METODY WNIOSKOWANIA STATYSTYCZNEGO W BADANIACH EKONOMICZNYCH

Studia Ekonomiczne

ZESZYTY NAUKOWE WYDZIAŁOWE

UNIWERSYTETU EKONOMICZNEGO W KATOWICACH

METODY WNIOSKOWANIA STATYSTYCZNEGO W BADANIACH EKONOMICZNYCH

Redaktorzy naukowi
Józef Kolonko
Grzegorz Kończak

Katowice 2013

ISSN 2083-8611

Wersją pierwotną Studiów Ekonomicznych jest wersja papierowa

WYDAWNICTWO UNIWERSYTETU EKONOMICZNEGO W KATOWICACH
ul. 1 Maja 50, 40-287 Katowice, tel.: +48 32 257-76-35, faks: +48 32 257-76-43
www.wydawnictwo.ue.katowice.pl e-mail: wydawnictwo@ue.katowice.pl

SPIS TREŚCI

Stanisław Heilpern
Uniwersytet Ekonomiczny we Wrocławiu

PROCES RYZYKA Z ZALEŻNYMI OKRESAMI MIĘDZY WYPŁATAMI – ANALIZA PRAWDOPODOBIEŃSTWA RUINY[1]

Wprowadzenie

W pracy będzie rozpatrywany ciągły proces ryzyka, w którym okresy między poszczególnymi wypłatami mogą być zależnymi zmiennymi losowymi. W klasycznych procesach ryzyka, stanowiących podstawę teorii ruiny [Kaas et al. 2001; Ostasiewicz 2000; Rolski et al. 1999], przyjmuje się niezależność występujących procesów i zmiennych losowych. Założenie o niezależności jest wygodne z teoretycznego, matematycznego punktu widzenia, upraszcza rozważania, wiele faktów można udowodnić, jednak często jest zbyt idealistycznym podejściem. W praktyce okresy między wypłatami są zwykle w większym lub mniejszym stopniu zależne. Na badany proces wpływają często czynniki zewnętrzne, np. ekstremalne zjawiska, takie jak powodzie, pożary, trzęsienia ziemi, czy karambole na autostradach, kryzysy gospodarcze lub polityczne, wpływające jednocześnie na wszystkich uczestników procesu, powołując występowanie zależności.

Proces ryzyka będzie badany ze względu na prawdopodobieństwo ruiny, ze szczególnym uwzględnieniem wpływu stopnia zależności okresów między wypłatami na to prawdopodobieństwo. Rozpatrzono przykład ścisłej zależności okresów oraz gdy struktura ich zależności jest opisana archimedesową funkcją łączącą (ang. *copula*). W obydwu założono, że zarówno okresy między wypłatami, jak i wypłaty mają rozkład wykładniczy.

Pracę można traktować jako kontynuację artykułu [Heilpern 2010], w którym był rozpatrywany proces ryzyka z zależnymi wypłatami. Otrzymane tam wyniki wskazywały na istotną zależność wpływu stopnia zależności wypłat na

[1] Praca naukowa finansowana ze środków na naukę w latach 2010-2012 jako projekt badawczy nr 3361/B/H03/2010/38.

prawdopodobieństwo ruiny, osiągania największych i najmniejszych prawdopodobieństw ruiny, od wartości kapitału początkowego. Podobne wyniki zostały osiągnięte w niniejszej pracy.

Obliczenia związane z wyznaczeniem prawdopodobieństwa ruiny zostały wykonane za pomocą programu Mathematica 6 oraz arkusza kalkulacyjnego Excel.

1. Proces ryzyka

Podstawą rozważań będzie następujący ciągły proces ryzyka [Kaas et al. 2001; Ostasiewicz 2000]:

$$U(t) = u + ct - \sum_{i=1}^{N(t)} X_i,$$

gdzie $u \geq 0$ jest kapitałem początkowym, $c > 0$ intensywnością napływu składki, $N(t) = \min\{n \geq 0: T_{n+1} > t\}$ procesem liczącym wypłaty $X_i > 0$, a T_i momentem pojawienia się i-tej wypłaty.

Przyjęto, że wypłaty są niezależne oraz mają ten sam rozkład z dystrybuantą $F_X(x)$ i wartością oczekiwaną $m = E(X_i)$, oraz że proces $N(t)$ generuje okresy między wypłatami $W_i = T_i - T_{i-1}$ o tym samym rozkładzie F_W. W pracy tej mogą być one zależnymi zmiennymi losowymi. Ponadto założono, że zmienne X_i, W_i są nawzajem niezależne. W przypadku gdy okresy W_i są niezależne, otrzymuje się tzw. model Sparre Andersena [Rolski et al. 1999].

Głównym tematem niniejszych rozważań będzie prawdopodobieństwo ruiny [Kaas et al. 2001; Ostasiewicz 2000]:

$$\psi(u) = P(T < \infty | \, U(0) = u),$$

gdzie T jest momentem zajścia ruiny

$$T = \min\{t: U(t) < 0\},$$

czyli zdarzenia, że proces ryzyka będzie ujemny w nieskończonym horyzoncie czasu. Prawdopodobieństwo ruiny można również wyznaczyć na podstawie znajomości nadwyżki wypłat $Y_i = X_i - cW_i$. Wtedy zachodzi zależność [Rolski et al. 1999]:

$$\psi(u) = P\left(\max_n \sum_{k=1}^{n} Y_i > u\right).$$

W przypadku gdy zmienne W_i są niezależne (model Sparre Andersena), a wypłaty X_i mają rozkład wykładniczy z parametrem $1/m$, to prawdopodobieństwo ruiny wyraża się wzorem [Rolski et al. 1999]:

$$\psi(u) = (1 - Rm)e^{-Ru},$$

gdzie współczynnik dopasowania R jest nieujemnym rozwiązaniem równania

$$\hat{m}_Y(s) = \hat{m}_X(s)\hat{m}_W(-cs) = 1,$$

a $\hat{m}_Y(s) = E(e^{sY})$ jest funkcją generującą momenty zmiennej losowej Y. Powyższy wzór na prawdopodobieństwo ruiny będzie wykorzystywany w dalszej części pracy.

W klasycznym modelu ryzyka przyjmuje się, że proces liczący wypłaty $N(t)$ jest procesem Poissona [Kaas et al. 2001; Ostasiewicz 2000; Rolski et al. 1999]. Wtedy okresy między wypłatami W_i są niezależne i mają rozkład wykładniczy z parametrem λ, gdzie λ jest intensywnością procesu Poissona. W przypadku małej intensywności napływu składki c, tzn. gdy zachodzi warunek $c \leq \lambda m$, zajście ruiny jest zdarzeniem pewnym dla każdej wartości kapitału początkowego u, czyli otrzymujemy $\psi(u) = 1$.

Dla skrajnych wartości kapitału początkowego u, prawdopodobieństwa ruiny przyjmują prostą postać:

$$\psi(0) = \frac{\lambda m}{c}, \lim_{u\to\infty} \psi(u) = \psi(\infty) = 0.$$

Dla dowolnych wartości kapitału początkowego u na ogół nie ma natomiast jawnych wzorów na prawdopodobieństwo ruiny. Jedynie w przypadku gdy wypłaty mają tzw. rozkład fazowy [Rolski et al. 1999] można podać konkretny wzór na to prawdopodobieństwo. Przykładowo, gdy wypłaty X_i mają rozkład wykładniczy z parametrem $1/m$ (szczególny przypadek rozkładu fazowego), prawdopodobieństwo ruiny wyznaczamy stosując wzór:

$$\psi(u) = \frac{\lambda m}{c} exp\left(-\frac{c - \lambda m}{cm}u\right).$$

Współczynnik dopasowania wynosi wtedy $R = \frac{c-\lambda m}{cm}$. Również w przypadku dyskretnych rozkładów wypłat istnieje kombinatoryczny wzór na prawdopodobieństwo ruiny [Kaas et al. 2001].

2. Silna zależność

Na początku rozpatrzmy skrajny przypadek, gdy okresy między wypłatami W_i są silnie zależne. Opisane są one wtedy przez tą samą zmienną losową W. Dla ustalonej wartości w tej zmiennej otrzymujemy proces ryzyka $U_w(t)$ o stałych, deterministycznych okresach między wypłatami o długości w:

$$U_w(t) = u + ct - \sum_{i=1}^{[t/w]} X_i,$$

gdzie $[x]$ jest częścią całkowitą x.

Prawdopodobieństwo ruiny w przypadku silnej zależności okresów między wypłatami wyznaczamy jako mieszankę:

$$\psi(u) = \int_0^\infty \psi_w(u) dF_W(w)$$

prawdopodobieństw ruiny $\psi_w(u)$ dla deterministycznych okresów. Pamiętając, że nierówność $w \leq \frac{m}{c}$ pociąga za sobą $\psi_w(u) = 1$, otrzymujemy:

$$\psi(u) = \int_{m/c}^\infty \psi_w(u) dF_W(w) + F_W\left(\frac{m}{c}\right).$$

Widzimy, że nawet dla nieskończenie dużego kapitału początkowego u, prawdopodobieństwo ruiny może być w tym przypadku dodatnie, równe $\psi(\infty) = F_W\left(\frac{m}{c}\right)$, w przeciwieństwie do przypadku niezależnych okresów, gdzie otrzymujemy zerowe prawdopodobieństwo ruiny.

Zajmijmy się teraz przypadkiem, gdy okresy między wypłatami są opisane tą samą zmienną losową W o rozkładzie wykładniczym z parametrem λ. Będzie to przeciwstawna sytuacja do klasycznego procesu ryzyka, gdy okresy te są niezależne. Jeśli dodatkowo przyjmiemy, że wypłaty są również wykładnicze z parametrem $1/m$, to prawdopodobieństwo ruiny dla ustalonej wartości $W = w$ jest określone wzorem:

$$\psi_w(u) = (1 - R_w m) e^{-R_w u},$$

gdzie współczynnik dopasowania $R_w > 0$ jest rozwiązaniem równania $e^{-scw} = 1 - ms$. Wtedy wzór na prawdopodobieństwo ruiny, gdy okresy między wypłatami są ściśle zależne o rozkładzie wykładniczym i wykładniczych wypłatach, przyjmuje postać:

$$\psi(u) = \lambda \int_{m/c}^{\infty} (1 - R_w m) e^{-(R_w u + \lambda w)} dw + 1 - e^{-\lambda m/c}.$$

Dla nieskończenie dużego kapitału początkowego prawdopodobieństwo ruiny jest dodatnie, równe:

$$\psi(\infty) = 1 - e^{-\lambda m/c}.$$

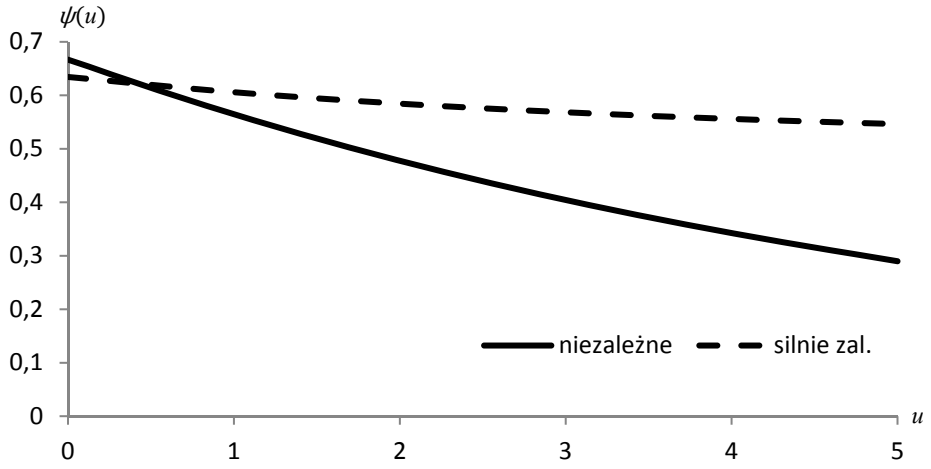
Przykład 1. Niech wartość oczekiwana wypłat $m = 2$, intensywność napływu składki $c = 3$, a okresy między wypłatami mają rozkład wykładniczy z parametrem $\lambda = 1$. W tabeli 1 zostały podane wartości prawdopodobieństwa ruiny dla ściśle zależnych i niezależnych okresów między wypłatami oraz dla różnych wartości kapitału początkowego u. Prawdopodobieństwa te zostały również przedstawione na rysunku 1.

Tabela 1

Prawdopodobieństwa ruiny dla ściśle zależnych i niezależnych okresów między wypłatami

u	niezależne	silnie zal.	u	niezależne	silnie zal.	u	niezależne	silnie zal.
0	0,666667	0,634336	9	0,148753	0,523385	18	0,033191	0,505515
1	0,564321	0,605764	10	0,125917	0,519991	19	0,028096	0,504529
2	0,477688	0,584422	11	0,106586	0,517135	20	0,023783	0,50364
3	0,404354	0,568260	12	0,090224	0,514707	25	0,010336	0,500248
4	0,342278	0,555848	13	0,076373	0,512621	30	0,004492	0,497978
5	0,289732	0,546180	14	0,064648	0,510814	35	0,001952	0,496354
6	0,245253	0,538543	15	0,054723	0,509236	40	0,000848	0,495134
7	0,207602	0,532428	16	0,046322	0,507846	50	0,00016	0,493426
8	0,175731	0,527465	17	0,039211	0,506614	100	3,85E-08	0,490005

Ponadto prawdopodobieństwo ruiny dla nieskończenie dużego kapitału początkowego i ściśle zależnych okresów między wypłatami wynosi $\psi(\infty) = 0{,}486583$.



Rys. 1. Prawdopodobieństwa ruiny dla ściśle zależnych i niezależnych okresów między wypłatami

Widzimy, że dla zerowego i dla małego kapitału początkowego prawdopodobieństwo ruiny dla niezależnego przypadku jest większe niż dla ściśle zależnego. Dla większych wartości kapitału u otrzymujemy natomiast relację odwrotną. Przypadek ściśle zależnych okresów między wypłatami jest „gorszy", daje nam większe prawdopodobieństwo ruiny. Ponadto różnice między prawdopodobieństwami ruiny dla różnych wartości kapitału początkowego są w tym przypadku niewielkie.

3. Archimedesowe funkcje łączące

Rozpatrywany powyżej przypadek, gdy okresy między wypłatami są ściśle zależne, jest wybitnie skrajną i sztuczną sytuacją. W praktyce zależność między okresami nie jest tak duża. Stopień zależności, mierzony np. współczynnikami korelacji τ Kendalla, zwykle jest istotnie mniejszy od jedynki. W niniejszej pracy do modelowania pośrednich, bardziej realistycznych zależności, wykorzystano archimedesowe funkcje łączące.

Funkcja łącząca C (ang. *copula*) jest łącznikiem między rozkładem łącznym a rozkładami brzegowymi [Nelsen 1999; Heilpern 2007]:

$$P(W_1 > w_1, \dots, W_n > w_n) = C\left(\bar{F}_{W_1}(w_1), \dots, \bar{F}_{W_n}(w_n)\right),$$

gdzie $\bar{F}_W(w) = 1 - F_W(w)$ jest funkcją przetrwania zmiennej losowej W. Funkcję łączącą można zdefiniować za pomocą dystrybuant, a nie funkcji przetrwania jak w tym przypadku, jednak dla nas postać ta jest wygodniejsza. Należy też pamiętać, że funkcja łącząca nie zależy od rozkładów brzegowych i gdy rozkłady brzegowe są ciągłe, jest ona jednoznacznie wyznaczona.

Archimedesowe funkcje łączące są indukowane jednowymiarowym generatorem g i przyjmują prostą, quasi-addytywną postać [Nelsen 1999; Heilpern 2007]:

$$C(u_1, \ldots, u_n) = g^{-1}(g(u_1) + \ldots + g(u_n)).$$

Generator g: (0, 1] $\rightarrow$ R_+ jest malejącą funkcją ciągłą taką, że $\lim_{n\to\infty} g(u) = \infty, g(1) = 0$. Funkcją g^{-1} odwrotna do generatora powinna być całkowicie monotoniczną funkcją, tzn. spełniać warunek:

$$(-1)^k(g^{-1})^{(k)}(t) \geq 0,$$

gdzie $f^{(k)}$ jest pochodną k-tego rzędu funkcji f, dla każdego $k = 0, 1, 2, \ldots$ oraz $t > 0$. Jest więc transformatą Laplace'a pewnej nieujemnej zmiennej losowej Θ o dystrybuancie F_Θ [Nelsen 1999].

Można pokazać [Frees i Valdez 1998; Heilpern 2007], że dla ustalonej wartości θ indukowane przez archimedesową funkcję łączącą C zmiennej Θ zmienne losowe W_i są warunkowo niezależne, tzn. zachodzi zależność:

$$P(W_1 > w_1, \ldots, W_n > w_n | \Theta = \theta) = P(W_1 > w_1 | \Theta = \theta) \cdot \ldots \cdot P(W_n > w_n | \Theta = \theta).$$

Jest to pożyteczna własność. Umożliwia ona stosowanie dla ustalonej wartości indukowanej zmiennej losowej Θ znanych metod klasycznej teorii ruiny opartej na niezależności. Zmienna ta generuje wtedy warunkowe zmienne losowe $W_{i|\Theta}$ o funkcji przetrwania [Frees i Valdez 1998; Heilpern 2007]:

$$\bar{F}(w|\theta) = exp(-\theta g(\bar{F}_W(w)))$$

i wartości oczekiwanej:

$$E(W_{i|\theta}) = \int_0^\infty \bar{F}(w|\theta)dw,$$

która jest malejącą funkcją θ.

Dla ustalonej wartości θ indukowanej zmiennej losowej Θ otrzymujemy warunkowy proces ryzyka U_θ z niezależnymi wypłatami X_i i niezależnymi okresami między wypłatami $W_{i|\Theta}$, czyli proces Sparee Andersena. Warunkowe prawdopodobieństwo ruiny tak określonego procesu ryzyka będziemy oznaczać symbolem $\psi_\theta(u)$. Wtedy bezwarunkowe prawdopodobieństwo ruiny możemy wyznaczyć korzystając z mieszanki warunkowych prawdopodobieństw ruiny ze zmienną mieszającą Θ:

$$\psi(u) = \int_0^\infty \psi(u|\theta) dF_\Theta(\theta).$$

Niech θ_0 spełnia zależność $cE\left(W_{i|\theta}\right) = m$, wtedy dla $\theta \geq \theta_0$ warunkowa ruina jest zdarzeniem pewnym, tzn. $\psi_\theta(u) = 1$, a bezwarunkowe prawdopodobieństwo ruiny jest określone wzorem:

$$\psi(u) = \int_0^{\theta_0} \psi(u|\theta) dF_\Theta(\theta) + \bar{F}_\Theta(\theta_0).$$

Widzimy, że gdy $\bar{F}_\Theta(\theta_0) > 0$, to nawet dla nieskończenie dużego kapitału początkowego prawdopodobieństwo ruiny jest dodatnie, podobnie jak w przypadku ścisłej zależności okresów.

Rozpatrzmy teraz przypadek, gdy zarówno wypłaty W_i, jak i okresy między wypłatami W_i mają rozkład wykładniczy. Ponadto założymy, że struktura zależności okresów W_i jest opisana funkcją łączącą Claytona, określoną wzorem:

$$C_\alpha(u_1, \ldots, u_n) = (u_1^{-\alpha} + \cdots + u_n^{-\alpha})^{-1/\alpha},$$

gdzie $\alpha > 0$. Jej generatorem jest funkcja $g(u) = u^{-\alpha} - 1$, a parametr α oddaje stopień zależności. Współczynnik korelacji τ Kendalla jest w tym przypadku określony prostym wzorem [Nelsen 1999]:

$$\tau = \frac{\alpha}{\alpha + 2}.$$

W granicy, gdy parametr α dąży do zera otrzymujemy niezależność, a gdy dąży do nieskończoności ścisłą zależność. Wraz ze wzrostem wartości tego parametru rośnie natomiast stopień zależności.

Funkcja łącząca Claytona indukuje zmienną losową Θ o rozkładzie gamma Ga(1/α, 1). Warunkowe rozkłady okresów między wypłatami są wtedy określone funkcją przetrwania postaci:

$$\bar{F}(w|\theta) = exp\left(-\theta\left(e^{\alpha w\lambda} - 1\right)\right).$$

Warunkowe prawdopodobieństwa ruiny są natomiast określone wzorem:

$$\psi_\theta(u) = (1 - R_\theta m)e^{-R_\theta u},$$

gdzie współczynnik dopasowania $R_\theta > 0$ jest rozwiązaniem równania:

$$1 - ms = \int_0^\infty e^{-csw} dF(w|\theta).$$

Przykład 2 (cd. przykładu 1). Niech struktura zależności jest opisana funkcją łączącą Claytona. W tabeli 2 są podane prawdopodobieństwa ruiny dla różnych wartości kapitału początkowego u i pięciu wartości parametru α: 0; 0,2; 2; 10 oraz ∞. Odpowiadają one wartością współczynnika korelacji τ Kendalla równym: 0 (niezależność); 0,091; 0,5; 0,833 oraz 1 (ścisła zależność).

Tabela 2

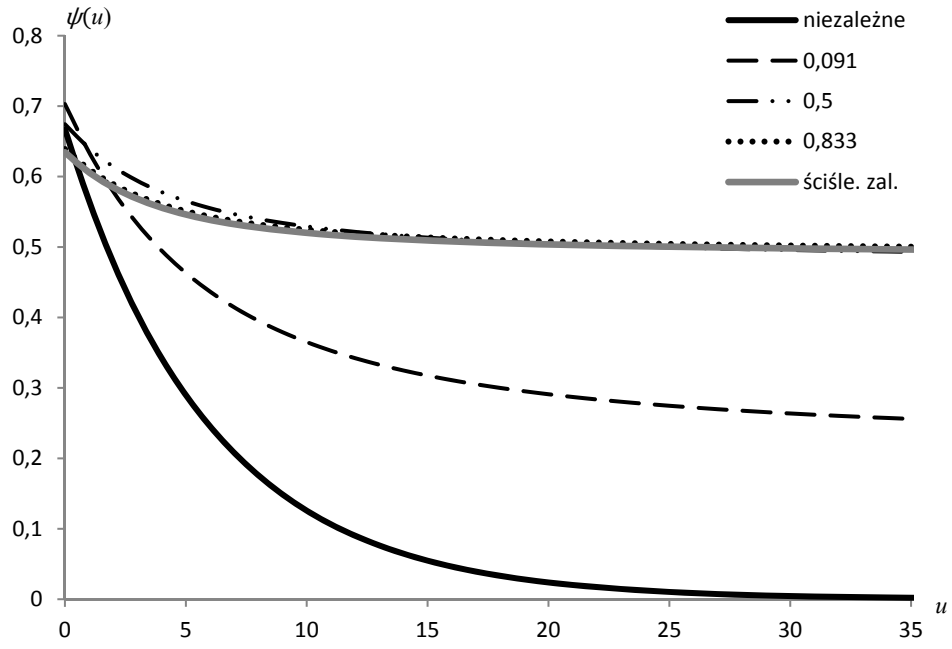
Prawdopodobieństwa ruiny dla wybranych wartości parametru α i kapitału początkowego u

u	niezależne	0,2	2	10	ściśle zal.
1	2	3	4	5	6
0	0,666667	**0,702899**	0,674263	0,639041	***0,634335***
0,4241	***0,621171***	**0,672066**	0,659001	0,626159	***0,621171***
1	***0,564321***	0,634356	**0,640168**	0,610523	0,605764
2	***0,477688***	0,578522	**0,614075**	0,589191	0,584423
3	***0,404354***	0,532698	**0,593839**	0,573012	0,568261
4	***0,342278***	0,494808	**0,577937**	0,560568	0,555849
5	***0,289732***	0,463248	**0,565275**	0,550861	0,546182
6	***0,245253***	0,436772	**0,555064**	0,543181	0,538545
7	***0,207602***	0,414403	**0,546727**	0,537022	0,532430
8	***0,175731***	0,395375	**0,539838**	0,532016	0,527467
9	***0,148753***	0,379083	**0,534083**	0,527896	0,523388

cd. tabeli 2

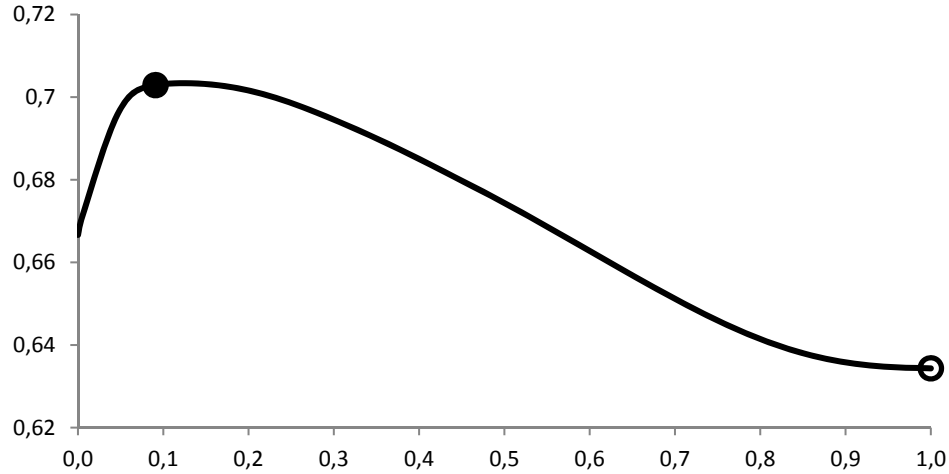
1	2	3	4	5	6
10	***0,125917***	0,365043	**0,529223**	0,524464	0,519994
20	***0,023783***	0,290999	0,504702	**0,507850**	0,503645
50	***0,000160***	0,242140	0,488346	**0,497383**	0,493437
100	***3,85E-08***	0,226576	0,482704	**0,493889**	0,490023
∞	***0***	0,211821	0,476998	**0,490350**	0,486583

Skrajne wartości, największe i najmniejsze, zostały wyróżnione w tabeli. Można zaobserwować brak regularności, monotoniczności. Położenie skrajnych wartości prawdopodobieństwa ruiny zależy istotnie od wartości kapitału początkowego u. Największe prawdopodobieństwo ruiny nigdy nie jest osiągalne dla skrajnych przypadków zależności, niezależności oraz ścisłej zależności okresów między wypłatami. Najmniejsze prawdopodobieństwa ruiny zachodzą natomiast wyłącznie dla skrajnych przypadków. Dla małych wartości kapitału początkowego, mniejszych od 0,4241, najmniejsze prawdopodobieństwo ruiny otrzymujemy dla ściśle zależnych okresów między wypłatami, a dla wartości $u > 0{,}4241$ dla niezależnych okresów. Prawdopodobieństwa ruiny są również przedstawione na rysunku 2.

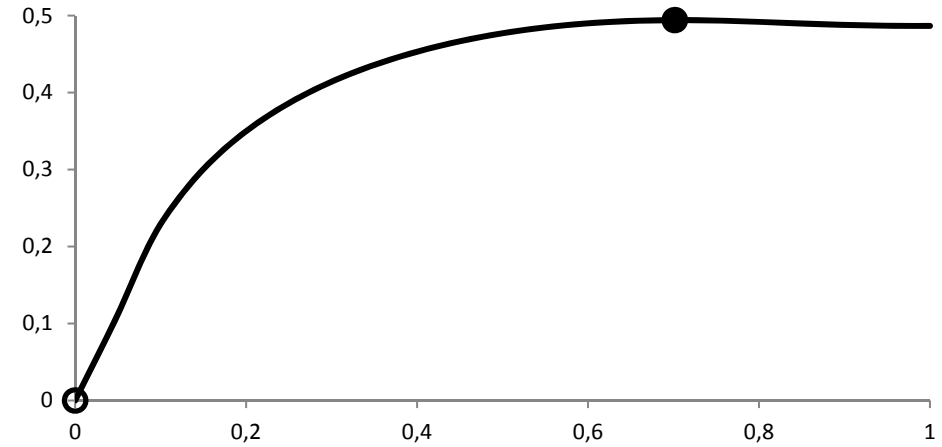


Rys. 2. Prawdopodobieństwo ruiny dla różnych wartości u i τ

Na rysunkach 3 i 4 są odpowiednio przedstawione wykresy prawdopodobieństwa ruiny dla zerowego oraz nieskończenie dużego kapitału początkowego i różnych wartości stopnia zależności okresów między wypłatami, mierzonych współczynnikiem τ Kendalla. Widzimy, że dla zerowej wartości kapitału początkowego prawdopodobieństwo ruiny najpierw rośnie wraz ze wzrostem stopnia zależności, osiąga maksimum dla współczynnika korelacji Kendalla przyjmującego wartość około 0,091, a następnie powoli maleje, przyjmując minimum w przypadku ścisłej zależności między wypłatami.



Rys. 3. Prawdopodobieństwo ruiny dla $u = 0$



Rys. 4. Prawdopodobieństwo ruiny dla $u = \infty$

W przypadku nieskończenie dużego kapitału początkowego sytuacja jest trochę inna. Najmniejsza wartość prawdopodobieństwa ruiny, równa zero, jest osiągana dla niezależnych okresów między wypłatami. Następnie prawdopodobieństwo to rośnie wraz ze wzrostem stopnia zależności i osiąga maksimum dla $\tau = 0{,}706$. Dla większych wartości współczynnika korelacji Kendalla prawdopodobieństwo ruiny nieznacznie spada. Podobna sytuacja zachodzi dla pośrednich większych niż 0,4241, wartości kapitału początkowego. Jedynie maksimum prawdopodobieństwa ruiny jest osiągane dla mniejszych stopni zależności. Przykładowo, dla $u = 5$ największe prawdopodobieństwo ruiny otrzymujemy dla współczynnika Kendalla przyjmującego wartość około 0,5.

Podsumowanie

W pracy przeprowadzono analizę wpływu stopnia zależności okresów między wypłatami na prawdopodobieństwo ruiny. Przyjęto bardziej realistyczne założenie, że badane okresy mogą być zależne w odróżnieniu od klasycznych założeń przyjmujących ich niezależność. Pokazano, że wartości stopnia zależności okresów, dla których są osiągane skrajne wartości prawdopodobieństwa ruiny, istotnie zależą od wielkości kapitału początkowego. Prawidłowość tę wyraźnie widać zwłaszcza w przypadku największych wartości prawdopodobieństwa ruiny. Wartości te są osiągane dla pośrednich wartości stopnia zależności, a nie dla wartości skrajnych, dotyczących niezależności, czy silnej zależności.

Praca jest kontynuacją artykułu [Heilpern 2010], w którym były rozpatrywane zależne wypłaty oraz była badana zależność prawdopodobieństwa ruiny od wielkości stopnia zależności wypłat. Następne prace autora związane z tą tematyką będą poświęcone procesowi ryzyka, w których mogą być zależne zarówno wypłaty, jak i okresy między nimi oraz badaniu zależności prawdopodobieństwa ruiny od stopnia zależności oraz od intensywności napływu składki.

Literatura

Frees E.W., Valdez E.A. (1998): *Understanding Relationships using Copulas*. „North Amer. Actuarial Journal", No. 2.

Heilpern S. (2007): *Funkcje łączące*. Wydawnictwo AE, Wrocław.

Heilpern S. (2010): *Wyznaczanie prawdopodobieństwa ruiny, gdy struktura zależności wypłat opisana jest Archimedesowi funkcją łączącą*. W: *Zagadnienia Aktuarialne. Teoria i Praktyka*. Red. W. Otto. Wydawnictwo Uniwersytetu Warszawskiego, Warszawa.

Kaas R., Goovaerts M., Dhaene J., Denuit M. (2001): *Modern Actuarial Risk Theory*. Kluwer, Boston.
Nelsen R.B. (1999): *An Introduction to Copulas*. Springer, New York.
Ostasiewicz W., red. (2000): *Modele aktuarialne*. Wydawnictwo AE, Wrocław.
Rolski T., Schmidli H., Schmidt V., Teugels J.L. (1999): *Stochastic Processes for Insurance and Finance.* Willey, New York.

RISK PROCESS WITH DEPENDENT INTERCLAIM TIMES – ANALYSIS OF PROBABILITY OF RUIN

Summary

The paper is devoted to the risk process with dependent interclaim times. The influence of degree of dependence of interclaims on the probability of ruin is investigated. The case of the strict dependence and the case when the dependence structure is described by the Archimedean copula is studied. The localization of the extreme values of the probability of ruin essentially depends on the value of initial capital. The most values of the probability of ruin are attain for the middle values of degree of dependence.

Zofia Rusnak
Uniwersytet Ekonomiczny we Wrocławiu

WYBRANE METODY POMIARU CECH JAKOŚCIOWYCH W ANALIZACH UBÓSTWA

Wprowadzenie

Jednym z głównych celów polityki społecznej jest dążenie do ograniczenia zasięgu ubóstwa i wykluczenia społecznego. Ubóstwo oznacza bowiem nie tylko brak środków finansowych na zaspokojenie podstawowych potrzeb, ale ogranicza również możliwość dokonania swobodnego wyboru dotyczącego stylu życia, co w konsekwencji może prowadzić do wykluczenia społecznego[1].

W analizach ubóstwa decydujący wpływ na identyfikację i pomiar sfery ubóstwa mają m.in.: określenie determinant ubóstwa, czyli tych czynników, które zwiększają ryzyko zagrożenia ubóstwem, wybór metody wyznaczania granic ubóstwa oraz wybór określonych wskaźników służących do oceny i porównań przestrzenno-czasowych zasięgu i głębokości tej sfery.

Głównym celem tej pracy jest próba zastosowania wybranych metod pomiaru cech jakościowych do określenia determinant ubóstwa i wskazania, które z zaproponowanych czynników i w jaki sposób wpływają na prawdopodobieństwo tego, że gospodarstwa domowe określonego typu znajdą się w sferze ubóstwa relatywnego.

Drugi cel pracy to ocena i porównanie sfery ubóstwa relatywnego w ujęciu regionalnym z wykorzystaniem m.in. podstawowych i najczęściej stosowanych wskaźników ubóstwa, do których należą stopa ubóstwa relatywnego i średnia luka wydatkowa ubogich.

Analiza sfery ubóstwa wymaga w pierwszym rzędzie ustalenia granicy ubóstwa (linii ubóstwa). Różnorodność stosowanych w praktyce metod wyznaczania

[1] Unia Europejska ogłosiła rok 2010 Europejskim Rokiem Walki z Ubóstwem i Wykluczeniem Społecznym. Podstawowe cele Roku to zwiększenie świadomości opinii publicznej na temat ubóstwa i wykluczenia społecznego oraz zwiększenie zaangażowania politycznego Unii Europejskiej i państw członkowskich w walkę z tymi problemami.

granic ubóstwa jest rezultatem m.in. traktowania ubóstwa jako kategorii absolutnej lub względnej oraz charakteru danych, stanowiących podstawę wyznaczania granic ubóstwa obiektywnego lub subiektywnego.

W tej pracy analiza dotyczy ubóstwa relatywnego, które oznacza względny (relatywny) brak środków finansowych na utrzymanie w gospodarstwie domowym. Jako wskaźnik zamożności gospodarstw przyjęto wydatki gospodarstw domowych, przy czym dla zachowania porównywalności sytuacji gospodarstw o różnej wielkości i różnym składzie demograficznym obliczono wydatki ekwiwalentne, stosując oryginalną skalę ekwiwalentności OECD typu 0,7/0,5[2]. Jako granicę ubóstwa relatywnego przyjęto połowę średnich wydatków ekwiwalentnych wyznaczonych dla zbiorowości wszystkich gospodarstw domowych objętych badaniami budżetów gospodarstw domowych BBGD w 2008 r.[3].

Podstawę wszystkich obliczeń stanowiły, zakupione w tym celu, dane jednostkowe z badania budżetów gospodarstw domowych (BBGD), realizowanego przez GUS w 2008 r.

1. Wybrane metody pomiaru cech jakościowych

Niniejsza analiza i ocena tego, które spośród zaproponowanych cech charakteryzujących gospodarstwa domowe mają istotny wpływ na prawdopodobieństwo uznania gospodarstwa domowego za ubogie, opiera się na wybranych metodach stosowanych do pomiaru cech jakościowych. Należą do nich m.in.: względne ryzyko, iloraz szans, model logitowy. Podstawy tych metod zostaną zaprezentowane poniżej.

Załóżmy, że zmienna Y jest dychotomiczną zmienną o wartościach 1 i 0, przy czym wartość 1 oznacza, że zaszło interesujące nas zdarzenie, a wartość 0 w przeciwnym przypadku. Na przykład Y jest zmienną odpowiedzi o wartościach 1 (tak), 0 (nie), przy czym odpowiedzi są udzielane przez k grup respondentów $R = [r_1, r_2, ..., r_k]$.

Do porównań dwóch grup respondentów ze względu na zmienną Y można wykorzystać:

- względne ryzyko (ang. *relative risk*),
- iloraz szans (ang. *odds ratio*).

---

[2] Zgodnie z tą skalą pierwszej osobie dorosłej w gospodarstwie domowym przypisuje się liczbę równą 1, następnej 0,7 oraz liczbę 0,5 każdemu dziecku w wieku do 14 lat.

[3] Taką granicę przyjmuje GUS na potrzeby analiz krajowych sfery ubóstwa relatywnego. Relatywną linię ubóstwa stosowaną przez EUROSTAT dla celów porównań międzynarodowych wyznacza się jako pewien procent (zwykle 60%) mediany dochodów ekwiwalentnych, do obliczenia których wykorzystuje się skalę zmodyfikowaną OECD typu 0,5/0,3.

$$\textit{Względne ryzyko} \quad \frac{P(Y=1/R=r_1)}{P(Y=1/R=r_2)} = \frac{P_{1/1}}{P_{1/2}}. \tag{1}$$

Gdy np. względne ryzyko wynosi 2, tzn. że prawdopodobieństwo odpowiedzi 1 (tak) jest 2 razy większe w pierwszej grupie respondentów niż w drugiej grupie.

$$\textit{Iloraz szans} \quad \Theta = \frac{P1}{P2}, \tag{2}$$

gdzie:

$$P1 = \frac{P(Y=1/R=r_1)}{P(Y=0/R=r_1)} = \frac{P_{1/1}}{P_{0/1}} \text{ i } P2 = \frac{P(Y=1/R=r_2)}{P(Y=0/R=r_2)} = \frac{P_{1/2}}{P_{0/2}}.$$

$P1$ – jest to szansa, że dla respondentów z pierwszej grupy odpowiedź będzie $Y = 1$, a nie $Y = 0$,

$P2$ – szansa, że dla respondentów z drugiej grupy odpowiedź będzie $Y = 1$, a nie $Y = 0$,

P1 > 1 i P2 > 1 gdy odpowiedź *tak* (z kolumny pierwszej) jest bardziej prawdopodobna niż odpowiedź *nie* (z drugiej kolumny).

Jeśli *P1 = P2,* to zmienne Y i R są niezależne.

Jeśli $1<\Theta<\infty$, to większa szansa wyboru odpowiedzi *tak* (Y=1) jest dla respondentów w grupie 1, niż w grupie 2.

Np. $\Theta = 4$ oznacza, że szansa odpowiedzi *tak* w porównaniu z szansą odpowiedzi *nie* w pierwszej grupie respondentów jest cztery razy większa niż w grupie drugiej.

Model logitowy służy do badania zależności między zmienną binarną Y, przyjmującą tylko dwie wartości, oznaczane symbolicznie 1, 0; a zmiennymi X_1, X_2, ... , Xm, które mogą być zmiennymi zarówno ilościowymi, jak i jakościowymi.

Chcemy znaleźć zależność prawdopodobieństwa wyboru wartości $Y = 1$ od wartości zmiennych objaśnianych X_j. Niech $p = P(Y = 1)$, a x_j będzie wartością zmiennej X_j. Najprostszą zależnością jest zależność liniowa:

$$\mathrm{P}(Y=1) = p = a_0 + \sum_{j=1}^{m} a_j x_j. \tag{3}$$

Parametry a_0, a_1, ... , a_m można wyznaczyć wtedy metodą najmniejszych kwadratów (MNK). Jednakże dla niektórych wartości x_j prawdopodobieństwo p może leżeć poza przedziałem [0, 1], co jest sprzeczne z podstawową własnością prawdopodobieństwa.

Aby nie było takich sprzeczności, wartości prawdopodobieństwa poddaje się transformacji. Najczęściej spotykaną transformacją jest funkcja logitowa:

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right), \tag{4}$$

czyli logarytm szansy, w wyniku której otrzymuje się model logitowy:

$$\text{logit}(p) = a_0 + \sum_{j=1}^{m} a_j x_j = X^T \cdot A, \tag{5}$$

gdzie A oznacza wektor parametrów modelu $A=[a_0, a_1, \ldots, a_m]$, zaś X^T wektor zmiennych objaśniających. Stosując np. metodę największej wiarygodności (NW), można oszacować wektor parametrów A, a następnie obliczyć prawdopodobieństwo p według wzoru:

$$P(Y=1) = p = \frac{e^{X^T \cdot A}}{1 + e^{X^T \cdot A}}. \tag{6}$$

Parametr kierunkowy a_j ma następującą interpretację: jeśli wartość x_j wzrośnie o 1 jednostkę, to szansa, że $Y = 1$ wzrośnie e^{a_j} razy.

2. Wskaźniki ubóstwa

Problem określenia rozmiaru sfery ubóstwa sprowadza się do wyboru odpowiednich wskaźników ubóstwa, które są konstruowane na podstawie różnych linii ubóstwa i pozwalają ocenić zasięg, głębokość, dotkliwość czy intensywność odpowiedniego rodzaju ubóstwa.

Do najprostszych i najczęściej stosowanych wskaźników określających sferę ubóstwa należą stopa ubóstwa i średnia luka wydatkowa albo dochodowa gospodarstw ubogich lub wszystkich gospodarstw objętych badaniem [por. GUS 1998].

Jeśli X jest zmienną losową oznaczającą wydatki lub dochód gospodarstwa domowego o dystrybuancie F(x) i wartości przeciętnej E(X)= μ , a x* odpowiednią granicą ubóstwa, wówczas wskaźnikiem charakteryzującym zasięg ubóstwa jest stopa ubóstwa P_0, którą można wyrazić wzorem:

$$P_0 = F(x^*). \tag{7}$$

Stopa ubóstwa określająca frakcję gospodarstw (czy osób) ubogich ma podstawową wadę, polegającą na tym, że nie pozwala ocenić, w jakim stopniu zjawisko to dotyka gospodarstw uznanych za ubogie: czy są to gospodarstwa o poziomie dochodów bliskim granicy ubóstwa, czy też dochody ich są praktycznie na poziomie zerowym. Ponadto wskaźnik ten jest niewrażliwy na spadek dochodów gospodarstw uznanych za ubogie, jak również na transfery dochodów między gospodarstwami ubogimi i transfery dochodów od gospodarstw ubogich do zamożniejszych.

Pewnym rozwiązaniem jest wskaźnik średniej luki wydatkowej (lub dochodowej) ubogich, który wyraża się wzorem:

$$P_1 = \frac{1}{n_p \cdot x^*} \sum_{i=1}^{n_p} \left(x^* - x_{iek} \right), \tag{8}$$

gdzie n i n_p oznaczają odpowiednio: liczbę wszystkich gospodarstw (lub osób) objętych badaniem i liczbę gospodarstw (lub osób) ubogich, zaś

x_{iek} – dochody lub wydatki ekwiwalentne i-tego gospodarstwa zaliczanego do ubogich.

