mathbf{a}}_{J-1}$, które spełniają warunek wzajemnej ortogonalności i są wektorami własnymi macierzy $\mathbf{W}^{-1}\mathbf{B}$, odpowiadającymi uporządkowanym malejąco wartościom własnym.

Drugie podejście związane jest z metodą preceptronową Rosenblatta (1958), która ma charakter iteracyjny. Metoda ta minimalizuje wartość kryterium:

$$Q(\mathbf{a}) = -\sum_{\mathbf{x}_i \in E} \mathbf{a}\mathbf{x}_i^T\,, \tag{17}$$

gdzie $E = \{\mathbf{x}_i : \mathbf{a}\mathbf{x}_i^T < 0\}$ jest zbiorem obserwacji błędnie sklasyfikowanych. Ponieważ funkcja (17) jest ciągła, do jej minimalizacji można wykorzystać klasyczną gradientową metodę spadku (*gradient descent*). Gradient funkcji (17) wynosi:

$$\frac{\partial Q(\mathbf{a})}{\partial \mathbf{a}} = -\sum_{\mathbf{x}_i \in E} \mathbf{x}_i\,. \tag{18}$$

W pierwszym kroku wybierane są losowo początkowe wartości wektora parametrów $\mathbf{a}^{(0)}$, a w następnych krokach następuje jego modyfikacja:

$$\mathbf{a}^{(r+1)} = \mathbf{a}^{(r)} + \lambda_r \sum_{\mathbf{x}_i \in E} \mathbf{x}_i \tag{19}$$

w kierunku przeciwnym do gradientu. Parametr λ_r określa długość kroku i nazywany jest także współczynnikiem uczenia. Inna często stosowana odmiana reguły (19) wykorzystuje pojedynczą, błędnie sklasyfikowaną obserwację:

$$\mathbf{a}^{(r+1)} = \mathbf{a}^{(r)} + \lambda_r \mathbf{x}_i . \tag{20}$$

Opisane sposoby modyfikacji nie zmieniają wartości wektora wag dla obserwacji poprawnie sklasyfikowanych.

Metoda perceptronowa jest zbieżna w przypadku liniowej separowalności klas. Naturalnym kryterium stopu jest więc wartość $Q(\mathbf{a}) = 0$. Wynik poszukiwania zależy od wyboru wartości początkowej wektora wag, dlatego procedurę optymalizacji można przeprowadzić kilka razy, dla różnych losowo wybranych wektorów wag początkowych $\mathbf{a}^{(0)}$, a następnie wybrać rozwiązanie najlepsze.

3. Kwadratowe funkcje dyskryminacyjne

Jak już zaznaczono, podejście regresyjne w analizie dyskryminacyjnej polega na tym, że w każdej klasie C_j budowany jest osobny model:

$$f_j(\mathbf{X}) = \alpha_{j0} + \sum_{l=1}^{L} \alpha_{jl} X_l \tag{21}$$

dla $j = 1, \ldots, J$, a położenie hiperpłaszczyzny separującej klasy, np. C_j, C_k, wyznacza się za pomocą równania:

$$\hat{f}_j(\mathbf{x}) = \hat{f}_k(\mathbf{x}) . \tag{22}$$

Jeżeli chodzi o etap klasyfikacji, to obserwacja $[\mathbf{x}_i, y_i]$ jest przydzielana do tej klasy, dla której wartość teoretyczna funkcji (21) jest największa:

$$\hat{y}_i = \arg\max_j \hat{f}_j(\mathbf{x}_i) . \tag{23}$$

Można także szukać prawdopodobieństw *a posteriori* w klasach, tj. $p(C_j \mid \mathbf{X})$. Jeżeli $g_j(\mathbf{X})$ jest funkcją gęstości w klasie C_j, to ze wzoru Bayesa wynika, że:

$$p(C_j \mid \mathbf{X}) = \frac{g_j(\mathbf{X}) p(C_j)}{\sum_{k=1}^{J} g_k(\mathbf{X}) p(C_k)} . \tag{24}$$

Zakładając, że $g_j(\mathbf{X})$ jest funkcją gęstości wielowymiarowego rozkładu normalnego:

$$g_j(\mathbf{X}) = \frac{1}{(2\pi)^{L/2}\left|\mathbf{\Sigma}_j\right|^{1/2}} e^{-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu}_j)^T \mathbf{\Sigma}_j^{-1}(\mathbf{X}-\boldsymbol{\mu}_j)}, \tag{25}$$

gdzie $\boldsymbol{\mu}_j = [\mu_{j1}, \ldots, \mu_{jL}]$ jest środkiem ciężkości klasy C_j, tj. $\mu_{jl} = \frac{1}{N_j}\sum_{i=1}^{N_j} x_{il}$.

Wtedy, w przypadku dwóch klas C_j, C_k okazuje się, że logarytm ilorazu wiarygodności ma postać:

$$\log\frac{p(C_j \mid \mathbf{X})}{p(C_k \mid \mathbf{X})} = \log\frac{p(C_j)}{p(C_k)} - \frac{1}{2}(\boldsymbol{\mu}_j + \boldsymbol{\mu}_k)^T \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_k) + \mathbf{X}^T \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_k). \tag{26}$$

Wartość logarytmu (26) wyznaczono przy założeniu, że macierze wariancji i kowariancji w obu klasach są równe, tj. $\mathbf{\Sigma}_j = \mathbf{\Sigma}_k$.

W przypadku gdy wszystkie macierze wariancji i kowariancji w klasach są równe:

$$\mathbf{\Sigma} = \mathbf{\Sigma}_1 = \ldots = \mathbf{\Sigma}_J, \tag{27}$$

to obserwacje w każdej klasie tworzą hipersferyczne skupienia tej samej wielkości i otrzymujemy dla każdej klasy liniową funkcję dyskryminacyjną (LDA):

$$f_j(\mathbf{X}) = \mathbf{X}^T \mathbf{\Sigma}^{-1} \boldsymbol{\mu}_j - \frac{1}{2}\boldsymbol{\mu}_j^T \mathbf{\Sigma}^{-1} \boldsymbol{\mu}_j + \log p(C_j). \tag{28}$$

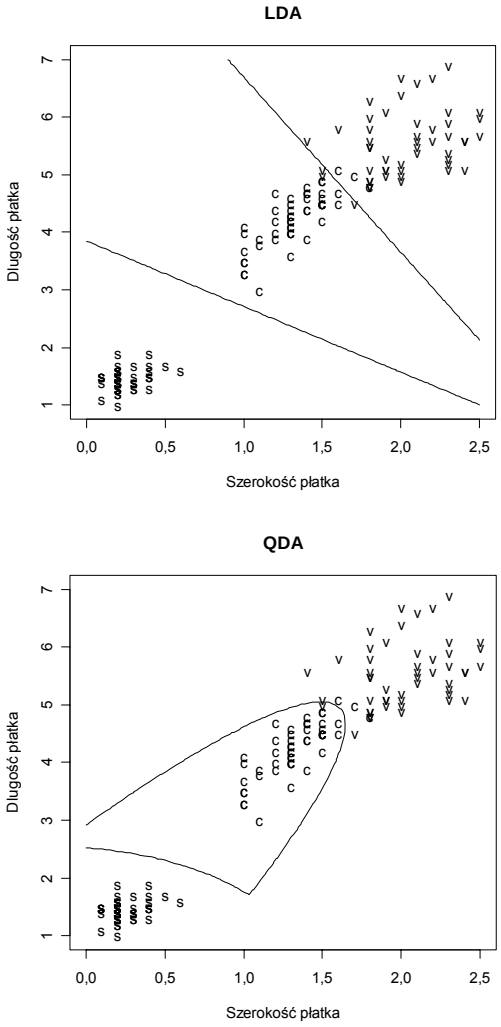
Jeżeli zmienne $\mathbf{X} = [X_1, \ldots, X_L]$ są niezależne i macierz $\mathbf{\Sigma}$ jest diagonalna, to można udowodnić, że hiperpłaszczyzna rozdzielająca dwie klasy C_j i C_k jest prostopadła do odcinka łączącego środki ciężkości obu tych klas. Ponadto, jeśli prawdopodobieństwa *a priori* dla tych klas są jednakowe, hiperpłaszczyzna przechodzi przez środek tego odcinka. W przeciwnym razie hiperpłaszczyzna jest „przesunięta" w stronę środka ciężkości tej klasy, która ma mniejsze prawdopodobieństwo *a priori* $p(C_j)$. Mówiąc obrazowo, obserwacja $\mathbf{x}_i$ ze zbioru rozpoznawanego będzie przydzielona do klasy C_j, której środek ciężkości $\boldsymbol{\mu}_j$ leży najbliżej w sensie odległości euklidesowej.

Jeżeli macierz wariancji i kowariancji $\mathbf{\Sigma}$ nie jest diagonalna, to wynik klasyfikacji obserwacji $\mathbf{x}_i$ ze zbioru rozpoznawanego zależy od odległości Mahalanobisa do środka ciężkości najbliższej klasy.

Gdy warunek (27) nie jest spełniony, to otrzymujemy kwadratowe funkcje dyskryminacyjne (QDA) w klasach:

$$f_j(\mathbf{X}) = -\frac{1}{2}\log\left|\mathbf{\Sigma}_j\right| - \frac{1}{2}(\mathbf{X}-\boldsymbol{\mu}_j)^T \mathbf{\Sigma}_j^{-1}(\mathbf{X}-\boldsymbol{\mu}_j) + \log p(C_j). \quad (29)$$

Na rys. 2 pokazano położenie liniowych i kwadratowych funkcji dyskryminacyjnych dla zbioru IRYS w przestrzeni dwuwymiarowej ($L = 2$). Wybrano dwie zmienne o największej zdolności dyskryminacyjnej, tj. długość płatka (*dp*) i szerokość płatka (*sp*).



Rys. 2. Liniowe i kwadratowe funkcje dyskryminacyjne dla zbioru IRYS

Źródło: Ibid.

Friedman (1989) zaproponował rozwiązanie kompromisowe pomiędzy liniowymi i kwadratowymi funkcjami dyskryminacyjnymi. Metoda ta nazywana jest regularyzowaną analizą dyskryminacyjną (*regularized discriminant analysis*) oraz polega na przekształceniu macierzy wariancji i kowariancji:

$$\boldsymbol{\Sigma}_j(\delta) = \delta \cdot \boldsymbol{\Sigma}_j + (1-\delta)\boldsymbol{\Sigma}. \tag{30}$$

Parametr regularyzacji $\delta \in [0,1]$ pozwala zbudować model dyskryminacyjny w postaci pośredniej pomiędzy liniową ($\delta = 1$) i kwadratową ($\delta = 0$). W praktyce parametr δ ustalany jest eksperymentalnie na podstawie zbioru testowego lub w wyniku sprawdzania krzyżowego.

Dwuparametrową rodzinę macierzy kowariancji można otrzymać, wstawiając do wzoru (30) zamiast $\boldsymbol{\Sigma}$:

$$\boldsymbol{\Sigma}_y = \gamma \boldsymbol{\Sigma} + (1-\gamma)\sigma^2 \mathbf{I}, \tag{31}$$

gdzie $\boldsymbol{I}$ jest macierzą jednostkową, $\gamma \in [0,1]$ jest parametrem regularyzacji, zaś $\sigma^2 \boldsymbol{I}$ to diagonalna macierz wariancji i kowariancji wyznaczona na podstawie próby.

4. Drzewa klasyfikacyjne

W opozycji do klasycznych, parametrycznych metod dyskryminacji, powstały metody nieparametryczne, niewymagające spełnienia przedstawionych w poprzedniej części artykułu wymagań. Należą do nich m.in. metoda *K*-najbliższych sąsiadów i metoda drzew klasyfikacyjnych.

Ta ostatnia polega na sekwencyjnym podziale *L*-wymiarowej przestrzeni zmiennych $\mathbf{X}^L$ na podprzestrzenie R_k (segmenty), aż do chwili, gdy zmienna zależna Y osiągnie w każdej z nich minimalny poziom zróżnicowania (mierzony za pomocą odpowiedniej funkcji straty). Metoda ta nazywana jest metodą rekurencyjnego podziału (*recursive partitioning*) i była stosowana w statystyce już przez Morgana i Sonquista (1963). Jej wykorzystanie w analizie dyskryminacyjnej i regresji przedstawili Breiman i in. (1984), proponując algorytm CART. W języku polskim wyczerpującą monografią poświęconą zagadnieniom budowy modeli w postaci drzew klasyfikacyjnych i regresyjnych jest praca Gatnara (2001).

Przebieg procedury rekurencyjnego podziału najlepiej reprezentuje drzewo, tj. graf spójny i bez cykli; stąd nazwa metody – drzewa klasyfikacyjne* (*classification trees*). W ramach omawianej metody model jest tworzony nie globalnie,

* W istocie prawidłowa nazwa w języku polskim powinna brzmieć: drzewa dyskryminacyjne.

lecz poprzez złożenie modeli lokalnych o najprostszej postaci (tj. stałej), budowanych w każdym z K rozłącznych segmentów, na jakie dzielona jest wielowymiarowa przestrzeń zmiennych:

$$f(\mathbf{x}_i) = \sum_{k=1}^{K} \alpha_k I(\mathbf{x}_i \in R_k), \tag{32}$$

gdzie R_k ($k = 1, \ldots, K$) to podprzestrzenie (segmenty) przestrzeni $\mathbf{X}^L$, α_k – parametry modelu, zaś I jest funkcją wskaźnikową.

Każdy z obszarów R_k jest definiowany poprzez jego granice w przestrzeni $\mathbf{X}^L$, które dla zmiennych metrycznych $X_1, \ldots, X_L$, można przedstawić jako:

$$I(\mathbf{x}_i \in R_k) = \prod_{l=1}^{L} I\left(v_{kl}^{(d)} \le x_{il} \le v_{kl}^{(g)}\right), \tag{33}$$

gdzie wartości $v_{kl}^{(d)}$ oraz $v_{kl}^{(g)}$ oznaczają odpowiednio jego górną i dolną granicę w *l*-tym wymiarze przestrzeni.

Gdy zmienne $X_1, \ldots, X_L$ mają charakter niemetryczny, to podprzestrzeń R_k można zdefiniować jako:

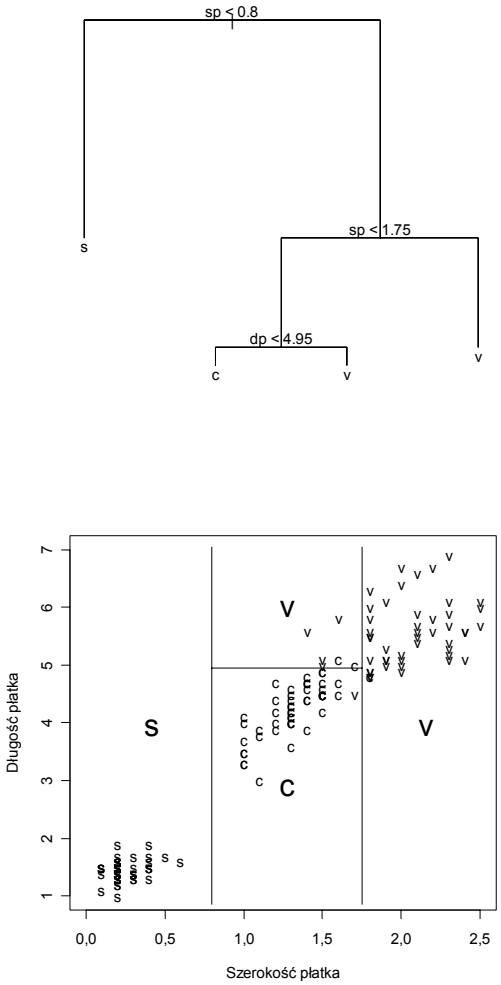
$$I(\mathbf{x}_i \in R_k) = \prod_{l=1}^{L} I\left(x_{il} \in B_{kl}\right), \tag{34}$$

gdzie B_{kl} to podzbiór zbioru kategorii zmiennej X_l, tj. $B_{kl} \subseteq V_l$.

Jeżeli zmienna zależna Y w modelu (32) jest zmienną nominalną, to model ten nazywany jest dyskryminacyjnym i reprezentuje go drzewo klasyfikacyjne. Wtedy parametry α_k modelu (32) są wyznaczane jako:

$$\alpha_k = \arg\max_j p(C_j \mid \mathbf{x}_i \in R_k). \tag{35}$$

Model w postaci drzewa klasyfikacyjnego dla zbioru IRYS, wykorzystujący dwie zmienne objaśniające: długość płatka *(dp)* i szerokość płatka *(sp)*, oraz odpowiadający mu podział przestrzeni dwuwymiarowej na 4 segmenty, zostały pokazane na rys. 1. Jak widać, segment oznaczony literą *S* jest homogeniczny, ponieważ należą do niego wszystkie kwiaty z gatunku *Setosa*. Jego granice wyznacza formuła: $0 < sp < 0{,}8$. Z kolei segment oznaczony literą *C* zawiera większość kwiatów z gatunku *Versicolor*, a jego granicami w pierwszym wymiarze jest $0{,}8 < sp < 1{,}75$ oraz w drugim – $0 < dp < 4{,}95$.



Rys. 3. Drzewo klasyfikacyjne oraz podział przestrzeni zmiennych dla zbioru IRYS

Źródło: Ibid.

Do oceny stopnia zróżnicowania podprzestrzeni R_k można wykorzystać jedną z następujących miar:

– błąd klasyfikacji (*misclassification error*):

$$Q(R_k) = 1 - \arg\max_j p(C_j \mid R_k), \tag{36}$$

– wskaźnik Giniego (*Gini index*):

$$Q(R_k) = 1 - \sum_{j=1}^{J} \left(p(C_j \mid R_k) \right)^2 , \tag{37}$$

– entropia:

$$Q(R_k) = -\sum_{j=1}^{J} p(C_j \mid R_k) \log_2 p(C_j \mid R_k) . \tag{38}$$

Omówienie własności przedstawionych powyżej miar oraz charakterystyka innych, nieco mniej znanych, znajduje się w pracy Gatnara (2001).

Powyższe miary homogeniczności są wykorzystywane do kontroli procesu podziału przestrzeni zmiennych. Stosowana jest w tym celu strategia wspinaczki (*hill climbing*), pozwalająca dokonać podziału, który jest optymalny w sensie lokalnym. Nie gwarantuje to oczywiście osiągnięcia rozwiązania optymalnego w sensie globalnym.

W każdym kroku ocena jakości podziału podprzestrzeni R na segmenty $R_1, \ldots, R_K$ odbywa się za pomocą kryterium:

$$\Delta Q(R) = Q(R) - \sum_{k=1}^{K} Q(R_k)\, p(k) , \tag{39}$$

gdzie $p(k)$ oznacza frakcję obserwacji w segmencie R_k. Kryterium (39) podlega maksymalizacji, tj. szukany jest taki podział, który zapewni jak największą jednorodność uzyskanych podprzestrzeni, czyli osiągnięcie minimum przez $Q(R_k)$ dla $k = 1, \ldots, K$.

Breiman i in. (1984) wykorzystali w swojej pracy do oceny homogeniczności segmentów wskaźnik Giniego (37). Ma on jednak pewną wadę, ponieważ osiąga maksimum również wtedy, gdy segmenty R_k zawierają jednakową liczbę obserwacji. Z kolei Quinlan (1993) w swoim algorytmie C4.5 stosuje entropię (38), której główna wada polega na tym, że preferuje ona taki podział, który generuje maksymalną liczbę segmentów R_k. Aby tego uniknąć, można zastosować normalizację, uzyskując tzw. względny przyrost informacji (*gain ratio*):

$$\Delta Q^*(R) = \frac{\Delta Q(R)}{-\sum_{k=1}^{K} p(k) \log_2 p(k)} . \tag{40}$$

Podział przestrzeni $\mathbf{X}^L$ na podprzestrzenie odbywa się za pomocą hiperpłaszczyzn równoległych do osi (gdy zmienne $X_1, \ldots, X_L$ są zmiennymi metrycznymi). Równanie takiej hiperpłaszczyzny ma wtedy postać $X_l = c$, gdzie zarówno wybór zmiennej X_l, jak i wartości c kontroluje miara (39).

Aby wyznaczyć stałą c, należy obliczyć wartość kryterium (39) dla wszystkich możliwych wariantów podziału zbioru wartości $V_l = \{v_{l1}, \ldots, v_{lT}\}$ zmiennej X_l:

$$c = \frac{v_{lt} + v_{lt+1}}{2}. \tag{41}$$

Zawsze uzyskuje się w ten sposób dwa zbiory obserwacji: $\{\mathbf{x}_i : x_{il} \le c\}$ oraz $\{\mathbf{x}_i : x_{il} > c\}$. Inaczej mówiąc, dokonywana jest dyskretyzacja zmiennej X_l, której rezultatem jest powstanie drzewa binarnego, w którym z każdego węzła wychodzą dwie krawędzie.

W procesie budowy modelu w postaci drzewa klasyfikacyjnego najpierw każda zmienna metryczna poddawana jest dyskretyzacji*, a następnie wybierana jest ta spośród nich, dla której kryterium (39) osiąga maksimum.

Jeżeli zmienna X_l ma charakter niemetryczny, to zbiór jej kategorii $V_l = \{v_{l1}, \ldots, v_{lT}\}$ jest dzielony na dwa podzbiory (w przypadku drzewa binarnego), tak aby wartość kryterium (39) była jak największa (takich podziałów jest 2^T dla zmiennych porządkowych oraz $2^{T-1} - 1$ dla zmiennych nominalnych). Najczęściej punktem wyjścia jest podział V_l na T podzbiorów $\{v_{l1}\}, \ldots, \{v_{lT}\}$, a następnie te podzbiory są stopniowo łączone. W metodzie CHAID, którą zaproponował Kaas (1980) tym procesem łączenia steruje statystyka χ^2.

W przypadku modeli w postaci drzew klasyfikacyjnych pojawia się problem wyboru takiej postaci modelu, by jego błąd predykcji był jak najmniejszy. Spośród metod wykorzystywanych w celu wyeliminowania tego zjawiska i zmniejszenia stopnia złożoności modelu, najczęściej** stosuje się tzw. przycinanie krawędzi (*pruning*). Zabieg ten powoduje redukcję wielkości drzewa poprzez usunięcie niektórych jego fragmentów, co może oznaczać, że z modelu zostaną wyeliminowane niektóre zmienne.

Breiman i in. (1984) zaproponowali pewną formę regularyzacji, która pozwala uzyskać kompromis pomiędzy złożonością modelu i jego jakością w postaci kryterium:

$$S_\lambda(D) = Q(D) + \lambda \cdot K, \tag{42}$$

---

* W pracy Gatnara (2001) omówiono także metody podziału zbioru wartości zmiennej X_l na trzy i więcej przedziałów (*multiway split*), w rezultacie czego powstają drzewa niebinarne. To zagadnienie jest jednak jeszcze bardziej złożone.

** Gatnar (2001) omawia także rzadziej stosowaną metodę skracania krawędzi drzewa (ang. *shrinking*), które są proporcjonalne do stopnia homogeniczności w węzłach.

które podlega minimalizacji. W powyższej formule $Q(D) = \sum_{k=1}^{K} Q(R_k)\,p(k)$ to miara jakości modelu D w postaci drzewa, K oznacza liczbę liści i jest oceną złożoności modelu, zaś λ to tzw. parametr złożoności ($\lambda \geq 0$). Duże wartości parametru λ oznaczają podział na niewiele segmentów (małe drzewa), zaś małe wartości – drzewa bardziej rozbudowane, o dużej liczbie liści. W przypadku gdy $\lambda = 0$, powstaje drzewo maksymalne (pełne) D_0.

5. Kierunki rozwoju analizy dyskryminacyjnej

Wzrost możliwości przetwarzania dużych zbiorów danych przez współczesne komputery oraz dostępność zaawansowanego oprogramowania statystycznego powoduje, że spośród metod analizy dyskryminacyjnej najszybciej rozwijają się metody nieparametryczne, wykorzystywane w systemach *business intelligence* i *data mining*.

Należą do nich np.: metoda K-najbliższych sąsiadów oraz drzewa klasyfikacyjne. Ważną zaletą tej ostatniej klasy modeli jest możliwość klasyfikacji danych niepełnych, tj. zawierających obserwacje, dla których nie można określić wartości pewnych zmiennych, np. gdy są one trudne do zmierzenia. Metoda ta jest również odporna na występowanie wartości nietypowych. Ponadto w modelu dyskryminacyjnym zmiennymi objaśniającymi mogą być zmienne mierzone zarówno na skalach mocnych, jak i na skalach słabych, bez konieczności dokonywania ich transformacji.

Od dłuższego czasu utrzymuje się zainteresowanie także takimi nieklasycznymi metodami dyskryminacji, jak: sieci neuronowe (*neural networks*) oraz metoda wektorów nośnych SVM (*Suport Vector Machines*). Sieć neuronowa to model złożony z wielu modeli liniowych znajdujących się w poszczególnych warstwach sieci, które przetwarzają dane wejściowe i korygują („uczą się"), w czasie tysięcy powtórzeń, parametry poszczególnych modeli składowych (Rosenblatt, 1958).

Z kolei metoda SVM, zaproponowana przez Vapnika (1995), polega na transformacji obserwacji z pierwotnej przestrzeni zmiennych objaśniających w przestrzeń o wiele większym wymiarze, w której klasy są łatwiej separowalne. Obserwacje definiujące położenie funkcji oddzielających klasy nazywane są wektorami nośnymi.

Warto także wspomnieć o metodach łączenia modeli dyskryminacyjnych, które powodują znaczące zwiększenie dokładności klasyfikacji. Zalety tego podejścia, nazywanego wielomodelowym, przedstawił wyczerpująco w swojej monografii Gatnar (2008).

Duży wysiłek badawczy jest wkładany obecnie w poszukiwanie metod dyskryminacji dla dużych i wielowymiarowych zbiorów danych. Należą do nich przede wszystkim zbiory danych o genotypach (*gene microarray data*).

Bibliografia

Breiman L., Friedman J., Olshen R., Stone C. (1984): *Classification and Regression Trees*. CRC Press, London.

Bryan J.G. (1951): *The Generalized Discriminant Function: Mathematical Foundation and Computational Routine*. Harvard Education Review, 21, s. 90-95.

Duda R. O., Hart P. E., Storck G. E. (2001): *Pattern Classification*. John Wiley & Sons, New York.

Fisher L.A. (1936): *The Use of Multiple Measurements in Taxonomic Problems*. Annals of Eugenics, t. 7, s. 179-188.

Friedman J. H. (1989): *Regularized Discriminant Analysis*. „Journal of the American Statistical Association", 84, s. 165-175.

Gatnar E. (1998): *Symboliczne metody klasyfikacji danych*. Wydawnictwo Naukowe PWN, Warszawa.

Gatnar E. (2001): *Nieparametryczna metoda dyskryminacji i regresji*. Wydawnictwo Naukowe PWN, Warszawa.

Gatnar E. (2008): *Podejście wielomodelowe w zagadnieniach dyskryminacji i regresji*. Wydawnictwo Naukowe PWN, Warszawa.

Hastie T., Tibshirani R., Friedman J. (2001): *The Elements of Statistical Learning*. Springer Series in Statistics, Springer, Berlin.

Huberty G. J. (1995): *Applied Discriminant Analysis*. John Wiley & Sons, New York.

Jajuga K. (1990): *Statystyczna teoria rozpoznawania obrazów*. Wydawnictwo Naukowe PWN, Warszawa.

Jajuga K. (1993): *Statystyczna analiza wielowymiarowa*. Państwowe Wydawnictwo Naukowe, Warszawa.

Kaas G. V. (1980): *An Exploratory Technique for Investigating Large Quantities of Categorical Data*. „Applied Statistics", 29, s. 119-127.

Kolonko J. (1980): *Analiza dyskryminacyjna w badaniach ekonomicznych*. PWN, Warszawa.

McLachlan G.J. (1992): *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley & Sons, New York.

Morgan J.N., Sonquist J.A. (1963): *Problems in the Analysis of Survey Data: A Proposal*. „Journal of the American Statistical Association", 58, s. 417-434.

Nilsson N.J. (1965): *Learning Machines: Foundations of Trainable Pattern-Classifying Systems*. McGraw-Hill.

Quinlan J.R. (1983): *Learning Efficient Classification Procedures and their Application to Chess and Games*. W: R. Michalski, J. Carbonell, T. Mitchell (eds.): *Machine Learning. An Artificial Intelligence Approach*. Tioga, Palo Alto, s. 126-142.

Quinlan J.R. (1986): *Induction of decision trees*, Machine Learning, 1, s. 81-106.

Quinlan J.R. (1993): *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA.

Rao C. (1948): *The Utilisation of Multiple Measurements in Problems of Biological Classification*. „Journal of the Royal Statistical Society B", 10, s. 159-203.

Ripley B.D. (1996): *Pattern Recognition and Neural Networks*. Cambridge University Press. Cambridge.

Rosenblatt F. (1958): *The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain*. „Psychological Review", 65(6), s. 386-408.

Titterington D.M., Murray G.D., Murray L.S., Spiegelhalter D.J., Skene A.M., Habbema J.D., Gelpke G.J. (1981): *Comparison of Discriminant Techniques Applied to Complex Data Sets of Head Injured Patients*. „Journal of the Royal Statistical Society, Series A", 144, s. 145-175.

Vapnik V. (1995): *The Nature of Statistical Learning Theory*. Springer, Berlin.

Vapnik V. (1998): *Statistical Learning Theory*. John Wiley and Sons, New York.

Wernecke K.-D. (1992): *A Coupling Procedure for Discrimination of Mixed Data*. „Biometrics", 48, s. 497-506.

DISCRIMINANT ANALYSIS – STATE OF THE ART AND FUTURE DEVELOPMENTS

Summary

The aim of the discriminant analysis is to partition the multivariate feature space into subspaces in order to separate observations belonging to different classes. In other words, its task is to find a model that can give class descriptions on the basis of a set containing previously classified observations. Then the model is applied to classify new ones with a minimum error. Founded in 1936 by Fisher, the discriminant analysis had become an important part of multivariate statistical analysis. It has many applications and is an obligatory procedure in many available data mining systems.

In Poland prof. Józef Kolonko has been one of the pioneering statisticians interested in discriminant analysis. He published his book on discriminant analysis in 1980, based on cybernetics. Therefore the analysis had a broader meaning, including both supervised and unsupervised classification.

Nowadays, in statistical literature, discriminant analysis is considered only as a set of methods that can discover class descriptions on the basis of the training set, i.e. supervised classification.

This article is devoted to review the existing approaches and new developments in discriminant analysis starting from its roots, i.e. Fisher's approach based on regression analysis, and concluding with classification trees as a nonparametric discriminant analysis techmique. We also mention new approaches recently proposed in statistical literature, such as: neural nets, K-nearest neighbors and support vector machines.

Ewa Genge
Uniwersytet Ekonomiczny w Katowicach

ROLA KOBIET W POLSKIM SPOŁECZEŃSTWIE – ANALIZA EMPIRYCZNA Z WYKORZYSTANIEM MODELI KLAS UKRYTYCH DLA DANYCH JAKOŚCIOWYCH

Wprowadzenie

Modele klas ukrytych (ang. *latent class models*), zwane również analizą klas ukrytych (ang. *latent class analysis*) należą do tzw. modeli ze zmiennymi ukrytymi (ang. *latent variable models*), w których ukrytą zmienną jest klasa. Modele te można zaliczyć również do tzw. podejścia modelowego w taksonomii (ang. *model-based clustering*), gdzie wykorzystywana jest idea mieszanek rozkładów (zob. Domański, Pruska, 2000; Witek, 2009). W odróżnieniu od heurystycznych metod taksonomicznych (tj. metod hierarchicznych, iteracyjno-aglomeracyjnych), w których podstawą klasyfikacji obiektów do klas są różnego rodzaju miary odległości, w podejściu modelowym obiekty klasyfikowane są na podstawie prawdopodobieństw.

Istotą modelowania klas ukrytych jest badanie związków między kategoriami zmiennych nominalnych i porządkowych. Wykorzystuje ona dane zawarte w tablicy kontyngencji. Metoda ta została wprowadzona przez Lazarsfelda (1950) w latach 50. XX w., a w kolejnych latach rozwijana przez Goodmana (1970), który przyczynił się do rozwinięcia algorytmu pozwalającego otrzymać parametry funkcji największej wiarygodności, oraz Habermana (1979), który pokazał związek pomiędzy modelami klas ukrytych oraz modelami logarytmiczno-liniowymi. Metoda ta nadal cieszy się dużym zainteresowaniem i rozwijana jest m.in. przez uczonych, takich jak Hagenaars (2002), Vermunt (2010), Linzer i Lewis (2011).

1. Model klas ukrytych – definicja

Rozważa się zbiór n obiektów, charakteryzowanych za pomocą zmiennych dychotomicznych lub politomicznych, zwanych zmiennymi obserwowanymi (ang. *manifest variables*) o wielu kategoriach $l_1, \ldots, l_m$ (zob. Bąk, 2011). Zbiór

wszystkich obiektów można więc zapisać za pomocą wektora $\mathbf{x}_i = (x_{ijh}; j = 1, \ldots, m; h = 1, \ldots, l_j; i = 1, \ldots, n)$, gdzie $x_{ijh} = 1$ oznacza *i*-tą obserwację na *j*-tej zmiennej o *h*-tej kategorii. Przyjmując, że liczba wszystkich kategorii jest równa $l = \sum_{j=1}^{m} l_j$, zbiór określany jest za pomocą macierzy o wymiarach $n \times m$.

Model klas ukrytych dla danych jakościowych można zapisać jako mieszankę rozkładów wielomianowych, w której zakłada się, że każda obserwacja $\mathbf{x}_i$ pochodzi z mieszanki wielowymiarowych rozkładów wielomianowych (ang. *mixture of multivariate multinominal distributions*), określonej jako:

$$f(\mathbf{x}_i|\boldsymbol{\Theta}) = \sum_{s=1}^{u} \tau_s f_s(x_i|\boldsymbol{\Theta}_s), \tag{1}$$

gdzie:

f_s – funkcja gęstości ukrytej klasy P_s, (*s*-tego rozkładu składowego mieszanki),

$\mathbf{x}_i$ – wektor realizacji zmiennych obserwowanych $\mathbf{x}_i = [x_{i1,\ldots,}x_{im_1}]$,

$\boldsymbol{\Theta}_s$ – wektor parametrów ukrytej klasy P_s,

$\boldsymbol{\Theta}$ – wektor wszystkich parametrów mieszanki rozkładów, $\boldsymbol{\Theta} = (\tau_s, \boldsymbol{\Theta}_s)$,

τ_s – prawdopodobieństwo *a priori* – wartość prawdopodobieństwa, że dana obserwacja należy do klasy P_s $(\tau_s \geq 0 \wedge \sum_{s=1}^{u} \tau_s = 1), \boldsymbol{\Theta}_\mathbf{s} \neq \boldsymbol{\Theta}_l \forall s \neq l.$

Rozkłady składowe można zaś zapisać jako:

$$f_s(\mathbf{x}_i|\boldsymbol{\Theta}_s) = \prod_{j=1}^{m} \prod_{h=1}^{l_j} (\Theta_{sjh})^{x_{ijh}}, \tag{2}$$

gdzie $\boldsymbol{\Theta}_s = (\Theta_{sjh}; j = 1, \ldots, m; h = 1, \ldots, l_j)$. Równanie (2) rozumiane jest jako iloczyn m niezależnych rozkładów wielomianowych o parametrach $\boldsymbol{\Theta}_{sj}$. Parametry mieszanki oznaczone są za pomocą wektora $\boldsymbol{\Theta} = (\tau_{1,\ldots,u}, \boldsymbol{\Theta}_{1,\ldots,u})$.

Dla danych estymatorów $\hat{\tau}_s$ i Θ_{sjh} prawdopodobieństwa *a posteriori* przynależności obiektów do poszczególnych klas mogą być obliczone za pomocą wzoru Bayes'a:

$$P(s|\mathbf{x}_i, \boldsymbol{\Theta}) = \frac{\hat{\tau}_s f(\mathbf{x}_i, \hat{\boldsymbol{\Theta}}_s)}{\sum_{q=1}^{u} \hat{\tau}_q f(\mathbf{x}_i, \hat{\boldsymbol{\Theta}}_q)}. \tag{3}$$

Należy zauważyć, że liczba szacowanych niezależnie parametrów modelu klas ukrytych wzrasta wraz z liczbą klas, zmiennych i ich kategorii. Liczba szacowanych parametrów wynosi $u\sum_j (l_j - 1) + (u - 1)$. Jeżeli liczba ta przekroczy liczebność zbioru lub łączną liczbę komórek w tablicy kontyngencji dla zmiennych obserwowanych, wtedy model klas ukrytych stanie się modelem nieidentyfikowalnym.

2. Model klas ukrytych z zmiennymi towarzyszącymi

Model klas ukrytych oprócz zmiennych obserwowanych może zawierać jeszcze tzw. zmienne towarzyszące (ang. *covariates* lub *concomitant variables*), mające wpływ na przynależność obiektów do klas – wpływ na prawdopodobieństwa *a priori* (zob. np. Dayton i Macready, 1988; Hagenaars i McCutcheon, 2002). Zmienne towarzyszące wraz ze zmiennymi $X_1, \ldots, X_m$ biorą udział w szacowaniu parametrów modelu klas ukrytych, na podstawie którego będzie można dokonać klasyfikacji nowych obiektów bez udziału zmiennych obserwowanych. Zmienne towarzyszące wykorzystywane są często w badaniach marketingowych, ekonomicznych, psychologicznych, w których pozyskanie zmiennych obserwowanych jest bardzo kosztowne (por. Witek, 2011).

Najczęściej parametry zmiennych towarzyszących szacowane są wraz z pozostałymi parametrami modelu klas ukrytych (jednocześnie). Ten sposób estymacji zwany jest jednokrokową techniką estymacji parametrów zmiennych towarzyszących (ang. *one-step technique for estimating the effects of covariates*) (zob. np. Dayton i Macready 1988; Hagenaars i McCutcheon, 2002). Alternatywnym sposobem estymacji parametrów zmiennych towarzyszących jest tzw. podejście trzykrokowe (ang. *three-step approach*), w którym szacowane są parametry klasycznego modelu klas ukrytych (1), następnie obliczane są prawdopodobieństwa *a posteriori* (3). W kroku trzecim szacowane są parametry równania regresji, gdzie prawdopodobieństwa te traktowane są jako zmienne zależne, a zmienne towarzyszące jako zmienne objaśniające. Jednakże Bolck, Crown i Hagenaars (2004) udowodnili, że w wyniku szacunku parametrów trzykrokową metodą estymacji, estymatory parametrów takiego modelu są obciążone.

Włączając do modelu klas ukrytych zmienne towarzyszące, zakładamy, że mają one wpływ na prawdopodobieństwa *a priori*. W klasycznym modelu klas ukrytych (bez zmiennych towarzyszących) zakładamy, że każda obserwacja ma takie samo prawdopodobieństwo przynależności do klasy ukrytej.

W przypadku gdy zmienne towarzyszące mają wpływ na prawdopodobieństwa przynależności obiektów do klas (τ_s), model klas ukrytych zapisać można jako:

$$f(\mathbf{x}_i, \mathbf{z}_i | \boldsymbol{\Theta}) = \sum_{s=1}^{u} \tau_s(\mathbf{z}_i, \boldsymbol{\alpha}) f_s(x_i | \boldsymbol{\Theta}_s), \tag{4}$$

gdzie: $\mathbf{z}_i$ – wektor realizacji zmiennych towarzyszących, $\mathbf{z}_i = [z_{i1,\ldots,}x_{im_2}]$.

Nadal jednak spełniony musi być warunek, że $(\tau_s(\mathbf{z}_i, \boldsymbol{\alpha}) \geq 0 \wedge \sum_{s=1}^{u} \tau_s(\mathbf{z}_i, \boldsymbol{\alpha}) = 1), \boldsymbol{\Theta}_\mathbf{s} \neq \boldsymbol{\Theta}_l \forall s \neq l.$ Wpływ zmiennych towarzyszących na prawdopodobieństwa *a priori* wyrażany jest za pomocą wielomianowej funkcji logitowej (zob. Agresti, 2002).

Jeżeli w szacowaniu parametrów modelu klas ukrytych biorą udział zmienne towarzyszące, zazwyczaj pierwsza z klas jest tzw. klasą referencyjną. Zakłada się wtedy, że iloraz szans prawdopodobieństw *a priori* dla klas ukrytych, w porównaniu do tej klasy (klasy referencyjnej) jest liniową funkcją zmiennych towarzyszących. Dla m_2 zmiennych towarzyszących, wektor parametrów tych zmiennych $\boldsymbol{\alpha}_s$ ma długość $m_2 + 1$ (dla każdej zmiennej towarzyszącej i wyrazu wolnego). Ponieważ pierwsza klasa jest klasą referencyjną, z definicji $\boldsymbol{\alpha}_1 = 0$. Wtedy:

$$\ln(\tau_{2i} / \tau_{1i}) = \mathbf{z}_i \boldsymbol{\alpha}_2 \tag{5}$$

$$\ln(\tau_{3i} / \tau_{1i}) = \mathbf{z}_i \boldsymbol{\alpha}_3 \tag{6}$$

$$\vdots$$

$$\ln(\tau_{ui} / \tau_{1i}) = \mathbf{z}_i \boldsymbol{\alpha}_u \tag{7}$$

W wyniku kilku przekształceń otrzymujemy:

$$\tau_{si} = \tau_s(\mathbf{z}_i; \boldsymbol{\alpha}) = \frac{e^{\mathbf{z}_i \boldsymbol{\alpha}_s}}{\sum_{q=1}^{u} e^{\mathbf{z}_i \boldsymbol{\alpha}_q}}. \tag{8}$$

W modelu klas ukrytych z udziałem zmiennych towarzyszących, szacowanych jest więc $u-1$ wektorów $\boldsymbol{\alpha}_s$, a także warunkowych prawdopodobieństw przynależności obiektów do klas ukrytych. Mając dane estymatory $\hat{\boldsymbol{\alpha}}_s$ i Θ_{sjh}, prawdopodobieństwa *a posteriori* i przynależności obiektów do klas uzyskiwane są poprzez zastąpienie τ_s w równaniu (3) funkcją $\tau_s(\mathbf{z}_i; \boldsymbol{\alpha})$ z równania (8):

$$P(s | \mathbf{x}_i, \mathbf{z}_i) = \frac{\hat{\tau}_s(\mathbf{z}_i; \hat{\boldsymbol{\alpha}}) f(\mathbf{x}_i, \hat{\boldsymbol{\Theta}}_s)}{\sum_{q=1}^{u} \hat{\tau}_q(\mathbf{z}_i; \hat{\boldsymbol{\alpha}}) f(\mathbf{x}_i, \hat{\boldsymbol{\Theta}}_q)}. \tag{9}$$

Liczba szacowanych parametrów takiego modelu klas ukrytych jest równa $u\sum_j (l_m - 1) + (s+1)(u-1)$.

3. Estymacja parametrów

Estymacja modelu klas ukrytych polega m.in. na oszacowaniu liczby i wielkości poszczególnych klas. Metodą największej wiarygodności szacowane są parametry modelu klas ukrytych (4). Funkcja największej wiarygodności określona jest wzorem:

$$\ln L = \sum_{i=1}^{n} \ln \sum_{s=1}^{u} \tau_s(\mathbf{z}_i;\boldsymbol{\alpha}) \prod_{j=1}^{m} \prod_{h=1}^{l_j} (\Theta_{sjh})^{x_{ijh}} . \tag{10}$$

Popularną metodą szacowania parametrów największej wiarygodności jest algorytm EM (Dempster et al., 1977). W pakiecie poLCA wykorzystywana jest zmodyfikowana wersja algorytmu EM (zob. Bandeen-Roche et al., 1977). Proces estymacji zapoczątkowany jest przez wartości startowe dla $\widehat{\boldsymbol{\alpha}}'_s$ i Θ'_{sjh}, dzięki którym wyznaczone są prawdopodobieństwa *a posteriori* $P(s|\mathbf{x}_i, \mathbf{z}_i)$ dane wzorem (9). Parametry zmiennych towarzyszących szacowane (i uaktualniane) są zgodnie z formułą:

$$\widehat{\boldsymbol{\alpha}}_s = \widehat{\boldsymbol{\alpha}}'_s + (-\mathbf{D}_\alpha^2 \log L)^{-1} \mathbf{D}_\alpha \log L, \tag{11}$$

gdzie $\widehat{\boldsymbol{\alpha}}'_s$ to wektor estymatorów parametrów zmiennej towarzyszącej, $\mathbf{D}_\alpha$ to gradient, zaś $\mathbf{D}_\alpha^2$ hesjan macierzy z parametrem $\boldsymbol{\alpha}$. Nowe wartości parametrów Θ_{sjh} wyznaczane są za pomocą formuły:

$$\Theta_{sj} = \frac{\sum_{i=1}^{n} \mathbf{x}_{ij} P(s|\mathbf{x}_i, \mathbf{z}_i)}{\sum_{i=1}^{n} P(s|\mathbf{x}_i, \mathbf{z}_i)}. \tag{12}$$

Kroki algorytmu powtarzane są dopóty, dopóki przyrost funkcji wiarygodności nie będzie mniejszy niż zadana wartość graniczna lub nie zostanie osiągnięta maksymalna liczba iteracji. Wzory oraz szczegółowe informacje dotyczące gradientu $\mathbf{D}_\alpha$ oraz hesjanu $\mathbf{D}_\alpha^2$ można znaleźć w pracy Bandeen-Roche et al. (1997).