Wskaźnik ten informuje o ile procent przeciętne wydatki (dochody) gospodarstw domowych uznanych za ubogie są niższe od wartości przyjętej za granicę ubóstwa. Przy tym im uboższe jest gospodarstwo, tym większy jest jego udział w pomiarze głębokości ubóstwa[4].

Bezpośrednim uogólnieniem luki wydatkowej czy dochodowej jest miara ubóstwa zależna od dodatniego parametru α w sposób następujący[5]:

[4] Wskaźnik średniej luki dochodowej jest wrażliwy na transfery dochodów od gospodarstw ubogich do gospodarstw, które znajdują się powyżej linii ubóstwa przed takimi transferami lub po takich transferach [por. Rusnak 2007].

[5] Miara została opisana w pracy [Foster, Greer i Thorbecke 1984].

$$P^{\alpha} = \frac{1}{n}\sum_{i=1}^{n_p}\left(1 - \frac{x_{iek}}{x^*}\right)^{\alpha} . \tag{9}$$

Jeśli $\alpha = 0$, to miara ta jest równa stopie ubóstwa, jeśli $\alpha = 1$, to jest identyczna ze średnią luką wydatkową wszystkich gospodarstw[6]. Jeśli wartość $\alpha = 2$, to otrzymuje się miarę wrażliwą na rozkład dochodów wśród ubogich, wykorzystywaną do oceny dotkliwości ubóstwa. Miara P^2 nadaje większe wagi tym gospodarstwom ubogim, których dochód (czy wydatki) są bardziej oddalone od granicy ubóstwa. Miara ta z uwagi na swoją addytywną strukturę spełnia własność dekompozycji[7], co oznacza, że jej wartość obliczona dla całej zbiorowości gospodarstw domowych jest ważoną sumą wartości obliczonych dla podgrup gospodarstw domowych, gdzie wagami są udziały tych podgrup w całej zbiorowości.

3. Determinanty ubóstwa

Przedmiotem rozważań tej części artykułu jest analiza zależności między ryzykiem zagrożenia ubóstwem gospodarstw domowych a różnymi cechami charakteryzującymi te gospodarstwa. Zmienna zależna Y jest zdefiniowana następująco:

$$Y = \begin{cases} 1 & \text{kiedy gospodarstwo domowe jest ubogie} \\ 0 & \text{kiedy gospodarstwo domowe nie jest ubogie} \end{cases}$$

Korzystając z dostępnych danych i klasyfikacji stosowanych w badaniach budżetów gospodarstw domowych (BBGD) w 2008 r., uwzględniono jako zmienne objaśniające następujące cechy jakościowe i przypisane im kategorie:

---

[6] Wskaźniki P^{α} (dla $\alpha = 0,1,2$) są stosowane w analizach sfery ubóstwa prowadzonych przez Bank Światowy.

[7] Indeksem uwzględniającym zarówno zasięg i głębokość ubóstwa, jak i nierówności w rozkładzie luk dochodowych badanych gospodarstw jest wskaźnik Sena-Shorrocksa-Thona służący jako miara intensywności ubóstwa. Indeks ten został zaproponowany przez A. Sena, a następnie zmodyfikowany przez A.F. Shorrocksa i D. Thona [Za: Panek, red., 2007].

- zmienną TS określającą typ społeczno-ekonomiczny gospodarstwa domowego, gdzie:
  - TS1 oznacza gospodarstwa pracowników,
  - TS2 gospodarstwa rolników,
  - TS3 gospodarstwa pracujących na własny rachunek,
  - TS4 gospodarstwa emerytów i rencistów,
  - TS5 gospodarstwa utrzymujące się z niezarobkowych źródeł,
- zmienną M, która oznacza klasę miejscowości, przy czym:
  - M1 oznacza duże miasta liczące przynajmniej 100 tys. mieszkańców,
  - M2 miasta do 100 tys. mieszkańców,
  - M3 oznacza wieś,
- zmienną R oznaczającą region, w którym znajduje się gospodarstwo domowe, gdzie:
  - R1 oznacza region centralny obejmujący województwa: łódzkie i mazowieckie,
  - R2 to region południowy, do którego należą województwa: małopolskie i śląskie,
  - R3 to region wschodni obejmujący województwa: lubelskie, podkarpackie, świętokrzyskie i podlaskie,
  - R4 to region północno-zachodni z województwami wielkopolskim, zachodnio-pomorskim i lubuskim,
  - R5 oznacza region południowo-zachodni z województwami dolnośląskim i opolskim,
  - R6 to region północny obejmujący województwa: kujawsko-pomorskie, warmińsko-mazurskie i pomorskie.

Ponadto zostały uwzględnione dwie cechy ilościowe, oznaczajace:

X_1 – wielkość gospodarstwa mierzoną liczbą osób w gospodarstwie, przy czym $X_1 = \{1,2,3,4,5,6+]$, gdzie 6+ oznacza gospodarstwo liczące 6 lub więcej osób,

X_2 – liczbę dzieci w gospodarstwie domowym w wieku do 14 lat, $X_2 = [0,1,2,3,4+]$, gdzie 4+ oznacza gospodarstwo, w którym jest przynajmniej czwórka dzieci w wieku do 14 lat.

Struktury gospodarstw domowych ze względu na wymienione w analizie cechy oraz fakt uznania gospodarstwa domowego za ubogie prezentuje tabela 1.

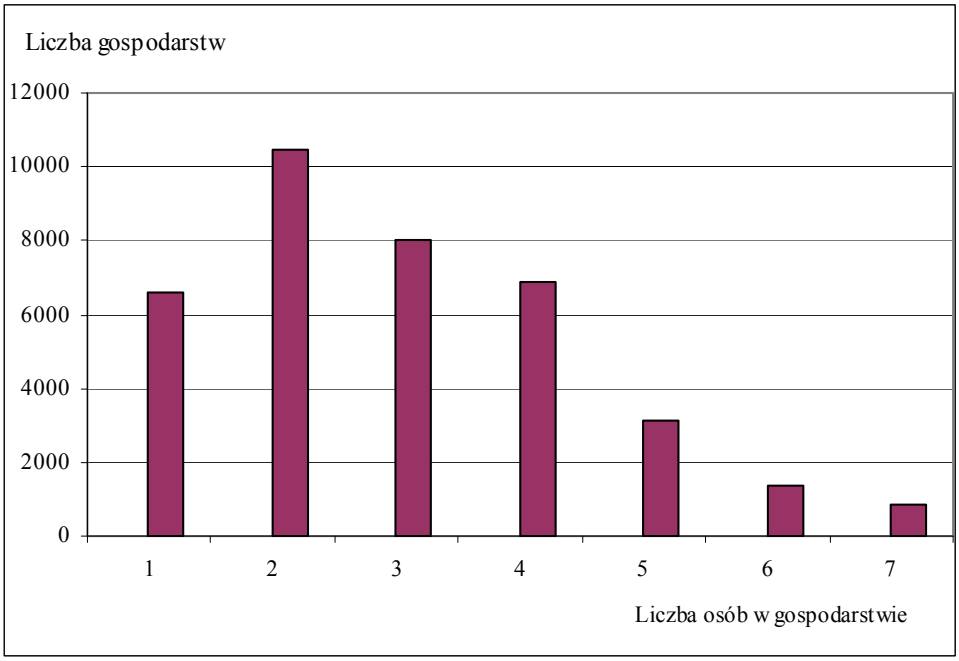
Strukturę gospodarstw domowych objętych BBGD ze względu na wielkość gospodarstwa oraz ze względu na liczbę dzieci w wieku do 14 lat prezentują natomiast rysunki 1 i 2.

Tabela 1

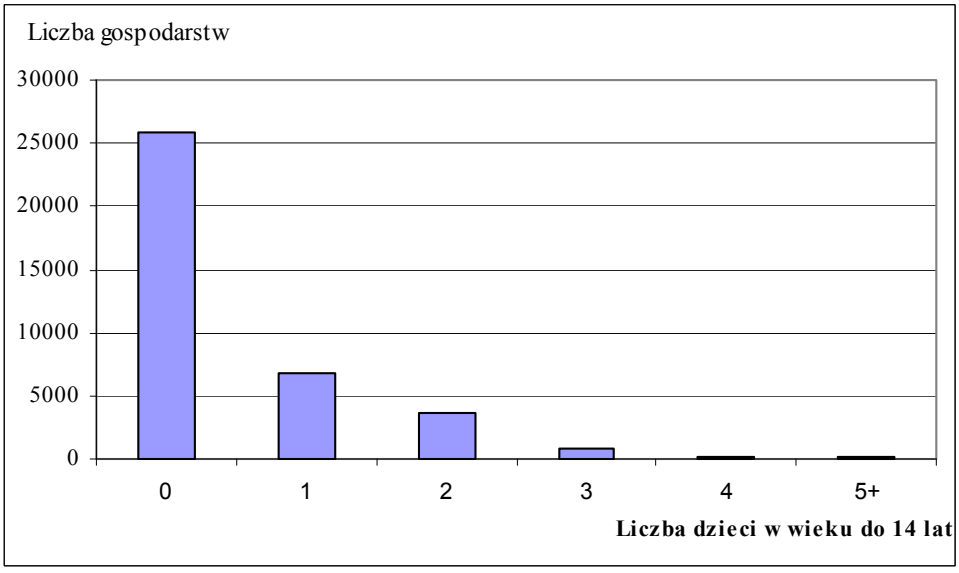
Struktura zbiorowości gospodarstw domowych objętych badaniem i zbiorowości gospodarstw ubogich ze względu na różne cechy społeczno-ekonomiczne

Klasy gospodarstw domowych wyróżnione ze względu na	Odsetek gospodarstw w [%]	Odsetek gospodarstw ubogich w [%]
Typ gospodarstwa TS	**100,00**	
TS1	49,96	13,8
TS2	5,36	22,68
TS3	6,63	7,79
TS4	34,35	13,13
TS5	3,70	31,88
Klasę miejscowości M	**100,00**	
M1	29,06	6,71
M2	28,81	11,89
M3	42,13	21,23
Region	**100,00**	
R1	21,53	10,06
R2	20,09	13,44
R3	17,78	19,39
R4	15,49	13,66
R5	10,68	12,43
R6	14,43	17,72
Liczba gospodarstw w BBGD	**37358**	

Źródło: Na podstawie danych z BBGD.



Rys. 1. Struktura gospodarstw domowych ze względu na wielkość gospodarstwa



Rys. 2. Struktura gospodarstw domowych ze względu na liczbę dzieci w wieku do 14 lat

Z danych prezentowanych w tabeli 1 oraz na rysunkach 1 i 2 wynika, że wśród gospodarstw objętych BBGD:

- dwie najliczniejsze grupy gospodarstw to gospodarstwa pracowników oraz emerytów i rencistów, łącznie stanowiły one prawie 85% całej zbiorowości gospodarstw,
- w większości (prawie 58%) były to gospodarstwa miejskie,
- największy odsetek stanowiły gospodarstwa z regionu centralnego i południowego,
- dominowały gospodarstwa bez dzieci w wieku do 14 lat (prawie 70%),
- większość stanowiły gospodarstwa małe, liczące nie więcej niż 3 osoby i były to głównie gospodarstwa bez dzieci w wieku poniżej 14 lat.

Podstawę uznania gospodarstwa za ubogie stanowiła granica ubóstwa relatywnego wyznaczona na poziomie 50% przeciętnych wydatków ekwiwalentnych gospodarstw domowych. Granica ta wyznaczona przy użyciu oryginalnej skali OECD typu 0,7/0,5 na podstawie danych z BBGD w 2008 r. wyniosła 575,2 PLN.

Gospodarstwa, których wydatki rzeczywiste przeliczone na jednostkę ekwiwalentną były niższe od wyznaczonej granicy ubóstwa zostały uznane za gospodarstwa ubogie, należące do sfery ubóstwa relatywnego.

Z danych prezentowanych w tabeli 1 wynika, że ze względu na główne źródło utrzymania największy odsetek gospodarstw ubogich stanowiły gospodarstwa utrzymujące się z niezarobkowych źródeł oraz gospodarstwa rolników (łącznie prawie 55%). Biorąc pod uwagę region czy klasę miejscowości, można zaobserwować, że wśród gospodarstw ubogich przeważały gospodarstwa zamieszkujące regiony północny i wschodni (łącznie ponad 37%) oraz gospodarstwa zamieszkujące na wsi.

Dane zawarte w tabelach 2 i 3 oraz prezentowane na rysunkach 3 i 4 opisują dostatecznie dokładnie sytuację dochodową i wydatkową analizowanych gospodarstw domowych.

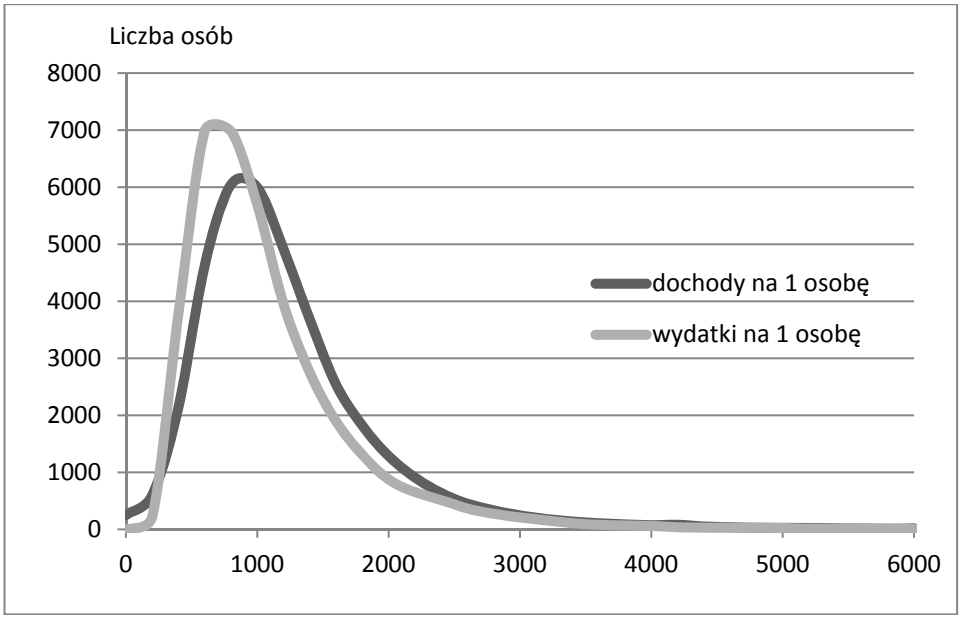
Tabela 2

Średni dochód gospodarstw domowych objętych BBGD w 2008 r.

Liczba gospodarstw w BBGD	37358
Średni miesięczny dochód / 1 gospodarstwo [w PLN]	3007,11
Średni miesięczny dochód / 1 osobę [w PLN]	1163,82
Średnie miesięczne wydatki / 1 gospodarstwo [w PLN]	2602,19
Średnie miesięczne wydatki / 1 osobę [w PLN]	1025,36

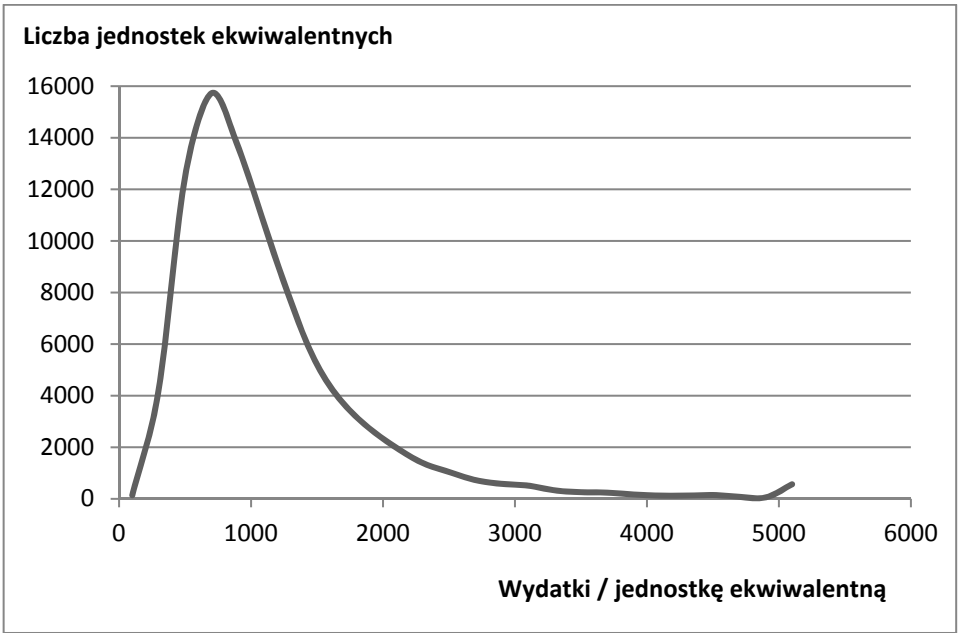
Źródło: Na podstawie BBGD.

Rozkłady dochodów i wydatków przeciętnych na 1 osobę oraz podstawowe charakterystyki tych rozkładów prezentują rysunek 3 i tabela 3.



Rys. 3. Rozkłady przeciętnych miesięcznych dochodów i wydatków na 1 osobę w zbiorowości gospodarstw domowych objętych BBGD w 2008 r.

Rozkłady obydwu charakterystyk, niezależnie od tego czy jednostką badaną jest gospodarstwo, osoba czy jednostka ekwiwalentna (por. rysunek 4) wykazują bardzo silną asymetrię prawostronną, co oznacza, że większość (prawie 65%) gospodarstw osiągało przeciętny dochód miesięczny poniżej średniej arytmetycznej, tzn. poniżej 3007,11 PLN oraz przeciętne wydatki poniżej 2602,19 PLN, zaś w przeliczeniu na 1 osobę są to kwoty odpowiednio poniżej 1163,82 PLN i 1025,36 PLN. Pozycyjne parametry rozkładu dochodów i wydatków prezentowane w tabeli 4, podobnie jak wartości średnie, wskazują na stosunkowo niski poziom zamożności badanych gospodarstw. Świadczy o tym również fakt, że 99,3% gospodarstw objętych badaniem stanowiły gospodarstwa o dochodach poniżej 5000 PLN.



Rys. 4. Rozkład przeciętnych miesięcznych wydatków ekwiwalentnych
Źródło: Na podstawie danych z BBGD.

Na przykład kwartyl Q_3 oznacza, że 75% gospodarstw cechuje poziom miesięcznych dochodów nieprzekraczający 3766 PLN oraz poziom wydatków niższy od 3188 PLN, a w przypadku wydatków na jednostkę ekwiwalentną jest to kwota poniżej 1363 PLN.

Tabela 3

Parametry opisowe rozkładu przeciętnych miesięcznych dochodów i wydatków w przeliczeniu na 1 gospodarstwo oraz na 1 osobę w 2008 r.

Parametry opisowe rozkładów	Rozkład dochodów na		Rozkład wydatków na	
	1 gospodarstwo	1 osobę	1 gospodarstwo	1 osobę
Średnia arytmetyczna	3007,11	1163,82	2602,187	1025,36
Mediana Q_2	2518,59	978,48	2144,61	820,68
Kwartyl pierwszy Q_1	1624,2	668,17	1415,3	556,11
Kwartyl trzeci Q_3	3765,92	1400,00	3177,37	1224,61
Współczynnik zmienności	0,984	1,09	0,796	0,879

Tabela 4

Parametry opisowe rozkładu wydatków ekwiwalentnych w 2008 r.

Parametry opisowe rozkładów	Rozkład wydatków na jednostkę ekwiwalentną
Średnia arytmetyczna	1150,4
Mediana Q_2	938,2
Kwartyl pierwszy Q_1	655,98
Kwartyl trzeci Q_3	1362,3
Współczynnik zmienności	0,807

Źródło: Na podstawie danych z BBGD.

Wysokie wartości współczynników zmienności wskazują z kolei na znaczne zróżnicowanie zarówno dochodów, jak i wydatków gospodarstw domowych i to również niezależnie od stosowanej skali.

Tablice dwudzielcze utworzone z danych jednostkowych z BBGD stanowiły podstawę do testowania hipotez o niezależności uznania gospodarstwo za ubogie od cech opisujących gospodarstwa domowe, uwzględnionych w tabeli 5. Wartości statystyki testowej χ^2, wynoszące dla poszczególnych cech 566,17; 1177,41; 327,58 i 387,24 były znacznie większe od wartości krytycznych, odpowiadających możliwym poziomom istotności, a zatem przemawiały za odrzuceniem hipotez o niezależności i przyjęciem hipotezy alternatywnej, mówiącej o tym, że istnieje zależność między uznaniem gospodarstwa domowego za ubogie a typem gospodarstwa domowego, klasą miejscowości i regionem, w którym gospodarstwo zamieszkuje. W celu określenia tej zależności zostały obliczone miary omówione w podrozdziale 1.

W pierwszym rzędzie obliczono względne ryzyko i iloraz szans uznania za ubogie gospodarstw różnego typu społeczno-ekonomicznego wyróżnionych spośród wszystkich gospodarstw objętych badaniem. Odpowiednie obliczenia zawiera tabela 5.

Tabela 5

Względne ryzyko i iloraz szans uznania gospodarstw domowych wyróżnionych typów za ubogie

Prawdopodobieństwa warunkowe, że gospodarstwo jest ubogie, jeśli jest ono określonego typu	Względne ryzyko dla wybranych typów gospodarstw	Szanse warunkowe	Iloraz szans dla wybranych typów gospodarstw
$P(Y = 1/TS1) = P_{1/1} =$ **0,138**	$P_{1/1}/P_{1/3} = 1,767$	$P1 = P_{1/1}/P_{0/1} = 0,16$	P2/P1 = 1,832
$P(Y = 1/TS2) = P_{1/2} = 0,227$	$P_{1/2}/P_{1/3} = 2,910$	$P2 = P_{1,2}/P_{0/2} = 0,293$	P5/P1 = 2,923
$P(Y = 1/TS3) = P_{1/3} =$ **0,078**	$P_{1/4}/P_{1/3} = 1,679$	$P3 = P_{1/3}/P_{0/3} = 0,084$	P1/P3 = 1,905
$P(Y = 1/TS4) = P_{1/4} =$ **0,131**	$P_{1/5}/P_{1/1} = 2,312$	$P4 = P_{1/4}/P_{0/4} = 0,151$	P5/P3 = 5,571
$P(Y = 1/TS5) = P_{1/5} = 0,319$	$P_{1/5}/P_{1/3} = 4,089$	$P5 = P_{1/5}/P_{0/5} = 0,468$	P2/P3 = 3,488

Jak można zaobserwować, najbardziej zagrożone ubóstwem relatywnym są gospodarstwa utrzymujące się z niezarobkowych źródeł (TS5) oraz gospodarstwa rolników (TS2), a najmniej gospodarstwa pracujących na własny rachunek (TS3). Zarówno prawdopodobieństwa warunkowe, względne ryzyko, jak i iloraz szans uznania za ubogie gospodarstw typu TS5 i TS2 w porównaniu do pozostałych typów gospodarstw przyjmują największe wartości. Na przykład największe wartości względnego ryzyka 4,089[8] i 2,91 oznaczają, że prawdopodobieństwo przynależności gospodarstwa typu TS5 do tej sfery ubóstwa jest ponad 4-krotnie większe, a dla gospodarstw typu TS2 prawie 3-krotnie większe niż dla gospodarstw pracujących na własny rachunek (TS3). Iloraz szans równy 5,571 oznacza z kolei, że szansa uznania gospodarstwa typu TS5 za ubogie w porównaniu z szansą, że nie będzie ono uznane za ubogie jest prawie 5,6-krotnie większa niż w przypadku gospodarstw typu TS3. W analogicznych porównaniach dla gospodarstw typu TS2 krotność ta wynosi 3,488. Podobnie można interpretować pozostałe wyniki prezentowane w tabeli 5.

W tabelach 6 i 7 zostały przedstawione wyniki obliczeń dotyczących względnego ryzyka oraz ilorazów szans uznania gospodarstw domowych za ubogie z uwzględnieniem klasy miejscowości i regionu, w którym zamieszkują gospodarstwa domowe.

[8] Bardzo duże liczebności w odpowiednich tablicach dwudzielnych pozwoliły na oszacowanie prawdopodobieństw za pomocą zgodnych i nieobciążonych estymatorów, jakimi są częstości występowania odpowiednich zdarzeń.

Tabela 6

Względne ryzyko i iloraz szans wpadania do strefy ubóstwa relatywnego dla gospodarstw zamieszkujących w określonych klasach miejscowości

Prawdopodobieństwa warunkowe, że gospodarstwo jest ubogie, jeśli jest z danej klasy miejscowości	Względne ryzyko	Szanse warunkowe	Iloraz szans
$P(Y = 1/M1) = P_{1/1} = 0,067$	$P_{1/3}/P_{1/1} = 3,165$	$P1 = P_{1/1}/P_{0/1} = 0,072$	P2/P1 = 1,875
$P(Y = 1/M2) = P_{1/2} = 0,119$	$P_{1/3}/P_{1/2} = 1,782$	$P2 = P_{1,2}/P_{0/2} = 0,135$	P3/P1 = 3,708
$P(Y = 1/M3) = P_{1/3} = 0,212$	$P_{1/2}/P_{1/1} = 1,776$	$P3 = P_{1/3}/P_{0/3} = 0,267$	P3/P2 = 1,978

Tabela 7

Względne ryzyko i iloraz szans wpadania do strefy ubóstwa relatywnego dla gospodarstw zamieszkujących w poszczególnych regionach

Prawdopodobieństwa warunkowe, że gospodarstwo jest ubogie, jeśli jest z określonego regionu	Względne ryzyko dla wybranych typów gospodarstw	Szanse warunkowe	Iloraz szans dla wybranych typów gospodarstw
$P(Y = 1/R1) = P_{1/1} = 0,101$	$P_{1/2}/P_{1/1} = 1,327$	$P1 = P_{1/1}/P_{0/1} = 0,112$	P2/P1 = 1,384
$P(Y = 1/R2) = P_{1/2} = 0,134$	$P_{1/3}P_{1/1} = 1,921$	$P2 = P_{1,2}/P_{0/2} = 0,155$	P3/P1 = 2,152
$P(Y = 1/R3) = P_{1/3} =$ **0,194**	$P_{1/4}/P_{1/1} = 1,356$	$P3 = P_{1/3}/P_{0/3} = 0,241$	P4/P1 = 1,420
$P(Y = 1/R4) = P_{1/4} = 0,137$	$P_{1/5}/P_{1/1} = 1,228$	$P4 = P_{1/4}/P_{0/4} = 0,159$	P5/P1 = 1,268
$P(Y = 1/R5) = P_{1/5} = 0,124$	$P_{1/6}/P_{1/1} = 1,752$	$P5 = P_{1/5}/P_{0/5} = 0,142$	P6/P1 = 1,920
$P(Y = 1/R6) = P_{1/6} =$ **0,177**	$P_{1/3}/P_{1/6} = 1,096$	$P6 = P_{1/6}/P_{0/6} = 0,215$	P3/P6 = 1,121

Analiza wyników przedstawionych w tabelach 6 i 7 pozwala na sformułowanie następujących wniosków:

- najbardziej zagrożone ubóstwem relatywnym są gospodarstwa zamieszkujące na wsi, względne ryzyko dla tych gospodarstw jest ponad 3-krotnie większe niż w gospodarstwach zamieszkujących miasta liczące powyżej 100 tys. mieszkańców i ponad 1,7 razy większe niż w miastach liczących do 100 tys. ludności,
- iloraz szans 3,71 wskazuje, że również szansa uznania gospodarstwa domowego za ubogie (w porównaniu z szansą, że nie będzie uznane za ubogie) jest w gospodarstwach wiejskich 3,7-krotnie większa niż w gospodarstwach zamieszkujących w dużych miastach,

– biorąc pod uwagę region, w którym zamieszkuje gospodarstwo, można zauważyć, że w najmniejszym stopniu zagrożone ubóstwem relatywnym są gospodarstwa zamieszkujące region centralny, natomiast najbardziej zagrożone są gospodarstwa z regionu wschodniego i północnego, wskazują na to wartości względnego ryzyka i ilorazów szans zawarte w tabeli 7.

Analiza zagrożenia ubóstwem relatywnym gospodarstw domowych została również przeprowadzona za pomocą modelu regresji logistycznej, w którym prawdopodobieństwo uznania gospodarstwa za ubogie jest uzależnione od typu gospodarstwa (zmienna TS), miejsca zamieszkania (obejmującego klasę miejscowości M i region R) oraz wielkości gospodarstwa (zmienna X1) i liczby dzieci w gospodarstwie wieku do 14 lat (zmienna X2).

Gospodarstwa referencyjne stanowiły jednoosobowe gospodarstwa domowe pracowników, bez dzieci w wieku poniżej 14 lat, zamieszkujące w regionie centralnym, w miastach liczących do 100 tys. mieszkańców. Wyniki estymacji modelu logitowego zawiera tabela 8.

Tabela 8

Wyniki estymacji modelu regresji logistycznej dla prawdopodobieństwa uznania gospodarstwa domowego za ubogie

Zmienne objaśniające	Ocena parametru a_i	Iloraz szans
Stała	-3,738	0,024
TS2	0,076	1,027
TS3	-0,626	0,535
TS4	0,563	1,754
TS5	1,650	**5,209**
M1	-0,528	0,589
M3	0,463	1,589
R2	0,306	1,358
R3	0,420	1,522
R4	0,178	1,195
R5	0,183	1,201
R6	0,555	1,741
X1	0,392	1,480
X2	0,066	1,068
Miary dopasowania	χ^2	Całkowita stara
	3444,5	13618,82

Źródło: Na podstawie danych z BBGD.

Wszystkie oceny parametrów są statystycznie istotne, co oznacza, że uwzględnione w modelu zmienne mają istotny wpływ na prawdopodobieństwa osiągnięcia przez gospodarstwo domowe statusu gospodarstwa ubogiego. Przy określonej powyżej grupie gospodarstw referencyjnych, dodatnie wartości ocen parametrów przy odpowiednich zmiennych wskazują, że większe prawdopodobieństwa uznania gospodarstwa za ubogie w porównaniu z gospodarstwem referencyjnym cechuje gospodarstwa typu TS4 i TS5, zamieszkujące na wsi, w dowolnym regionie z wyjątkiem regionu centralnego. Prawdopodobieństwo to rośnie wraz z rosnącą liczbą osób w gospodarstwie oraz rosnącą liczbą dzieci w wieku do 14 lat.

Analizując ilorazy szans zaprezentowane w tabeli 7, można bowiem stwierdzić, że np.:

- jeśli gospodarstwa są tego samego typu i miejscem zamieszkania jest ta sama klasa miejscowości, w tym samym regionie, to szansa uznania gospodarstwa, w którym jest o jedną osobę więcej, rośnie prawie 1,5-krotnie, zaś zwiększenie liczby dzieci w wieku do 14 powoduje zwiększenie tej szansy o 6,8 p.p,
- jeśli gospodarstwa są tej samej wielkości, z taką samą liczbą dzieci (do 14 lat) i są to gospodarstwa zamieszkujące w tym samym regionie, w miejscowości należącej do tej samej klasy, to najbardziej zagrożone ubóstwem są gospodarstwa typu TS5 (iloraz szans wynosi 5,209),
- jeśli gospodarstwa różnią się tylko klasą miejscowości, w której zamieszkują, to szansa osiągania statusu gospodarstwa ubogiego jest ponad 1,5-krotnie większa dla gospodarstw wiejskich w porównaniu z gospodarstwem zamieszkującym w mieście do 100 tys. ludności,
- ujemne wartości ocen parametrów przy pozostałych zmiennych wskazują, że zmniejszenie prawdopodobieństwa zagrożenia ubóstwem jest spowodowane m.in. tym, że gospodarstwo jest gospodarstwem pracującym na własny rachunek i zamieszkuje w mieście liczącym powyżej 100 tys. mieszkańców. Ilustrują to prawdopodobieństwa zagrożenia ubóstwem dla wybranych grup gospodarstw domowych obliczone na podstawie oszacowanego modelu logitowego prezentowane w tabeli 9. Ujemne wartości ocen parametrów przy zmiennych TS3 i M1 znajdują odzwierciedlenie w najmniejszych prawdopodobieństwach zagrożenia ubóstwem gospodarstw opisanych tymi zmiennymi, zawartymi w tabeli 9. Analizując prawdopodobieństwa w tabeli 9, można zaobserwować, że wraz ze wzrostem liczby osób rośnie prawdopodobieństwo uznania gospodarstwa za ubogie.

Tabela 9

Prawdopodobieństwa uznania gospodarstwa za ubogie na podstawie modeli logitowych, dla wybranych typów gospodarstw

Region	Klasa miejscowości	Typ gospodarstwa domowego	P(Y=1)		
			jeśli $X_1 = 1$ i $X_2 = 0$	jeśli $X_1 = 2$ i $X_2 = 0$	jeśli $X_1 = 2$ i $X_2 = 1$
R3	M1	TS1	0,031	0,045	0,048
		TS2	0,033	0,043	0,051
		TS3	0,017	0,024	0,026
		TS4	0,053	0,076	0,081
		TS5	0,141	0,196	0,207
R3	M3	TS1	0,078	0,112	0,119
		TS2	0,084	0,120	0,127
		TS3	0,044	0,063	0,067
		TS4	0,130	0,181	0,191
		TS5	0,307	0,396	0,412
R6	M1	TS1	0,035	0,051	0,054
		TS2	0,038	0,055	0,058
		TS3	0,019	0,028	0,030
		TS4	0,060	0,086	0,091
		TS5	0,159	0,218	0,230

Źródło: Na podstawie danych z BBGD.

Podobne obliczenia przeprowadzone dla gospodarstw o różnym składzie demograficznym, pokazały, że największe prawdopodobieństwo uznania gospodarstwa za ubogie odpowiada gospodarstwom utrzymującym się z niezarobkowych źródeł, mieszkającym na wsi, w regionie wschodnim. Najmniejsze prawdopodobieństwo zagrożenia ubóstwem odnosi się natomiast do gospodarstw pracujących na własny rachunek, mieszkających w mieście powyżej 100 tys. mieszkańców, w regionie centralnym.

4. Ocena sfery ubóstwa w ujęciu regionalnym

W analizie ubóstwa relatywnego przedstawionej w tym artykule jako miernik zamożności gospodarstw domowych zostały przyjęte wydatki konsumpcyjne tych gospodarstw. Sytuację wydatkową gospodarstw domowych w poszczególnych regionach opisują dane zawarte w tabeli 10. W 2008 r. najwyższy poziom wydatków, niezależnie od stosowanej skali, cechował region centralny, natomiast najniższy regiony wschodni i północny.

Tabela 10

Średnie wydatki w zbiorowości wszystkich gospodarstw z BBGD

Region	Średnie wydatki w zbiorowości wszystkich gospodarstw objętych BBGD przypadające na:		
	1 gospodarstwo	1 osobę	1 jednostkę ekwiwalentną
Polska	**2602,18**	**1025,36**	**1150,40**
Centralny	**2883,12**	**1038,33**	**1335,58**
Południowy	2562,73	884,24	1144,59
Wschodni	***2428,31***	***760,39***	***1001,73***
Płn.-Zachodni	2558,56	848,11	1105,59
Płd.-Zachodni	2631,40	954,22	1225,48
Północny	***2482,13***	***831,49***	***1087,95***

Źródło: Na podstawie danych z BBGD.

Granica ubóstwa relatywnego wyznaczona na poziomie połowy średnich wydatków ekwiwalentnych stanowiła podstawę do wyznaczenia stopy ubóstwa relatywnego oraz średniej luki wydatkowej gospodarstw ubogich dla zbiorowości gospodarstw zamieszkujących w poszczególnych regionach – por. wzory (7) i (8). Wyniki obliczeń prezentuje tabela 11.

Jak można zaobserwować, region centralny charakteryzują najniższe wartości wskaźników dotyczących tak zasięgu, jak i głębokości ubóstwa. Najbardziej zagrożone ubóstwem są gospodarstwa domowe zamieszkujące regiony: wschodni i północny, przy czym w największym stopniu zjawisko to dotyka dzieci w wieku do 14 lat.

Tabela 11

Stopy ubóstwa relatywnego i średnia luka wydatkowa w regionach Polski w 2008 r.

Region	Stopa ubóstwa P_0 określająca [%] ubogich:			Średnia luka wydatkowa ubogich P_1	Wskaźnik P^2 jako miara dotkliwości ubóstwa
	gospodarstw	osób	dzieci w wieku do 14 lat		
Polska	14,3	18,7	24,9	0,218	0,073
Centralny	**10,1**	**13,4**	**18,7**	**0,215**	**0,07**
Południowy	13,4	17,2	23,5	0,204	0,064
Wschodni	***19,4***	***24,0***	***28,3***	***0,235***	***0,081***
Płn.-Zachodni	13,7	18,0	23,9	0,208	0,068
Płd.-Zachodni	12,4	16,3	21,2	0,210	0,068
Północny	***17,7***	***23,4***	***32,6***	***0,224***	***0,077***

Źródło: Na podstawie danych z BBGD.

Regiony wschodni i północny cechują wysokie wartości wszystkich rodzajów stóp ubóstwa i luki wydatkowej ubogich. Małe zróżnicowanie miary P^2 świadczy o tym, że stopień dotkliwości ubóstwa jest w poszczególnych regionach na podobnym poziomie.

Dane w tabeli 12 prezentują ranking regionów ze względu na stopę ubóstwa P_0, średnią lukę wydatkową ubogich P_1, średnią lukę wydatkową P^1 i wskaźnik P^2.

Tabela 12

Ranking regionów ze względu na wskaźniki ubóstwa

Region	*P*o	Pozycja	*P*1	Pozycja	P^1	Pozycja	P^2	Pozycja
Centralny	**0,101**	6	0,215	3	0,022	6	0,071	3
Południowy	0,134	4	0,204	6	0,027	4	0,064	6
Wschodni	***0,194***	**1**	***0,235***	**1**	***0,046***	**1**	***0,081***	**1**
Płn.-Zachodni	0,137	3	0,208	5	0,028	3	0,068	5
Płd.-Zachodni	0,124	5	0,210	4	0,026	5	0,068	4
Północny	***0,177***	***2***	***0,224***	***2***	***0,040***	***2***	***0,077***	***2***

Źródło: Na podstawie danych z BBGD.

Dane te wskazują, że gospodarstwa w regionach wschodnim i północnym są najbardziej zagrożone ubóstwem (pozycje 1 i 2), ponieważ wskaźniki służące do oceny zasięgu, głębokości i dotkliwości sfery ubóstwa w tych regionach przyjmują największe wartości. W najmniejszym stopniu zagrożone ubóstwem są gospodarstwa domowe z regionu centralnego.

Przyjmując zatem jako wskaźnik zamożności gospodarstw średnie wydatki, można stwierdzić, że zarówno w zbiorowości wszystkich badanych gospodarstw, jak i w zbiorowości gospodarstw uznanych za ubogie, w najlepszej sytuacji finansowej znajdują się gospodarstwa z regionu centralnego, zaś w najgorszej z regionów: wschodniego i północnego.

Podsumowanie

Podstawowym celem niniejszej pracy była ocena istotności wpływu wybranych cech społeczno-ekonomicznych, charakteryzujących gospodarstwa domowe w Polsce na prawdopodobieństwo tego, że gospodarstwa domowe określonego typu znajdą się w sferze ubóstwa relatywnego. Do realizacji tego celu zostały wykorzystane pewne metody pomiaru cech jakościowych, takie jak: względne ryzyko, iloraz szans czy model regresji logistycznej. Otrzymane wyniki pozwoliły na sformułowanie m.in. następujących wniosków:

- wszystkie uwzględnione w analizach ubóstwa relatywnego zmienne miały istotny wpływ na prawdopodobieństwo uznania gospodarstwa domowego za ubogie,
- do cech zwiększających ryzyko zagrożenia ubóstwem należy zaliczyć wielkość gospodarstwa oraz liczbę dzieci w wieku do 14 lat,
- najbardziej zagrożone ubóstwem w 2008 r. były gospodarstwa domowe utrzymujące się z niezarobkowych źródeł, zamieszkujące na wsi w regionie wschodnim lub północnym,
- najmniejsze ryzyko uznania gospodarstw za ubogie dotyczy jednoosobowych gospodarstw pracujących na własny rachunek, które zamieszkują w miastach powyżej 100 tys. ludności, w regionie centralnym.

W pracy dokonano również oceny i porównań sfery ubóstwa relatywnego w ujęciu regionalnym z wykorzystaniem podstawowych oraz najczęściej stosowanych wskaźników ubóstwa, do których należą stopa ubóstwa relatywnego, średnia luka wydatkowa ubogich, a także wskaźnik dotkliwości ubóstwa.

Przeprowadzona analiza zasięgu i głębokości ubóstwa potwierdza otrzymane wcześniej wyniki, bowiem wartości wszystkich miar ubóstwa wskazują na najgorszą sytuację gospodarstw domowych z regionów: wschodniego i północnego oraz najkorzystniejszą gospodarstw z regionu centralnego.

Literatura

Foster J., Greer J., Thorbecke E. (1984): *A Class of Decomposable Poverty Measures.* „Econometrica", Vol. 52, No. 3.
GUS (1998): *Warunki życia ludności w 1997 r.* GUS, Warszawa.
Panek T., red. (2007): *Statystyka społeczna.* PWE, Warszawa.
Rusnak Z. (2007): *Statystyczna analiza dobrobytu ekonomicznego gospodarstw domowych.* AE im. O. Langego, Wrocław.

CHOSEN METHODS OF MEASURING QUALITATIVE CHARACTERISTICS IN POVERTY ANALYSES

Summary

In classical approach to poverty spheres' analyses – both objective and subjective – one uses poverty indicators that characterize mostly the range and the depth of this phenomenon. One of the most basic aspects in a multivariate approach is to determine these factors that increase the risk of poverty.

The main aim of this paper is to characterize the poverty determinants as well as to estimate the risk of households becoming threatened by this phenomenon. Chosen methods of measuring qualitative characteristics will be used to achieve this aim.

The second goal of this paper is an attempt at estimating and comparing poverty spheres in regional approach by means of most important poverty indicators.

The source of the data in both cases is unidentifiable unitary data from household budget research carried out by CSO in 2008 and made available for academic research.

Katarzyna Ostasiewicz
Uniwersytet Ekonomiczny we Wrocławiu

MODELS OF WILLINGNESS TO PAY FOR SUSTAINABLE DEVELOPMENT

Introduction

Growing consciousness of dangers that overexploatation and overpollution of natural environment bring to the human beings causes growing careness of natural resources, involving increasing interest in the questions of sustainable development. According to the definition of World Commission on Environment and Development (better known as Brundtland Commission), sustainable development is such one that "(…) meets the needs of the present without compromising the ability of future generations to meet their own needs" [United Nations 1987].

From economic point of view one of the problems is, that these natural goods have no obvious price and are difficult to be involved into economic accounts. Thus instead of asking about price of clear water or fresh air, economists rather ask people, which amount of money are they ready to pay for obtaining some defined desired state. This is called a *willingness to pay*, and has to be estimated in any attempt to introduce changes, that will lead to the better state of natural environment. It is obvious, that new "clear" technologies will cost more than traditional ones, at least in the early stage of their introduction. Thus, there are needed tools for investigating whether the society is ready to bring additional costs for introducing new technologies – and the level of these acceptable additional costs.

In this paper a certain simple model of willingness to pay for public ecological goods will be proposed.

1. Utility function and willingness to pay

There are two approaches to modeling willingness to pay [Haab and McConnell 2003]. One is based on the random utility function. The other one,

which is possible only in the case of a binary choice, is to model willingness to pay directly, not through the medium of utility function. Both approaches have advantages: the first one is more universal while the second enables to model a bounded willingness to pay. Let s briefly summarize both of these approaches.

Within approach based on a random utility function it is assumed, that each individual has a defined utility of each possible scenario. The utility of scenario i for individual j, u_{ij}, depends on a vector of individual preferences and characteristics of an individual j, $\bar{z}_j$; his/her income, y_j (this particular characteristic of an individual is excluded from the vector $\bar{z}_j$ for reasons, which will become obvious soon); and also on a random variable, ε_{ij}:

$$u_{ij} \equiv f_i(y_j, \bar{z}_j, \varepsilon_{ij}). \tag{1}$$

Mostly, it is assumed, that deterministic and random parts of utility function are additive:

$$u_{ij} = v_i(y_j, \bar{z}_j) + \varepsilon_{ij}. \tag{2}$$

Individuals are assumed to choose this option, of which his/her utility has greater value. Thus, if an individual j is proposed to pay a certain sum of money, t_j, for a scenario 1 to be realized, he/she will agree conditioned that:

$$v_1(y_j - t_j, \bar{z}_j) + \varepsilon_{1j} > v_0(y_j, \bar{z}_j) + \varepsilon_{0j}. \tag{3}$$

Note, that within scenario 1 the income of an individual is lessen by a proposed sum of money t_j – and that is a cause of excluding income from the vector of individual characteristics and taking it explicitly into regard.

As utility function involve a random term, also the decisions of individuals will have a probabilistic character. Denoting a difference of random variables of two scenarios by ε_j:

$$\varepsilon_j \equiv \varepsilon_{0j} - \varepsilon_{1j}, \tag{4}$$

the probability, that an individual j will agree on a proposed payment t_j (what will be denoted by $\omega_j = 1$, while $\omega_j = 0$ will mean answer "no") may be expressed in terms of cumulative distribution function of this random variable ε_j:

$$\begin{aligned} \Pr(\omega_j = 1) &= \Pr\left(\varepsilon_{0j} - \varepsilon_{1j} < v_1(y_j - t_j, \bar{z}_j) - v_0(y_j, \bar{z}_j)\right) \\ &= \mathrm{F}\left(v_1(y_j - t_j, \bar{z}_j) - v_0(y_j, \bar{z}_j)\right), \end{aligned} \tag{5}$$

where:

$\mathrm{F}(x) = \Pr(\varepsilon_j < x)$.

Parameters of functions v_0 and v_1 may be estimated by means of maximizing the likelihood function:

$$L = \prod_j \left[\mathrm{F}\left(v_1(y_j - t_j, \bar{z}_j) - v_0(y_j, \bar{z}_j)\right)\right]^{\omega_j} \left[1 - \mathrm{F}\left(v_1(y_j - t_j, \bar{z}_j) - v_0(y_j, \bar{z}_j)\right)\right]^{1-\omega_j} \quad (6)$$

The willingness to pay of an individual j for scenario 1, WTP_j, is defined as such a sum of money, for which utility of scenarios 1 and 0 are equal:

$$v_1(y_j - WTP_j, \bar{z}_j) + \varepsilon_{1j} = v_0(y_j, \bar{z}_j) + \varepsilon_{0j}. \quad (7)$$

Transforming (7) one may obtain an expression for willingness to pay:

$$WTP_j(\varepsilon_j) = y_j - v_1^{-1}(v_0 - \varepsilon). \quad (8)$$

In general, it may be a complicated function of a random variable ε_j. However, if ε_j has an infinite distribution (like normal or logistic ones) also WTP_j has an infinite distribution.

There are at least two problems with random utility function approach. The first is the possible infinity of distribution of willingness to pay, mentioned already. In some cases it may turn out, that the probabilities of willingness to pay less than zero or exceeding income are quite significant. In special situations it may be so, that an individual will accept a proposed scenario only conditioned that he/she will be paid for it (willingness to pay less than zero), however, the willingness to pay exceeding income still makes sense in no circumstances.

The second problem is in comparing utilities. Most of ecological goods belong to the classes of so-called common-pool resources and public goods (according to Elinor Ostrom classification [Ostrom 2009]). Both these classes are characterized by a "high difficulty of excluding potential beneficiaries" [Ostrom 2009]. In the context of this kind of goods there appear a problem of "free riding" [Wiser 2007; Champ and et al. 1997]. Paying for any "private good", e.g. food or clothing, one always gets what he/she has paid for. On the

contrary, due to high difficulty of excluding potential beneficiaries, one can be never sure of the effect of his/her paying for any common-pool resource or public good. Thus it may seem more adequate to take into regard the willingness to pay in such cases, instead of the utility of scenario which may never be realized. The general form of willingness to pay function reads:

$$WTP_j \equiv WTP_j(\bar{z}_j, \varepsilon_j), \tag{9}$$

where $\bar{z}_j$ denotes a vector of individual preferences and characteristics (may also include income – note, that here there is no necessity of including it explicitly as a function variable), and ε_j is a random variable with a distribution to be specified.

In attempts to model willingness to pay directly, one should take also the first of the problems mentioned above into regard and choose the model function so to assure, that willingness to pay will range from zero to income:

$$0 \leq WTP_j \leq y_j.$$

The natural form of such function will in general read:

$$WTP_j(\bar{z}_j, \varepsilon_j) = G_j(\bar{z}_j, \varepsilon_j) y_j, \tag{10}$$

with $0 \leq G_j(\bar{z}_j, \varepsilon_j) \leq 1$.

Estimation of parameters of function G proceed as usually, by means of finding maximum of likelihood function.

Later on we will propose a simple definite form of willingness to pay function. First, let us ground the choice of key variables that will appear in the proposed function.

2. Social interactions and altruism

A common sense and life knowledge have never negated the fact, that people's behavior strongly depend on the behavior of the others. This is the base of so-called "social norms" (which may differ from society to society), without which the social harmony wouldn't be possible. Moreover, in the context of modern knowledge the "herd behavior" of individuals – no matter whether judged positively or negatively – cannot be denied [Aronson 2003]. However,

economists for a long time tried to get around, treating all social interactions as mediated by the market. This approach has turned out to be unsatisfactory in explanations of phenomena of real social world, and thus there are arising consecutive models, that are taking social interactions explicitly into account [Ostasiewicz et al. 2008; Blume 1995; Brock and Durlauf 2001].

Some of them are based on the concept of Mark Granovetter, who was the author of widely known threshold model [Granovetter 1997]. The main idea of the threshold model is that a given individual will join a certain action ($\omega_j = 1$) conditioned that a certain percent of the others has already joined it:

$$P(\omega_j = 1) = \begin{cases} 1 & \text{if} \quad m \geq Th_i \\ 0 & \text{if} \quad m < Th_i \end{cases}. \tag{11}$$

This percent, Th_i, may be different for different individuals (in particular, it may equals 0), and is called a threshold of the individual. For the action to be initialized there must exist at least one person with the zero threshold.

Dependence of choices of individuals on the choices of the others implies a dynamical character of the model. This dynamics may be called a kind of a "domino effect", as action of some individuals results in actions of some other individuals and so on. The percent of participants of the action changes from time step to next time step according to the following expression:

$$m(t) = F_{Th}\big(m(t-1)\big), \tag{12}$$

where F_{Th} denotes distribution of thresholds across the whole population.

Stationary states, that is, such states that are reached within long enough time and do not evolve in time any more, denoted by m^*, may be obtained from:

$$m^* = F_{Th}(m^*). \tag{13}$$

In the quasi continuous version expression (12) takes the form:

$$\frac{dm}{dt} = F_{Th}(m) - m, \tag{14}$$

while stationary state condition remains the same, see (13).

The evolution of percent of participants of an action and finding a stationary state may be easily pictured, as on an exemplary graph below.

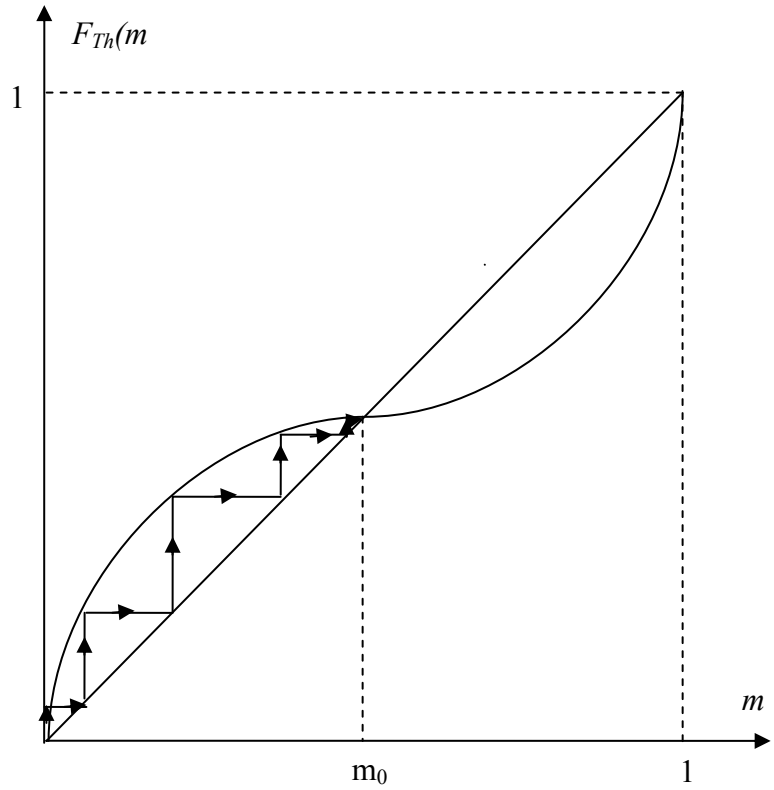


Fig. 1. An example of evolution of percent of participants and obtaining a stationary state in threshold model

Let us also define a so-called "potential", $V(m)$, which is a counterpart of physical potential. Let us define it as follows:

$$\frac{dm}{dt} = -\frac{dV(m)}{dm}. \tag{15}$$

This construct is very useful for determining the character of stationary points (there may be stable or unstable stationary states). The properties of potential defined by (15) are as follows:

1) it has an extremum in the stationary point, what follows from the condition:

$$\frac{dm}{dt} = 0 \Rightarrow \frac{dV(m)}{dm} = 0;$$

2) this extremum is a minimum for stable stationary point and maximum for unstable one.

Thus, having a potential one may easily decide, which solutions of stationary state condition are stable and which are unstable ones, imagining a ball, rolling in the valley of the shape of potential. On the other hand, apart from the definition (15) the potential may be expressed as follows (by substituting (14) to definition (15)):

$$V(m) = \frac{m^2}{2} - \int_0^m F_{Th}(m')dm'. \tag{16}$$

As an example, the shape of the potential for the system shown on Figure 1 is presented below.

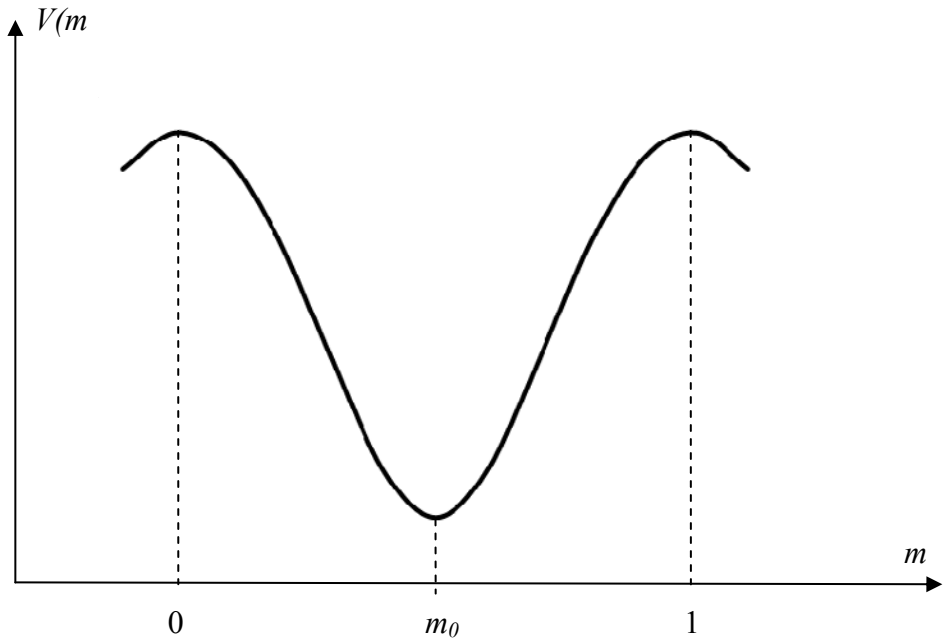


Fig. 2. Potential corresponding to the system presented on Fig. 1

The second key variable in the model that will be presented in next sections is a degree of altruism. It may be a matter of discussion, whether an altruist feelings are an artifact of social interactions (inner social norms), product of mirror neurons or even something transcendental. Anyway, it is the fact, that in many situations many individuals behave altruistically, even if sometimes it may be perceived as not a "pure" altruism but rather a reciprocal one or forced by social pressure. Otherwise, no cooperation would be possible and all common-pool re-

sources would be already destroyed, as forecasted by Garrett Hardin [Hardin 1968]. However, as Elinor Ostrom has definitely proved [Ostrom 2009], there are many circumstances in which people are willing to cooperate and behave not accordingly to their sole economical profit. This kind of altruism may be viewed as a "long-range" rationality. However, there are also many situations, when individuals behave altruistically not having in perspective a reciprocity of long-time economical gain. Daniel Kahneman [Kahneman and Knetsch 1992] stated, that people are willing to pay not only for measureable economical goods but are willing also to "buy" a moral satisfaction of behaving "good".

The degree of altruism may be measured by some standard questionnaires of this feature of human character, and will be taken in what follows as ranging from 0 to 1. It will be also assumed, that that greater degree of altruism the less temptation of free riding of an individual.

3. Willingness to pay model

Taking social nature of human beings and temptation of "free riding" into account, the general form of model of willingness to pay for public ecological goods will be as follows:

$$WTP_j \equiv WTP_j\left(\bar{z}_j, m, a_j, \varepsilon_j\right), \tag{17}$$

where m denotes a percent of the whole population which is also willing (or is thought/expected by a given individual j to be willing) to pay; a_j denotes the degree of altruism of an individual j (as measured by some standard questionnaires for measuring this feature), $a_j \in \langle 0,1 \rangle$; $\bar{z}_j$ is a vector of measures of other characteristics of an individual and ε_j is a random variable.

The dependence of willingness to pay on the percent of others that are also willing to pay enables reinterpreting this model in terms of threshold model [Ostasiewicz et al. 2008]. Indeed, conditioned that:

$$\frac{\partial WTP_j\left(\bar{z}_j, m, \varepsilon_j\right)}{\partial m} \geq 0$$

and

$$\frac{\partial WTP_j\left(z_j, m, \varepsilon_j\right)}{\partial \varepsilon_j} \geq 0$$

the probability that the willingness to pay will exceed a proposed sum of money t_j, may be expressed in terms of probability that the percent of participants will exceed a certain number, which may be called a threshold:

$$P\left(WTP_j \geq t_j\right) = P\left(m \geq Th_j\left(t_j, \varepsilon_j\right)\right) = \begin{cases} 1 & \text{if} \quad m \geq Th_j\left(t_j, \tilde{\varepsilon}_j\right) \\ 0 & \text{if} \quad m < Th_j\left(t_j, \tilde{\varepsilon}_j\right) \end{cases},$$

where $\tilde{\varepsilon}_j$ denotes a realization of a random variable ε_j.

However, one has to keep in mind that this similarity of willingness to pay model to threshold model (11) holds only in the probabilistic sense and not in all situations. Note, that within WTP model the threshold is a random variable itself and decision of even a single one individual has a probabilistic character. On the other hand, within the simple version of threshold model presented in the previous section, a value of threshold of a given individual is constant and there is no uncertainty. Probabilistic approach may be applied to this threshold model only in the case of a large population and values of thresholds that may be treated as realizations of some random variable. Then, if random variables of different individuals within willingness to pay model are identical, $\varepsilon_j \equiv \varepsilon$, but characteristics of individuals are different, it is possible to treat both kinds of models as equivalent in probabilistic sense. Then dynamical equations (12,14) and stationary state condition (13) may be applied to describe willingness to pay model.

In the case of so-called mean-field approach, that is, such approximation, within which each individual is "replaced" by an "averaged" one (or in the case, when all characteristics of all individuals are in fact identical) there may not be a real dynamics. If all individuals are the same, then all of them should make the same decision – any differences in decisions will be only a result of a randomness, which will cause fluctuations around some state. Nevertheless, the condition for stationary states (13) will still holds and graphical way of obtaining them is still valid. Potential (16) may be still used to examine stability properties of obtained stationary states.

4. An example

Let us propose a simple example. We want to investigate the properties of the model (17) with the willingness to pay function dependent only on the percent of the population which is willing to pay, m, the degree of altruism, a_j, and

income, y_j, as characteristics of an individual. We want to find a function that will have the following properties:

1) the values of the function range from zero to income of an individual, $0 \leq WTP_j \leq y_j$;
2) it is an increasing function of m, a_j and ε_j;
3) for zero altruism and $m \neq 1$ it tends to zero, $\lim_{a_j \to 0} WTP_j = 0$;
4) for maximum altruism it does not depend on percent of population, $WTP_j(m, 1, \varepsilon) \equiv WTP_j(1, \varepsilon)$.

It may be checked, that the following function is fulfilling all the required properties:

$$WTP_j(m, a_j, \varepsilon) = \frac{y_j}{1+\exp\left[\alpha m^{\beta \ln a_j} - \varepsilon\right]} \tag{18}$$

conditioned that $\alpha > 0$, $\beta > 0$. Note, that random variables are assumed here to be identical for all individuals.

1) As $\exp(x) \geq 0$ for $x \in \langle -\infty, +\infty \rangle$ thus:

$$\left[1/\left(1 + \exp\left[\alpha m^{\beta \ln a_j} - \varepsilon\right]\right)\right] \in \langle 0,1 \rangle$$

and:

$$0 \leq WTP_j \leq y_j.$$

2) For $\frac{\partial WTP_j(m,a_j,\varepsilon)}{\partial m} \geq 0$ it is sufficient $\frac{\partial \exp\left[\alpha m^{\beta \ln a_j}\right]}{\partial m} \leq 0$ to be fulfilled. Then:

$$\frac{\partial \exp\left[\alpha m^{\beta \ln a_j}\right]}{\partial m} = \exp\left[\alpha m^{\beta \ln a_j}\right] \alpha \beta \ln a_j m^{\beta \ln a_j - 1}.$$

As $\exp(x) \geq 0$ for any x, for $m \in \langle 0,1 \rangle$: $m^x \geq 0$ for any x, and for $a_j \in \langle 0,1 \rangle$: $\ln a_j \leq 0$, thus for $\alpha > 0$, $\beta > 0$ condition $\frac{\partial \exp\left[\alpha m^{\beta \ln a_j}\right]}{\partial m} \leq 0$ indeed holds, and the function (18) increases with increasing m.

Similarly, for $\frac{\partial WTP_j(m,a_j,\varepsilon)}{\partial a_j} \geq 0$ it is sufficient $\frac{\partial \exp\left[\alpha m^{\beta \ln a_j}\right]}{\partial a_j} \leq 0$ to be fulfilled. As

$$\frac{\partial \exp\left[\alpha m^{\beta \ln a_j}\right]}{\partial a_j} = \alpha \beta \frac{\exp\left[\alpha m^{\beta \ln a_j}\right]}{a_j} m^{\beta \ln a_j} \ln m \leq 0 \text{ for } 0 \leq m \leq 1$$

thus the function (18) increases with increasing a_j.

As for ε it is obvious, that $\exp[-\varepsilon]$ is a decreasing function of ε and thus WTP_j an increasing function of ε.

3) $\lim_{a_j\to 0} WTP_j(m, a_j, \varepsilon) = \lim\limits_{a_j\to 0} \frac{y_j}{1+\exp\left[\alpha m^{\beta \ln a_j} - \varepsilon\right]} = 0$

as $\lim\limits_{a_j\to 0} \ln a_j = -\infty$

and thus $\lim\limits_{a_j\to 0} \exp\left[\alpha m^{\beta \ln a_j} - \varepsilon\right] = \infty$.

4) $WTP_j(m, 1, \varepsilon) = \frac{y_j}{1+\exp\left[\alpha m^{\beta \ln 1} - \varepsilon\right]} = \frac{y_j}{1+\exp[\alpha - \varepsilon]} = WTP_j(1, \varepsilon)$.

Let us examine the stationary states of the system in the mean-field approximation. Within this approximation each individual is replaced by an "averaged individual", that is: $y_j \equiv \bar{y}$, $a_j \equiv \bar{a} = a$. Within this approximation the condition for stationary states reads:

$$m = 1 - F\left(-\ln\frac{1-t}{t} + \alpha m^{\beta \ln a}\right) \tag{19}$$

where F is a cumulative distribution function of a random variable ε and t proposed percent of an income to be paid.

If distribution of a random variable ε is symmetrical, then $F(x) = 1 - F(-x)$ holds and the condition (19) may be rewritten as:

$$m = F\left(\ln\frac{1-t}{t} - \alpha m^{\beta \ln a}\right). \tag{20}$$

Let us assume as a distribution of a random variable a logistic one with standard values of parameters:

$$F(x) = \frac{1}{1+\exp\left(-\frac{x}{\sigma}\right)}. \tag{21}$$

Substituting (21) into (20) one gets a stationary state condition:

$$m = 1 \Big/ \left[1 + \left(\frac{t}{1-t}\right)^{\frac{1}{\sigma}} \exp\left(\frac{\alpha}{\sigma} m^{\beta \ln a}\right)\right]. \tag{22}$$

Let us examine the existence and numbers of stationary states defined by the condition (22). In order to shorten notation let us introduce functions $g(m)$ and $h(m)$ defined as a right-hand-side and left-hand-side, respectively, of condition (22), i.e.:

$$g(m) \equiv 1 \Big/ \left[1 + \left(\frac{t}{1-t}\right)^{\frac{1}{\sigma}} \exp\left(\frac{\alpha}{\sigma} m^{\beta \ln a}\right)\right], \tag{23a}$$

$$h(m) \equiv m. \tag{23b}$$

As $m \in \langle 0,1 \rangle$ we are interested in the crossing of the curve $y = g(m)$ and a line $y = m$ within this range of variable. In the lower limit $g(m)$starts from zero, $g(0) = 0$ and then it is increasing to the value of the upper limit,

$$g(1) = 1 \Big/ \left(1 + \left(\frac{t}{1-t}\right)^{\frac{1}{\sigma}} \exp\left(\frac{\alpha}{\sigma}\right)\right) \in (0, 1\rangle.$$

As $h(0) = 0$ there always exists at least one intersection of curves $g(m)$ and $h(m)$, for $m = 0$. As $h(1) = 1 \geq g(1)$ thus there may exists only this one intersection. On the other hand, the maximum possible number of intersections equals three. The actual number of solutions of stationary state condition depends on the values of parameters α, β and σ, as well as on the proposed fraction of income, t, and degree of altruism, a.

Let us present a simple (and artificial) example of dependence of stationary states on average altruism and proposed percent of income to be paid. Here we will take $\alpha = 1$, $\beta = 18$, $\sigma = 1$.

First, let us fix the proposed percent of income to be paid, $t = 0.01$. For average altruism less than 0.6 there are no stable stationary states apart from zero. For values of altruism greater than 0.6 there appear a second stationary and stable state with $m > 0$ (see Figure 3 for intersections of curves $g(m)$ and $h(m)$ – and thus existence of stationary points – and Figure 4 for corresponding potentials to examine stability of existing stationary points). The stability of this stationary state and the value of final percent of participants that are willing to pay are growing with growing value of a (see Figure 4 for degree of stability – the depth of the well of potential corresponding to the given stationary point, and Figure 5 for the value of m within this possible final state).

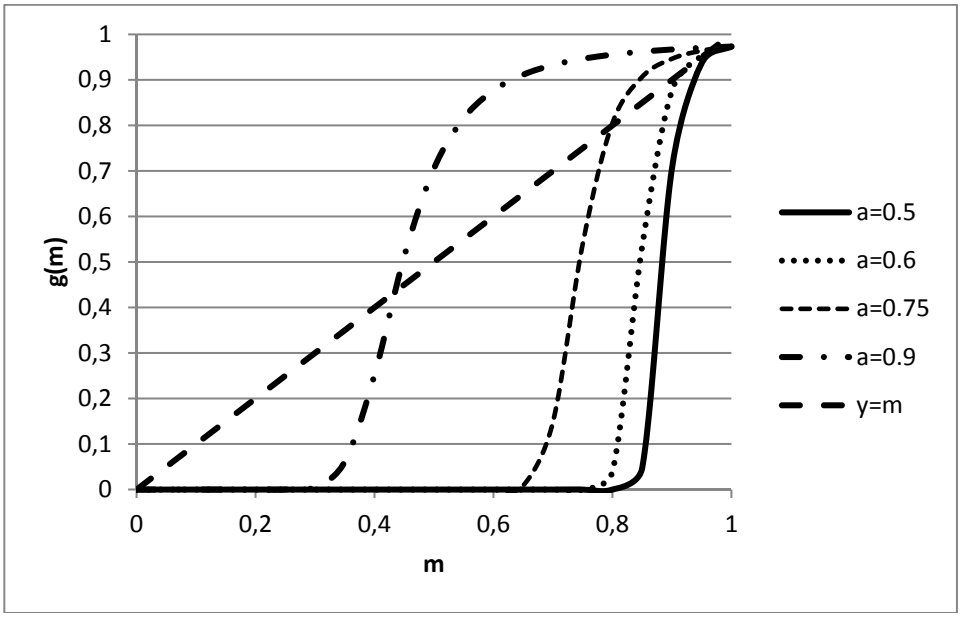


Fig. 3. Plots of $g(m)$ and $h(m)$ for different values of average altruism

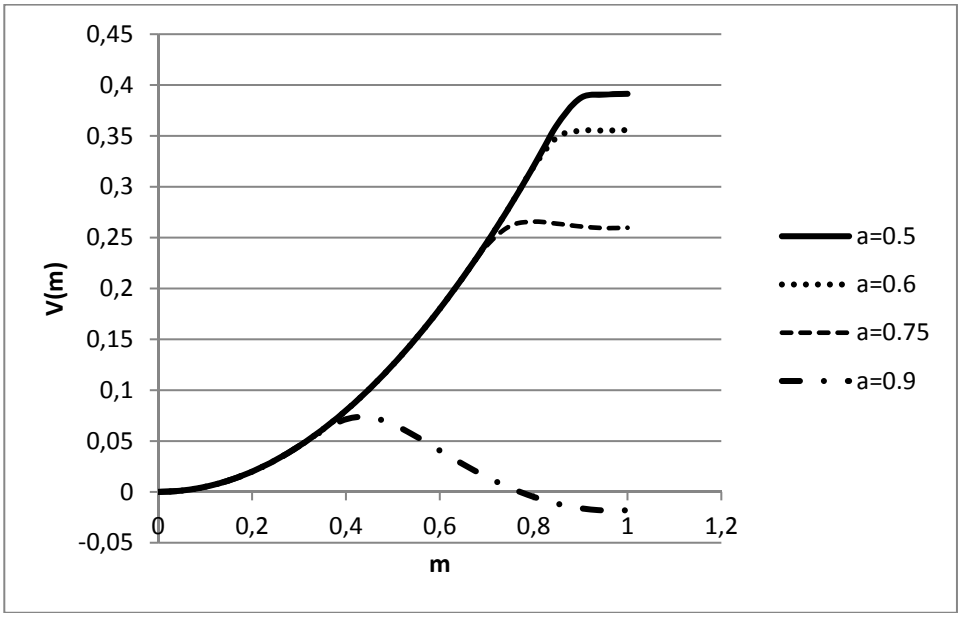


Fig. 4. Potential of the system for different values of average altruism, corresponding to Fig. 3

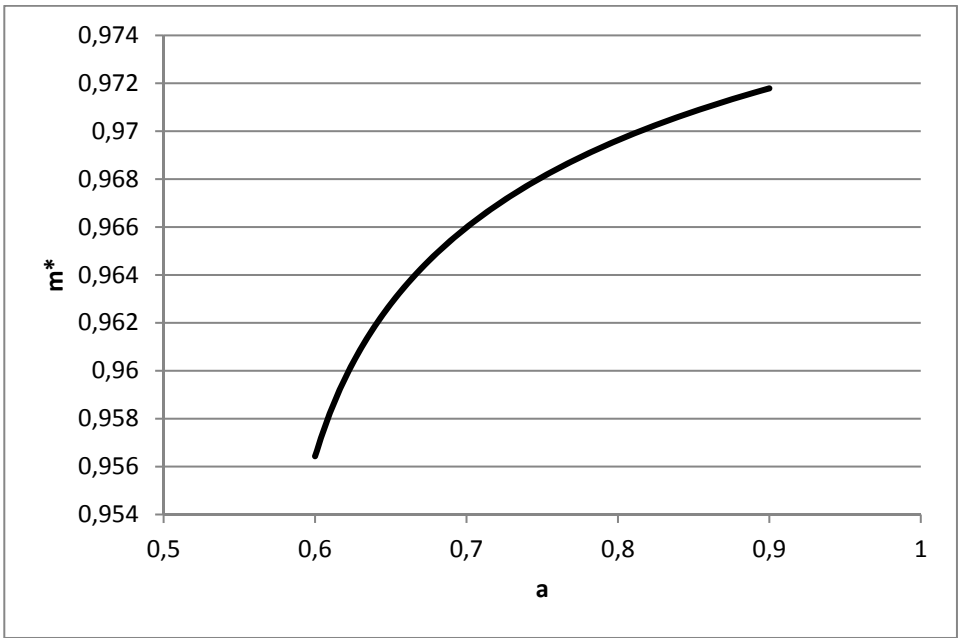


Fig. 5. Dependence of nonzero stable stationary state on average degree of altruism

Now let us fix the value of average altruism on $a = 0.9$ and examine the dependence of stationary states on proposed percent of income to be paid. For $t > 0.056$ there does not exist a stationary state apart from $m^* = 0$. For $t = 0.056$ there appears second stable stationary state, which stability and value of m^* within it increases with decreasing t (see Figure 6 for existence of stationary states, Figure 7 for stability properties of these eventual states and Figure 8 for dependence of value of $m^* > 0$ on proposed payment t).

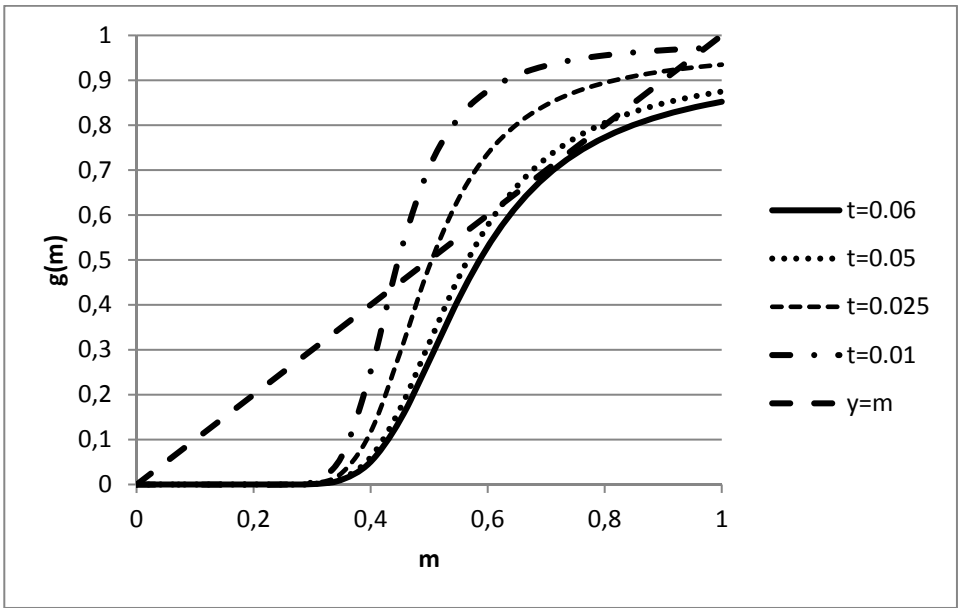


Fig. 6. Plots of $g(m)$ and $h(m)$ for different values of proposed payment

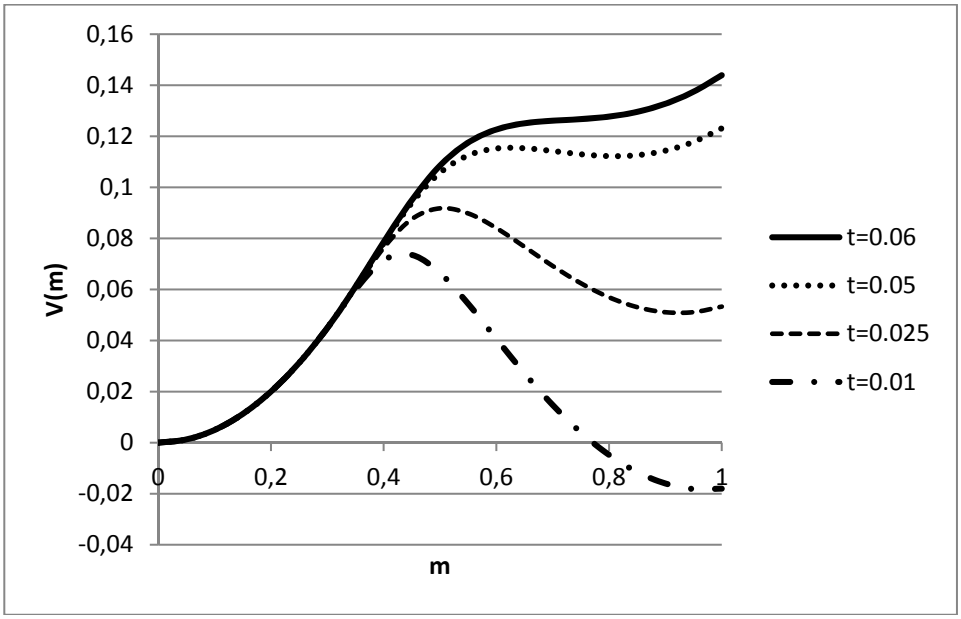


Fig. 7. Potential of the system for different values of proposed payment, corresponding to Fig. 6

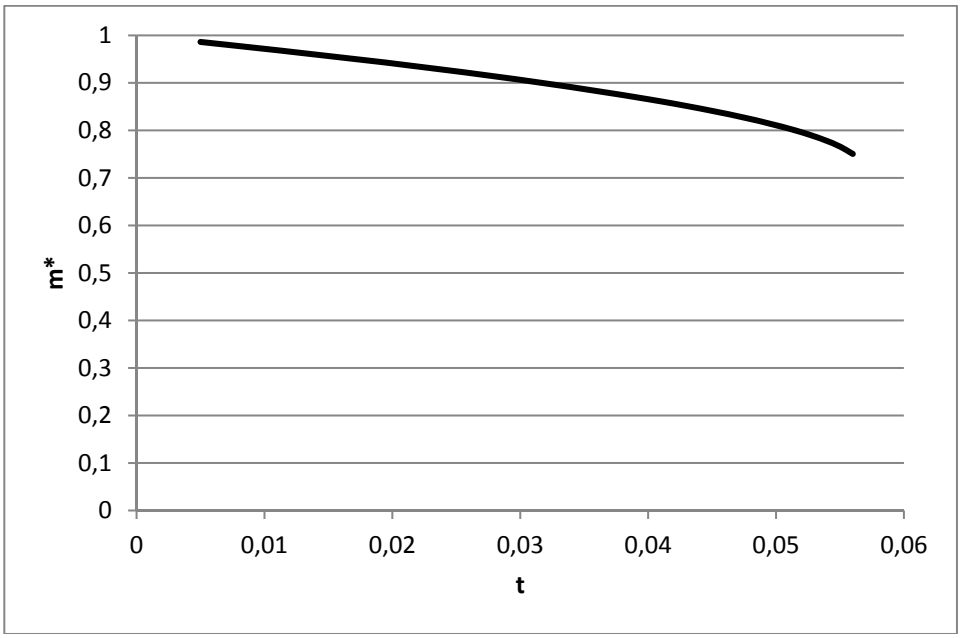


Fig. 8. Dependence of nonzero stable stationary state on proposed payment

The above results are intuitive ones, as one expects, that percent of donators for any public good will increase with increasing altruism and decrease with increasing sum of donation.

To obtain results with any reference to a certain real society one has to estimate the parameters of willingness to pay function and random variable basing on real data collected from empirical studies.

In order to estimate values of parameters α, β and σ one has to construct a likelihood function and then maximize it. In what follows we assume, that random variables are independent and identical for all individuals, and that each individual may be proposed to pay different percent of his/her income than the others individuals. The probability, that an individual j will agree on a proposed payment t_j reads (having in mind that t_j denotes a percent of income):

$$P\left(WTP_j \geq t_j y_j\right) = P\left(\frac{y_j}{1+\exp\left[\alpha m^{\beta \ln a_j} - \varepsilon\right]} \geq t_j y_j\right) =$$

$$= P\left(\varepsilon \geq \ln\frac{t_j}{1-t_j} + \alpha m^{\beta \ln a_j}\right). \tag{24}$$

As ε is a symmetric distribution (a logistic one) thus (24) may be rewritten as:

$$P\left(WTP_j \geq t_j y_j\right) = P\left(\varepsilon \geq \ln\frac{t_j}{1-t_j} + \alpha m^{\beta \ln a_j}\right) = F\left(-\ln\frac{t_j}{1-t_j} - \alpha m^{\beta \ln a_j}\right). \quad (25)$$

Substituting (21) for $F(x)$:

$$P\left(WTP_j \geq t_j y_j\right) = \frac{1}{1+\exp\left[\frac{1}{\sigma}\ln\frac{t_j}{1-t_j} + \frac{\alpha}{\sigma} m^{\beta \ln a_j}\right]}. \quad (26)$$

Thus the likelihood function reads:

$$L(\alpha, \beta, \sigma) = \prod_i \left[\frac{1}{1+\exp\left[\frac{1}{\sigma}\ln\frac{t_j}{1-t_j} + \frac{\alpha}{\sigma} m^{\beta \ln a_j}\right]}\right]^{\omega_j} \left[\frac{1}{1+\exp\left[-\frac{1}{\sigma}\ln\frac{t_j}{1-t_j} - \frac{\alpha}{\sigma} m^{\beta \ln a_j}\right]}\right]^{1-\omega_j}. \quad (27)$$

Maximizing this function basing on collected real data will allow to get a final form of the model, which can be used to predict behavior of a given society stated in front of decision of paying or not for a certain ecological innovation.

Conclusions

As the challenge of slowing down degradation of our natural environment is a very short-time one, we need tools for investigating the state of minds of participants of our community, whose agreement for some steps toward this aim is needed in the democratic society. A standard willingness to pay approach seems insufficient unless it takes into regard the dependence of choices of individuals on the choices of the others. Including percent of participants as a variable of willingness to pay function impose an inner dynamics on the model. That is rather realistic, as ecological thinking seems to spread across populations like fashions (what does not necessarily suggest that believes and attitudes are nothing more than fashions or conformist acting).

In this paper general arguments and approach to including dependence individuals on the others is presented and a simple concrete model is proposed. Its properties has been shortly investigated within a mean-field approach. However, it is possible to go beyond this approximation and get more realistic (and much

more complicated) dynamical equations. Further analysis would reveal detailed dependence of behavior of the model on a specific distribution of degree of altruism and incomes across the population.

Literature

Aronson E. (2003): *The Social Animal*. Worth Publishers, Bedford.

Blume L.E. (1995): *The Statistical Mechanics of Best-Response Strategy Revisions*. "Games and Economic Behavior", No. 11.

Brock W., Durlauf S.N. (2001): *Discrete Choice with Social Interactions*. "Review of Economic Studies", No. 68.

Champ P.A., Bishop R.C., Brown T.C., McCollum D.W. (1997): *Using Donation Mechanisms to Value Nonuse Benefits from Public Goods*. "Journal of Environmental Economics and Management", No. 33.

Granovetter M. (1979): *Threshold Models of Collective Behavior*. "American Journal of Sociology", No. 83.

Haab T.C., McConnell K.E. (2003): *Valuing Environmental and Natural Resources*. Edward Elgar Publishing, Cheltenham.

Hardin G. (1968): *The Tragedy of The Commons*. "Science", No. 162.

Kahneman D., Knetsch J.L. (1992): *Valuing Public Goods: The Purchase of Moral Satisfaction*. "Journal of Environmental Economics and Management", No. 22.

Ostasiewicz K., Tyc M.H., Radosz A., Magnuszewski P., Goliczewski P., Hetman P., Sendzimir J. (2008): *Multistability of Impact, Utility and Threshold Concepts of Binary Choice Models.* "Physica A", No. 387.

Ostrom E. (2009): *Beyond Markets and States: Polycentric Governance of Complex Economic Systems*. Nobel Prize Lecture.

United Nations (1987): *Report of the World Commission on Environment and Development*. General Assembly Resolution 42/187, 11 December.

Wiser R.H. (2007): *Using Contingent Valuation to Explore Willingness to Pay for Renewable Energy: A Comparison of Collective and Voluntary Payment Vehicles*. "Ecological Economics", No. 62.

MODELOWANIE GOTOWOŚCI DO PŁACENIA NA RZECZ ZRÓWNOWAŻONEGO ROZWOJU

Streszczenie

Postępująca degradacja środowiska naturalnego jest palącym problem współczesności. Jednym z elementów koniecznych do jego rozwiązania jest ekologiczna świadomość obywateli. Z tego względu, istotne jest badanie postaw ludzi wobec dobrowolnego

ponoszenia zwiększonych kosztów działań zachowujących środowisko naturalne w dobrym stanie. Standardowe modele skłonności do płacenia na rzecz ekologii wydają się niekompletne, gdyż nie uwzględniają zależności postaw jednostek od postaw ich otoczenia. Dbałość o ekologię jest bowiem nastawieniem szerzącym się w społecznościach na podobieństwo innych wzorców kulturowych i zachowań społecznych. Włączenie mechanizmu naśladownictwa do modelu nadaje mu zatem automatycznie charakter dynamiczny.

W pracy zaprezentowano ogólny model gotowości do płacenia, uwzględniający zależność wyborów jednostek od wyborów innych osób. Następnie, jako przykład, został zaproponowany i poddany analizie bardziej szczegółowy model. Omówiono jego właściwości zarówno w przybliżeniu średniego pola, jak i w ujęciu dynamicznym. Pokazano zależność rezultatów od rozkładu osobniczego stopnia altruizmu oraz dochodów.