4. Wybór modelu i ocena jakości dopasowania

Jedną z głównych zalet modeli klas ukrytych jest to, że w odróżnieniu od popularnych metod taksonomicznych (tj. k-średnich, metody Warda), istnieje kilka statystycznych miar służących wyborowi i ocenie ich jakości dopasowania. Najczęściej w różnego rodzaju badaniach empirycznych na początku sprawdza się dopasowanie dla $s = 1$. W kolejnych krokach zwiększa się liczbę klas o jeden, tak długo aż model osiągnie najlepsze dopasowanie. Należy jednak pamiętać, że wraz z dodatkową liczbą klas, liczba szacowanych parametrów wzrasta o $1 + \sum_j (l_j - 1)$, dlatego najczęściej wykorzystywane są kryteria informacyjne, będące wyrazem kompromisu pomiędzy jakością dopasowania a złożonością modelu. Do najbardziej popularnych kryteriów informacyjnych zaliczane są: Bayesowskie kryterium informacyjne Schwarza BIC (*Bayesian Information Criterion*), kryterium informacyjne Akaike AIC (*Akaike Information Criterion*). Kryteria te mogą dawać niejednoznaczne wskazania co do oceny modeli klas ukrytych.

Istnieje kilka formuł zapisu wspomnianych kryteriów oceny dopasowania modeli klas ukrytych. W pakietach programu **R** najczęściej wykorzystywane są kryteria podlegające minimalizacji. Można je przestawić na pomocą następujących wzorów:

$$BIC_s = -2 \log P(\mathbf{x}_i \left| \hat{\boldsymbol{\Theta}}_s, M_s \right.) + v_s \log(n), \tag{13}$$

$$AIC_s = -2 \log P(\mathbf{x}_i \left| \hat{\boldsymbol{\Theta}}_s, M_s \right.) + 2v_s, \tag{14}$$

gdzie:

$\log P(\mathbf{x}_i \left| \hat{\boldsymbol{\Theta}}_s, M_s \right.)$ – logarytm funkcji wiarygodności dla oszacowanego wektora parametrów modelu,

M_s, v_s – liczba parametrów modelu,

n – liczba obserwacji.

Pierwsza część powyższych równań odpowiada za wybór modeli o najwyższej dobroci dopasowania, zaś część druga odrzuca modele z nadmierną liczbą parametrów. Porównania różnych kryteriów informacyjnych można znaleźć m.in. w pracach: McLachlan i Peel (2000), Biernacki et al. (1999), Bozdogan (2000). W części empirycznej pracy wykorzystano dwa najbardziej popularne kryteria, tj. BIC oraz AIC. Kryteria te stosowane są w celach porównawczych modeli o różnej liczbie klas. Im niższa wartość kryteriów, tym lepsza jakość dopasowania danego modelu.

5. Analiza empiryczna

Analizę klas ukrytych przeprowadzono na podstawie danych uzyskanych z bezpłatnej bazy danych Polskiego Generalnego Sondażu Społecznego (PGSS) 1992-2008*. W niniejszym artykule rozważano dane z 2008 r. Analiza została przeprowadzona z uwzględnieniem sześciu zmiennych i z pominięciem odpowiedzi „nie wiem" („trudno powiedzieć"). Badana próba liczyła 986 osób.

W przykładzie wykorzystano sześć zmiennych obserwowanych $X_1 - X_6$. W nawiasie podano oryginalne nazwy ze zbioru PGSS 2008.

1. X_1 (q5): Kobiety nie nadają się do polityki (1 – zgadzam się; 2 – nie zgadzam się);
2. X_2 (q6): Rządzenie krajem pozostawić mężczyznom (1 – zgadzam się; 2 – nie zgadzam się);
3. X_3 (q7a): Pracująca matka może zapewnić ciepło (1 – zgadzam się; 2 – nie zgadzam się);
4. X_4 (q7b): Żona niech zapewni mężowi karierę (1 – zgadzam się; 2 – nie zgadzam się);
5. X_5 (q7c): Praca matki szkodzi dziecku (1 – zgadzam się; 2 – nie zgadzam się)
6. X_6 (q7d): Lepiej gdy mężczyzna zarabia/kobieta w domu (1 – zgadzam się; 2 – nie zgadzam się).

Uwzględniono również następujące zmienne towarzyszące:

a) Z_1: płeć respondenta (1 – mężczyzna, 2 – kobieta);
b) Z_2: stan cywilny: kawaler, konkubinat, żonaty, rozwiedziony, separacja, wdowiec;
c) Z_3: wykształcenie: zawodowe (niepełne podstawowe, podstawowe, zasadnicze zawodowe), średnie (niepełne średnie, średnie ogólnokształcące, średnie zawodowe, policealne/pomaturalne, nieukończone studia wyższe), wyższe (ukończone studia licencjackie, ukończone studia magisterskie).

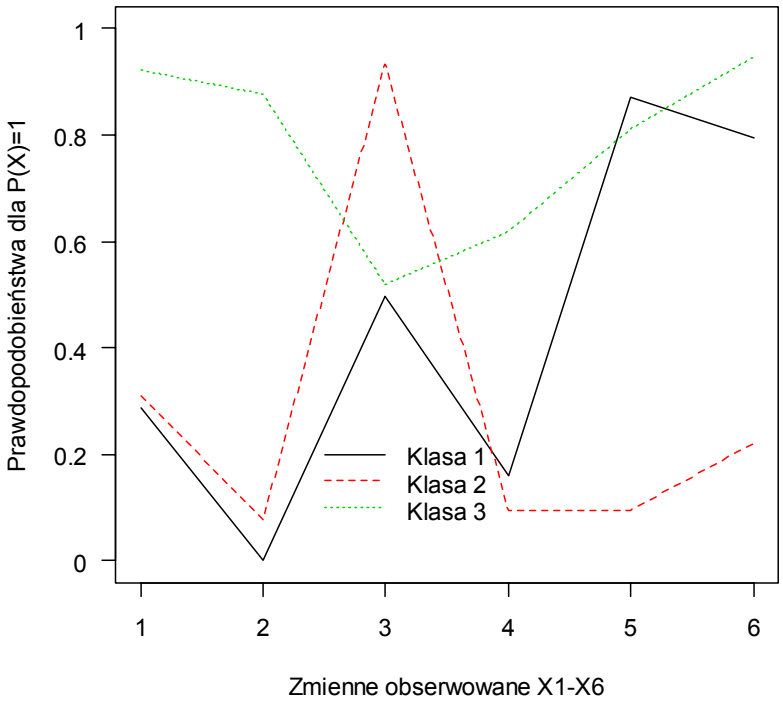
W badaniach wykorzystano pakiet poLCA programu **R**.

Aby wybrać optymalną liczbę klas ukrytych (ukrytą liczbę składowych modelu), obliczono wartości kryteriów informacyjnych AIC oraz BIC dla liczby klas $s = 1, \ldots, u$ dla tzw. modelu podstawowego, tj. bez udziału zmiennych towarzyszących (ang. *base model*), (zob. np. Collins i Lanza, 2011). W przypadku analizowanego zbioru danych kryteria wskazały minimalną wartość dla liczby klas równej cztery. Niewiele większą wartość otrzymano dla trzech klas. W takich sytu-

* Dane dostępne na stronie: http://pgss.iss.uw.edu.pl.

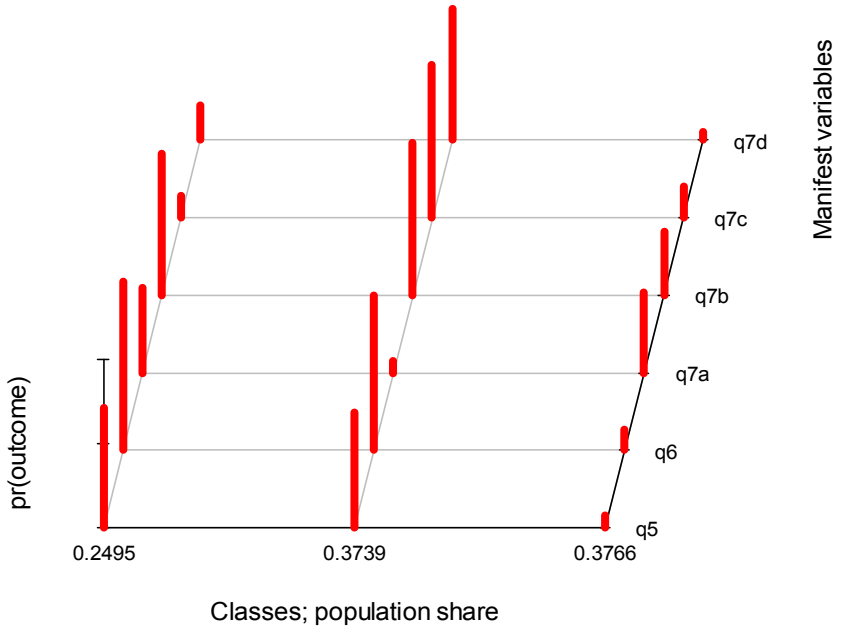
acjach często wybierane są modele mniej złożone (zob. np. Collins i Lanza, 2011), dlatego też w dalszej części pracy analizowano model o trzech klasach ukrytych.

Następnie szacowano modele klas ukrytych dla zmiennych $X_1 - X_6$ i różnych zestawach zmiennych towarzyszących (np. $Z_1 + Z_2$, $Z_1 + Z_3$). Rozważano również interakcje pomiędzy zmiennymi towarzyszącymi, ale wszystkie z nich okazały się nieistotne. Na podstawie analizy przeprowadzonych obliczeń (analiza kryteriów informacyjnych oraz badania istotności parametrów za pomocą testu t-Studenta) przyjęto ostateczny podział badanej próby respondentów na trzy klasy, z wykorzystaniem trzech zmiennych towarzyszących. Dla wybranego modelu przedstawiono prawdopodobieństwa przyjmowania przez zmienne obserwowane wartości 1 („zgadzam się") w klasie pierwszej, drugiej i trzeciej (rys. 1).



Rys. 1. Prawdopodobieństwo wyboru wartości 1 dla zmiennych $X_1 - X_6$

Na rys. 2 przedstawiono prawdopodobieństwa wyboru pierwszej kategorii dla zmiennych $X_1 - X_6$ (odpowiedź na „tak") dla każdej z klas. Wysokość słupków oznacza prawdopodobieństwa odpowiedzi „tak/zgadzam się". Widoczne są także prawdopodobieństwa *a priori* (wagi) dla poszczególnych klas.



Rys. 2. Wyniki segmentacji respondentów

W klasie pierwszej, najmniej licznej ($\tau_1 = 0{,}25$), 28% respondentów twierdzi, że kobiety nie nadają się do polityki. Bardzo mały procent (0,07%) w tej klasie stanowią osoby zgadzające się z tym, że rządzenie krajem należy pozostawić mężczyznom. Prawie 50% zgadza się z opinią, że pracująca matka może zapewnić ciepło. 16% twierdzi, że żona jest odpowiedzialna za karierę męża. Największy odsetek w tej grupie (87%) stanowią respondenci przekonani, że praca matki szkodzi dziecku. Niewiele mniej (79%) respondentów uważa, że lepiej, gdy zarabia mężczyzna.

Klasa druga jest klasą liczniejszą – należy do niej 37% wszystkich ankietowanych. W klasie tej 31% respondentów uważa, że kobiety nie nadają się do polityki, a 7% zgodziło się z opinią, że rządzenie krajem należy pozostawić mężczyznom. W klasie drugiej jest największy (w porównaniu z klasą pierwszą i trzecią) udział osób (93%), które sądzą, że pracująca matka może zapewnić ciepło. Tylko 9% ankietowanych uważa, że żona powinna zapewnić karierę mężowi. Taki sam procent stanowią osoby, które twierdzą, że praca matki szkodzi dziecku. W klasie tej jest najmniej osób (w porównaniu do klasy pierwszej i trzeciej), tj. 22%, które sądzą, iż lepiej jest, gdy o utrzymanie rodziny troszczy się mężczyzna.

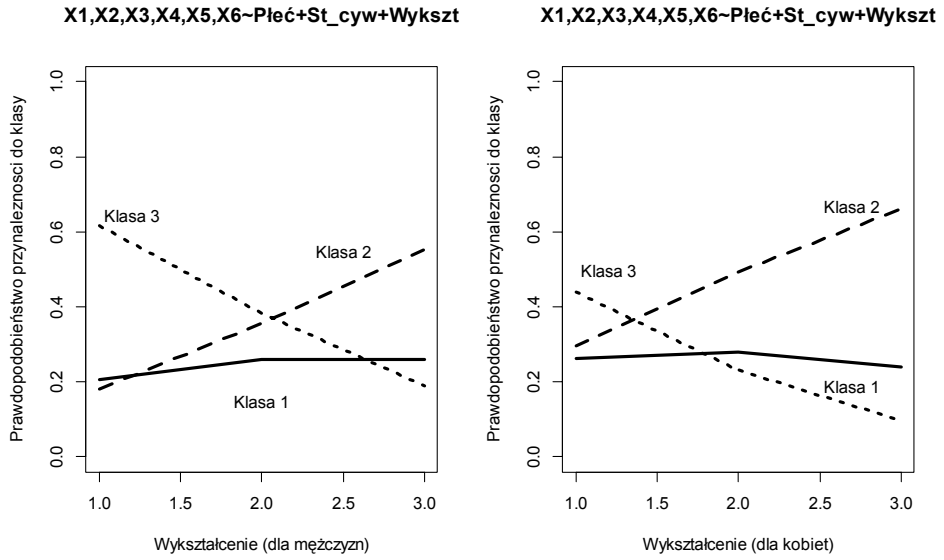
Klasa trzecia jest tak samo liczna, jak klasa druga (τ_3 = 0,37). Ponad 90% osób zgadza się z opinią, że kobiety nie nadają się do polityki. Nieco mniej (87%) uważa, że rządzenie krajem należy pozostawić mężczyznom. Ponad połowa ankietowanych tej klasy jest zdania, że pracująca kobieta może zapewnić rodzinne ciepło, ale na pytanie: „Czy praca matki szkodzi dziecku?” aż 81%

odpowiedziało twierdząco. W klasie tej aż 95% osób uważa, że lepszym rozwiązaniem dla rodziny jest tylko zawodowa praca mężczyzny.

W kolejnej części pracy dokonano analizy wpływu zmiennych towarzyszących na przynależność analizowanych obiektów do klas. Jeżeli chodzi o zmienną „płeć", okazuje się, że dla mężczyzn występuje najwyższe prawdopodobieństwo przynależności do klasy trzeciej, a najniższe w przypadku klasy drugiej. Z kolei udział kobiet w klasie drugiej jest najwyższy i wynosi prawie 50%, kolejno w klasie pierwszej oraz trzeciej.

Dokonując analizy wpływu zmiennej towarzyszącej „stan cywilny" (dla mężczyzn z średnim wykształceniem), prawdopodobieństwo przynależności do klasy pierwszej jest prawie takie samo dla osób o różnym stanie cywilnym. W klasie drugiej największe prawdopodobieństwo występuje w przypadku kawalerów, następnie panów żyjących w konkubinacie oraz żonatych (najniższe dla wdowców). Prawdopodobieństwo przynależności do klasy trzeciej („konserwatywnej") jest najwyższe dla wdowców, następnie osób żyjących w separacji i rozwiedzionych.

Jeżeli chodzi o zmienną towarzyszącą „wykształcenie", to dla mężczyzn, żonatych o wykształceniu zawodowym, najwyższe jest prawdopodobieństwo przynależności do klasy trzeciej. Prawdopodobieństwo przynależności do tej klasy spada wraz z lepszym wykształceniem respondentów. Z kolei prawdopodobieństwo przynależności do klasy drugiej wzrasta wraz z lepszym wykształceniem. Jeśli chodzi o klasę pierwszą, to prawdopodobieństwo przynależności do tej klasy jest prawie takie samo dla osób o różnym poziomie wykształcenia. Wpływ wykształcenia na przynależność do klas dla kobiet jest bardzo podobny (rys. 3). Ze względu na ograniczenia objętościowe na rys. 3 zamieszczono tylko wykres dla zmiennej towarzyszącej Z_3 (wykształcenie).



Rys. 3. Wykres przynależności kobiet (strona lewa) i mężczyzn (strona prawa) do trzech klas

Dla zmiennej towarzyszącej „wykształcenie" sporządzono oddzielne wykresy dla kobiet i mężczyzn, przyjmując, że zmienne jakościowe są równe kategorii występującej najczęściej (stan cywilny – zamężna/żonaty). W podobny sposób sporządzono wykresy i dokonano interpretacji dla zmiennej towarzyszącej „płeć" i „stan cywilny" (zob. np. Linzer i Lewis, 2011; Witek, 2011).

Podsumowanie

W artykule przedstawiono przykład zastosowania modeli klas ukrytych do oceny roli kobiet w polskim społeczeństwie. Analiza klas ukrytych umożliwiła segmentację respondentów na podstawie odpowiedzi udzielonych w badaniu Polskiego Generalnego Sondażu Społecznego. Wyodrębniono trzy klasy o podobnych wzorcach zachowań i postaw dla polskich respondentów. Dokonano również oceny wpływu zmiennych demograficznych na ich przynależność do klas.

Do klasy pierwszej zaliczono najmniej osób przeciwnych temu, by kobiety zajmowały się polityką (zarówno jeśli chodzi pełnienie różnych funkcji politycznych, jak i rządzenie krajem). W przypadku pracy zawodowej panuje tu raczej przekonanie, by kobieta została w domu. Respondenci klasy drugiej są przekonani, że kobiety jak najbardziej powinny realizować się zawodowo, a rodzina na tym nie ucierpi. Nie mają również przeciwwskazań, by kobiety pełniły funkcje polityczne. Klasa trzecia jest klasą osób „konserwatywnych", będących zdania, że kobieta po prostu powinna przebywać w domu (ani nie pracować, ani nie angażować się w życie polityczne naszego kraju).

Bibliografia

Agresti A. (2002): *Categorical Data Analysis*. John Wiley & Sons, Hoboken.

Bandeen-Roche K., Miglioretti D.L., Zeger S.L., Rathouz P.J. (1997): *Latent Variable Regression for Multiple Discrete Outcomes*. „Journal of the American Statistical Association", No. 92(40), s. 123-135.

Bąk A. (2011), *Modele klas ukrytych dla danych jakościowych*. W: *Analiza danych jakościowych i symbolicznych z wykorzystaniem programu R*. Red. E. Gatnar, M. Walesiak. C.H. Beck, Warszawa, s. 204-222.

Biernacki C., Celeux G., Govaert G. (1999): *Choosing Models in Model-Based Clustering and Discriminant Analysis*. „Journal of Statistical Computation and Simulation", No. 64, s. 49-71.

Bolck A., Croon M., Hagenaars J. (2004): *Estimating Latent Structure Models with Categorical Variables: One-step Versus Three-step Estimators*. „Political Analysis", No. 12(1), s. 3-27.

Bozdogan H. (2000): *Akaike's Information Criterion and Recent Developments in Information Criterion*. „Journal of Mathematical Psychology", No. 44, s. 62-91.

Collins L.M., Lanza S.T. (2011): *Latent Class and Latent Transition Analysis with Applications in the Social, Behavioral, and Health Sciences*. John Wiley & Sons, Wiley, s. 100-103; 151, 177.

Dayton C. M., Macready G.B. (1988): *Concomitant-variable Latent-class Models*. „Journal of the American Statistical Association", No. 83(401), s. 173-178.

Dempster A.P., Laird N.P., Rubin D.B. (1977): *Maximum Likelihood for Incomplete Data Via the EM Algorithm (with discussion)*. „Journal of the Royal Statistical Society", No. 39, ser.B, s. 1-38.

Domański C., Pruska K. (2000): *Nieklasyczne metody statystyczne*. PWE, Warszawa.

Goodman L. (1970): *The Multivariate Analysis of Qualitative Data: Interactions Among Multiple Classification*. „Journal of the American Statistical Association", No. 65, s. 226-256.

Haberman S.J. (1979): *Analysis of Qualitative Data, New Developments*. Academ ic Press, New York, No 2.

Hagenaars A.J., McCutcheon A.L. (2002): *Applied Latent Class Analysis*. Cambridge University Press, Cambridge.

Lazarsfeld P.F. (1950): *The Logical* and Mathematical Foundations of Latent Structure Analysis. W: *Measurement and Prediction*. Red. S.A. Stouffer. John Wiley & Sons, New York, s. 362-412.

Linzer D., Lewis J. (2011): *poLCA: An R Package for Polytomous Variable Latent Class Analysis*. „Journal of Statistical Software", No. 42(10), s. 1-29.

McLachlan G.J., Peel D. (2000): *Finite Mixture Models*. Wiley, New York, s. 81-116.

Vermunt, J.K. (2010): *Latent Class Modeling With Covariates: Two Improved Three-step Approaches*. Political Analysis, 18, s. 450-469.

Witek E. (2009): *Analiza skupień – podejście modelowe*. W: *Statystyczna analiza danych z wykorzystaniem programu R*. Red. M. Walesiak, E. Gatnar. Wydawnictwo Naukowe PWN, Warszawa, s. 434-462.

Witek E. (2011): *Modele mieszanek dla danych jakościowych*. W: *Analiza danych jakościowych i symbolicznych z wykorzystaniem programu R*. Red. E. Gatnar, M. Walesiak. C.H. Beck, Warszawa, s. 223-241.

A ROLE OF WOMEN IN POLISH SOCIETY – AN EMPIRICAL ANALYSIS WITH THE USE OF LATENT CLASS MODELS

Summary

The paper focuses on latent class models and it's application for quantitative data. Latent class modeling is one of a multivariate analysis techniques of the contingency table and can be viewed as a special case of model-based clustering, for multivariate discrete data. It is assumed that each observation comes from one of a number of subpopulations, with its own probability distribution.

We used latent class analysis for grouping and detecting inhomogeneities of Polish opinions on role of women in polish society. We analyzed data collected as part of the Polish General Social Survey (GSS) using poLCA package of R.

Krzysztof Jajuga
Uniwersytet Ekonomiczny we Wrocławiu

RYZYKO MODELU A MIARY RYZYKA

1. Ryzyko modelu i miary ryzyka – wprowadzenie

Nie ulega wątpliwości, iż modele matematyczne są często przydatne w analizie zjawisk ekonomicznych, a w konsekwencji we wspomaganiu decydentów. Wiadomo jednak, iż model jest przybliżeniem rzeczywistości, zaś jego praktyczna użyteczność zależy od tego, jak dobre jest to przybliżenie. Kluczowe dla użytkownika nie jest to, na ile model jest dobrym przybliżeniem obecnej (i przeszłej) rzeczywistości, lecz jak dobrym będzie przybliżeniem przyszłej rzeczywistości. To jednak może być zweryfikowane dopiero *ex post*. Pojawia się zatem pytanie, czy w momencie tworzenia modelu nie będzie jakiegokolwiek narzędzia weryfikacji modelu z punktu widzenia jego użyteczności w przyszłości.

Odpowiedź na to pytanie jest negatywna; pomocna tu może być analiza tzw. ryzyka modelu. Ogólna definicja tego pojęcia jest następująca: Ryzyko modelu jest to ryzyko wynikające z zastosowania błędnego modelu w świecie rzeczywistym.

Problematyka ryzyka modelu jest szczególnie widoczna w przypadku modeli dotyczących zjawisk i procesów finansowych, przede wszystkim tych zachodzących na rynkach finansowych. Znaczenie ryzyka modelu jako zjawiska, które może doprowadzić do poważnych problemów finansowych, zostało wyraźnie zidentyfikowane po raz pierwszy przy okazji upadku funduszu hedgingowego Long Term Capital Management (LTCM). Fundusz ten w latach 1994-1998 stosował bardzo zaawansowane modele wyceny instrumentów finansowych. Na początku decyzje podejmowane na podstawie tych modeli prowadziły do ponadprzeciętnych wyników inwestycyjnych funduszu, jednak w 1997 r. w trakcie kryzysu finansowego w Azji Południowo-Wschodniej fundusz poniósł straty, a jego upadek nastąpił w trakcie kryzysu finansowego w Rosji w 1998 r. Upadek LTCM pokazał, iż w warunkach dynamicznie zmieniających się rynków finansowych, precyzyjnie „skrojone" modele przestają działać skutecznie.

Inny przykład ryzyka modelu finansowego to ryzyko sprawozdania finansowego sporządzanego przez podmiot gospodarczy. Informacje zawarte w jego elementach składowych służą jako podstawa podejmowania decyzji, jednak

sprawozdanie finansowe jest uproszczonym modelem podmiotu gospodarczego. W ostatnich latach uproszczenie to jest coraz większe. Sprawozdawczość finansowa nie może sobie poradzić z takimi istotnymi współczesnymi zjawiskami gospodarczymi, jak: stosowanie instrumentów pochodnych, występowanie powiązań biznesowych czy też pomiar kapitału intelektualnego. Przykład sprawozdania finansowego holdingu Enron, który upadł w 2001 roku, to tylko jeden z wielu przypadków.

Celem tego artykułu jest wskazanie poprzez analizę dwóch przypadków, w jaki sposób ryzyko modelu objawia się w modelach finansowych, w których występują miary ryzyka.

Można wyróżnić trzy podstawowe rodzaje ryzyka modeli finansowych:
- ryzyko w zakresie struktury modelu,
- ryzyko estymacji modelu,
- ryzyko zastosowania modelu.

Zostaną przedstawione te trzy rodzaje ryzyka oraz działania, jakie tworzący model oraz jego użytkownik powinni podjąć w celu zmniejszenia ryzyka modelu.

1.1. Ryzyko w zakresie struktury modelu

Takie ryzyko dotyczy konstrukcji modelu, np. wadliwej postaci funkcyjnej modelu, nieuwzględnienia istotnych zmiennych oraz dynamiki w modelu. Przykładem jest model stóp zwrotu akcji, w którym założono rozkład normalny, co jest zbyt dużym uproszczeniem w przypadku istnienia grubych ogonów.

Analiza ryzyka w zakresie struktury modelu jest analizą jakościową. Polega ona na weryfikacji założeń modelu i jego postaci. Czasem może też być przydatna weryfikacja empiryczna modelu na danych historycznych – stosowana jest tu często nazwa testowanie wsteczne. Można też przedstawić kontrargument, iż funkcjonowanie modelu w przeszłości niekoniecznie może być dobrym wskaźnikiem co do tego, jak będzie działał w przyszłości. Jeśli jednak okaże się, że model źle funkcjonuje na danych historycznych, może to oznaczać, że nie uwzględnia on pewnych istotnych zmiennych – a to jest właśnie problem ryzyka w zakresie struktury modelu.

1.2. Ryzyko w zakresie ocen parametrów modelu

Dotyczy to estymacji modelu, która może być przeprowadzona za pomocą różnych technik i z zastosowaniem różnych zbiorów danych. Przykładem jest estymacja współczynnika beta za pomocą klasycznej metody najmniejszych kwadratów, w której nie bierze się pod uwagę możliwości występowania obserwacji nietypowych, jak również faktu niestacjonarności zmiennych danych w postaci szeregów czasowych.

Analiza ryzyka estymacji modelu może być przeprowadzona za pomocą narzędzi ilościowych. Można tu zaproponować proste podejście, polegające na analizie wrażliwości modelu (ściślej: zmiennej wyjściowej modelu) na zmiany wartości jego parametrów. Dzięki temu jest możliwa identyfikacja tych parametrów modelu, które mają duży wpływ na wyniki. Pomocne może też być wykorzystanie przedziałów ufności dla parametrów, tam gdzie są one znane.

1.3. Ryzyko zastosowania modelu w specyficznej sytuacji

Dotyczy to sytuacji, gdy prawidłowy (w sensie struktury i oszacowania) model jest stosowany w otoczeniu, w którym warunki są inne niż te, w jakich model się sprawdził. Przykładem jest zastosowanie modelu wyceny opcji na wschodzącym rynku, charakteryzującym się niską płynnością oraz niewielkim doświadczeniem uczestników rynku.

Analiza ryzyka zastosowania modelu jest zazwyczaj analizą jakościową, polegającą na sprawdzeniu założeń i warunków stosowania modelu oraz odniesieniu ich do rynku, na którym chcemy go stosować.

W bardzo wielu modelach finansowych występują miary ryzyka. Jeśli chodzi o miary ryzyka rynkowego, to najbardziej ogólna klasyfikacja wyróżnia:
- miary oparte na rozkładzie stóp zwrotu;
- miary wrażliwości.

Główna różnica między tymi dwiema grupami polega na tym, że w pierwszej analizowany jest „efekt ryzyka" (w postaci rozkładu statystycznego), zaś w drugiej analizowane są też czynniki ryzyka (poprzez analizę wrażliwości zmiennej wyjściowej na te czynniki).

W pierwszej grupie miar znajdują się:
- miary zmienności rozkładu,
- kwantyle rozkładu,
- wartości dystrybuanty rozkładu.

Zostaną przedstawione dwa podstawowe przypadki, w których występuje ryzyko modelu wynikające z faktu, że miara ryzyka podlega oszacowaniu. Rozpatrywana jest podstawowa miara ryzyka, mianowicie odchylenie standardowe stopy zwrotu. W modelach finansowych odchylenie standardowe zwykle nazywane jest po prostu zmiennością (*volatility*).

2. Model portfela dwuskładnikowego o najniższym ryzyku

Teoria portfela (por. Markowitz, 1952) jest to historycznie pierwszy model finansowy, w którym pojawiła się miara zmienności (wariancja stopy zwrotu). Jak wiadomo, ryzyko portfela (mierzone odchyleniem standardowym stopy

zwrotu) zależy od ryzyka składników tego portfela oraz od współczynników korelacji stóp zwrotu tych składników. Błędne oszacowanie tych elementów może prowadzić do niedoszacowania (bądź przeszacowania) ryzyka tego portfela.

W dalszej części będzie rozważane zagadnienie wyznaczania portfela o minimalnym ryzyku (*Minimum Variance Portfolio*), przy czym dla uproszczenia rozpatrywany jest jedynie portfel dwuskładnikowy.

Udziały składników w dwuskładnikowym portfelu o minimalnym ryzyku dane są następującymi wzorami:

$$w_1 = \frac{s_2^{\,2} - s_1 s_2 \rho_{12}}{s_1^{\,2} + s_2^{\,2} - 2 s_1 s_2 \rho_{12}} \tag{1}$$

$$w_2 = \frac{s_1^{\,2} - s_1 s_2 \rho_{12}}{s_1^{\,2} + s_2^{\,2} - 2 s_1 s_2 \rho_{12}} \tag{2}$$

W dalszych rozważaniach nie jest analizowane ryzyko modelu wynikające z oszacowania współczynnika korelacji. Problemowi temu poświęcony jest artykuł Jajugi (2010). Następnie będzie rozpatrywane jedynie ryzyko wynikające z estymacji odchylenia standardowego stóp zwrotu obu składników portfela.

W celu analizy tego ryzyka wykorzystano wzór na przedział ufności odchylenia standardowego dla dużej próby:

$$P\left(\frac{s}{1+u_\alpha/\sqrt{2n}} < \sigma < \frac{s}{1-u_\alpha/\sqrt{2n}}\right) = 1-\alpha \tag{3}$$

$$P(-u_\alpha < U < u_\alpha) = 1-\alpha \tag{4}$$

gdzie U jest zmienną o standaryzowanym rozkładzie normalnym.

Należy pamiętać, że podstawowe uproszczenie w tym podejściu wynika z założenia, iż rozpatrywana populacja ma rozkład normalny.

Można przyjąć, że dolny i górny kraniec przedziału ufności informują o niedoszacowaniu bądź przeszacowaniu odchylenia standardowego. Wskazują zatem na ryzyko modelu portfela dwuskładnikowego wynikające z estymacji odchylenia standardowego.

W celu zilustrowania tego zagadnienia rozważany jest przykład wskazujący na wpływ estymacji odchylenia standardowego na ryzyko portfela o minimalnym ryzyku (MVP). Ryzyko tego portfela wyznacza się za pomocą klasycznego wzoru na ryzyko portfela dwuskładnikowego z zastosowaniem udziałów danych wzorami (1) i (2).

Przyjmijmy, iż liczba obserwacji, na podstawie których oszacowano odchylenie standardowe, wynosi 50, zaś poziom ufności to 0,95. Po zastosowaniu wzorów (3) i (4) otrzymuje się przedział ufności dany wzorem:

$$0{,}836s < \sigma < 1{,}244s \tag{5}$$

W przykładzie rozpatrywane są dwa przypadki, jeśli chodzi o wartość odchylenia standardowego składników portfela:
- taka sama wartość odchylenia standardowego obu składników, równa 20%;
- istotnie różne wartości odchylenia standardowego, wynoszące odpowiednio 10% i 50%.

Analizowane są też dwa przypadki, jeśli chodzi o wartość współczynnika korelacji stóp zwrotu, mianowicie 0,1 oraz 0,5.

Tabela 1 przedstawia wartości odchylenia standardowego portfela MVP w pięciu scenariuszach:
- prawidłowo oszacowane oba odchylenia standardowe (symbol DD);
- niedoszacowane oba odchylenia standardowe (symbol NN);
- niedoszacowane pierwsze, a przeszacowane drugie odchylenie standardowe (symbol NP);
- przeszacowane pierwsze, a niedoszacowane drugie odchylenie standardowe (symbol PN);
- przeszacowane oba odchylenia standardowe (symbol PP).

Tabela 1

Ryzyko portfela MVP według różnych scenariuszy

Odchylenie standardowe pierwszego składnika	Odchylenie standardowe drugiego składnika	Współczynnik korelacji stóp zwrotu	Ryzyko portfela DD	Ryzyko portfela NN	Ryzyko portfela NP	Ryzyko portfela PN	Ryzyko portfela PP
20%	20%	0,1	14,83%	12,40%	14,50%	14,50%	18,45%
20%	20%	0,5	17,32%	14,48%	16,40%	16,40%	21,55%
10%	50%	0,1	9,95%	8,32%	8,36%	12,20%	12,38%
10%	50%	0,5	9,45%	7,90%	7,70%	12,11%	11,75%

Źródło: Obliczenia własne.

Z tab. 1 wynika, iż:
- niedoszacowanie (przeszacowanie) miar ryzyka powoduje niedoszacowanie (przeszacowanie) ryzyka portfela, co jest naturalne;
- wielkość błędu oszacowania ma większe znaczenie w przypadku przeszacowania niż niedoszacowania;
- współczynnik korelacji stóp zwrotu nie ma większego znaczenia dla niedoszacowania bądź przeszacowania ryzyka portfela.

3. Model wyceny opcji

Klasycznym modelem wyceny opcji europejskich jest model Blacka-Scholesa-Mertona (por. Black, Scholes, 1973; Merton, 1973). Wartość opcji ma istotne znaczenie dla praktyka, który podejmuje decyzje dotyczące zabezpieczenia przed ryzykiem bądź decyzje inwestycyjne z zastosowaniem tego instrumentu pochodnego.

Rozpatrywany jest tu najprostszy przypadek, w którym instrument podstawowy nie przynosi żadnych dochodów przed wygaśnięciem opcji. Występuje tu zatem klasyczny model wyceny Blacka-Scholesa, w którym wartość opcji zależy od pięciu czynników. Spośród nich cztery są znane: wartość instrumentu podstawowego, cena wykonania, czas do wygaśnięcia opcji oraz stopa wolna od ryzyka (za którą często przyjmuje się stopę z rynku międzybankowego). Oznacza to, że te cztery czynniki nie wpływają na ryzyko modelu wyceny opcji.

Inaczej jest w przypadku piątego czynnika, którym jest zmienność wartości instrumentu podstawowego, mierzona jako odchylenie standardowe stopy zwrotu tego instrumentu. Praktycy działający na rynku opcji zazwyczaj sięgają po jeden z dwóch sposobów estymacji zmienności.

Pierwszym jest zastosowanie tzw. zmienności historycznej, czyli estymacja odchylenia standardowego (bądź wariancji) za pomocą modeli ekonometrycznych i statystycznych, przy zastosowaniu danych historycznych dotyczących wartości instrumentu podstawowego. Główną wadą tego sposobu jest to, że zmienność w najbliższej przyszłości może się różnić od zmienności przeszłej.

Drugim sposobem jest zastosowanie tzw. zmienności implikowanej, która polega na „odwróceniu" modelu Blacka-Scholesa i wyznaczeniu zmienności na podstawie obecnych cen opcji. Zasadniczą wadą tego sposobu jest przyjęcie założenia, że rynek jest w stanie równowagi określanym modelem Blacka-Scholesa, a obecnie obserwowana zmienność jest dobrym oszacowaniem oczekiwań co do przyszłej zmienności.

Warto zatem zadać pytanie, jakie jest ryzyko modelu wyceny opcji wynikające z błędnego oszacowania zmienności, tzn. z tego, że zmienność zrealizowana w okresie do wygaśnięcia opcji będzie różna od oszacowanej zmienności.

Odpowiedź daje analiza wrażliwości wartości opcji na zmiany zmienności. Do tego celu można wykorzystać jeden z tzw. greckich współczynników, którym jest współczynnik *vega*. Standardowo wykorzystywany jest on jako miara ryzyka rynkowego opcji, natomiast tutaj posłuży on do oceny ryzyka modelu wyceny opcji.

Współczynnik *vega* (w klasycznym przypadku modelu Blacka-Scholesa) dany jest następującym wzorem (por. Hull, 2011):

$$vega = S\phi(d_1)\sqrt{T} \tag{6}$$

gdzie:
S – wartość instrumentu podstawowego,
T – czas do wygaśnięcia opcji,
ϕ – gęstość standaryzowanego rozkładu normalnego,
d_1 – argument funkcji, określony wzorem (por. Hull, 2011):

$$d_1 = \frac{\ln(\frac{S}{X}) + (r + \frac{\sigma^2}{2})T}{\sigma\sqrt{T}} \tag{7}$$

gdzie:
X – cena wykonania (tzw. *strike*),
r – stopa wolna od ryzyka,
σ – odchylenie standardowe stopy zwrotu instrumentu podstawowego.

Współczynnik *vega* wskazuje na zmiany wartości opcji przy jednoprocentowych zmianach zmienności. Standardowo wykorzystywany jest w różnych strategiach opcyjnych, m.in. prowadzących do uzyskania portfela opcji odpornego na zmiany zmienności instrumentu podstawowego.

Z punktu widzenia ryzyka modelu wyceny opcji istotne są dwie właściwości tego współczynnika (*ceteris paribus*):
- wartość *vega* jest najwyższa, gdy wartość instrumentu podstawowego jest równa cenie wykonania (tzw. opcja *at-the-money* – ATM);
- wartość *vega* jest tym wyższa, im dłuższy jest czas pozostały do wygaśnięcia.

W charakterze ilustracji przedstawiony jest przykład określenia współczynnika wrażliwości *vega* dla różnych okresów do wygaśnięcia opcji (7, 31, 91 i 182 dni) oraz różnego poziomu relacji między wartością instrumentu podstawowego, wynoszącą 100, a ceną wykonania (pięć różnych wartości: 80, 90, 100, 110, 120). W przykładzie tym stopa wolna od ryzyka wynosi 5% (dla wszystkich rozpatrywanych terminów).

Wartości współczynnika *vega* przedstawia tab. 2.

Tabela 2

Wartości współczynnika *vega* według różnych scenariuszy

Liczba dni do wygaśnięcia	Vega: Strike = 100	Vega: Strike = 90	Vega: Strike = 80	Vega: Strike = 110	Vega: Strike = 120
7	13,80	0	0	0	0
31	28,77	0,02	0	0,23	0
91	48,08	2,91	0	13,14	0,17
182	65,48	12,05	0,13	44,48	6,37

Źródło: Obliczenia własne.

Wyniki zawarte w tab. 2 potwierdzają teoretyczne właściwości. Wynika z nich, że problem ryzyka estymacji w klasycznych modelach wyceny opcji występuje w zasadzie jedynie w przypadku opcji ATM, tzn. takich, w których cena wykonania jest bliska wartości instrumentu podstawowego, ewentualnie w przypadku opcji o długim terminie do wygaśnięcia.

Podsumowanie

Przenoszenie danego modelu w czasie (zmiana funkcjonowania rynku) lub w przestrzeni (inny rynek) może prowadzić do jego niewłaściwego zastosowania. Jednym z podstawowych błędów twórców i użytkowników modeli jest dążenie do tego, aby model był precyzyjny oraz dobrze dopasowany do danych historycznych.

Modele stosowane na rynkach finansowych powinny spełniać dwa warunki brzegowe; być po pierwsze odporne na zmiany warunków rynkowych, a po drugie – przejrzyste dla użytkownika. Spełnienie pierwszego warunku może zmniejszyć ryzyko w zakresie obszaru zastosowań modelu. Z kolei spełnienie drugiego warunku oznacza zmniejszenie zagrożenia związanego z zastosowaniem dobrego modelu w niewłaściwy sposób z powodu niezrozumienia modelu przez użytkownika.

Dobrą praktyką powinno się stać uzupełnianie każdego modelu o informację na temat ryzyka tego modelu. Dotyczy to wszystkich modeli stosowanych w finansach – od modelu wyceny opcji po rachunek zysków i strat spółki.

Bibliografia

Black F., Scholes M. (1973): *The Pricing of Options and Corporate Liabilities*. „Journal of Political Economy", 81, s. 637-654.

Hull J. (2011): *Options, Futures and Other Derivatives*. Pearson, Upper Saddle River.

Jajuga K. (2010): *Assessment of Model Risk in Financial Markets*. „Optimum Studia Ekonomiczne", 48, s. 35-43.

Markowitz H.M. (1952): *Portfolio Selection*. „Journal of Finance", 7, s. 77-91.

Merton R.C. (1973): *Theory of Rational Option Pricing*. „Bell Journal of Economics and Management Science", 4, s. 141-183.

MODEL RISK AND RISK MEASURES

Summary

The paper discusses the problem of model risk, defined as risk resulting from the application of wrong model in real world. Three sources of model risk are distinguished: risk related to the structure of the model, risk of model estimation and risk connected with the application of the model.

The main part of the paper presents the measures that can be used to evaluate risk of model estimation. Two particular cases are solved. The first one is the construction of two stock portfolio with minimal risk, the second one is option pricing. In both cases estimation risk results from the fact that main parameter, which is volatility (standard deviation if returns), has to be estimated. Finally, the paper states important conditions to limit model risk.

Grzegorz Kończak
Uniwersytet Ekonomiczny w Katowicach

SPRAWDZANIE JEDNORODNOŚCI JAKOŚCI MATERIAŁÓW NIEKSZTAŁTNYCH Z WYKORZYSTANIEM ROZKŁADÓW WARTOŚCI EKSTREMALNYCH

Wprowadzenie

Klasyczne metody statystyczne bardzo często odwołują się do założenia normalnego rozkładu analizowanych charakterystyk oraz niezależności pomiarów. Tak szczególne znaczenie rozkładu normalnego jest związane m.in. z centralnym twierdzeniem granicznym. Przy stosowaniu metod statystycznych w kontroli jakości zwykle zakłada się dodatkowo, że przedstawiany do kontroli materiał jest jednorodny.

Teoria statystycznej kontroli jakości przedstawia wiele różnych metod sprawdzania jakości produktów. W zdecydowanej większości dotyczą one obiektów policzalnych, występujących w sztukach, opakowaniach itp. W artykule podjęto analizę zagadnienia badania jakości towarów materiałów niekształtnych.

Jako materiał niekształtny będzie określany materiał, z którego nie sposób bezpośrednio wyodrębnić poszczególne elementy, opakowania itp. Przykładem materiału niekształtnego może być np. zwał węgla kamiennego w punkcie sprzedaży lub ziarno w silosach. W takim rozumieniu materiał niekształtny może mieć postać regularną np. sześciennego bloku. W artykule przedstawiono propozycję metody pozwalającej na weryfikację hipotezy głoszącej, że sprawdzany materiał jest jednorodny ze względu na badaną charakterystykę.

1. Rozkłady wartości ekstremalnych

Rozkładem granicznym dla wielu spotykanych w praktyce rozkładów jest rozkład normalny. Tak jest np. z rozkładem średniej arytmetycznej z próby prostej, jeśli tylko wartość oczekiwana i wariancja badanej zmiennej są skończone. Wraz ze wzrostem liczebności próbki n rozkład średniej z próby prostej zmierza do rozkładu

normalnego. Często w badaniach kontroli jakości nie interesuje nas jednak przeciętny poziom jakości, lecz występowanie materiałów o szczególnie niskiej (lub wysokiej) wartości sprawdzanej charakterystyki.

1.1. Dokładny rozkład wartości maksymalnych

Niech $(X_1, X_2, \ldots, X_n)$ będzie *n*-elementową próbą losową pobraną niezależnie z populacji o dystrybuancie $F(x)$. Oznaczając przez X_M zmienną losową określoną następująco:

$$X_M = \max(X_1, X_2, \ldots, X_n) \tag{1}$$

dystrybuantę $G(x)$ tej zmiennej losowej można wyznaczyć następująco:

$$G(x) = P(X_M < x) = P(X_1 < x \wedge X_2 < x \wedge \ldots \wedge X_n < x) =$$
$$= P(X_1 < x) \cdot P(X_2 < x) \cdot \ldots \cdot P(X_n < x) = F^n(x)$$

W praktyce zwykle nie ma możliwości odwołania się do rozkładu dokładnego. Tak będzie np. w sytuacji, gdy nie jest znany rozkład analizowanej zmiennej lub jego złożona postać nie daje możliwości wyznaczenia rozkładu dokładnego na podstawie wzoru (1). W takich przypadkach niezbędne jest odwołanie się do rozkładu granicznego. Postać graniczna rozkładu wartości ekstremalnych (maksima lub minima) z *k* próbek o liczebności *n* może przyjmować wyłącznie jedną z trzech form: rozkładu Gumbela, rozkładu Frecheta lub rozkładu Weibulla (por. Castillo et al., 2005).