Janusz L. Wywiał
Uniwersytet Ekonomiczny w Katowicach

ON LIMIT DISTRIBUTION OF HORVITZ-THOMPSON STATISTIC UNDER POISSON SAMPLING DESIGN

Introduction

Let U_N be a fixed population of the size N, so, $N = 2, 3, ...$ The elements of the population are identified. So, the population can be represented by the set: $U_N = \{1, ..., N\}$. The observation of a variable under study is denoted by $y_{k,N}$, $k = 1, ..., N$, $N = 2, 3, ...$ So, the vector $\mathbf{y}_N = [y_{1,N}\ y_{2,N} \ldots y_{N,N}]$ is attached to the set U_N. Particularly, when we assume that $U_N \subset U_{N+1}$ then the observations of a variable in the population can be represented more simply by the vector: $\mathbf{y}_{N+1} = [\mathbf{y}_N y_{N+1}]$ where $\mathbf{y}_N = [y_1, y_2 \ldots y_N]$. A more particular case is as follows. Let $(y_i, w_i N_t)$ means that the value y_i is $w_i N_t$ times replicated in the population, where $\Sigma_{i=1}^{k} w_i = 1$ and $0 < w_1 < 1$ for $i = 1, 2, ..., k$, $t = 1, 2, ...$, and the vector $\mathbf{y}_N = [y_k\ y_k, ..., y_k]$ is fixed. The size N_t of the population U_{N_t} is determined in such a way that $w_i N_t$ is an integer for all $i = 1, ..., k$. So,

$$\mathbf{y}_{N_t} = [(y_1, w_1 N_t)(y_2, w_2 N_t) \ldots (y_k, w_k N_t)].$$

We assume that the all elements of the population can be selected for the sample with different probabilities. A k-th population element, $k \in U_N$, be selected for the sample with the inclusion probability $0 < \pi_{k,N} < 1$, $k = 1, ..., N$. More precisely, let $S_N = [S_{1,N} \ldots S_{N,N}]$ be the vector of independent binary random variables and

$$P(S_{k,N}=1)=\pi_{k,N}=1-P(S_{k,N}=0) \tag{1}$$

So, $s_N=[s_{1,N}\ldots s_{N,N}]$ is the realization of the random sample S.

The probability distribution of the random sample S is known as Poisson sampling design [see, e.g. Tille 2006]:

$$P(S_N=s_N)=\prod_{k=1}^{N}\pi_{k,N}^{s_{k,N}}(1-\pi_{k,N})^{1-s_{k,N}}.$$

The population total $\tilde{y}=\sum_{k\in U_N}y_{k,N}$ can be estimated on the basis of the Horvitz-Thompson statistic [1952]:

$$y_{HTS_N}=\sum_{k\in U_N}\frac{y_{k,N}S_{k,N}}{\pi_{k,N}}.$$

It is well known that $E(y_{HTS_N})=\tilde{y}$ if $\pi_{k,N}>0$ for all $k=1,...,N$. Because of $P(S_{k,N}=1,S_{h,N}=1)=\pi_{k,N}\pi_{h,N}$ for all $k=1,...,N$, $h=1,...,N$ and $k\neq h$ the variance of the statistic y_{HTS_N} is:

$$V\left(y_{HTS_N}\right)=\sum_{k\in U_N}\frac{y_{k,N}^2(1-\pi_{k,N})}{\pi_{k,N}}. \tag{2}$$

Its unbiased estimator is:

$$V\left(y_{HTS_N}\right)=\sum_{k\in U_N}\frac{y_{k,N}^2(1-\pi_{k,N})S_{k,N}}{\pi_{k,N}^2}. \tag{3}$$

Let

$$b_{3,N}=\frac{1}{N}\sum_{k\in U_N}\left|y_{k,N}\right|^3 \text{ and } v_{2,N}=\frac{1}{N}\sum_{k\in U_N}y_{k,N}^2 \quad v_{4,N}=\frac{1}{N}\sum_{k\in U_N}y_{k,N}^4.$$

The original version and the proof of the Lapunov's [2001] theorem, which is slightly less general, can be found in the monograph by Fisz [1963]. On the basis of the books by Billingsley [2009] or Jakubowski and Sztencel [2004], the following more general version of the theorem is presented.

Theorem 1. Let $Z_{k,N}$, $k = 1, ..., N$, $N = 1, 2, ...$ be a sequence of independent random variables and for some $\delta > 0$

$$\beta_N = \frac{B_N^{\delta}}{C_N^{2+\delta}} \to 0 \text{ if } N \to \infty \tag{4}$$

where:

$$B_N^{\delta} = \sum_{k=1}^{N} E\left|Z_{k,N} - E(Z_{k,N})\right|^{2+\delta}, \; C_N = \sqrt{\sum_{k=1}^{N} V(Z_{k,N})}. \tag{5}$$

Under these Lapunov's conditions, the random variable:

$$Z_N = \frac{\sum_{k=1}^{N}\left(Z_{k,N} - E(Z_{k,N})\right)}{C_N}$$

converges in distribution to the normal standard distribution if $N \to \infty$.

Hájek [1964] considered a limit distribution for the following statistic

$$\overline{H}_S = y_{HTS_N} - \frac{r}{N}\sum_{k=1}^{N}\left(\pi_k - S_k\right)$$

where:

$$r = \frac{\sum_{k=1}^{N} y_{k,N}\left(1-\pi_k\right)}{\sum_{k=1}^{N} \pi_k\left(1-\pi_k\right)}.$$

He proved that the probability distribution of the statistic $\overline{H}_S$ tends to the normal distribution because it fulfils the well known Lindeberg condition.

In the next section, the limit theorem for the estimator y_{HTS_N} will be considered.

1. Limit theorem

Firstly, let us formulate the following statistics and the theorem.

$$T_N = \frac{y_{HTS_N} - \widetilde{y}}{\sqrt{V(y_{HTS_N})}},$$

$$\hat{T}_N = \frac{y_{HTS_N} - \widetilde{y}}{\sqrt{V_{S_N}(y_{HTS_N})}}. \tag{6}$$

We say that $\pi_{k,N} = O(N^{-\alpha})$ if for all $0 \le \alpha < 1$ there exists such a_1 and a_0 that $0 \le a_1 \le a_0 < 1$ and

$$0 < a_1 N^{-\alpha} \le \max_{N=1,2,\ldots}\left\{\max_{k=1,\ldots,N}\left\{\pi_{k,N}\right\}\right\} \le a_0 N^{-\alpha} < 1 \tag{7}$$

or

$$0 < a_1 \le \frac{\max_{N=1,2,\ldots}\left\{\max_{k=1,\ldots,N}\left\{\pi_{k,N}\right\}\right\}}{N^{\alpha}} \le a_0 < 1$$

Particularly, if $\alpha = 0$,

$$0 < a_1 \le \max_{N=1,2,\ldots}\left\{\max_{k=1,\ldots,N}\left\{\pi_{k,N}\right\}\right\} \le a_0 < 1$$

Moreover, $\pi_{k,N}^{-1} = O(N^{\alpha})$ because for all $0 \le \alpha < 1$ there exists such $1 < c_1 \le \frac{1}{a_0} < \frac{1}{a_1} \le c_0$ that

$$1 < c_1 N^{\alpha} \le \max_{N=1,2,\ldots}\left\{\max_{k=1,\ldots,N}\left\{\frac{1}{\pi_{k,N}}\right\}\right\} \le c_0 N^{\alpha} \tag{8}$$

$\pi_{k,N}^{-1} - 1 = O(N^{\alpha})$ because for all $0 \le \alpha < 1$ there exists such $0 < d_1 \le c_1 - N^{-\alpha} < c_0 - N^{-\alpha} \le d_0$ that

$$0 < d_1 N^{\alpha} \le \max_{N=1,2,...} \left\{ \max_{k=1,...,N} \left\{ \frac{1}{\pi_{k,N}} \right\} \right\} \le d_0 N^{\alpha} \tag{9}$$

$O(N^{\alpha})O(N^{\gamma}) = O(N^{\alpha+\gamma})$ because for all $\alpha \ge 0$ and $\gamma \ge 0$ from the inequalities $0 < d_1 \le N^{\alpha} \le d_0$ and $0 < g_1 \le N^{\gamma} \le g_0$ results the following one

$$0 < e_1 \le d_1 g_1 \le N^{\alpha+\gamma} \le d_0 g_0 \le e_0 \tag{10}$$

Finally, $O(N^{\alpha})O(N^{-\gamma}) = O(N^{\alpha-\gamma})$ because for all $\alpha \ge 0$ and $\gamma \ge 0$ from the inequalities $0 < d_1 \le N^{\alpha} \le d_0$ and $0 < g_1 \le N^{\gamma} \le g_0$ results the following one

$$0 < l_1 \le \frac{d_1}{g_0} \le N^{\alpha-\gamma} \le \frac{d_0}{g_1} \le l_0 \tag{11}$$

Theorem 2. Let $\{\pi_{k,N}\} = O(N^{-\alpha})$ for all $0 \le \alpha < 1$ and $0 < v_0 \le v_{2,N} \le v_2 < \infty$ and $b_{3,N} \le b_3 < \infty$ for $N = 1, 2, ...$ When $N \to \infty$ then $T_N \xrightarrow[d]{} T \sim N(0,1)$. When additionally $v_{4,N} \le v_4 < \infty$ then $\hat{T}_N \xrightarrow[d]{} T \sim N(0,1)$.

Proof: On the basis of the theorem 1, it is sufficient to assume that $\delta = 1$. The Horvitz-Thompson statistic can be rewritten in the following way.

$$y_{HTS_N} = \sum_{k=1}^{N} Z_{k,N}$$

where:

$$Z_{k,N} = \frac{y_{k,N} S_{k,N}}{\pi_{k,N}} \quad k = 1, ..., N$$

On the basis of the expression (1), we have:

$$P(Z_{k,N} = z_{k,N}) = \begin{cases} \pi_{k,N} & \text{if } z_{k,N} = \dfrac{y_{k,N}}{\pi_{k,N}}, \\ 1 - \pi_{k,N} & \text{if } z_{k,N} = 0 \end{cases} \tag{12}$$

and

$$E(Z_{k,N}) = y_{k,N},\ E(Z_{k,N}^2) = \frac{y_{k,N}^2}{\pi_{k,N}},\ V(Z_{k,N}) = y_{k,N}^2\left(\frac{1}{\pi_{k,N}} - 1\right),\ k = 1, ..., N$$

$$E\left|Z_{k,N} - E(Z_{k,N})\right|^3 = \frac{\left|y_{k,N}\right|^3}{\pi_{k,N}^3} E\left|S_{k,N} - \pi_{k,N}\right|^3 = \frac{\left|y_{k,N}\right|^3}{\pi_{k,N}^3}\left(\left(1-\pi_{k,N}\right)^3 \pi_{k,N} +\right.$$

$$\left. + \pi_{k,N}^3\left(1-\pi_{k,N}\right)\right) = \frac{\left|y_{k,N}\right|^3}{\pi_{k,N}^2}\left(1-\pi_{k,N}\right)\left(\left(1-\pi_{k,N}\right)^2 + \pi_{k,N}^2\right) \le \frac{\left|y_{k,N}\right|^3}{\pi_{k,N}^2}\left(1-\pi_{k,N}\right).$$

This and the expression (5) for γ=1 lead to the following.

$$B_N = \sum_{k=1}^{N} \frac{\left|y_{k,N}\right|^3}{\pi_{k,N}^2}\left(1-\pi_{k,N}\right)\left(\left(1-\pi_{k,N}\right)^2 + \pi_{k,N}^2\right)$$

,\

$$C_N = \sqrt{\sum_{k=1}^{N} \frac{y_{k,N}^2\left(1-\pi_{k,N}\right)}{\pi_{k,N}}} = \sqrt{V(y_{HTS_N})}$$

,

On the basis of the expressions (7)-(11) we have:

$$\beta_N = \frac{B_N}{C_N^3} = \frac{\sum_{k=1}^{N} \frac{\left|y_{k,N}\right|^3}{\pi_{k,N}^2}(1-\pi_{k,N})\left(\left(1-\pi_{k,N}\right)^2 + \pi_{k,N}^2\right)}{\left(\sum_{k=1}^{N} y_{k,N}^2\left(\frac{1}{\pi_{k,N}} - 1\right)\right)^{\frac{3}{2}}} \le$$

$$\le \frac{\sum_{k=1}^{N}\left|y_{k,N}\right|^3 \frac{1}{\pi_{k,N}}\left(\frac{1}{\pi_{k,N}} - 1\right)}{\left(\sum_{k=1}^{N} y_{k,N}^2\left(\frac{1}{\pi_{k,N}} - 1\right)\right)^{\frac{3}{2}}} = \frac{\sum_{k=1}^{N}\left|y_{k,N}\right|^3 O\left(N^{\alpha}\right)\left(O\left(N^{\alpha}\right) - 1\right)}{\left(\sum_{k=1}^{N} y_{k,N}^2\left(O\left(N^{\alpha}\right) - 1\right)\right)^{\frac{3}{2}}} =$$

$$= \frac{O\left(N^{2\alpha}\right)\sum_{k=1}^{N}\left|y_{k,N}\right|^3}{\left(O\left(N^{\alpha}\right)\sum_{k=1}^{N} y_{k,N}^2\right)^{\frac{3}{2}}} = \frac{O\left(N^{2\alpha+1}\right)b_{3,N}}{O\left(N^{3(\alpha+1)/2}\right)v_{2,N}^{3/2}} \le \frac{O\left(N^{2\alpha+1}\right)b_3}{O\left(N^{3(\alpha+1)/2}\right)v_0^{3/2}} = O\left(N^{(\alpha-1)/2}\right).$$

It is easy to show that $\beta \to 0$, when $N \to 0$ to and $0 \le \alpha \le \alpha_1 < 1$. This and the theorem 1 lead to the conclusion that $T_N \to T \sim N(0,1)$.

In order to prove the second part of the theorem, we firstly show that $R_N = V_{S_N}(y_{HTS_N}) / V(y_{HTS_N})$ converge in probability to 1. The expression (3) leads to the variance of the sample variance of Horvitz-Thompson statistic:

$$V(V_{S_N}(y_{HTS_N})) = V\left(\sum_{k=1}^{N} \frac{y_k^2(1-\pi_{k,N})S_{k,N}}{\pi_k^2(N)}\right) =$$

$$= \sum_{k=1} \frac{y_k^4(1-\pi_{k,N})^2}{\pi_k^4(N)} V(S_{k,N}) = \sum_{k\in U} y_k^4 \left(\frac{1}{\pi_{k,N}} - 1\right)^3 = \sum_{k\in U} y_k^4 \left(O(N^\alpha) - 1\right)^3 =$$

$$= O\left(N^{(3\alpha)}\right) \sum_{k\in U} y_k^4 = O(N^{3\alpha+1}) v_{4,N}.$$

Hence, on the basis of the expression (2) we have

$$V(R_N) = \frac{V(V_{S_N}(y_{HTS_N}))}{V^2(y_{HTS_N})} = \frac{\sum_{k\in U_N} y_k^4 \left(\frac{1}{\pi_{k,N}} - 1\right)^3}{\left(\sum_{k\in U_N} y_k^2 \left(\frac{1}{\pi_{k,N}} - 1\right)\right)^2} = \frac{\sum_{k\in U} y_k^4 \left(O(N^\alpha) - 1\right)^3}{\left(\sum_{k\in U_N} y_k^2 \left(O(N^\alpha) - 1\right)\right)^2} =$$

$$= \frac{O(N^{3\alpha+1}) v_{4,N}}{O(N^{2(\alpha+1)}) v_{2,N}^2} \le \frac{O(N^{3\alpha+1}) v_4}{O(N^{2(\alpha+1)}) v_0^2} = O(N^{\alpha-1})$$

Hence, $V(R_N) = O(N^{\alpha-1}) \to 0$ when $N \to \infty$ and $0 \le \alpha < 1$. So, this and the well known Tchebyshev's inequality lead to the conclusion that that $R_N = V_{S_N}(y_{HTS_N}) / V(y_{HTS_N})$ converges in probability to 1 (in short: $R_N \xrightarrow[p]{} 1$ if $v_0 > 0$, $v_4 < \infty$ and $N \to \infty$. Let us note that

$$\hat{U}_N = \frac{U_N}{R_N}.$$

Hence, when $N \to \infty$ then $T_N \xrightarrow[d]{} T \sim N(0,1)$ and $R_N \xrightarrow[p]{} 1$. So, this and the well known Slucky's lemma, see e.g. Van der Vaart [2007], let us conclude that $\hat{T}_N \xrightarrow[d]{} T \sim N(0,1)$. So, the proof of Theorem 2 has been completed.

2. Applications

The Poisson sampling design is frequently used to model non-response. In this case, $\pi_{k,N}$ is the probability that a k-th population element will respond. The Poisson sampling design can be treated as a model of the Internet research. In this case, $\pi_{k,N}$ is the probability that a k-th Internet user will respond. Moreover, the Poisson sampling design can be considered in an audit sampling. Let us note that, in the cases mentioned the probabilities $\pi_{k,N}$, $k = 1, ..., N$, $N = 2, ...$ are usually defined as follows.

$$\pi_{k,N} = \mathrm{n}\frac{\mathrm{x_{k,N}}}{\sum_{\mathrm{i=1}}^{\mathrm{N}} x_{i,N}}$$

where n is the expected sample size and x_k is a value of a positive auxiliary variable x observed in all the population. Let us assume that $0 < \alpha \le x_{k,N} \le b < \infty$ and $n = wN$ for all $k = 1, ..., N$, $N = 2, ...$ where $0 < w < \frac{a}{b} \le 1$. So, in this case, the first assumption of the Theorem 2 is fulfilled, because

$$0 < a_0 = w\frac{a}{b} = \frac{na}{bN} \le \pi_{k,N} \le \frac{nb}{aN} = w\frac{b}{a} = a_1 < 1$$

Theorem 2 lets us construct the confidence interval for the mean value estimated by means of the Poisson-Horvitz-Thompson strategy. Let γ be the confidence level and let u_γ be such a quantile that $\phi(u_\gamma) = \frac{1+\gamma}{2}$ where $\phi(u)$ is the distribution function of the standard normal variable. When N is sufficient-

ly large, the confidence interval for the population mean is determined by the expression:

$$P\left(y_{HTS_N} - u_\gamma \sqrt{V_{S_N}(y_{HTS_N})} < \tilde{y} < y_{HTS_N} + u_\gamma \sqrt{V_{S_N}(y_{HTS_N})}\right) = \gamma .$$

It is possible to test the hypothesis on the population mean. The hypothesis $H_0 : \tilde{y} = \tilde{y}_0$ can be tested on the basis of the statistic defined by the expression (6) when N is sufficiently large.

Finally, let us note that if the Lapunov's condition is fulfilled, the Lindeberg's condition is fulfilled [see, e.g. Billingsley 2009], too. Hence, if the assumptions of the above theorem 2 are fulfilled, the assumptions of Hájek's theorem are fulfilled, too. Moreover, it seems that in our case the assumptions of the theorem 2 are verified more simply than the Lindeberg's ones.

Acknowledgements

The research was supported by the grant number N N111 434137 from the Ministry of Science and Higher Education.

Literature

Billingsley P. (2009): *Prawdopodobieństwo i miara* (*Probability and Measure)*. Wydawnictwo Naukowe PWN, Warszawa.

Fisz M. (1963): *Probability Theory and Mathematical Statistics*. Wiley and Sons, New York.

Hájek J. (1964): *Asymptotic Theory of Rejective Sampling with Varying Probabilities from a Finite Population*. "The Annals of Mathematical Statistics", No. 35, 4.

Horvitz D.G., Thompson D.J. (1952): *A Generalization of the Sampling without Replacement from Finite Universe*. "Journal of the American Statistical Association", No. 47.

Jakubowski J., Sztencel R. (2004): *Wstęp do teorii prawdopodobieństwa* (*Introduction to Probability Theory*). SCRIPT, Warszawa.

Lapunov A.M. (1901): *Nouvell forme du theorem sur la limite de probabilite.* „Mem. Acad. Sci. St. Pétersburg", No. 12.

Tillé Y. (2006): *Sampling Algorithms*. Springer, New York.

Van der Vaart A.W. (2007): *Asymptotic Statistic*. Cambridge University Press, Cambridge, New York, Melbourne, Madrit, Cape Town, Singapore, Sao Paulo.

O ROZKŁADZIE GRANICZNYM STATYSTYKI HORVITZA-THOMPSONA DLA PRÓBY DOBIERANEJ ZGODNIE Z PLANEM LOSOWANIA POISSONA

Streszczenie

W pracy na podstawie znanego twierdzenia centralnego Lapunowa jest wyprowadzany rozkład graniczny prawdopodobieństwa znanej statystyki Horvitza-Thompsona (HT). Okazało się, że jeśli określane przez plan losowania Poissona prawdopodobieństwa wylosowania do próby poszczególnych elementów populacji spełniają pewne założenia oraz rozmiar populacji rośnie nieograniczenie, to rozkład standardowej postaci statystyki HT zmierza do rozkładu normalnego standardowego. Taki sam wynik otrzymano przy dodatkowym założeniu narzuconym na prawdopodobieństwa wylosowania elementów populacji do próby, gdy w standardowej postaci statystyki HT jej odchylenie standardowe zastąpimy przez pierwiastek z nieobciążonego estymatora tej wariancji.

Rezultaty pracy znajdują zastosowania np. w pewnych typach badań ankietowych, a w szczególności internetowych, wykorzystujących wnioskowanie statystyczne, czyli estymację przedziałową lub testowanie hipotez statystycznych.

Wojciech Gamrot
Uniwersytet Ekonomiczny w Katowicach

ON A CLASS OF ESTIMATORS FOR A RECIPROCAL OF BERNOULLI PARAMETER

Introduction

The estimation of a reciprocal for the probability of an event is an issue of broad practical interest. Such a problem arises e.g. in empirical Horvitz--Thompson estimation for complex sampling designs considered by Fattorini [2006], Thompson and Wu [2008] and Fattorini [2009]. When sampling weights, defined as reciprocals of first order inclusion probabilities are too complex to compute exactly they are instead replaced with estimates evaluated in a simulation study. For this purpose, it is desired that the estimator of inverse probability always takes finite values, so that the Horvitz-Thompson statistic is also always finite. This effect may be achieved in many ways including the use of restricted maximum likelihood principle, the truncation of the binomial distribution considered by Stephan [1946], Rempała and Szekely [1998] and Marciniak and Wesołowski [1999] or bayesian estimation considered e.g. by Berry [1989], Marchand and MacGibbon [2000] and Marchand et al. [2005]. Another approach to this problem, proposed by Fattorini [2006] utilizes the reciprocal of the sampling fraction with a certain constant adjustment that prevents the denominator from reaching zero. In this paper, the original estimator of Fattorini is generalized by admitting varying values of the adjustment constant. This leads to a broader class of estimators. Expressions for the bias for some of these estimators are provided. The problem of setting adjustment constant is then considered.

1. Estimator of Fattorini

Let X denote the number of successes in n Bernoulli trials with success probability equal to p an let q = 1-p. Hence X has a binomial distribution with expectation E(X) = np and variance V(X) = np(1-p). Fattorini [2006] considers a consistent estimator of the probability p in the form:

$$\hat{p}_1 = \frac{X+1}{n+1}.$$

This leads to the estimator of the reciprocal:

$$\hat{d}_1 = \frac{n+1}{X+1}.$$

Fattorini [2006] derived exact formula for its bias:

$$B(\hat{d}_1) = \frac{-q^{n+1}}{p} \tag{1}$$

and by developing this estimator into factorial series he also gave an upper bound for its variance:

$$V(\hat{d}_1) \leq \frac{5}{(n+2)p^3}.$$

The bias of the estimator $\hat{d}_1$ may be intuitively viewed as a superposition of biases resulting from two different sources. First, introduction of adjustment constant into sampling fraction $\hat{p}_1$ to make it strictly positive makes it also positively biased. When $\hat{p}_1$ resides in a denominator of $\hat{d}_1$, this bias for the whole reciprocal becomes negative. On the other hand, from Jensen's inequality we have E(1/X) > 1/E(X) which means that taking inverse of any random variable is by itself a transformation introducing positive bias. Hence, it is reasonable that

at least to some extent these two components cancel each other. Meanwhile, the formula (1) is in general negative for $p \in (0,1)$, which means that they do not reduce to zero. This also suggests, that by choosing an adjustment constant different from unity one might perhaps reduce the bias totally. In the sequel such a possibility is investigated.

2. A more general estimator

We will now consider a more general statistic in the form:

$$\hat{p}_c = \frac{X + c}{n + c},$$

where c is some non-negative constant. In general, it is not required for c to be an integer, and hence we consider a class of estimators for $c \in \langle 0, +\infty \rangle$. It is easy to show that:

$$E(\hat{p}_c) = \frac{np + c}{n + c}.$$

And hence its bias is:

$$B(\hat{p}_c) = E\left(\frac{X + c}{n + c}\right) - p = \frac{cq}{n + c}.$$

The bias is positive and tends to zero when $n \to \infty$. Meanwhile, the variance is given by:

$$V(\hat{p}_c) = \frac{npq}{(n + c)^2},$$

which means that $\hat{p}_c$ is consistent. This lets us construct a consistent estimator for d in the form:

$$\hat{d}_c = \frac{n + c}{X + c}.$$

OBVIOUSLY, the statistic of Fattorini is a special case of the above estimator for c = 1. The statistic $\hat{p}_c$ takes discrete values from the interval [c / (n + c),1] and consequently, $\hat{d}_c$ takes values from [1,(n + c) / c]. For integer c > 0 and 0 < p < 1 we have (see Appendix 1 for a proof):

$$E\left(\frac{1}{X+c}\right)=\frac{1}{p^c}\frac{n!}{(n+c)!}\left(E_{n+c}\left(\prod_{r=1}^{c-1}(X-r)\right)+(-1)^c(c-1)!q^{n+c}\right), \qquad (2)$$

where the symbol $E_{n+c}(\ )$ represents expectation computed with respect to the binomial distribution B(n+c,p), as opposed to the symbol E( ) representing expectation calculated with respect to the B(n,p) binomial distribution. This notation assumes that the multiplication of elements in the empty set for c = 1 equals 1. Hence, for c = 1,2,3,4,... the expectation (2) depends on:

$E_{n+c}(1) = 1$

$E_{n+c}(X-1) = E_{n+c}(X) - 1$

$E_{n+c}((X-1)(X-2)) = E_{n+c}(X^2) - 3E_{n+c}(X) + 2$

$E_{n+c}((X-1)(X-2)(X-3)) = E_{n+c}(X^3) - 6E_{n+c}(X^2) + 11E_{n+c}(X) - 6$

.....

and hence on raw moments for the B(n + c,p) distribution which are easy to calculate from the moment generating function for B(n,p) that takes the well-known form:

$$M(t) = (1 - p + pe^t)^n.$$

Hence raw moments of the order r = 1,2,3,... for B(n,p) are computed as:

$$E(X^r) = \left.\frac{\partial^r M(t)}{\partial t^r}\right|_{t=0}.$$

This leads to:

$$E(X)=np$$

$$E(X^2)= np\,(q+np)$$

$$E(X^3) = np(1-3p+3np+2p^2-3np^2+n^2p^2)$$

and so on. Substituting (n + c) instead of n in above formulas one gets raw moments of the B(n + c,p) distribution. Hence for c = 1,2,3,... the expectations of the estimator $\hat{d}_c$ may be expressed as (see Appendix 2):

$$E(\hat{d}_1) = \frac{1-q^{n+1}}{p} \tag{3}$$

$$E(\hat{d}_2) = \frac{(n+2)p-1+q^{n+2}}{(n+1)p^2} \tag{4}$$

$$E(\hat{d}_3) = \frac{p(n+3)\,(np+2p-2)+2-2q^{n+3}}{(n+1)(n+2)p^3} \tag{5}$$

$$E(\hat{d}_4) = \frac{(n+4)p[6+3p+(n+4)p(np+p-3)+2p^2]-6+6q^{n+4}}{p^4(n+1)(n+2)(n+3)} \tag{6}$$

and so on. Obviously the formula (3) corresponding to $E(\hat{d}_1)$ is equivalent to the result of Fattorini [2006] given by (1). The other three are not. It is easy to verify that for c = 1,2,3,4 we have $\lim_{n\to\infty} E(\hat{d}_c) = d$ which confirms that for c = 1,2,3,4 the estimator $\hat{d}_c$ is asymptotically unbiased for d. Let us use these formulas to calculate the exact bias:

$$B(\hat{d}_c) = E(\hat{d}_c) - d\,,$$

for c = 1,2,3,4, for n = 1,...,2000 and for p = 0.01, 0.05, 0.1, 0.5. The results of calculations are shown on Figure 1.

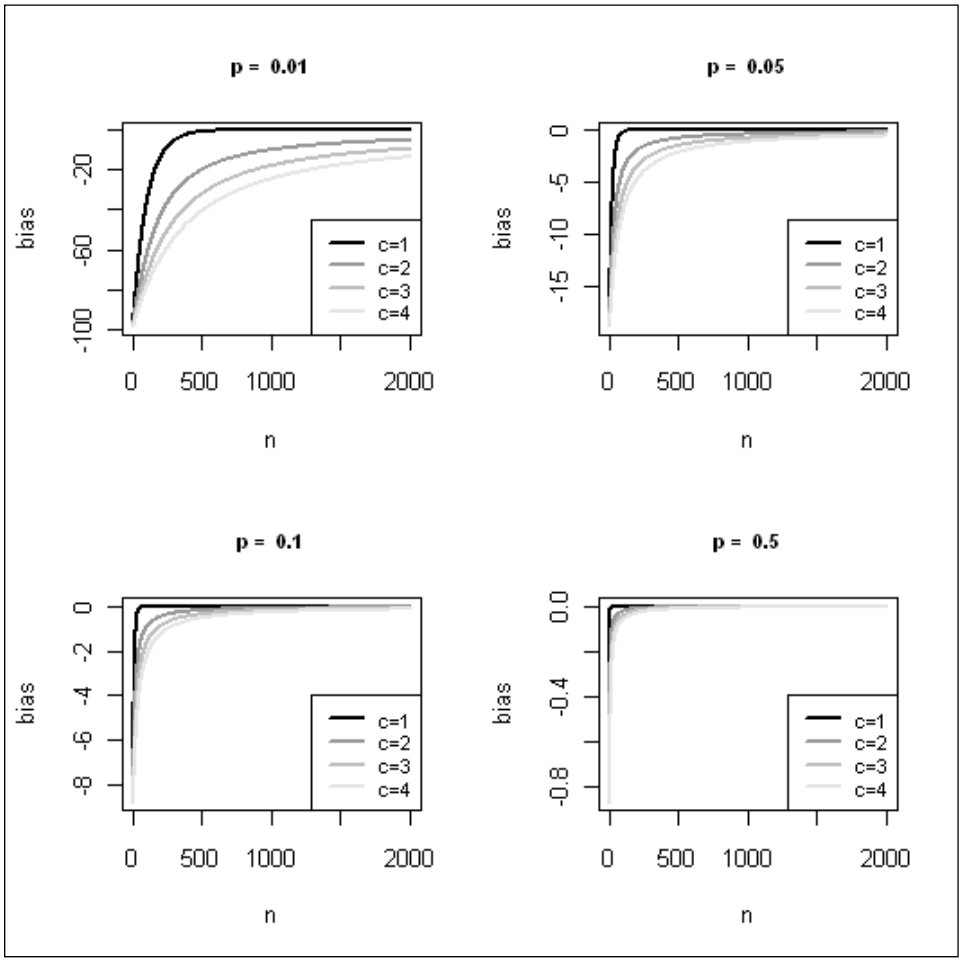


Fig. 1. Exact biases for $\hat{d}_c$, c = 1,2,3,4 and n = 1,...,2000.

For all presented values of c, n and p the bias of $\hat{d}_c$ is negative. Not surprisingly, it tends to zero when sample size grows. The lower the true probability p, the higher absolute values of bias (see the scale on the y-axis). The same tendency was also observed for p > 0.5 although is not illustrated with the graph. The absolute value of bias also grows with c. It is lowest for c = 1 and highest for c = 4. Results shown on Figure 1 suggest, that increasing the constant c to take integer values higher than one significantly increases the bias. However, this still does not preclude finding some non-integer values for c (perhaps in the vicinity of c = 1) that would provide lower bias.

3. Properties when c is not an integer

When n is small enough, the behavior of the proposed general estimator may be examined by computing the probabilities of all possible sample counts via probability function of binomial distribution. Stochastic properties of the estimator may then be computed explicitly using their definitions. Let us now present such a 'small-sample' study, carried out for n = 500 and varying values of c and p. The bias of the estimator as a function of c and p is shown on Figure 2. Its MSE as a function of c and p is shown on Figure 3. The Figure 4 presents the share of bias in the MSE as a function of c and p.

The bias of the estimator turns out to not always be negative, as the analysis above would suggest. Indeed, it increases when c decreases and for small values of c it grows above zero. Moreover, the bias strongly depends on p. Its absolute value seems to remain rather stable and close to zero for large p but it tends to increase dramatically when p takes values close to zero. Similar tendency is observed for the mean square error of the estimator. It seems to remain quite stable for large p, but increases very quickly when p closes to zero. The MSE also depends on c and for a constant p there is apparently a value of c that minimizes the MSE.

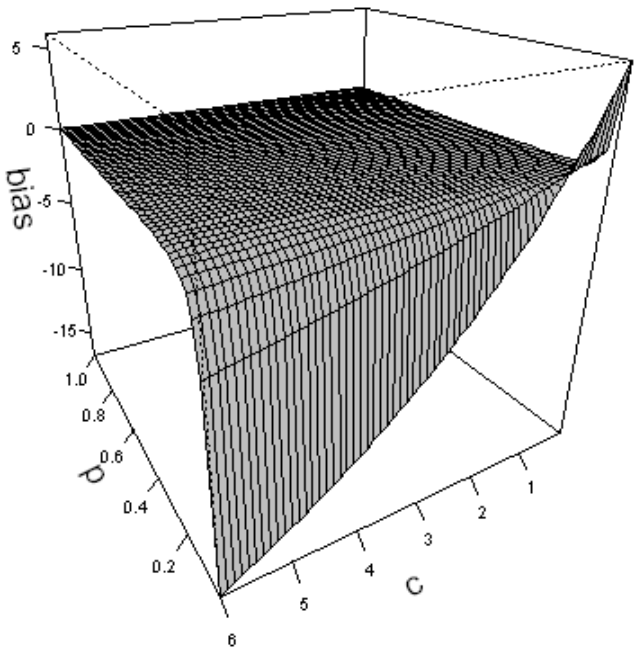


Fig. 2. The bias of the estimator $\hat{d}_c$ as a function of p and c for n = 500

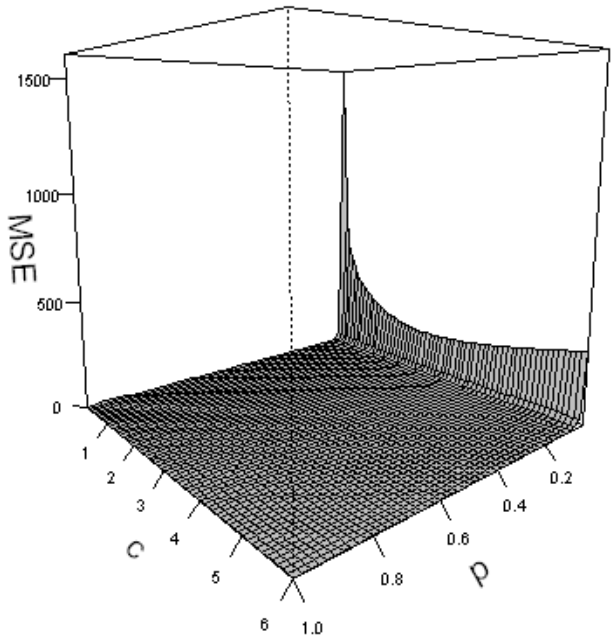


Fig. 3. The MSE of the estimator $\hat{d}_c$ as a function of p and c for n = 500

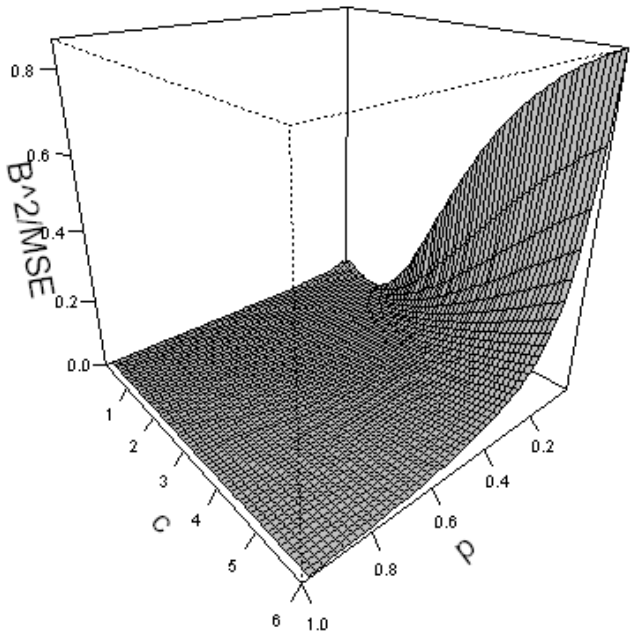


Fig. 4. The share of bias in the MSE as a function of p and c for n = 500

For large p the share of bias in the MSE is very modest, but for smaller p it grows dramatically. If c is also large then the bias dominates the MSE. The dependency of bias share on c resembles that of MSE itself: it seems that for a constant p there exists a value of c for which the share of bias in the MSE is minimized.

For a fixed n, one may be interested in choosing the optimal value of c that nullifies the bias or minimizes the MSE. However the probability p is not controlled by the sampler. Hence it is important to verify if such an optimal value of c depends on p or if it is perhaps constant. In order to do this the dependence of bias and MSE on c for certain values of p is plotted on Figures 5 and 6.

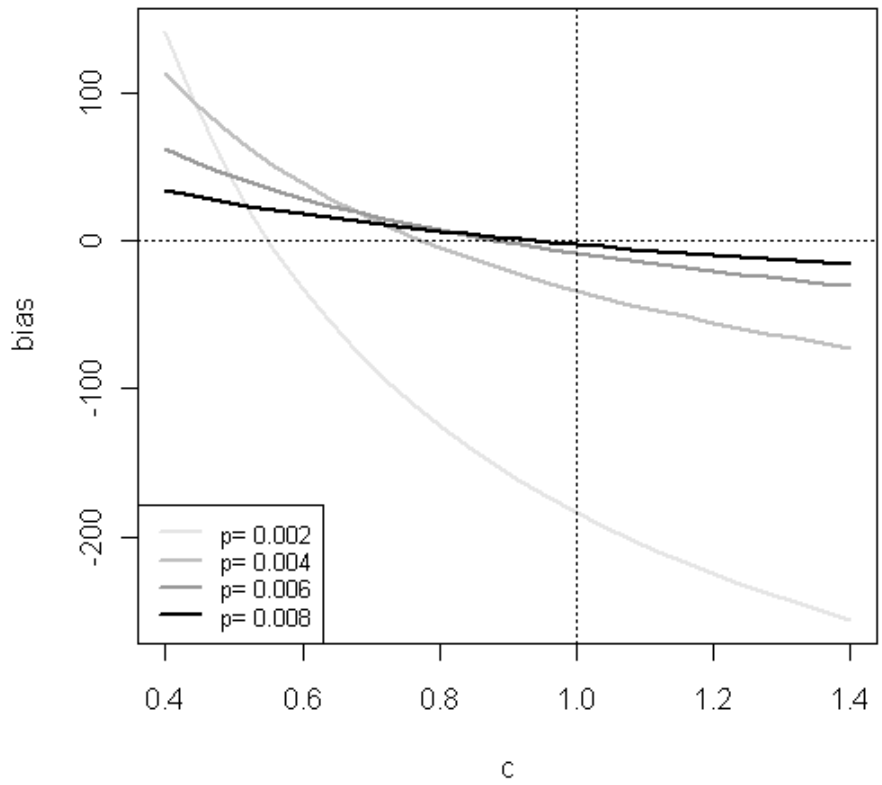


Fig. 5. The bias as a function of c for certain values of p and n = 500

The Figure 5 clearly demonstrates that unfortunately there is no universal value of c that would nullify the bias for any possible p. When p grows, the value of c for which the bias is equal to zero is shifting towards higher values (see

Figure 7), although it remains below one (even for much larger p's such as p = 0.99, since the bias curve very quickly looses steepness when p grows). It seems to be coherent with earlier observation of bias growing with c when c > 1 and sample size is large. This renders any value c above one unjustifiable. These results are also coherent with the observation that for the statistic of Fattorini [2006] based on c = 1 the bias is negative for any possible $p \in (0,1)$. On the other hand, while optimal c-values nullifying the bias are known to lie below one, setting c to some value which is too close to zero may result in strong positive bias.

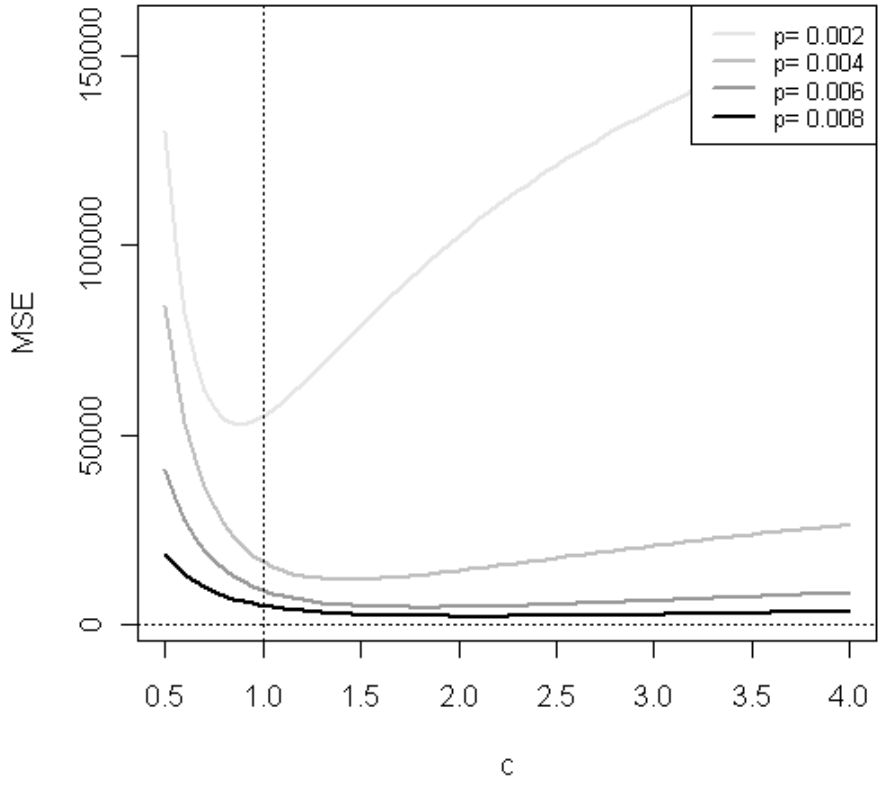


Fig. 6. The bias as a function of c for certain values of p and n = 500

The results for the MSE shown on Figure 6 indicate that unfortunately optimum values of c that minimize the MSE also change with p. They move upwards with growing p but do not seem to stabilize. The curve reflecting dependence of optimal c-values on p shown on Figure 7 is non-decreasing but its shape is somewhat complicated, with the second derivative changing sign at least two times (for $p \approx 0.02$ and $p \approx 0.96$). The value of c = 1 chosen by Fattorini [2006]

apparently minimizes MSE for some value of p inside the [0.002, 0.004] interval. Hence, if there is e.g. external knowledge available that implies $p > 0.004$ then it is reasonable to expect that MSE is minimized for some $c > 1$, and any c-value lower than one would not be advisable. However again, choosing a too large c-value might also increase the MSE instead of reducing it. On the other hand, one may also notice that although optimal values swing wildly with changing p, the MSE seems to be much more dependent on p itself, than on the choice of constant c.

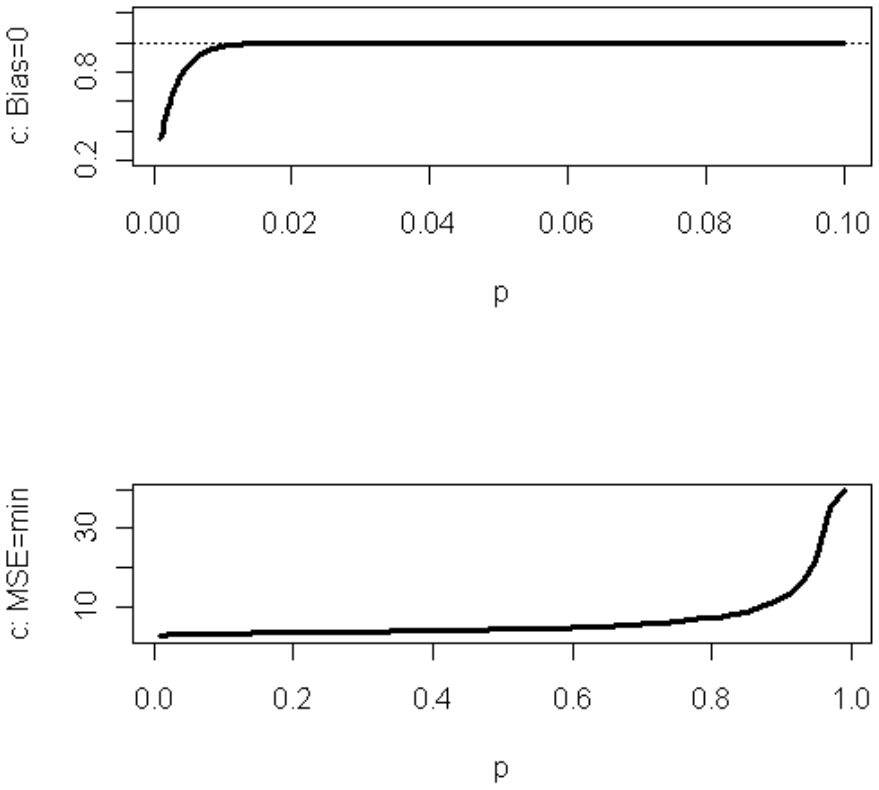


Fig. 7. The c-values nullifying the bias and minimizing the MSE for p = 0.01,0.02,...,0.99 and for n = 500

Conclusions

In this paper a class of estimators for inverse probability indexed by a parameter $c \in (0,+\infty)$ was considered. All estimators in the class are consistent.

They always take finite values, as opposed to the simple reciprocal of the sampling fraction. The class incorporates the well-known Fattorini's [2006] statistic for c = 1. The formulas for bias of these estimators were derived for c = 1,...,4 and a method for computing the bias for c = 5,6,... was suggested.

The bias of estimators depends on sample size n, parameter c and unknown probability p. It turns out that there is no single value of c that would nullify the bias or minimize the mean square error for all possible values of p. In other words, no estimator in the class dominates others in terms of accuracy for a fixed n and all values of p. This also applies to the Fattorini's statistic. However, it seems that values of c greater than one do not nullify bias for any possible p so they should be avoided.

If the exact formula (or upper bound) for MSE or absolute bias when c is not integer and the sample is large were known, then some partial knowledge on p taking e.g. form of inequality constraints might be explored to set the value of c in such a way that MSE or bias for most pessimistic (unfavorable) p is minimized. This justifies further efforts aimed at finding such a formula.

Literature

Berry C.J. (1989): *Bayes Minimax Estimation of a Bernoulli p in a Restricted Parameter Space*. "Communications in Statistics – Theory and Methods", No. 18(12).

Fattorini L. (2006): *Applying the Horvitz-Thompson Criterion in Complex Designs: A Computer-Intensive Perspective for Estimating Inclusion Probabilities*. "Biometrika", No. 93(2).

Fattorini L. (2009): *An Adaptive Algorithm for Estimating Inclusion Probabilities and Performing the Horvitz-Thompson Criterion in Complex Designs*. "Computational Statistics", No. 24.

Marciniak E., Wesołowski J. (1999): *Asymptotic Eulerian Expansions for Binomial and Negative Binomial Reciprocals*. "Proceedings of the American Mathematical Society", No. 127(11).

Marchand E., Perron F., Gueye R. (2005): *Minimax Estimation of a Constrained Binomial Proportion p When |p-1/2| is Small*. "Sankhya", No. 67(3).

Marchand E., MacGibbon B. (2000): *Minimax Estimation of a Constrained Binomial Proportion*. "Statistics & Decisions", No. 18.

Rempała G., Szekely G.J. (1998): *On Estimation with Elementary Symmetric Polynomials*. "Random Operations and Stochastic Equations", No. 6.

Stephan F.F. (1946): *The Expected Value and Variance of the Reciprocal and Other Negative Powers of a Positive Bernoullian Variate*. "Annals of Mathematical Statistics", No. 16.

Thompson M.E., Wu C. (2008): *Simulation-based Randomized Systematic PPS Sampling under Substitution of Units*. "Survey Methodology", No. 34 (1).

APPENDIX 1

Derivation of the formula (2):

$$E\left(\frac{1}{X+c}\right)=\sum_{k=0}^{n}\frac{1}{k+c}\binom{n}{k}p^k q^{n-k}=\sum_{k=0}^{n}\frac{1}{k+c}\frac{n!}{k!(n-k)!}p^k q^{n-k}=$$

$$=\frac{1}{p^c}\frac{n!}{(n+c)!}\sum_{k=0}^{n}\frac{(k+c-1)!}{k!}\binom{n+c}{k+c}p^{k+c}q^{n+c-(k+c)}=$$

$$=\frac{1}{p^c}\frac{n!}{(n+c)!}\sum_{k=c}^{n+c}\frac{(k-1)!}{(k-c)!}\binom{n+c}{k}p^k q^{n+c-k}=$$

$$=\frac{1}{p^c}\frac{n!}{(n+c)!}\sum_{k=c}^{n+c}\left(\prod_{r=1}^{c-1}(k-r)\right)\binom{n+c}{k}p^k q^{n+c-k}=$$

$$=\frac{1}{p^c}\frac{n!}{(n+c)!}\left(E_{n+c}\left(\prod_{r=1}^{c-1}(X-r)\right)-\sum_{k=0}^{c-1}\left(\prod_{r=1}^{c-1}(k-r)\right)\binom{n+c}{k}p^k q^{n+c-k}\right)=$$

$$=\frac{1}{p^c}\frac{n!}{(n+c)!}\left(E_{n+c}\left(\prod_{r=1}^{c-1}(X-r)\right)-(0-1)(0-2)(0-(c-1))\binom{n+c}{0}p^0 q^{n+c}-\right.$$

$$\left.-\sum_{k=1}^{c-1}(k-1)(k-2)(k-(c-1))\binom{n+c}{k}p^k q^{n+c-k}\right)=$$

$$=\frac{1}{p^c}\frac{n!}{(n+c)!}\left(E_{n+c}\left(\prod_{r=1}^{c-1}(X-r)\right)+(-1)^c(c-1)!q^{n+c}\right)$$

APPENDIX 2

Derivation of formulas (3)-(6) representing expectations for $\hat{d}_c$ when c = 1,2,3,4.

$$E(\hat{d}_1)=E\left(\frac{n+1}{X+1}\right)=(n+1)\frac{1}{p^1}\frac{n!}{(n+1)!}\left(E_{n+1}(1)+(-1)^1(1-1)!q^{n+1}\right)=\frac{1-q^{n+1}}{p}$$

$$E(\hat{d}_2) = E\left(\frac{n+2}{X+2}\right) = \frac{n+2}{p^2}\frac{n!}{(n+2)!}\left(E_{n+2}(X) - 1 + (-1)^2(2-1)!q^{n+2}\right) =$$

$$= \frac{1}{n+1}\frac{(n+2)p - 1 + q^{n+2}}{p^2}$$

$$E(\hat{d}_3) = E\left(\frac{n+3}{X+3}\right) =$$

$$= (n+3)\frac{1}{p^3}\frac{n!}{(n+3)!}\left(E_{n+3}(X^2) - 3E_{n+3}(X) + 2 + (-1)^3(3-1)!q^{n+3}\right) =$$

$$= \frac{(n+3)p\,(q + (n+3)p) - 3(n+3)p + 2 - 2q^{n+3}}{(n+1)(n+2)p^3} =$$

$$= \frac{p(n+3)\,(np + 2p - 2) + 2 - 2q^{n+3}}{(n+1)(n+2)p^3}$$

For c = 4 let us note that:

$E_{n+4}(X^3)$-6 $E_{n+4}(X^2)$+ 11$E_{n+4}(X)$-6 =(n+4)p(1-3p+3(n+4)p+2p^2-

-3(n+4)p^2+(n+4)$^2p^2$) -6 (n+4)p (1-p+(n+4)p)+11(n+4) p -6 =

= (n+4)p [1 -3p + (n+4)p [3-3p+(n+4)p-6] +2p^2 -6(1-p) +11] -6 =

= (n+4)p [6 +3p + (n+4)p [np+p-3] +2p^2] -6

and consequently:

$$E\left(\frac{n+4}{X+4}\right) = (n+4)\frac{1}{p^4}\frac{n!}{(n+4)!}\left(E_{n+4}\left(\prod_{r=1}^{4-1}(X-r)\right) + (-1)^4(4-1)!q^{n+4}\right) =$$

$$= \frac{1}{p^4}\frac{E_{n+4}(X^3) - 6E_{n+4}(X^2) + 11E_{n+4}(X) - 6 + 6q^{n+4}}{(n+1)(n+2)(n+3)} =$$

$$= \frac{(n+4)p[6 + 3p + (n+4)p(np + p - 3) + 2p^2] - 6 + 6q^{n+4}}{p^4(n+1)(n+2)(n+3)}$$

O KLASIE ESTYMATORÓW ODWROTNOŚCI PRAWDOPODOBIEŃSTWA

Streszczenie

Powszechnie znany estymator odwrotności prawdopodobieństwa zaproponowany przez Fattoriniego uogólniono poprzez dopuszczenie zmian wartości jednostkowej stałej występującej we wzorze definiującym go. Prowadzi to do konstrukcji klasy estymatorów odwrotności prawdopodobieństwa. Własności tych estymatorów zbadano analitycznie. Zaproponowano metodę wyznaczania asymptotycznego obciążenia dla całkowitych wartości zmodyfikowanej stałej. Własności estymatorów dla małych prób wyznaczono numerycznie, korzystając z własności rozkładu dwumianowego.

Tomasz Żądło
Uniwersytet Ekonomiczny w Katowicach

ON SOME PROBLEMS OF PREDICTION OF DOMAIN TOTAL IN LONGITUDINAL SURVEYS WHEN AUXILIARY INFORMATION IS AVAILABLE

Introduction

In the survey sampling the problem of estimation or prediction of subpopulations' (domains') characteristics has become very important issue. What is more, in the case of longitudinal surveys there is a possibility to increase the accuracy of the estimators or predictors by using information from other periods or even to estimate or predict subpopulation's characteristic for the period when the number of sampled domain elements equals zero. Domains with small or zero sample sizes are called small areas. In small area estimation empirical versions of Henderson's [1950] best linear unbiased predictors (BLUP) are widely used under different longitudinal area level models (see e.g. Rao [2003] chapter 8.3 and Rao and Yu [1994]). In the paper the class of unit level longitudinal models with auxiliary variables is proposed assuming that the population and the domains affiliation may change in time. In the paper the predictors which are empirical versions of Royall's [1976] BLUP under some special cases of the proposed model are derived. They can be used to predict the domain total based on any longitudinal data (including e.g. random and purposive samples, panel data and rotating samples) for any (including future) periods. Their mean squared errors (MSEs) and MSEs' estimators are also derived. In the Monte Carlo simulation study the problems of the accuracy of the predictor and biases of the MSE estimators are analyzed based on real data including several cases of model misspecification. The results of the simulation show that the proposed predictor and the proposed MSE estimator may perform very well even in some cases of model misspecification.

1. Basic notations

Let us introduce some notation presented earlier by Żądło [2009b]. In the paper longitudinal data for periods $t = 1,...,M$ are considered. In the period t the population of size N_t is denoted by Ω_t. The population in the period t is divided into D disjoint domains (subpopulations) Ω_{dt} of size N_{dt}, where $d = 1,...,D$. Let the set of population elements for which observations are available in the period t be denoted by s_t and its size by n_t. The set of domain elements for which observations are available in the period t is denoted by s_{dt} and its size by n_{dt}. Let: $\Omega_{rdt} = \Omega_{dt} - s_{dt}$, $N_{rdt} = N_{dt} - n_{dt}$.

Let M_{id} denotes the number of periods when the ith population element may be potentially observed in the dth domain (when the ith population element belongs to the dth domain). Let us denote the number of periods when the ith population element (which belongs to the dth domain) is observed by m_{id}. Let $m_{rid} = M_{id} - m_{id}$. We assume that the population may change in time and that one population element may change its domain affiliation in time (from technical point of view observations of some population element which change its domain affiliation are treated as observations of new population element). It means that i and t completely identify domain affiliation but additional subscript d will be needed as well. More about this assumptions will be written at the end of the next section.

The set of elements which belong at least in one of periods $t = 1,...,M$ to sets Ω_t is denoted by Ω and its size by N. Similarly, sets Ω_d, s, s_d, Ω_{rd} of sizes N_d, n, n_d, N_{rd} respectively are defined as sets of elements which belong at least in one of periods $t = 1,...,M$ to sets Ω_{dt}, s_t, s_{dt}, Ω_{rdt} respectively. The d^*th domain of interest in the period of interest t^* will be additionally denoted by a symbol * in the subscript i.e. $\Omega_{d^*t^*}$, and the set of elements which belong at least in one of periods $t = 1,...,M$ to sets $\Omega_{d^*t^*}$ will be denoted by Ω_{d^*}.

Values of the variable of interest are realizations of random variables Y_{idj} for the ith population element which belongs to the dth domain in the period t_{ij}, where $i = 1,...,N$, $j = 1,...,M_{id}$, $d = 1,...,D$. The vector of size $M_{id} \times 1$ of random variables Y_{idj} for the ith population element which belongs to the dth domain will be denoted by $\mathbf{Y_{id}} = \left[Y_{idj} \right]$, where $j = 1,..., M_{id}$. Let us consider values of

the variables of interest $Y_{i'd'j'}$ for the i'th population element which belongs to the d'th domain observed in periods $t_{i'j'}$, where $i' = 1,...,n$, $j' = 1,...,m_{i'd'}$, $d' = 1,...,D$. The vector of random variables $Y_{i'd'j'}$ (where $i' = 1,...,n$, $j' = 1,...,m_{i'd'}$, $d' = 1,...,D$) of size $m_{i'd'} \times 1$ will be denoted by $\mathbf{Y_{s\,i'd'}} = \left[Y_{i'd'j'} \right]$, where $j' = 1,...,m_{i'd'}$. The vector of random variables $Y_{i"d"j"}$ of size $m_{ri"d"} \times 1$ for the $i"$th population element which belongs to the $d"$th domain for observations which are not available in the sample is denoted by $\mathbf{Y_{r\,i"d"}} = \left[Y_{i"d"j"} \right]$, where $j" = 1,...,m_{ri"d"}$.

The proposed approach may be used to predict the domain total for any (past, current and future) periods. If the problem of prediction of the domain total for the future period is considered, the number of periods M includes future period or periods. What is more, in this case the division of the population into domains and values of the auxiliary variables in the future are assumed to be known.

2. Superpopulation model

We consider some class of superpopulation models (studied earlier by Żądło [2009b]) used for longitudinal data (compare Verbeke, Molenberghs, [2000]; Hedeker, Gibbons [2006]) which are – what is important for further considerations – special cases of the General Linear Model (GLM) and the General Linear Mixed Model (GLMM). The following two-stage model is assumed. Firstly:

$$\mathbf{Y_{id}} = \mathbf{Z_{id}}\boldsymbol{\beta}_{\mathbf{id}} + \mathbf{e_{id}}, \tag{1}$$

where $i = 1,...,N$; $d = 1,...,D$, $\mathbf{Y_{id}}$ is a random vector of size $M_{id} \times 1$, $\mathbf{Z_{id}}$ is known matrix of size $M_{id} \times q$, $\boldsymbol{\beta}_{\mathbf{id}}$ is a vector of unknown parameters of size $q \times 1$, $\mathbf{e_{id}}$ is a random component vector of size $M_{id} \times 1$. Vectors $\mathbf{e_{id}}$ ($i = 1,...,N$; $d = 1,...,D$) are independent with $\mathbf{0}$ vectors of expected values and variance-covariance matrices $\mathbf{R_{id}}$. Although $\mathbf{R_{id}}$ may depend on i it is often assumed that $\mathbf{R_{id}} = \sigma_e^2 \mathbf{I}_{\mathbf{M_{id}}}$ where $\mathbf{I}_{\mathbf{M_{id}}}$ is the identity matrix of rank M_{id}. Secondly, we assume that:

$$\boldsymbol{\beta}_{\mathbf{id}} = \mathbf{K}_{\mathbf{id}}\boldsymbol{\beta} + \mathbf{v}_{\mathbf{id}}, \tag{2}$$

where $i = 1,...,N$; $d = 1,...,D$, $\mathbf{K}_{\mathbf{id}}$ is known matrix of size $q \times p$, $\boldsymbol{\beta}$ is a vector of unknown parameters of size $p \times 1$, $\mathbf{v}_{\mathbf{id}}$ is a vector of random components of size $q \times 1$. It is assumed that vectors $\mathbf{v}_{\mathbf{id}}$ ($i = 1,...,N$; $d = 1,...,D$) are independent with $\mathbf{0}$ vectors of expected values and variance-covariance matrix $\mathbf{G}_{\mathbf{id}} = \mathbf{H}$ what means that $\mathbf{G}_{\mathbf{id}}$ does not depend on i.

Similar assumptions to (1) and (2) are presented by Verbeke, Molenberghs [2000, p. 20] but there are two differences. Firstly, in the book assumptions are made for profiles defined by elements. In this paper assumptions are made for profiles defined by elements and domains affiliation i.e. $\mathbf{Y}_{\mathbf{id}}$ (of size $M_{id} \times 1$) what allows to take the possibility of population changes in time into account. Secondly, in the book the assumptions are made only for the sampled elements (i.e. $i = 1,...,n$). In this paper they are made for all of population elements ($i = 1,...,N$).

Based on (1) and (2) it is obtained that:

$$\mathbf{Y}_{\mathbf{id}} = \mathbf{X}_{\mathbf{id}}\boldsymbol{\beta} + \mathbf{Z}_{\mathbf{id}}\mathbf{v}_{\mathbf{id}} + \mathbf{e}_{\mathbf{id}}, \tag{3}$$

where $i = 1,...,N$; $d = 1,...,D$, $\mathbf{X}_{\mathbf{id}} = \mathbf{Z}_{\mathbf{id}}\mathbf{K}_{\mathbf{id}}$ is known matrix of size $M_{id} \times p$. Let $\mathbf{V}_{\mathbf{id}} = D^2_{\xi}(\mathbf{Y}_{\mathbf{id}})$. Hence,

$$\mathbf{V}_{\mathbf{id}} = \mathbf{Z}_{\mathbf{id}}\mathbf{H}\mathbf{Z}^T_{\mathbf{id}} + \mathbf{R}_{\mathbf{id}}. \tag{4}$$

Let $\mathbf{A}_d$ be a column vector and $col_{1\le d\le D}(\mathbf{A}_d) = \left[\mathbf{A}_{\mathbf{1}}^T \quad ... \quad \mathbf{A}_d^T \quad ... \quad \mathbf{A}_D^T\right]^T$ be a column vector obtained by stacking $\mathbf{A}_d$ vectors. Note that by stacking $\mathbf{Y}_{\mathbf{id}}$ vectors (i.e. $\mathbf{Y} = col_{1\le d\le D}(col_{1\le i\le N_d}(\mathbf{Y}_{\mathbf{id}}))$ from (3) we obtain the formula of the GLMM. Let $\mathbf{V} = D^2_{\xi}(\mathbf{Y})$. Hence,

$$\mathbf{V} = diag_{1\le d\le D} diag_{1\le i\le N_d}(\mathbf{V}_{\mathbf{id}}) \tag{5}$$

Unknown elements of $\mathbf{V}$ will be denoted by $\boldsymbol{\delta}$. Let, $\mathbf{Y}_{\mathbf{s}} = col_{1\le d\le D} col_{1\le i\le n_d}(\mathbf{Y}_{\mathbf{sid}})$, $\mathbf{V}_{\mathbf{ss}} = D^2_{\xi}(\mathbf{Y}_{\mathbf{s}})$, $\mathbf{V}_{\mathbf{ssid}} = D^2_{\xi}(\mathbf{Y}_{\mathbf{sid}})$. Hence,

$$\mathbf{V_{ss}} = diag_{1\le d\le D} diag_{1\le i\le n_d}(\mathbf{V_{ssid}}) = diag_{1\le d\le D} diag_{1\le i\le n_d}(\mathbf{Z_{sid}} \mathbf{H} \mathbf{Z}_{\mathbf{sid}}^{\mathbf{T}} + \mathbf{R_{sid}}) \quad (6)$$

where $\mathbf{Z_{sid}}$ is known matrix of size $m_{id} \times q$, $\mathbf{R_{sid}} = D_{\xi}^2(\mathbf{e_{sid}})$ and $\mathbf{e_{sid}}$ is $m_{id} \times 1$ random components vector.

3. EBLUP, its MSE AND MSE estimator

At the beginning let us compare BLUPs proposed by Henderson [1950] and Royall [1976]. Firstly, Royall derived the BLUP assuming the GLM which is generalization of the GLMM assumed by Henderson. Secondly, Royall predicts linear combination of $\mathbf{Y}$ given by $\theta = \boldsymbol{\gamma}^T \mathbf{Y}$ what is more general then linear combination of $\boldsymbol{\beta}$ and $\mathbf{v}$ given by $\theta_s = \mathbf{l}^{\mathbf{T}}\boldsymbol{\beta} + \mathbf{m}^{\mathbf{T}}\mathbf{v}$ studied by Henderson. Thirdly, in both cases linear predictors are considered: $\hat{\theta} = \mathbf{g}_{\mathbf{s}}^{\mathbf{T}}\mathbf{Y}_{\mathbf{s}}$ by Royall [1976] and $\hat{\theta}_s = \mathbf{a}^{\mathbf{T}}\mathbf{Y}_{\mathbf{s}} + b$ by Henderson, which forms are equivalent because $b = 0$ under unbiasedness. Hence, Royall's BLUP may be treated as the generalization of Henderson's BLUP. In the paper the BLUP proposed by Royall is studied (and its empirical version – EBLUP) where the element k of the $\boldsymbol{\gamma}$ vector is given by:

$$\gamma_k = \begin{cases} 0 & \text{if} \quad i \notin \Omega_{d^*t^*} \\ 1 & \text{if} \quad i \in \Omega_{d^*t^*} \end{cases} \quad (7)$$

To obtain the BLUP of d^* th domain total in t^* th period and its MSE for model (3) general formulae proposed by Royall should be used with (7) and block-diagonal form of variance-covariance matrix (5). If the unknown parameters in the formula of the BLUP proposed by Royall are replaced by their estimates, two-stage predictor called the EBLUP is obtained. Kackar and Harville [1981] prove unbiasedness of empirical version of the BLUP proposed by Henderson under some weak assumptions. The proof of unbiasedness of empirical version of the BLUP proposed by Royall under similar weak assumptions (inter alia symmetric but not necessarily normal distribution of random components for the model assumed for the whole population), is presented in Żądło [2004]. The approximation of the MSE and its estimator for the empirical version of the BLUP proposed by Henderson are derived inter alia

by Prasad and Rao [1990] and Datta and Lahiri [2000]. The approximation of the MSE and its estimator for the empirical version of the BLUP proposed by Royall are derived in Żądło [2009a] based on results presented in Datta and Lahiri [2000].

4. Special cases of superpopulation model

In the section we consider two special cases of the model (3). The first model is longitudinal random regression coefficient model similar to the one proposed in Dempster, Rubin and Tsutakawa [1981] and studied later e.g. in Moura and Holt [1999] and for one auxiliary variable in Prasad and Rao [1990]. Unlike the proposed longitudinal model, these authors only consider a model with domain-specific random effects (and for one period). We assume that:

$$Y_{idj} = (\beta_d + v_{id})x_{idj} + e_{idj} = \beta_d x_{idj} + v_{id} x_{idj} + e_{idj}, \tag{8}$$

where $i = 1,2,...,N$; $d = 1,2,...,D$, $j = 1,2,...,M_{id}$. Special case of (8) where

$$\forall_d \ \beta_d = \beta \tag{9}$$

will also be considered. What is more (similarly to Verbeke, Molenberghs [2000]), we assume that e_{idj} and v_{id} are mutually independent and $e_{idj} \sim (0, \sigma_e^2)$ and $v_{id} \sim (0, \sigma_v^2)$. Hence,

$$Cov_\xi(Y_{idj}, Y_{i'j'd'}) = \begin{cases} 0 & \text{if} \quad i \neq i' \vee d \neq d' \\ \sigma_e^2 + x_{idj}^2 \sigma_v^2 & \text{if} \quad i = i' \wedge j = j' \\ x_{idj} x_{i'j'd'} \sigma_v^2 & \text{if} \quad i = i' \wedge d = d' \wedge j \neq j' \end{cases}, \tag{10}$$

The second model is nested error regression models similar to the one proposed in Battese, Harter and Fuller [1988]. Unlike the proposed longitudinal model, these authors only consider a model with domain-specific random effects (and for one period). We assume that:

$$Y_{idj} = \mathbf{x_{idj}}\boldsymbol{\beta}_\mathbf{d} + v_{id} + e_{idj}, \tag{11}$$

where $\mathbf{x_{idj}} = \left[x_{idj1} \quad x_{idj2} \quad ... \quad x_{idjp} \right]$, e_{idj} and v_{id} are mutually independent and $e_{idj} \sim (0, \sigma_e^2)$ and $v_{id} \sim (0, \sigma_v^2)$. Special case of (11) where:

$$\forall_d \ \boldsymbol{\beta}_\mathbf{d} = \boldsymbol{\beta} \tag{12}$$

will also be considered. Hence,

$$Cov_\xi(Y_{idj}, Y_{i'j'd'}) = \begin{cases} 0 & \text{if} \quad i \neq i' \vee d \neq d' \\ \sigma_e^2 + \sigma_v^2 & \text{if} \quad i = i' \wedge j = j' \\ \sigma_v^2 & \text{if} \quad i = i' \wedge d = d' \wedge j \neq j' \end{cases}. \tag{13}$$

For all of the superpopulation models presented in this section the vector of unknown variance parameters will be denoted by $\boldsymbol{\delta} = \left[\sigma_e^2 \quad \sigma_v^2 \right]^T$.

We have assumed that the population and the domain affiliation of population elements may change in time. Observations of new element of the population or observations of the population element after the change of its domain affiliation are treated as realizations of new profile (3). Hence, because of the covariance structure (5) where nonzero covariances are only within profiles, we assume the lack of correlation of observations for some population element before and after the change of the domain affiliation.

5. Prediction under a longitudinal random regression coefficient model

Based on Royall's theorem [1976], it is possible to derive the BLUP of the d^* th domain total in the t^* th (past, current or future) period and its MSE under longitudinal simple random regression coefficient model (8). They are given by:

$$\hat{\theta}_{BLU} = \sum_{i \in s_{d^*t^*}} Y_{id^*t^*} + \hat{\beta}_{d^*} \sum_{i=1}^{N_{rd^*t^*}} x_{id^*t^*} + \sigma_v^2 \sum_{i=1}^{N_{rd^*t^*}} x_{id^*t^*} b_{id^*}^{-1} \sum_{j=1}^{m_{id^*}} x_{id^*j} (Y_{id^*j} - x_{id^*j} \hat{\beta}_{d^*}) \tag{14}$$

where $\hat{\beta}_{d^*} = \left(\sum_{i=1}^{n_{d^*}} b_{id^*}^{-1} \sum_{j=1}^{m_{id^*}} x_{id^*j}^2 \right)^{-1} \left(\sum_{i=1}^{n_{d^*}} b_{id^*}^{-1} \sum_{j=1}^{m_{id^*}} Y_{id^*j} x_{id^*j} \right)$, $b_{id^*} = \sigma_e^2 + \sigma_v^2 \sum_{j=1}^{m_{id^*}} x_{id^*j}^2$

and

$$MSE_{\xi}(\hat{\theta}_{BLU}) = g_1(\boldsymbol{\delta}) + g_2(\boldsymbol{\delta}) \tag{15}$$

where:

$$g_1(\boldsymbol{\delta}) = N_{rd^*t^*}\sigma_e^2 + \sigma_v^2 \sum_{i=1}^{N_{rd^*t^*}} x_{id^*t^*}^2 - \sigma_v^4 \sum_{i=1}^{N_{rd^*t^*}} x_{id^*t^*}^2 b_{id^*}^{-1} \sum_{j=1}^{m_{id^*}} x_{id^*j}^2 \tag{16}$$

$$g_2(\boldsymbol{\delta}) = \left(\sum_{i=1}^{N_{rd^*t^*}} x_{id^*t^*} - \sigma_v^2 \sum_{i=1}^{N_{rd^*t^*}} x_{id^*t^*} b_{id^*}^{-1} \sum_{j=1}^{m_{id^*}} x_{id^*j}^2 \right)^2 \left(\sum_{i=1}^{n_{d^*}} b_{id^*}^{-1} \sum_{j=1}^{m_{id^*}} x_{id^*j}^2 \right)^{-1}. \tag{17}$$

Let the unknown variance parameters in (14) be replaced by their maximum likelihood (ML) or restricted maximum likelihood (REML) estimates under normality. Hence, we obtain the two-stage predictor called EBLUP. Using general theorems proved in Żądło [2009a] it is possible to derive the formula of the MSE of the EBLUP and its estimators. Firstly, under assumptions presented in Żądło [2009a] (including the GLMM with block-diagonal variance-covariance matrix and normality of random components) the MSE in this case is given by:

$$MSE_{\xi}(\hat{\theta}_{EBLU}) = g_1(\boldsymbol{\delta}) + g_2(\boldsymbol{\delta}) + g_3^*(\boldsymbol{\delta}) + o(D^{-1}) \tag{18}$$

where $g_1(\boldsymbol{\delta})$ and $g_2(\boldsymbol{\delta})$ are given by (16) and (17) respectively and

$$g_3^*(\boldsymbol{\delta}) = \sum_{i=1}^{N_{rd^*t^*}} x_{id^*t^*}^2 b_{id^*}^{-3} \sum_{j=1}^{m_{id^*}} x_{id^*j}^2 \left(I_{vv}^{(-1)}\sigma_v^4 - 2I_{ve}^{(-1)}\sigma_e^2\sigma_v^2 + I_{ee}^{(-1)}\sigma_e^4 \right) \tag{19}$$

and

$$I_{vv}^{(-1)} = 2b^{-1} \sum_{d=1}^{D} \sum_{i=1}^{n_d} b_{id}^{-2} \left(\sum_{j=1}^{m_{id}} x_{idj}^2 \right)^2, \tag{20}$$

$$I_{ve}^{(-1)} = -2b^{-1}\sum_{d=1}^{D}\sum_{i=1}^{n_d} b_{id}^{-2}\left(\sum_{j=1}^{m_{id}} x_{idj}^2\right), \tag{21}$$

$$I_{ee}^{(-1)} = 2b^{-1}\sum_{d=1}^{D}\sum_{i=1}^{n_d}\left((m_{id}-1)\sigma_e^{-4} + b_{id}^{-2}\right), \tag{22}$$

and

$$b = \left(\sum_{d=1}^{D}\sum_{i=1}^{n_d}\left((m_{id}-1)\sigma_e^{-4} + b_{id}^{-2}\right)\right)\left(\sum_{d=1}^{D}\sum_{i=1}^{n_d} b_{id}^{-2}\left(\sum_{j=1}^{m_{id}} x_{idj}^2\right)^2\right) +$$

$$-\left(\sum_{d=1}^{D}\sum_{i=1}^{n_d} b_{id}^{-2}\left(\sum_{j=1}^{m_{id}} x_{idj}^2\right)\right)^2$$

Secondly, under general assumptions presented in Żądło [2009a] (including the GLMM with block-diagonal variance-covariance matrix and normality of random components) the approximately unbiased (its bias is $o(D^{-1})$) estimator of the MSE (18) for REML estimators of $\boldsymbol{\delta}$ in this case is given by:

$$\hat{MSE}_{\xi}(\hat{\theta}_{EBLU}) = g_1(\boldsymbol{\delta}) + g_2(\boldsymbol{\delta}) + 2g_3^*(\boldsymbol{\delta}) \tag{23}$$

and for ML estimators of $\boldsymbol{\delta}$ by

$$\hat{MSE}_{\xi}(\hat{\theta}_{EBLU}) = g_1(\hat{\boldsymbol{\delta}}) + g_2(\hat{\boldsymbol{\delta}}) + 2g_3^*(\hat{\boldsymbol{\delta}}) +$$

$$-\frac{1}{2}\left[\mathbf{I}_{\boldsymbol{\delta}}^{-1}(\hat{\boldsymbol{\delta}})col_{1\le k\le q}tr\left[\mathbf{I}_{\boldsymbol{\beta}}^{-1}(\hat{\boldsymbol{\delta}})\frac{\partial}{\partial\delta_k}\mathbf{I}_{\boldsymbol{\beta}}(\hat{\boldsymbol{\delta}})\right]\right]^T\frac{\partial g_1(\hat{\boldsymbol{\delta}})}{\partial\boldsymbol{\delta}} \tag{24}$$

where $g_1(\hat{\boldsymbol{\delta}})$, $g_2(\hat{\boldsymbol{\delta}})$, $g_3^*(\hat{\boldsymbol{\delta}})$ are given by (16), (17), (19) respectively where $\boldsymbol{\delta}$ is replaced by $\hat{\boldsymbol{\delta}}$, $\mathbf{I}_{\boldsymbol{\delta}}^{-1} = \begin{bmatrix} I_{vv}^{(-1)} & I_{ve}^{(-1)} \\ I_{ve}^{(-1)} & I_{ee}^{(-1)} \end{bmatrix}$, where $I_{vv}^{(-1)}$, $I_{ve}^{(-1)}$, $I_{ee}^{(-1)}$ are given by (20), (21), (22) respectively, $col_{1\le k\le q}tr\left[\mathbf{I}_{\boldsymbol{\beta}}^{-1}(\hat{\boldsymbol{\delta}})\frac{\partial}{\partial\delta_k}\mathbf{I}_{\boldsymbol{\beta}}(\hat{\boldsymbol{\delta}})\right]$ and $\frac{\partial g_1(\hat{\boldsymbol{\delta}})}{\partial\boldsymbol{\delta}}$ are given by

$$col_{1\le k\le q} tr\left[\mathbf{I}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\delta})\frac{\partial}{\partial\delta_k}\mathbf{I}_{\boldsymbol{\beta}}(\boldsymbol{\delta})\right] = -\begin{bmatrix} \sum_{d=1}^{D}\left(\sum_{i=1}^{n_d} b_{id}^{-1}\sum_{j=1}^{m_{id}} x_{idj}^2\right)^{-1}\left(\sum_{i=1}^{n_d} b_{id}^{-2}\sum_{j=1}^{m_{id}} x_{idj}^2\right) \\ \sum_{d=1}^{D}\left(\sum_{i=1}^{n_d} b_{id}^{-1}\sum_{j=1}^{m_{id}} x_{idj}^2\right)^{-1}\left(\sum_{i=1}^{n_d} b_{id}^{-2}\left(\sum_{j=1}^{m_{id}} x_{idj}^2\right)^2\right) \end{bmatrix} \quad (25)$$

and

$$\frac{\partial g_1(\boldsymbol{\delta})}{\partial\boldsymbol{\delta}} = \left[\frac{\partial g_1(\boldsymbol{\delta})}{\partial\sigma_e^2} \quad \frac{\partial g_1(\boldsymbol{\delta})}{\partial\sigma_v^2}\right]^T =$$

$$= \begin{bmatrix} N_{rd^*t^*} - \sigma_v^4 \sum_{i=1}^{N_{rd^*t^*}} x_{id^*t^*}^2 b_{id^*}^{-2} \sum_{j=1}^{m_{id^*}} x_{id^*j}^2 \\ \sum_{i=1}^{N_{rd^*t^*}} x_{id^*t^*}^2 + \sigma_v^4 \sum_{i=1}^{N_{rd^*t^*}} x_{id^*t^*}^2 b_{id^*}^{-1}\left(2\sum_{j=1}^{m_{id^*}} x_{id^*j}^2 - b_{id^*}^{-1}\left(\sum_{j=1}^{m_{id^*}} x_{id^*j}^2\right)^2\right) \end{bmatrix} \quad (26)$$

respectively, where $\boldsymbol{\delta}$ is replaced by $\hat{\boldsymbol{\delta}}$.

Under assumptions (8) and (9) the equations presented above remain true but $\hat{\beta}_{d^*}$ in (14) should be replaced by

$$\hat{\beta} = \left(\sum_{d=1}^{D}\sum_{i=1}^{n_d} b_{id}^{-1}\sum_{j=1}^{m_{id}} x_{idj}^2\right)^{-1}\left(\sum_{d=1}^{D}\sum_{i=1}^{n_d} b_{id}^{-1}\sum_{j=1}^{m_{id}} Y_{idj}x_{idj}\right),$$

$g_2(\boldsymbol{\delta})$ given by (17) should by replaced by

$$g_2(\boldsymbol{\delta}) = \left(\sum_{i=1}^{N_{rd^*t^*}} x_{id^*t^*} - \sigma_v^2 \sum_{i=1}^{N_{rd^*t^*}} x_{id^*t^*} b_{id^*}^{-1} \sum_{j=1}^{m_{id^*}} x_{id^*j}^2\right)^2 \left(\sum_{d=1}^{D}\sum_{i=1}^{n_d} b_{id}^{-1}\sum_{j=1}^{m_{id}} x_{idj}^2\right)^{-1}$$

and $col_{1\le k\le q} tr\left[\mathbf{I}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\delta})\frac{\partial}{\partial\delta_k}\mathbf{I}_{\boldsymbol{\beta}}(\boldsymbol{\delta})\right]$ given by (25) should be replaced by:

$$col_{1\le k\le q} tr\left[\mathbf{I}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\delta})\frac{\partial}{\partial\delta_k}\mathbf{I}_{\boldsymbol{\beta}}(\boldsymbol{\delta})\right] = -\left[\begin{array}{c}\left(\sum_{d=1}^{D}\sum_{i=1}^{n_d} b_{id}^{-1}\sum_{j=1}^{m_{id}} x_{idj}^2\right)^{-1}\left(\sum_{d=1}^{D}\sum_{i=1}^{n_d} b_{id}^{-2}\sum_{j=1}^{m_{id}} x_{idj}^2\right)\\ \left(\sum_{d=1}^{D}\sum_{i=1}^{n_d} b_{id}^{-1}\sum_{j=1}^{m_{id}} x_{idj}^2\right)^{-1}\left(\sum_{d=1}^{D}\sum_{i=1}^{n_d} b_{id}^{-2}\left(\sum_{j=1}^{m_{id}} x_{idj}^2\right)^2\right)\end{array}\right]$$

6. Prediction under a longitudinal nested error regression model

Based on Royall's theorem [1976], it is possible to derive the BLUP of the d^* th domain total in t^* th (past, current or future) period and its MSE under longitudinal nested error regression coefficient model (11). The BLUP is given by:

$$\hat{\theta}_{BLU} = \sum_{i\in s_{d^*t^*}} Y_{id^*t^*} + \sum_{i=1}^{N_{rd^*t^*}} \mathbf{x}_{\mathbf{id^*t^*}}\hat{\boldsymbol{\beta}}_{\mathbf{d^*}} + \sigma_v^2\sum_{i=1}^{N_{rd^*t^*}} b_{id^*}^{-1}\sum_{j=1}^{m_{id^*}}(Y_{id^*j} - \mathbf{x}_{\mathbf{id^*j}}\hat{\boldsymbol{\beta}}_{\mathbf{d^*}}), \qquad (27)$$

where $b_{id^*} = \sigma_e^2 + \sigma_v^2 m_{id^*}$, $\hat{\boldsymbol{\beta}}_{\mathbf{d^*}} = \left(\sum_{i=1}^{n_{d^*}} b_{id^*}^{-1}\mathbf{X}_{\mathbf{sid^*}}^{\mathbf{T}}\mathbf{X}_{\mathbf{sid^*}}\right)^{-1}\left(\sum_{i=1}^{n_{d^*}} b_{id^*}^{-1}\mathbf{X}_{\mathbf{sid^*}}^{\mathbf{T}}\mathbf{Y}_{\mathbf{sid^*}}\right)$ and $\mathbf{X}_{\mathbf{sid^*}}$ is $m_{id^*}\times p$ known matrix of auxiliary variables. The MSE of the BLUP (27) is given by general formula (15), where

$$g_1(\boldsymbol{\delta}) = N_{rd^*t^*}\left(\sigma_e^2 + \sigma_v^2\right) - \sigma_v^4\sum_{i=1}^{N_{rd^*t^*}} b_{id^*}^{-1}m_{id^*}, \qquad (28)$$

$$g_2(\boldsymbol{\delta}) = \left(\sum_{i=1}^{N_{rd^*t^*}} \mathbf{x}_{\mathbf{id^*t^*}} - \sigma_v^2\sum_{i=1}^{N_{rd^*t^*}} b_{id^*}^{-1}\sum_{j=1}^{m_{id^*}} \mathbf{x}_{\mathbf{id^*j}}\right)\left(\sum_{i=1}^{n_{d^*}} b_{id^*}^{-1}\mathbf{X}_{\mathbf{sid^*}}^{\mathbf{T}}\mathbf{X}_{\mathbf{sid^*}}\right)^{-1}\times$$

$$\times\left(\sum_{i=1}^{N_{rd^*t^*}} \mathbf{x}_{\mathbf{id^*t^*}} - \sigma_v^2\sum_{i=1}^{N_{rd^*t^*}} b_{id^*}^{-1}\sum_{j=1}^{m_{id^*}} \mathbf{x}_{\mathbf{id^*j}}\right)^{\mathbf{T}} \qquad (29)$$

If the unknown variance parameters in (27) are replaced by their ML or REML estimates under normality we obtain the EBLUP with the MSE given by general formula (18), where $g_1(\boldsymbol{\delta})$ and $g_2(\boldsymbol{\delta})$ are given by (28) and (29) respectively and

$$g_3^*(\boldsymbol{\delta}) = \sum_{i=1}^{N_{rd^*}^{**}} b_{id^*}^{-3} m_{id^*} \left(I_{vv}^{(-1)} \sigma_v^4 - 2 I_{ve}^{(-1)} \sigma_e^2 \sigma_v^2 + I_{ee}^{(-1)} \sigma_e^4 \right) \quad (30)$$

where

$$I_{vv}^{(-1)} = 2b^{-1} \sum_{d=1}^{D} \sum_{i=1}^{n_d} b_{id}^{-2} m_{id}^2, \quad (31)$$

$$I_{ve}^{(-1)} = -2b^{-1} \sum_{d=1}^{D} \sum_{i=1}^{n_d} b_{id}^{-2} m_{id}, \quad (32)$$

$$I_{ee}^{(-1)} = 2b^{-1} \sum_{d=1}^{D} \sum_{i=1}^{n_d} \left((m_{id} - 1) \sigma_e^{-4} + b_{id}^{-2} \right), \quad (33)$$

and $b = \left(\sum_{d=1}^{D} \sum_{i=1}^{n_d} \left((m_{id} - 1) \sigma_e^{-4} + b_{id}^{-2} \right) \right) \left(\sum_{d=1}^{D} \sum_{i=1}^{n_d} b_{id}^{-2} m_{id}^2 \right) - \left(\sum_{d=1}^{D} \sum_{i=1}^{n_d} b_{id}^{-2} m_{id} \right)^2$

The approximately unbiased (its bias is $o(D^{-1})$) estimator of the MSE of the EBLUP for the REML estimators of $\boldsymbol{\delta}$ is given by (23) and for the ML estimators of $\boldsymbol{\delta}$ by (24) where $g_1(\hat{\boldsymbol{\delta}})$, $g_2(\hat{\boldsymbol{\delta}})$, $g_3^*(\hat{\boldsymbol{\delta}})$ are given by (28), (29), (30) respecively where $\boldsymbol{\delta}$ is replaced by $\hat{\boldsymbol{\delta}}$, $\mathbf{I}_{\boldsymbol{\delta}}^{-1} = \begin{bmatrix} I_{vv}^{(-1)} & I_{ve}^{(-1)} \\ I_{ve}^{(-1)} & I_{ee}^{(-1)} \end{bmatrix}$, where $I_{vv}^{(-1)}$, $I_{ve}^{(-1)}$, $I_{ee}^{(-1)}$ are given by (31), (32), (33) respectively, $col_{1 \le k \le q} tr \left[\mathbf{I}_{\boldsymbol{\beta}}^{-1}(\hat{\boldsymbol{\delta}}) \frac{\partial}{\partial \delta_k} \mathbf{I}_{\boldsymbol{\beta}}(\hat{\boldsymbol{\delta}}) \right]$ and $\frac{\partial g_1(\hat{\boldsymbol{\delta}})}{\partial \boldsymbol{\delta}}$ are given by

$$col_{1\le k\le q} tr\left[\mathbf{I}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\delta})\frac{\partial}{\partial \delta_k}\mathbf{I}_{\boldsymbol{\beta}}(\boldsymbol{\delta})\right]=$$

$$=-\begin{bmatrix}\sum_{d=1}^{D} tr\left(\sum_{i=1}^{n_d} b_{id}^{-1}\mathbf{X}_{\mathbf{sid}}^{\mathbf{T}}\mathbf{X}_{\mathbf{sid}}\right)^{-1}\left(\sum_{i=1}^{n_d} b_{id}^{-2}\mathbf{X}_{\mathbf{sid}}^{\mathbf{T}}\mathbf{X}_{\mathbf{sid}}\right)\\ \sum_{d=1}^{D} tr\left(\sum_{i=1}^{n_d} b_{id}^{-1}\mathbf{X}_{\mathbf{sid}}^{\mathbf{T}}\mathbf{X}_{\mathbf{sid}}\right)^{-1}\left(\sum_{i=1}^{n_d} b_{id}^{-2} m_{id}\mathbf{X}_{\mathbf{sid}}^{\mathbf{T}}\mathbf{X}_{\mathbf{sid}}\right)\end{bmatrix} \tag{34}$$

and

$$\frac{\partial g_1(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}}=\left[\frac{\partial g_1(\boldsymbol{\delta})}{\partial \sigma_e^2}\quad \frac{\partial g_1(\boldsymbol{\delta})}{\partial \sigma_v^2}\right]^T=\begin{bmatrix} N_{rd^*t^*}-\sigma_v^4\sum_{i=1}^{N_{rd^*t^*}} b_{id^*}^{-2} m_{id^*}\\ N_{rd^*t^*}+\sigma_v^4\sum_{i=1}^{N_{rd^*t^*}} b_{id^*}^{-1} m_{id^*}\left(2-b_{id^*}^{-1} m_{id^*}\right)\end{bmatrix} \tag{35}$$

respectively, where $\boldsymbol{\delta}$ is replaced by $\hat{\boldsymbol{\delta}}$.