1.2. Rozkład Gumbela

Rozkład Gumbela występuje bardzo często w praktyce, a w szczególności podczas obserwacji np. wartości maksymalnych lub minimalnych różnych pomiarów. Funkcja gęstości zmiennej losowej X o rozkładzie Gumbela dla wartości minimalnych z parametrami λ i δ jest zadana wzorem:

$$f(x) = \frac{1}{\delta} e^{\frac{-(x-\lambda)}{\delta} - e^{\frac{-(x-\lambda)}{\delta}}} \quad \text{dla } x \in R \tag{2}$$

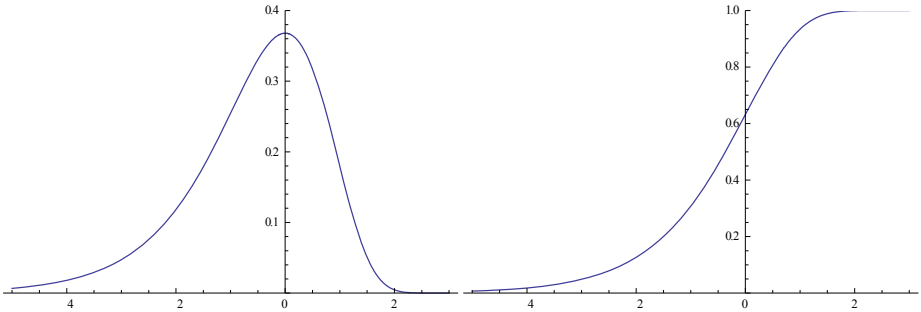
Parametry λ i δ to odpowiednio dominanta oraz parametr rozproszenia. Dystrybuanta tej zmiennej losowej wyraża się następującym wzorem:

$$F(x) = 1 - e^{-e^{\frac{-(x-\lambda)}{\delta}}} \quad \text{dla } x \in R \tag{3}$$

Wartość oczekiwana μ i wariancja σ^2 zmiennej losowej X są następujące:

$$\mu = \lambda - 0{,}57772\,\delta \text{ oraz } \sigma^2 = \frac{\pi^2 \delta^2}{6} \tag{4}$$

Funkcję gęstości i dystrybuantę zmiennej losowej o rozkładzie Gumbela dla wartości minimalnych z parametrami $\lambda = 0$ i $\delta = 1$ przedstawia rys. 1.



Rys. 1. Gęstość prawdopodobieństwa i dystrybuanta zmiennej losowej o rozkładzie Gumbela wartości minimalnych z parametrami $\lambda = 0$ i $\delta = 1$

Funkcja gęstości zmiennej losowej X o rozkładzie Gumbela dla wartości maksymalnych z parametrami λ i δ jest zadana wzorem:

$$f(x) = \frac{1}{\delta} e^{\frac{(x-\lambda)}{\delta} - e^{\frac{(x-\lambda)}{\delta}}} \quad \text{dla } x \in R \tag{5}$$

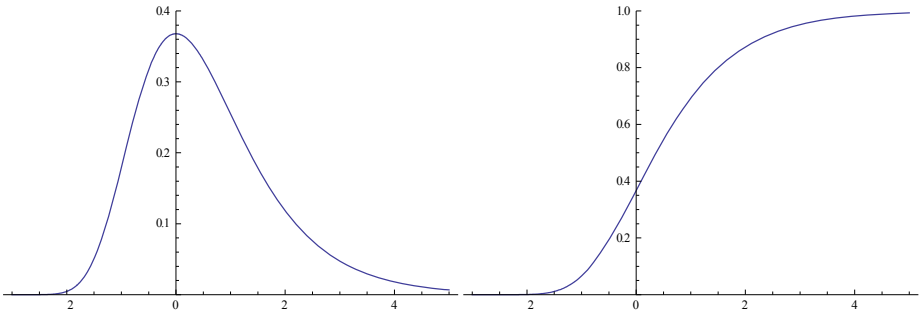
Dystrybuanta tej zmiennej losowej wyraża się następującym wzorem:

$$F(x) = e^{-e^{\frac{-(x-\lambda)}{\delta}}} \quad \text{dla } x \in R \tag{6}$$

Wartość oczekiwana i wariancja zmiennej losowej X są następujące:

$$\mu = \lambda + 0{,}57772\delta \text{ oraz } \sigma^2 = \frac{\pi^2 \delta^2}{6} \tag{7}$$

Funkcję gęstości i dystrybuantę zmiennej losowej o rozkładzie Gumbela dla wartości maksymalnych z parametrami $\lambda = 0$ i $\delta = 1$ przedstawia rys. 2.



Rys. 2. Gęstość prawdopodobieństwa i dystrybuanta zmiennej losowej o rozkładzie Gumbela wartości maksymalnych z parametrami $\lambda = 0$ i $\delta = 1$

1.3. Rozkład Weibulla

Zmienna losowa X ma rozkład Weibulla z parametrami λ, δ i β, jeśli jej gęstość prawdopodobieństwa wyraża się następującym wzorem:

$$f(x) = \frac{\beta}{\delta}\left(\frac{x-\lambda}{\delta}\right)^{1-\beta} e^{-\left(\frac{x-\lambda}{\delta}\right)^{\beta}} \text{ dla } x > \lambda \tag{8}$$

Dystrybuantę tej zmiennej losowej przedstawia wzór:

$$F(x) = 1 - e^{-\left(\frac{x-\lambda}{\delta}\right)^{\beta}} \text{ dla } x \geq \lambda \tag{9}$$

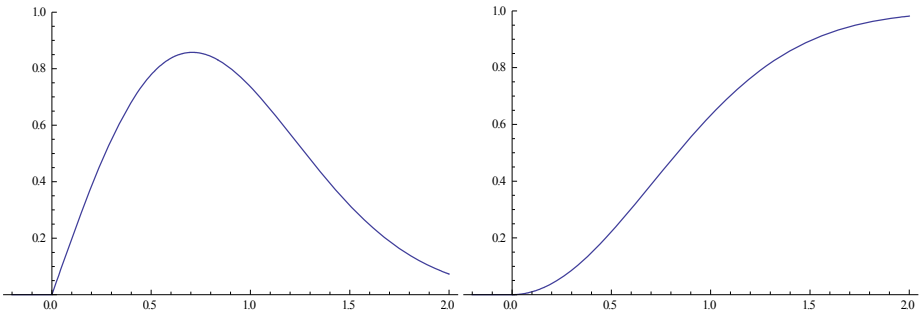
Wartość oczekiwana i wariancja zmiennej losowej X są następujące:

$$\mu = \lambda + \delta\Gamma\left(1+\frac{1}{\beta}\right) \text{ oraz } \sigma^2 = \delta^2\left[\Gamma\left(1+\frac{2}{\beta}\right) - \Gamma^2\left(1+\frac{2}{\beta}\right)\right] \tag{10}$$

gdzie Γ(x) jest funkcją Eulera, czyli

$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t}\, dt\,.$$

Funkcję gęstości i dystrybuantę zmiennej losowej o rozkładzie Weibulla zaprezentowano na rys. 3.



Rys. 3. Gęstość prawdopodobieństwa i dystrybuanta zmiennej losowej o rozkładzie Weibulla z parametrami $\lambda = 0$, $\delta = 1$ oraz $\beta = 2$

1.4. Rozkład Frecheta

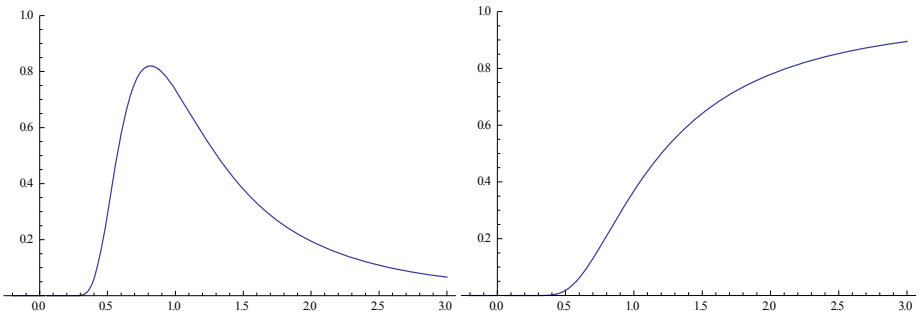
Zmienna losowa X ma rozkład Frecheta z parametrami λ, δ i β, jeśli jej gęstość prawdopodobieństwa jest zadana wzorem:

$$f(x) = \frac{\beta\delta}{(x-\lambda)^2}\left(\frac{\delta}{x-\lambda}\right)^{\beta-1} e^{-\left(\frac{\delta}{x-\lambda}\right)^{\beta}} \quad \text{dla } x > \lambda \tag{11}$$

Dystrybuanta tej zmiennej losowej wyraża się następującym wzorem:

$$F(x) = e^{-\left(\frac{\delta}{x-\lambda}\right)^{\beta}} \quad \text{dla } x > \lambda \tag{12}$$

Funkcję gęstości i dystrybuantę zmiennej losowej o rozkładzie Frecheta przedstawia rys. 4.



Rys. 4. Gęstość prawdopodobieństwa i dystrybuanta zmiennej losowej o rozkładzie Frecheta z parametrami $\lambda = 0$, $\delta = 1$ oraz $\beta = 2$

W zależności od typu rozkładu badanej zmiennej X graniczny rozkład maksimów lub minimów będzie zawsze rozkładem Gumbela, Weibulla lub Frecheta. Odpowiednie zależności dla wartości maksymalnych i minimalnych przedstawia tab. 1.

Tabela 1

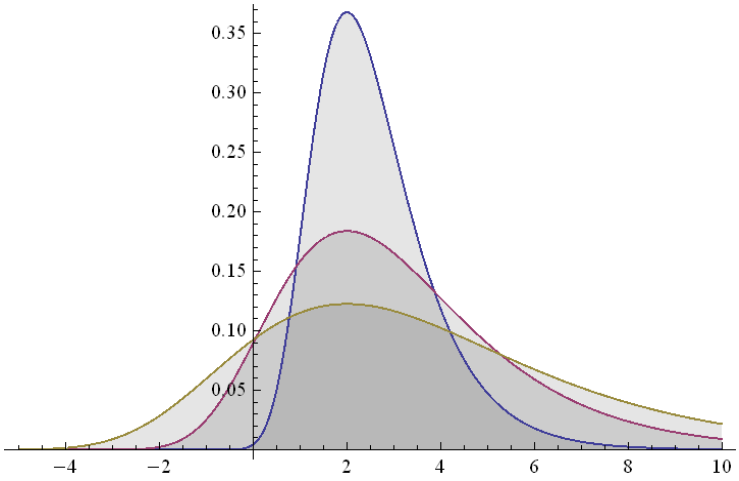
Typ rozkładu granicznego maksimów i minimów w zależności od rozkładu badanej zmiennej

Rozkład badanej zmiennej	Rozkład graniczny	
	Maksimów	Minimów
Normalny	Gumbela	Gumbela
Wykładniczy	Gumbela	Weibulla
Log-normalny	Gumbela	Gumbela
Gamma	Gumbela	Weibulla
Jednostajny	Weibulla	Weibulla
Cauchy'ego	Frecheta	Frecheta
Pareto	Frecheta	Weibulla

Źródło: Na podstawie (Castillo et al., 2005).

W dalszych rozważaniach będzie rozważany wyłącznie rozkład Gumbela dla wartości maksymalnych jako opisujący rozkład zmiennej losowej $X_M = \max_{i=1,2,\ldots,n} (X_1, X_2, \ldots X_n)$ gdzie $(X_1, X_2, \ldots X_n)$ jest próbą prostą pochodzącą z rozkładu, dla którego granicznym rozkładem wartości maksymalnych

jest rozkład Gumbela. Gęstości rozkładu Gumbela dla wartości maksymalnych dla różnych wartości parametrów przedstawia rys. 5.



Rys. 5. Rozkład Gumbela – gęstości w zależności od wartości parametrów rozkładu

2. Sprawdzenie jednorodności jakości materiałów niekształtnych

Niech do oceny jakości przedstawiany będzie pewien materiał niekształtny. Obserwowane są wartości zmiennej losowej X charakteryzującej jakość badanego materiału. W poniższych rozważaniach przyjęto, że badana zmienna ma rozkład, dla którego granicznym rozkładem wartości maksymalnych jest rozkład Gumbela (np. rozkład normalny, logarytmiczno-normalny, wykładniczy, gamma). Przyjęto, że występuje swobodny dostęp do dowolnej części badanego materiału niekształtnego. Celem jest ocena jednorodności materiału ze względu na wyróżnioną badaną charakterystykę. Wiele procedur statystycznej kontroli jakości wymaga, aby badany materiał charakteryzował się jednorodnością. Problemem jednorodności danych dla mieszanek zajmował się m.in. R.L. Schaeffer (1971). Często w analizach jednorodności danych wykorzystywane są metody repróbkowania, jak bootstrap (por. np. Boss i Brownie, 1989 oraz Kończak, 2006) czy testy permutacyjne (Good, 1994). W poniższych rozważaniach jednorodności danych wykorzystane zostaną własności rozkładów ekstremalnych.

Jeżeli materiał ze względu na badaną charakterystykę jest jednorodny, to wyznaczane wartości maksymalne z k próbek o jednakowych liczebnościach n, przy powyższych założeniach, będą obserwacjami z rozkładu Gumbela. Przebieg procesu sprawdzenia jednorodności badanego materiału jest następujący:

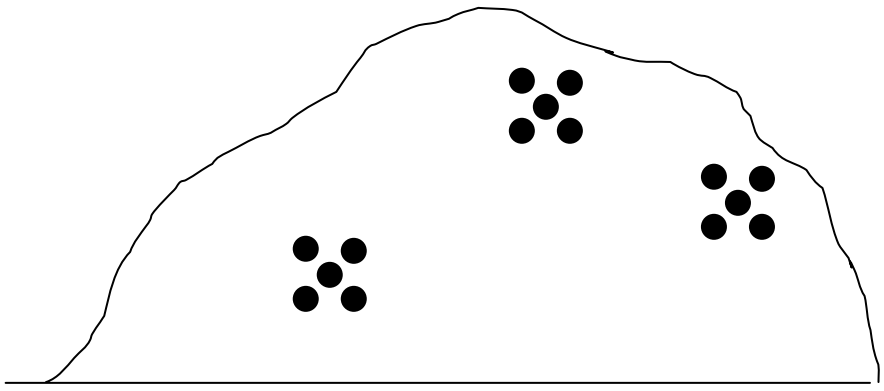
1. Wyznaczane jest losowo k punktów* (obszarów) w materiale niekształtnym.
2. Z każdego otoczenia wyróżnionych punktów pobierana jest próbka o liczebności n (pobieranie próbek o liczebności $n = 5$ z $k = 3$ obszarów schematycznie przedstawiono na rys. 6).
3. Dla każdej z próbek wyznaczana jest wartość badanej charakterystyki. Otrzymane dla i-tego ($i = 1, 2, \ldots, k$) obszaru wyniki są oznaczane przez $(x_{i1}, x_{i2}, \ldots, x_{in})$.
4. Dla każdej z próbek wyznaczana jest maksymalna zaobserwowana wartość $\hat{x}_i = \max\{x_{i1}, x_{i2}, \ldots x_{in}\}$ dla $i = 1, 2 \ldots, k$ badanej charakterystyki. Tym sposobem otrzymywany jest ciąg wartości maksymalnych $(\hat{x}_1, \hat{x}_2, \ldots \hat{x}_k)$.
5. Obliczana jest średnia oraz odchylenie standardowe z otrzymanych wartości maksymalnych

$$\bar{x} = \frac{1}{k}\sum_{i=1}^{k}\hat{x}_i \quad \text{oraz} \quad S(x) = \sqrt{\frac{1}{k-1}\sum_{i=1}^{k}(\hat{x}_i - \bar{x})^2}$$

6. Estymowane są parametry rozkładu Gumbela λ i δ. Oszacowanie jest przeprowadzane na podstawie wzoru (7), czyli

$$\delta \approx \frac{\sqrt{6}S}{\pi} \text{ oraz } \lambda \approx \bar{x} - 0{,}57772\delta \approx \bar{x} - 0{,}57772\frac{\sqrt{6}S}{\pi}$$

7. Testem Kołmogorowa-Smirnowa (Kanji, 2006) weryfikowana jest hipoteza o zgodności danych z próby (maksima) z rozkładem Gumbela dla wartości maksymalnych.



Rys. 6. Schemat pobierania próbek o liczebności $n = 5$ z $k = 3$ obszarów materiału niekształtnego

* W artykule poprzez losowy wybór punktu P rozumiany jest przypadkowy wybór, realizowany poprzez losowanie z rozkładu jednostajnego, współrzędnych (x_1,x_2, x_3) w przestrzeni trójwymiarowej.

3. Analiza symulacyjna

Funkcjonowanie przedstawionej procedury zostało poddane analizie symulacyjnej. W symulacjach uwzględniono dwa rodzaje rozkładu badanej zmiennej: normalny i logarytmiczno-normalny. W obu przypadkach rozkładem granicznym wartości ekstremalnych jest rozkład Gumbela wartości maksymalnych (por. tab. 1). W analizach symulacyjnych rozważano próbki pobierane z losowych obszarów sześciennego bloku o boku 1 z wymienionych rozkładów charakteryzujących się jednorodnością (oznaczenia – N_0 i LN_0) oraz niejednorodnością (oznaczenia – N_1, LN_1, N_2 i LN_2) o specyfikacji jak poniżej:

N_0 – rozkład normalny $N(10; 1)$

LN_0 – rozkład log-normalny $LN(0; 1)$

N_1 – rozkład normalny $N(10 + x_1 + x_2 + x_3, 1)$

LN_1 – rozkład log-normalny $LN(x_1 + x_2 + x_3, 1)$

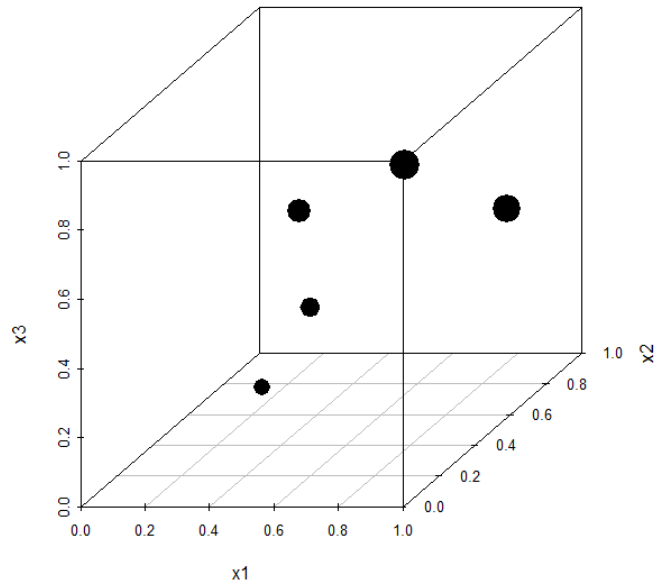
N_2 – rozkład normalny $N(10 + 2x_1 + 2x_2 + 2x_3, 1)$

LN_2 – rozkład log-normalny $LN(2x_1 + 2x_2 + 2x_3, 1)$

gdzie $x_1, x_2, x_3 \in (0,1)$ są współrzędnymi punktu badanego bloku sześciennego. Blok sześcienny został przedstawiony na rys. 7. Schematycznie zaprezentowano również niejednorodności badanej charakterystyki. Większe wartości zmiennej symbolicznie oznaczono na rys. 7 większymi kropkami.

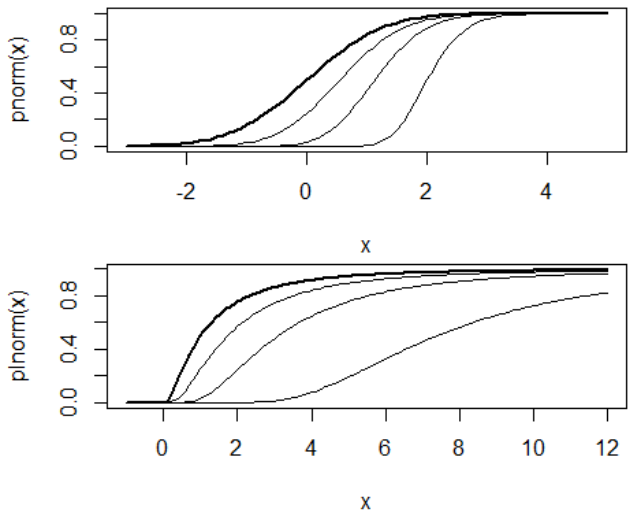
Wszystkie analizy symulacyjne wykonano w programie R (WWW1). We wszystkich przypadkach pobierano losowo k = 5, 10 oraz 30 próbek o liczebnościach n = 5.

We wszystkich rozważanych przypadkach w analizach symulacyjnych sprawdzano zgodność rozkładu próbkowego z dwoma postaciami rozkładów: granicznym rozkładem, czyli rozkładem Gumbela dla wartości maksymalnych, oraz rozkładem dokładnym. Dystrybuanty dokładnych rozkładów wartości maksymalnych dla próbek o liczebnościach n = 1, 2, 5 i 30 dla danych pochodzących z rozkładu normalnego oraz logarytmiczno-normalnego wyznaczono na podstawie wzoru (1) i przedstawiono na rys. 8.



Rys. 7. Schemat materiału, z którego symulacyjnie pobierano próbki z symbolicznie zaznaczoną niejednorodnością badanej charakterystyki (dot. N_1, LN_1, N_2 i LN_2)

Wyniki analizy symulacyjnej oceny prawdopodobieństw odrzucenia hipotezy głoszącej, że rozkład próbkowy jest zgodny z rozkładem granicznym Gumbela przedstawia tab. 2, a że jest zgodny z rozkładem dokładnym – tab. 3. Wyniki te zostały również zobrazowane na rys. 9.



Rys. 8. Dystrybuanty rozkładów dokładnych wartości maksymalnych dla próbek o liczebnościach n = 1, 2, 5 i 30 pochodzących z rozkładów normalnego (u góry) i logarytmiczno-normalnego (u dołu).

W przypadkach gdy badany materiał jest jednorodny (N_0 oraz LN_0), odwołanie się do dokładnego rozkładu wartości maksymalnych pozwala na uzyskanie właściwego rozmiaru testu (prawdopodobieństwa odrzucenia hipotezy o zgodności rozkładów są zbliżone do $\alpha = 0{,}05$). Właściwego rozmiaru testu nie zapewnia porównanie wyników z próby z rozkładem granicznym. W przypadku występowania niejednorodności materiału zarówno poprzez wykorzystanie rozkładu dokładnego, jak i granicznego postaci, czyli rozkładu Gumbela otrzymujemy wyniki podobne. W większości przypadków większą mocą charakteryzuje się test, gdzie wykorzystywany jest rozkład dokładny wartości maksymalnych.

Tabela 2

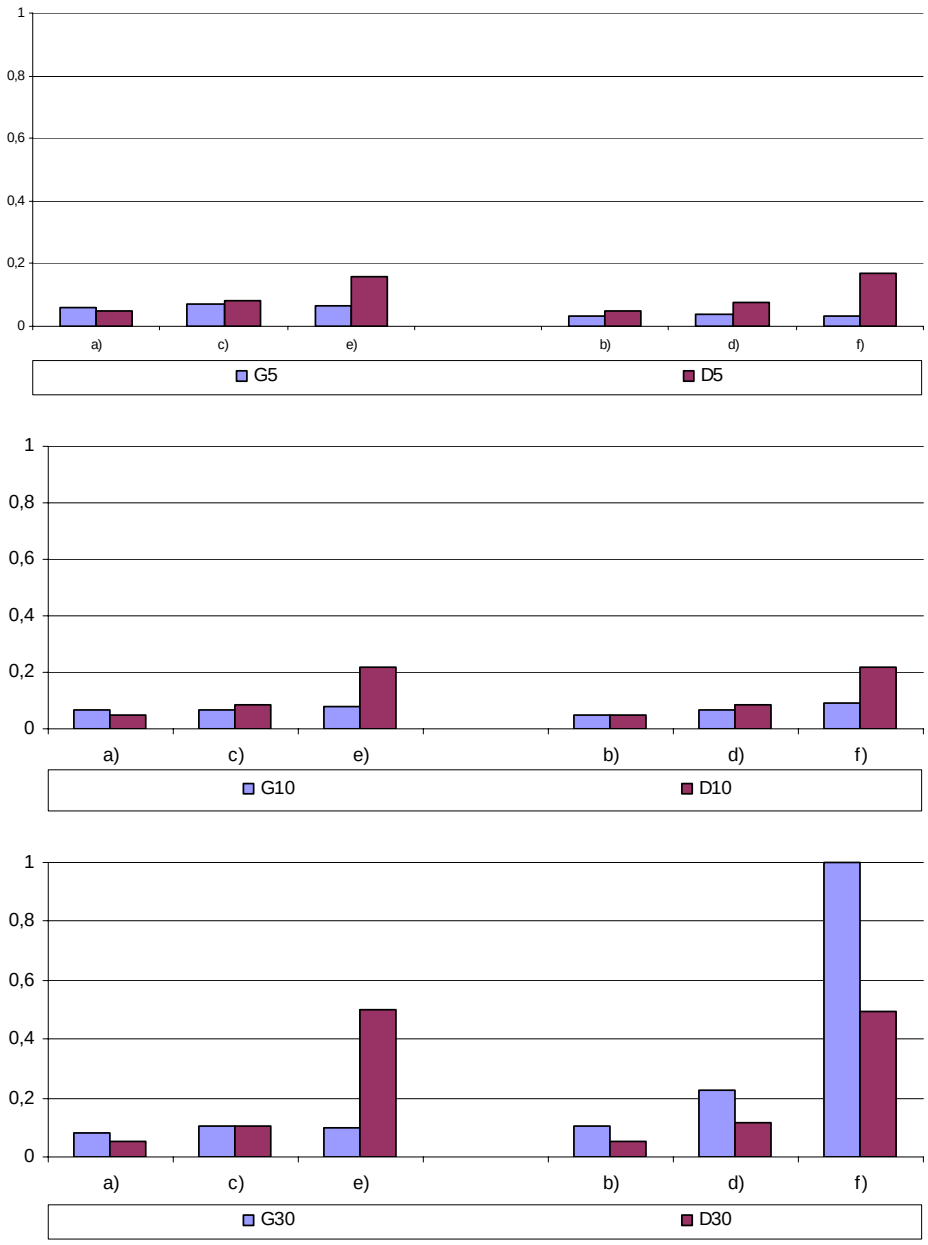
Oceny prawdopodobieństw (*p*) oraz błąd oceny (*sp*) odrzucenia hipotezy o zgodności badanego rozkładu z rozkładem Gumbela

Rozkład	$k = 5$		$k = 10$		$k = 30$	
	p	*sp*	*p*	*sp*	*P*	*sp*
N_0	0,058	0,008	0,068	0,008	0,079	0,007
LN_0	0,032	0,004	0,049	0,009	0,103	0,008
N_1	0,070	0,004	0,067	0,006	0,105	0,004
LN_1	0,037	0,007	0,065	0,009	0,226	0,010
N_2	0,066	0,007	0,079	0,006	0,101	0,011
LN_2	0,032	0,006	0,092	0,009	1,000	0,000

Tabela 3

Oceny prawdopodobieństw (*p*) oraz błąd oceny (*sp*) odrzucenia hipotezy o zgodności badanego rozkładu z rozkładem dokładnym wartości maksymalnych

Rozkład	$k = 5$		$k = 10$		$k = 30$	
	p	*sp*	*p*	*sp*	*p*	*sp*
N_0	0,048	0,006	0,051	0,007	0,050	0,006
LN_0	0,049	0,007	0,051	0,006	0,050	0,008
N_1	0,082	0,006	0,086	0,010	0,107	0,008
LN_1	0,077	0,011	0,085	0,012	0,116	0,013
N_2	0,157	0,018	0,218	0,013	0,501	0,009
LN_2	0,167	0,012	0,216	0,013	0,497	0,009



Rys. 9. Oceny prawdopodobieństw odrzucenia hipotezy o jednorodności badanego materiału dla prób o liczebnościach $n = 5$ (u góry), $n = 10$ (w środku) i $n = 30$ (na dole) dla rozkładu Gumbela (G) oraz rozkładu dokładnego (D)

Podsumowanie

W badaniach dotyczących jakości próbek przedstawianych do kontroli zwykle zakłada się jednorodność materiału ze względu na badane charakterystyki. Nie zawsze przyjęcie takiego założenia jest uzasadnione. Bez potwierdzenia jednorodności sprawdzanego materiału niezasadne jest przeprowadzanie różnych klasycznych analiz jakości. W artykule rozważano problem weryfikacji hipotezy głoszącej, że przedstawiany do kontroli materiał jest jednorodny. W tym celu weryfikowano hipotezę o zgodności rozkładu wartości maksymalnych z rozkładem dokładnym oraz z rozkładem Gumbela, który przy założonych rozkładach charakterystyk jest rozkładem granicznym wartości maksymalnych. Proponowana procedura ma na celu wykrycie istnienia obszarów o różnych wartościach badanej charakterystyki w materiałach niekształtnych. Nieco większe możliwości podjęcia prawidłowej decyzji zapewnia porównanie rozkładu próbkowego z rozkładem dokładnym, zwłaszcza dla prób o niewielkich liczebnościach. Jednak porównanie z rozkładem dokładnym wartości maksymalnych możliwe jest jedynie wówczas, gdy dysponujemy postacią funkcyjną (dystrybuantą) rozkładu badanej zmiennej. W przeciwnym przypadku jedynym wyjściem jest porównanie rozkładu próbkowego wartości maksymalnych z rozkładem granicznym (np. rozkładem Gumbela). Przeprowadzone analizy potwierdziły, że większą mocą charakteryzuje się test porównujący rozkład próbkowy z rozkładem dokładnym, ale wykorzystanie rozkładu Gumbela również przynosi dobre rezultaty.

Bibliografia

Boos D.D., Brownie C. (1989): *Bootstrap Methods for Testing Homogeneity of Variances*. „Technometrics", Vol. 31, No. 1.

Castillo E., Hadi A.S., Balakrishnan N., Sarabia J.M. (2005): *Extreme Value and Related Models with Applications in Engineering and Science*. John Wiley & Sons, New Jersey.

Efron B., Tibshirani R. (1993): *An Introduction to the Bootstrap*. Chapman & Hall. New York.

Good P.I. (1994): *Permutation Tests: A Practical Guide for Testing Hypotheses*. Springer-Verlag, New York.

Kanji G.K. (2006): *100 Statistical Tests*. Sage Publications, London.

Kończak G. (2006): *O teście równości wariancji wykorzystującym metodę bootstrap*. W: *Taksonomia 13. Klasyfikacja i analiza danych – teoria i zastosowania*. Red. K. Jajuga, M. Walesiak. Wydawnictwo Akademii Ekonomicznej, Wrocław, s. 543-552.

Scheaffer R.L. (1971): *A Test for the Homogeneity of a Mixture*. „Technometrics", Vol. 13, No. 2.

(WWW1): http://www.r-project.org.

THE USE OF EXTREME VALUE DISTRIBUTIONS IN CHECKING THE QUALITY SHAPELESS MATERIALS HOMOGENEITY

Summary

For quality control it is essential that the control samples are homogeneous. In practice this is impossible, and the requirement can be reduced to the condition that the samples were taken from the same population. The study presented in the paper is an analysis of the issue of testing the quality of the shapeless material. As a shapeless material is referred to the material from which it is impossible to directly extract the individual elements, packages, etc. This paper proposes a method to verify the hypothesis that the tested material is homogeneous due to the observed characteristics.

Walenty Ostasiewicz
Uniwersytet Ekonomiczny we Wrocławiu

UWAGI O PREHISTORII STATYSTYKI

Wprowadzenie

Historia statystyki jako nauki rozpoczyna sie w 1660 r. (zaokrąglona data początku działalności arytmetyków politycznych). Korzenie statystyki sięgają jednak odległych czasów starożytności. W tym artykule pobieżnie prześledzono tę długą drogę rozwoju statystyki. Rozpoczynając od spisów biblijnych Mojżesza i Dawida, poprzez reformy Solona i króla Tuliusza Serwiusza, został przedstawinony spis króla Wilhelma Zdobywcy i jego słynny kataster o nazwie *Domesday Book*. Artykuł kończy krótka prezentacja pisarstwa średniowiecznego.

To co w obecnych czasach określa się mianem statystyki ma dość skomplikowaną historię. Jeśli chodzi o przedział czasowy, to jest on stosunkowo krótki, gdyż obejmuje zaledwie około 350 lat. Jeśli zaś uwzględnia się źródła, z których ona wyrosła, to sięgają one tysięcy lat. Na samym początku XX w. M. Poznański pisał w „Ekonomiście", że historia statystyki to historia sporów o jej treść i zakres. Jak żadna inna nauka, statystyka szczyci się swoją własną humorystyką. Do jej kanonów należy „statystyka określeń statystyki".

Przyjmijmy że historia statystyki rozpoczyna się wraz z rozwojem nauki nowożytnej, czyli w XVII w. Cały jej rozwój do tego momentu określimy mianem prehistorii.

Początki rozwoju statystyki były związane z potrzebami fiskalno-obronnymi państwa. Każdy władca potrzebował pieniędzy nie tylko na budowę fortyfikacji obronnych i utrzymania wojska, ale też na budowę swoich pałaców oraz rezydencji. Pieniądze były potrzebne też na utrzymanie wdów i sierot, których mężowie i ojcowie zginęli w wojnach. Od najdawniejszych czasów najważniejsze źródło pieniędzy stanowili obywatele lub mieszkańcy danego kraju. Problem władcy polegał na określeniu wielkości pobieranych danin. Były one pobierane w zależności od zamożności płatników. Należało więc przede wszystkim posiadać informacje o liczebności kraju i zamożności jego obywateli. Uzyskiwano je w wyniku przeprowadzonych spisów ludności oraz poprzez sporządzanie rejestrów nieruchomości – gruntów oraz budynków.

W dalszej części pobieżnie omówiono trzy spisy opisane w Biblii. Dwa z nich przeprowadził Mojżesz, trzeci zaś, król Dawid. Oprócz tego przedstawione są trzy spisy dokonane przez władców świeckich.

Spisy ludności i opis majętności całego kraju przeprowadzone na polecenie władców nazywano spisami urzędowymi. Dały one początek temu, co w obecnych czasach nazywa się statystyką publiczną lub oficjalną, czyli urzędową.

Niezależnie od takich zlecanych opisów, opisy państw dokonywano także z pobudek czysto poznawczych, czyli naukowych. Opisy te, w odróżnieniu od urzędowych, Z. Korzyński – polski statystyk z XIX w. – nazywa opisami prywatnymi. Były one dokonywane nie w celu zaspokojenia praktycznych potrzeb władców, lecz aby wyjaśnić stan rzeczy.

Pierwszego opisu prywatnego dokonał Arystoteles. Zgodnie z rozwiniętą przez siebie filozofią uważał, że opis czegokolwiek, co istnieje, czyli bytu, polega głównie na opisie zasad, które nadają określoną strukturę wszystkiemu temu, co istnieje. Zasady te Arystoteles nazwał przyczynami i wyróżnił ich cztery: formalne, materialne, sprawcze oraz celowe.

W wiekach średnich zaczęto „z urzędu" prowadzić rejestry kościelne. Oprócz tego duże znaczenie miały „sprawozdania" dyplomatów i urzędników, jakie musieli składać ze swych wojaży zagranicznych.

Na szczególną uwagę zasługują sprawozdania włoskie, czyli *Relazioni*. W 1296 r. gubernatorowie weneckich prowincji byli zobowiązani zawiadamiać swoje rządy o wszystkim, co na uwagę zasługuje.

Pod koniec wieków średnich na Półwyspie Apenińskim było wiele lokalnych państw rządzonych przez kilka możnych rodów. Do najbardziej znanych należał ród Viscontich, Medyceuszy i Sforzów. Warto zauważyć, że drugą żoną Zygmunta I Starego była Bona Sforza d'Aragona (1494-1557), która pochodziła z jednego z tych rodów.

1. Spisy biblijne

Do najstarszych, a jednocześnie dobrze udokumentowanych, należą biblijne spisy Izraelitów. Ich prezentacja wymaga choćby bardzo pobieżnej znajomości historii Izraelitów. Praojcem Izraelitów był Abraham, który urodził się, według rachuby biblijnej, w 1948 r. po stworzeniu świata. Osiedlił się w krainie wybrzeża Morza Śródziemnego, którą wówczas nazywano Kanaanem; dzisiaj są to tereny Palestyny. W krainie tej wykupił na własność pieczarę o nazwie Makrela, w której pochował swoją żonę Sarę. Znajdują się tam również groby Patriarchów, w szczególności samego Abrahama i jego wnuka Jakuba, od którego pochodzi nazwa całego narodu, Jakub bowiem po całonocnej walce z aniołem uzy-

skał nowe imię – Izrael – co oznacza mniej więcej tego, który „walczy z Bogiem" lub który „walczy o Boga", ewentualnie „sprawiedliwy przed Bogiem" (por. Johnson, 1998). Jakub-Izrael był ojcem 12 plemion, z których składał się cały naród, w Biblii nazywany Izraelem.

Prawdopodobnie z powodu niedostatku żywności naród ten wyemigrował do Egiptu. H. Graetz, znany profesor Uniwersytetu Wrocławskiego, w 1888 r. emigrację tę ujął następująco: „W siedemdziesiąt tylko dusz wywędrowawszy czasu głodu z kraju Kanaan, odwiecznego siedliska praojców, do Egiptu i przyjaźnie zrazu przyjęci, rozmnożyli się nadzwyczajnie w Egipcie północnym (ziemi Gosen)" (Graetz, 1929, s. 5-6).

Z upływem czasu życie Izraelitów w Egipcie było coraz bardziej uciążliwe. W szczególności za czasów panowania Ramzesa II (1304-1237 r. p.n.e) byli wyjątkowo uciskani. Po nieudanych pertraktacjach z faraonem zdecydowali się opuścić Egipt.

„Pierwszego dnia drugiego miesiąca w drugim roku po wyjściu z ziemi egipskiej przemówił Pan do Mojżesza na pustyni Synaj w Namiocie Zgromadzenia tymi słowy:

Zróbcie spis całego zboru synów izraelskich według ich szczepów i rodów, spis imienny wszystkich mężczyzn, głowa po głowie" (Księga Liczb, 1-2). W wyniku tego polecenia został dokonany pierwszy spis Mojżeszowy. Po upływie 40 lat „(…) rzekł Pan do Mojżesza i do Eleazara, syna Aarona, kapłana, mówiąc: sporządź spis całego zboru synów izraelskich od dwudziestego roku wzwyż według ich rodów wszystkich zdatnych do służby wojskowej w Izraelu" (Księga Liczb 26, 1-2).

Ich wyjście z Egiptu opisane jest w drugiej księdze biblijnej o nazwie Exodus. Wyjście to miało miejsce około 1225 r. przed narodzeniem Chrystusa.

Oba te spisy były wielokrotnie poddawane różnym analizom. Są one przedstawione w poniższej tabeli.

Tabela 1

Spisy ludności plemion izraelskich

12 plemion Izraela	Pierwszy spis	Drugi spis
Ruben	46 500	43 730
Symeon	59 300	22 200
Czad	45 650	40 500
Juda	74 600	76 500
Issachar	54 400	64 300
Zabulon	57 400	60 500
Manasses	32 200	52 700
Efraim	40 500	32 500
Beniamin	35 400	45 600
Dan	62 700	64 400
Aser	41 500	53 400
Neftal	53 400	45 400
Razem	**603 550**	**601 730**

Źródło (Spisy Ludności…, 2002, s. 4).

2. Spis i reformy Solona

Rozwój każdej cywilizacji rozpoczyna się z chwilą kształtowania się struktur społecznych i form sprawowania władzy.

Podstawową formą organizacji społeczeństw, która dała początek cywilizacji nazywanej dzisiaj cywilizacją zachodnią była instytucja społeczna o nazwie polis. Greckie słowo πολις (polis) tłumaczone jest zwykle jako państwo lub miasto-państwo, jednak jak dowiadujemy się z głębszych analiz tego pojęcia (por. Kumaniecki, 1975), to polis nie miało charakteru tego, co dzisiaj rozumiemy pod pojęciem państwa. Nie miało bowiem policji ani systemu fiskalno-podatkowego. Pod pojęciem polis rozumiano pewien (niezbyt wielki) obszar geograficzny, na którym zamieszkała tam ludność sama stanowiła o swoich sprawach. Nie wszyscy mieli jednakowe prawa, a niektórzy nie mieli ich wcale; byli to niewolnicy, stanowiący własność ludzi wolnych. Ludzie wolni dzielili się na dwie klasy. Jedną stanowili arystokraci, określani greckim słowem *avistoi* (najlepsi), drugą klasę stanowił lud, określany słowem *kakoi* (źli).

Nie trzeba nikogo przekonywać, że przyjaźni między lepszymi i gorszymi nigdy nie było. Wraz z upływem czasu pogłębiała się natomiast przepaść między lepszymi, którzy byli bogaci, oraz gorszymi – biedniejszymi. Biedni, żeby wyżyć, często zapożyczali się u bogatych, stawiając w zastaw swoją własną osobę. W ten sposób ubywającą społeczność niewolników uzupełniano nie tylko poprzez ich zakup, ale niewolnik stanowił również formę spłaty pożyczki.

Tę bardzo niezdrową sytuację, grożącą rozpadem państwa, postanowił uzdrowić Solon, zaliczany według tradycji do słynnego grona siedmiu mędrców starożytności, wśród których są również: Tales z Miletu, Pittakos z Mityleny, Bias z Prieny, Periandros z Koryntu, Kleobulos z Lindos, Chilon ze Sparty.

W 594 r. p.n.e. Solon został wybrany archontem Aten, czyli przywódcą wojskowym i politycznym. W tym samym roku przeprowadził reformę społeczną. Przede wszystkim postanowił umorzyć wszystkie długi, w szczególności te najokrutniejsze – w postaci niewolnictwa. Zniósł również przywileje rodowe, a uprawnienia do sprawowania urzędów i obowiązki łożenia na rzecz utrzymania i funkcjonowania wszystkich instytucji państwowych zostały uzależnione od statusu majątkowego.

Przeprowadzając spis wszystkich obywateli, Solon dokonał ich klasyfikacji. Według słów samego Arystotelesa zrobił to następująco: „Według majątku podzielił ich na cztery klasy, które już i przedtem istniały: *pentakosiomedimnoi*, *hippeis*, zeugitów i thetów. Prawo sprawowania wszystkich urzędów: dziewięciu archontów, skarbników, poletów, Jedenastu i kola kretów przyznał obywatelom z klasy *pentakosiomedimnoi, hippeis* i zeugitów, przydzielając przy tym urzędy każdemu stosownie do wysokości jego majątku; należącym do klasy thetów przyznał jedynie prawo udziału w Zgromadzeniu i w sądach" (por. Arystoteles, 2001).

Wysokość majątku mierzono wielkością rocznych plonów uzyskiwanych z posiadanych gospodarstw. Jednostkę miary stanowił jeden medymnos (gr. μεδιμνος), uważa się, że stanowił on pojemność ok. 52-53 litrów.

Pierwszą klasę stanowili pięćsetmedymnowcy, czyli ci, których dochód roczny był nie mniejszy niż 500 medynów. Drugą klasę stanowili rycerze i jeźdźcy (gr. *hippeis*), których majątek przekraczał 300 medymnów. W czasie wojny oni mieli obowiązek dostarczania jeźdźców. Trzecią klasę stanowili rzemieślnicy i chłopi. Do tej klasy należeli ci, których roczny dochód szacowano na 200 medymnów. Oni stanowili główny trzon wojska. Czwartą klasę stanowili tzw. Teci (gr. *thetes*), czyli parobkowie i najemnicy. Oni nie posiadali stałego dochodu. W razie potrzeby (wojny) teci odbywali służbę na wodzie.

3. Cenzus Serwiusza Tuliusza

Dzisiejsze słowa cenzor, cenzura czy cenzus pochodzą z czasów szóstego z kolei króla rzymskiego. Aby lepiej zrozumieć pochodzenie i sens tych słów, warto przyjrzeć się okresowi historii Rzymu, określanego jako archaiczny.

Jest to okres od legendarnych czasów ósmego wieku do wieku piątego przed narodzeniem Chrystusa według tradycji Rzym został założony w 753 r. p.n.e. Od tej legendarnej daty liczono lata historii Rzymu (*ab urbe condita*). Na początku swego istnienia było to królestwo, czyli taki ustrój państwa-miasta, w którym najwyższą władzę sprawuje król. Zgodnie z tradycją było siedmiu królów: Romulus, Numa Pompiliusz, Tullus Hostiliusz, Hankus Marcjusz, Tarkwiniusz Priskus, Sergiusz Tuliusz i Tarkwiniusz Superbus. Historia potwierdza istnienie tylko trzech ostatnich królów z tej listy.

Ze względu na historię statystyki (statystyki urzędowej) warta uwagi jest jedynie sylwetka szóstego z wymienionych – Sergiusza Tuliusza. Podobnie jak w Grecji, podstawę organizacji życia społecznego w państwie, którym rządził król Tuliusz stanowił ród (*gens*). Rodów w owym czasie było 300, które dzieliły się na 30 kurii (*curia*), w każdej kurii były zaś 3 plemiona (*tribus*). Podstawową jednostkę społeczną stanowiła rodzina, w której główną rolę odgrywał ojciec rodziny zwany *pater familius*. Ojcowie rodzin stanowili radę starszych określaną mianem senat (od łacińskiego słowa senatu, starszy). Potomków ojców rodzin nazywano *patricić* czyli patrycjuszami. Była to klasa uprzywilejowana. Drugą klasę, pozbawioną przywilejów, stanowili plebejusze od łacińskiego słowa *plebei* oznaczającego mniej więcej napełnianie, czyli byli to ci, którzy napełniali państwo siłą roboczą. Oprócz patrycjuszy i plebejuszy, podobnie jak w Grecji, dużą grupę społeczną stanowili niewolnicy, których traktowano jako narzędzie mówiące; po łacinie określano ich mianem *instrumentum vocale*.

Aby osłabić władzę rodów i ułatwić rządzenie państwem, Tuliusz przeprowadził w 550 r. p.n.e. reformę społeczną, dokonując przy okazji spisu powszechnego wolnych mieszkańców swego kraju.

Każdy mieszkaniec Rzymu miał zgłosić się do urzędnika określanego mianem *cenzora*. Funkcję jaką on wykonywał nazywano *cenzusem*. Jedną z wielu czynności było dokonanie spisu wszystkich obywateli, z których każdy był przypisywany do jednej z pięciu klas według jego statusu majątkowego. Jeśli nie posiadał żadnego majątku, to zaliczany był do klasy szóstej. Stanowili ją więc obywatele będący poza klasyfikacją (łac. *infra classem*). Klasę tę określano mianem *proletarii*, czyli proletariuszy. Nazwę taką obrano dlatego, że głównym zadaniem członków tej klasy było dostarczanie potomstwa, (łacińskie słowo *proli* oznacza latorośl, potomka). Proletariusz nie miał więc nic, ale miał rodzić dzieci.

4. Kataster Wilhelma Zdobywcy

Około 911 r. wojowniczy lud Skandynawii, który nazywano Normanami napadł na Francję i w jej zachodniej części założył księstwo znane jako Normandia. Około 1028 r. księciu Normandii urodził się nieślubny syn Wilhelm. Był on podobno dalekim krewnym króla Anglii Edwarda Wyznawcy, dlatego też kiedy król zmarł nie zostawiwszy następcy, Wilhelm postanowił sięgnąć po koronę brytyjską. Pomocne okazało się nie rzekome pokrewieństwo, lecz wojsko. Podbiwszy Anglię, będąc księciem Normandii, Wilhelm stał się też królem Anglii. Do historii przeszedł jako Wilhelm I Zdobywca. Nazywano go także Wilhelmem Bękartem, mimo iż nie krył swego urodzenia, to jednak surowo karał tych, którzy go tak nazywali. Aby się zorientować o stanie posiadania podbitego kraju, Wilhelm I zarządził sporządzenie katastra gruntowego. To nieco dziwnie brzmiące słowo nie ma jasnej etymologii; wyprowadza się je od starogreckiego słowa *katastichon*, oznaczającego rejestr, spis. Słowo to w wersji włoskiej przyjęło formę *catastico*. Kataster sporządzony na polecenie Wilhelma Zdobywcy został sporządzony w 1086 r. Jest on uważany za najbardziej znaczący dokument statystyczny w historii Europy, zawiera opis typograficzny całej Anglii. Szczegółowo opisano 13 418 miejscowości, podając ich granice, rodzaje upraw, pogłowie bydła itp.

Cały dokument nosi nazwę *Domesday Book*, czyli w tłumaczeniu na język polski – *Księga Sądu Ostatecznego*. Nazwa taka wynika z tego, że decyzje urzędników dokonujących spisu nie podlegały żadnej apelacji, czyli tak, jak ma być w Dniu Sądu Ostatecznego.

5. *Politeja Ateńska* Arystotelesa

Politeja Ateńska jest pierwszym zachowanym do czasów obecnych opisem państwa ateńskiego, dokonanym przez Arystotelesa. Arystoteles urodził się w 384/383 r. p.n.e. w Stagirze. Mając osiemnaście lat przybył do Aten i od razu wstąpił do Akademii Platońskiej.

W latach 335/334 p.n.e. wynajął kilka budynków w pobliżu świątyni poświęconej Lykejosowi i założył własną szkołę, którą nazwano Lykeion, od tej nazwy pochodzi polskie słowo liceum. Arystoteles jest autorem wielu pism; wszystkie one zostały ostatnio pięknie wydane, z wybornymi komentarzami, przez Polskie Wydawnictwo Naukowe. W tomie 6 „Dzieł wszystkich" wydanych w 2001 r. zawarte jest dzieło pt. *Politeja ateńska* (w oryginale ma ono tytuł Αυηναιων πολιτεια).

Całe dzieło składa się z 69 rozdziałów zgrupowanych w dwie części. W pierwszej części obejmującej rozdziały od 1 do 41 opisane są dzieje Aten od czasów legendarnych do 403 r. p.n.e. Część druga obejmująca rozdziały od 42 do 69 poświęcona jest ustrojowi czasów współczesnych Arystotelesowi. Ustroje te zmieniały się jedenaście razy. Na samym początku był to ustrój rodowy, w którym podstawową jednostkę społeczną stanowił ród, na czele którego stał naczelnik zwany basilensem. Rady łączyły się we fratrie, a te z kolei w plemiona, nazywane fylami. Początkowo państwo ateńskie składało się z czterech plemion, czyli czterech fyl, z których każda dzieliła się na trzy trytie.

Formy ustrojowe zmieniały się jedenaście razy. Krótkie ich streszczenie zawarte jest w ostatnim (41) rozdziale części pierwszej. Niżej przytoczony jest on w całości (por. Arystoteles, 2001, s. 757-759). „Twórcą pierwszej formy ustroju był Ion oraz ci, którzy się wraz z nim osiedlili, albowiem dopiero wówczas przeprowadzono podział na cztery fyle i ustanowiono królów fyl. Druga zmiana, ale pierwsza zasługująca na nazwę ustroju, przeprowadzona została za Tezeusza, kiedy ograniczono nieco władzę królewską. Następnym – był ustrój Drakona, połączony z pierwszym spisaniem praw, trzecim – wprowadzony po walkach ustrój Solona, który położył podwaliny demokracji; czwartym – tyrania Pizystrata, piątym po obaleniu tyranów ustrój Klejstenesa, bardziej demokratyczny aniżeli Solona. Szósty nastał po wojnach medyjskich, kiedy to Rada Areopagu ujęła ster rządów. Do siódmego z kolei utorował drogę Arystoteles, a doprowadził do końca Efialtes, pozbawiwszy znaczenia Radę Areopagitów. W tym okresie państwo pod wpływem demagogów najwięcej popełniło błędów w związku z panowaniem na morzu. Ósma zmiana ustroju nastąpiła z ustanowieniem rządów Czterystu, a następny, dziewiąty, był znów demokratyczny. Dziesiątym była tyrania Trzydziestu i Dziesięciu, a jedenastym ustrój wprowadzony po powrocie powstańców z Fyle i Pireusu, który przetrwał aż dotąd, przy czym

lud zdobywa coraz większe wpływy. Sam bowiem zrobił siebie panem wszystkiego i zarządza wszystkim przez uchwały Zgromadzeń i przez sądy, w których ma głos decydujący. Przecież nawet uprawnienia sądowe Rady przeszły na lud. I dobrze, zdaje się, zrobili, gdyż małą garstkę ludzi łatwiej jest pozyskać czy to widokami zysku, czy osobistymi względami. Początkowo nie zgodzono się na wprowadzenie wynagrodzenia za udział w Zgromadzeniu. Ponieważ jednak obywatele nie przychodzili na zgromadzenia, a prytanowie na próżno wymyślali różne sposoby, żeby zebrać ludzi dla powzięcia uchwały, wprowadził Agyrrios wynagrodzenie w wysokości jednego obola; Heraklejdes z Kladzomenaj, nazywany «królem», podniósł je następnie do wysokości dwóch oboli, i znów Agirrios podwyższył do trzech".

Charakterystyka ustroju z czasów Arystotelesa zawarta jest w drugiej części jego dzieła. Zostanie przytoczony początkowy jego fragment (Arystoteles, 2001, s. 759-761): „Obecny ustrój polityczny przedstawia się następująco: Prawa polityczne posiadają ci, których oboje rodzice są obywatelami, wpisuje się ich w poczet członków demu, gdy ukończą lat osiemnaście. Kiedy zgłaszają się do zapisania, democi, złożywszy przysięgę, stwierdzają przez głosowanie, czy kandydaci posiadają wiek wymagany prawem, jeśli zaś okaże się, że tak nie jest, przechodzą z powrotem do liczby małoletnich; następnie sprawdzają, czy kandydat jest człowiekiem wolnym i urodził się w związku spełniającym wymogi prawa; jeżeli uznają, że nie jest wolny, może się odwołać do sądu, a członkowie demu wybierają ze swego grona pięciu oskarżycieli. Jeśli się okaże, że bezprawnie domaga się wpisania {na listę obywateli}, państwo sprzedaje go; jeśli zaś wygra sprawę, członkowie demu są zmuszeni wpisać go na listę. Potem wpisanych bada Rada i jeśli się okaże, że ktoś nie ma osiemnastu lat, wymierza grzywnę tym demotom, którzy wpisanie przeprowadzili.

Po dokimazji efebów, zbierają się ich ojcowie według fyl i po złożeniu przysięgi wybierają trzech członków z fyli, w wieku powyżej czterdziestu lat, których uznają za najlepszych i najodpowiedniejszych do kierowania efebami; lud zaś wybiera spośród nich, po jednym z każdej fyli, sofronistę i spośród ogółu Ateńczyków kosmetę, który jest naczelnym kierownikiem wszystkich. Ci wybrani, zebrawszy efebów, obchodzą z nimi najpierw świątyni, a następnie wyruszają do Pireusu i jedni z efebów pełnią straż w Munichii, a inni na Akte. Również Zgromadzenie Ludowe przez głosowanie wybiera dla nich dwóch pajdotrybów oraz nauczycieli, którzy szkolą ich w walce w pełnym uzbrojeniu, w strzelaniu z łuku, rzucaniu oszczepem i miotaniu pocisków z procy. Na wyżywienie dają każdemu sofroniście po drachmie, a efebom po cztery obole na głowę. Każdy sofronista otrzymuje pieniądze od członków swej fyli, zakupuje, co trzeba, dla wszystkich razem (jedzą bowiem wspólnie według fyl) i w ogóle troszczy się o wszelkie inne potrzeby.

W ten sposób spędzają rok pierwszy. W następnym roku na posiedzeniu eklezji, które odbywa się w teatrze, popisują się przed ludem zręcznością w ćwiczeniach wojskowych, otrzymują od polis tarczę i włócznię, a potem rozchodzą się po kraju i pełnią służbę wojskową w strażnicach. Pełnią tę służbę strażniczą przez dwa lata, noszą wtedy chlamidy i są wolni od wszelkich obowiązków obywatelskich. I nie może ich nikt pozywać do sądu ani też oni nie mogą nikogo oskarżać, aby nie mieli pretekstu do oddalania się, wyjąwszy wypadki, gdy chodzi o spadek i poślubienie spadkobierczyni, albo jeżeli któremuś z nich przypadnie związana z rodem godność kapłańska. Po upływie tych dwóch lat wchodzą już w szeregi innych obywateli".

6. Średniowieczne opisy państw

Zostaną wymienione w porządku chronologicznym ważniejsze opisy państw dokonane u schyłku wieków średnich:

1. Eneasz Silvio Piccolomini (Aeneas Silvius), późniejszy papież Pius II w 1531 r. jeszcze jako kardynal napisał dzieło *Asiae Europaeque descriptio*, znane także jako *Cosmographia*, w którym opisane są kraje Azji i Europy.
2. Sebastian Münster (1488-1552) napisał pierwszą w języku niemieckim pracę dotyczącą opisu całego świata. Pierwsza część tej pracy pod tytułem *Cosmographia*, wydana została w 1536 r., w całości zaś w 1544 r. W języku niemieckim wydawana była aż 19 razy. Ostatnie niemieckie wydanie ma datę 1628, a więc wydano je 76 lat po śmierci autora. Praca ta była przetłumaczona na języki łaciński, angielski, francuski, włoski, a nawet czeski. Składa się ona z 6 ksiąg. W pierwszej opisany jest cały świat według Ptolemeusza. W kolejnych księgach opisane są kraje Europy, Azji i Afryki.
3. Francesco Sansovino (1521-1586) w 1562 r. wydał pracę pt. *Dal governo ed amministrazione di diversi regni ed republiche*, w którym opisuje ustroje 20 państw. Oprócz tego opisuje też państwa starożytne: Rzym, Spartę i Ateny, a nawet utopijne państwo Platona.
4. Giovanni Botero (1544-1617) w wieku 15 lat wstąpił do zakonu Jezuitów, kształcił się w wielu ich kolegiach, ale z powodu jednego kazania doktrynalnie błędnego został z zakonu wykluczony w 1580 r. Nie przeszkodziło mu to jednak być osobistym sekretarzem C. Borromeo, biskupa Mediolanu.

   W 1589 r. wydał on pracę w języku włoskim *Le ralazioni universali divisi in quatro parti,* w języku łacińskim jej tytuł jest nieco inny: *Relationes Universales de viribus, opijus et regimine principium Europae, Asiae et Africae*. W dziele tym szczegółowo opisane są w sposób porównawczy ustroje 20 państw. Stanowiło ono podstawowe źródło wiedzy, z którego jeszcze prawie

po stu latach obficie korzystał H. Conring. Dzieło to było przetłumaczone na język polski przez Łęczyckiego, kapłana zakonu Ojców Bernardynów. Część dotycząca Polski podobno „nic dla Polaka ważnego nie zawiera" (por. Kobrzyński, 1870). Oprócz tego Botero napisał jeszcze dwa inne ważne dzieła: *Ragione di stato* (Racja państwa) i *Cause della grandezza Della csittá* (O przyczynach wielkości miast).

5. Marcin Kromer (1512-1589) napisał pracę pt. *Polonia, seve de situ, populis, moribus, magistratibus et republica regni Polonici libri duo auctore Martino Cromero*, która została wydana w 1577 r. Na język polski została przetłumaczona w 1853 r. jako *Polska, czyli o położeniu, obyczajach, urzędach Rzeczypospolitej Królewstwa Polskiego.*
6. Jan Krasiński (1550-1612) był kanonikiem krakowskim i gnieźnieńskim. W 1574 r. wydał w Bolonii w języku łacińskim pracę *Joanis Crassinii Polonia* z dedykacją *Ad Serenissimum et Potentissimum Henricum I Valesium, Poloniqe Regem.* W 1582 r. praca ta została przetłumaczona na język polski. Powyższa dedykacja jest ciekawa; praca była wydana dokładnie w roku koronacji Henryka Walezego na króla Polski, który jeszcze w tym samym roku uciekł z Polski. Był on synem Henryka de Valois (stąd Walezy) i Katarzyny Medycejskiej, uważanej za sprawczynię Nocy św. Bartłomieja, w której tuż przed wyborem na króla Polska brał udział Henryk Walezy.

Bibliografia

Arystoteles (2001): *Dzieła wszystkie*. Tom 6. Wydawnictwo Naukowe PWN, Warszawa.

Cotterell A., red. (1990): *Cywilizacje starożytne*. Wydawnictwo Łódzkie.

Graetz H. (1929): *Czasy prastare. Historia Żydów*. Wydawnictwo „Judaica", Warszawa.

Johnson P. (1998): *Historia Żydów*. Platan, Kraków.

Kobrzyński Z. (1870): *Wstęp do teorii statystyki*. Warszawa.

Kumaniecki K. (1975): *Historia kultury starożytnej Grecji i Rzymu*. PWN, Warszawa.

Lexikon der Antike (1977). VEB, Leipzig.

Mały atlas biblijny (1990). KUL, Lublin.

Matthew D. (1996): *Europa wieków średnich*. Świat Książki, Warszawa.

Meitzen A. (1891): *History, Theory, and Technique of Statistics*. American Academy of Political and Social Science, Philadelphia.

Nussbaum H. (1893): *Przewodnik judaistyczny*. Warszawa.

Poznański M. (1907): *Kilka uwag o metodologicznych zadaniach statystyki*. „Ekonomista", tom 1, s. 257-269.

Reale G. (2005): *Historia filozofii starożytnej*. Tom II. KUL, Lublin.

Sedlaček T. (2012): *Ekonomia dobra i zła*. Studio Emka, Warszawa.

Spisy ludności Rzeczypospolitej Polskiej. Polskie Towarzystwo Demograficzne i Główny Urząd Statystyczny, Warszawa 2002.

Wipszycka E., red. (1985): *Vademecum historyka starożytnej Grecji i Rzymu*. PWN, Warszawa.

REMARKS ON THE PREHISTORY OF STATISTICS

Summary

The history of Statistics as a science is dated to 1660. The roots of Statistics are however in the antiquity. In this paper there are presented some, the most significant, events on the long way of the development of this science. We start with the Biblical censuses of Moses, and David, then, the reforms of Solon and Tulius Servius are shortly reviewed, as well as the census performed by William the Conqueror, and his famous Doomsday Book. The paper is closed by the short list of medieval writers.

Jean H.P. Paelinck*
George Mason University

SPATIAL ECONOMETRICS: A PERSONAL OVERVIEW

Introduction

The paper is written on the basis of my lecture given at the Academy of Economics, University of Katowice, June 2012.

Spatial econometrics is a fast-growing field in the series of quantitative disciplines, auxiliaries of economics and related social sciences.

The subject matter is all the more complex as it has a twin brother, spatial statistics, which in many aspects is complementary of today's subject.

We will essentially treat spatial econometrics as we have experienced it, and are still practicing, referring to other materials to cover spatial statistics proper; a recent paper with references is Griffith and Paelinck (2004), a condensed version appearing in Griffith and Paelinck (2011).

Next section is devoted to an obscure period of our discipline, when essentially "spatial econometrics without space" was practiced. Then, progressively, that space emerged, and fragments of real spatialized exercises could be excavated from dispersed articles and books, as section 3 testifies.

In 1979, with our colleague Leo Klaassen, we defined a number of principles that, we hoped, would guide future work along lines that tried to identify the specificity of spatial econometrics, totally respecting the teachings of general econometrics; the same applies to theoretical spatial economics, which integrates all the principles of general theoretical economics.

Looking ahead is a more dangerous undertaking, for which we will rely on our recent work; statements resulting from this exercise will probably be controversial, but we hope they will stimulate a fruitful discussion on some delicate points of the discipline.

Conclusions and references follow, as usual.

---

* Distinguished Visiting Professor, Doctor honoris causa of University of Economics in Katowice.

1. Prehistory

Let us start with mentioning the fact that we entered in contact with spatial problems on the occasion of treating a problem in regional development. In those times, very little statistical material happened to be available, so reasoning had to proceed in a still theoretical way, the basis being a theory of polarized regional development. The concepts set out there (Paelinck, 1963) already invited a *joint* consideration of regional developments, which became later the principle of asymmetric inter-regional linkages.

Econometric exercises published over that period were of the most classical type, relating only variables possessing the same regional index. Examples are Thompson and Mattila (1959, see the comments on this study in Paelinck and Klaassen, 1979, p. 6; Vanhove, 1963, we used the latter in an exercise of calculating the parameters of an optimal multiregional policy (Paelinck, 1967), already noticing there that the underlying model used was inadequate to represent the correct spatial workings of the economy, which would be reflected in the policy outcomes. To quote from Paelinck (1967, pp. 57-58): "(…) les résultats de l'économétrie régionale, telle qu'elle est souvent pratiquée, sont fortement affectés par la négligence de deux facteurs essentiels:

- la localisation relative des régions faisant l'objet de l'étude;
- la localisation intra-régionale des activités sur lesquelles porte l'analyse.

(…) De statique et indifférenciée, l'économétrie doit devenir *dynamique et différenciée*; des modèles adaptés au *caractère spécifique de l'analyse régionale* doivent être mis au point".

The same criticism, levied against regional macro-models, was repeated in Paelinck (1973).

One can justifiably ask the question: why has no attention being given to spatially entwined activities? On the one hand, time series analysis knew the notion of serial correlation, reflected on in terms of lagged variables, exogenous and stochastic. On the other hand, input-output analysis was starting to develop multi-regional variants, though typical topological variables were absent. Furthermore, in studies of inter-regional and international movements (inter-regional migration or inter-zonal commuting, international trade) gravity models were in use, which implied the use of certain distance measures.

The answer lies probably in the fact that the bridge between spatial analysis and econometrics proper had still to be built. How progressively that has happened will become clear from the next section.

2. Interim period

Things would change in the late sixties and seventies. In 1966 (see also Lebart, 1969), a colleague of ours at the "Institut d'Etude du Développement Economique et Social" of Paris University, Ludovic Lebart, applied the Geary statistic (Geary, 1954) to French departmental data, generalizing it and studying spatial autocorrelation at different degrees of contiguity, in fact measuring distance effects, as the degree of contiguity is a fully-fledged metric.

A similar study was published in Paelinck and Nijkamp (1975, pp. 223-234 and 243-246) using five different variables (number of inhabitants per square kilometer, rate of growth of population, employment per square kilometer, per capita gross value added at factor prices, per capita available income); the first variable showed systematic positive spatial autocorrelation, the second and third ones a negative one, the last two variables again positive autocorrelation, except at order two. These findings, for 1960 and 1965, gave an acceptable insight in the spatial structure of the variables taken up in the study.

Moran's I (1948, 1950) gained popularity over the same period, the Cliff and Ord volume (1973) on spatial autocorrelation completing the picture.

What was missing though was a complete integration of theoretical spatial economics and econometrics proper; this was what "Spatial Econometrics" (1979) meant to do.

3. Recent times

The initial vision on what spatial econometrics could be was expressed in our General Address to the Dutch Statistical Association on the occasion of its Annual Meeting on May 2nd 1974, held at the city of Tilburg.

The integration just mentioned was laid down in the form of five principles which we would like to remember briefly.

The first one was already introduced by the previous considerations, to wit spatial interdependence. The new vision however was to derive that interdependence from the workings of spatial economies. Two models were essentially staged, an income generating model and a so-called attraction model; the first started from a spatial generalization of the Keynesian macro-economic consumption-cum-investment model, introducing spatial shopping behavior and so generating spatial consumption propensities, later to be estimated; the second was a generalized Weberian sectoral location model, based on locational choices as a function of expected profits. As an obiter dictum, let it be said that I still used those models to train Ph.D. students at the School of Public Policy of George Mason

University. Without going into detail, it should be mentioned that the classical econometric problems – specification, dimensional homogeneity, identification, estimation, testing – are systematically taken up.

The second principle was that of spatial asymmetry in the measures of spatial interdependence; this was already known to be the case for trade flows, but the idea had to be generalized to parameters. An example, modeled and estimated, was that of spatial consumption behavior in terms of urban and non-urban spatial unit within the spatial income generating model mentioned above.

Both these principles show a stark contrast with the simple measurement of spatial moving average or spatial autocorrelation process parameters.

Principle three was called "allotopy", from the Greek words αλλος and τοπος, meaning respectively "other" and "site"; it alludes to the influence at a distance of exogenous variables, the Weberian location model, already mentioned, being a perfect example of this.

Choice problems in space lead more than often to non-linear solutions, so this too should be reflected in our model specification, especially if the model intends to picture ex-ante choices; this is the case of locational models (see e.g. the European FLEUR-model: Ancot and Paelinck, 1983), though resulting ex-post behavior (for instance, transport flows) could still be treated linearly.

Finally, topological variables (locations, distances, densities, etc.) should not be absent; as said earlier, spatial models should not be "without space"! It should explicitly be said here that the choice of a distance measure is a strategic moment in the model specification, too little attention being often given to the problems of metric topology involved.

As spatial econometrics *is* about economics, much stress was laid on the specification of the underlying model, the first moments of the distributions, so to say (see: Paelinck and Klaassen, 1979, conclusions to chapter 2, pp. 42-43); we will come back to this important point in our further developments.

It was very fortunate that shortly afterwards a stormy evolution took place, of which we will only mention some recent volumes devoted to spatial econometrics, skipping but not neglecting the articles, and insisting also on the value of the contributions in the complementary field of spatial statistics (for references, see again Griffith and Paelinck, 2004); one is referred to Anselin (1988), Anselin and Florax (1995), Morena Serrano and Vayá Valcarce (2000), Anselin, Florax, and Rey (2004), Getis, Mur, and Zoller (2004), Lesage, Page and Tiefelsdorf (2004), Tivez a.o. (2005), Arbia (2006), Arbia and Fingleton (2008), Mur, and Lopez (2010), Lepage and Pace (2009), Griffith and Paelinck (2011). The crowding over the last years is, at the least, impressive.

No wonder that a Spatial Econometric Association (SEA) was founded at Rome, on May 26th 2006, on the occasion of another international workshop on

spatial econometrics. Article 2 of its statutes stipulated that: ”The purpose of the Association is to promote the development of the theoretical instruments and practical applications of spatial econometrics – including spatial statistics and spatial data analysis. Spatial econometrics should be regarded in abroad sense and inclusive of the developments of statistic models and instruments to analyze externalities, interactions, etc. in different areas such as, without limitation, economics, geography and regional sciences. The mission of the Association is to disseminate and encourage knowledge and good practice among universities and research institutions and in the community in general, becoming a reference point for operators in the field”.

The activity of SEA has led to further initiatives; the important fact is indeed hat a specific SE platform is now present, platform from which diffusion can proceed. And what has to come will probably be still more exciting.

4. Looking ahead

Some challenges the future will be faced with, can be grouped under five headings: spatial bias, spatial specification, spatial estimation, spatial complexity, isomorphisms; these themes will be taken up in succession in what follows. They lead straightforwardly to what we have called “non-standard spatial econometrics” (see again Griffith and Paelinck, 2011).

4.1. Spatial bias

Taking an econometric view of the MAUP (Modifiable Areal Unit Problem), Paelinck (2000) proved that in principle all spatial econometric models will inevitably show up an aggregation bias; his findings were summarized in the following terms: „The important result is that in general econometric aggregation, if only one macro-aggregate is considered, just one parameter bias is present in the macro-model; in meso-aggregation, as it took place here, every meso-area has its own specific aggregation bias, which leads to parameter variability between meso-areas, and this might result, in econometric estimation, in some sort of (biased) average value, depending on the characteristics of the sample being investigated and the particular spatial aggregation specification”.

The consequence of this is that there is an extra element to simple spatial heterogeneous behavior, and how to isolate the two aspects has to be seriously considered (Griffith and Paelinck, 2011, chapter 17). So only heterogeneity calls for extra treatments: composite parameters (Ancot et al., 1971), differential spa-

tial regimes (Griffith and Paelinck, 2011, chapter 12); special attention is called for, and further fundamental investigation of this problem is highly advocated.

4.2. Spatial specification

Given the characteristics of spatial inter-linkages, special attention should be devoted to the specification of spatial models; scanning five collective volumes on spatial econometrics, we found out that only 14% of the papers were devoted to that problem (Griffith and Paelinck, 2004).

As an example we would like to mention here the well-known problem of spatial convergence. A simple beta-convergence model is, in our opinion, not the right tool to tackle the problem; in Arbia and Paelinck (2003 a and b) an alternative is proposed and presented, to wit a non-linear – this lon-linearity being recently refined (see: Griffith and Paelinck, 2011, chapter 11) – fully-fledged inter-regional model of the Lotka-Volterra type. It allowed to separate the problems of *mathematical* and *economic* convergence, and arrived at the conclusion that in a system of 119 European regions, if mathematical convergence was present, economic convergence was simply lacking, something in fact already confirmed by some theoretical and empirical studies.

But this is still not the end of the journey; indeed, an even more important problem is that of the algebraic structure to be given to the model under construction. For that we studied (Paelinck, 2001) the possibilities of model specification based on a so-called "min-algebra"; that algebra, in fact, generalizes the specification of the European FLEUR-model (Ancot et Paelinck, 1983), the latter being based on the idea of a "growth threshold". In a min-algebra, one (or several) explanatory terms (variables with their reaction coefficients) *of minimal value* determine the value of the endogenous variable(s). So, instead of considering a (linear or non-linear) combination of endogenous, exogenous or predetermined variables, one will only consider one (or a limited number) of explanatory variables in each equation; for instance, the development of a region could be hampered by the absence of a strategic factor, such as technologically highly trained manpower.

A bad specification of regional models becomes really dramatic when they are used to derive "baskets" of regional policy measures; indeed, the logic of the chosen algebra produces a linear programming solution with more non-zero decision variables than the number of constraints, whereas under a "classical" algebra would in general produce only that number of non-zero decision variable.

Still another specifications is that of a finite automaton, which can be seen as an "if"-specification, for example, it could read as follows: if $\alpha x_i + \beta < \gamma z_i + \delta$, the values of the left-hand terms obtains, otherwise those of the right-hand one. To

submit such a finite automaton model to an empirical test in a well-documented case, gross regional product figures for the Netherlands have been investigated. They were divided in two macro-regional sets, one for the western provinces (Noord-Holland, Zuid-Holland, Utrecht, the so-called "Rimcity"), the other one comprising the data for the remaining provinces.

The curious thing, at first sight, was the behavior of the growth rate values for the non-Rimcity provinces: whatever the state of the location factors attractiveness, they follow the ups and downs of the Rimcity growth rates. This is completely in line with the fact that the Rimcity is indeed the "motor" of the Dutch economy (Paelinck, 1973, pp. 25-40, especially pp. 37-40), imposing its evolutionary rhythm to the other regions. This finding has recently been confirmed with a Lotka-Volterra finite automaton specification (Griffith and Paelinck, 2011, chapter 13).

A last remark is on the specification of spatial lags, in the endogenous and/or exogenous variables, the *W*-matrix problem to call it that way. Several suggestions have been made, some of them purely "mechanical"; again, as spatial econometrics is about economics in pre-geographical space, some economic background is desirable. One possibility is the use of a flexible potential function (Ancot and Paelinck, 1983); another one is the following procedure: estimate the model without spatial lags and without constant(s); inspect the residuals – the "doggy-bag principle" (Griffith and Paelinck, 2011, chapter 14) which should not add up to zero; relatively high and/or low, positive and/or negative instances should be inspected trying to generate assumptions (e.g., competition and/or cooperation could be present at short or long distances: distance can be "hampering" or "protecting"). Examples are known where mapped locations of residuals led to the right complementary variable missing in the model.

This does not end the list of relevant specification problems; the main point is that the model should correspond to the workings of the spatial economies investigated, and that the consequences for their ulterior use – or uses – should be explicitly considered.

4.3. Estimators

The fundamental structure of spatial models invites at developing different types of estimators adapted to the special situations encountered; indeed, the established software does not always fit the specific estimation problem encountered.

Paelinck (1990) proposes various estimators especially fit to treat particular circumstances.

A first example is a simultaneous dynamic least squares estimator (SDLS), perfectly well adapted for use in the type models treated above, i.e. models with simultaneous spatial and temporal interdependencies. Instead of minimizing

residuals between observed endogenous values and values estimated from the latter – even if they are predetermined – it minimizes a norm between observed and *endogenously computed* values. This takes into account future uses of the model, like forecasting or computing policy impacts.

Originally the algorithm consisted of computation by iterative OLS, the method integrating the derivation of optimal – spatial and temporal – „starting points" for endogenous simulations. The estimator is consistent, and the probability limit of its variance-covariance matrix is known (in fact, it is the usual OLS matrix), but we will come back to this remark.

In recent research, the estimating procedure has been endogenized (again Griffith and Paelinck, 2011, chapter 11), resulting in the parameters and the estimated – endogenous – variables being produced in the same process; the method can be applied to spatial models – with their typical spatial lags – both static and dynamic. Recently also the method has been applied to the estimation of composite parameters; not implying an inverse, but resting on mathematical programming, the problem of non-identifiability could be side-stepped successfully.

As already mentioned, most spatial models are inherently non-linear, so that after appropriate specification adequate estimation methods should be used. Let only be mentioned here a recent estimation method for Box-Cox transformations (Paelinck and van Gastel, 1995; Griffith, Paelinck, and van Gastel, 1998), the estimation proceeding from the (partial) elasticities of the transformed function.

Some form of semi-parametric estimation should also be considered, by semi-parametric estimation being understood here that a second order differential expansion (derived from a second order MacLaurin expansion) is used, allowing of changing the coefficients of the linear terms by adding periodically the coefficients of the quadratic terms, which, in production functions e.g., express the changes in marginal productivities.

Applying this technique, originally developed for time series, to a problem in *spatial* econometrics, raises the difficulty originating from the difference between „time's arrow" and the non-oriented presentation of spatial data; a solution is presented in Griffith and Paelinck (2004). The specification does not privilege any direction in space, and allows for increases or decreases of the reaction parameter between couples of regions with the same separating distances.

Finally, min-algebraic and finite automata parameters can be estimated and the specification tested; one is again referred to Griffith and Paelinck (2011, chapter 13).

One important remark to conclude this topic: we consider the estimation problem not as one centering only around the parameters and their properties (see also Valavanis, 1959, especially p. 47), but on the whole model and its future uses, the parameters being in fact auxiliaries to allow us a relevant representation of the spatial-economic reality. An example is the SDLS-estimator treated earlier, with its recent specification and applications.

4.4. Complexity

To get an idea of the relevance of one or another specification, one should look into the complexity of the problem, by which we mean *computational complexity* of the series to be explained (Chaitin, 1975; Wolfram, 2002, pp. 557--559); this complexity, which we will call *conditional complexity* – due to the presence of exogenous variables – can be expressed through the number of parameters necessary to fit to an endogenous variable a polynomial in the exogenous ones. An indicator on [0,1] is (Getis and Paelinck, 2003)*:*

$$c = (n_p - 1)/(n_{pm} - 1),$$

where n_p is the number of non-zero parameters, and n_{pm} their maximum number (equal to the length of the series of endogenous variables, i.e. the size of the sample). Let it be said that we first considered the endogenous variables as void of measurement errors, as the observed values are the only ones of which we avail; but at the same time as spatial bias could be filtered out, those errors too can be treated appropriately (again Griffith and Paelinck, 2011, chapter 17).

Just to give an example, in such a pre-test of the model to be specified, a series of observations lead to a complete cubic equation, with a value $c = 1$, the ensuing test between a simple linear model and a finite automaton one selecting the latter; when errors and spatial bias were filtered out, the simple linear model submerged, the complexity of the data having been reduced by two thirds.

4.5. Isomorphism

Isomorphism, in the broad sense of specification similarity can be a serious help in solving some classical problems in spatial econometrics. Take the extended SAR model

$$\mathbf{y} = \mathbf{Ay} + \mathbf{Xb} + \boldsymbol{\varepsilon}$$

The first fact to be noted is that the y variables have all the same definition, GRP, for instance, contrary to the classical non-spatial model $\mathbf{Ay+Xb=\varepsilon}$, where the vector **y** consists of different variables. A second fact is that the model just presented is isomorphic to the classical input-output model

$$\mathbf{y} = \mathbf{Ay} + \mathbf{f}$$

where **A** is the input coefficients matrix, **f** the final demand vector.

The difficulty with the spatial econometric model, compared to the input-output one, is that the input coefficients are known from statistical observation, which is not the case of the **A** matrix in model in the extended SAR model; for instance, from the input-output equation total relative inputs can be computed as **a' = i'A**; and this has indeed ledto a useful suggestion for solving an identification problem of the spatial econometric model.

Conclusions

"Spacetime" economics – to borrow an expression from theoretical physics – is indeed characterized by great complexity; we only refer here to some studies in potentialized partial differential equations we did with a former Erasmus University colleague (Kaashoek and Paelinck, 1994, 1996, 1998 and 2002) which lead to unexpected spatial patterns, which have been empirically verified Coutrot et al., 2009). Interpreting those patterns start with theoretical spatial economics and should flow over into spatial econometrics, if we want our theories to confront the facts, possibly to be contradicted, at times to be relegated to the waiting room of theories pending there ever uncertain status.

Anyhow, the real challenge for the future is to create beauty from garbage, as the expression goes.

References

Ancot J.-P., Paelinck J.H.P. (1983): *The Spatial Econometrics of the European FLEUR-Model*. In: *Evolving Geographical Structures*. Eds. D.A. Griffith, A. Lea Martinus. Nijhoff Publishers, The Hague, pp. 220-246.

Ancot J.-P., Chevailler J.-Cl., Paelinck J.H.P., Smit H., Stijnen H. (1971): *Parameter Component Models in Spatial Econometrics*. The Econometrics of Panel data. Annales de l'INSEE, 30-4/31, Paris, pp. 53-98.

Anselin L. (1988): *Spatial Econometrics: Methods and Models*. Kluwer Academic Publishers, Dordrecht et al. loc.

Anselin L., Florax R.J.G.M., eds. (1995): *New Directions in Spatial Econometrics*. Springer, Berlin et al. loc.

Anselin L., Florax R.J.G.M., Rey S., eds. (2004): *Advances in Spatial Econometrics. Methodology, Tools and Applications*. Springer, Berlin et al. loc.

Arbia G. (2006): *Spatial Econometrics*. Springer, Berlin-Heidelberg.

Arbia G., Fingleton B., eds. (2008): *New Spatial Econometric Techniques and Applications in Regional Science*. Special issue of „Papers in Regional Science", Vol. 87, No. 3.

Arbia G., Paelinck J.H.P. (2003a): *Economic Convergence or Divergence? Modeling the Interregional Dynamics of EU Regions, 1985-1999*. „Journal of Geographical Systems", Vol. 5, No. 3, pp. 291-314.

Arbia G., Paelinck, J.H.P. (2003b)*: Spatial Econometric Modeling of Regional Convergence in Continuous Time*. „International Regional Science Review", Vol. 26, No. 3, pp. 342-362.

Chaitin G.J. (1975): *Randomness and Mathematical Proof*. „Scientific American", Vol. 232, No. 5, pp. 47-52.

Cliff A.D., Ord J.K. (1973): *Spatial Autocorrelation*. Pion Press, London.

Coutrot B., Sallez A., Paelinck J.H.P., Sutter R. (2009): *On Potentialized Partial Finite Difference equations: Analyzing the Complexity of Knowledge-Based Spatial Economic Activities*. „Région et Développement", No. 29, pp. 201-228.

Geary R.C. (1954): *The Contiguity Ratio and Statistical Mapping*. „The Incorporated Statistician", Vol.5, pp.115-145.

Getis A., Mur J., Zoller H.G., eds. (2004): *Spatial Econometrics and Spatial Statistics*. Palgrave-Macmillan, Basingstoke.

Getis A., Paelinck J.H.P. (2004): On *Analytical Descriptions of Geographical Patterns*. „L'Espace Géographique", No. 1, pp.61-68.

Griffith D.A., Paelinck J.H.P, van Gastel M.A.J.J. (1998): *The Box-Cox Transformation: New Computation and Interpretation Features of the Parameters*. in D.A.

Griffith D.A., Amrhein C., Huriot J.-M., eds. (1998): *Econometric Advances in Spatial Modeling and Methodology*. Kluwer, Dordrecht, Advanced Studies in Theoretical and Applied Econometrics, Vol. 35, pp. 45-58.

Griffith D.A., Paelinck J.H.P. (2004): *An Equation by Any Other Name is Still the Same: Spatial Statistics and Spatial Econometrics*. Paper Presented at the Annual Meetings of the NARSA, Seattle, November.

Griffith D.A, Paelinck J.H.P. (2011):, *Non-Standard Spatial Statistics and Spatial Econometrics*. Springer, Berlin and Heidelberg.

Kaashoek J.F., Paelinck J.H.P. (1994): *On Potentialised Partial Differential Equations in Theoretical Spatial Economics*. In: Special issue of *Chaos, Solitons and Fractals* on „Non-Linear Dynamics in Urban and Transport Analysis". Ed. D.S. Dendrinos. Vol. 4, No. 4, pp. 585-594.

Kaashoek J.F., Paelinck J.H.P. (1996): *Studying the Dynamics of Pre-geographical Space by Means of Space and Time-Potentialised Partial Differential Equations*. „Geographical Systems", Vol.3, pp. 259-277.

Kaashoek J.F., Paelinck J.H.P. (1998): *Potentialised Partial Differential Equations in Economic Geography and Spatial Economics: Multiple Dimensions and Control*. „Acta Mathematica Applicanda", Vol. 51, pp. 1-23.

Kaashoek, J.F. Paelinck J.H.P. (2001): *Potentialised Partial Differential Equations in Spatial Economics: Some Further Results on the Potentialising Function*. „The Annals of Regional Science", Vol. 35, No. 3, pp. 463-482.

Lebart L. (1966): *Les variables socio-économiques départementales et regionales*, I, Méthodes statistiques de l'étude, IEDES, Paris.

Lebart L. (1969): *Analyse statistique de la contiguité*. Publications de l'Institut de Statistique de l'Université de Paris, 18, pp. 81-112.

LeSage J., Page R.K., Tiefelsdorf M., eds. (2004): Special Issue on Methodological Developments in Spatial Econometrics and Statistics of *Geographical Analysis*. Vol. 36, No. 2.

LeSage J., Pace R.K. (2009): *Introduction to Spatial Econometrics*. CRC Press, Boca Raton et al. loc.

Moran P.A.P. (1948): *The Interpretation of Statistical Maps*. „Journal of the Royal Statistical Society B", Vol. 10, pp. 243-251.

Moran P.A.P. (1950): *A Test for Serial Correlation of Residuals*. „Biometrica", Vol. 37, pp. 178-181.

Moreno Serrano R., Vayá Valcarce E. (2000): *Técnicas econométricas para el tratamiento de datos espaciales: la econometría espacial*. Edicions Universitat de Barcelona, UG44, Manuals, Barcelona.

Mur J., Lopez F.A., eds. (2010): *Modeling Spatio-Temporal Data*. special issue of „Journal of Geographical Systems", Vol. 12, No. 2.

Paelinck J.H.P. (1963): *La teoría del desarrollo regional polarizado*. *Revista Latino-americana de Economía* (Caracas), No. 9, pp. 175-229 (reproduced in *Cahiers de l'ISEA*, Série L (Economies Régionales), March 1965, pp. 5-47,with title: *La théorie du développement régional polarisé*, and in Italian with title: *La teoria dello sviluppo regionale polarizzato*, in A. Testi (a cura di): *Sviluppo e pianificazione regionale*. Giulio Einaudi, Torino 1977.

Paelinck J.H.P. (1967): *L'efficacité des mesures de politique économique régionale*. Rapport introductif, Faculté des Sciences Economique, Centre de Recherches, Namur.

Paelinck J.H.P. (1973): *Hoe doelmatig kan regionaal en sectoraal beleid zijn?* Stenfert Kroese, Series: Bedrijfseconomische Verkenningen, Leiden.

Paelinck J.H.P., Nijkamp P. (1975): *Operational Theory and Method in Regional Economics*. Saxon House, Farnborough.

Paelinck J.H.P., Klaassen L.H. (1979): *Spatial Econometrics*. Saxon House Farnborough

Paelinck J.H.P. (1990): *Some New Estimators in Spatial Econometrics*. In: *Spatial Statistics: Past, Present and Future*. Ed. D.A. Griffith. Syracuse University, Institute of Mathematical Geography, Monograph No. 12, pp. 163-181.

Paelinck J.H.P., van Gastel M.A.J.J. (1995): *Computing Box-Cox Transform Parameters: A New Method and its Application to Spatial Econometrics*. In: *New Dimensions in Spatial Econometrics*. Eds. L. Anselin, R.J.G.M. Florax. Springer, Berlin, pp. 136-155.

Paelinck J.H.P. (2000): *On Aggregation in Spatial Econometric Modelling*. „Journal of Geographical Systems", Vol. 2, No. 2, pp. 157-165.

Paelinck J.H.P. (2001): *A Multiple Gap Approach to Spatial Econometrics*. „The Annals of Regional Science", Vol. 36, No. 2, pp. 219-228.

Thompson W.R., Mattila J.M. (1959): *An Econometric Model of Postwar State Industrial Development*. Wayne State University Press, Detroit.

Trivez F.J. et al., eds. (2005): *Contributions to Spatial Econometrics*. Copy Center Digital, Zaragoza.

Valavanis S. (1959): *Econometrics, An Introduction to Maximum Likelihood Methods.* McGraw-Hill, New York et al. loc.

Vanhove N. (1963): *De doelmatigheid van het regionaal-economisch beleid in Nederland*. Gent and Hilversum.

Wolfram S. (2002): *A New Kind of Science*. Wolfram Media, Champaign (IL).

SPATIAL ECONOMETRICS: A PERSONAL OVERVIEW

Summary

The paper is based on the invited lecture given at the Katowice University of Economics in June 2012.

In the paper the beginnings of subdiscipline called spatial econometrics are presented. The history of spatial statistical analysis and its influence on economic surveys is considered. Author's contribution to this area is presented together with personal overview of the developments of the subdiscipline. Moreover, some future challenges connected with "non-standard spatial econometrics" are analyzed including spatial bias, spatial specification, spatial estimation, spatial complexity and isomorphisms.

Józef Pociecha
Uniwersytet Ekonomiczny w Krakowie

WYBRANE METODY KLASYFIKACYJNE ORAZ ICH EFEKTYWNOŚĆ W PROGNOZOWANIU UPADŁOŚCI FIRM

Wprowadzenie

„Tworzenie pojęć oraz reguł pozwalających klasyfikować obiekty na klasy podobieństwa odpowiadające tym pojęciom towarzyszy rozwojowi każdej dyscypliny naukowej. Proces badania, czy dowolny obiekt należy do danej klasy, nazywamy rozpoznawaniem obiektu" (Kolonko, 1982, s. 7). Takim stwierdzeniem rozpoczyna się podstawowe dzieło profesora Józefa Kolonki, wprowadzające czytelnika w szeroką, a w tym czasie mało znaną problematykę klasyfikacji zadań, metod klasyfikacji oraz zagadnień redukcji opisu. Ta niewątpliwie nowatorska w tym czasie praca dała impuls do intensywnego rozwoju metodologii klasyfikacji danych i stała się inspiracją do publikowania dalszych prac z zakresu taksonomii, metod dyskryminacyjnych oraz metod rozpoznawania obrazów. Praca prof. Kolonki wskazywała jednocześnie na szerokie możliwości wykorzystania proponowanej metodologii badawczej do wielorakich analiz ekonomicznych, m.in. do normowania kosztów, konstrukcji grupowych modeli konsumpcji, analizy danych przekrojowo-czasowych oraz porównywania i porządkowania obiektów wielocechowych. Mimo upływu lat od czasu jej opublikowania ciągle inspiruje do rozwoju metod klasyfikacji danych oraz ich zastosowań w badaniach społeczno-ekonomicznych.

Artykuł ten nawiązuje do problematyki metodologicznej pracy prof. Kolonki oraz jest kolejnym przykładem możliwości zastosowania metod klasyfikacji danych we współczesnych analizach ekonomicznych. Przedmiotem artykułu jest przedstawienie najpopularniejszych metod prognozowania bankructwa, będących w istocie statystycznymi lub iteracyjnymi metodami klasyfikacji danych, oraz ocena ich efektywności jako narzędzia oceny ryzyka upadłości firm.