Under assumptions (11) and (12) the equations presented above remain true but $\hat{\boldsymbol{\beta}}_{\mathbf{d}^*}$ in (27) should be replaced by

$$\hat{\boldsymbol{\beta}}=\left(\sum_{d=1}^{D}\sum_{i=1}^{n_d} b_{id}^{-1}\mathbf{X}_{\mathbf{sid}}^{\mathbf{T}}\mathbf{X}_{\mathbf{sid}}\right)^{-1}\left(\sum_{d=1}^{D}\sum_{i=1}^{n_d} b_{id}^{-1}\mathbf{X}_{\mathbf{sid}}^{\mathbf{T}}\mathbf{Y}_{\mathbf{sid}}\right),$$

and $g_2(\boldsymbol{\delta})$ given by (29) should be replaced by

$$g_2(\boldsymbol{\delta})=\left(\sum_{i=1}^{N_{rd^*t^*}}\mathbf{x}_{\mathbf{id^*t^*}}-\sigma_v^2\sum_{i=1}^{N_{rd^*t^*}} b_{id^*}^{-1}\sum_{j=1}^{m_{id^*}}\mathbf{x}_{\mathbf{id^*j}}\right)\left(\sum_{d=1}^{D}\sum_{i=1}^{n_d} b_{id}^{-1}\mathbf{X}_{\mathbf{sid}}^{\mathbf{T}}\mathbf{X}_{\mathbf{sid}}\right)^{-1}\times$$

$$\times\left(\sum_{i=1}^{N_{rd^*t^*}}\mathbf{x}_{\mathbf{id^*t^*}}-\sigma_v^2\sum_{i=1}^{N_{rd^*t^*}} b_{id^*}^{-1}\sum_{j=1}^{m_{id^*}}\mathbf{x}_{\mathbf{id^*j}}\right)^{\mathbf{T}}$$

and $col_{1\le k\le q} tr\left[\mathbf{I}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\delta})\frac{\partial}{\partial\delta_k}\mathbf{I}_{\boldsymbol{\beta}}(\boldsymbol{\delta})\right]$ given by (34) should be replaced by:

$$col_{1\le k\le q} tr\left[\mathbf{I}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\delta})\frac{\partial}{\partial\delta_k}\mathbf{I}_{\boldsymbol{\beta}}(\boldsymbol{\delta})\right] =$$

$$= -\begin{bmatrix} tr\left(\left(\sum_{d=1}^{D}\sum_{i=1}^{n_d} b_{id}^{-1}\mathbf{X}_{\mathbf{sid}}^{\mathbf{T}}\mathbf{X}_{\mathbf{sid}}\right)^{-1}\sum_{d=1}^{D}\sum_{i=1}^{n_d} b_{id}^{-2}\mathbf{X}_{\mathbf{sid}}^{\mathbf{T}}\mathbf{X}_{\mathbf{sid}}\right) \\ tr\left(\left(\sum_{d=1}^{D}\sum_{i=1}^{n_d} b_{id}^{-1}\mathbf{X}_{\mathbf{sid}}^{\mathbf{T}}\mathbf{X}_{\mathbf{sid}}\right)^{-1}\sum_{d=1}^{D}\sum_{i=1}^{n_d} b_{id}^{-2}m_{id}\mathbf{X}_{\mathbf{sid}}^{\mathbf{T}}\mathbf{X}_{\mathbf{sid}}\right) \end{bmatrix}$$

7. Simulation analyses

The limited Monte Carlo simulation analyses are based on real data on N = 314 Polish poviats (what is NUTS 4 level) excluding cites with poviat's rights for M = 4 years 2005-2008. Data are available at the website of the Polish Central Staistical Office – www.stat.gov.pl. The problem is to estimate subpopulations (domains) totals for D = 6 regions (NTS 1 level) in 2008. The variable of interest is poviats' own incomes (in PLN) and the auxiliary variable is the population size in poviats (in persons). Two simulations are conducted using R (R Development Core Team [2011]). In the simulations the accuracy of the proposed predictor is compared with accuracies of two calibration estimators [Rao 2003, pp. 17-18] which will be denoted by GREG1 and GREG2. Both calibration estimators are of the form $\hat{\theta}_{d^*t^*}^{GREG} = \sum_{i\in s_{d^*t^*}} w_{sit^*} y_{i^*t^*}$ but weights w_{sit^*} are solutions for GREG1 of

$$\begin{cases} \sum_{i\in s_{d^*t^*}} \dfrac{(w_{sit^*} - 1/\pi_{it^*})^2}{1/\pi_{it^*}} \to \min \\ \sum_{i\in s_{d^*t^*}} w_{sit^*}\mathbf{x}_{\mathbf{id^*t^*}} = \sum_{i\in\Omega_{d^*t^*}} \mathbf{x}_{\mathbf{id^*t^*}} \end{cases}$$

and for GREG2 are subsequence of weights obtained as a solution of

$$\begin{cases} \sum_{i \in s_{t^*}} \frac{(w_{sit^*} - 1/\pi_{it^*})^2}{1/\pi_{it^*}} \rightarrow \min \\ \sum_{i \in s_{t^*}} w_{sit^*} \mathbf{x}_{\mathbf{id^*t^*}} = \sum_{i \in \Omega_{t^*}} \mathbf{x}_{\mathbf{id^*t^*}} \end{cases}$$

where π_{it^*} are inclusion probabilities in the period t^*. These calibration estimators are classic model-assisted estimators which are known as good alternatives for model-based methods especially in the case of possible model misspecification.

The first simulation is design-based. In this case a sample of size n = 79 elements (c.a. 25% of population size) is balanced panel sample (each sampled element is observed in all of 4 periods), which is drawn at random in the first period with inclusion probabilities proportional to the value of the auxiliary variable in this period. For this sample size it was possible to estimate all of domain totals in each iteration even using direct estimators GREG1 and GREG2. The number of samples drawn in the simulation equals 10 000.

The second simulation is model based. In this case one sample is drawn using the sample design described above what gives the division of the population into the sampled and unsampled part. Then 10 000 populations are generated using model (11) – (with one auxiliary variable and constant) with parameters computed (REML) based on real, whole population data and random components with the following distributions: in the model denoted in the simulation as N case – normal distributions of both random components, U case – uniform distributions of both random components and E case – "shifted" exponential distribution of both random components. What is more, to study the problem of model misspecification, equations for linear model are used but 10 000 population are generated based on modified model (11) where instead of the auxiliary variable its natural logarithm is used. Both random components have the following distributions: Nm case – normal distribution, Um – uniform distribution and Em – "shifted" exponential distribution.

What is important, the predictor presented for the model (11) simplifies to the BLUP (i.e. does not depend on unknown variance parameters) for the balanced sample. Hence, in the equation of the MSE estimator the $g_3^*(\boldsymbol{\delta})$ element is omitted.

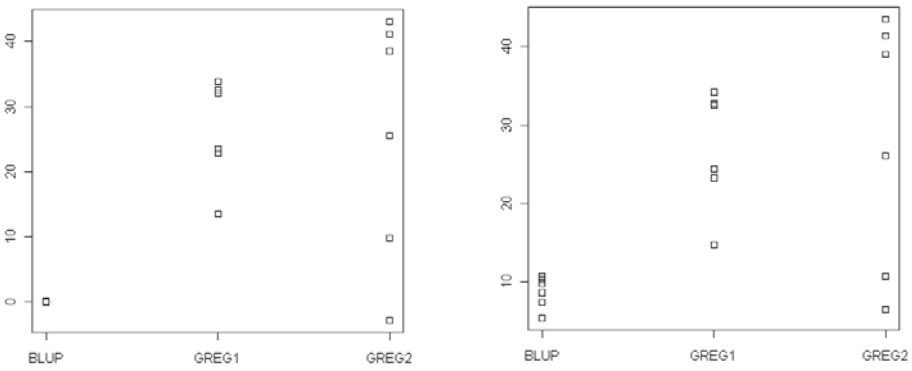


Fig. 1. Relative model-biases (on the left) and relative model RMSE (on the right) for N case (in %) for six domains

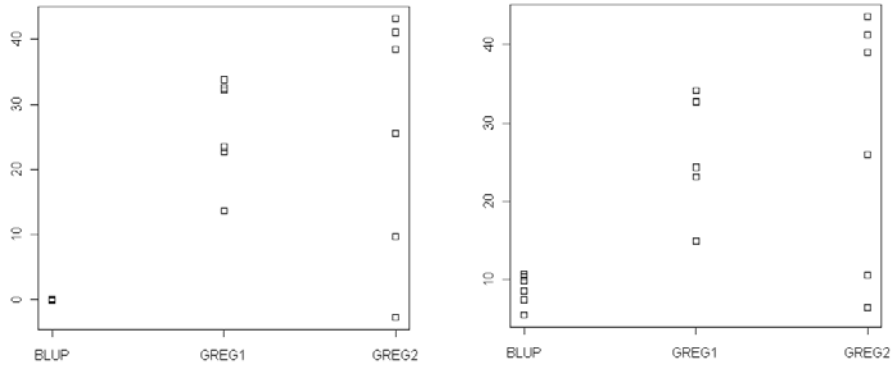


Fig. 2. Relative model-biases (on the left) and relative model RMSE (on the right) for U case (in %) for six domains

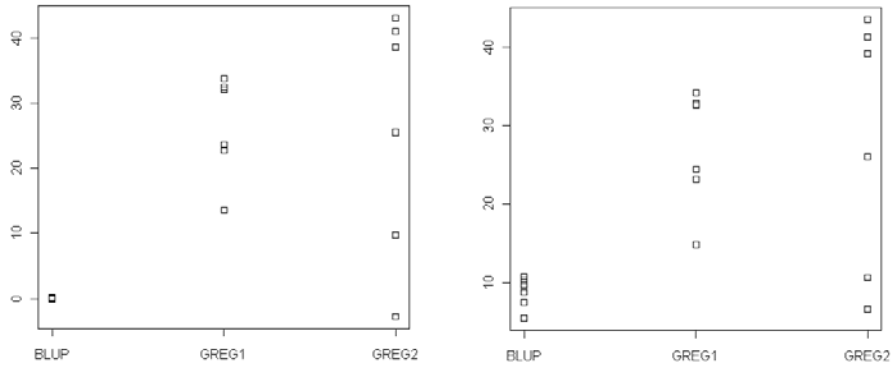


Fig. 3. Relative model-biases (on the left) and relative model RMSE (on the right) for E case (in %) for six domains

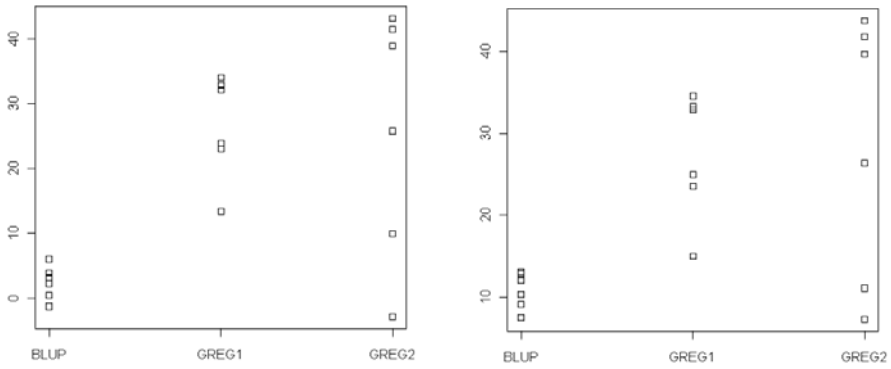


Fig. 4. Relative model-biases (on the left) and relative model RMSE (on the right) for Nm case (in %) for six domains

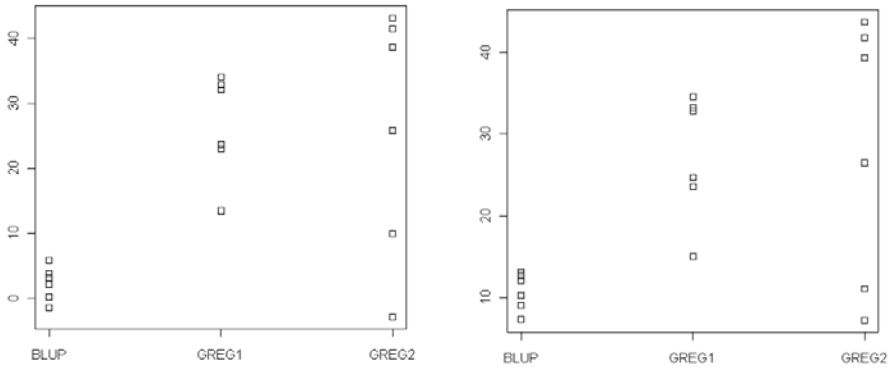


Fig. 5. Relative model-biases (on the left) and relative model RMSE (on the right) for Um case (in %) for six domains

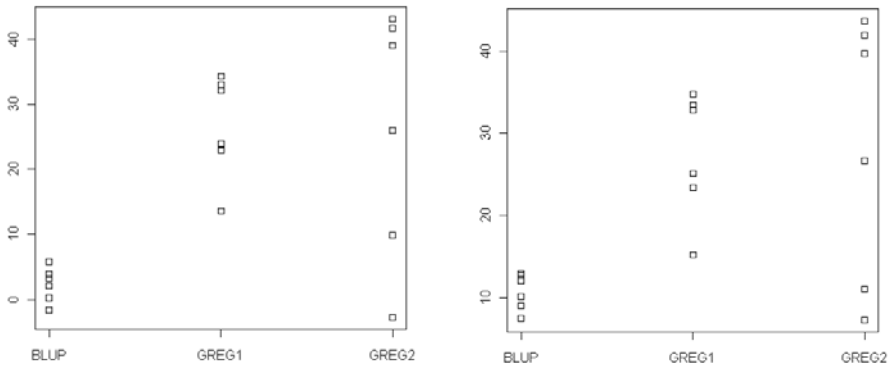


Fig. 6. Relative model-biases (on the left) and relative model RMSE (on the right) for Em case (in %) for six domains

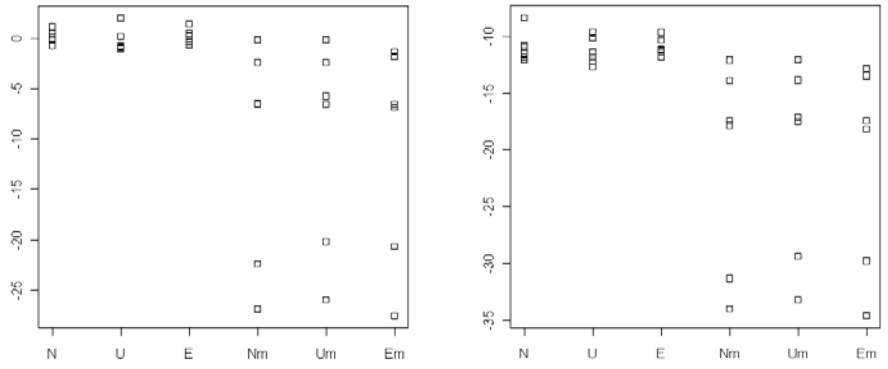


Fig. 7. Relative biases of MSE estimators of BLUP for REML (on the left) and ML (on the right) estimates of $\boldsymbol{\delta}$ (in %) for six domains

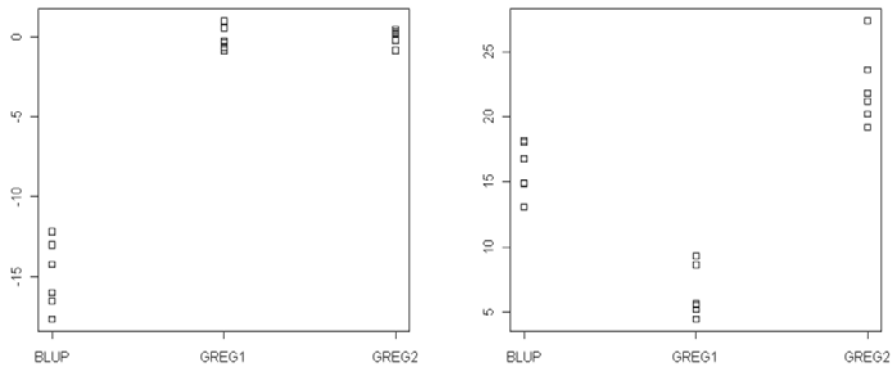


Fig. 8. Relative design-biases (on the left) and relative design RMSE (on the right) for six domains in %

Each point on the figures presents value of some statistic for one out of $D = 6$ domains. Comparing prediction accuracy of the BLUP and GREG1 and GREG2 (see Figures 1-6) it should be noted that the BLUP is better then both GREG estimators even in the cases of model misspecification. The absolute value of its relative bias does not exceed 10% for all of the considered cases of model misspecification (see Figures 4-6). The bias of the considered MSE estimator under normality is $o(D^{-1})$ (as proved in Żądło [2009a]) but for the data interesting case is studied – the number of domains is very small, it equals $D = 6$. It is known that for the big number of domains D the differences between

biases of REML and ML MSE esimators (given by general equations (23) and (24) respectively) are small. In the simulation study the REML MSE estimator is (see Figure 7) less biased in all of the considered cases, what may have occured due to the relatively big (comparing with D) loss of degrees of freedom when ML method is used instead of REML to estimate $\boldsymbol{\delta}$. Let us limit further consideration to the REML MSE estimator. The absolute values of relative biases of MSE estimators (see Figure 7) are small not only under normality assumptions (N case), under which they were derived, but also for U and E cases. For the proof of robustness of some MSE estimators of the EBLUP of the form of Henderson's BLUP in some cases of model misspecification see Lahiri and Rao [1995]. The maximum absolute value of the relative biases of MSE estimators are less then 2,1% for these cases. When the true model is non-linear (the Nm, Um, Em cases) the biases obtained in the simulation are larger.

Results presented on the Figure 8 show that the design bias of the BLUP is larger than the design bias of both GREG1 and GREG2. Comparing the design accuracy for the data and large sample size (n = 79), the BLUP is better than GREG2 but worse than GREG1.

Conclusions

In the paper the EBLUP for longitudinal data is proposed. The predictor allows to predict the domain total for any (past, current, future) period assuming that population and domain affiliation of population elements may change in time. Its MSE is also derived and some MSE estimator is proposed. Its accuracy is analyzed for real data in the Monte Carlo simulation study.

Literature

Battese G.E., Harter R.M., and Fuller W.A. (1988): *An Error-components Model for Prediction of County Crop Areas Using Survey and Satellite Data*. "Journal of the American Statistical Association", No. 83.

Datta G.S., Lahiri P. (2000): *A Unified Measure of Uncertainty of Estimated Best Linear Unbiased Predictors in Small Area Estimation Problems*. "Statistica Sinica", No. 10.

Dempster A.P., Rubin D.B., Tsutakawa R.K. (1981): *Estimation in Covariance Components Models*. "Journal of the American Statistical Association", Vol. 76, No. 374.

Hedeker D., Gibbons R.D. (2006): *Longitudinal Data Analysis.* John Wiley and Sons, New Jersey.

Henderson C.R. (1950): *Estimation of Genetic Parameters (Abstract)*. "Annals of Mathematical Statistics", No. 21.

Kackar R.N., Harville D.A. (1981): *Unbiasedness of Two-stage Estimation and Prediction Procedures for Mixed Linear Models*. "Communications in Statistics" Ser. A,10.

Lahiri P., Rao J.N.K. (1995): *Robust Estimation of Mean Squared Error of Small Area Estimators*. "Journal of the American Statistical Association", No. 90.

Moura F.A.S., Holt D. (1999): *Small Area Estimation Using Multilevel Models.* "Survey Methodology", No. 25.

Prasad N.G.N, Rao J.N.K. (1990): *The Estimation of Mean the Mean Squared Error of Small Area Estimators*. "Journal of the American Statistical Association", No. 85.

R Development Core Team (2011): *A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna.

Rao J.N.K. (2003): *Small Area Estimation*. John Wiley and Sons, New York.

Rao J.N.K., Yu M. (1994): *Small-area Estimation by Combining Time-series and Cross-sectional Data*. "The Canadian Journal of Statistics", No. 22.

Royall R.M. (1976): *The Linear Least Squares Prediction Approach to Two-stage Sampling.* "Journal of the American Statistical Association", No. 71.

Verbeke G., Molenberghs G. (2000): *Linear Mixed Models for Longitudinal Data.* Springer, New York.

Żądło T. (2004): *On Unbiasedness of Some EBLU Predictor*. In: *Proceedings in Computational Satistics 2004.* Ed. J. Antoch. Physica-Verlag, Heidelberg-New York.

Żądło T. (2009a): *On MSE of EBLUP.* Statistical Papers. Springer, 50.

Żądło T. (2009b): *On Prediction of Domain Totals Based on Unbalanced Longitudinal Data*. In: *Survey Sampling in Economic and Social Research.* Eds. J. Wywiał, T. Żądło. Wydawnictwo AE, Katowice.

O PEWNYCH PROBLEMACH PREDYKCJI WARTOŚCI GLOBALNEJ W DOMENIE W BADANIACH WIELOOKRESOWYCH, GDY SĄ DOSTĘPNE INFORMACJE O ZMIENNYCH DODATKOWYCH

Streszczenie

W artykule wyprowadzono postacie najlepszych liniowych nieobciążonych predyktorów przy założeniu pewnych modeli będących uogólnieniami na przypadek danych

przekrojowo-czasowych modeli znanych z literatury statystyki małych obszarów. Ponadto wyprowadzono postacie błędów średniokwadratowych empirycznych wersji tych predyktorów oraz zaproponowano ich estymatory. W symulacji Monte Carlo porównywano dokładność zaproponowanego predyktora z dwoma ogólnymi estymatorami regresyjnymi po planie losowania i po modelu nadpopulacji (także w różnych przypadkach złej specyfikacji modelu). Ponadto analizowano obciążenia zaproponowanych estymatorów błędu średniokwadratowego.

Grzegorz Kończak
Magdalena Chmielińska
Uniwersytet Ekonomiczny w Katowicach

ZASTOSOWANIE METOD SYMULACYJNYCH W ANALIZIE WIELOWYMIAROWYCH TABLIC WIELODZIELCZYCH

Wprowadzenie

W ostatnich latach w badaniach naukowych znacznie wzrosła rola metod symulacji komputerowej. Początki tych metod sięgają lat 40. XX w. Właśnie wtedy zespół pracujący w amerykańskiej bazie wojskowej w Los Alamos kierowany przez J. von Neumana zastosował metody Monte Carlo do symulacji funkcjonowania elektrowni jądrowych oraz wybuchów atomowych. Ze względu na brak możliwości zastosowania złożonych technik obliczeniowych dynamiczny wzrost zastosowań tych metod przypada jednak dopiero na końcowe lata XX w. Do najczęściej wykorzystywanych metod symulacyjnych w badaniach statystycznych należy zaliczyć bootstrap, algorytmy Gibbsa i Metropolisa oraz próbkowanie ważone. Wszystkie wymienione metody zasadniczo są związane z analizą danych o charakterze ilościowym. Stosunkowo rzadko metody symulacyjne są wykorzystywane do analizy danych o charakterze jakościowym, czyli danych rejestrowanych na skalach nominalnych lub porządkowych. W opracowaniu przedstawiono propozycję metody symulacyjnej pozwalającej na zbadanie zależności dla grupy zmiennych nominalnych. Proponowana metoda wykorzystuje test permutacyjny. Przeprowadzono również porównania otrzymanych rezultatów z wynikami klasycznego testu niezależności chi-kwadrat. W przedstawionych przykładach zastosowano proponowaną metodę do analizy danych pochodzących z Polskiego Generalnego Sondażu Społecznego[1].

[1] Badania prowadzone od 1992 r. w Instytucie Studiów Społecznych Uniwersytetu Warszawskiego (http://pgss.iss.uw.edu.pl).

1. Wnioskowanie statystyczne dla danych w tablicach wielodzielczych

Załóżmy, że badaniu poddano *n* jednostek ze względu na dwie cechy nominalne *X* i *Y* przyjmujące odpowiednio *r* oraz *s* wartości (wariantów cech). Wyniki takiego badania mogą zostać zapisane w tablicy kontyngencji o *r* wierszach oraz *s* kolumnach. Do sprawdzenia hipotezy o niezależności tych zmiennych można wykorzystać statystykę [Aczel 2000]:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \tag{1}$$

gdzie:
O_{ij} – liczebności obserwowane,
E_{ij} – liczebności oczekiwane.

Na podstawie otrzymanych danych dwuwymiarowych można skonstruować tablicę wielodzielczą. Przykład tablicy dla danych pochodzących z badania Polski Generalny Sondaż Społeczny przedstawia tabela 1. Przedstawiono w niej rozkład odpowiedzi na pytanie „jak ważna jest własna rodzina i dzieci?" w zależności od stanu cywilnego ankietowanych. Cała tabela składa się z 35 komórek. W tabeli 1 dodatkowo szarym tłem wyróżniono komórki tablicy, dla których liczebności oczekiwane są mniejsze niż 5.

Statystyka (1) przy założeniu niezależności zmiennych *X* i *Y* ma asymptotycznie rozkład chi-kwadrat o $(r-1)\cdot(s-1)$ stopniach swobody. Wnioskowanie statystyczne jest uzasadnione, jeśli liczebności oczekiwane są nie mniejsze niż 5 we wszystkich komórkach tablicy wielodzielczej [Blalock 1975]. W tabeli 1 występuje aż 14 komórek o liczebnościach oczekiwanych mniejszych od 5. Jest to istotną przeszkodą w wykorzystaniu testu niezależności chi-kwadrat. W takich przypadkach zwykle łączy się wybrane klasy. Połączenie wierszy „rozwiedziony" oraz „separacja" zmniejszy rozmiar tablicy wielodzielczej i liczbę komórek z liczebnościami oczekiwanymi mniejszymi od 5 do 8. Wyróżnienie dwóch kategorii wagi rodziny: „nieważna" (wskazania 1-4) oraz „ważna" (wskazania 5-7) pozwoliłoby na uniknięcie komórek z liczebnościami oczekiwanymi mniejszymi od 5. Takie rozwiązanie jest jednak związane ze stratą informacji o natężeniu roli przypisywanej przez respondentów rodzinie.

Tabela 1

Jak ważna jest własna rodzina i dzieci w zależności od stanu cywilnego respondenta

Stan cywilny	Jak ważna jest własna rodzina i dzieci?						
	nieważne						bardzo ważne
	1	2	3	4	5	6	7
zamężna/żonaty/konkubinat	13	5	3	30	54	205	5609
wdowiec/wdowa	5	1	2	8	16	58	997
rozwiedziony(a)	3	0	3	6	15	21	301
separacja	1	0	1	2	2	4	80
kawaler/panna	20	6	12	38	71	152	1043

Źródło: Na podstawie danych z PGSS.

Powyższe rozważania można rozszerzyć na większą niż dwie liczbę zmiennych X_1, X_2, ..., X_h. W takim przypadku mamy do czynienia z tablicami wielodzielczymi wielowymiarowymi. W przypadku trzech zmiennych nominalnych ($h = 3$) przyjmujących odpowiednio r, s oraz t wartości tablica wielodzielcza jest faktycznie kostką trójwymiarową. W tym przypadku statystyka testowa przyjmie następującą postać [Sheskin 2004]:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t}\frac{(O_{ijk} - E_{ijk})^2}{E_{ijk}}, \tag{2}$$

gdzie:
O_{ijk} – liczebności obserwowane,
E_{ijk} – liczebności oczekiwane.

Dla niezależnych zmiennych statystyka (2) ma asymptotyczny rozkład chi-kwadrat o $rst - r - s - t + 2$ stopniach swobody. Zapis tablicy dla trzech wymiarów nie jest już tak naturalny jak w przypadku tablicy dwuwymiarowej. Dla danych zamieszczonych w tabeli 1 uwzględniając dodatkowo zmienną „płeć" odpowiednia tablica wielodzielcza miałaby dwie warstwy. Poszczególne warstwy takiej tablicy mogą zostać przedstawione w oddzielnych tablicach dwuwymiarowych. Wyniki można jednak również przedstawić w formie jak w tabeli 2.

W miarę wzrostu liczby zmiennych h i liczby kategorii w_i (i = 1, 2, ..., h) dla zmiennych coraz trudniej zapewnić spełnienie warunku, aby liczebności oczekiwane w każdej komórce tabeli wynosiły przynajmniej 5. W tablicy wielodzielczej uwzględniającej płeć respondenta (por. tabela 2) jest 70 komórek. Ponad połowa z nich (38 komórek) ma liczebności oczekiwane mniejsze od 5. Biorąc pod uwagę fakt, że badana próba liczy 8787 osób łatwo zauważyć, że dla mniejszych prób często praktycznie nie będzie możliwości odwołania się do testu niezależności chi-kwadrat.

Tabela 2

Tablica wielodzielcza dla trzech zmiennych klasyfikujących

Jak ważna jest własna rodzina i dzieci w zależności od stanu cywilnego i płci respondenta

Stan cywilny	Jak ważna jest własna rodzina i dzieci?													
	Kobieta							Mężczyzna						
	nieważna				bardzo ważna			nieważna				bardzo ważna		
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
zamężna/ żonaty/ konkubinat	11	2	2	13	17	80	2915	2	3	1	17	37	125	2694
wdowiec /wdowa	3	1	2	7	10	46	841	2	0	0	1	6	12	156
rozwie-dziony(a)	1	0	1	2	8	8	222	2	0	2	4	7	13	79
separacja	0	0	0	0	1	1	58	1	0	1	2	1	3	22
kawaler/ panna	9	4	8	15	27	58	501	11	2	4	23	44	94	542

Źródło: Na podstawie danych z PGSS.

Jeżeli dla każdej ze zmiennych liczba kategorii jest jednakowa i wynosi w ($w_1 = w_2 = ... = w_h = w$), to liczba komórek w wielowymiarowej tablicy wielodzielczej wynosi w^h. Biorąc pod uwagę, że w każdej z tych kratek liczebność oczekiwana powinna wynosić przynajmniej 5, to minimalna liczebność próby wynosi $5w^h$. W tabeli 3 przedstawiono przykładowe wartości minimalnych liczebności próby dla ustalonej ilości zmiennych i wariantów dla każdej cechy.

Tabela 3

Minimalne liczebności dla wybranych rozmiarów tablic

Liczba zmiennych h	Liczba wariantów każdej zmiennej w			
	2	3	4	5
2	20	45	80	125
3	40	135	320	625
4	80	405	1280	3125
5	160	1215	5120	15625

Minimalne liczebności przedstawione w tabeli 3 dotyczą szczególnego przypadku, gdy realizacja każdego z wariantów dla wszystkich zmiennych jest jednakowo prawdopodobna. Jeżeli prawdopodobieństwa realizacji dla różnych kategorii nie są jednakowe, to minimalne liczebności próby będą większe niż przedstawione w tabeli 3. Na podstawie danych zamieszczonych w tabeli 3 widoczne jest, że już dla kilku zmiennych przy paru różnych wariantach każdej z cech praktycznie niemożliwe jest przeprowadzenie badania oraz otrzymania uogólnienia na całą populację poprzez bezpośrednie wykorzystanie testu niezależności chi-kwadrat.

Ze względu na potrzeby praktyczne analizy danych w tablicach wielodzielczych przy niespełnieniu warunku na liczebność oczekiwaną zaproponowano różne modyfikacje. W dużej mierze modyfikacje te dotyczą przypadku tablic o wymiarach 2 x 2. Jeżeli nie wszystkie liczebności oczekiwane wynoszą przynajmniej 5, to można wykorzystać modyfikacje statystyki chi-kwadrat uwzględniające poprawki na ciągłość Yatesa lub poprawkę Dandekara [Rao 1982]. Statystyka chi-kwadrat z poprawką Yatesa ma postać:

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{\left(\left|O_{ij} - E_{ij}\right| - 0{,}5\right)^2}{E_{ij}}. \tag{3}$$

Postać statystyki zaproponowanej przez Dandekara jest następująca:

$$\chi_c^2 = \chi_0^2 - \frac{\chi_0^2 - \chi_{-1}^2}{\chi_1^2 - \chi_{-1}^2}(\chi_1^2 - \chi_0^2), \tag{4}$$

gdzie χ_0^2, χ_1^2, χ_{-1}^2 oznaczają wartości statystyki (1) wyznaczone po dodaniu odpowiednio 0, 1 lub -1 do liczebności n_{11} (pierwszy wiersz i pierwsza kolumna tablicy wielodzielczej).

Powyższe korekty nie powinny być jednak stosowane w przypadkach, gdy w więcej niż jednej komórce tablicy wielodzielczej liczebności oczekiwane są mniejsze od 5. Wyjściem w takich sytuacjach jest zastosowanie testu dokładnego Fishera [Agresti 1996]. Test ten, podobnie jak powyższe modyfikacje, dotyczy tablic o wymiarach 2 x 2. W tym teście są obliczane prawdopodobieństwa wystąpienia wszystkich możliwych układów liczebności w tablicy wielodzielczej przy zachowaniu ustalonych liczebności brzegowych. Rozważmy przypadek otrzymania wyników badania jak w tabeli 4.

Tabela 4

Przykład danych w tablicy o wymiarach 2*x*2

Zmienna X	Zmienna Y	
	Y_1	Y_2
X_1	4	0
X_2	0	4

Wprowadźmy oznaczenia dla poszczególnych liczebności w komórkach tabeli 4: $a = 4$, $b = 0$, $c = 0$, $d = 4$. Niech ponadto $r_1 = 4$, $r_2 = 4$ oznaczają liczebności brzegowe w wierszach, a $c_1 = 4$ oraz $c_2 = 4$ liczebności brzegowe w kolumnach. Przy założeniu ustalonych powyższych liczebności brzegowych wszystkich możliwych tablic wielodzielczych jest 5 (por. rys. 1).

4	0
0	4

3	1
1	3

2	2
2	2

1	3
3	1

0	4
4	0

Rys. 1. Wszystkie tablice wielodzielcze o liczebnościach brzegowych $r1 = r_2 = c_1 = c_2 = 4$

Uwzględniając jednak, które elementy zostały zakwalifikowane do poszczególnych komórek tablicy wielodzielczej liczbę wszystkich możliwych układów tworzących taką tablicę, można wyrazić następująco:

$$K = \sum_{x=0}^{4} \binom{4}{x} \cdot \binom{4}{4-x} = 70.$$

Prawdopodobieństwo wystąpienia realizacji tablicy wielodzielczej o kolejnych liczebnościach a, b, c i d w komórkach wyraża się wzorem:

$$p = \frac{r1!r2!c1!c2!}{n!a!b!c!d!}. \tag{5}$$

Wszystkie możliwe wartości liczebności a (przy ustalonym a i danych liczebnościach brzegowych wartości b, c i d są wyznaczone jednoznacznie) dla zmiennych niezależnych z przykładu z tabeli 4 przy zachowaniu liczebności brzegowych podaje tabela 5.

Tabela 5

Prawdopodobieństwa i wartości statystyki chi-kwadrat

a	p	χ^2
4	0,014	8,00
3	0,229	2,00
2	0,514	0,00
1	0,229	2,00
0	0,014	8,00

Wysokie wartości statystyki χ^2 świadczą przeciw hipotezie o niezależności zmiennych. Dla przedstawionego w tabeli 4 przypadku wartość statystyki χ^2 wynosi 8. Przy założeniu niezależności zmiennych prawdopodobieństwo wystąpienia tak dużej lub większej wartości wynosi 0,028. Dla poziomu istotności $\alpha = 0{,}05$ należy odrzucić hipotezę o niezależności zmiennych.

2. Weryfikacja hipotezy o niezależności h ($h > 2$) zmiennych nominalnych

Weryfikacja hipotezy o niezależności $h = 3$ zmiennych może być przeprowadzona z wykorzystaniem statystyki (2). W przypadku większej liczby zmiennych wzór (2) należy uzupełnić o odpowiednie sumy i indeksy liczebności. Po-

ważnym problemem jest zapewnienie odpowiednich liczebności oczekiwanych w komórkach nawet dla prób o stosunkowo dużych liczebnościach. W takich przypadkach skuteczna może być procedura wykorzystująca test permutacyjny [Good 2005]. Na podstawie wylosowanej próby jest obliczana wartość statystyki testowej T_0. Jako statystykę można przyjąć np. χ^2. Następnie zbiór danych jest N-1 razy permutowany i każdorazowo jest wyznaczana wartość statystyki T_i (i = 1, 2, ..., N-1). Wartość statystyki T_0 jest porównywana z kwantylem rzędu $1-\alpha$ empirycznego rozkładu statystyki T.

2.1. Procedura obliczeniowa

Przebieg procedury obliczeniowej zostanie przedstawiony dla przypadku tablicy trójwymiarowej. Wszystkie rozważania można w analogiczny sposób stosować do tablic h-wymiarowych. W obliczeniach zostanie wykorzystany test permutacyjny. Struktura analizowanych danych została schematycznie przedstawiona na rys. 2. X_{i1}, X_{i2}, …X_{in} oznaczają warianty zmiennej X. Odpowiednio zostały oznaczone warianty zmiennych Y i Z.

X	Y	Z
X_{i1}	Y_{i1}	Z_{i1}
X_{i2}	Y_{i2}	Z_{i2}
…	…	…
X_{in}	Y_{in}	Z_{in}

Rys. 2. Struktura danych wykorzystana w opisie algorytmu

Procedura algorytmu zastosowania testu permutacyjnego do analizy tablic wielodzielczych wielowymiarowych jest następująca:

1. Pobierana jest próbka losowa. Na podstawie próby losowej jest konstruowana tablica wielodzielcza.
2. Dla otrzymanej tablicy wielodzielczej jest obliczana wartość statystyki chi-kwadrat (2). Otrzymaną wartość oznaczmy przez T_0.
3. Dla pobranej próbki kolumny 2-3 (por. rys. 2) są losowo, niezależnie permutowane. Dla tak otrzymanej próby jest obliczana wartość statystyki chi-kwadrat.

4. Krok 3 jest wykonywany N razy. Otrzymujemy wartości statystyki T_1, T_2, ..., T_N.
5. Obliczana jest wartość *ASL* (ang. *achieving significance level*) [Efron i Tibshirani 1993]:

$$ASL = \frac{card(i : T_i \geq T_0)}{N}, \text{ gdzie } i = 0, 1, ..., N\text{-}1. \quad (6)$$

Jeżeli *ASL* jest mniejsze od przyjętego poziomu istotności α, to odrzucamy hipotezę H_0.

Przedstawiona procedura pozwala na weryfikację hipotez statystycznych dotyczących zależności pomiędzy pewną liczbą zmiennych nominalnych. Procedura symulacyjna pozwala na sprawne przeprowadzenie weryfikacji hipotezy nawet dla tablic wielodzielczych o bardzo dużych rozmiarach również w przypadku stosunkowo niewielkiej liczby obserwacji.

2.2. Wyniki analizy symulacyjnej

Dla porównania własności opisywanego testu permutacyjnego i klasycznego testu chi-kwadrat przeprowadzono analizy na podstawie danych pochodzących z Polskiego Generalnego Sondażu Społecznego (PGSS). Celem Polskiego Generalnego Sondażu Społecznego jest systematyczny pomiar trendów i skutków zmian społecznych w Polsce. Problematyka PGSS obejmuje badanie indywidualnych postaw, cenionych wartości, orientacji i zachowań społecznych, jak również pomiar zróżnicowania społeczno-demograficznego, zawodowego, edukacyjnego i ekonomicznego reprezentatywnych grup i warstw społecznych w Polsce [Cichomski et al. 2009]. Systematyczną analizę trendów społecznych umożliwia cykliczne powtarzanie badań, które zachowują porównywalne standardy metodologiczne i identyczne wskaźniki.

W tabeli 6 przedstawiono analizowane zbiory zmiennych oraz wartości statystyki chi-kwadrat. Zamieszczono również p-wartość dla testu chi-kwadrat. Wielkości te należy traktować wyłącznie orientacyjnie ze względu na niespełnienie założenia dotyczącego minimalnych wartości liczebności oczekiwanych w komórkach. W tabeli 6 przedstawiono również wartości *ASL* dla zastosowanego testu permutacyjnego.