Według przeprowadzonych badań (Aziz, Dar, 2004) jako narzędzia prognozowania upadłości firm najczęściej stosowane są modele dyskryminacyjne (30,3% publikowanych prac), na drugim miejscu pod względem częstości zastosowań znajdują się modele logitowe (21,3%), na trzecim miejscu sieci neurono-

we (9,0%) a na czwartym miejscu – drzewa klasyfikacyjne (5,6%). Te też narzędzia zostaną pokrótce scharakteryzowane, a w dalszej kolejności zostaną przytoczone ich pierwotne zastosowania znane z literatury światowej oraz wybrane zastosowania dla celów predykcji bankructwa firm we współczesnej gospodarce polskiej. W końcowej części pracy przedstawiono ocenę efektywności przedstawionych metod jako narzędzi prognozowania bankructwa. Ich efektywność została zmierzona zdolnością do poprawnej klasyfikacji firm do zbioru bankrutów i niebankrutów na zbiorze uczącym oraz zbiorze testowym. W zakończeniu sformułowano pewne ogólne warunki określające przydatność prezentowanych metod do predykcji bankructwa firm.

1. Liniowa funkcja dyskryminacyjna

Liniowa funkcja dyskryminacyjna, jako narzędzie klasyfikacji danych, została sformułowana w 1936 r. przez R.A. Fishera (1936) i zastosowana w badaniach przyrodniczych. Poniżej zostanie przedstawiona jej zasadnicza idea. W ujęciu klasycznym* przyjmujemy, że mamy dwie populacje. Załóżmy że $X_1 = [x_1, x_2,..., x_{n1}]$ jest wektorem n_1 obserwacji z populacji 1 a $X_2 = [x_{n1+1}, x_{n1+2},..., x_{n1+n2}]$ jest wektorem n_2 obserwacji, pochodzącym z populacji 2. Są to wektory o wymiarach $p \times 1$, gdzie p jest liczbą zmiennych dyskryminujących. Fisher zaproponował liniową funkcję (dyskryminator) dla zaklasyfikowania wylosowanej obserwacji do jednej z dwóch wyróżnionych populacji. Metoda ta polega na znalezieniu liniowej transformacji oryginalnych zmiennych:

$$l(X) = a^t X \tag{1}$$

tak aby było możliwe ich maksymalne odseparowanie w dwóch populacjach. Zaproponował znalezienie wektora $\hat{a}$ maksymalizującego funkcję separacji $|S(a)|$, gdzie:

$$S(a) = \frac{\bar{y}_1 - \bar{y}_2}{S_y}, \tag{2}$$

$\bar{y}_1$ oraz $\bar{y}_2$ są średnimi zmiennych transformowanychY_1 z populacji 1 oraz Y_2 z populacji 2,

* Spośród wielu prac z tego zakresu w języku angielskim patrz np.: (Christensen, 1991; Giri, 1996 lub Rencher, 1998). Wprowadzenie do tej problematyki wraz z zastosowaniami do prognozowania bankructwa można znaleźć m.in. w pracach: (Pociecha, 2006 lub Pociecha, 2007).

$$S_y = \left[\frac{\sum_{i=1}^{n_1} (y_i - \bar{y}_1)^2 + \sum_{i-n1+1}^{n1+n2} (y_i - \bar{y}_2)^2}{n_1 + n_2 - 2} \right]^{\frac{1}{2}} \tag{3}$$

oraz

$$y_i = a^t x_i, \; i = 1, 2, ..., n_1 + n_2. \tag{4}$$

S(a) dane wzorem (2) mierzy różnicę pomiędzy przekształconymi średnimi $\bar{y}_1 - \bar{y}_2$ relatywnie do wielkości odchylenia standardowego z prób (3). Jeśli zmienne przekształcone $y_1, y_2,...,y_{n1}$ oraz $y_{n1+1}, y_{n1+2}, y_{n1+n2}$ byłyby całkowicie odseparowane, to $|\bar{y}_1 - \bar{y}_2|$ powinno być możliwie duże.

Wektor $\hat{a}$ maksymalizujący separację $|S(a)|$ jest następujący:

$$\hat{a} = S^{-1}(\bar{x}_1 - \bar{x}_2) \tag{5}$$

gdzie:

$$S = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2} \tag{6}$$

$$S_1 = \frac{\sum_{i=1}^{n1} (x_i - \bar{x}_1)(x_i - \bar{x}_2)}{n_1 - 1} \tag{7}$$

$$S_2 = \frac{\sum_{n1+1}^{n1+n2} (x_i - \bar{x}_1)(x_1 - \bar{x}_2)}{n_2 - 1} \tag{8}$$

oraz $\bar{x}_1$, $\bar{x}_2$ są średnimi wektora 1 i 2.

Przyjmijmy że mamy obserwację x_0. Opierając się na funkcji dyskryminacyjnej (1), możemy alokować tę obserwację do odpowiedniej populacji na podstawie następujących reguł klasyfikacyjnych:

– alokuj x_0 do populacji 1, jeśli:

$$\hat{y}_0 = (\bar{x}_1 - \bar{x}_2)^t S^{-1} x_0 \geq \frac{1}{2}(\bar{x}_1 - \bar{x}_2)^t S^{-1}(\bar{x}_1 + \bar{x}_2) \tag{9}$$

– alokuj x_0 do populacji 2, jeśli:

$$\hat{y}_0 = (\overline{x}_1 - \overline{x}_2)^t S^{-1} x_0 < \frac{1}{2}(\overline{x}_1 - \overline{x}_2)^t S^{-1}(\overline{x}_1 + \overline{x}_2) \tag{10}$$

Innymi słowy, jeśli $\hat{y}_0$ znajduje się po prawej stronie $\frac{\overline{y}_1 + \overline{y}_2}{2}$ (bliżej do $\overline{y}_1$), zaklasyfikuj x_0 do populacji 1 i odwrotnie.

Analiza dyskryminacyjna jest od przeszło 40 lat wykorzystywana do prognozowania upadłości firm.

2. Model logitowy

Modele logitowe należą do klasycznych modeli klasyfikacji binarnej, tj. takich, w których zmienna objaśniana przyjmuje tylko dwie wartości. Przy analizowaniu zagrożenia upadłością przedsiębiorstwa badamy wpływ zmiennych objaśniających, które mogą być cechami jakościowymi lub ilościowymi, na zmienną objaśnianą o charakterze jakościowym, przy założeniu logistycznej postaci analitycznej natężenia poszczególnych zmiennych objaśniających*. Jeśli zmienna objaśniana ma charakter zero-jedynkowy (przedsiębiorstwo o dobrej kondycji finansowej, przedsiębiorstwo o złej kondycji), to mamy do czynienia z modelem dychotomicznym. Gdy zmienna objaśniana jest cechą jakościową wielowariantową (bankrut, średnia kondycja finansowa, dobra kondycja finansowa, to mamy do czynienia z modelem wielomianowym uporządkowanym.

Wartości zmiennej objaśnianej wskazują na wystąpienie lub brak wystąpienia pewnego zdarzenia, które chcemy prognozować. Regresja logistyczna pozwala na obliczanie prawdopodobieństwa tego zdarzenia jako prawdopodobieństwa sukcesu według wzoru:

$$P(x) = \frac{e^x}{1+e^x} \tag{11}$$

lub

$$P(x) = \frac{1}{1+e^{-x}} \tag{12}$$

czyli funkcji logistycznej podającej prawdopodobieństwo bankructwa firmy.

Model regresji logistycznej należy do grupy uogólnionych modeli liniowych (GLM), w którym wykorzystano formułę logitu jako funkcji wiążącej (McCulloch, Searle, Neuhaus, 2009). Logitem nazywamy funkcję przekształcającą prawdopodobieństwo na logarytm ilorazu szans, czyli:

* Na temat modelu ligitowego oraz możliwości jego zastosowania patrz np.: (Pociecha, 2012).

$$L = \text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \ln(p) - \ln(1-p) \tag{13}$$

Dystrybuanta F prawdopodobieństwa sukcesu jest wzięta z rozkładu logistycznego, czyli:

$$p = F(\beta_0 + \beta_1 + ... + \beta_k + \varepsilon) = \frac{\exp(\beta_0 + \beta_1 X_1 + ... + \beta_k X_k + \varepsilon)}{1 + \exp(\beta_0 + \beta_1 X_1 + ... + \beta_k X_k + \varepsilon)} \tag{14}$$

Wartość funkcji odwrotnej do dystrybuanty F jest logitem (L) określonym wzorem (13). Logit jest więc logarytmem ilorazu szans bankructwa i „niebankructwa" firmy. Jeżeli te szanse są jednakowe, czyli p = 0,5, to logit jest równy zero, dla p > 0,5 logit jest dodatni, a gdy p < 0,5 – jest on ujemny.

Po przekształceniu logitowym można przystąpić do badania zależności pomiędzy wartościami logitu a zmiennymi objaśniającymi, będącymi odpowiednimi wskaźnikami finansowymi, przyjmując najczęściej model liniowy o postaci:

$$L = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + \varepsilon \tag{15}$$

Parametrów powyższego modelu nie można szacować klasyczną metodą najmniejszych kwadratów, gdyż nie jest spełnione założenie stałości wariancji dychotomicznej zmiennej objaśnianej. W tym przypadku można stosować uogólnioną metodę najmniejszych kwadratów, a obecnie, szczególnie przy małej liczbie obserwacji stosuje się metodę największej wiarygodności.

3. Sieć neuronowa

Sztuczna sieć neuronowa jest techniką informatyczną wzorowaną na strukturze i sposobie działania układów nerwowych organizmów żywych*. Sztuczny neuron jest modelem swojego rzeczywistego odpowiednika. Jego zasadniczym celem jest przetworzenie informacji wejściowej, dostarczanej w postaci wektora skończonej liczby sygnałów wejściowych $x_1, \ldots, x_n$ w wartość wyjściową y. Przyjmuje się, że zarówno wartości wejściowe neuronu, jak i wartość wyjściowa mają postać liczb rzeczywistych. Z każdym wejściem neuronu związany jest współczynnik nazywany wagą. Współczynniki wagowe neuronu są podstawowymi parametrami wpływającymi na sposób funkcjonowania sztucznej komórki nerwowej.

* Charakterystyka sieci neuronowych i ich zastosowania dla celów prognozowania bankructwa została przedstawiona m.in. w pracy: (Pociecha, 2010b).

Oprócz n wejść oraz jednego wyjścia (y), występuje także wartość stała dla każdego neuronu w_0 (wyraz wolny), niezależna od danych wejściowych, nazywana biasem. W najprostszym przypadku, tj. neuronu liniowego, przetwarzanie sygnałów wejściowych odbywa się na podstawie obliczenia sumy ważonej wejść o wagach odpowiednio w_i, czyli:

$$\overline{y} = w_0 + \sum_{i=1}^{n} w_i x_i = w_0 + \mathbf{w}^T \mathbf{x} \tag{16}$$

gdzie:

$\mathbf{x} = [x_i]$ – wektor $[n \times 1]$ sygnałów wejściowych,

$\mathbf{w} = [w_i]$ – wektor $[n \times 1]$ wag, które z jednej strony wyrażają stopień ważności informacji przekazywanej i-tym wejściem, a z drugiej strony stanowią swojego rodzaju pamięć neuronu, zapamiętują bowiem związki zachodzące pomiędzy sygnałami wejściowymi a sygnałem wyjściowym.

W pewnych zastosowaniach przyjmowane są inne formuły agregacji, np. w postaci kwadratu odległości Euklidesa pomiędzy wektorem wejściowym **x** a wektorem wag **w**; wtedy takie neurony są nazywane neuronami radialnymi (Witkowska, 2002).

Drugim etapem przekształcenia otrzymanej wartości pobudzenia jest wyznaczenie sygnału (wartości) wyjściowej. Elementem odpowiedzialnym za wykonanie tej czynności jest funkcja aktywacji:

$$y = \varphi(\overline{y}) \tag{17}$$

Funkcja aktywacji może być funkcją liniową lub nieliniową.

Spośród znanych funkcji aktywacji najczęściej przyjmuje się funkcję logistyczną:

$$\varphi(\overline{y}) = \frac{1}{1 + \exp(-\beta \overline{y})} \tag{18}$$

lub tangens hiperboliczny:

$$\varphi(\overline{y}) = \frac{\exp(\beta \overline{y}) - \exp(-\beta \overline{y})}{\exp(\beta \overline{y}) + \exp(-\beta \overline{y})} \tag{19}$$

W celu stworzenia sieci neuronowej łączy się neurony w określony sposób. Zwykle neurony wchodzące w skład sieci tworzą warstwy, z których pierwsza nosi nazwę warstwy wejściowej, ostatnia – warstwy wyjściowej, zaś wszystkie znajdujące się pomiędzy nimi określane są jako warstwy ukryte. Wartości wejściowe sieci wprowadzane są na wejścia neuronów warstwy wejściowej. Następnie, poprzez istniejące połączenia, wartości wyjściowe neuronów jednej warstwy przekazywane są na wejścia elementów przetwarzających kolejnej war-

stwy. Wartości uzyskane na wyjściach neuronów ostatniej warstwy są wartościami wyjściowymi sieci.

Zwolennicy sieci neuronowych jako narzędzia prowadzenia badań ekonomicznych wskazują na ich przewagę w stosunku do klasycznych modeli regresji, modeli dyskryminacyjnych oraz klasycznych modeli tendencji rozwojowych. Z tego też względu sieci neuronowe zalecane są także jako narzędzie prognozowania bankructwa. Jednakże sposób funkcjonowania sieci neuronowej, gwarantujący prawidłowe rozwiązywanie postawionych przed nią problemów, uzależniony jest od dwóch podstawowych czynników (Lula, 1999):

- wartości współczynników wagowych neuronów składających się na sieć,
- struktury (topologii) sieci, która określana jest przez liczbę warstw, liczbę neuronów w poszczególnych warstwach, sposób połączeń neuronów oraz przyjęty model neuronu (sposób agregacji danych wejściowych, rodzaj zastosowanej funkcji aktywacji).

Ze względu na architekturę sieci neuronowych można wyróżnić trzy jej główne grupy. Pierwszą z nich stanowią sieci jednokierunkowe. Typowym przykładem sieci jednokierunkowej jest perceptron wielowarstwowy. Drugą grupę stanowią sieci rekurencyjne, w których dopuszcza się występowanie cykli. Dynamika tego typu sieci jest zdecydowanie bardziej skomplikowana niż w przypadku sieci jednokierunkowych. Typowym przykładem sieci rekurencyjnej jest sieć Hopfielda (Tadeusiewicz, 1993). Trzecią grupę sieci neuronowych stanowią sieci komórkowe. W tej grupie sieci łączone są w dowolny sposób, ale tylko w obrębie sąsiedzkich węzłów. Typowym przykładem tego typu sieci jest sieć SOM Kohonena.

Po podjęciu decyzji co do wyboru właściwej architektury sieci neuronowej i jej zbudowaniu, należy rozpocząć proces przygotowania sieci do prawidłowego jej działania, zwany uczeniem sieci. Sieć uczy się prawidłowo działać na podstawie prezentowanych jej przykładów realizacji badanych obiektów lub zjawisk. Opierając się na przedstawionych rzeczywistych przypadkach, sieć stara się odkryć i zapamiętać ogólne prawidłowości charakteryzujące te obiekty lub kierujące przebiegiem badanych zjawisk. Rozpoznanie reguły sztuczna sieć neuronowa przechowuje w postaci zakodowanej w wartościach współczynników wagowych neuronów.

Zbiór danych wykorzystywany w trakcie uczenia sieci nazywamy zbiorem uczącym. Proces uczenia sieci uwalnia nas od uciążliwego tworzenia i zapisywania algorytmu wymaganego dla przetwarzania danych wejściowych, tak aby uzyskać pożądany wynik końcowy. Nie odbywa się to jednak bez kosztów, gdyż ceną jest długotrwały i wymagający dużych mocy obliczeniowych proces uczenia. Co więcej, proces uczenia sieci jest zawsze procesem indeterministycznym,

czyli wynik uczenia nie jest nigdy całkowicie pewny. Najczęściej stosowanym algorytmem uczenia z nauczycielem jest algorytm wstecznej propagacji błędu oparty na regule *delta* (Korbicz, Obuchowicz, Uciński, 1994).

W przypadku uczenia bez nauczyciela zbiór uczący zawiera tylko wartości zmiennych wejściowych. Uczenie polega na cyklicznym prezentowaniu danych uczących i na stopniowej, systematycznej modyfikacji wag, prowadzącej w efekcie do wytworzenia w sieci pewnej wiedzy o ogólnych cechach i właściwościach zbiorowości sygnałów wejściowych. Sieci uczone w trybie bez nauczyciela stosowane są do rozwiązywania zadań klasyfikacji bezwzorcowej, mającej na celu rozpoznanie struktury analizowanego zbioru obiektów lub identyfikacji jednorodnych fragmentów szeregów czasowych. Podstawowym algorytmem treningu sieci w trybie bez nauczyciela jest reguła Hebba (Ossowski, 1996).

4. Drzewo klasyfikacyjne

Drzewo klasyfikacyjne jest jednym z obrazów podziału rekurencyjnego badanego zbioru. Polega on na stopniowym podziale wielowymiarowej przestrzeni cech na rozłączne podzbiory, aż do uzyskania ich homogeniczności ze względu na wyróżnioną cechę (y). Następnie w każdym z uzyskanych segmentów budowany jest lokalny model tej zmiennej. Graficzną prezentacją metody podziału rekurencyjnego jest drzewo decyzyjne. Jeśli takie drzewo odnosi się do cechy y będącej cechą nominalną, to reprezentujące ją drzewo nazywane jest drzewem klasyfikacyjnym, a jeśli jest to zmienna ciągła, to takie drzewo nazywamy drzewem regresyjnym (Por. Gatnar, 2001, s. 26). W przypadku prognozowania bankructwa nasza cecha y jest cechą binarną (bankrut, niebankrut), dlatego też mówimy o drzewach klasyfikacyjnych jako narzędziu bankructwa.

Przedmiotem podziału rekurencyjnego (Por. Gatnar, 2001, s. 13) jest N-elementowy zbiór obiektów, w naszym przypadku firm, scharakteryzowanych przez wektor $M+1$ cech, będących wskaźnikami ich kondycji finansowej (płynność, zadłużenie, efektywność działania, rentowność): $[\mathbf{x},y]$, gdzie:

$$\mathbf{X} = [x_1, x_2, \ldots, x_M]. \tag{20}$$

Wielowymiarowe obserwacje wskaźników finansowych w analizowanym zbiorze firm można zapisać w postaci macierzy:

$$[\mathbf{x}_n, y_n]_{N\times M+1} = \begin{bmatrix} x_{11} & \cdots & x_{1M} & y_1 \\ x_{21} & \cdots & x_{2M} & y_2 \\ \cdots & \cdots & \cdots & \cdots \\ x_{N1} & \cdots & x_{NM} & y_N \end{bmatrix} \tag{21}$$

Zgodnie z tradycyjnymi określeniami zmienne $x_1, x_2, \ldots, x_M$ są nazywane predykatorami, a zmienna *y* jest zmienną objaśnianą. Jest ona zmienną zero-jedynkową (0 – niebankrut, 1 – bankrut). Dysponując danymi z macierzy (21), należy znaleźć taką relację pomiędzy zmienną *y* a zmiennymi $x_1, x_2, \ldots, x_M$, aby na podstawie znajomości wartości predykatorów można było określić wartość zmiennej *y*. Szuka się więc funkcji *f*, takiej że:

$$y = f(\mathbf{x}, \boldsymbol{\alpha}) + \boldsymbol{\varepsilon} \tag{22}$$

W praktyce rozpatruje się jedynie model addytywny o postaci:

$$y = \alpha_0 + \sum_{k=1}^{K} \alpha_k g_k(\mathbf{x}, \boldsymbol{\beta}) \tag{23}$$

Stosując metodę rekurencyjnego podziału, otrzymujemy aproksymację modelu (23) w postaci funkcji:

$$y = a_0 + \sum_{k=1}^{2} a_k I\ (\mathbf{x} \in R_k) \tag{24}$$

gdzie:

R_k (dla k – 1,2) – rozłączne segmenty (bankrut, niebankrut) w wielowymiarowej przestrzeni cech,

a_k – parametry modelu.

Drzewo klasyfikacyjne jest graficzną reprezentacją modelu (24)[*].

Model (24) nie jest tworzony globalnie, lecz poprzez złożenie modeli lokalnych, budowanych w każdym z K rozłącznych segmentów, na jakie dzielona jest wielowymiarowa przestrzeń zmiennych $\mathbf{X}^m$. Każdy z obszarów jest definiowany przez jego granice w przestrzeni cech i w przypadku, gdy zmienne $x_1, x_2, \ldots, x_M$ mają charakter ilościowy (są wskaźnikami finansowymi), można go przedstawić jako iloczyn:

$$I(\mathbf{x} \in R_k) = \prod_{m=1}^{M} I(v_{km}^{(d)} \leq x_m \leq v_{km}^{(g)}) \tag{25}$$

gdzie wartości $v_{km}^{(d)}$ oraz $v_{km}^{(g)}$ oznaczają odpowiednio górną i dolną granicę odcinka w *m*-tym wymiarze przestrzeni, a *I* jest funkcją wskaźnikową:

$$I(q) = \begin{cases} 1 \text{ gdy } q \text{ jest prawdziwe,} \\ 0 \text{ w przeciwnym przypadku} \end{cases} \tag{26}$$

Problem oddzielenia (odseparowania) obiektów należących do różnych klas jest rozważany na podstawie statystycznej analizy wielowymiarowej w ramach zagadnienia klasyfikacji. W tym celu konieczne jest posiadanie zbioru obiektów *S*, nazywanego zbiorem uczącym, w którym przynależność do określonej klasy jest znana, tj. wartość zmiennej *y* dla każdego z nich została poprawnie zaklasy-

[*] (Gatnar, Walesiak, 2004, s. 107).

fikowana. Celem analizy jest znalezienie charakterystyk klas, tak aby wykorzystać je do klasyfikacji zbioru rozpoznawanego.

Zmienna *y* w modelu (24) jest cechą nominalną, wobec tego model ten jest modelem klasyfikacyjnym, a parametry a_k tego modelu są wyznaczane zgodnie z zasadą majoryzacji:

$$a_k = \arg\max_i\{p(l|k)\} \tag{27}$$

gdzie $p(l|k)$ oznacza prawdopodobieństwo, że pewien obiekt z segmentu R_k należy do klasy *l*. Innymi słowy, formuła (27) mówi o tym, że w segmencie R_k zmienna *y* przyjmuje tę wartość *l*, która występuje najczęściej.

Model (24) jest konstruowany drogą krokowej procedury doboru zmiennych, a więc dzielenia przestrzeni $\mathbf{X}^m$ hiperpłaszczyznami równoległymi do osi poszczególnych zmiennych. Klasyczna procedura rekurencyjnego podziału składa się z następujących kroków (Gatnar, Walesiak, 2004, s. 108):

1. Mając przestrzeń $\mathbf{X}^m$ (w której znajduje się zbiór uczący S), sprawdź, czy jest ona jednorodna ze względu na wartości zmiennej zależnej *y* (lub spełniony został inny warunek stopu). Jeżeli tak, to zakończ pracę.
2. W przeciwnym razie rozważ wszystkie możliwe sposoby podziału przestrzeni $\mathbf{X}^m$ na rozłączne segmenty $R_1, \ldots, R_K$ (z uwzględnieniem wartości kolejno wybieranych zmiennych objaśniających).
3. Dokonaj oceny jakości każdego z tych podziałów zgodnie z przyjętym kryterium homogeniczności i wybierz najlepszy z nich.
4. Podziel przestrzeń $\mathbf{X}^m$ w wybrany sposób.
5. Wykonaj kroki 1-4 rekurencyjnie dla każdego z segmentów $R_1, \ldots, R_K$.
6. Procedura podziału się kończy, jeżeli zostało osiągnięte jedno z kryteriów stopu. Najczęściej jest nim jednorodność obiektów w segmencie $R_1, \ldots, R_K$ lub (ze względów praktycznych) określona minimalna liczba obiektów w uzyskanych segmentach (podzbiorach).

Przebieg procesu podziału rekurencyjnego najlepiej reprezentuje graf spójny bez cykli, czyli drzewo. Stąd wynika popularna nazwa tej metody – drzewa decyzyjne, a w sytuacji gdy zmienna *y* ma charakter jakościowy – drzewa klasyfikacyjne. Metody podziału rekurencyjnego oraz ich prezentacja w postaci drzew decyzyjnych są stosowane do wykrywania związków i relacji występujących w dużych zbiorach danych. Są więc one technikami stosunkowo młodej dyscypliny naukowej, jaką jest *Data Mining*[*].

[*] W literaturze polskiej wyczerpujące informacje dotyczące różnych wariantów metody drzew klasyfikacyjnych i regresyjnych znaleźć można m.in. w pracach: (Gatnar, 2001; Gatnar, 2008 lub Gatnar, Walesiak, 2004).

5. Wybrane przykłady zastosowań przedstawionych metod do przewidywania bankructwa

W niniejszym punkcie zostaną przytoczone pierwsze zastosowania na świecie prezentowanych powyżej metod jako narzędzi predykcji upadłości firm oraz reprezentatywne przykłady ich zastosowania w prognozowaniu bankructwa firm polskich.

Pierwszym, który w 1968 r. wykorzystał liniową funkcję dyskryminacyjną Fishera dla celów prognozowania bankructwa był Altman (1968). Wyodrębnił on dwie grupy firm amerykańskich: bankrutów i niebankrutów. Jego próba wynosiła 66 korporacji, spośród których 33 stanowiły przedsiębiorstwa, jakie zbankrutowały w latach 1946-1965, a 33 to wylosowane firmy działające na rynku amerykańskim, które nie zbankrutowały w badanym okresie. Źródłem danych były sprawozdania finansowe tych przedsiębiorstw (dla bankrutów z roku poprzedzającego bankructwo). Wstępnie rozpatrywał on 22 wskaźniki finansowe reprezentujące klasyczne ich grupy: płynność finansową, rentowność, wspomaganie finansowe, wypłacalność i aktywność ekonomiczną. Po wstępnej analizie ekonomicznej, opartej na doświadczeniu profesjonalnym audytorów, do modelu wybranych zostało 5 wskaźników finansowych.

Klasyczny model Altmana, nazwany przez autora „Z-Score model" jest następujący (Za: Pociecha, 2006, s. 206):

$$Z = 1{,}2\ X_1 + 1{,}4\ X_2 + 3{,}3\ X_3 + 0{,}6\ X_4 + 1{,}0\ X_5$$

gdzie:
X_1 – kapitał pracujący / majątek ogółem,
X_2 – zysk zatrzymany / majątek ogółem,
X_3 – zysk przed opodatkowaniem (EBIT) /majątek ogółem,
X_4 – wartość rynkowa kapitału akcyjnego / wartość księgowa zadłużenia,
X_5 – przychody ze sprzedaży / majątek ogółem.

Wartość krytyczna (progowa) wyniosła 2,675. Według badań Altmana zdolność poprawnej klasyfikacji w prezentowanym modelu wynosiła 95%, przy czym błąd klasyfikacji I typu (zaklasyfikowanie bankruta jako niebankruta) wynosił 6%, a błąd II typu (zaklasyfikowania niebankruta jako bankruta) – 3%. W rezultacie badań dotyczących minimalizacji prawdopodobieństwa błędnej klasyfikacji, Altman przyjął następujące wartości graniczne (Z) dla prognozowania bankructwa:
1,81 lub mniej – duże prawdopodobieństwo bankructwa firmy (strefa I – wszystkie firmy w próbie zbankrutowały),

3,00 lub więcej – niskie prawdopodobieństwo bankructwa firmy (strefa II – żadna firma nie zbankrutowała),
$1{,}81 < Z < 2{,}67$ – niepewna przyszłość firmy.

Model ten, mający już historyczne znaczenie, stał się punktem odniesienia do wszystkich późniejszy modeli prognozowania bankructwa. Altman później wielokrotnie modyfikował prezentowane przez siebie modele i testował je na różnych próbach firm*.

Pierwszym przykładem zastosowania metody dyskryminacyjnej do prognozowania bankructwa polskich firm jest praca Mączyńskiej (1994), natomiast za najbardziej rozbudowany przykład zastosowania wielowymiarowej analizy dyskryminacyjnej do prognozowania upadłości w polskich realiach gospodarczych można uznać pracę Hołdy (2006). Zostanie przedstawiony przykład liniowej funkcji dyskryminacyjnej zbudowanej i szacowanej przez autora niniejszego artykułu. Model został zbudowany na podstawie wskaźników płynności i rentowności.

W literaturze przyjmuje się, że zdolność klasyfikacyjna i predyktywna modelu dyskryminacyjnego jest tym lepsza, im więcej obszarów analizy finansowej reprezentują zmienne włączane do modelu. Punktem wyjścia do konstrukcji tego modelu był zbiór 13 zmiennych dyskryminujących. Ich dobór do modeli dyskryminacyjnych odbywał się drogą selekcji krokowej z poszczególnych grup wskaźników. W rezultacie otrzymano następujący model zbudowany na podstawie wskaźników płynności i rentowności:

$$Y_1 = -1{,}843 - 0{,}486\ \mathrm{ROE} + 1{,}116\ \mathrm{WKOP} - 0{,}500\ \mathrm{WPB}$$

gdzie:
ROE – rentowność kapitału własnego,
WKOP – wskaźnik kosztów operacyjnych,
WPB – wskaźnik płynności bieżącej.

Model ten charakteryzuje się 92,50-proc. zdolnością do poprawnej klasyfikacji oraz 86,25-proc. zdolnością predyktywną. W próbie podstawowej błąd I rodzaju wynosi 6,67% (4/56), a błąd II rodzaju – 8,33% (5/55). W próbie testowej błąd I rodzaju wynosi 22,50% (9/31), a błąd II rodzaju – 5,00% (2/38). Wystarczyło przyjęcie trzech wymienionych zmiennych dyskryminujących, gdyż włączenie któregokolwiek z pozostałych wskaźników płynności czy rentowności nie poprawiało sprawności modelu.

Pierwszym na świecie, który zastosował model logitowy do prognozowania bankructwa był Ohlson (1980). Zmienne do swojego modelu wybierał, opierając się na wskazaniach literatury z zakresu analizy finansowej. W efekcie uwzględ-

* Podsumowanie doświadczeń z budową i zastosowaniami zaproponowanych modeli zawarł on w pracy: Altman: Predicting Financial Distress of Companies: Revisiting the Z-Score and ZETA® Models, http://pages.stern.nyu.edu/~ealtman/Zscores.PDF.

nił w modelu 9 wskaźników finansowych. Jego próba składała się ze 105 bankrutów i 2058 dobrze funkcjonujących firm w latach 1970-1976. Model Ohlsona (w formie zlinearyzowanej) przedstawiał się następująco:

$$Y = -1{,}3 - 0{,}4\,X_1 + 6{,}0\,X_2 - 1{,}4\,X_3 + 0{,}1\,X_4 - 2{,}4\,X_5 - 1{,}8\,X_6 + 0{,}3\,X_7 - 1{,}7\,X_8 - 0{,}5\,X_9$$

gdzie:
X_1 – log (aktywa ogółem /ogólny indeks cen),
X_2 – zobowiązania ogółem/aktywa ogółem,
X_3 – kapitał pracujący/aktywa ogółem,
X_4 – zobowiązania bieżące/aktywa obrotowe,
X_5 = 1, gdy zobowiązania ogółem przekraczają aktywa ogółem, 0 w przeciwnym przypadku;
X_6 – zysk netto/aktywa ogółem,
X_7 – przychody/zobowiązania ogółem,
X_8 = 1, gdy firma przynosiła straty w ciągu ostatnich dwóch lat, w przeciwnym przypadku 0,
X_9 – miara zmian zysku netto.

Od tego momentu nastąpił gwałtowny rozwój zastosowań modelu logitowego dla celów prognozowania bankructwa. W Polsce jako pierwsze zastosowanie modelu logistycznego do predykcji upadłości polskich firm przedstawione zostało w pracy Hołdy (2000).

Spośród wielu publikacji z tego obszaru w Polsce należy zwrócić uwagę na wyniki uzyskane przez Wędzkiego (2005). Przedstawił on wiele modeli logitowych upadłości w gospodarce polskiej. Przykładowo zaprezentowano tutaj model M3 – wielogałęziowy model dla przedsiębiorstw przemysłowych:

$$Y = 1{,}0 - 5{,}0\,WB + 4{,}721\,UKON + 3{,}598\,WZO - 0{,}334\,WUO + 0{,}048\,IDF + 0{,}021\,CN + 0{,}061\,RIR$$

gdzie:
WB – bieżący wskaźnik płynności,
UKON – udział kapitału obrotowego w aktywach,
WZO – wskaźnik zadłużenia ogólnego,
WUO – wskaźnik udziału odsetek,
IDF – indeks dźwigni finansowej,
CN – cykl należności,
RIR – stopa zysku rezydualnego.
Próba przedsiębiorstw wynosiła: 40 bankrutów i 40 firm dobrze prosperujących.

Obszerny przegląd modeli regresji logistycznej oszacowanych dla celów prognozowania upadłości firm polskich jest zawarty m.in. w pracy Prusaka (2005).

Sieci neuronowe jako narzędzie prognozowania bankructwa były wykorzystywane w świecie od lat 90. ubiegłego wieku*. Szczególnie interesująca jest praca Odoma i Shardy (1990). Autorzy podjęli próbą zastosowania sieci neuronowych dla 5 wskaźników finansowych uwzględnionych w klasycznym, dyskryminacyjnym modelu Altmana prognozowania bankructwa. Posłużyli się oni danymi finansowymi, dotyczącymi 128 firm, pochodzących z okresu na rok przed bankructwem. Zdolność do poprawnej klasyfikacji przez model neuronowy dla firm „bankrutów" mieściła się w granicach 77,8%-81,5%, a dla firm „niebankrutów" – 78,6%-58,7%.

Jako pierwszy w Polsce propozycję zastosowania sztucznych sieci neuronowych dla celów prognozowania bankructwa polskich firm przedstawił Michaluk (2000). Przyjął on zasadę, że liczba neuronów w warstwie ukrytej nie może być większa od liczby neuronów w warstwie wejściowej. Do swojego badania wykorzystał dane dotyczące 259 przedsiębiorstw, w tym 79 bankrutów oraz 180 firm o dobrej kondycji finansowej. Rozpatrywał on 12 podstawowych wskaźników finansowych, charakteryzujących kondycję finansową przedsiębiorstw. Podstawowe parametry konstrukcji tej sieci były następujące:

- 12 neuronów w warstwie wejściowej (wskaźniki finansowe $X_1, \ldots, X_{12}$),
- 12 neuronów w warstwie ukrytej,
- 1 neuron wyjściowy (RB – bankrut = 0; NBR – niebankrut = 1),
- sieć była uczona algorytmem wstecznej propagacji błędu.

Autor sieci przyjął trzy próby, różniące się ze względu na proporcje: niebankrut/bankrut, przyjmując odpowiednio: proporcje 1:1 (próba ucząca 40/40, próba testowa 39/39), 3:1 oraz 10:1 (próba ucząca 100/10; próba testowa 69/69). Zdolność do poprawnej klasyfikacji nie była całkiem zadowalająca, gdyż najmniejszy błąd I typu (procent bankrutów, które zostały nieprawidłowo zakwalifikowane do zbioru firm kontynuujących działalność) wynosił 10%, natomiast błąd drugiego typu (procent firm które kontynuują działalność, nieprawidłowo zakwalifikowanych do zbioru bankrutów) – 2%. Ogólna zdolność do poprawnej klasyfikacji (procent prawidłowo zaklasyfikowanych firm) była wysoka i wynosiła 97,27%, jednak generowała stosunkowo wysokie błędy I typu. W praktyce przyjmuje się, że koszty, jakie są ponoszone przez błędy I typu (uznanie firm, które później zbankrutowały za firmy zdrowe) są o wiele wyższe niż koszty błędów II typu.

Bogate wyniki badań prognostycznych wykonanych przy użyciu sieci neuronowych zaprezentowane zostały w pracy Korola i Prusaka (2005). Do kon-

* Pierwsze prace z tego zakresu to: (Bell, Ribar, Verchio, 1990 oraz Odom, Sharda, 1990).

strukcji modelu sieci neuronowej wykorzystano dane finansowe dotyczące 180 polskich przedsiębiorstw produkcyjnych z lat 1998-2001. Próba ucząca składała się z 39 firm będących w dobrej kondycji oraz 39 firm zagrożonych upadłością. Firmy znajdujące się w dobrej sytuacji wybierano drogą parowania z potencjalnymi bankrutami, tak aby one pochodziły zawsze z tej samej branży. Próby uczące konstruowano dla danych na rok i na dwa lata przed złożeniem wniosku o upadłość. Próba testowa również składała się z 39 przedsiębiorstw zdrowych oraz 39 bankrutów. Do budowy sieci neuronowej autorzy przyjęli 27 wskaźników finansowych charakteryzujących płynność finansową, zadłużenie, sprawność działania i rentowność firm. Dodatkowo autorzy wprowadzili zmienną pozafinansową, jaką był region działania danej spółki. Wykorzystano tutaj mapę intensywności zagrożenia upadłością firm w poszczególnych województwach. Sieć miała więc pierwotnie 28 sygnałów wejściowych, których zbiór autorzy oznaczyli jako K1. Alternatywnie zaproponowano redukcję sygnałów wejściowych (wskaźników finansowych) na podstawie wyników analizy współczynników korelacji pomiędzy przyjętymi pierwotnie cechami. Doprowadziło to do redukcji sygnałów wejściowych do czterech. Zbiór ten oznaczono jako K2. W trzecim wariancie (K3) przyjęto pięć wskaźników z modelu Altmana. Przyjmowano także wiele wariantów architektury sieci, różniących się liczbą neuronów w warstwie ukrytej oraz dla prób uczących i testowych na rok oraz na dwa lata przed upadłością. Rozważano również różne proporcje niebankrut/bankrut w strukturze prób.

Wyniki badań potwierdziły, że „zachodnie" wskaźniki finansowe, jakie były przyjmowane m.in. w modelu Altmana, nie sprawdzają się w warunkach gospodarki polskiej, gdyż podejście K3 dawało najgorsze rezultaty. Natomiast efektywność prognostyczną pozostałych modeli badano na próbach testowych. Skuteczność prognostyczna (procent prawidłowo zaklasyfikowanych firm w zbiorze testowym) była następująca: w wariancie K1 na rok przed upadkiem firmy wynosiła 98,72%, a na dwa lata przed upadkiem – 91,03%. W wariancie K2 na rok przed upadkiem wynosiła 97,44% a na dwa lata przed upadkiem – 87,18%. Jak wynika z powyższego, najlepsze rezultaty otrzymano dla najbardziej rozbudowanej sieci neuronowej.

Metodologia drzew klasyfikacyjnych bardzo wcześnie została zastosowana dla celów prognozowania bankructwa. Pierwszą pracą z tego zakresu był artykuł Frydman, Altmana i Kao (1985). Autorzy zastosowali algorytm RPA *(Recursive Partitioning Algorithm)* zaproponowany przez Breimana i in. (1984) do badania zdolności kredytowej i upadłości firm. Jedno z przedstawionych tam drzew klasyfikacyjnych stworzonych przez Frydman, Altmana i Kao charakteryzowało się wysoką zdolnością do poprawnej klasyfikacji ogółem, wynoszącą 94,00%. Praca

Frydman Altmana i Kao stała się później punktem odniesienia do badania efektywności innych metod prognozowania bankructwa (Por. Barniv, McDonald, 1999).

Pierwszym, który w badaniach upadłości polskich firm zastosował metodologię drzew decyzyjnych w wersji algorytmu CART, był Hołda (2006). Zbudował on drzewo decyzyjne dla firm branży produkcyjnej oraz porównał zdolność do poprawnej klasyfikacji oraz zdolność prognostyczną zbudowanego drzewa klasyfikacyjnego z siecią neuronową zbudowaną dla tego samego zestawu wskaźników finansowych. Zdolność do poprawnej klasyfikacji (na zbiorze uczącym) wynosiła 93,24%, a na zbiorze testowym – 83,87%. Zbudowana dla tych samych wskaźników finansowych sieć neuronowa miała następującą efektywność: zdolność do poprawnej klasyfikacji na zbiorze uczącym wynosiła 92,57% a na zbiorze testowym – 77,42%. Drzewo klasyfikacyjne okazało się więc nieco bardziej efektywnym narzędziem przewidywania bankructwa niż sieć neuronowa, dla tego samego zestawu wskaźników finansowych.

Drzewa klasyfikacyjne upadłości firm przynależnych do innych branż, budowane metodą CART, zostały przedstawione w pracy Hołdy i Pociechy (2009). Dla firm branży budowlanej istotnymi okazały się dwa wskaźniki:

- rentowności – uwzględniający zysk/stratę netto zwiększoną o niepodzielny wynik z lat poprzednich oraz podatek dochodowy w stosunku do majątku ogółem,
- przepływów pieniężnych z działalności inwestycyjnej (środki pieniężne netto z działalności inwestycyjnej/ majątek ogółem).

Zdolność do poprawnej klasyfikacji tego drzewa była bardzo wysoka i wynosiła 93,75%, ale zdolność prognostyczna na grupie testowej była dość niska i wynosiła 65,00%.

Dla branży handlowo-usługowej drzewo klasyfikacyjne obejmowało następujące wskaźniki finansowe:

- zadłużenia (zysk/strata netto w stosunku do zobowiązań krótkookresowych),
- rentowności – uwzględniający zysk/stratę netto zwiększoną o niepodzielny wynik z lat poprzednich oraz podatek dochodowy w stosunku do majątku ogółem,
- rentowności – uwzględniający zysk/stratę netto w stosunku do kosztów ogółu działalności,
- zadłużenia – jako stosunek zobowiązań oraz rezerw do sumy bilansowej.

Zdolność do poprawnej klasyfikacji firm do grupy bankrutów i niebankrutów była wysoka i wynosiła 91,76%. Wyraźnie niższa była jednak zdolność prognostyczna drzewa na zbiorze testowym – wynosiła 67,86%.

Jak widać z przytoczonych przykładów, metody statystyki wielowymiarowej, a także komputerowe metody iteracyjne są często wykorzystywane w praktyce badań ekonomicznych nad problemami upadłości firm w Polsce i na świecie.

6. Uwarunkowania efektywności metod klasyfikacyjnych w prognozowaniu bankructwa

Z przedstawionego w poprzednim punkcie przeglądu wynika, że różni autorzy preferowali odmienne podejścia metodologiczne do prognozowania bankructwa. W związku z tym rodzi się pytanie, czy istnieją metody, które z reguły dają bardziej precyzyjne prognozy bankructwa niż inne. Z szerokich badań porównawczych przeprowadzonych na świecie (Patrz np. Bellovary, Giacomino, Akers, 2007), a także przeglądów prowadzonych przez autora (Pociecha, 2010a lub Pociecha, Pawełek, 2011) wynika, że nie można mówić o metodach „lepszych" lub „gorszych" dla celów prognozowania upadłości firm.

Jako kryterium sprawności prognostycznej metod wymienionych w poprzednim punkcie przyjmuje się następujące miary zdolności do poprawnej klasyfikacji:

1. Sprawność I rodzaju – udział (procent) firm, które zbankrutowały, prawidłowo zakwalifikowanych przez model do zbioru bankrutów.
2. Błąd I rodzaju – procent bankrutów, które zostały nieprawidłowo zakwalifikowane do zbioru firm kontynuujących działalność.
3. Sprawność II rodzaju – procent firm kontynuujących swoją działalność (niebankrutów), prawidłowo rozpoznanych przez model.
4. Błąd II rodzaju – procent firm które kontynuują działalność, nieprawidłowo zakwalifikowanych do zbioru bankrutów.
5. Sprawność ogólna – procent prawidłowo zaklasyfikowanych firm,
6. Błąd ogólny – procent nieprawidłowo zaklasyfikowanych firm.

Najpopularniejsze polskie modele przewidywania upadłości firm (Pociecha, 2010, s. 56-57) miały na zbiorze uczącym sprawność ogólną dla modeli dyskryminacyjnych w granicach 78,6%-93,2%, dla modeli logitowych: 89,0%-91,9%, dla sieci neuronowych: 93,9%-96,2%, dla drzew klasyfikacyjnych: 91,8%-93,8%. W tym przeglądzie widać, że nieco bardziej precyzyjne wyniki otrzymywano z sieci neuronowych, lecz w literaturze podawano także przykłady, gdzie liniowa funkcja dyskryminacyjna może dawać lepsze rezultaty niż zbyt skomplikowana sieć neuronowa. Zdolność prognostyczną bada się na zbiorze testowym, który na ogół jest wyodrębnioną częścią pierwotnego zbioru danych. Tutaj też nie stwierdzono wyraźnych różnic w zdolności prognostycznej omawianych typów modeli. Reasumując, można stwierdzić, że precyzja prognozy bankructwa nie zależy od typu modelu prognostycznego.

Wobec braku rozstrzygnięć, jakiego typu modele są najodpowiedniejsze dla celów prognozowania bankructwa, należy sobie postawić fundamentalne pyta-

nie: Jakie są źródła błędów popełnianych w procesie prognozowania bankructwa? (Pociecha, 2011, s. 128-129).

Jednym z nich jest wartościowy charakter wskaźników finansowych. Istnieją krajowe i międzynarodowe standardy sprawozdawczości finansowej, ale daleko jeszcze do ujednolicenia sposobu pomiaru wielkości finansowych, szczególnie w skali międzynarodowej. Precyzja pomiaru wskaźników finansowych jako zmiennych klasyfikujących do zbioru bankrutów lub niebankrutów nie jest więc zbyt wysoka.

Drugie możliwe źródło błędów to metoda doboru prób. W klasycznym ujęciu np. metody dyskryminacyjnej, próby z badanych populacji są wybierane drogą losową. W praktyce doboru prób nie przeprowadza się w sposób nielosowy. Uwzględnia się na ogół wszystkie firmy upadłe w badanym okresie, a do niej nielosową metodą parowania dobiera się przedsiębiorstwa dobrze funkcjonujące. Nie można więc mówić o doborze losowym, w sensie klasycznym, a więc także o błędzie próbkowania. Testowany błąd klasyfikacji nie wynika z tego, że operujemy próbami losowymi.

Istotnym elementem błędów w prognozowaniu bankructwa jest tzw. bankructwo z przyczyn strategicznych. Zarządcy lub właściciele firmy dobrze prosperującej mogą celowo doprowadzić firmę do bankructwa, wyprowadzając nieco wcześniej jej aktywa np. do „rajów podatkowych". Żaden model predykcji bankructwa nie uwzględnia celowego działania zarządców firm w kierunku celowego doprowadzenia do bankructwa.

Kolejnym źródłem błędów jest niestabilny charakter badanych populacji. Populacje bankrutów i przedsiębiorstw dobrze funkcjonujących w sytuacji koniunktury gospodarczej nie są identyczne z tymi populacjami w okresie kryzysu gospodarczego. Błąd prognozy może więc zależeć od tego, że model zbudowany został dla danych z okresu koniunktury, a prognoza budowana jest dla firmy w okresie recesji.

Podsumowując, warto jeszcze raz zwrócić uwagę na pionierski charakter pracy prof. J. Kolonki (1980), która w środowisku polskich statystyków przez lata motywowała do rozwijania teorii i zastosowań metod klasyfikacji danych, czego wyrazem jest ich użyteczność m.in. dla celów prognozowania upadłości firm.

Bibliografia

Altman E.I. (1968): *Financial Ratios, Discriminant Analysis and Prediction of Corporate Bankruptcy*. „The Journal of Finance", Vol. 23, September.

Altman E.I. (2000): *Predicting Financial Distress of Companies*: Revisiting the Z-Score and ZETA® Models,, http://pages.stern.nyu.edu/~ealtman/Zscores.PDF.

Aziz M.A., Dar H.A. (2004): *Predicting Corporate Bankruptcy, Whither do We Stand?* 3rd Annual Meeting of the European Economics and Finance Society "Word Economy and European Integration", University of Gdańsk, 13-16 May.

Barniv R, McDonald J.B. (1999): *Review of Categorical Models for Classification Issues In Accounting and Finance*. „Review of Quantitative Finance and Accounting", 13.

Bell T.B., Ribar G.S., Verchio J. (1990): *Neural Nets Versus logistic Regression: A Comparison of Each Model's Ability to Predict Commercial Bank Failures*. In: *Proceedings of the 1990 Deloitte and Touché/University of Kansas Symposium of Auditing Problems*. Ed. R.P. Srivastava.

Bellovary J., Giacomino D., Akers M. (2007): *A Review of Bankruptcy Prediction Studies: 1930 to Present*. „Journal of Financial Education", Vol. 33, Winter.

Breiman L., Friedman J., Olshen R., Stone C. (1984): *Classification and regression trees*. CRC Press, London.

Christensen R. (1991): *Linear Models for Multivariate, Time Series, and Spatial Data*. Springer, New York.

Fisher R.A. (1936): *The Use of Multiple Measurements in Taxonomic Problems*. „Annals of Eugenics", 7.

Frydman H., Altman E.I., Kao D. (1985): *Introducing Recursive Partitioning for Financial Classification: The Case of Financial Distress*. „Journal of Finance", Vol. 40.1.

Gatnar E. (2001): *Nieparametryczna metoda dyskryminacji i regresji*. Wydawnictwo Naukowe PWN, Warszawa.

Gatnar E. (2008): *Podejście wielomodelowe w zagadnieniach dyskryminacji i regresji*. Wydawnictwo Naukowe PWN, Warszawa.

Gatnar E., Walesiak M. (2004): *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych,* Wyd. AE, Wrocław.

Giri N.C. (1996): *Multivariate Statistical Analysis*. Dekker, New York.

Hołda A. (2000): *Optymalizacja i model zastosowania procedur analitycznych w rewizji sprawozdań finansowych*. Praca doktorska, Akademia Ekonomiczna, Kraków.

Hołda A. (2006): *Zasada kontynuacji działalności i prognozowanie upadłości w polskich realiach gospodarczych*. Wydawnictwo Akademii Ekonomicznej, seria specjalna nr 174, Kraków.

Hołda A., Pociecha J. (2009): *Probabilistyczne metody badania sprawozdań finansowych*. Wydawnictwo Uniwersytetu Ekonomicznego, Kraków.

Kolonko J. (1980): *Analiza dyskryminacyjna i jej zastosowania w ekonomii*. PWN, Warszawa.

Korbicz J., Obuchowicz A., Uciński D. (1994): *Sztuczne sieci neuronowe. Podstawy i zastosowania*. Akademicka Oficyna Wydawnicza PLJ, Warszawa.

Korol T., Prusak B. (2005): *Upadłość przedsiębiorstw a wykorzystanie sztucznej inteligencji*. CeDeWu, Warszawa.

Lula P. (1999): *Jednokierunkowe sieci neuronowe w modelowaniu zjawisk ekonomicznych*. Wydawnictwo AE, Kraków.

Mączyńska E. (1994): *Ocena kondycji przedsiębiorstwa (Uproszczone metody)*. „Życie gospodarcze", nr 38.

McCulloch C.E., Searle S.R., Neuhaus J.M. (2009): *Generalized, Linear, and Mixed Models*. Wiley, New York.

Michaluk K. (2000): *Efektywność modeli prognozujących upadłość przedsiębiorstw*. Praca doktorska, Uniwersytet Szczeciński, Szczecin.

Odom M.D., Sharda R. (1990): *A Neural Network Model for Bankruptcy Prediction*. Proceedings of IEEE International Conference on Neural Networks, Vol. 2, San Diego.

Ohlson J. (1980): *Financial Ratios and the Probabilistic Prediction of Bankruptcy*. „Journal of Accounting Research", Spring.

Ossowski S. (1996): *Sieci neuronowe w ujęciu algorytmicznym*. WNT, Warszawa.

Pociecha J. (2006): *Funkcja dyskryminacyjna jako narzędzie prognozowania bankructwa firmy – metoda oraz rezultaty praktyczne*. W. *Matematyka język uniwersalny*. Wyd. AE, Kraków.

Pociecha J. (2007): *Problemy prognozowania bankructwa firmy metodą analizy dyskryminacyjnej*. „Acta Universitatis Lodziensis, Folia Oeconomica", nr 205.

Pociecha J. (2010a): *Metodologiczne problemy prognozowania bankructwa*. „Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu", „Taksonomia 17, Klasyfikacja i analiza danych – teoria i zastosowania", Wrocław.

Pociecha J. (2010b): *Zastosowania sztucznych sieci neuronowych w prognozowaniu bankructwa firm*. W: *Nauki ekonomiczne wobec wyzwań współczesnej gospodarki światowej*. Uniwersytet Ekonomiczny, Kraków.

Pociecha J. (2011): *Modele prognozowania bankructwa w systemie wczesnego ostrzegania przedsiębiorstw*. W: *Społeczna rola statystyki*. Red. W. Ostasiewicz. „Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu", nr 165, Wrocław.

Pociecha J. (2012): *Model logitowy jako narzędzie prognozowania bankructwa. Jego zalety i ograniczenia*. W: *Spotkania z królową nauk*. Wydawnictwo Uniwersytetu Ekonomicznego, Kraków.

Pociecha J., Pawełek B. (2011): *Prognozowanie bankructwa a koniunktura gospodarcza*. „Metody analizy danych", Zeszyty Naukowe Uniwersytetu Ekonomicznego w Krakowie, nr 873, Kraków.

Prusak B. (2005): *Nowoczesne metody prognozowania zagrożenia finansowego przedsiębiorstwa*. Difin, Warszawa.

Rencher A.C. (1998): *Multivariate Statistical Inference and Applications*. Wiley, New York.

Tadeusiewicz R. (1993): *Sieci neuronowe*. Wydawnictwo Naukowe PWN, Warszawa.

Wędzki D. (2005): *Bankruptcy Logit Model for Polish Economy*. „Argumenta Oeconomica Cracoviensia", nr 3.

Witkowska D. (2002): *Sztuczne sieci neuronowe i metody statystyczne. Wybrane zagadnienia finansowe*. C.H. Beck, Warszawa.

SELECTED CLASSIFICATION METHODS AND THEIR EFFECTIVENESS IN FIRMS' COLLAPSING PREDICTION

Summary

Classification methods are recognised as useful tool for bankruptcy prediction. Among them the most popular are: linear discriminant function, Logit model, neural network and classification tree. The ideas and basic formulas of these methods are presented in the paper. Some examples of application those procedures, which were published in world and Polish literature, are mentioned in the following parts of the paper. Some effectiveness conditions of presented methods are discussed. In the conclusion it has been stressed, that the precision of bankruptcy prediction not strictly depend on classification method which has been used. Sources of errors in bankruptcy prediction has been discussed on the end of the paper. Among them important are: valuated character of financial ratios, as an impute variables in such models, problems in samples selection, which usually hasn't random character and unstable character of considered populations. Probability of firms' collapse strongly depends on the stage of business cycle.

Dorota Rozmus
Uniwersytet Ekonomiczny w Katowicach

PORÓWNANIE STABILNOŚCI TAKSONOMII SPEKTRALNEJ ORAZ ZAGREGOWANYCH ALGORYTMÓW TAKSONOMICZNYCH

Wprowadzenie

Stosując metody taksonomiczne w jakimkolwiek zagadnieniu klasyfikacji, ważną kwestią jest zapewnienie wysokiej poprawności wyników grupowania. Od niej bowiem zależeć będzie skuteczność wszelkich decyzji podjętych na ich podstawie. Stąd też w literaturze wciąż proponowane są nowe rozwiązania, które mają przynieść poprawę dokładności grupowania w stosunku do tradycyjnych metod (np. *k*-średnich, metod hierarchicznych). Przykładem mogą tu być metody polegające na zastosowaniu podejścia zagregowanego oraz algorytmy spektralne.

Podejście zagregowane w taksonomii można sformułować następująco: mając wyniki wielokrotnie przeprowadzonego grupowania, należy znaleźć zagregowany podział ostateczny. Taksonomia spektralna natomiast polega na zastosowaniu wartości własnych pochodzących ze spektralnej dekompozycji macierzy podobieństwa, opisującej badane obiekty.

Pożądaną cechą algorytmu taksonomicznego jest, by był on odporny na niewielkie zmiany w zbiorze danych czy też w wartościach parametrów tych metod (np. losowo wybierane zalążki skupień w metodzie *k*-średnich). Wyniki empiryczne pokazują, że podejście zagregowane jest stabilniejsze niż klasyczne metody taksonomiczne. Celem tego artykułu natomiast jest porównanie stabilności zagregowanych i spektralnych algorytmów taksonomicznych.

1. Taksonomia zagregowana

Idea podejścia zagregowanego pojawiła się w taksonomii w ostatnich latach jako próba przeniesienia koncepcji podejścia wielomodelowego z zagadnień dyskryminacji i regresji. Zasadniczo podejście zagregowane w taksonomii pole-

ga na połączeniu wyników wielokrotnie przeprowadzonego grupowania i ma za zadanie przede wszystkim podnieść dokładność rozpoznawania rzeczywistej struktury klas, zwiększyć odporność oraz zmniejszyć zmienność wyników grupowania (Fern, Brodley, 2003; Fred, 2002; Fred, Jain, 2002; Kuncheva, Vetrov, 2006; Strehl, Gosh, 2002). W ostatnich latach liczne badania w tej dziedzinie ugruntowały już nowy obszar w tradycyjnej taksonomii. Istnieje wiele możliwości zastosowania idei podejścia zagregowanego w taksonomii, wśród których do najważniejszych należy zaliczyć:

a) łączenie wyników grupowania uzyskanych za pomocą różnych metod;
b) uzyskanie różniących się między sobą podziałów z zastosowaniem różnych podzbiorów danych, np. poprzez losowanie bootstrapowe;
c) zastosowanie różnych podzbiorów zmiennych (łącznych lub rozłącznych);
d) wielokrotne zastosowanie określonego algorytmu z różnymi wartościami parametrów lub punktami startowymi (np. losowo wybranymi zalążkami skupień w metodzie *k*-średnich).

Do ciekawych propozycji metod agregacji w dziedzinie taksonomii należy zaliczyć algorytmy oparte na idei metody *bagging* przedstawione przez Leischa (1999), Duidoit i Fridlyand (2003), Hornika (2005) oraz podejście oparte na macierzy współwystąpień przedstawione przez Fred i Jain (2002).

Metoda *bagging* w taksonomii jest pewną ogólną ideą, w ramach której narodziło się kilka szczegółowych rozwiązań. Została ona zaczerpnięta z dyskryminacji (Breiman, 1996) i generalnie polega na losowaniu *B* prób bootstrapowych i dokonywaniu ich grupowania w celu uzyskania podziałów składowych, które będą łączone. Różnice w poszczególnych rozwiązaniach w taksonomii polegają na zastosowaniu różnych operatorów agregacji.

1.1. Metoda *bagging* w taksonomii – propozycja Leischa

Leisch (1999) zaproponował połączenie metod iteracyjno-optymalizacyjnych z hierarchicznymi Na podstawie każdej próby bootstrapowej określane są rezultaty grupowania przy zastosowaniu tzw. bazowej metody taksonomicznej, którą jest jedna z metod iteracyjno-optymalizacyjnych, np. metoda *k*-średnich. W kolejnym etapie ostateczne centra skupień przekształcane są w nowy zbiór danych, który poddawany jest podziałowi za pomocą metod hierarchicznych.

Algorytm zaproponowany przez Leischa przebiega w następujących krokach:

1. Z pierwotnego *n*-elementowego zbioru $G = \{x_1,...,x_n\}$ należy wylosować *B* prób bootstrapowch $G_n^1, G_n^2,..., G_n^B$, losując *n* obserwacji przy wykorzystaniu schematu losowania ze zwracaniem.
2. Na podstawie każdego podzbioru za pomocą metod iteracyjno-optymalizacyjnych (np. *k*-średnich) dokonuje się podziału na grupy obserwacji podobnych do siebie,

uzyskując w ten sposób $B \times K$ zalążków skupień $c_{11}, c_{12}, ..., c_{1K}, c_{21}, ..., c_{BK}$, gdzie K oznacza liczbę skupień w metodzie bazowej, a c_{bk} jest k-tym zalążkiem znalezionym na podstawie podpróby G_n^b.

3. Niech zalążki skupień uzyskane na podstawie kolejnych prób bootstrapowych utworzą nowy zbiór danych $C^B(K) = \{c_{11}, ..., c_{BK}\}$.
4. Do tak skonstruowanego zbioru należy zastosować hierarchiczną metodę taksonomiczną, uzyskując w ten sposób dendrogram.
5. Niech $c(x_i)$ oznacza zalążek skupienia znajdujący się najbliżej obserwacji x_i, $i = 1, \ldots, n$. Podział na grupy pierwotnego zbioru danych określany jest w ten sposób, że dendrogram uzyskany na podstawie zbioru $C^B(K)$ jest cięty na określonym przez badacza poziomie, co prowadzi do uzyskania grup obiektów podobnych $C_1^B, ..., C_m^B$, gdzie $1 \le m \le BK$. Każda obserwacja x_i z pierwotnego zbioru danych G jest przydzielana do tej grupy, w której znajduje się najbliżej leżący zalążek $c(x_i)$.

1.2. Metoda *bagging* w taksonomii – propozycja Dudoit i Fridlyand

Metoda *bagging* w wersji zaproponowanej przez Dudoit i Fridlyand (2003) wykorzystuje algorytm iteracyjno-optymalizacyjny do oryginalnego zbioru danych i do poszczególnych prób bootstrapowych. Następnie, po dokonaniu permutacji etykiet grup w poszczególnych podpróbach, tak by zachodziła jak największa zbieżność z podziałem obiektów z oryginalnego zbioru danych, stosuje głosowanie majoryzacyjne w celu określenia ostatecznego grupowania zagregowanego.

Kroki zaproponowanego przez nich algorytmu można ująć nastepująco.

Dla założonej liczby klas K:

1. Zastosuj iteracyjno-optymalizacyjny algorytm taksonomiczny T do pierwotnego zbioru danych $G = \{x_1, ..., x_n\}$, uzyskując w ten sposób etykiety klas $T(x_i, G) = \hat{y}_i$ dla każdej obserwacji x_i, $i = 1, \ldots, n$.
2. Skonstruuj b-tą próbę bootstrapową $G_n^b = \{x_1^b, ..., x_n^b\}$.
3. Zastosuj metodę taksonomiczną T do skonstruowanej próby bootstrapowej G_n^b, uzyskując podział na klasy: $T(x_i^b, G_n^b)$ dla każdej obserwacji w zbiorze G_n^b.
4. Dokonaj permutacji etykiet klas przyznanych obserwacjom w próbie bootstrapowej G_n^b, tak by zachodziła jak największa zbieżność z podziałem obiektów z oryginalnego zbioru danych G. Niech PR_K oznacza zbiór wszystkich permutacji zbioru liczb całkowitych *1*, … , *K*. Znajdź permutację $\tau^b \in PR_K$ maksymalizującą:

$$\sum_{i=1}^{n} I(\tau(T(x_i^b, G_n^b)) = T(x_i^b, G)), \tag{1}$$

gdzie $I(\cdot)$ to funkcja wskaźnikowa, równa 1, gdy zachodzi prawda, natomiast 0 w przypadku przeciwnym.

5. Powtórz kroki 2-4 B razy. Ostatecznie zaklasyfikuj i-tą obserwację, stosując głosowanie majoryzacyjne, zatem przydzielając ją do tej grupy, dla której zachodzi:

$$\arg\max_{1\le k\le K} \sum_{b:x_i\in G_n^b} I(\tau^b(T(x_i, G_n^b)) = k). \tag{2}$$

1.3 Metoda *bagging* w taksonomii – propozycja Hornika

W metodzie tej (Hornik, 2005) po skonstruowaniu B prób bootstrapowych i zastosowaniu do nich algorytmu taksonomicznego, uzyskuje się podziały składowe. Grupowanie zagregowane natomiast jest uzyskiwane za pomocą tzw. podejścia optymalizacyjnego, które ma za zadanie zminimalizować funkcję o postaci:

$$\sum_{b=1}^{B} dist(c, c_b)^2 \Rightarrow \min_{c\in C}, \tag{3}$$

gdzie:
C – zbiór wszystkich możliwych podziałów zagregowanych,
dist – odległość Euklidesowa,
$(c_1, ..., c_B)$ – grupowania wchodzące w skład podziału zagregowanego.

1.4. Podejście zagregowane oparte na macierzy współwystąpień

Innym rozwiązaniem jest zaproponowana przez Fred i Jain (2002) idea łączenia wyników wielokrotnie dokonanego grupowania w celu konstrukcji macierzy współwystąpień. Biorąc pod uwagę wystąpienie pary obiektów w tej samej grupie jako wskazówkę istnienia związku między nimi, wyniki wielokrotnie przeprowadzonego podziału są przekształcane w $n \times n$-wymiarową macierz opisującą podobieństwo między obiektami. W dalszym kroku macierz ta może zostać potraktowana bądź jako macierz odległości, która jest podstawą do przeprowadzenia grupowania (np. za pomocą metod hierarchicznych), albo może też zostać potraktowana jako macierz opisująca zbiór danych.

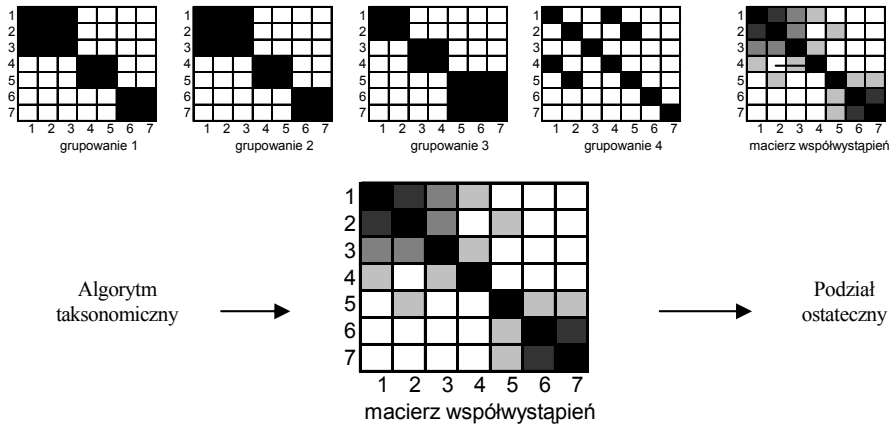
Szczegółowo kroki służące konstrukcji macierzy współwystąpień mogą zostać sformułowane następująco:

1. Wielokrotna klasyfikacja. Dla założonej liczby składowych S macierzy współwystąpień dokonaj grupowania obiektów np. za pomocą metody k-średnich, uzyskując różniące się miedzy sobą rezultaty dzięki losowo wybranym zalążkom skupień.
2. Agregacja. U podstaw tego podejścia leży założenie, że obiekty należące do tej samej grupy najprawdopodobniej będą lokowane w tej samej grupie wśród tych S podziałów. Biorąc zatem współwystąpienie pary obiektów w tej samej grupie jako wskazówkę istnienia związku między nimi, wyniki podziałów uzyskane dzięki wielokrotnie zastosowanej metodzie k-średnich przekształcane są w macierz współwystąpień o wymiarach $n \times n$:

$$co\_assoc(a,b) = votes_{ab}, \tag{4}$$

gdzie $votes_{ab}$ zlicza, ile razy para obiektów a i b zaliczona została do tej samej grupy, wśród tych S składowych podziałów.
3. Ostateczny podział. W celu określenia podziału ostatecznego zastosuj dowolny algorytm taksonomiczny do skonstruowanej macierzy współwystąpień.

Konstrukcja macierzy współwystąpień jest zilustrowana na rys. 1.



Rys. 1. Konstrukcja macierzy współwystąpień i ostateczny podział

2. Taksonomia spektralna

Taksonomia spektralna polega na zastosowaniu wartości własnych pochodzących ze spektralnej dekompozycji macierzy podobieństwa opisującej badane obiekty. Następnie największe wartości własne oraz odpowiadające im wektory własne są wykorzystywane do ostatecznego podziału obserwacji. W literaturze zaproponowano kilka metod spektralnych, a w każdej z nich w nieco inny spo-

sób stosuje się wektory własne (Kannan et al., 2004; Ng et al., 2001; Shi i Malik, 2000). W niniejszym badaniu zastosowana zostanie metoda zaproponowana przez Ng et al. (2001).

Dany jest zbiór obserwacji $G = \{x_1,...,x_n\}$ w przestrzeni R^l, który należy podzielić na *k* grup:

1. Skonstruuj macierz podobieństwa (ang. *affinity matrix*) $A \in R^{n \times n}$, której elementy są zdefiniowane jako:
$$A_{ij} = \exp\left(-\left\|x_i - x_j\right\|^2 / 2\sigma^2\right), \tag{5}$$
gdy $i \neq j$ oraz $A_{ii} = 0$. σ to parametr skalujący dobierany przez badacza.
2. Zdefiniuj ***D*** jako macierz diagonalną, której element (*i*, *i*) jest sumą *i*-tego wiersza macierzy ***A*** i na jej podstawie skonstruuj macierz:
$$\mathbf{L} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}. \tag{6}$$
3. Znajdź *k* pierwszych wektorów własnych $(\mathbf{z_1}, \mathbf{z_2}, ..., \mathbf{z_k})$ macierzy ***L*** i zestawiając je w kolumny, skonstruuj macierz:
$$\mathbf{Z} = \left[\mathbf{z_1}, ..., \mathbf{z_k}\right] \in R^{n \times k}. \tag{7}$$
4. Skonstruuj macierz ***Y*** poprzez normalizację każdego wiersza macierzy ***Z*** tak, by miały jednakową długość, tj.:
$$y_{ij} = z_{ij} / \left(\sum\nolimits_j z_{ij}^2\right)^{1/2}. \tag{8}$$
5. Traktując każdy wiersz macierzy ***Y*** jako punkt w przestrzeni R^k, podziel je na *k* grup z zastosowaniem metody *k*-średnich (lub innej).
6. Ostatecznie przydziel każdą pierwotną obserwację x_i do *j*-tej grupy wtedy i tylko wtedy, gdy *i*-ty wiersz macierzy ***Y*** został przydzielony do *j*-tej grupy.

3. Badania empiryczne

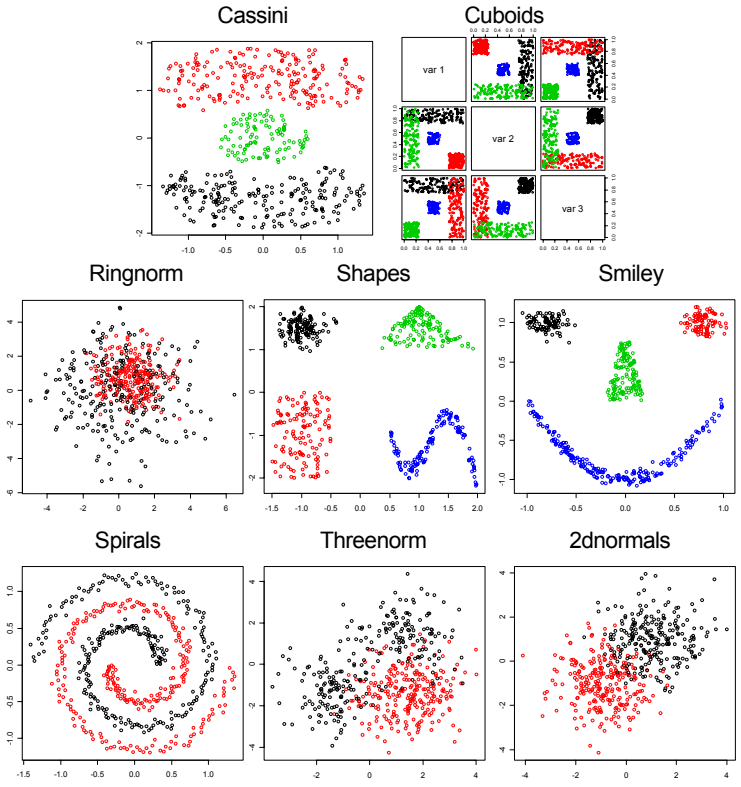
W badaniach empirycznych zastosowano sztucznie generowane zbiory danych, które standardowo wykorzystywane są w badaniach porównawczych w taksonomii*. Są to takie zbiory, w których przynależność obiektów do grup jest z góry znana. Ich krótka charakterystyka znajduje się w tab. 1, natomiast struktura przedstawiona jest na rys. 2.

* Zbiory zostały zaczerpnięte z pakietu mlbench z programu **R**.

Tabela 1

Charakterystyka zastosowanych zbiorów danych

Zbiór danych	Liczba obiektów	Liczba cech	Liczba klas
Cassini	500	2	3
Cuboids	500	3	4
Ringnorm	500	2	2
Shapes	500	2	4
Smiley	500	2	4
Spirals	500	2	2
Threenorm	500	2	2
2dnormals	500	2	2



Rys. 2. Struktura zastosowanych zbiorów danych

W metodzie *bagging* według Leischa jako metodę bazową zastosowano metodę *k*-średnich, natomiast ostatecznego grupowania dokonano z zastosowaniem metod*: najbliższego sąsiedztwa (bc_single), najdalszego sąsiedztwa

* W nawiasach podano notację stosowaną na rys. 3.

(bc_complete), średniej odległości między skupieniami (bc_average), środka ciężkości (bc_centroid), mediany (bc_median), warda (bc_ward). W metodzie Dudoit i Fridlyand oraz Hornika utworzono 50 prób bootstrapowych oraz na ich podstawie określano podziały składowe z zastosowaniem metody *c*-średnich i *k*-średnich; natomiast agregacja przebiegała z zastosowaniem równania 2 w metodzie Dudoit i Fridlyand oraz 3 w metodzie Hornika*. Macierz współwystąpień była konstruowana za pomocą dwóch metod, tj. metody *c*-średnich i *k*-średnich, a jej późniejszego podziału dokonano za pomocą metod: *c*-średnich, *k*-średnich, *k*-medoidów (pam) oraz clara**.

W taksonomii spektralnej macierz $\boldsymbol{Y}$ grupowana była z zastosowaniem metody *k*-średnich (specc).

W celu porównania stabilności grupowania badanych metod zastosowano miarę opartą na Indeksie Randa. Definicja Indeksu Randa (R) jest następująca (Rand, 1971). Niech U i V będą wynikami dwóch różnych podziałów zbioru G mającego n elementów. Przez a oznaczymy liczbę obiektów znajdujących się w tej samej grupie w podziale U i w tej samej grupie w podziale V (*pary zgodne*), b – liczbę obiektów znajdujących się w różnych grupach w podziale U i w różnych grupach w podziale V (*pary zgodne*), c – liczbę obiektów znajdujących się w tej samej grupie w podziale U i w różnych grupach w podziale V (*pary niezgodne*) oraz przez d – liczbę obiektów znajdujących się w różnych grupach w podziale U i w tej samej grupie w podziale V (*pary niezgodne*). Wtedy Indeks Randa dany jest wzorem:

$$R(U,V) = \frac{a+b}{a+b+c+d} = \frac{a+b}{n(n-1)/2}. \tag{9}$$

Miara Randa mierzy odsetek par obiektów zgodnych w obydwu podziałach U i V w ogólnej liczbie par obiektów określonych na zbiorze obiektów G.

Miarę stabilności można zdefiniować jako:

$$Stab = \frac{2}{Z \cdot (Z-1)} \sum_{\substack{1 \le z,l \le Z \\ z<l}}^{Z} R(P_z, P_l), \tag{10}$$

gdzie:
Z – liczba badanych grupowań,
R – Indeks Randa,

* Na rys. 4 i 5 stosowano skróty cl_bagg_k i cl_consensus_k, jeżeli grupowania składowe określane były z zastosowaniem metody *k*-średnich oraz cl_bagg_c i cl_consensus_c, gdy wykorzystywano metodę *c*-średnich.

** Na rys. 6 i 7 pierwszy człon nazwy odnosi się do sposobu konstrukcji macierzy współwystąpień, a drugi – do sposobu jej późniejszego podziału.

P_z – podział na podstawie z-tego grupowania,

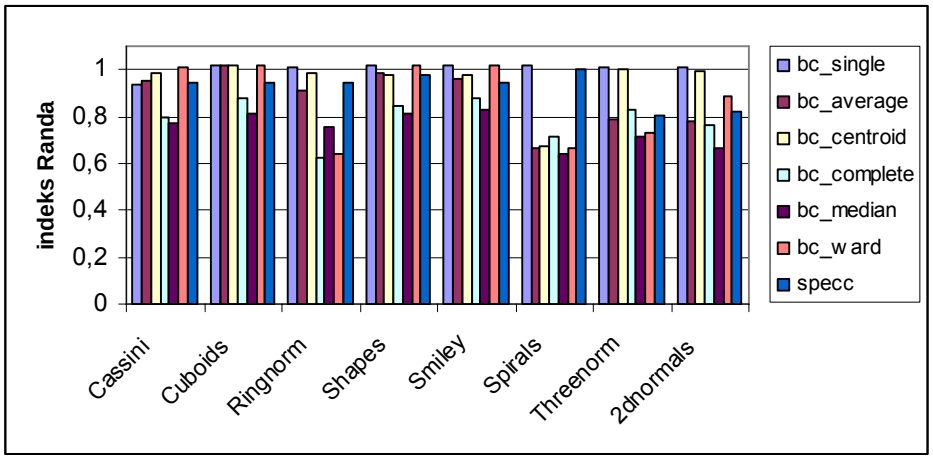
P_l – podział na podstawie l-tego grupowania.

Miara ta ocenia stabilność poprzez pomiar podobieństwa wyników podziałów, które na ich podstawie zostały uzyskane.

4. Wyniki badań empirycznych

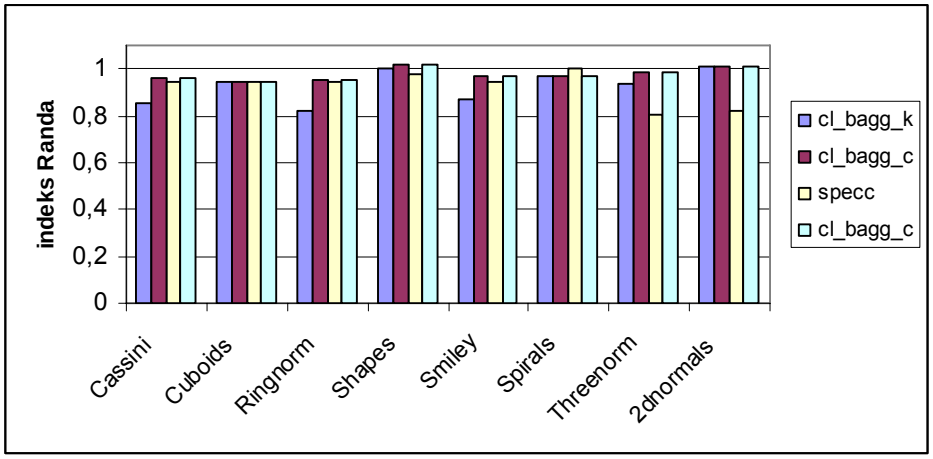
W przypadku wszystkich metod dla zbiorów *Threenorm* i 2*dnormals* taksonomia spektralna daje niższą stabilność niż podejście zagregowane (rys. 3-7). Jedynymi wyjątkami są warianty bc_median, kmeans_kmeans i cmeans_kmeans, a także dodatkowo bc_ward dla zbioru *Threenorm* i bc_complete dla zbioru *2dnormals*. Te dwa zbiory zostaną pominięte w dalszej analizie.

Porównując wyniki dla metody *bagging* wg Leischa (rys. 3), można stwierdzić, że taksonomia spektralna zawsze daje wyższą stabilność niż warianty zagregowane bc_complete, bc_median, ale niższą niż bc_single, bc_average i bc_centroid (za wyjątkiem wariantu bc_average dla zbioru *Ringnorm*).



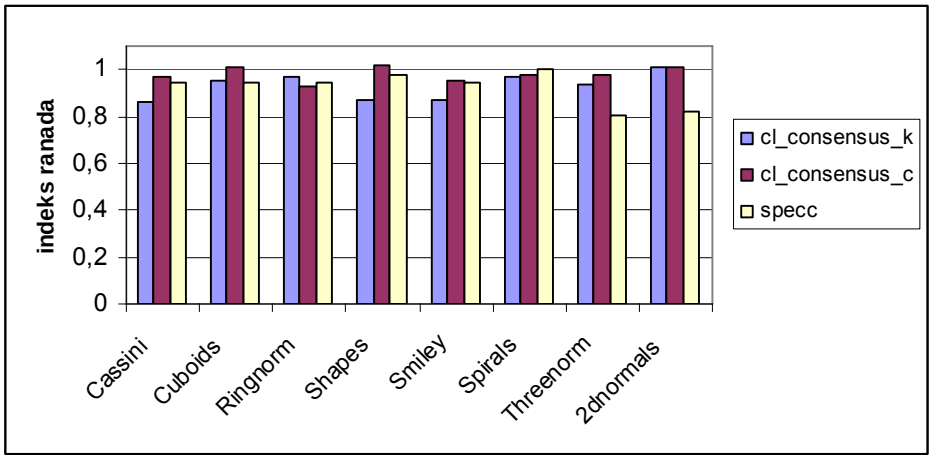
Rys. 3. Porównanie stabilności metody *bagging* według Leischa oraz podejścia spektralnego

W przypadku wyników dla metody *bagging* według Dudoit i Friydland i taksonomii spektralnej (rys. 4) można zauważyć, że obydwa podejścia dają zbliżoną stabilność, za wyjątkiem wariantu cl_bagg_k dla zbiorów *Cassini*, *Ringnorm* i *Smiley*.

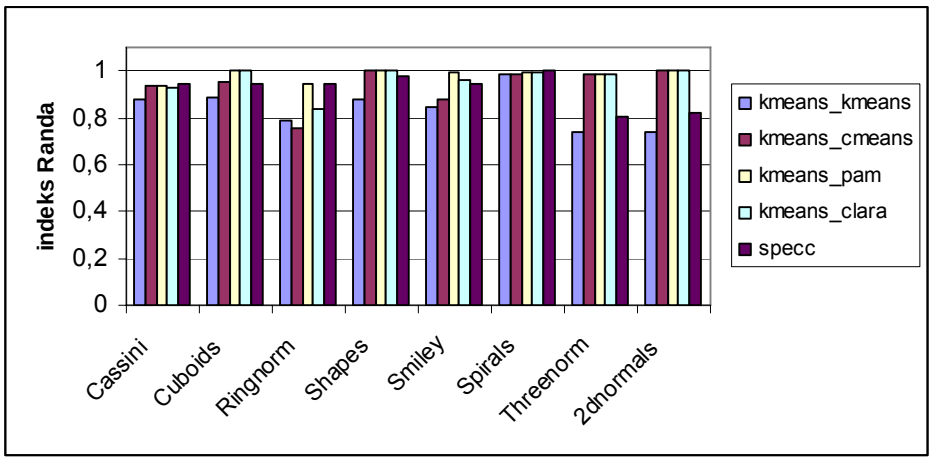


Rys. 4. Porównanie stabilności metody *bagging* według Dudoit i Fridlyand oraz podejścia spektralnego

Wyniki dla metody *bagging* według Hornika i taksonomii spektralnej (rys. 5) pokazują, że dla zbiorów *Cassini*, *Shapes* i *Smiley* metody spektralne zawsze dają nieco niższą stabilność niż cl_consesnsus_c, ale znacznie wyższą niż cl_consensus_k. Dla zbioru *Spirals* metody spektralne wydają się najbardziej stabilne, a dla zbioru *Cuboids* – najmniej stabilne.

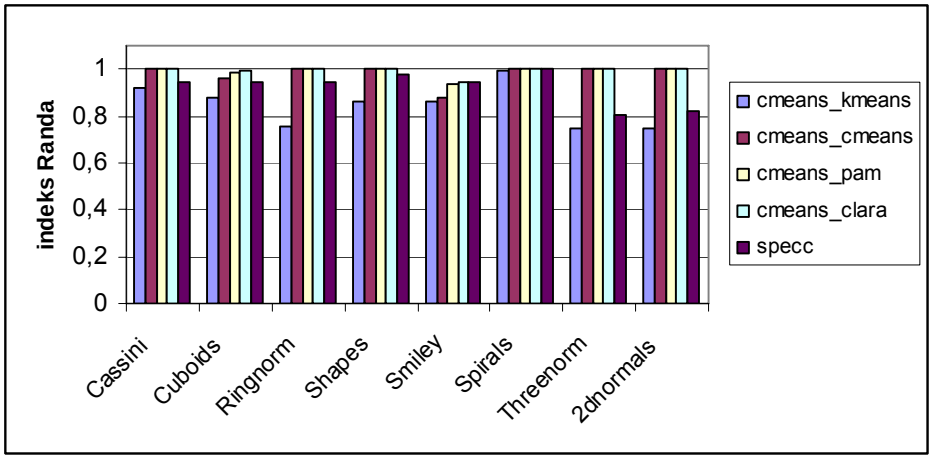


Rys. 5. Porównanie stabilności metody *bagging* według Hornika oraz podejścia spektralnego



Rys. 6. Porównanie stabilności taksonomii spektralnej oraz taksonomii zagregowanej opartej na macierzy współwystąpień konstruowanej za pomocą metody *k*-średnich

W przypadku metody zagregowanej opartej na macierzy współwystąpień konstruowanej za pomocą metody *k*-średnich (rys. 6) można zauważyć, że dla zbiorów *Cassini* i *Spirals* taksonomia spektralna daje bardzo zbliżone rezultaty, jak podejście zagregowane. Dla zbiorów *Cuboids*, *Shapes* i *Smiley* taksonomia spektralna daje niższą stabilność niż warianty zagregowane kmeans_pam i kmeans_clara, ale wyższą niż kmeans_kmeans. W przypadku zbioru *Ringnorm* zastosowanie metod spektralnych nie wydaje się dobrym rozwiązaniem w porównaniu z podejściem zagregowanym.



Rys. 7. Porównanie stabilności taksonomii spektralnej oraz taksonomii zagregowanej opartej na macierzy współwystąpień konstruowanej za pomocą metody *c*-średnich

Dla metody zagregowanej opartej na macierzy współwystąpień konstruowanej za pomocą metody *c*-średnich (rys. 7), w przypadku zbiorów *Cassini*, *Cuboids*, *Ringnorm* i *Shapes* taksonomia spektralna daje niższą stabilność niż warianty zagregowane cmeans_cmeans, cmeans_pam i cmeans_clara, ale wyższą niż cmeans_kmeans. Dla zbiorów *Smiley* i *Spirals* metody spektralne dają podobną bądź nieco wyższą stabilność niż rozpatrywane warianty metody opartej na macierzy współwystąpień.

Podsumowanie

Podsumowując całość przeprowadzonych badań, można stwierdzić, że w przypadku zbiorów danych z trudno separowalnymi grupami (np. *Threenorm*, *2dnormals*) taksonomia spektralna może dawać niższą stabilność niż podejście zagregowane. Podejście spektralne jest zawsze bardziej stabilne niż warianty zagregowane bc_median, bc_complete, kmeans_kmeans i cmeans_kmeans, ale równie bądź mniej stabilne niż bc_single, kmeans_pam i cmeans_clara.

Bilbiografia

Breiman L. (1996): *Bagging predictors*. „Machine Learning", 26(2).

Dudoit S., Fridlyand J. (2003): *Bagging to Improve the Accuracy of a Clustering Procedure*. „Bioinformatics", 19(9).

Fern X.Z., Brodley C.E. (2003): *Random Projection for High Dimensional Data Clustering: A Cluster Ensemble Approach*. „Proceedings of the 20th International Conference on Machine Learning".

Fred A. (2002): *Finding Consistent Clusters in Data Partitions*. „Proceedings of the International Workshop on Multiple Classifier Systems".

Fred N.L., Jain A.K. (2002): *Combining Multiple Clusterings Using Evidence Accumulation*. „IEEE Transactions on PAMI", 27(6).

Hornik K. (2005): *A CLUE for CLUster Ensembles*. „Journal of Statistical Software", 14.

Kannan R., Vempala S., Vetta A. (2000): *On Clusterings – Good, Bad, and Spectral*. „Proceedings of the 41st Annual Symposium on Foundations of Computer Science".

Kuncheva L., Vetrov D. (2006): *Evaluation of Stability of k-means Cluster Ensembles with Respect to Random Initialization*. „IEEE Transactions on Pattern Analysis and Machine Intelligence", Vol. 28, No. 11.

Leisch F. (1999): *Bagged Clustering*. „Adaptive Information Systems and Modeling in Economics and Management Science", Working Paper 51.

Ng A.Y., Jordan M.I, Weiss, Y. (2001): *On Spectral Clustering: Analysis and an Algorithm*. „Advances in Neural Information Processing Systems".

Shi J., Malik J. (2000): *Normalized Cuts and Image Segmentation*. „IEEE Transactions on Pattern Analysis and Machine Intelligence", 22(8).

Strehl A., Ghosh J. (2002): *Cluster Ensembles - A Knowledge Reuse Framework for Combining Multiple Partitions*. „Journal of Machine Learning Research", 3.

COMPARISON OF SPECTRAL CLUSTERING AND CLUSTER ENSEMBLES STABILITY

Summary

High accuracy of the results is very important task in any grouping problem (clustering). It determines effectiveness of the decisions based on them. Therefore in the literature there are proposed methods and solutions that main aim is to give more accurate results than traditional clustering algorithms (e.g. k-means or hierarchical methods). Examples of such solutions can be cluster ensembles or spectral clustering algorithms. A desirable quality of any clustering algorithm is also stability of the method with respect to small perturbations of data (e.g. data subsampling, small variations in the feature values) or the parameters of the algorithm (e.g. random initialization). Empirical results shown that cluster ensembles are more stable than traditional clustering algorithms. Here, we carry out an experimental study to compare stability of spectral clustering and cluster ensembles.

Małgorzata Szerszunowicz
Uniwersytet Ekonomiczny w Katowicach

ANALIZA ZDOLNOŚCI PROCESU O ZALEŻNYCH CHARAKTERYSTYKACH

Wprowadzenie

Statystyczna kontrola jakości ma na celu doskonalenie procesu produkcyjnego poprzez wykorzystanie metod probabilistycznych i statystycznych. Wśród najważniejszych narzędzi statystycznej kontroli jakości należy wyróżnić karty kontrolne, planowanie eksperymentów, analizę zdolności procesu i plany odbiorcze.

Aby określić stopień zgodności wymogów produkcyjnych z rezultatami procesu produkcyjnego, wykorzystuje się analizę zdolności procesu. Jako „zdolność" procesu należy rozumieć zdolność technologiczną realizowanego w praktyce procesu produkcyjnego do spełnienia wymagań projektowo-konstrukcyjnych produktu określonych przez wartości zadanego zbioru parametrów. W literaturze określenie „zdolności" zastępuje się czasami pojęciem „wydolności" bądź „zdolności projektowej" procesu. Zatem analiza zdolności procesu jest badaniem zgodności pomiędzy wymaganiami projektu produktu a możliwościami procesu technologicznego. Celem niniejszego artykułu jest analiza problemu pomiaru zdolności procesów wielowymiarowych o zależnych charakterystykach, wykorzystująca konstrukcję pewnego wielowymiarowego wskaźnika zdolności procesu.

1. Wskaźniki zdolności procesu

Analiza zdolności procesu ma istotny wpływ na jakość wytwarzanego produktu, o czym świadczą m.in. korzyści wynikające z poprawnej analizy zdolności procesu. Do podstawowych zalet stosowania tej analizy można zaliczyć możliwość (Montgomery, 1997):

- oceny stopnia spełnienia wymagań obserwowanego procesu odnośnie do granic tolerancji,
- wprowadzenia ewentualnych modyfikacji projektu,

- podjęcia działań w celu zmniejszenia wariancji charakterystyk procesu produkcyjnego,
- określenia wymagań odnośnie do rezultatów przy zastosowaniu nowych technologii,
- umożliwienia wyboru najlepszego produktu spośród oferowanych przez różnych dostawców,
- poprawnego ustalenia porządku produkcji, wówczas gdy poszczególne etapy procesu mają wpływ na tolerancję,
- ustalenia właściwych odstępów pomiędzy pobieraniem poszczególnych próbek z bieżącego procesu.