Tabela 6

Charakterystyka analizowanych zmiennych i wyniki obliczeń

Lp.	Zmienne	Liczba komórek tablicy *rst*	Liczebność próby	Test chi-kwadrat		Test permutacyjny
				χ^2	*p* wartość	*ASL*
1	X_2, X_3	35	8787	459,7	*)	*)
2	X_1, X_2, X_3	70	8787	1046,5	*)	*)
3	X_4, X_5	30	15934	632,1	*)	*)
4	X_1, X_4, X_5	60	15924	734,0	*)	*)
5	X_6, X_7	40	2445	15,6	0,971	0,974
6	X_2, X_6, X_7	200	2435	193,7	0,297	0,318

*) – wartość mniejsza lub równa 0,001
X_1 – płeć (2 wartości: mężczyzna, kobieta)
X_2 – stan cywilny (5 wartości: żonaty/zamężna/konkubinat, wdowiec/wdowa, rozwiedziony(a), separacja, kawaler/panna)
X_3 – jak ważna jest własna rodzina i dzieci (7 wartości: 1 – nieważne, ..., 7 – bardzo ważne)
X_4 – zadowolenie z własnej sytuacji finansowej (3 wartości: 1 – zadowolony, 2 – mniej więcej zadowolony, 3 – niezadowolony)
X_5 – skala chęci do życia (10 wartości: 1 – „w ogóle nie chce mi się żyć", ..., 10 – „bardzo mocno chce mi się żyć")
X_6 – region zamieszkania (8 wartości)
X_7 – większy % podatku od bogatych (5 wartości: 1 – „znacznie większy %, ..., 5 – „znacznie mniejszy %")

Z analizy wyników przedstawionych w tabeli 6 można wysnuć wniosek, iż decyzje odnośnie do testowanej hipotezy o niezależności zmiennych przy zastosowaniu klasycznego testu niezależności chi-kwadrat (p-wartości) oraz testu permutacyjnego (wartości *ASL*) są w analizowanych przypadkach identyczne. Wyniki otrzymane z zastosowaniem testu niezależności chi-kwadrat nie mogą jednak być przedstawiane jako wiarygodne ze względu na niespełnienie założeń dotyczących minimalnych liczebności oczekiwanych w komórkach tablicy wielodzielczej. Metoda symulacyjna nie wymaga spełnienia takich założeń i może być stosowana nawet dla prób o niewielkich liczebnościach. Prezentowana metoda symulacyjna może okazać się szczególnie przydatna w przypadku wielo-

wymiarowych tablic wielodzielczych, gdzie wnioskowanie statystyczne o zależności pomiędzy badanymi zmiennymi na podstawie klasycznego testu niezależności chi-kwadrat może okazać się niezasadne ze względu na małe liczebności oczekiwane w komórkach tablicy.

Podsumowanie

Ograniczenia klasycznych metod statystycznych sprawiają, iż metody symulacyjne są coraz chętniej wykorzystywane w badaniach naukowych. Metody symulacyjne, mimo iż obecnie wykorzystywane są głównie do analizy danych o charakterze ilościowym, z powodzeniem mogą być również stosowane do analizy danych o charakterze jakościowym. Zaprezentowana w niniejszym artykule metoda oparta na teście permutacyjnym, pozwala na weryfikację hipotezy o niezależności dla grup zmiennych nominalnych. Zaletą omawianej metody symulacyjnej jest brak założeń dotyczących minimalnych liczebności oczekiwanych w komórkach tablicy wielodzielczej. Nabiera to szczególnego znaczenia w przypadku analizy zależności pomiędzy wieloma zmiennymi, z których każda posiada po kilka wariantów wartości. W takich przypadkach duża liczba zmiennych i kategorii dla zmiennych w praktyce uniemożliwia wnioskowanie o niezależności pomiędzy tymi zmiennymi za pomocą testu niezależności chi-kwadrat.

Literatura

Aczel A. (2000): *Statystyka w zarządzaniu*. Wydawnictwo Naukowe PWN, Warszawa.

Agresti A. (1996): *An Introduction to Categorical Data Analysis*. John Wiley & Sons, New York.

Blalock H.M. (1975): *Statystyka dla socjologów*. PWN, Warszawa.

Cichomski B., Jerzyński T., Zieliński M. (2009): *Polskie Generalne Sondaże Społeczne: struktura skumulowanych wyników badań 1992-2008*. Instytut Studiów Społecznych, Uniwersytet Warszawski, Warszawa,.

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

Good P. (2005): *Permutation, Parametric and Bootstrap Tests of Hypotheses*. Springer Science Business Media, New York.

Rao C.R (1982): *Modele liniowe statystyki matematycznej*. PWN, Warszawa.

Sheskin D.J. (2004): *Handbook of Parametric and Nonparametric Statistical Procedures*. Chapman & Hall/CRC, Boca Raton.

ON THE USE OF THE MONTE CARLO METHODS IN THE MULTIDIMENSIONAL CONTINGENCY TABLES ANALYSIS

Summary

Recently the role of computer simulation methods in scientific research has significantly increased. In this paper the proposal of the use simulation methods to test for independence of some categorical variables with contingency tables is presented. The permutation test is used in this method. The comparison of obtained results and results of classically chi-square test of independence has been done. In presented examples the data from Polish General Public Opinion Poll have been used.

Dorota Rozmus
Uniwersytet Ekonomiczny w Katowicach

PORÓWNANIE STABILNOŚCI ZAGREGOWANYCH ALGORYTMÓW TAKSONOMICZNYCH OPARTYCH NA IDEI METODY BAGGING

Wprowadzenie

Pierwotnie podejście zagregowane (wielomodelowe) z dużym powodzeniem było stosowane w dyskryminacji i regresji w celu podniesienia dokładności predykcji. Zasadnicza idea tego podejścia polega na tym, że w pierwszym kroku są budowane liczne różniące się między sobą pojedyncze modele, które następnie za pomocą różnych operatorów są łączone w model zagregowany. W dyskryminacji najczęściej stosowanym operatorem jest głosowanie majoryzacyjne, co oznacza, że jest wybierana ta klasa, która najczęściej była wskazywana przez pojedyncze modele; natomiast w regresji najczęściej stosuje się uśrednianie wartości teoretycznych zmiennej y. Wśród najbardziej znanych metod agregacji należy wymienić: *bagging* [Breiman 1996], który jest oparty na losowaniu kolejnych prób boostrapowych oraz *boosting* [Freund 1999] polegający na nadawaniu wyższych wartości wag błędnie sklasyfikowanym obiektom.

W ostatnich latach analogiczne propozycje pojawiły się także w taksonomii, aby zapewnić większą poprawność i stabilność wyników grupowania [Fern i Brodley 2003; Fred 2002; Fred i Jain 2002; Strehl i Gosh 2002]. Zagadnienie agregacji w taksonomii może zostać sformułowane następująco: mając wyniki wielokrotnie przeprowadzonej klasyfikacji, znajdź zagregowany podział ostateczny o lepszej jakości. Liczne badania w tej dziedzinie ustanowiły już nowy obszar w tradycyjnej taksonomii. Istnieje wiele możliwości zastosowania idei podejścia zagregowanego w dziedzinie uczenia bez nauczyciela, wśród których jako najpopularniejsze należy wymienić:

1. Łączenie wyników grupowania uzyskanych za pomocą różnych metod.
2. Uzyskanie różniących się między sobą klasyfikacji z zastosowaniem różnych podzbiorów danych, np. poprzez losowanie bootstrapowe.
3. Stosowanie różnych podzbiorów zmiennych.
4. Wielokrotne zastosowanie tego samego algorytmu z różnymi wartościami parametrów lub punktami startowymi (np. losowo wybranymi zalążkami skupień w metodzie *k*-średnich).

Algorytm taksonomiczny powinien charaktryzować się stabilnością, a więc powinien być odporny na niewielkie zmiany w zbiorze danych, czy też wartości parametrów tego algorytmu. Wiadomo jednakże również, że kluczem do sukcesu podejścia zagregowanego jest zróżnicowanie klasyfikacji składowych. Klasyfikacja zagregowana, która została zbudowana na różniących się między sobą elementach składowych jest bardziej dokładna i stabilna niż pojedyncze metody taksonomiczne. W niniejszym badaniu uwaga zostanie skupiona na stabilności metod taksonomicznych. Głównym celem tego artykułu jest porównanie stabilności zagregowanych algorytmów taksonomicznych, a także relacji między stabilnością i dokładnością; przy czym pod uwagę zostanie wzięta tylko specyficzna klasa metod agregacji, które są oparte na idei metody *bagging*.

1. Metoda *bagging* w taksonomii

Metoda *bagging* jest pewną ogólną koncepcją, w ramach której narodziły się szczegółowe rozwiązania zaproponowane m.in. przez Hornika [2005], Dudoid i Fridlyand [2003] oraz Leischa [1999]. Pierwszy krok we wszystkich tych metodach jest taki sam: polega na konstrukcji *B* prób bootstrapowych i zastosowaniu do nich pojedynczego algorytmu taksonomiczego w celu uzyskania klasyfikacji składowych wchodzących w skład klasyfikacji ostatecznej. Poszczególne warianty tej metody różnią się natomiast w kroku drugim, czyli w kroku agregacji wyników.

Propozycja Leischa

Leisch [1999] zaproponował, by w pierwszym kroku na podstawie każdej podpróby bootstrapowej dokonać grupowania przy zastosowaniu tzw. bazowej metody taksonomicznej, którą jest jedna z metod iteracyjno-optymalizacyjnych, np. algorytm *k*-średnich. W kolejnym etapie ostateczne centra skupień są prze-

kształcane w nowy zbiór danych obejmujący $B \times K$ obserwacji (K to liczba skupień w metodzie bazowej), który jest poddawany podziałowi za pomocą metod hierarchicznych. Uzyskany dendrogram jest podstawą ostatecznego podziału – obserwacje z pierwotnego zbioru są przydzielane do tej grupy, której środek ciężkości znajduje się w minimalnej odległości Euklidesowej.

Algorytm zaproponowany przez Leischa przebiega w następujących krokach:

1. Z pierwotnego N-elementowego zbioru G 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 zbioru 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 $C^B = C^B(K) = \{c_{11}, ..., c_{BK}\}$.
4. Do tak skonstruowanego zbioru należy zastosować hierarchiczną metodę taksonomiczną, uzyskując w ten sposób dendrogram.
5. Niech $c(x_i)$ oznacza zalążek znajdujący się najbliżej obserwacji x_i, i = 1, …, n. Podział na grupy pierwotnego zbioru danych jest określany w ten sposób, że dendrogram uzyskany na podstawie zbioru C^B 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 $c(x_i)$.

Propozycja Duidoid i Fridlyand

Metoda *bagging* w wersji zaproponowanej przez Dudoid i Fridlyand [2003] stosuje algorytmy iteracyjno-optymalizacyjne do oryginalnego zbioru danych i poszczególnych prób bootstrapowych, a po dokonaniu permutacji etykiet klas w wynikach grupowania uzyskanych na podstawie każdej podpróby tak, by zachodziła jak największa zbieżność z klasyfikacją obiektów z oryginalnego zbioru danych, stosuje głosowanie majoryzacyjne w celu określenia ostatecznej klasyfikacji zagregowanej.

Kroki zaproponowanego przez nich algorytmu można ująć według następującego schematu. Dla założonej liczby klas *K*:

1. Zastosuj iteracyjno-optymalizacyjny algorytm taksonomiczny *T* do pierwotnego zbioru danych *G*, 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 algorytm taksonomiczny *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 klasyfikacją 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, 0 w przypadku przeciwnym.

5. Powtórz kroki 2-4 *B* razy. Ostatecznie zaklasyfikuj *i*-tą obserwację, stosując głosowanie majoryzacyjne, zatem przydzielając ją do tej klasy, dla której zachodzi:

$$\arg\max_{1 \le k \le K} \sum_{b: x_i \in G_n^b} I(\tau^b(T(x_i, G_n^b)) = k). \tag{2}$$

Propozycja Hornika

W metodzie tej po skonstruowaniu *B* prób bootstrapowych i zastosowaniu do nich pojedynczego algorytmu taksonomicznego uzyskuje się klasyfikacje składowe. Klasyfikacja zagregowana natomiast jest uzyskiwana 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 klasyfikacji zagregowanych,
dist – odległość Euklidesowa,
$(c_1,...,c_B)$ – klasyfikacje wchodzące w skład klasyfikacji zagregowanej.

2. Miary stabilności i dokładności

W celu zbadania stabilności i dokładności zastosowano koncepcję miar zaproponowanych przez Kunchevę i Vetrova [2006]. Mierniki te są oparte na skorygowanym indeksie Randa (*AR*), którego definicja jest następująca [Hubert i Arabie 1985]: niech A i B będą wynikami dwóch różnych klasyfikacji zbioru Z posiadającego N elementów. Przez l_A oznaczmy liczbę klas w klasyfikacji A, natomiast przez l_B – liczbę klas w klasyfikacji B; N_{ij} to liczba obiektów znajdujących się w klasie i w grupowaniu A i w klasie j w klasyfikacji B; $N_{i\bullet}$ to liczba obserwacji w klasie i w klasyfikacji A, natomiast $N_{\bullet j}$ to liczba obserwacji w klasie j w klasyfikacji B. Skorygowany indeks Randa jest dany wzorem:

$$AR(A,B) = \frac{\sum_{i=1}^{l_A}\sum_{j=1}^{l_B}\binom{N_{ij}}{2} - t_3}{\frac{1}{2}(t_1 + t_2) - t_3}, \tag{4}$$

gdzie:

$$t_1 = \sum_{i=1}^{l_A}\binom{N_{i\bullet}}{2}, \tag{5}$$

$$t_2 = \sum_{j=1}^{l_B}\binom{N_{\bullet j}}{2}, \tag{6}$$

$$t_3 = \frac{2t_1 t_2}{N(N-1)}. \tag{7}$$

1. STABILNOŚĆ DLA PAR KLASYFIKACJI ZAGREGOWANYCH (ang. *pairwise ensemble stability*):

$$S_{agr} = \frac{2}{Z \cdot (Z-1)} \sum_{\substack{1 \le z,l \le Z \\ z<l}}^{Z} AR(P_z^{agr}, P_l^{agr}), \tag{8}$$

gdzie:
Z – liczba klasyfikacji zagregowanych,
AR – skorygowany indeks Randa,
P_z^{agr} – klasyfikacja na podstawie z-tej klasyfikacji zagregowanej,
P_l^{agr} – klasyfikacja na podstawie l-tej klasyfikacji zagregowanej.

Miara ta ocenia stabilność klasyfikacji zagregowanych poprzez ocenę podobieństwa wyników grupowania, które na ich podstawie zostały uzyskane.

2. PRZECIĘTNA DOKŁADNOŚĆ KLASYFIKACJI ZAGREGOWANEJ (ang. *average ensemble accuracy*):

$$A_{agr} = \frac{1}{Z} \sum_{z=1}^{Z} AR(P_z^{agr}, P^T), \tag{9}$$

gdzie: P^T – rzeczywiste etykiety klas.

Miara ta jest uśrednioną po wszystkich klasyfikacjach zagregowanych miarą dokładności i mierzy podobieństwo między ostateczną klasyfikacją zagregowaną a prawdziwymi etykietami klas.

3. Badania empiryczne

W badaniach zastosowano sztucznie generowane zbiory danych, które standardowo są wykorzystywane w badaniach porównawczych w taksonomii[1]. Są to takie zbiory, w których przynależność obiektów do klas jest znana. Ich krótka

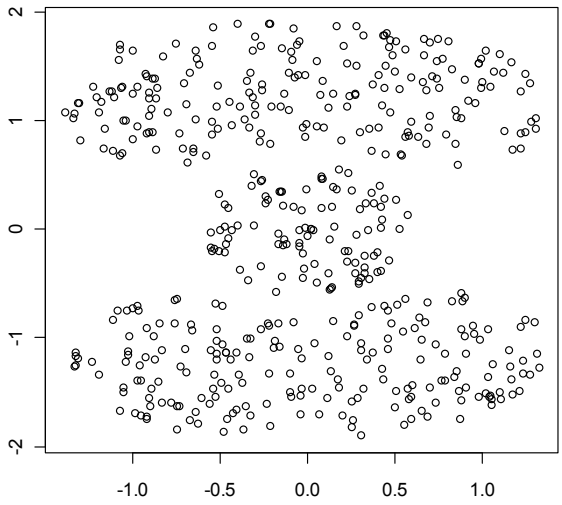
[1] Zbiory zaczerpnięte zostały z pakietu `mlbench` z programu **R**.

charakterystyka znajduje się w tabeli 1, natomiast struktura jest pokazana na rys. 1-8. Zbiory *Cassini*, *Cuboids*, *Shapes*, *Smiley* oraz *Spirals* należą do zbiorów o wyraźnie separowalnych klasach, natomiast *2dnormals*, *Ringnorm* i *Threenorm* posiadają nakładające się na siebie, trudno separowalne klasy.

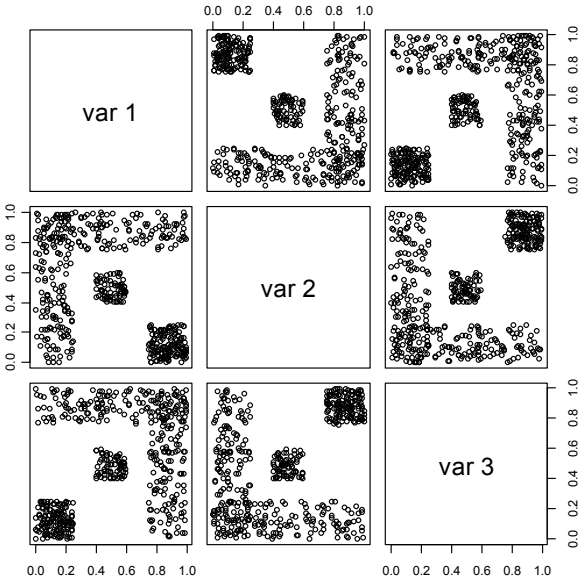
Tabela 1

Charakterystyka zastosowanych zbiorów danych

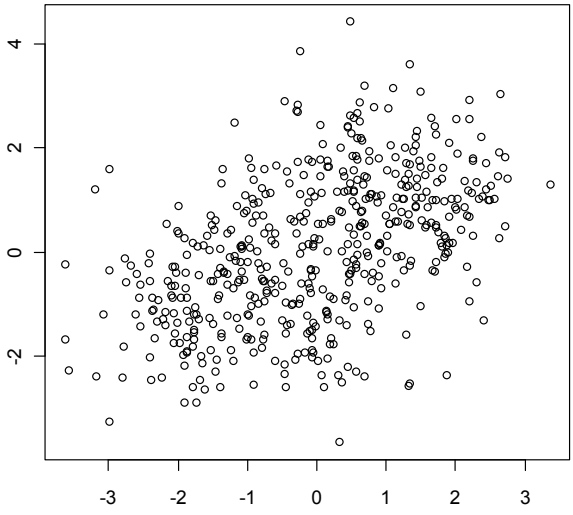
Zbiór danych	Liczba obiektów	Liczba cech	Liczba klas
Cassini	500	2	3
Cuboids	500	3	4
2dnormals	500	2	2
Ringnorm	500	2	2
Shapes	500	2	4
Smiley	500	2	4
Spirals	500	2	2
Threenorm	500	2	2



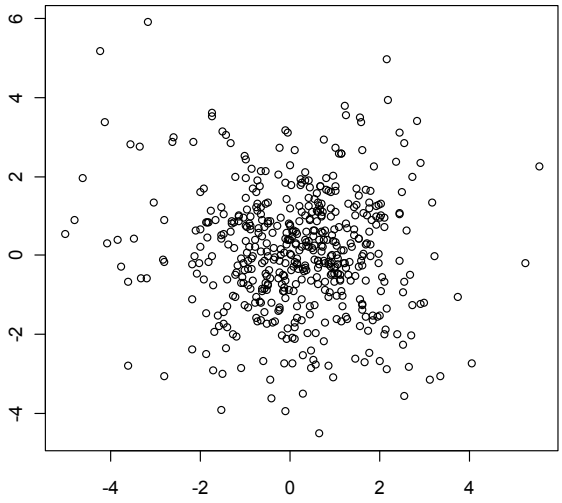
Rys. 1. Zastosowane zbiory danych – zbiór *Cassini*



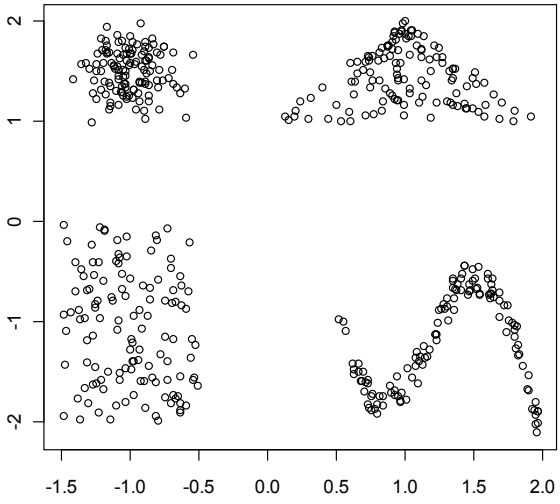
Rys. 2. Zastosowane zbiory danych – zbiór *Cuboids*



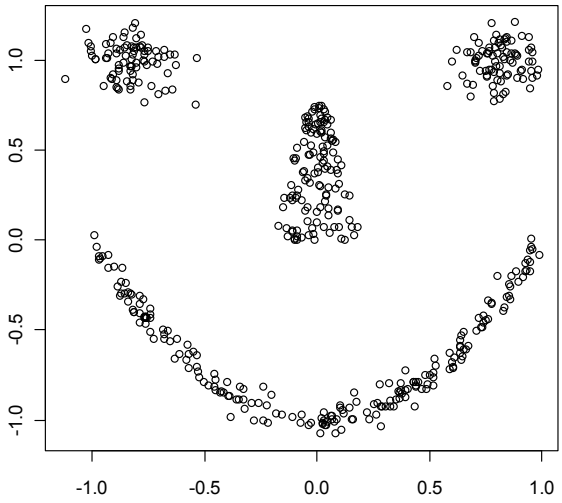
Rys. 3. Zastosowane zbiory danych – zbiór *2dnormals*



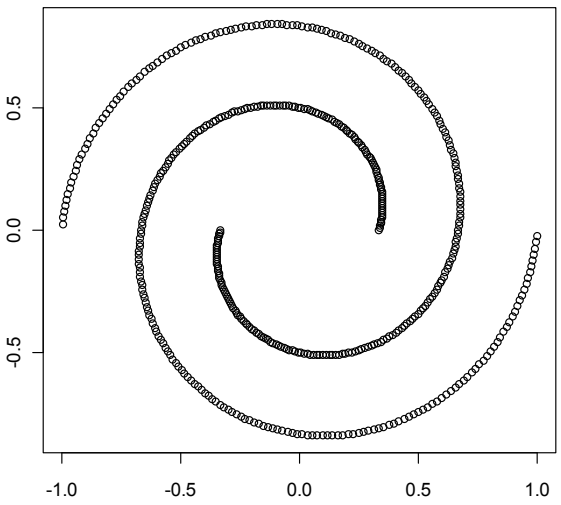
Rys. 4. Zastosowane zbiory danych – zbiór *Ringnorm*



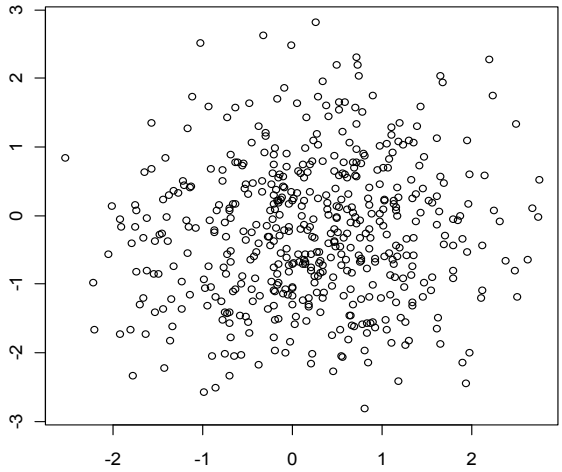
Rys. 5. Zastosowane zbiory danych – zbiór *Shapes*



Rys. 6. Zastosowane zbiory danych – zbiór *Smiley*



Rys. 7. Zastosowane zbiory danych – zbiór *Spirals*



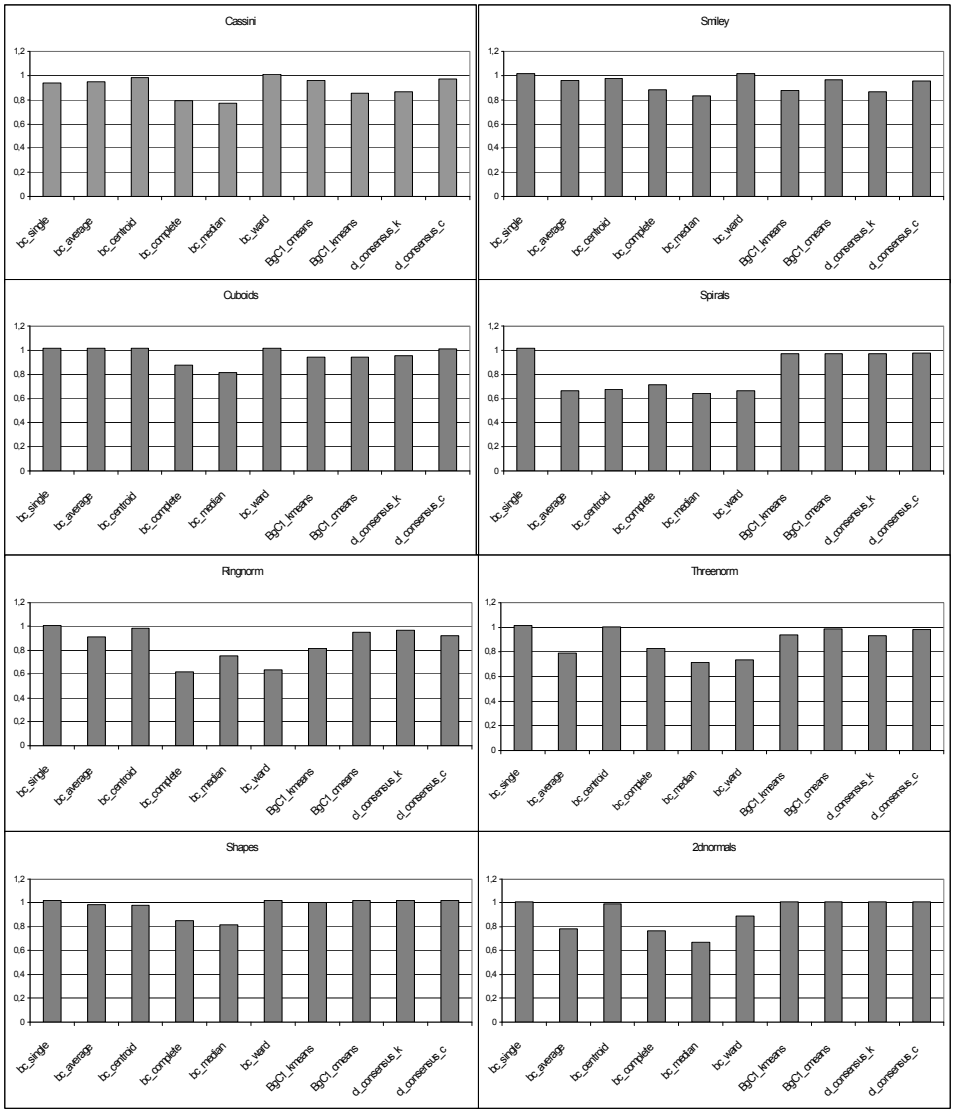
Rys. 8. Zastosowane zbiory danych – zbiór *Threenorm*

W badaniach empirycznych zastosowano 50 klasyfikacji zagregowanych, a wszystkie obliczenia zostały powtórzone 50 razy, by uzyskać bardziej dokładne i wiarygodne rezultaty. W metodzie *bagging* zaproponowanej przez Leischa po skonstruowaniu 10 prób bootstrapowych jako bazowy iteracyjno-optymalizacyjny algorytm taksonomiczny zastosowano metodę *k*-średnich z wartością parametru $k = 50$[2], a po przekształceniu ostatecznych zalążków skupień do postaci zbioru danych obejmującego 500 obserwacji dokonano podziału za pomocą następujących hierarchicznych metod taksonomicznych[3]: najbliższego sąsiedztwa (bclust_single), najdalszego sąsiedztwa (bclust_complete), centroidy (bclust_centroid), mediany (bclust_median), średniej odległości (bclust_mean), warda (bclust_ward). Obliczenia zostały wykonane w programie **R** z zastosowaniem funkcji `bclust` z pakietu `e1071`.

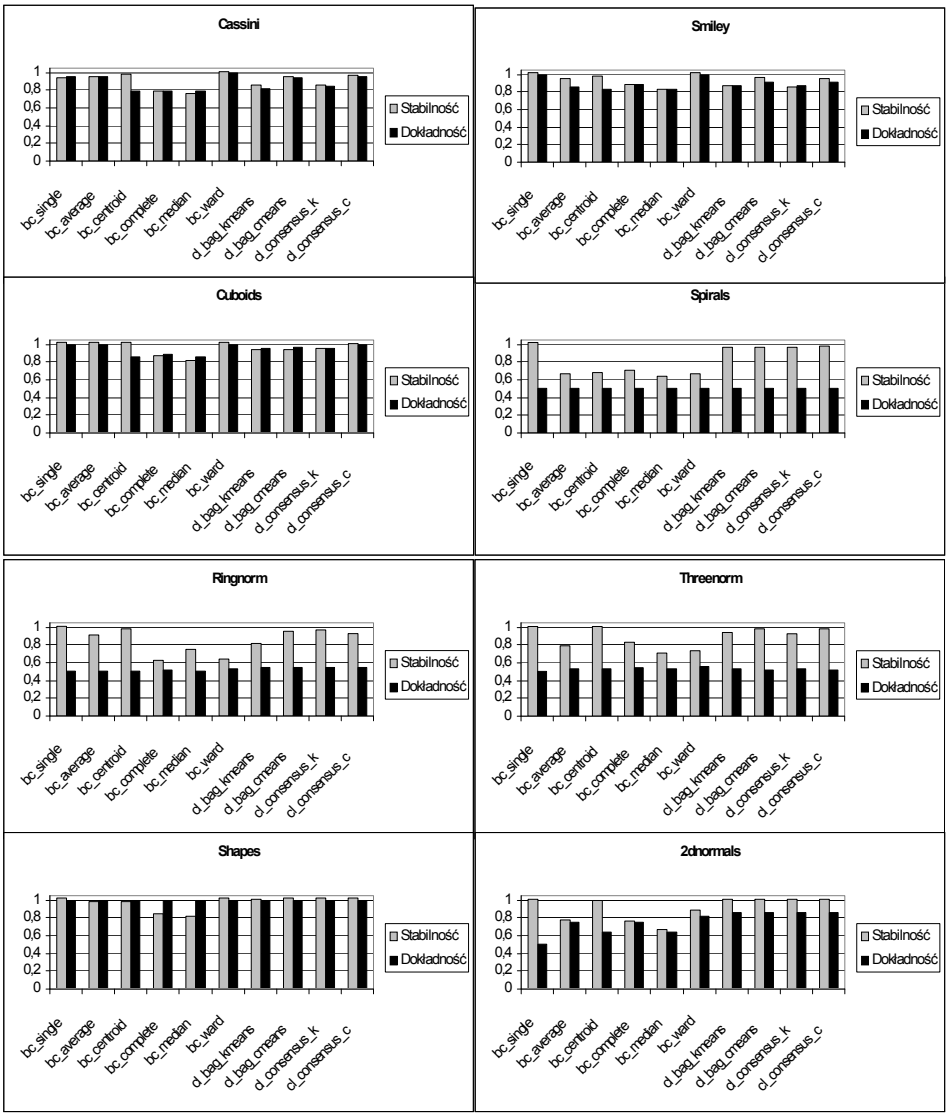
W metodzie *bagging* w wersji zaproponowanej przez Dudoid i Fridlyand oraz przez Hornika po skonstruowaniu 25 prób bootstrapowych zastosowano dwa algorytmy, a mianowicie metodę *k*-średnich oraz *c*-średnich, która jest rozmytą wersją metody *k*-średnich opracowaną przez Bezdeka [1981]. Metoda Dudoid i Fridlyand jest oprogramowana w programie **R** pod nazwą funkcji `cl_bag` w pakiecie `clue` (na rysunkach zastosowano nazwy cl_bag_kmeans oraz cl_bag_cmeans), natomiast metodę Hornika można znaleźć w tym samym pakiecie pod nazwą `cl_consensus` (na rysunkach oznaczenie cl_consensus_k odnosi się do metody agregacji, gdzie na poszczególnych próbach bootstrapowych była stosowana metoda *k*-średnich, a cl_consensus_c – metoda *c*-średnich).

[2] Autor metody zaleca, by wartość tego parametru była większa niż rzeczywista liczba skupień.
[3] W nawiasach zostały podane skróty nazw stosowane na rysunkach.

Rezultaty obliczeń widoczne na rys. 9 pozwalają stwierdzić, że w prawie wszystkich przypadkach najmniej stabilną okazała się metoda bclust_complete oraz bclust_median. Najwyższą stabilnością w przypadku większości zbiorów danych charakteryzują się metody: bclust_single, bclust_average oraz bclust_centroid (z wyjątkiem metod bclust_average oraz bclust_centroid dla zbioru *Spirals* oraz metody bclust_average dla zbioru *Threenorm* i *2dnormals*). Całkiem stabilne rezultaty można także zaobserwować dla reszty badanych metod z wyjątkiem metody bclust_ward dla zbiorów *Ringnorm*, *Threenorm* oraz *Spirals*.



Rys. 9. Stabilność poszczególnych metod opartych na idei *bagging* dla różnych zbiorów danych



Rys. 10. Relacje między stabilnością a dokładnością dla poszczególnych metod opartych na idei *bagging* dla różnych zbiorów danych

Wykresy na rys. 10 pokazujące relacje zachodzące między miarami stabilności i dokładności pozwalają stwierdzić brak generalnie obowiązującej zależności. Na przykład dla zbioru *Cassini* oraz *Cuboids* miary stabilności i dokładności osiągają niemalże ten sam poziom (z wyjątkiem metody bc_centroid). Podobnie miary te kształtują się także dla zbiorów *Shapes* oraz *Smiley* (z wyjąt-

kiem metod bclust_complete i bclust_median dla zbioru *Shapes* oraz metody bclust_median dla zbioru *Smiley*). Już dla zbioru *Ringnorm, Threenorm* oraz *Spirals* można jednak zaobserwować, że miary dokładności kształtują się na niemalże tym samym poziomie, natomiast miary stabilności zachowują się różnie dla różnych metod[4]. Na przykład dla cl_bag_cmeans, cl_bag_kmeans, cl_consensus_kmeans i cl_consensus_cmeans przyjmują dosyć duże wartości, a dla bclust_ward – stosunkowo niskie.

Podsumowanie

Przechodząc do sformułowania uwag końcowych, należy na wstępie zauważyć, że wybór dobrego algorytmu taksonomicznego jest znacznie trudniejszy niż wybór dobrego algorytmu dyskryminacyjnego. Wynika to przede wszystkim z faktu, że w klasyfikacji wzorcowej mamy do czynienia z zagadnieniem uczenia z nauczycielem. W taksonomii natomiast nie znamy klas, do których należą obiekty, a tym samym brak jest określonej z góry struktury, która powinna zostać rozpoznana przez algorytm. W związku z tym, by ominąć ryzyko wyboru niewłaściwego algorytmu taksonomicznego, można zastosować podejście zagregowane celem połączenia wyników klasyfikacji różnych algorytmów. Każdy z nich ma swoje mocne i słabe strony, ale wydaje się, że ich łączne zastosowanie przyniesie efekt kompensacji.

Drugą zaletą podejścia zagregowanego jest uniezależnienie wyników od wybranej metody, czy też wartości pewnych parametrów tych metod (np. początkowo wybranych zalążków skupień w metodzie *k*-średnich), a także zwiększenie odporności algorytmów taksonomicznych na szum i obserwacje oddalone. Agregacja wyników pozwala zatem na stabilizację rezultatów grupowania.

Wspomniane zalety powodują, że podejście to jest warte uwagi i tego, by spróbować zbadać relacje zachodzące między stabilnością i dokładnością zagregowanych algorytmów taksonomicznych. W przypadku gdyby między nimi zachodził wyraźny związek, mierniki stabilności mogłyby posłużyć jako wskazówka pomagająca wybrać najlepszą metodę podziału.

Z przeprowadzonych badań nad stabilnością zagregowanych metod taksonomicznych opartych na metodzie *bagging* wynika, że najbardziej stabilne okazały się metody: bclust_single, bclust_average, bclust_centroid, cl_bag_cmeans, cl_bag_kmeans, cl_consensus_kmeans oraz cl_consensus_cmeans. Najmniej

[4] Głównym punktem zainteresowania badań jest stabilność zagregowanych algorytmów taksonomicznych, dlatego przedstawiono wyniki nawet wtedy, gdy dokładność klasyfikacji nie osiągała wysokich wartości.

stabilne okazały się natomiast metody bclust_centroid oraz bclust_median; podczas gdy metoda bclust_ward dla niektórych zbiorów była bardzo stabilna (np. dla zbiorów *Cassini, Cuboids, Shapes* i *Smiley*), a dla niektórych stabilność była stosunkowo niska.

Z badań nad relacją między stabilnością i dokładnością w algorytmach opartych na metodzie *bagging* wynika, że nie da się sformułować jasnej i ogólnie obowiązującej zasady. Dla niektórych zbiorów danych stabilność i dokładność kształtuje się na zbliżonym do siebie poziomie, a dla niektórych stwierdza się brak jakiegokolwiek związku między nimi.

Literatura

Bezdek J.C. (1981): *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum, New York.

Breiman L. (1996): *Bagging Predictors*. „Machine Learning", No. 26(2).

Dudoit S., Fridlyand J. (2003): *Bagging to Improve the Accuracy of a Clustering Procedure*. „Bioinformatics", Vol. 19, No. 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 of 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", No. 27(6).

Freund Y. (1999): *An Adaptive Version of the Boost by Majority Algorithm*. „Proceedings of the 12th Annual Conference on Computational Learning Theory".

Hornik K. (2005): *A CLUE for CLUster Ensembles*. „Journal of Statistical Software", No. 14.

Hubert L.., Arabie P. (1985): *Evaluating Object Set Partitions: Free Sort Analysis and Some Generalizations*. „Journal of Verbal Learning and Verbal Behaviour", No. 15.

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.

Strehl A., Ghosh J. (2002): *Cluster Ensembles – A Knowledge Reuse Framework for Combining Multiple Partitions*. „Journal of Machine Learning Research", No. 3.

COMPARISON OF STABILITY OF CLUSTER ENSEMBLES BASED ON BAGGING IDEA

Summary

Ensemble approach has been successfully applied in the context of supervised learning to increase the accuracy and stability of classification. One of the most popular method is bagging based on bootstrap samples. Recently, analogous techniques for cluster analysis have been suggested in order to increase classification accuracy, robustness and stability of the clustering solutions. Research has proved that, by combining a collection of different clusterings, an improved solution can be obtained.

A desirable quality of the method is the stability of a clustering algorithm with respect to small perturbations of data (e.g., data subsampling or resampling, small variations in the feature values) or the parameters of the algorithm (e.g., random initialization). Here, we look at the stability of the ensemble and carry out an experimental study to compare stability of cluster ensembles based on bagging idea.

Ewa Genge
Uniwersytet Ekonomiczny w Katowicach

THE MULTINOMIAL MIXTURE MODEL – THE ANALYSIS OF STUDENTS' ATTITUDE TO THE SILESIA REGION

Introduction

Many statistical models involve mixture distributions in some way or other. In mixture distributions a population made up of *u* subgroups, mixed at random in proportion to the relative group sizes is considered. The interest lies in some random variable X which is heterogeneous across and homogeneous within the subgroups. Due to heterogeneity, X has a different probability distribution in each group, usually assumed to arise from the same parametric family, however, with the vector of parameter $\mathbf{\Theta}_s$ differing across the groups (*s*).

An overview of mixture models is given in Titterington et al. [1985] or McLachlan and Peel [2000, p. 81-116]. The most popular are multivariate normal mixture models (Gaussian mixture models). They are used in a lot of different areas such as astronomy, biology, economic, marketing or medicine [see i.e. Fraley and Raftery 2002, p. 611-631; Wedel and DeSarbo 1995, p. 21-55; Witek 2010a, p. 615-624; 2010b, p. 63-72]. Since the mixture of multinomial distributions is applied in the empirical part of this article we present the definition of this kind of mixture below.

1. The multinomial mixture model – definition

The data of n objects described by categorical variables $l_1,\ldots,l_m$ is considered. The data can be represented by the vector of objects $\mathbf{x}_i = (x_{ijh}; j = 1,\ldots,m; h = 1,\ldots,l_j; i = 1,\ldots,n)$ where $x_{ijh} = 1$ if the object i

belongs to the category h of the variable j. The total number of categories is given by $l = \sum_{j=1}^{m} l_j$, then the data is defined by the n by m matrix.

In the multinomial mixture model it is assumed that each observation $\mathbf{x}_i$ arises independently from a mixture of multivariate multinomial distributions defined by:

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

where:

f_s – density function of component *s*,

$\mathbf{x}_i$ – the vector of objects,

$\boldsymbol{\Theta}_s$ – the component specific parameter vector for the density function f_s,

$\boldsymbol{\Theta}$ – the vector of all parameters for the mixture density function, $\boldsymbol{\Theta} = (\tau_s, \boldsymbol{\Theta}_s)$,

τ_s – the prior probability of component *s*;

$(\tau_s \geq 0 \wedge \sum_{s=1}^{u} \tau_s = 1), \boldsymbol{\Theta}_\mathbf{s} \neq \boldsymbol{\Theta}_l \forall s \neq l.$

The *s*th component of the mixture can be given as:

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

where $\boldsymbol{\Theta}_s = (\Theta_{sjh}; j = 1,\ldots,m; h = 1,\ldots,l_j)$ and (2) formula is a product of m conditionally independent multinomial distributions of parameters $\boldsymbol{\Theta}_{sj}$.

Banfield and Raftery [1993, p. 803-821] proposed to constrain the covariances in the mixture of multivariate normal distributions, which resulted in 14 Gussian mixture models. Similarly, Celeux and Govaert [2008] imposed some constraints on the parameters of the mixture of multinomial distributions ($\boldsymbol{\Theta}$) and received 5 multinomial models.

The basic idea of this proposition is to impose the vector of components on distributions parameters $\boldsymbol{\Theta}_{sj} = (\Theta_{sj1},\ldots,\Theta_{sjl_j})$ to take the form $(\beta_{sj},\ldots,\beta_{sj},\gamma_{sj},\beta_{sj},\ldots,\beta_{sj})$, with $\gamma_{sj} > \beta_{sj}$. Since $\sum_{h=1}^{l_j} \Theta_{sjh} = 1$, we have:

$$(l_j - 1)\beta_{sj} + \gamma_{sj} = 1, \tag{3}$$

$$\beta_{sj} = (1 - \gamma_{sj})/(l_j - 1). \tag{4}$$

The constraint $\gamma_{sj} > \beta_{sj}$ can be finally written as $\gamma_{sj} > 1/l_j$. Then the vector $\mathbf{\Theta}_{sj}$ can be split into the following parameters:

- $\mathbf{a}_{sj} = (a_{sj1}, \ldots, a_{sjl_j})$, where $a_{sjh} = 1$ if h is equal γ_{sj}, $a_{sjh} = 0$ otherwise,
- $\varepsilon_{sj} = 1 - \gamma_{sj}$ corresponds to the probability that the data $\mathbf{x}_i$ arising from the sth component, such that $x_{ijh(s,j)} \neq 1$.