Należy zauważyć, że powyższe efekty stosowania analizy zdolności procesu odnoszą się do różnych etapów procesu produkcyjnego, co świadczy o jej istotnym znaczeniu.

Zdolność procesu produkcyjnego jest ściśle związana z ogólną oceną jakości produktu, na którą wpływ ma naturalna zmienność procesu w danym punkcie czasowym oraz zmienność obserwowanej charakterystyki w czasie (Kończak, 2007). Weryfikacja tak rozumianej zmienności procesu może się odbywać poprzez analizę histogramu, wykresu pudełkowego czy karty kontrolnej, co może być uznane za wstępną analizę zdolności procesu. Natomiast w praktyce przedsiębiorstw produkcyjnych powszechnie stosowane są liczbowe oceny zdolności procesu, a mianowicie wskaźniki zdolności procesu (ang. *process capability index*).

Wskaźniki zdolności procesu stanowią wiarygodną ocenę zdolności procesu wówczas, gdy spełnione są następujące założenia (Kończak, 2007):

- badany proces jest uregulowany i stabilny,
- charakterystyka opisująca badany proces ma rozkład normalny,
- pobrana próbka jest reprezentatywna,
- obserwacje do próby pobierane są w sposób niezależny.

Jeżeli powyższe założenia nie są jednocześnie spełnione, stosowanie klasycznych wskaźników zdolności procesu nie jest zasadne.

Najczęściej wykorzystywanym wskaźnikiem zdolności procesu jest znormalizowany wskaźnik C_p, który można opisać za pomocą następującego ilorazu:

$$C_p = \frac{\text{długość przedziału tolerancji}}{\text{długość przedziału obejmującego 99,73 \% wartości badanej cechy}}.$$

Symbolicznie wskaźnik C_p można przedstawić w postaci

$$C_p = \frac{USL - LSL}{6\sigma} \tag{1}$$

gdzie:
USL – górna granica przedziału tolerancji,
LSL – dolna granica przedziału tolerancji,
σ – odchylenie standardowe obserwowanej zmiennej o rozkładzie normalnym.
Obszar zmienności cech, w którym produkt uznaje się za zgodny z wymaganiami jakościowymi wyznaczony jest przez wartości USL i LSL ustalone przez zleceniodawcę. W zależności od wartości wskaźnika C_p zdolność procesu można określić następująco (Iwasiewicz, 1999):

- $C_p < 1$ – niska względna zdolność procesu,
- $1 < C_p < 1{,}33$ – średnia względna zdolność procesu,
- $1{,}33 < C_p$ – wysoka względna zdolność procesu.

Zatem im większą wartość przyjmuje wskaźnik C_p, tym większa jest względna zdolność badanego procesu, przy czym należy podkreślić, że obecnie wymagania dotyczące zdolności procesów są znacznie większe, dlatego w wielu przypadkach umowne granice należałoby określić inaczej.

Wśród licznych modyfikacji wskaźnika zdolności C_p należy wyróżnić wskaźniki C_{pk} i C_{pm}. Wskaźnik C_{pk} dzięki swojej konstrukcji umożliwia określenie różnicy pomiędzy wartością średnią procesu a nominalnym poziomem przeciętnym, natomiast wskaźnik C_{pm} informuje o rozmieszczeniu wartości nominalnej procesu pomiędzy granicami obszaru tolerancji.

2. Analiza zdolności procesów produkcyjnych o wielu charakterystykach

Ocena zdolności procesu produkcyjnego zazwyczaj opiera się na analizie pojedynczej zmiennej najlepiej charakteryzującej badany proces. Jednak często rozważany proces produkcyjny wymaga weryfikacji spełnienia wymagań projektowo-konstrukcyjnych przez większą liczbę charakterystyk. Wówczas pełniejszy obraz przebiegu procesu dostarcza jego analiza wielowymiarowa.

W literaturze rozważa się liczne sposoby analizy zdolności procesu produkcyjnego opisanego przez więcej niż jedną charakterystykę. Zazwyczaj rozważania te sprowadzają się do przedstawienia postaci wielowymiarowych odpowiedników klasycznych wskaźników zdolności procesu (Kotz, Jonson, 1993).

Konstrukcja wielowymiarowych wskaźników zdolności procesu odnosi się bezpośrednio do jednowymiarowych wskaźników zdolności procesu, a więc do zależności pomiędzy obszarem specyfikacji a obszarem zawierającym 99,73%

obserwacji. Zatem konstrukcję wielowymiarowego wskaźnika zdolności procesu można opisać za pomocą następującego ilorazu:

$$C_p^{Mult} = \frac{\text{miara obszaru tolerancji}}{\text{miara obszaru obejmującego 99,73 \% wartości badanej cechy}}.$$

Ponadto charakterystyki procesu muszą spełniać wcześniej wspomniane założenia, w szczególności założenie dotyczące niezależności poszczególnych obserwacji. Niezależność w ujęciu wielowymiarowym może być rozumiana jako:
- niezależność pomiędzy poszczególnymi obserwacjami każdej ze zmiennych,
- niezależność pomiędzy rozkładami rozważanych charakterystyk.

Jeżeli zachodzi zależność pomiędzy poszczególnymi obserwacjami rozważanych charakterystyk, wtedy nawet jednowymiarowa analiza zdolności nie jest zasadna. Interesującym zagadnieniem jest analiza zdolności procesu charakteryzowanego przez zmienne zależne, kiedy to analiza zdolności polegająca na jednowymiarowej weryfikacji zdolności procesu może okazać się błędna ze względu na zależność pomiędzy zmiennymi. Wówczas w literaturze rozważa się wskaźnik zdolności procesu, będący ilorazem pola zmodyfikowanego obszaru specyfikacji i obszaru obejmującego 99,73% obserwacji badanych charakterystyk (Zahid, Sultana, 2008).

Niech proces produkcyjny będzie opisany za pomocą dwóch zmiennych mających rozkład normalny o parametrach μ_i, σ_i, dla i = 1,2 i zadanym współczynnikiem korelacji ρ_{12}. Wówczas w celu oceny zdolności badanego procesu należy określić obszar specyfikacji dla rozważanych charakterystyk. Niech obszar ten będzie zadany w postaci elipsy obejmującej q% obserwacji pochodzących z procesu uregulowanego, przy czym wartość q% jest ustalona przez zleceniodawcę. Dwuwymiarowy wskaźnik zdolności procesu można określić w postaci ilorazu:

$$C_p^M = \frac{\text{pole obszaru tolerancji obejmującego } q\% \text{ wartości badanej cechy z procesu uregulowanego}}{\text{pole obszaru obejmującego 99,73 \% wartości badanej cechy z procesu badanego}},$$

gdzie obszar obejmujący 99,73% wartości badanej cechy z procesu badanego również jest elipsą. Tak określony wskaźnik zdolności procesu pozwoli na weryfikację spełnienia wymagań projektowo-konstrukcyjnych określonych na podstawie danych pochodzących z procesu uregulowanego, w stosunku do badanego procesu.

3. Analiza zdolności dwuwymiarowego procesu produkcyjnego

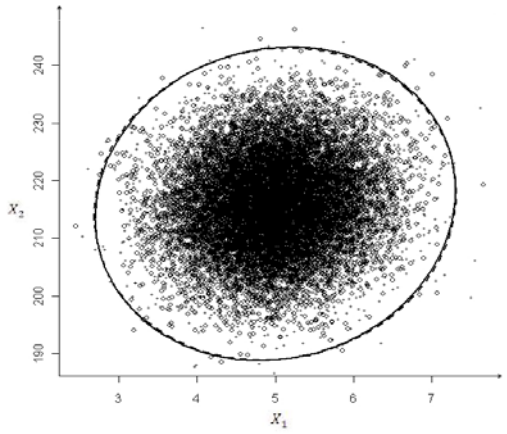
Rozważmy proces produkcyjny polegający na wytworzeniu elementów ceramicznych (Zahid, Sultana, 2008). Proces ten charakteryzowany jest przez zmienne pochodzące z dwuwymiarowego rozkładu normalnego: X_1 – wysokość

i X_2 – waga. Wiadomo, że tak określony, uregulowany proces produkcyjny ma następujące parametry:

$$\mu = [5,216], \quad \sigma = [0,67,\ 7,92], \quad \rho_{12} = 0,08 .$$

Ponadto zleceniodawca określił przedziały zmienności dla wysokości i wagi elementu, które są odpowiednio postaci (4, 6) oraz (190, 242).

Niech ustalony obszar specyfikacji określony na podstawie procesu uregulowanego będzie elipsą obejmującą 99,73% obserwacji. Pole tak określonego obszaru dla zadanego procesu produkcyjnego wynosi 9,61. Jeżeli z procesu uregulowanego zostanie pobrana odpowiednio liczna próba oraz określony zostanie obszar obejmujący 99,73% jej obserwacji, wówczas wskaźnik zdolności tak określonego procesu będzie bliski wartości 1, co potwierdza rys. 1.



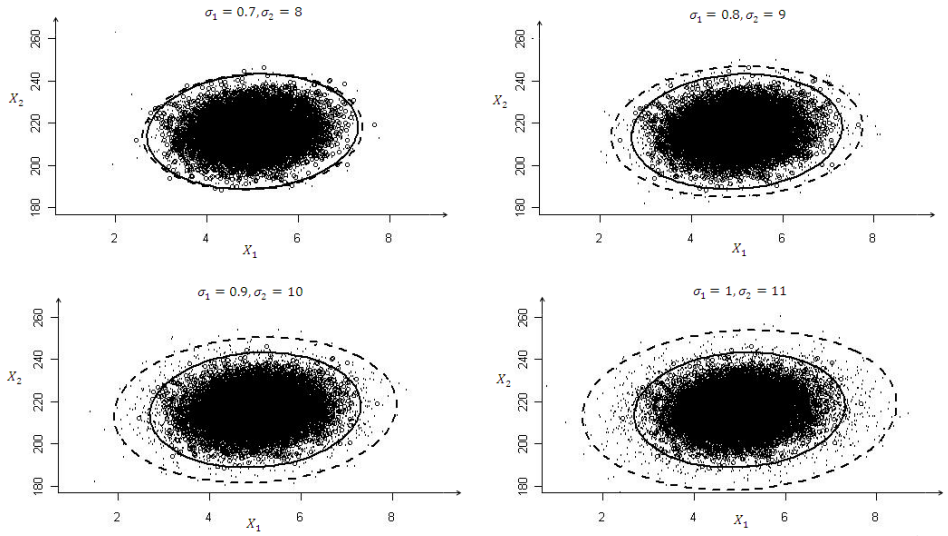
Rys. 1. Obszar specyfikacji obejmujący 99,73% obserwacji procesu uregulowanego wraz z obszarem obejmującym 99,73% obserwacji próby pochodzącej z procesu uregulowanego

Jednym z możliwych rozregulowań procesu produkcyjnego jest zmiana wartości odchyleń standardowych poszczególnych zmiennych, czyli ich rozrzutu. Zatem rozważmy proces, w którym zarówno rozrzut pierwszej, jak i drugiej zmiennej rośnie. Wartości dwuwymiarowego wskaźnika zdolności procesu w zależności od wartości odchyleń standardowych przedstawia tab. 1. Zauważmy, że wraz ze wzrostem odchyleń standardowych poszczególnych zmiennych względna zdolność badanego procesu istotnie maleje, co potwierdza graficzna prezentacja obserwacji pochodzących z procesu badanego i procesu uregulowanego (rys. 2).

Tabela 1

Wartości dwuwymiarowego wskaźnika zdolności procesu w zależności od rosnących wartości odchyleń standardowych procesu

σ_1	0,7	0,8	0,9	1
σ_2	8	9	10	11
C_p^M	0,9841	0,9256	0,8881	0,8586



Rys. 2. Graficzna prezentacja obszaru specyfikacji i obszarów obejmujących 99,73% obserwacji pochodzących z procesów o większym rozrzucie

W przypadku gdy rozrzut obserwacji badanego procesu maleje, wartość wskaźnika zdolności procesu rośnie, co potwierdzają obliczenia zamieszczone w tab. 2.

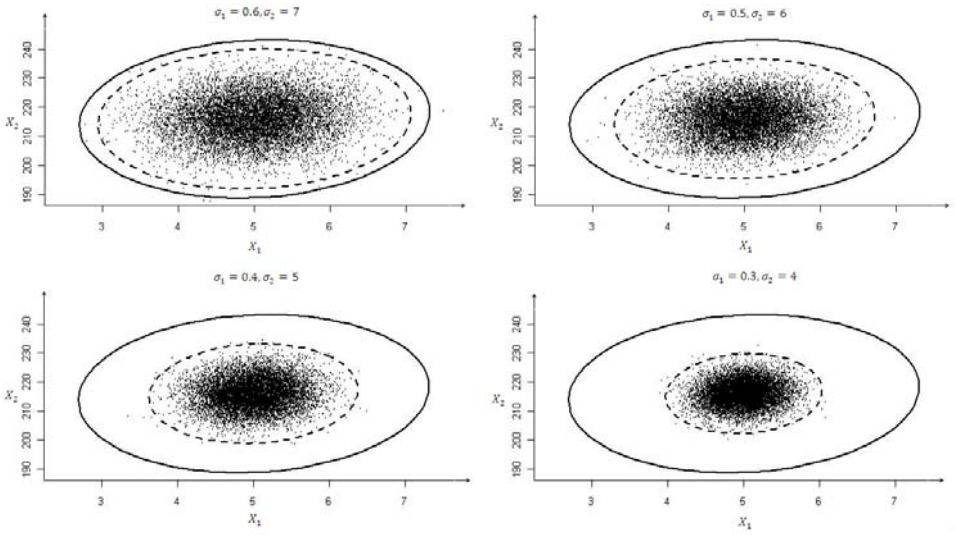
Tabela 2

Wartości dwuwymiarowego wskaźnika zdolności procesu w zależności od malejących wartości odchyleń standardowych procesu

σ_1	0,6	0,5	0,4	0,3
σ_2	7	6	5	4
C_p^M	1,0351	1,1387	1,2452	1,4201

Istotnie, wartość miernika zdolności procesu maleje, ponieważ pole obejmujące 99,73% obserwacji pochodzących z procesu rozregulowanego maleje, co przedstawia rys. 3.

Jednym z istotnych rozregulowań procesu produkcyjnego jest zależność pomiędzy badanymi zmiennymi, zatem wartość współczynnika korelacji ma wpływ również na ocenę zdolności procesu. Rozważmy proces produkcyjny o parametrach μ_1, σ_1, μ_2, σ_2 ustalonych dla procesu uregulowanego, natomiast o różnych wartościach współczynnika korelacji. Wówczas, w zależności od wartości współczynnika korelacji, wskaźnik zdolności procesu przyjmuje wartości przedstawione w tab. 3.



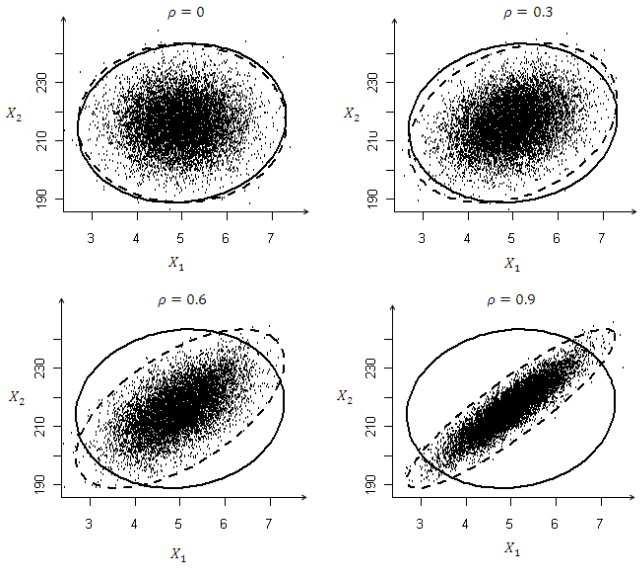
Rys. 3. Graficzna prezentacja obszaru specyfikacji i obszarów obejmujących 99,73% obserwacji pochodzących z procesów o mniejszym rozrzucie

Tabela 3

Wartości dwuwymiarowego wskaźnika zdolności procesu w zależności od wartości współczynnika korelacji

ρ	0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
C_p^M	1,006	1,014	1,006	1,024	1,027	1,040	1,053	1,077	1,121	1,208

Łatwo można dostrzec, że wraz ze wzrostem współczynnika korelacji wzrasta wartość wskaźnika zdolności procesu i co za tym idzie względna zdolność procesu rośnie. Graficzną prezentację danych dla wybranych współczynników korelacji badanych zmiennych przedstawia rys. 4.



Rys. 4. Graficzna prezentacja obszaru specyfikacji i obszaru obejmującego 99,73% obserwacji pochodzących z procesów o różnych wartościach współczynnika korelacji

Warto podkreślić, że istotnym elementem w prowadzonych rozważaniach jest uwzględnianie zależności pomiędzy zmiennymi.

Podsumowanie

Mierniki zdolności procesu produkcyjnego stanowią istotny element analizy zdolności procesu. Zazwyczaj są one określane dla zmiennych o rozkładach normalnych, zarówno w aspekcie jedno-, jak i wielowymiarowym. W przedstawionych rozważaniach dotyczących szczególnego przypadku dwuwymiarowego procesu produkcyjnego wykorzystanie proponowanego wskaźnika zdolności procesu pozwala na weryfikację spełnienia wymagań projektowo-konstrukcyjnych, również wtedy, gdy badane zmienne są zależne.

Bibliografia

Iwasiewicz A. (1999): *Zarządzanie jakością*. Wydawnictwo Naukowe PWN, Warszawa – Kraków.

Kończak G. (2007): *Metody statystyczne w sterowaniu jakością produkcji*. Wydawnictwo Akademii Ekonomicznej, Katowice.

Kotz S., Johnson N.L. (1993): *Process Capability Indices*. Chapman & Hall, London.

Montgomery D.C. (1997): *Introduction to Statistical Quality Control*. John Wiley & Sons, New York.

Zahid A., Sultana A. (2008): *Assesment and Comparison of Multivariate Process Capability Indices in Ceramic Industry*. „Journal of Mechanical Engineering", Vol. ME39, No. 1.

PROCESS CAPABILITY ANALYSIS FOR PROCESSES WITH DEPENDENT CHARACTERISTICS

Summary

One of the most important tools of statistical quality control is a process capability analysis. In order to measure process capability the most used are indices constructed for one–dimensional characteristic of production process. However often production process is described by more than one characteristic. One should then conduct a multidimensional assessment of process capability with appropriately designed indicators.

The aim of this article is to analyze the problem of process capability measure of multi-dimensional process with dependent characteristics, using a proposed multi-dimensional process capability index.

Grażyna Trzpiot
Uniwersytet Ekonomiczny w Katowicach

WYBRANE STATYSTYKI ODPORNE

Wprowadzenie

Obserwacje oddalone (*outliers*) są takimi obserwacjami w próbie, które mogą powodować zakłócenia w ocenie relacji w próbie. Nie jest to termin o znaczeniu pejoratywnym; obserwacje oddalone mogą być poprawne, ale powinny być identyfikowane dla oceny błędów. Poczynając od 60., zaproponowano wiele metod silnych i odpornych *(robust* and *resistant)* mniej wrażliwych na obserwacje oddalone. Mogą one konkurować, a nawet wygrywać ze standardowymi statystycznymi metodami. Omawiana tematyka jest przedmiotem wcześniejszych prac autorki zawsze w kontekście zastosowań w ekonomii (Trzpiot 2009, 2011a, 2011b). Artykuł ten ma charakter opisowy w powiązaniu z przygotowywanym podręcznikiem.

1. Statystyki jednowymiarowe położenia i skali

Średnia z próby może być załamana przez pojedynczą obserwację. Jeżeli dowolna obserwacja ma wartość taką, że $y_i \to \pm\infty$, wówczas średnia z próby $\bar{y} \to \pm\infty$, w przeciwieństwie do mediany z próby, która nie jest wrażliwa na pojedyncze wartości zmierzające do nieskończoności. Mówimy, że mediana jest odporna na duże błędy, podczas gdy średnia nie. Faktycznie mediana może znieść do 50% dużych błędów zanim będzie arbitralnie duża; mówimy, że ma *punkt załamania* 50-proc., podczas gdy dla średniej mamy odpowiednio 0%.

Średnia jest efektywnym estymatorem parametru położenia dla rozkładu normalnego, dlatego może być wykorzystywana jako estymator parametru położenia dla rozkładów zbliżonych do normalnych. Metody odporne powinny mieć wysoką efektywność w otoczeniu zakładanego modelu statystycznego.

Dlaczego nie jest wystarczające przesianie danych i odrzucenie obserwacji odstających? Należy rozważyć wiele aspektów metodologicznych:

1. Praktycy, nawet eksperci statystycy, nie zawsze przeglądają zbiory danych.
2. Ostre decyzje, czy zachować, czy odrzucić obserwacje mogą być niezbyt trafne. Proponujemy nadać wagi wątpliwym obserwacjom. Możemy również odrzucić kompletnie złe obserwacje.

3. Może być zadaniem trudnym lub wręcz niemożliwym umiejscowienie obserwacji odstających w wielowymiarowym lub mocno zrestrukturyzowanym zbiorze danych.
4. Odrzucenie obserwacji odstających wpływa na rozkład teoretyczny (zmiennej losowej), który musi być skorygowany. W szczególności wariancja będzie niedoszacowania w „wyczyszczonym" zbiorze.

Dla ustalonego rozkładu definiujemy *relatywną efektywność* estymatora. Efektywność estymatora $\hat{\theta}$ względem innego estymatora $\tilde{\theta}$ możemy zmierzyć, posługując się następującą miarą efektywności:

$$RE(\tilde{\theta},\hat{\theta}) = \frac{D^2(\hat{\theta})}{D^2(\tilde{\theta})} \tag{1.1}$$

Granicę $RE(\tilde{\theta},\hat{\theta})$ przy rosnącej do nieskończoności wielkości próby nazywamy *efektywnością asymptotyczną*:

$$ARE(\tilde{\theta},\hat{\theta}) = \lim_{n\to\infty} RE(\tilde{\theta},\hat{\theta}) \tag{1.2}$$

Estymatorem asymptotycznie najefektywniejszym jest estymator, którego asymptotyczna efektywność równa się jedności. Można problem zdefiniować również w odniesieniu do asymptotycznych wariancji. Jeżeli estymator $\hat{\theta}$ nie jest znany, wówczas zakładamy, że jest efektywnym estymatorem. Pojawiają się trudności z obciążonymi estymatorami, których wariancja jest mała lub wynosi zero.

Proponowanym w literaturze rozwiązaniem jest wykorzystanie błędów średniokwadratowych, innym – przeskalowanie $\theta/E(\hat{\theta})$. Iglewicz (1983) proponuje wykorzystanie wariancji logarytmu estymatora $\hat{\theta}$: $D^2(\log\hat{\theta})$ jako estymatora parametru skali*. Zastosujmy podejście *ARE* do oceny średniej i mediany (Venables, Ripley, 2002). Dla rozkładu normalnego

$$ARE(\text{mediana, średnia}) = \frac{D^2(\textit{średnia})}{D^2(\textit{mediana})} = 2/\pi \approx 64\%$$

Dla rozkładów o innych wartościach rozkładów w ogonach mediana ma lepsze własności. Przykładowo, dla rozkładu t-Studenta z pięcioma stopniami swobody, a to jest często rozkład zgodny z rozkładem błędów modeli, *ARE* (mediana, średnia) ≈ 96%.

* jest niezależna od skali

Kolejny przykład podał Tukey (1960). Zakładamy, że mamy n obserwacji $Y_i \sim N(\mu, \sigma^2)$ dla $i = 1,..., n$ oraz chcemy estymować wartość wariancji σ^2. Rozważmy dwa estymatory $\hat{\sigma}^2 = s^2$ oraz $\tilde{\sigma}^2 = d^2\pi/2$, gdzie:

$$d = \frac{1}{n}\sum_i \left|Y_i - \overline{Y}\right|$$

oraz stała jest wybrana tak, że dla rozkładu normalnego $d \to \sqrt{2/\pi}\sigma$. Wówczas $ARE(\tilde{\sigma}^2, s^2) = 0{,}876$.

Załóżmy, że dla każdego Y_i mamy obserwacje z rozkładu $N(\mu, \sigma^2)$ z prawdopodobieństwem $1 - \varepsilon$ oraz wartości z rozkładu $N(\mu, 9\sigma^2)$ z prawdopodobieństwem ε. Zauważmy, że obydwie wariancje dla wszystkich obserwacji oraz wariancja niezakłóconego rozkładu obserwacji są proporcjonalne do σ^2. Otrzymujemy dane zawarte w tab. 1.

Tabela 1

Wartości *ARE* dla wybranych wartości ε

ε (%)	$ARE(\tilde{\sigma}^2, s^2)$
0	0,876
0,1	0,948
0,2	1,016
1	1,44
5	2,04

Źródło: Na podstawie (Venables, Ripley, 2002).

Mieszanka rozkładów z zakłóceniem $\varepsilon = 1\%$ jest nieodróżnialna od rozkładu normalnego, zwłaszcza w praktycznych zastosowaniach, dlatego optymalność s^2 jest bardzo wrażliwa. Mówimy o braku *odporności efektywności* estymatora.

Znajdujemy odmienne estymatory parametru σ niż $d\sqrt{\pi/2}$ (mają punkt załamania 0%). Dwa proponowane rozwiązania przyjmowane jako estymatory są porównywalne:

$$IQR = X_{(3n/4)} - X_{(n/4)} \tag{1.3}$$

oraz

$$MAD = \underset{i}{\text{mediana}} \{|Y_i - \underset{j}{\text{mediana}}(Y_j)|\} \tag{1.4}$$

Przykładowo, dla rozkładu normalnego otrzymujemy odpowiednio następujące wyniki:

$MAD \rightarrow$ mediana $\{|Y - \mu|\} \approx 0{,}6745\sigma$,
$IQR \rightarrow \sigma[\Phi{-}1(0{,}75) - \Phi{-}1(0{,}25)] \approx 1{,}35\sigma$

Obydwa estymatory są efektywne, ale bardzo odporne na obserwacje oddalone w zbiorze danych. Dla estymatora MAD i dla rozkładu normalnego mamy $ARE = 37\%$ (Staudte, Sheather, 1990, s. 123).

W kolejnym kroku rozważań zakładamy, że mamy n niezależnych obserwacji Y_i z rodziny z parametrem położenia o funkcji gęstości $f(y-\mu)$, oraz funkcja f jest symetryczna względem zera. Zatem μ jest wartością centralną (mediana, średnia, jeżeli istnieje) dystrybuanty Y_i. Rozważamy rozkład niewiele różniący się od rozkładu normalnego. Mamy wiele estymatorów wartości μ. Wśród tego zbioru estymatorów znajdujemy średnią z próby, medianę z próby i estymatory wyznaczane metodą największej wiarygodności (MNW). Dodatkowo rozpatrujemy *średnią uciętą*, która jest średnią dla $1-2\alpha$ wartości rozkładu, zatem αn obserwacji jest usuniętych z każdego końca badanego rozkładu (największych i najmniejszych).

2. M-estymatory parametrów położenia i skali

Rozważymy jako estymatory parametru położenia znane z literatury *M-estymatory*. Nazwa pochodzi od sformułowania „prawie" MNW estymatory (‘*MLElike’ estimators*).

Analizując funkcję gęstości f, możemy zdefiniować funkcję $\rho = -\log f$. Wówczas estymator największej wiarygodności wyznaczamy jako:

$$\min_{\mu} \sum_{i} -\log f(y_i - \mu) = \min_{\mu} \sum_{i} \rho(y_i - \mu) \tag{2.1}$$

Powyższe przekształcenie jest użyteczne, jeżeli funkcja ρ nie jest funkcją gęstości. Zapiszmy, jako $\psi = \rho'$ (jeżeli ta pochodna istnieje), wówczas otrzymujemy:

$$\sum_{i} \psi(y_i - \hat{\mu}) = 0 \text{ lub } \sum_{i} w_i(y_i - \hat{\mu}) = 0 \tag{2.2}$$

gdzie:
$w_i = \psi(y_i - \hat{\mu})/(y_i - \hat{\mu})$.

To sugeruje iteracyjne metody rozwiązania, przy czym wagi uaktualniamy przy każdej kolejnej iteracji.

Przykłady M-estymatorów

Średnia z próby odpowiada funkcji $\rho(x) = x^2$, mediana z próby odpowiada funkcji $\rho(x) = |x|$. Dla dowolnego n mediana jest rozwiązaniem zapisanego problemu.

Funkcja

$$\psi(x) = \begin{cases} x, & |x| < c \\ 0, & |x| \geq c \end{cases}$$

odpowiada *uciętej metryce*; duże odległości pomiędzy wartościami nie mają żadnego wpływu.

Funkcja*

$$\psi(x) = \begin{cases} -c, x < -c \\ x, & |x| < c \\ c, & x > c \end{cases}$$

odpowiada metryce Winsora i obejmuje wartości ekstremalne obserwacji jako $\mu \pm c$.

Odpowiednia funkcja $\rho = -\log f$ jest następująca:

$$\rho(x) = \begin{cases} x^2, & |x| < c \\ c(2|x| - c), & |x| \geq c \end{cases}$$

i wyznacza funkcję gęstości o rozkładzie Gaussa w centrum rozkładu, mającą podwójnie wykładnicze ogony. Ten estymator zdefiniował Huber (1981). Zauważmy, że jeżeli $c \to 0$, w granicy otrzymujemy medianę, oraz jeżeli $c \to \infty$, wówczas granicą jest średnia. Wartość $c = 1{,}345$ zapewnia 95% efektywności dla rozkładu normalnego.

Funkcja podwójnie ważąca Tukey'a ma postać:

$$\psi(t) = t[1 - \left(\frac{t}{R}\right)^2]_+^2$$

gdzie $[.]_+$ oznacza dodatnie wartości. To jest, jak zwykło się określać, „łagodne" (*soft*) ucinanie. Wartość $R = 4{,}685$ zapewnia 95% zgodności efektywności dla rozkładu normalnego (Venables, Ripley, 2002).

Kolejny przykład to funkcja ψ Hampela (1986), która jest kawałkami liniowa:

* Pojęcie określone przez Charlesa P. Winsora (por. Dixon, 1960).

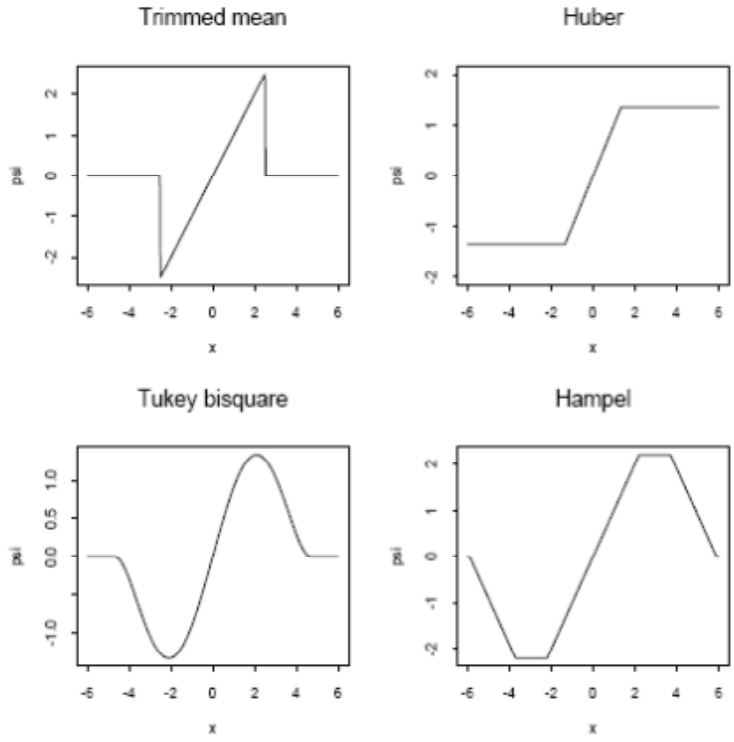
$$\psi(x) = \begin{cases} |x|, & 0 < |x| < a \\ a, & a < |x| < b \\ a(c-|x|)/(c-b), & b < |x| < c \\ 0, & c < |x| \end{cases}$$

Ilustracja omówionych estymatorów (rys. 1) wymagała przyjęcia umownych wartości parametrów: $a = 2{,}2s$, $b = 3{,}7s$, $c = 5{,}9s$. Zauważamy oczywiście problem skali. W czterech ostatnich przypadkach mamy współczynniki skali (c, R lub s).

Możemy zastosować estymator do przeskalowania rezultatów:

$$\min_{\mu} \sum_{i} \rho\left(\frac{y_i - \mu}{s}\right) \tag{2.3}$$

dla współczynnika skali s, przykładowo estymator MAD. Alternatywnie, możemy estymować s w podobny sposób.



Rys. 1. Przykłady funkcji ważących dla M-estymatorów

Wykorzystując MNW dla gęstości $s^{-1}f((x-\mu)/s)$, otrzymujemy równanie

$$\sum_i \psi\left(\frac{y_i-\mu}{s}\right)\left(\frac{y_i-\mu}{s}\right) = n \tag{2.4}$$

które nie jest odporne (oraz obciążone dla rozkładu normalnego). Trzeba to równanie zmodyfikować do

$$\sum_i \chi\left(\frac{y_i-\mu}{s}\right) = (n-1)\gamma \tag{2.5}$$

dla ograniczonej funkcji χ, gdzie γ jest wybierane tak, aby uzyskać zgodność z rozkładem normalnym, zatem $\gamma = E\chi(N)$.

Przykładem niech będzie następna propozycja Hubera:

$$\chi(x) = \psi(x)^2 = \min(|x|, c)^2 \tag{2.6}$$

W bardzo małych próbach należy skupić uwagę dodatkowo na zmienności $\hat{\mu}$ w przypadku zastosowania metryki Winsora (Huber, 1981). Jeżeli położenie μ jest znane, możemy zastosować ten estymator, zastępując $n-1$ przez n, celem estymacji jedynie współczynnika skali s.

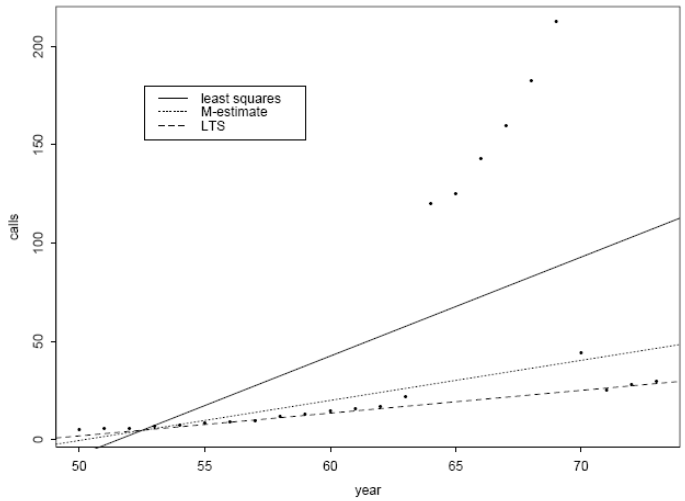
3. Własności modeli regresji

Omówimy koncepcję *odpornej regresji* w zakresie modeli liniowych, która mówi o niezagrażających zachowaniach w bieżących niewłaściwych wartościach danych. W terminologii, którą wprowadzimy, *odporna regresja* ma wysoki punkt załamania – proponujemy 50%.

Rozważymy zamianę metody najmniejszych kwadratów (MNK) przez jeden dwóch z metod:

1. LMS – najmniejsze medianowe kwadraty: minimalizują medianę kwadratów reszt. Bardziej ogólnie LQS – minimalizuje pewien kwantyl (przykładowo 80%) kwadratów reszt.
2. LTS – najmniejsze ucięte kwadraty: minimalizują sumę kwadratów najmniejszych q reszt. Oryginalnie q zawiera trochę powyżej 50%.

Omówione podejścia wymagają znacznie więcej obliczeń numerycznych niż metoda najmniejszych kwadratów, ponieważ nie mamy zachowanej różniczkowalności. Obydwie metody wychwytują efekt wielowymiarowych obserwacji oddalonych i koncentrują się na dobrym dopasowaniu, do co najmniej powyżej 50% danych. W konsekwencji są mniej efektywne w przypadku braku obserwacji odstających (LMS bardziej niż LTS). Aby zilustrować pewne problemy, rozważmy przykład. Rousseeuw i Leroy (1987) rozpatrują roczne dane liczby połączeń telefonicznych w Belgii (rys. 2). Zaprezentowano liniową funkcję regresji (MNK – least squares), regresję z M-estymacją oraz najmniejsze ucięte kwadraty reszt (LTS).



Rys. 2. Miliony połączeń telefonicznych w Belgii, 1950-1973

Źródło: (Rousseeuw, Leroy, 1987).

Linia LQS jest następująca: $\hat{y} = 56{,}16 + 1{,}16$ t (rok). Wykonane badania pokazują, że dla lat 1964–1969 powinna być badana całkowita długość połączeń (w minutach) w miejsce liczby połączeń (jak to było wykonywane w latach 1963-1970).

4. Odporna regresja*

Regresję odporną zdefiniowano w latach 80. XX w. (Huber, 1981). Pierwsza najbardziej znana regresja była określona następująco:

$$\min_{b} \underset{i}{mediana} |y_i - x_i b|^2$$

jako najmniejsze medianowe kwadraty (*least median of squares* – LMS).

Uzasadnieniem dla kwadratów reszt jest następująca obserwacja, gdy n jest parzyste, wówczas wybierana jest mediana. To jest bardzo odporna metoda regresji, dodatkowo nie wymaga estymacji parametru skali. Jest jednak bardzo nieefektywna pokrywa, co najwyżej $1/\sqrt[3]{n}$ danych. Dodatkowo, cechuje się wrażliwością na obserwacje centralne w zbiorze danych (Hettmansperger, Sheather, 1992; Davies, 1993, *§*2.3).

* *Resistant regression*.

Rousseeuw (1987) sugeruje regresję najmniejszych uciętych kwadratów (*least trimmed squares* – LTS):

$$\min_b \sum_i |y_i - x_i b|^2_{(i)} \tag{4.1}$$

Ta metoda jest bardziej efektywna, ale oddziela same krańcowe obserwacje. Rekomendowana suma kwadratów reszt nie powinna przekraczać wartości $q = [(n + p + 1)/2]$.

Następnie wprowadzono *S-estymatory*, dla których współczynniki równania są wyznaczane jako rozwiązanie zadania

$$\sum_{i=1}^{n} \chi\left(\frac{y_i - x_i b}{c_o s}\right) = (n-p)\beta \tag{4.2}$$

z najmniejszym parametrem skali s. W równaniu (4.2) jako funkcja χ jest zazwyczaj przyjmowana całkowalna podwójnie ważąca funkcja Tukey'a

$$\chi(u) = \begin{cases} u^6 - 3u^4 + 3u^2, & |u| \le 1 \\ 1, & |u| \ge 1 \end{cases}$$

Wartości $c_0 = 1{,}548$ i $\beta = 0{,}5$ są wyznaczane celem spełnienia warunku zgodności, jeżeli rozkład błędów jest rozkładem normalnym. To daje efektywność 28,7% przy rozkładzie normalnym, która jest niska, ale lepsza niż LMS i LTS.

Jedynie w kilku specjalnych przypadkach (LMS dla jednowymiarowej regresji ze stałą) możemy ten problem optymalizacyjny rozwiązać dokładnie, wykorzystując aproksymacyjne metody kolejnych przybliżeń (Marazzi, 1993). Wiele tych metod wykorzystuje podejście metody najmniejszych kwadratów, proponując dopasowanie najmniejszych kwadratów dla wybranych q punktów ze zbioru danych. Następne losowo sprawdzają duże próby dla tego dopasowania.

5. Mocna regresja

W modelu regresji mamy dwa podstawowe źródła błędów: wartości obserwacji y_i oraz odpowiadający wektor p^* wartości zmiennych objaśniających $\boldsymbol{x}_i$ (*regressors*). Większość metod regresji rozważa jedynie pierwszy rodzaj źródła błędów. W pewnych okolicznościach (przykładowo planowanie eksperymentów) błędy zmiennych objaśniających mogą być ignorowane. Tak jest w przypadku M-estymatorów, którymi zajmiemy się w tym punkcie.

---

* n obserwacji (y_i, x_{i1},..., x_ip)

Rozważmy problem regresji dla n przypadków (y_i, $\boldsymbol{x}_i$) z modelu

$$y = \boldsymbol{x}\beta + \varepsilon \tag{5.1}$$

dla p-wymiarowego wektora $\boldsymbol{x}$.

M-estymatory

Przyjmujemy skalowanie dla funkcji gęstości $f(e/s)/s$ dla ε oraz przyjmujemy $\rho = -\log f$, wówczas estymator maksymalnej wiarygodności minimalizuje

$$\sum_{i=1}^{n} \rho\left(\frac{y_i - x_i b}{s}\right) + n \log s \tag{5.2}$$

Załóżmy, że s jest znane oraz funkcja $\psi = \rho'$. Wówczas w MNW, wyznaczając b celem estymacji β, rozwiązujemy nieliniowe równanie:

$$\sum_{i=1}^{n} x_i \psi\left(\frac{y_i - x_i b}{s}\right) = 0 \tag{5.3}$$

Zapiszmy reszty jako: $r_i = y_i - \boldsymbol{x}_i b$. Rozwiązanie równania (5.3) lub minimalizacja względem (5.2) definiuje M-estymatory względem współczynników β.

Znaną metodą rozwiązania (5.3) jest metoda iteracyjna ważonych najmniejszych kwadratów, z wagami określonymi następująco:

$$w_i = \psi\left(\frac{y_i - \mu}{s}\right) \Big/ \left(\frac{y_i - \mu}{s}\right) \tag{5.4}$$

Iteracja jest zbieżna jedynie dla wypukłych (*convex*) funkcji ρ oraz dla niemalejących (Tukey, 1960), a równanie (5.3) może mieć wiele pierwiastków. W takich przypadkach należy wybrać dobry punkt startowy i uważnie przeprowadzić iterację.

W zastosowaniach współczynnik skali s jest nieznany. Łatwy i odporny estymator współczynnika skali to MAD (względem pewnego przyjętego centrum). Można go zastosować dla reszt bliskich zero, również dla pozostających w pewnym w otoczeniu albo dla reszt z odpornego dopasowania. Alternatywnie, możemy estymować s, wykorzystując „prawie" MNW estymatory (MLE-*like way*). Znajdując punkt stacjonarny równania (5.2) względem s, otrzymujemy:

$$\sum_i \psi\left(\frac{y_i - x_i b}{s}\right)\left(\frac{y_i - x_i b}{s}\right) = n \tag{5.5}$$

Rozwiązanie nie jest odporne oraz obciążone dla rozkładu normalnego (Venables, Ripley, 2002).

W przypadku jednowymiarowym możemy to równanie zmodyfikować przekształcając do postaci:

$$\sum_i \chi\left(\frac{y_i - x_i b}{s}\right) = (n-p)\gamma \tag{5.6}$$

MM-estymacja

Możliwe jest połączenie odporności oraz efektywności M-estymatorów. Takim rozwiązaniem jest MM-estymator zaproponowany przez Yohai, Stahel i Zamar (1991)*. MM-estymator to M-estymator, który wykorzystuje współczynniki wyznaczone na pierwszym etapie przez S-estymator oraz stały współczynnik skali dany przez S-estymator. To pozwala uzyskać (dla $c > c_0$) wysoki punkt załamania S-estymatorów oraz wysoką efektywność dla rozkładu normalnego. Przy znacznych kosztach obliczeń otrzymujemy to, co najlepsze z obydwu omówionych podejść.

Podsumowanie

W przedstawionym artykule omówiono wybrane statystyki odporne podstawowych parametrów wraz z ich własnościami. W szczególności omówiono wybrane estymatory parametrów położenia i skali. Zwrócono uwagę na podstawowe uwarunkowania odpornej regresji. Stosując klasyczne estymatory, nie wracamy do założeń, które towarzyszą metodom wyznaczania tych estymatorów. Brak spełnienia tych założeń powoduje trudności w wyznaczaniu rozwiązań formułowanych zadań. Estymatory odporne wymagają zastosowania metod przybliżonych, iteracyjnych. Celem efektywnego wyznaczenia wartości tych estymatorów ważne jest spojrzenie na własności numeryczne metod iteracyjnych stosowanych do rozwiązań zapisanych zadań. Wiele programów komputerowych wspomagających procesy analizy danych, takich jak S-Plus czy *Statistica* lub SAS, mają funkcje powiązane ze statystycznymi metodami odpornymi. Badanie porównawcze efektywności tych metod iteracyjnych jest odrębnym zadaniem powiązanym ze statystyką odporną.

Bibliografia

Davies P.L. (1993): *Aspects of Robust Linear Regression*. „Annals of Statistics", 21, s. 1843-1899.

Dixon W.J. (1960): *Simplified Estimation for Censored Normal Samples*. „Annals of Mathematical Statistics", 31, s. 385-391.

* (Zob. Marazzi, 1993).

Hampel F.R., Ronchetti E.M., Rousseeuw P.J., Stahel W.A. (1986): *Robust Statistics. The Approach Based on Influence Functions*. John Wiley and Sons, New York.

Hettmansperger T.P., Sheather S.J. (1992): *A Cautionary Note on the Method of Least Median Squares*. „American Statistician" 46, s. 79-83

Huber P.J. (1981): *Robust Statistics*. John Wiley and Sons, New York.

Iglewicz B. (1983): Robust *Scale Estimators and Confidence Intervals for Location*. W: *Understanding Robust and Exploratory Data Analysis*. Eds. D.C. Hoaglin, F. Mosteller, J.W. Tukey. John Wiley and Sons, New York, s. 405-431.

Marazzi A. (1993): *Algorithms, Routines and* S *Functions for Robust Statistics*. Wadsworth and Brooks/Cole. Pacific Grove, CA.

Rousseeuw, P. J., Leroy, A.M. (1987): *Robust Regression and Outlier Detection*. John Wiley and Sons, New York.

Staudte R.G., Sheather S.J. (1990): *Robust Estimation and Testing*. John Wiley and Sons, New York.

Trzpiot G. (2009): *Extreme Value Distributions and Robust Estimation*. Acta Universitatis Lodziensis. Folia Economica 228, Łódź, s. 85-92.

Trzpiot G. (2011a): *Odporna analiza szeregów czasowych*. Prace Naukowe nr 165, 171-179 Uniwersytet Ekonomiczny, Wrocław.

Trzpiot G. (2011b): *Wybrane odporne metody estymacji beta*. W: *Modelowanie preferencji a ryzyko '11*. Red. T. Trzaskalik. Wydawnictwo Uniwersytetu Ekonomicznego, Katowice, s. 133-148.

Trzpiot G. (2012): *Odporna regresja kwantylowa*. W: *Dylematy ekonometrii 2*. Red. J. Biolik. Wydawnictwo Uniwersytetu Ekonomicznego, Katowice, s. 147-158.

Tukey J.W. (1960): *A Survey of Sampling from Contaminated Distributions*. W: *Contributions to Probability and Statistics*. Eds I. Olkin, S. Ghurye, W. Hoeffding, W. Madow, H. Mann. Wiley and Sons, New York.

Yohai V., Stahel W.A., Zamar R.H. (1991): *A Procedure for Robust Estimation*. John Wiley and Sons, New York.

Venables W.N., Ripley B.D. (2002): *Modern Applied Statistics with S-PLUS*. Springer-Verlag.

SOME ROBUST STATISTICAL METHODS

Summary

Outliers are sample values that cause surprise in relation to the majority of the sample. This is not a pejorative term; outliers may be correct, but they should always be checked for transcription errors. Many *robust* and *resistant* methods have been developed since 1960 to be less sensitive to outliers. This methods can be used instead or be even better than classical one. Robust methods were used early in me works (Trzpiot 2009, 2011a, 2011b) as an application in finance and economy. This article has a descriptive character, connected with new book for students.

Janusz L. Wywiał
Uniwersytet Ekonomiczny w Katowicach

SAMPLING DESIGNS PROPORTIONATE TO NON-NEGATIVE FUNCTIONS OF TWO QUANTILES OF AUXILIARY VARIABLE

1. Sampling design

Let U be a fixed population of size N. The observation of a variable under study and an auxiliary variable are denoted by y_i and x_i, $i = 1, \ldots, N$, respectively. Moreover, let $x_i \leq x_{i+1}$, $i = 1, \ldots, N-1$. Our problem is estimation of the population average $\bar{y} = \frac{1}{N}\sum_{k\in U} y_k$.

Let us consider the sample space $\mathbf{S}$ of the samples s of the fixed effective size $1 < n < N$. The sampling design is denoted by $P(s)$. We assume that $P(s) \geq 0$ for all $s \in \mathbf{S}$ and $\sum_{s\in\mathbf{S}} P(s) = 1$.

Let $(X_{(j)})$ be the sequence of the order statistics of observations of auxiliary variable in the sample s. It is well known that the sample quantile of order a $\alpha \in (0;1)$ is defined as follows: $Q_{s,\alpha} = X_{(r)}$ where $r = [n\alpha]+1$, the function $[n\alpha]$ means the integer part of the value $[n\alpha]$, $r = 1, 2, \ldots, n$. Let us note that $X_{(r)} = Q_{s,\alpha}$ for $\frac{r-1}{n} \leq \alpha < \frac{r}{n}$. In this paper it will be more conveniently to consider the order statistic than the quantile.

Let $G(r,u,i,j) = \{s : X_{(r)} = x_i, X_{(u)} = x_j\}$, $r = 1, \ldots, n-1$; $u = 2, \ldots, n$, $r < u$ be the set of all samples whose r-th and u-th order statistics of the auxiliary variable are equal to x_i and x_j, respectively where $r \leq i < j \leq N-n+u$. Moreover,

$$\bigcup_{i=r}^{\mathbf{N-n+r}} \bigcup_{j=i+u-r}^{N-n+u} G(r,u,i,j) = \mathbf{S} \tag{1}$$

The size of the set $G(r,u,i,j)$ is denoted by $g(r,u,i,j) = Card(G(r,u,i,j))$ and

$$g(r,u,i,j) = \binom{i-1}{r-1}\binom{j-i-1}{u-r-1}\binom{N-j}{n-u} \tag{2}$$

Let $f(x_j, x_j, c)$ be non-negative function of values of the order statistics $X_{(r)}$ and $X_{(r)}$ and

$$g(r,u,i,j) = \binom{i-1}{r-1}\binom{j-i-1}{u-r-1}\binom{N-j}{n-u} \tag{3}$$

The straightfoward generalization of the Wywiał's (2009) sampling design is as follows.

Definition 1.1. The conditional sampling design proportional to the non-negative functions of the order statistics $X_{(u)}, X_{(r)}$ is as follows:

$$P_{r,u}(s \mid c) = \frac{f(X_{(u)}, X_{(r)}, x)}{z(r,u,c)} \tag{4}$$

where $i < j$ and $r \le i \le N-n+r$ and $r < u \le j \le N-n+u$.

Particularly, let $f(x_j, x_i, c) = x_j - x_i$ and

$$f(x_j, x_i, c) = \begin{cases} x_j - x_i \ \textit{for}\ x_j - x_i \ge c, \\ 0 \ \textit{for}\ x_j - x_i < c. \end{cases} \tag{5}$$

We say that the above sampling design is (conditional) unconditional when $(c > 0)$ $c = 0$. In general this concept is agree with definition of the conditional sampling design considered by Tillé (1999; 2006).

As it is well known the inclusion probability of the first order is determined by the equations: $\pi_k(r,u,c) = P_{r,u}(s : k \in s) = \sum_{\{s:k\in s\}} P_{r,u}(s \mid c)$, $k = 1, \ldots, N$. We assume that if $x \le 0$ then $\delta(x) = 0$ otherwise $\delta(x) = 1$. Let us note that $\delta(x)\delta(x-1) = \delta(x-1)$.

Theorem 1.1. Under the sampling design $P_{r,u}(s)$ the inclusion probabilities of the first degree are as follows. If $k < r$,

$$\pi_k(r,u,c) = \frac{\delta(r-1)\delta(r-k)}{z(r,u,c)} \sum_{i=r}^{N-n+r} \sum_{j=i+u-r}^{N-n+u} \binom{i-2}{r-2}\binom{j-i-1}{u-r-1}\binom{N-j}{n-u} f(x_j, x_i, c) \tag{6}$$

If $r \le k \le N-n+u$,

$$\pi_k(r,u,c) = \frac{\delta(r-1)\delta(r-k)}{z(r,u,c)} \cdot$$
$$\cdot \Bigg(\delta(k-u)\delta(n-u) \sum_{i=r}^{k-u+r-1} \sum_{j=i+u-r}^{k-1} \binom{i-1}{r-1}\binom{j-i-1}{u-r-1}\binom{N-j-1}{n-u-1} f(x_j, x_i, c) +$$
$$+ \delta(k-u+1)\binom{N-k}{n-u} \sum_{i=r}^{k-u+r} \binom{i-1}{r-1}\binom{k-i-1}{u-r-1} f(x_j, x_i, c) +$$
$$+ \delta(k-r)\delta(N-n+u-k)\delta(u-r-1) \sum_{i=r}^{k-1} \sum_{j=k+1}^{N-n+u} \binom{i-1}{r-1}\binom{j-i-1}{u-r-1}\binom{N-j}{n-u} f(x_j, x_i, c) + \quad (7)$$
$$+ \delta(N-n+r-k+1)\binom{k-1}{r-1} \sum_{j=k+u-r}^{N-n+u} \binom{j-k-1}{u-r-1}\binom{N-j}{n-u} f(x_j, x_k, c) +$$
$$+ \delta(N-n+r-k)\delta(r-1) \sum_{i=k+1}^{N-n+r} \sum_{j=i+u-r}^{N-n+u} \binom{i-2}{r-2}\binom{j-i-1}{u-r-1}\binom{N-j}{n-u} f(x_j, x_i, c) \Bigg)$$

If $k > N-n+u$,

$$\pi_k(r,u,c) = \frac{\delta(k-N+n-u)}{z(r,u,c)} \sum_{i=r}^{N-n+r} \sum_{j=i+u-r}^{N-n+u} \binom{i-1}{r-1}\binom{j-i-1}{u-r-1}\binom{N-j-1}{n-u-1} f(x_j, x_i, c) \quad (8)$$

The inclusion probabilities of the second order are defined by

$$\pi_{k,t}(r,u,c) = \pi_{t,k}(r,u,) = P_{r,u}(s : k \in s, t \in s) = \sum_{\{s:k \in s, t \in s\}} P_{r,u}(s \mid c)$$

where $k < t = 1, ..., N$.

Theorem 1.2. The inclusion probabilities of the second degree of the sampling design $P_{r,u}(s \mid c)$ are as follows.

$$\pi_{k,t}(r,u,c) = P(k,t \in s_3) + P(X_{(u)} = x_k, t \in s_3) + P(k \in s_2, t \in s_3) +$$
$$+ P(X_{(r)} = x_k, t \in s_3) + P(k \in s_1, t \in s_3) + P(k \in s_2, X_{(u)} \in x_t) + \quad (9)$$
$$+ P(X_{(r)} = x_k, X_{(u)} = x_t) + P(k \in s_1, X_{(u)} = x_t) + P(k,t \in s_2) +$$
$$+ P(X_{(r)} = x_k, t \in s_2) + P(k \in s_1, t \in s_2) + P(k \in s_1, X_{(r)} \in x_t) + P(k,t \in s_1)$$

where

$$P(k,t\in s_3)=\frac{\delta(n-u-1)}{z(r,u,c)}\big(\delta(k-N+n-u)\cdot$$
$$\cdot\sum_{i=r}^{N-n+r}\sum_{j=i+u-r}^{N-n+u}\binom{i-1}{r-1}\binom{j-i-1}{u-r-1}\binom{N-j-2}{n-u-2}f(x_j,x_i,c)+\delta(N-n+u-k+1\_\delta(k-u)\cdot$$
$$\cdot\sum_{i=r}^{N-u+r}\sum_{j=k-1}^{N-n+u}\binom{i-1}{r-1}\binom{j-i-1}{u-r-1}\binom{N-j-2}{n-u-2}f(x_j,x_i,c)\Big), \tag{10}$$

$$P(X_{(u)}=x_k,t\in s_3)=\frac{\delta(N-k)\delta(n-u)\delta(N-n+u-k+1)}{z(r,u,c)}\cdot$$

$$\cdot\binom{N-k-1}{n-u-1}\sum_{i=r}^{k-n+r}\binom{i-1}{r-1}\binom{k-i-1}{u-r-1}f(x_j,x_i,c), \tag{11}$$

$$P(k\in s_2,t\in s_3)=\frac{\delta(u-r-1)\delta(n-u)\delta(t-k-u+r+1)}{z(r,u,c)}\cdot$$
$$\cdot\Big(\delta(N-n+u-t+1\sum_{i=r}^{k-1}\sum_{j=i+u-r}^{t-1}\binom{i-1}{r-1}\binom{j-i-2}{u-r-2}\binom{N-j-1}{n-u-1}f(x_j,x_i,c)+ \tag{12}$$
$$+\delta(t-N+n-u)\sum_{i=r}^{k-1}\sum_{j=i+u-r}^{N-n+u}\binom{i-1}{r-1}\binom{j-i-2}{u-r-2}\binom{N-j-1}{n-u-1}f(x_j,x_i,c)\Big),$$

$$P(X_{(r)}\in x_k,t\in s_3)=\frac{\delta(n-u)\delta(N-k)\delta(k-r+1)}{z(r,u,c)}\cdot$$
$$\cdot\Big(\delta(t-k-u+r)\delta(N-n+u-t+1)\binom{k-1}{r-1}\sum_{j=k+u-r}^{t-1}\binom{j-k-1}{u-r-1}\binom{N-j-1}{n-u-1}f(x_j,x_i,c)+ \tag{13}$$
$$+\delta(N-n-k+r+1)\delta(t-N+n-u)\binom{k-1}{r-1}\sum_{j=k+u-r}^{N-n+u}\binom{j-k-1}{u-r-1}\binom{N-j-1}{n-u-1}f(x_j,x_k,c)\Big),$$

$$P(X\in s_1,t\in s_3)=\frac{\delta(r-1)\delta(n-u)}{z(r,u,c)}\big(\delta(r-k)\delta(t-N+n-u)\cdot$$
$$\cdot\sum_{i=r}^{N-n+r}\sum_{j=i+u-r}^{N-n+u}\binom{i-2}{r-2}\binom{j-i-1}{u-r-1}\binom{N-j-1}{n-u-1}f(x_j,x_i,c)+\delta(r-k)\delta(N-n+u-t+1)\delta(t-u)\cdot$$
$$\cdot\sum_{i=r}^{t-u+r-1}\sum_{j=i+u-r}^{t-1}\binom{i-2}{r-2}\binom{j-i-1}{u-r-1}\binom{N-j-1}{n-u-1}f(x_j,x_i,c)+ \tag{14}$$
$$+\delta(k-r+1)\delta(t-N+n-u)\delta(N-n+r-k)\delta(t-u)\cdot$$
$$\cdot\sum_{i=k+1}^{t-u+r-1}\sum_{j=i+u-r}^{t-1}\binom{i-2}{r-2}\binom{j-i-1}{u-r-1}\binom{N-j-1}{n-u-1}f(x_j,x_i,c)\Big),$$

$$P(X \in s_2, X_{(u)} = x_t) = \frac{\delta(t-u)\delta(u-r)}{z(r,u,c)} \binom{N-t}{n-u} \sum_{i=r}^{t-u+r-1} \binom{i-1}{r-1} \binom{t-i-1}{u-r-1} f(x_j, x_i, c) \cdot$$
$$\cdot \sum_{i=r}^{t-u+r-1} \sum_{j=i+u-r}^{t-1} \binom{i-2}{r-2} \binom{j-i-1}{u-r-1} \binom{N-j-1}{n-u-1} f(x_j, x_i, c) + \tag{15}$$
$$+ \delta(k-r+1)\delta(t-N+n-u)\delta(N-n+r-k)\delta(t-u) \cdot$$
$$\cdot \sum_{i=k+1}^{t-u+r-1} \sum_{j=i+u-r}^{t-1} \binom{i-2}{r-2} \binom{j-i-1}{u-r-1} \binom{N-j-1}{n-u-1} f(x_j, x_i, c) \Bigg),$$

$$P(X_{(r)} = x_k, X_{(u)} \in x_t) =$$
$$= \frac{\delta(k-r+1)\delta(N-n+u-t+1)}{z(r,u,c)} \binom{k-1}{r-1} \binom{t-k-1}{u-r-1} \binom{N-t}{n-u} f(x_j, x_i, c), \tag{16}$$

$$P(k \in s_1, X_{(u)} = x_t) = \frac{\delta(r-1)\delta(u-r)}{z(r,u,c)} \cdot$$
$$\cdot \Bigg(\delta(r-k)\delta(t-u+1) \binom{N-t}{n-u} \sum_{i=r}^{t-u+r} \binom{i-2}{r-2} \binom{t-i-1}{u-r-1} f(x_j, x_i) + \tag{17}$$
$$\cdot \delta(k-r+1)\delta(t-u+r-k) \binom{N-t}{n-u} \sum_{i=k+1}^{t-u+r} \binom{i-2}{r-2} \binom{t-i-1}{u-r-1} f(x_j, x_i) \Bigg),$$

$$P(k, t \in s_2) = \frac{\delta(u-r-t+k-1)\delta(u-r-2)}{z(r,u,c)} \cdot$$
$$\cdot \sum_{i=r}^{k-1} \sum_{j=t+1}^{N-n+u} \binom{i-1}{r-1} \binom{j-i-3}{u-r-3} \binom{N-j}{n-u} f(x_j, x_i, c), \tag{18}$$

$$P(X_{(r)} \in x_k, t \in s_2) = \frac{\delta(N-n-k+r+1)\delta(u-r-1)}{z(r,u,c)} \cdot$$
$$\cdot \binom{k-1}{r-1} \sum_{j=k+u-r}^{N-n+u} \binom{j-k-2}{u-r-2} \binom{N-j}{n-u} f(x_j, x_i, c), \tag{19}$$

$$P(k \in s_1, t \in s_2) = \frac{\delta(r-1)\delta(u-r-1)\delta N-n+u-t)}{z(r,u,c)} \cdot$$

$$\cdot \left(\delta(r-k) \sum_{i=r}^{t-1} \sum_{j=t+1}^{N-n+u} \binom{i-2}{r-2} \binom{j-i-2}{u-r-2} \binom{N-j}{n-u} f(x_j, x_i, c) + \right. \tag{20}$$

$$\left. + \delta(k-r+1) \sum_{i=k+1}^{t-1} \sum_{j=t+1}^{N-n+u} \binom{i-2}{r-2} \binom{j-i-2}{u-r-2} \binom{N-j}{n-u} f(x_j, x_i, c) \right),$$

$$P(k \in s_1, X_{(r)} = x_t) = \frac{\delta(r-1)\delta(k-1)\delta(N-n+u-t)}{z(r,u,c)} \cdot$$

$$\cdot \binom{k-2}{r-2} \sum_{j=t+1}^{N-n+u} \binom{j-k-1}{u-r-1} \binom{N-j}{n-u} f(x_j, x_k, c), \tag{21}$$

$$P(k, t \in s_1) = \frac{\delta(r-2)}{z(r,u,c)} \left(\delta(r-t) \sum_{i=r}^{N-n+r} \sum_{j=i+u-r}^{N-n+u} \binom{i-3}{r-3} \binom{j-i-1}{u-r-1} \binom{N-j}{n-u} f(x_j x_i, c) + \right.$$

$$\left. + \delta(t-r-1) \sum_{i=t+1}^{N-n+r} \sum_{j=i+u-r}^{N-n+u} \binom{i-3}{r-3} \binom{j-i-1}{u-r-1} \binom{N-j}{n-u} f(x_j x_i, c). \right. \tag{22}$$

2. Sampling scheme

The sampling scheme implementing the sampling design $P_{r,u}(s \mid c)$ P is as follows. Firstly, population elements are ordered according to increasing values of the auxiliary variable. Let $s = s_1 \cup \{i\} \cup s_3 \cup \{j\} \cup s_3$, $s_1 = \{k : k \in U, x_k < x_i\}$, $s_2 = \{k : k \in U, x_j > x_k > x_i\}$ and $s_3 = \{k : k \in U, x_k > x_j\}$. Moreover, let $U = U(1, i-1) \cup \{i\} \cup U(i+1, j-1) \cup \{j\} \cup U(j+1, N)$ where $U(1, i-1) =$ $= (1, \ldots, i-1)$, $U(i+1, j-1) = (i+1, \ldots, j-1)$, $U(j+1, N) = (j+1, \ldots, N)$. Let $S(U(1, i-1); s)$ be sample space of the sample s_1 of size $r-1, S(U(i+1, j-1); s)$ be sample space of the sample s_2 of size $u-r-1, S(U(j+1, N); s)$ be sample space of the sample s_2 of size $n-u$. Similarly, $S = S(U, s)$.

The sampling scheme is given by the following of probabilities:

$$P_{r,u}(s \mid c) = P_1(s_1) p_{r,u}(i \mid c) P_2(s_2) p'_{r,u}(j \mid c) P_3(s_3) \tag{23}$$

where

$$P_1(s_1)=\binom{i-1}{r-1}^{-1},P_2(s_2)=\binom{j-i-1}{u-r-1}^{-1},P_3(s_3)=\binom{N-j}{n-u}^{-1} \tag{24}$$

$$P_{r,u}(i\mid c)=P(X_{(r)}=x_i\mid X_{(u)}=x_j,c)=\frac{P_{r,u}(X_{(r)}=x_i,X_{(u)}=x_j,c)}{P_{r,u}(X_{(u)}=x_j,c)} \tag{25}$$

$$p'_{r,u}(j\mid c)=P_{r,u}(X_{(u)}=x_j,c)=\frac{1}{z(r,u,)}\sum_{i=r}^{N-n+r}f(x_jx_i,c)g(r,u,i,j) \tag{26}$$

$$P_{r,u}(X_{(r)}=x_i\mid X_{(u)}=x_j,c)=\sum_{s\in G(r,u,i,j)}P_{r,u}(s\mid c)\frac{f(x_jx_i,c)g(r,u,i,j)}{z(r,u,c)} \tag{27}$$

In order to select the sample s, firstly the j-th element of the population should be selected, according to the probability function $p'_{r,u}=(j\mid c)$. Next, the i-the element of the population should be drawn according to the probability function $p'_{r,u}=(i\mid c)$. Finally, the samples s_1, s_2 and s_3 should be selected, according to the sampling designs $P_1(s_1)$, $P_2(s_2)$ and $P_3(s_3)$, respectively.

3. Some sampling strategies

The well known Horvitz-Thompson (1952) estimator is as follows:

$$\bar{y}_{HT,s}=\frac{1}{N}\sum_{k\in s}\frac{y_k}{\pi_k} \tag{28}$$

The statistic is unbiased estimator of the population mean value if $\pi_k>0$ for $k=1,...,N$. The variance and its estimator are determined by the expressions (32) and (34), respectively.

The particular case of the above estimator is the well known sampling design of the simple sample drawn without replacement is as follows: $P_0(s)=\binom{N}{n}^{-1}$. The variance of the mean from the simple sample

$\bar{y}_s = \frac{1}{n}\sum_{k\in s} y_k$ drawn without replacement is $D^2(\bar{y}_s, P_0(s)) = \frac{N-n}{nN} v_y$ where $v_y = \frac{1}{N-1}\sum_{k=1}^{N}(y_k - \bar{y})^2$.

The results of the previous chapter lead to construction of the regression sampling strategy for the population mean $\bar{y}_s = \frac{1}{n}\sum_{ks} y_i$. We assume that $y_i = a + bx_i + e_i$ for all $i \in U$, $\sum_{i\in U} e_i = 0$ and the residuals of that linear regression function are not correlated with the auxiliary variable. The linear correlation coefficient between the variables y and x will be denoted by ρ . Let $(X_{(r)}, Y_r)$ be two dimensional random variable where $X_{(r)}$ is the *r*-th order statistic of an auxiliary variable and Y_r is the variable under study. Wywiał (2009) considered the following estimator:

$$\bar{y}_{r,u,s} = \bar{y}_{HT,s} + b_{r,u,s}(\bar{x} - \bar{x}_{HT,s}) \tag{29}$$

where

$$b_{r,u,s} = \frac{Y_u - Y_r}{X_{(u)} - X_{(r)}} \tag{30}$$

Wywiał (2009) showed that under the sampling design stated in the definition 1.1 the parameters of the following strategies $\bar{y}_{r,u,s}, P_{r,u}(s \mid c))$, are approximately as follows.

$$E\left(\bar{y}_{r,u,s}, P_{r,u}(s \mid c)\right) \approx \bar{y}$$

$$D^2(\bar{y}_{r,u,s}, P_{r,u}(s \mid c)) \approx D^2(\bar{y}_{HT,s}, P_{r,u}(s \mid c)) - 2bCov(\bar{y}_{HT,s}, \bar{x}_{HT,s}, P_{r,u}(s \mid c)) + \\ + b^2 D^2(\bar{x}_{HT,s}, P_{r,u}(s \mid c)) \tag{31}$$

where

$$Cov(\bar{y}_{HT,s}, \bar{x}_{HT,s}, P_{r,u}(s \mid c)) = \frac{1}{N^2}\left(\sum_{k\in U}\sum_{l\in U} \Delta_{k,l} \frac{y_k}{\pi_k}\frac{x_l}{\pi_l}\right) \tag{32}$$

$$\Delta_{k,l} = \pi_{k,l} - \pi_k, \pi_l,$$
$$D^2(\bar{x}_{HT,s}, P_{r,u}(s \mid c)) = Cov(\bar{x}_{HT,s}, \bar{x}_{HT,s}, P_{r,u}(s \mid c)),$$
$$D^2 = \bar{y}_{HT,s}, P_{r,u}(s \mid c)) = Cov(\bar{y}_{HT,s}, \bar{y}_{HT,s}, P_{r,u}(s \mid c)).$$

The approximately unbiased estimator of the variance: $D^2 = \bar{y}_{r,u,s}, P_{r,u}(s \mid c))$ is as follows

$$\hat{D}^2(\bar{y}_{r,u,s}, P_{r,u}(s \mid c)) = \hat{D}^2(\bar{y}_{HT,s}, P_{r,u}(s \mid c)) - 2b_{r,u,s}\hat{Cov}(\bar{y}_{HT,s}, \bar{x}_{HT,s}P_{r,u}(s \mid c)) + \\ + b_{r,u,s}^2 \hat{D}^2(\bar{x}_{HT,s}, P_{r,u}(s \mid c)) \tag{33}$$

where

$$\hat{Cov}(\bar{y}_{HT,s}, \bar{x}_{HT,s}, P_{r,u}(s \mid c)) = \frac{1}{N^2}\left(\sum_{k\in s}\sum_{l\in s}\Delta_{*,k,l}\frac{y_k}{\pi_k}\frac{x_l}{\pi_l}\right) \tag{34}$$

$$\Delta_{*k,l} = \frac{\Delta_{k,l}}{\pi_{k,l}},\ \hat{D}^2(\bar{x}_{HT,s}, P_{r,u}(s \mid c)) = \hat{Cov}(\bar{x}_{HT,s}, \bar{x}_{HT,s}, P_{r,u}(s \mid c)),$$

$$\hat{D}^2(\bar{y}_{r,u,s}, P_{r,u}(s \mid c)) = \hat{Cov}(\bar{y}_{HT,s}, \bar{x}_{HT,s}P_{r,u}(s \mid c))$$

The next regression type estimator is given by:

$$\tilde{y} = \bar{y}_{HT,s} + \bar{b}_s(\bar{x} - \bar{x}_{HT,s}) \tag{35}$$

$$\tilde{b}_s = \frac{\tilde{v}_{x,y,s}}{\tilde{v}_{x,s}} \tag{36}$$

where

$$\tilde{v}_{x,y,s} = \frac{1}{\tilde{N}-1}\sum\frac{(x_k - \tilde{x}_{HT,s})(y_k - \tilde{y}_{HT,s})}{\pi_k},\ \tilde{v}_{x,s} = \tilde{v}_{x,x,s},$$

$\tilde{N} = \sum_{k\in s}\frac{1}{\pi_k}$, $\tilde{x}_{HT,s} = N\bar{x}_{HT,s}/\tilde{N}$, $\tilde{y}_{HT,s} = N\bar{y}_{HT,s}/\tilde{N}$. The statistics $\tilde{v}_{x,y,s}$ and $\tilde{v}_{x,s}$ are the consistent estimators of the population covariance and variance, $v_{xy} = \frac{1}{N-1}\sum_{k\in U}(x_k - \bar{x})(y_k - \bar{y})$, respectively (see e.g. Särndal et al., 1997, p. 187).

The strategy $(\tilde{y}_s, P_{r,u}(s \mid c))$ has approximately the same parameters as the above considered strategy $(\tilde{y}_{r,u,s}, P_{r,u}(s \mid c))$.

Let us remember that the ordinary regression estimator is given by:

$$\hat{y}_s = \bar{y}_s + b_s(\bar{x} - \bar{x}_s) \tag{37}$$

where

$$b_s = \frac{\sum_{k\in s}(x_k - \bar{x}_s)(y_k - \bar{y}_s)}{\sum_{k\in s}(x_k - \bar{x}_s)^2} \tag{38}$$

The approximate value of the variance is as follows.

$$D^2(\hat{y}_s, P_0(s)) = \frac{N-n}{Nn} v_y (1-\rho^2) \tag{39}$$

where $\rho = \frac{v_{x,y}}{\sqrt{v_x v_y}}$.

The approximately unbiased estimator of the variance is as follows.

$$D^2(\hat{y}_s, P_0(s)) = \frac{N-n}{Nn} v_{y,s} (1-r_s^2) \tag{40}$$

where $r_s = \frac{v_{x,y,s}}{\sqrt{v_{x,s} v_{y,s}}}$, $v_{xy,s} = \frac{1}{n-1}\sum_{k\in s}(x_k - \bar{x}_s)(y_k - \bar{y}_s)$, $v_{x,s} = v_{xx,s}$, $v_{y,s=v_{yy,s}}$

The strategy $(\hat{y}_s, P_0(s))$ is asymptotically unbiased for the population mean $\bar{y}$. It is well known that the strategy $(\hat{y}_s, P_{S,S}(s))$ is unbiased for $\bar{y}$ where

$$P_{S,S}(s) = \frac{v_{x,s}}{\binom{N}{n} v_x} \tag{41}$$

is the sampling design of Singh and Srivastava (1980). The strategies $\hat{y}_s, P_{SS}(s)$ and $(t_{HT,s}, P_{SS}(s))$ are unbiased for the population mean $\bar{y}$. The variances of the strategy $(\hat{y}_s, P_{S,S}(s))$ is approximately equal to the right side of the equation (39).

4. Simulation analysis of strategies accuracy

Let $MSE(t, P(s))$ be the mean square error of the strategy $(t, P(s))$ used to estimate the population mean $\bar{y}$. The coefficient of the relative efficiency we define as follows:

$$e(t, P(s)) = \frac{MSE(t, P(s))}{D^2(\bar{y}_s, P_0(s))} 100\%$$

where $\left(\bar{y}_s, P_0(s)\right)$ is ordinary simple sample mean. Let $e1, e2, e3, e4, e5$ and $e6$ be the relative efficiency coefficients of the strategies $\left(\hat{y}_s, P_0(s)\right)$, $\left(\hat{y}_s, P_{SS}(s)\right)$, $\left(\bar{y}_{HT,s}, P_{SS}(s)\right)$, $\left(\bar{y}_{r,u,s}, P_{r,u}(s \mid c)\right)$, $\left(\bar{y}_s, P_{r,u}(s \mid c)\right)$ and $\left(\bar{y}_{HT,s}, P_{r,u}(s \mid c)\right)$, respectively.

The population consists of the municipalities in Sweden. The auxiliary variable x is: 1975 municipal population (in thousands) and the variable under study y is: 1985 municipal taxation revenues (in millions of kronor). Their observations have been published by Srndal, Swenson and Wretman (1992). The size of this population is 284 municipalities. There are three outlier observations of the variables, see Figure 1. Let d and β_3 be the standard deviation and the skewnees coefficient, respectively, of the auxiliary variable in the population. In the case of data without outliers (size of the population N = 281) $\bar{x}$ = 24, 263, d = 24,153 and β_3 = 0, 043. In the case of data with outliers (size of the population N = 284) $\bar{x}$ = 28,810, d = 52, 873 and β_3 = 8, 427. The samples were replicated 1000 times.

In general, Tables 1, 2, 3 let us infer that the regression type strategies are the best among the considered ones.

Table 1

The relative efficiency coefficients (%) of the strategies

N :	281			284		
n	e1	e2	e3	e1	e2	e3
2 (0,7%)	80,5	6,9	30,6	5,0	5,8	10,0
3 (1%)	6,6	2,8	22,7	1,7	5,3	9,3
4 (1,5%)	4,7	2,6	24,8	1,8	4,0	8,6
6 (2%)	4,2	2,7	28,9	1,7	3,7	9,6
9 (3%)	4,1	2,7	35,0	1,9	3,3	12,2
11 (4%)	3,4	2,6	42,6	2,4	3,3	13,1
14 (5%)	3,8	2,6	45,0	2,6	3,3	14,6
29 (10%)	3,4	2,6	61,0	4,1	4,4	21,3

On the basis of the Table 1 we conclude that the regression strategy $(\bar{y}_{HT,s}, P_{SS}(s))$ is worse than the strategies $(\hat{y}_{HT,s}, P_0(s))$, and $(\hat{y}_S, P_{SS}(s))$. In the case of the population without outliers the Singh-Srivastava regression strategy is better than ordinary regression strategy. But in the case of the population with outliers there is an opposite situation because $(\hat{y}_S, P_0(s))$, is better than $(\hat{y}_s, P_{SS}(s))$.

The Horvitz-Thompson strategy for Singh and Srivastava's strategy $(\bar{y}_{HT,S}, P_{SS}(s))$ is better than the unconditional strategy $(\bar{y}_{HT,S}, P_{1,n}(s))$ when the outliers do not exists. But in the case when the population is with the outliers the strategy $(\bar{y}_{HT,S}, P_{1,n}(s))$ is better than $(\bar{y}_{HT,S}, P_{SS}(s))$ only for small sample sizes $n \leq 4$.

The simulation analysis of the accuracy of the conditional strategies $(\bar{y}_{r,u,s}, P_{r,u}(s \mid c))$, $(\tilde{y}_s, P_{r,u}(s \mid c))$ and $(\bar{y}_{HT,s}, P_{r,u}(s \mid c))$ leads to the conclusion that they are the best for the sampling design proportional to the difference of the last and the first order statistics. That is why the Tables 2 and 3 deals only with strategies dependent on the conditional sampling design $P_{1,n}(s \mid c)$.

Table 2

The relative efficiency coefficients (%) of the conditional strategies for $P_{1,n}(s \mid c)$, $c = kd$, $k = 0, 1, 2, 3$ (The population with outliers, N = 284)

c = kd	0			d			2d			3d		
n	e4	e5	e6	e4	e5	e6	e4	e5	e6	e4	e5	e6
2	3,3	1,1	9,7	3,5	1,7	11,4	8,2	2,2	23,2	11,6	2,1	30,7
3	1,4	1,4	5,1	1,6	1,4	6,1	4,9	1,8	13,5	5,5	1,9	16,6
4	1,6	1,5	7,2	1,8	1,6	6,1	3,1	1,8	10,4	5,2	2,0	15,7
6	1,9	1,9	9,8	1,7	1,6	6,7	3,1	1,8	9,7	4,5	2,1	14,6
9	2,3	2,1	14,0	2,1	1,9	10,0	3,1	2,1	10,2	4,2	2,2	13,0
11	2,6	2,4	16,1	2,6	2,4	11,9	3,1	2,1	10,1	4,1	2,0	12,7
14	3,4	2,9	20,1	2,9	2,7	14,8	3,6	2,6	10,5	4,2	2,3	12,4
29	4,5	3,9	27,1	4,7	3,8	26,8	5,0	3,8	17,0	4,8	3,0	15,7

In the both case of the population with outliers and without them the Horvitz-Thompson type strategy $(\bar{y}_{HT,s}, P_{1,n}(s \mid c))$ is less accurate than $(\bar{y}_{1,n,s}, P_{1,n}(s \mid c))$ and $(\bar{y}_s, P_{1,n}(s \mid c))$.

The regression strategy $(\tilde{y}_s, P_{1,n}(s \mid c))$ is better than $(\bar{y}_{1,n,s}, P_{1,n}(s))$ for all considered sample sizes and values of the parameter c (see Tables 2 and 3 or Figures 2 and 3).

It is possible to observe that for $n \geq 6$ and some conditional strategies are more accurate then the unconditional appropriate ones. For instance, the strategy $(\bar{y}_{HT,s}, P_{1,14}(s \mid 3d))$ is better than the unconditional strategy $(\bar{y}_{HT,s}, P_{1,14}(s)) = (\bar{y}_{HT,s}, P_{1,14}(s \mid 0))$. This situation explain Figures 4 and Figure 5, where the distribution spread of the unconditional strategy is gray and the conditional one is black. The distribution of the unconditional strategy $(\bar{y}_{HT,s}, P_{1,14}(s))$ has

some large outliers and that is why the variance of this strategy is larger than the variance of the strategy $(\bar{y}_{HT,s}, P_{1,14}(s \mid 3d))$. Moreover, the conditional strategy neglect some small values of the unconditional one.

Table 3

The relative efficiency coefficients (%) of the conditional strategies for $P_{1,n}(s \mid kd)$, $k = 0, 1, 2, 3$ (The population without outliers, N = 281)

c = kd	0			d			2d			3d		
n	e4	e5	e6	e4	e5	e6	e4	e5	e6	e4	e5	e6
2	3,3	1,1	9,7	3,5	1,7	11,4	8,2	2,2	23,2	11,6	2,1	30,7
3	1,4	1,4	5,1	1,6	1,4	6,1	4,9	1,8	13,5	5,5	1,9	16,6
4	1,6	1,5	7,2	1,8	1,6	6,1	3,1	1,8	10,4	5,2	2,0	15,7
6	1,9	1,9	9,8	1,7	1,6	6,7	3,1	1,8	9,7	4,5	2,1	14,6
9	2,3	2,1	14,0	2,1	1,9	10,0	3,1	2,1	10,2	4,2	2,2	13,0
11	2,6	2,4	16,1	2,6	2,4	11,9	3,1	2,1	10,1	4,1	2,0	12,7
14	3,4	2,9	20,6	2,9	2,7	14,8	3,6	2,6	10,5	4,2	2,3	12,4
29	4,5	3,9	27,1	4,7	3,8	26,8	5,0	3,8	17,0	4,8	3,0	15,7

Conclusions

The inclusion probabilities of the conditional sampling design proportionate to the difference of two order statistics are presented. They let determine the variance of the Horvitz-Thompson statistic as well as its estimator. The construction of the sampling design can be easy modified in such a way that definition of the sampling design proportionate to e.g. the sum of two order statistics is straightforward.

The simulation analysis let us infer that in general the considered regression type strategies are the best among the considered ones. Moreover, the relative efficiencies of the regression ones are similar and the best among them is the strategy $(\bar{y}_s, P_{1,n}(s \mid 3d))$. In some cases the conditional strategies could be slightly better than the appropriate unconditional ones but this conclusion will be developed in separate and more deep studies.

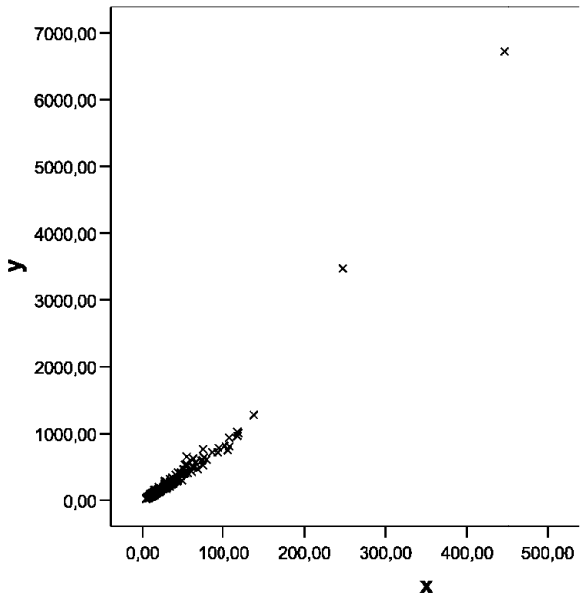


Figure 1. Spread of variables x and y

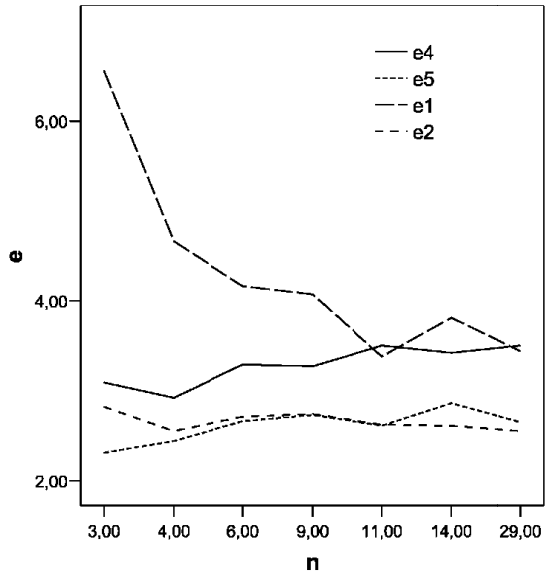


Figure 2. Efficiencies of the strategies in the case of the population without outliers

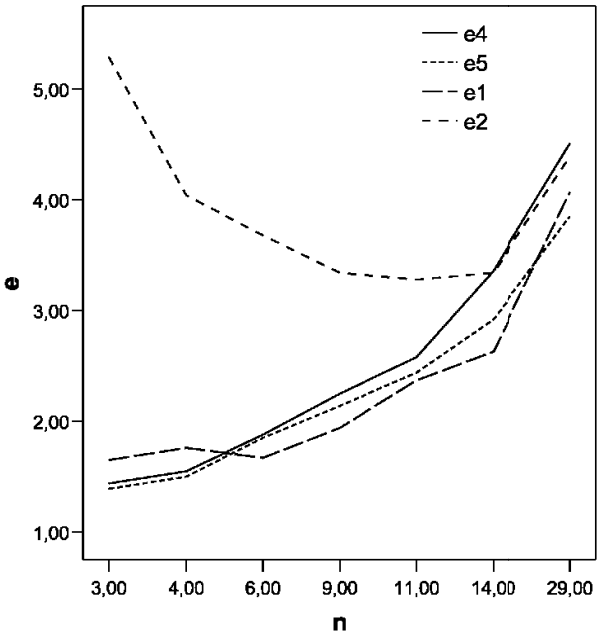


Figure 3. Efficiencies of the strategies in the case of the population with outliers

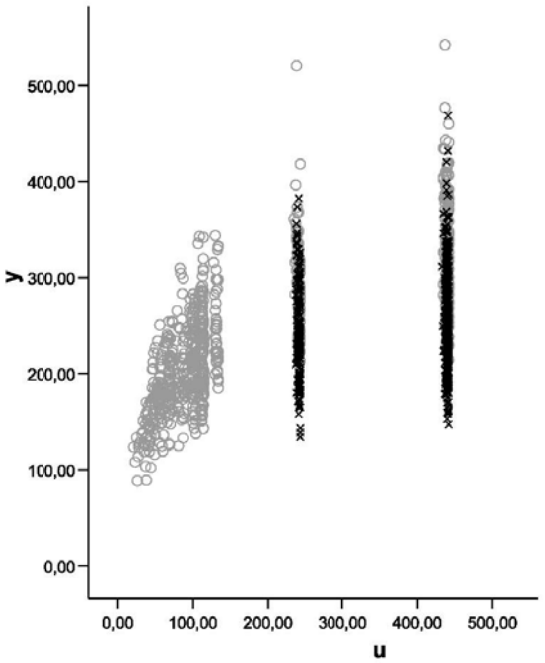


Figure 4. Distribution spreads of the strategy $u = (\bar{y}_{HT,s}, P_{r,u}(s \mid d))$ on $x = x_{(n)} - x_{(n)}$ for $d = 0$ or $d = 3$ and $N = 284$, $n = 14$

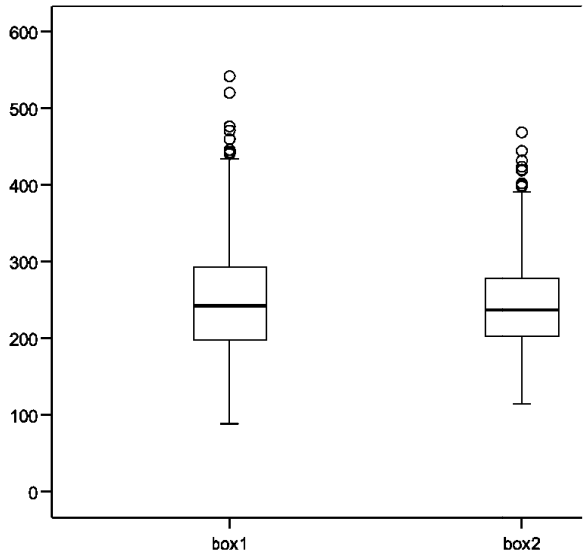


Figure 5. Distribution boxplots of the strategy $u = (\bar{y}_{HT,s}, P_{r,u}(s \mid d))$ for $d = 0$ or $d = 3$ and $N = 284$, $n = 14$

Acknowledgement

The research was supported by the grant number N N111 434137 from the Ministry of Science and Higher Education.

References

Horvitz D.G., Thompson D.J. (1952): *A Generalization of the Sampling Without Replacement from Finite Universe*. „Journal of the American Statistical Association", 47, s. 663-685.

Särndal C.E., Swensson B., Wretman J. (1992): *Model Assisted Survey Sampling*. Springer Verlag, New York – Berlin – Heidelberg.

Singh P., Srivastava A.K. (1980): *Sampling Schemes Providing Unbiased Regression Estimators*. „Biometrika" 67 (1), s. 205-9.

Tillé Y. (1999): E*stimation in Surveys Using Conditional Inclusion Probabilities: Complex Design*. „Survey Methodology", Vol. 25, No, 1, s. 57-66.

Tillé Y. (2006): *Sampling algorithms*. Springer, New York.

Wywiał J.L. (2003): *On Conditional Sampling Strategies*. „Statistical Papers", Vol. 44, 3, s. 397-419.

Wywiał J.L. (2007): *Simulation Analysis of Accuracy Estimation of Population Mean on the Basis of Strategy Dependent on Sampling Design Proportionate to the Order Statistic of an Auxiliary Variable*. „Statistics in Transition – New Series", Vol. 8, No. 1, s. 125-137.

Wywiał J.L. (2008): *Sampling Design Proportional to Order Statistic of Auxiliary Variable*. „Statistical Papers", Vol. 49, No. 2, s. 277-289.

Wywiał J.L. (2009): *Performing Quantiles in Regression Sampling Strategy*. „Model Assisted Statistics and Applications", Vol. 4, No. 2, s. 131-142.

SAMPLING DESIGNS PROPORTIONATE TO NON-NEGATIVE FUNCTIONS OF TWO QUANTILES OF AUXILIARY VARIABLE

Summary

Estimation of the population average in a finite population by means of sampling strategies dependent on sample quantiles of an auxiliary variable are considered. The sampling design proportional to an order statistic of an auxiliary variable was defined by Wywiał (2007, 2008). It was generalized into case of the sampling design proportional to the difference of two order statistics by Wywiał (2009), too. In this paper those results are generalized on the case of a conditional sampling design. Several strategies including the Horvitz-Thompson statistic and regression estimators are considered. Their accuracy is analyzed on the basis of computer simulation experiments.