In other words, the multinomial distribution associated with the jth variable of the sth component is reparameterized by a center $\mathbf{a}_{sj}$ and the dispersion parameter ε_{sj}, which allows a interpretation similar to the center and the variance matrix used for continuous data in the Gaussian mixture models.

The relationship between the initial and new distribution parameters can be written as:

$$\Theta_{sjh} = \begin{cases} 1 - \varepsilon_{sj} & if \quad h = h(s,j), \\ \varepsilon_{sj}/(l_j - 1) & if \quad h \neq h(s,j). \end{cases} \tag{5}$$

Equation (2) can be for $\mathbf{a}_s = (\mathbf{a}_{sj}, j = 1, \ldots, m)$ and $\varepsilon_s = (\varepsilon_{sj}, j = 1, \ldots, m)$ rewritten as:

$$f_s(\mathbf{x}_i | \mathbf{\Theta}_s) = \bar{f}_s(\mathbf{x}_i | \mathbf{a}_s, \varepsilon_s) = \prod_{j=1}^{m} \prod_{h=1}^{l_j} ((1 - \varepsilon_{sj})^{a_{sjh}} (\varepsilon_{sj}/(l_j - 1))^{1 - a_{sjh}})^{x_{ijh}}. \tag{6}$$

This model will be denoted as $[\varepsilon_{sj}]$, in the following. On the basis of (6), three other models can be deduced:

- $[\varepsilon_s]$ – the model where ε_{sj} is independent of the variable j,
- $[\varepsilon_j]$ – the model where ε_{sj} is independent of the sth component,
- $[\varepsilon]$ – the model where ε_{sj} is independent both of the variable j and the sth component.

The most general model will also be denoted as $[\varepsilon_{sjh}]$. The number of the parameters associated with each models is given in Table 1, where $\sigma = 0$ in the case of equal prior probabilities and $\sigma = u - 1$ when prior probabilities are different for each class.

Table 1

The number of parameters of the 5 multinomial models

Model	Number of parameters
$[\varepsilon]$	$\sigma + 1$
$[\varepsilon_j]$	$\sigma + m$
$[\varepsilon_s]$	$\sigma + u$
$[\varepsilon_{sj}]$	$\sigma + um$
$[\varepsilon_{sjh}]$	$\sigma + u\sum_{j=1}^{m}(l_j - 1)$

Source: Celeux, Govaert [2008, p. 35].

2. Parameter estimation and model selection

The parameters of the mixture of multinomial models are usually estimated by maximum likelihood using the Expectation-Maximization (EM) algorithm [Dempster et al. 1977, p. 1-38]. Each EM iteration consists of two steps – an E-step and an M-step. In the M-step (for the a posteriori probabilities, obtained in E-step) new parameters of maximum likelihood given by (7) are obtained:

$$L(\mathbf{x}_i|\boldsymbol{\Theta}_s, \pi_s, z_{is}) = \sum_{i=1}^{n}\sum_{s=1}^{u} z_{is} \log\left[\tau_s f_s(\mathbf{x}_i|\boldsymbol{\Theta}_s)\right], \tag{7}$$

where $z_{is} = 1$ if $\mathbf{x}_i$ belongs to group s or $z_{is} = 0$ otherwise. Maximum likelihood estimators for each of the five models presented in Table 1 are given below. We adopt the notation:

$$e_{sjh} = n_s - \sum_{i=1}^{n} z_{is} x_{ijh}, \tag{8}$$

and $h(s, j)$ for the value which minimizes the difference given in (8).

For convenience, we assume that $e_{sj} = e_{sjh(s,j)}$.

1. Model $[\varepsilon_{sjh}]$:

$$\Theta_{sjh} = 1 - e_{sjh} / n_s . \tag{9}$$

2. Model $[\varepsilon_{sj}]$:

$$\Theta_{sjh} = \begin{cases} 1 - e_{sj} / n_s & if \quad h = h(s,j), \\ e_{sj} /(n_s(l_j - 1)) & if \quad h \neq h(s,j). \end{cases} \tag{10}$$

3. Model $[\varepsilon_{s}]$:

$$\Theta_{sjh} = \begin{cases} 1 - (\sum_j e_{sj}) / n_s m & if \quad h = h(s,j), \\ (\sum_j e_{sj}) /(n_s m(l_j - 1)) & if \quad h \neq h(s,j). \end{cases} \tag{11}$$

4. Model $[\varepsilon_{j}]$:

$$\Theta_{sjh} = \begin{cases} 1 - (\sum_s e_{sj}) / n_s & if \quad h = h(s,j), \\ (\sum_s e_{sj}) /(n(l_j - 1)) & if \quad h \neq h(s,j). \end{cases} \tag{12}$$

5. Model $[\varepsilon]$:

$$\Theta_{sjh} = \begin{cases} 1 - (\sum_{j,s} e_{sj}) /(nm) & if \quad h = h(s,j), \\ (\sum_{j,s} e_{sj}) /(nm(l_j - 1)) & if \quad h \neq h(s,j). \end{cases} \tag{13}$$

The M steps for each of five models ($[\varepsilon_{sjh}]$, $[\varepsilon_{sj}]$, $[\varepsilon_{s}]$, $[\varepsilon_{j}]$, $[\varepsilon]$) could also be written using the new parameterization $\mathbf{a}_s$ and ε_s. Then it is assumed that:

$$a_{sjh} = \begin{cases} 1 & if \quad h = h(s,j), \\ 0 & if \quad h \neq h(s,j). \end{cases} \tag{14}$$

$$\varepsilon_{sj} = 1 - \Theta_{sjh}(s,j). \tag{15}$$

The E and M steps are repeated until the likelihood improvement falls under a pre-specified threshold or a maximum number of iterations is reached [see Wang 1994 for more details].

In order to select the optimal clustering model several measures have been proposed [see i.e. McLachlan and Peel 2000, p. 81-116]. Four information criteria are available in `mixtools` package of **R**: BIC (*Bayesian Information Criterion*), AIC (*Akaike Information Criterion*), ICL (*Integrated Completed Likelihood*) and CAIC (*Consistent Akaike Information Criterion*). The performance of some of these criteria was compared by Biernacki et al. [1999, p. 49-71] and Bozdogan [2000, p. 62-91]. In general, BIC was found to be consistent under correct specification of the component densities [Kass and Raftery 1995, p. 928--934; Keribin 2000, p. 49-66] and has given good results in a range of applications [i.e. Fraley and Raftery 2002, p. 611-631; Stanford and Raftery 2000, p. 601-609]. The criteria used in further analysis are defined:

$$AIC_s = 2\log p(\mathbf{x}_i, y_i \left| \hat{\boldsymbol{\Theta}}_s, M_s \right.) - 2v_s, \tag{16}$$

$$BIC_s = 2\log p(\mathbf{x}_i, y_i \left| \hat{\boldsymbol{\Theta}}_s, M_s \right.) - v_s \log(n), \tag{17}$$

$$ICL_s = 2\log p(\mathbf{x}_i, y_i \left| \hat{\boldsymbol{\Theta}}_s, M_s \right.) + \frac{v_s}{2}\log(n), \tag{18}$$

$$CAIC_s = 2\log p(\mathbf{x}_i, y_i \left| \hat{\boldsymbol{\Theta}}_s, M_s \right.) - v_s(\log(n)+1), \tag{19}$$

where: $\log p(\mathbf{x}_i, y_i \left| \hat{\boldsymbol{\Theta}}_s, M_s \right.)$ – is the maximized loglikelihood for the model M_s, v_s is the number of parameters to be estimated in that model, n is the number of observations in the data.

The first term in criteria measures the goodness-of-fit, whereas the second term penalizes model complexity.

3. Example

In this example the data collected by the Marketing Department of University of Economics in Katowice in 2008 were analysed. The main goal of this sampling survey was to recognize students' attitudes to the Silesia region and its

promotion. The survey comprised different areas of the Silesia region: central, the Dabrowa Basin, south, north, south-west. The respondents studied at:

- the University of Economics in Katowice,
- the University of Economics in Katowice (Rybnik Centre),
- the University of Economics in Katowice (Bielsko Campus),
- the Katowice School of Economics (Katowice Piotrowice),
- the Katowice School of Finance and Banking,
- the Czestochowa University of Technology,
- the Czestochowa School of Linguistics,
- the Academy of Fine Arts in Katowice,
- the Higher School of Applied Sciences in Ruda Slaska.

Students were asked 12 questions about their background and their attitude to Silesia, its culture, tradition and promotion.

There were 627 polls collected. The main goal of the analysis was to find clusters with similar students' attitudes to our region. The mixture of multinomial distributions were applied. All computations in this paper were done in `mixtools` package of **R** and SPSS software. Some results of `mixtools` package of **R** are presented in Figure 1.

```
> x.new<-makemultdata(slask, cuts = 2)
> multmixmodel.sel(x.new$y, comps = c(1,2),
  epsilon = 1e-03)
number of iterations= 114
1 2 Winner
AIC -3244.819 -1764.462 2
BIC -3247.039 -1771.123 2
CAIC -3247.539 -1772.623 2
ICL -3247.039 -1770.603 2
Loglik -3243.819 -1761.462 2
```

Fig. 1. The results of `mixtools` package of R

The optimal number of the mixture components was chosen using four different information criteria. Figure 1 shows that the optimal number of components is 2 (for each of criterion). We estimated parameters of two components using EM algorithm. The mixture of multinomial distribution methodology outlined before yields two groups of students consisting of 255 and 372 students respectively.

The first group comprises students who feel a strong bond with Silesia. For question: "Do you feel ties with Silesia?", 58% chose answer "yes", 32% – "rather yes". There were no negative answer. Students are also rather intent on staying in Silesia: 61% of students are going to stay in Silesia, 34% have not decided yet and 5% are going to leave. The students in this group like Silesian traditions. The question "Do you like Silesian traditions?" elicited 37% "yes" answers and 46% "rather yes" answers. As far as the Polish Silesian dialect is concerned, the majority of students like it (38% "yes" answers and 28% "rather yes" answers). However, 33% of students do not like it too much (the percentage of students who chose answers: "neither yes, neither no"). High pollution is perceived as the main disadvantage of living in the Silesia area (64% "yes" and 28% "rather yes" answers). Nearly three-quarters of students polled believe in the improvement of the Silesia's image. However, as many as 75% of students did not observe any Silesia's promotion. There were different opinions concerning Silesia's promotion in our country: 38% think that the Silesia region should be promoted as a whole, 24% claim that the separate subregions should be promoted and 38% think that the separate subregions should be promoted but under the common logo of the Silesia region. Silesia is perceived as a region attractive for tourists by 42% of students, 26% think the opposite and 32% do not have any opinion. We can say that students of this group have a positive attitude towards Silesia. We can suppose that this kind of attitude and the sense of belonging to this region stem from students' background. 70% of students of this group were brought up here and their parents come from here, 21% of students have been living in Silesia for years, but their parents come from another part of Poland, only 8% of students polled came here just to study.

Quite a different attitude towards Silesia can be observed in the second group of students. The ties with Silesia are quite weak, i.e. only 39% of students feel strong ties with Silesia, 27% feel some kind of bond, 20% of the respondents feel no ties with Silesia, 13% haven't even thought about it. Only 46% of students have decided to stay here in the future, as many as 17% are intent on leaving and 37% haven't taken any decision on this issue yet. The students belonging to this group do not like Silesian traditions very much: 23% chose "yes" answers, 31% chose "rather yes" answers, 16% do not like the traditions at all. The last part of this group do not have any opinion (answer "neither yes, nor no"). The vast majority of this group do not like the Silesian dialect either. The question "Do you like the Silesian dialect?" elicited 30% "no" answers and 20% "rather no" answers. The positive attitude to the infrastructure development is almost at the same level in both groups. The air pollution in this region is also very negatively perceived in the second group of students. As far as the im-

provement of the image of the Silesia region is concerned, 5% less than in the first class believe that it is at all possible. Most of the students have not observed the new promotional campaign (64%), but there are also 12% of students who like it very much (16% have no opinion). There are also different opinions about the way of promoting the Silesia region, similarly to the first group. The vast majority of students (35%) think that the separate subregions should be promoted but under the common Silesian logo. A large part of this group perceives Silesia as unattractive for tourists (35%), 34.7% of students do not have any opinion. For 40% of the respondents, Silesia is as an industrial area, comprising an area of the former Katowice voivodship, for 28% of students Silesia is a region associated with the current area of this part of Poland. However, as many as 12% less than in the first group of students do perceive the Dabrowa Basin as a separate part of Silesia. We think that the reason of this split approach is that many people looking for a job came and settled down in this part of Silesia many years ago.

We think that the definitely skeptical attitude to the Silesia, its customs, dialects, tradition and different Silesian borders in this group is connected with students' and their parents' background. 59% of students and their parents come from Silesia, 29% of parents come from other regions of Poland and 12% of students came only to study here.

Conclusions

We have shown the use of the mixture models in the classification of students studying in different parts of Silesia. The mixture of multinomial models analysis yields two groups of students. The first group comprises students who feel strong ties with Silesia. The bond with Silesia in the second group of students is quite weak.

The mixture model analysis has confirmed that students' and their parents' background has the influence on those two different attitudes. The difference can be especially observed among students living/studying in the Dabrowa Basin. Administratively, they feel Silesian. They live in this region, but do not have the roots here, so they do not necessarily identify with everything that Silesia is connected with.

Literature

Banfield J.D., Raftery A.E. (1993): *Model-based Gaussian and Non-Gaussian Clustering.* "Biometrics", No. 49.

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

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

Celeux G., Govaert G. (2008): http://www.mixmod.org/IMG/pdf/statdoc_2_1_1.pdf.

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.

Fraley C., Raftery A.E. (2002): *Model-based Clustering, Discriminant Analysis, and Density Estimation.* "Journal of the American Statistical Association", No. 97.

Kass R.E., Raftery A.E. (1995): *Bayes Factors*. "Journal of the American Statistical Association", No. 90.

Keribin C. (2000): Consistent Estimation of the Order of Mixture Models. "Sankhya Indian Journal Statistics", No. 62.

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

Stanford D., Raftery A.E. (2000): *Principal Curve Clustering with Noise*. "IEEE Transactions on Pattern Analysis and Machine Intelligence", No. 22.

Titterington D.M., Smith A.F., Makov U.E. (1985): *Statistical Analysis of Finite Mixture Distribution*. John Wiley & Sons, San Diego.

Wang P. (1994): *Mixed Regression Models for Discrete Data*, *PhD thesis.* University of British Columbia, Vancouver.

Wedel M., DeSarbo W.S. (1995): *A Mixture Likelihood Approach for Generalized Linear Models.* "Journal of Classification", No. 12.

Witek E. (2010a): *Analysis of Massive Emigration from Poland – the Model-based Clustering Approach.* Proceedings of the 32nd Annual Conference of the Gesellschaft für Klassifikation, Springer.

Witek E. (2010b): *Wykorzystanie mieszanek rozkładów w regresji*. W: *Współczesne problemy modelowania i prognozowania zjawisk społeczno-gospodarczych*. Red. J. Pociecha. Wydawnictwo UE, Kraków.

MIESZANKI ROZKŁADÓW WIELOMIANOWYCH – ANALIZA POSTAW STUDENTÓW WOBEC WOJEWÓDZTWA ŚLĄSKIEGO

Streszczenie

Mieszanki rozkładów są stosowane wówczas, gdy zbiór obserwacji charakteryzuje się nadmiernym rozproszeniem. W literaturze najczęściej są spotykane mieszanki rozkładów normalnych (*model-based clustering*). W referacie zostaną przedstawione mieszanki rozkładów wielomianowych oraz wyniki ich zastosowań do podziału studentów o podobnych postawach wobec województwa śląskiego (jego tradycji, kultury, możliwości rozwoju itd.).

Badania zostaną przeprowadzone za pomocą pakietu `mixtools` programu komputerowego **R**.

Jacek Stelmach
Grzegorz Kończak
Uniwersytet Ekonomiczny w Katowicach

O PORÓWNYWANIU DWÓCH POPULACJI WIELOWYMIAROWYCH Z WYKORZYSTANIEM OBJĘTOŚCI ELIPSOID UFNOŚCI

Wprowadzenie

Statystyka dostarcza wielu różnorodnych metod pozwalających na porównanie dwóch populacji. Samo określenie porównanie dwóch populacji może dotyczyć porównywania np. wartości przeciętnych, wskaźników struktury, wariancji lub współczynników korelacji. W takim przypadku mamy do czynienia z wnioskowaniem parametrycznym. Do odpowiedzi na tak postawione zagadnienia wykorzystujemy testy parametryczne, jak np. test t, test z, test równości wskaźników struktury lub współczynników korelacji. Często dokonując porównań populacji, stawiamy hipotezę o jednakowych rozkładach w tych populacjach. W takim przypadku odwołujemy się do testów nieparametrycznych, jak np. test serii lub test Manna-Whitneya. W opracowaniu przedstawiono propozycję porównania rozkładów dwóch populacji wielowymiarowych. Podstawą porównań są objętości elipsoid ufności dla dwóch populacji oraz dla populacji będącej sumą mnogościową analizowanych populacji. W rozważaniach nie przyjmowano założenia o postaci rozkładu. Takie podejście nie daje możliwości wyznaczenia dokładnego rozkładu rozważanej statystyki i odczytania wartości krytycznej testu z tablic. Z tego powodu w analizach wykorzystano testy permutacyjne. Ze względu na możliwość porównania własności rozważanej procedury z testem T^2 Hotellinga w analizach symulacyjnych skoncentrowano się na porównaniu populacji o wielowymiarowych rozkładach normalnych.

1. Wybrane metody porównania dwóch populacji

1.1. Porównanie populacji jednowymiarowych

Do najczęściej wykorzystywanych testów pozwalających na porównanie dwóch populacji należy zaliczyć test *t*. Pozwala on na porównanie wartości oczekiwanych w dwóch populacjach na podstawie pobranych niezależnie prób. Należy on do grupy testów parametrycznych. Związane są z tym określone założenia. Próby powinny być pobrane z populacji o rozkładach normalnych, a w przypadku prób o dużych liczebnościach jest dopuszczalne, aby rozkład był zbliżony do normalnego. Dodatkowo, jeżeli znane jest odchylenie standardowe rozkładu σ, to jest wykorzystywany test, który w literaturze zwykle jest określany jako test *z* [Kanji 2006]. Jeżeli dysponujemy próbkami z dwóch populacji o rozkładach zero-jedynkowych, to dla porównania możemy wykorzystać test równości wskaźników struktury. Test dla równości wariancji (test *F*) pozwala na nieco inne spojrzenie na porównanie dwóch populacji. W teście jest wykorzystywana statystyka *F*. Pozwala on na porównanie wariancji w dwóch populacjach. Podobnie jak w przypadku testu *t* zakłada się, że próby pochodzą z populacji o rozkładach normalnych.

Inną grupę testów stanowią testy nieparametryczne. Do najważniejszych testów nieparametrycznych pozwalających na porównanie dwóch populacji należy zaliczyć test Manna-Whitneya i test serii Walda-Wolfowitza [Blalock 1974]. Oba testy nie wymagają spełnienia ostrych założeń, jednak jako testy nieparametryczne charakteryzują się niewielką mocą. Powszechnie stosowane testy parametryczne wymagają spełnienia ostrych założeń dotyczących postaci rozkładów. Testy nieparametryczne charakteryzują się natomiast niewielką mocą. Alternatywnym rozwiązaniem jest stosowanie testów permutacyjnych [Efron i Tibshirani 1993]. Testy te nie wymagają spełnienia założenia normalności rozkładów, co jest niezbędne przy testach parametrycznych, a mimo to charakteryzują się podobną mocą.

1.2. Porównanie populacji wielowymiarowych

Porównując populacje wielowymiarowe napotykamy problemy, które nie występowały przy podobnej analizie dla populacji jednowymiarowych. Dla pewnych zmiennych (współrzędnych wektora) wartości oczekiwane mogą być jednakowe, a dla innych mogą się znacznie różnić. Ważne w takim przypadku będzie nie tylko stwierdzenie, że wektory wartości oczekiwanych nie są jedna-

kowe, ale również wskazanie, dla których zmiennych występują różnice w wartościach oczekiwanych. Poza różnicami w wartościach oczekiwanych mogą występować różnice w macierzach wariancji-kowariancji, co w praktyce przekłada się na inne wielowymiarowe kształty populacji.

Dla porównania wektorów wartości przeciętnych w dwóch populacjach wielowymiarowych można wykorzystać np. test T^2 Hotellinga. Statystyka ta jest uogólnieniem statystyki t wykorzystywanej dla porównania populacji ze względu na jedną zmienną mierzalną. Dla stosowania tego testu jest wymagane spełnienie założenia, że próby pobrano z populacji o wielowymiarowych rozkładach normalnych [Rencher 2002].

Przyjmijmy, że pobrano dwie niezależne próby $\mathbf{x}_{11}, \mathbf{x}_{12}, ..., \mathbf{x}_{1n_1} : N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ oraz $\mathbf{x}_{21}, \mathbf{x}_{22}, ..., \mathbf{x}_{2n_2} : N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$. Zakładając, że macierze kowariancji są nieznane, ale jednakowe ($\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$) do weryfikacji hipotezy o równości wektorów średnich:

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$$

wykorzystywana jest statystyka [Rencher 2002]:

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T \mathbf{S}_{pl}^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2), \tag{1}$$

gdzie:

$$\mathbf{S}_{pl} = \frac{1}{n_1 + n_2 - 2} \left(\sum_{i=1}^{n_1} (\mathbf{x}_{1i} - \bar{\mathbf{x}}_1)(\mathbf{x}_{1i} - \bar{\mathbf{x}}_1)^T + \sum_{i=1}^{n_2} (\mathbf{x}_{2i} - \bar{\mathbf{x}}_2)(\mathbf{x}_{2i} - \bar{\mathbf{x}}_2)^T \right) \tag{2}$$

jest nieobciążonym estymatorem wariancji $\boldsymbol{\Sigma}$.

Statystyka (1) ma rozkład Hotellinga. Wartości krytyczne dla tej statystyki wyznaczamy wykorzystując fakt, że statystyka:

$$F = \frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)p} T^2 \tag{3}$$

ma rozkład F o p oraz n_1+n_2-p-1 stopniach swobody. Weryfikując hipotezę o równości wektorów wartości oczekiwanych, obszar krytyczny jest prawo-

stronny. Statystyka (1) może być wykorzystana do weryfikacji hipotezy o jednakowych wektorach wartości przeciętnych, gdy jest spełnione założenie wielowymiarowej normalności rozważanych wektorów. Nieco zmodyfikowana postać statystyki może być wykorzystana, jeżeli np. znana jest postać macierzy kowariancji $\mathbf{\Sigma}$.

A.C. Rencher przedstawia również test dla porównania wariancji dla dwóch populacji wielowymiarowych [Rencher 2002]. Do weryfikacji hipotezy o identyczności macierzy wariancji $H_0 : \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ jest wykorzystywana statystyka:

$$M = \frac{|\mathbf{S}_1|^{v_1/2} + |\mathbf{S}_2|^{v_2/2}}{|\mathbf{S}_{pl}|^{(v_1+v_2)/2}}, \tag{4}$$

gdzie $v_i = n_i - 1$, a $\mathbf{S}_i$ jest macierzą kowariancji dla i-tej próbki ($i = 1, 2$).

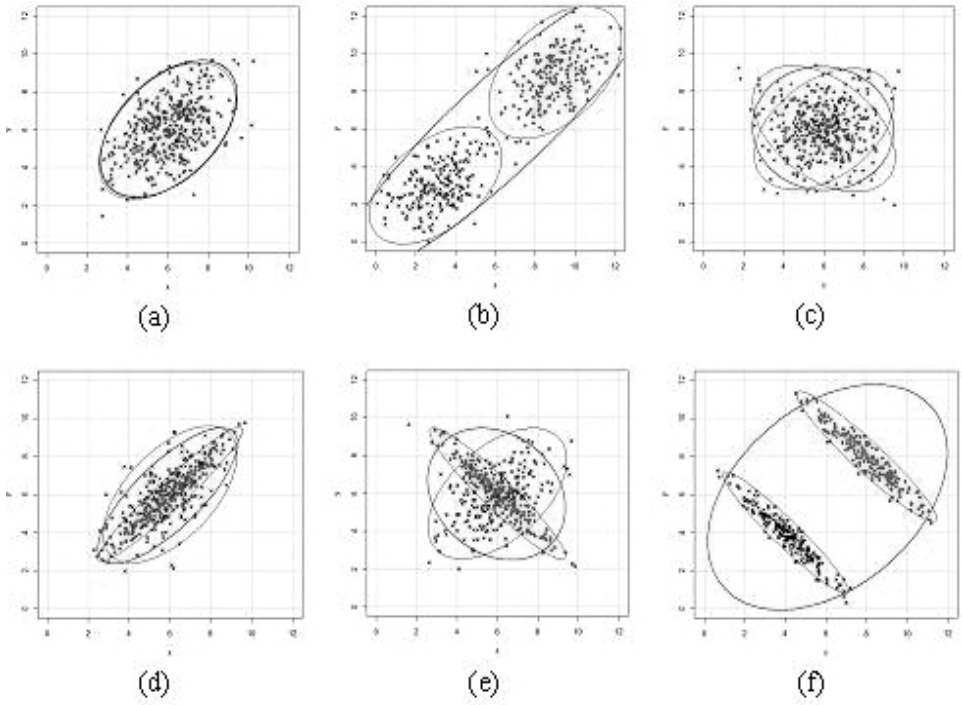
W praktyce badań statystycznych bardzo często nie ma podstaw do przyjęcia założenia o normalności rozkładów. W takich przypadkach niezbędne są narzędzia pozwalające na porównanie dwóch populacji wielowymiarowych nie przyjmując dodatkowych założeń np. odnośnie do postaci rozkładów.

2. Objętość elipsoid ufności – porównanie dwóch populacji wielowymiarowych

Rozważmy dwie populacje o dwuwymiarowych rozkładach normalnych. W kolejnych częściach rysunku 1 przedstawiono wykresy rozrzutu dla próbek wylosowanych z dwóch populacji o rozkładach normalnych. Parametry rozkładów normalnych były identyczne tylko w przypadku przedstawionym na rysunku 1a. W kolejnych częściach rysunku populacje różniły się wektorami wartości oczekiwanych i/lub macierzami kowariancji. Odpowiednie przypadki zostały schematycznie przedstawione na rysunkach 1a-f. Na tych rysunkach zostały również przedstawione elipsoidy ufności dla obu próbek oraz dla próby będącej mnogościową sumą tych prób. Tylko dla przypadku a) rozkłady w populacjach są jednakowe. W przypadkach b-f rozkłady nie są identyczne. Krótka charakterystyka rozważanych rozkładów została przedstawiona poniżej:

a) rozkłady identyczne (jednakowe wartości oczekiwane, jednakowe macierze wariancji),
b) różne wartości oczekiwane, jednakowe macierze wariancji,

c) jednakowe wartości oczekiwane, różne macierze wariancji (inne kształty),
d) jednakowe wartości oczekiwane, różne macierze wariancji (inny rozrzut),
e) jednakowe wartości oczekiwane, różne macierze wariancji (inny kształt i rozrzut)
f) różne wartości oczekiwane, jednakowe macierze wariancji.



Rys. 1. Dwie populacje – wzajemne położenie

Jak łatwo zauważyć, w przypadku gdy próbki pochodzą z populacji o różnych rozkładach, elipsoida ufności dla sumy prób charakteryzuje się większym polem (ogólnie: większą objętością) niż elipsoida dla każdej z dwóch prób. Dla przedstawionych graficznie 6 przypadków wzajemnego położenia tylko w pierwszym przypadku rozkłady w dwóch populacjach są jednakowe. Zastosowanie testu T^2 Hotellinga pozwoli wskazać na różnice w populacjach w przypadkach b) i f). W pozostałych wektory wartości przeciętnych są takie same, ale inne są macierze wariancji. Proponowany test powinien skutecznie

wskazywać występowanie różnic w rozkładach we wszystkich pięciu przypadkach (b-f), a jedynie w przypadku a) prowadzić do stwierdzenia braku podstaw do odrzucenia hipotezy. Do porównania rozkładów zostanie wykorzystana statystyka:

$$T = \frac{\max\{volA, volB, vol(A \cup B)\}}{\min\{volA, volB, vol(A \cup B)\}} \tag{5}$$

gdzie:

vol(A), vol (B) oznaczają objętość elipsoidy (w przypadku dwuwymiarowym pole elipsy) ufności otrzymaną na podstawie próby odpowiednio z populacji *A* oraz *B*,

$vol(A \cup B)$ oznacza objętość elipsoidy (w przypadku dwuwymiarowym pole elipsy) ufności otrzymaną na podstawie połączonych prób.

Statystyka (5) może przyjmować wartości z przedziału $[1, +\infty)$. W przypadku gdy rozkłady populacji będą identyczne, wartości statystyki będą bliskie wartości 1. Duże wartości statystyki *T* będą świadczyły o występujących różnicach w rozkładach populacji. Obszar krytyczny w proponowanym teście jest prawostronny.

3. Testy permutacyjne

Weryfikując hipotezę H_0, musimy znać wartości krytyczne, które określą obszar odrzucenia tej hipotezy. W rozważanym przypadku, rozkład statystyki testowej jest nieznany – a więc także nie jest on stablicowany. Nie jest więc możliwe odczytanie wartości krytycznych. W takich sytuacjach bardzo pomocne będzie zastosowanie testów permutacyjnych [Good 1994, s. 176]. Testy te nie wymagają żadnej wiedzy o rozkładzie wykorzystywanej statystyki. W eksperymencie zastosowano test permutacyjny, w którym z uwagi na czas obliczeń ograniczono się do $N = 1000$ permutacji (co dla większości przypadków jest ilością wystarczającą) – [Hesterberg et al. 2003, s. 18-60], polegających na utworzeniu podgrup przez losowanie bez zwracania z sumy mnogościowej prób pochodzących z populacji *A* i *B*.

Wartość *p-value* testu permutacyjnego wyznaczono z zależności:

$$ASL = \frac{card\{i \in \{1,2,...,N\} : T^* > T_i\}}{N}, \tag{6}$$

gdzie:

T^* – wartość statystyki obliczonej dla próby pierwotnej,

T_i – wartość statystyk obliczonych dla prób permutacyjnych.

Wartości *p-value* mniejsze od przyjętego poziomu istotności α prowadzą do odrzucenia hipotezy o identyczności rozkładów, a większe od α prowadzą do stwierdzenia o braku podstaw do odrzucenia hipotezy o identyczności rozkładów.

4. Analiza symulacyjna

Wszystkie analizy symulacyjne i obliczenia wykonano w programie R (http://www.r-project.org). W symulacjach rozważano populacje 5-wymiarowe. Dane o wektorach wartości oczekiwanych $\boldsymbol{\mu}_A$ i $\boldsymbol{\mu}_B$ oraz macierzach wariancji-kowariancji $\boldsymbol{\Sigma}_A$ i $\boldsymbol{\Sigma}_B$ populacji przedstawia tabela 1. Kolejne wiersze tabeli (a-f) korespondują z prezentacją graficzną na rysunku 1.

Tabela 1

Wektory wartości przeciętnych i macierze wariancji-kowariancji przypadków testowych

Przypadek	$\boldsymbol{\mu}_A$	$\boldsymbol{\mu}_B$	$\boldsymbol{\Sigma}_A$	$\boldsymbol{\Sigma}_B$
a)	[0, 0, 0, 0, 0]	[0, 0, 0, 0, 0]	$\mathbf{I}$	$\mathbf{I}$
b)	[0, 0, 0, 0, 0]	[1, 1, 1, 1, 1]	$\mathbf{I}$	$\mathbf{I}$
c)	[0, 0, 0, 0, 0]	[0, 0, 0, 0, 0]	$\boldsymbol{\Sigma}_1$	$\boldsymbol{\Sigma}_2$
d)	[0, 0, 0, 0, 0]	[0, 0, 0, 0, 0]	$\mathbf{I}$	$1{,}7\mathbf{I}$
e)	[0, 0, 0, 0, 0]	[0, 0, 0, 0, 0]	$\mathbf{I}$	$\boldsymbol{\Sigma}_3$
f)	[0, 0, 0, 0, 0]	[1, 1, 1, 1, 1]	$\boldsymbol{\Sigma}_2$	$\boldsymbol{\Sigma}_2$

Macierze występujące w tabeli 1 są zadane następującymi wzorami:

I – macierz jednostkowa o wymiarach 5 x 5.

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 1 & 0{,}6 & 0{,}6 & 0 & 0 \\ 0{,}6 & 1 & 0{,}6 & 0{,}6 & 0 \\ 0{,}6 & 0{,}6 & 1 & 0{,}6 & 0 \\ 0 & 0{,}6 & 0{,}6 & 1 & 0{,}6 \\ 0 & 0 & 0 & 0{,}6 & 1 \end{bmatrix}$$

$$\boldsymbol{\Sigma}_2 = \begin{bmatrix} 1 & -0{,}6 & 0{,}6 & 0 & 0 \\ -0{,}6 & 1 & -0{,}6 & 0{,}6 & 0 \\ 0{,}6 & -0{,}6 & 1 & -0{,}6 & 0 \\ 0 & 0{,}6 & -0{,}6 & 1 & -0{,}6 \\ 0 & 0 & 0 & -0{,}6 & 1 \end{bmatrix}$$

$$\boldsymbol{\Sigma}_3 = \begin{bmatrix} 1 & 0{,}7 & 0{,}2 & 0 & 0 \\ 0{,}7 & 1 & 0{,}7 & 0{,}2 & 0 \\ 0{,}2 & 0{,}7 & 1 & 0{,}7 & 0{,}2 \\ 0 & 0{,}2 & 0{,}7 & 1 & 0{,}7 \\ 0 & 0 & 0{,}2 & 0{,}7 & 1 \end{bmatrix}$$

Dla każdego z przedstawionych w tabeli 1 przypadków rozważano próbki o liczebnościach $n_1 = n_2 = 10$, 20, 30 i 50. Dla tych przypadków generowano 1000-krotnie próby z populacji A oraz B. Następnie przyjmując poziom istotności $\alpha = 0{,}05$ oraz wykorzystując statystykę (5), przeprowadzano test permutacyjny. Na tej podstawie wyznaczano oceny prawdopodobieństw odrzucenia hipotezy H_0.

Niezależnie od powyżej opisanych symulacji wykonano analizę Monte Carlo pozwalającą porównać własności rozważanego testu z klasycznym testem T^2 Hotellinga. W tym celu porównywano dwie populacje o dwuwymiarowych rozkładach normalnych o parametrach $\boldsymbol{\mu}_A = [0;0]$, $\boldsymbol{\Sigma}_A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ oraz

$\boldsymbol{\mu}_B = [x; x]$, $\boldsymbol{\Sigma}_B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, gdzie x = 0,1; 0,2; …;2. W symulacjach uwzględniono próbki o liczebnościach $n_1 = n_2 = 10$, 20, 30 i 50. Na podstawie $N = 1000$ symulacji dla każdej takiej pary prób przeprowadzono test permutacyjny oraz T^2 Hotellinga. W obu przypadkach wyznaczono liczbę odrzuceń hipotezy o równości wektorów wartości przeciętnych. Na tej podstawie otrzymano oceny prawdopodobieństw odrzucenia hipotezy H_0.

5. Wyniki

W analizach uwzględniono objętości elipsoid ufności pokrywających wielowymiarową przestrzeń z badanymi obserwacjami – z 95% prawdopodobieństwem, wykorzystując procedury pakietu R. Wyniki analiz (dane porównywanych populacji w tabeli 1) umieszczono w tabeli 2. Przedstawiono w niej oceny prawdopodobieństw odrzucenia hipotezy o identyczności rozkładów 5-wymiarowych populacji (dla rozkładów identycznych – przypadek testowy a) oznacza błąd pierwszego rodzaju, dla pozostałych, w których symulowano różnice w rozkładach – moc testu) – z poziomem istotności $\alpha = 0.05$.

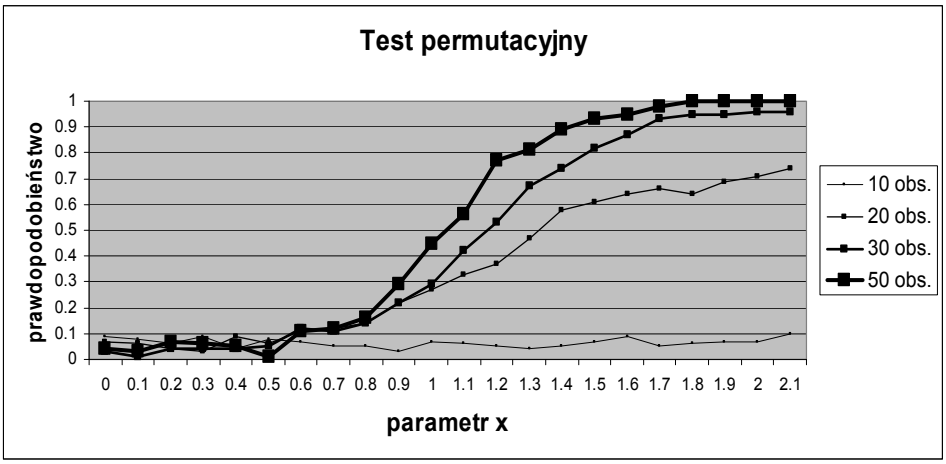
Tabela 2

Oceny prawdopodobieństw odrzucenia hipotezy o identyczności rozkładów

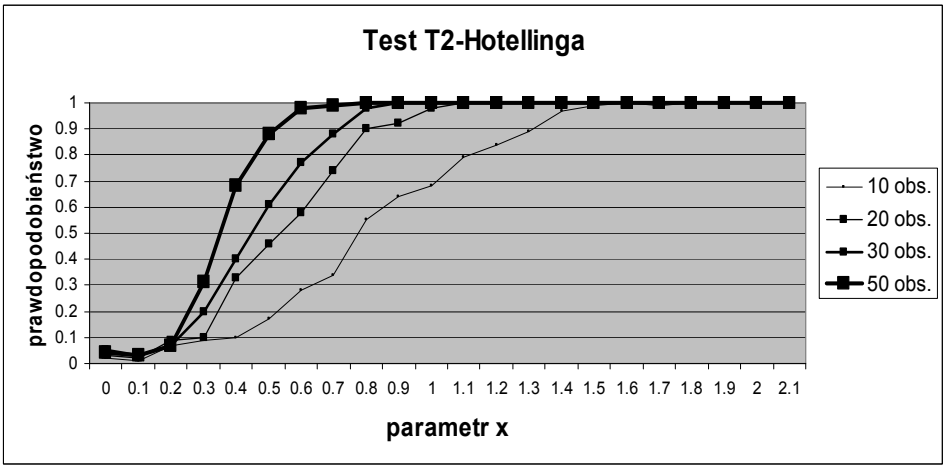
Przypadek	Liczebność próby			
	10	20	30	50
a)	0,036	0,037	0,048	0,037
b)	0,120	0,312	0,337	0,397
c)	0,897	1,000	1,000	1,000
d)	0,095	0,361	0,626	0,890
e)	0,486	0,969	1,000	1,000
f)	0,213	0,615	0,825	0,953

Dla przypadku identycznych rozkładów populacji (wariant a) rozmiar proponowanego testu jest nieco mniejszy od przyjętego poziomu istotności α. We wszystkich przypadkach, gdy rozkłady nie są identyczne, test dla wszystkich

rozważanych liczebności skutecznie wskazuje na występujące różnice. Szczególnie dobrze jest to widoczne dla prób o liczebności przynajmniej 30. Przeciętny błąd oceny szacowanych prawdopodobieństw we wszystkich analizowanych sytuacjach jest mniejszy od 0,016.



Rys. 2. Oceny prawdopodobieństwa odrzucenia hipotezy zerowej dla testu permutacyjnego w zależności od różnicy w wartościach współrzędnych *x*



Rys. 3. Oceny prawdopodobieństwa odrzucenia hipotezy zerowej dla testu T^2-Hotellinga w zależności od różnicy w wartościach współrzędnych *x*

Na rysunkach 2 i 3 zawarto wyniki analizy Monte Carlo – porównawczo test permutacyjny ze statystyką testową jak w (5) oraz test T^2 Hotellinga dla dwuwymiarowych prób o rozkładach różniących się wektorami średnich. Różnicę w wartościach współrzędnych tych wektorów opisuje parametr x wykresu leżący na osi odciętych. Badanie przeprowadzono dla różnych liczebności symulowanych prób. Analizując otrzymane w tej części wyniki, można zauważyć, że test T^2 Hotellinga skuteczniej odróżnia populacje różniące się wektorami wartości średnich. Porównań dokonano dla populacji o rozkładach normalnych, bo tylko wówczas jest uprawnione porównanie tych dwóch testów. Stosowanie testu T^2 Hotellinga wymaga spełnienia założenia o normalności rozkładu w badanych populacjach. Zaletą proponowanego testu permutacyjnego jest fakt, że może on być stosowany dla populacji o dowolnych rozkładach.

Podsumowanie

Proponowana statystyka testowa, oparta na analizie objętości elipsoid ufności obejmujących badane próby pozwala na weryfikację hipotez o identyczności rozkładów, także w przypadkach, w których test T^2 Hotellinga z uwagi na równość wartości średnich z badanych prób nie doprowadzi do odrzucenia hipotezy. Dodatkowo wykorzystanie testów permutacyjnych zwalnia z weryfikacji założenia o zgodności badanych rozkładów z rozkładem normalnym wielowymiarowym i nie wymaga tablicowania proponowanej statystyki testowej. Przeprowadzone badania symulacyjne wykazały zadowalające prawdopodobieństwo odrzucenia hipotezy zerowej dla rozkładów różniących się macierzą kowariancji czy wektorem wartości przeciętnych już dla prób o liczebności powyżej 30 obserwacji, a wysokie prawdopodobieństwo – dla liczebności ponad 50 obserwacji.

Przeprowadzona analiza Monte Carlo, porównująca moc testów permutacyjnego i T^2 Hotellinga, przeprowadzona dla rozkładów dwuwymiarowych normalnych, różniących się tylko wektorem wartości średnich (a więc dla rozkładów, do których jest predystynowany test parametryczny T^2 Hotellinga) wykazała większą moc testu parametrycznego. Niemniej jednak test permutacyjny cechował się porównywalną wielkością błędu pierwszego rodzaju, a zdolność rozpoznawania różniących się populacji osiągał dla różnicy wektorów wartości średnich na poziomie [1.0; 1.0].

Literatura

Blalock H.M. (1974): *Statystyka dla socjologów*. PWN, Warszawa.
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.
Hesterberg T., Monaghan S., Moore D.S., Clipson A., Epstein R. (2003): *The Practice of Business Statistics.* W.H. Freeman and Company, New York.
Kanji G.K. (2006): *100 Statistical Tests*. Sage Publications, London.
Rencher A.C. (2001): *Methods of Multivariate Analysis*. John Wiley & Sons, New York.

ON THE COMPARISON OF TWO MULTIDIMENSIONAL POPULATIONS USING THE CONFIDENCE ELLIPSOID VOLUMES

Summary

A comparison of two populations seems to be interesting and very common statistical problem. The most often way is to verify the hypothesis concerned the equality of certain, characteristic parameter i.e. mean, standard deviation or fraction with parametric or non-parametric tests. The authors propose to compare the distribution of two populations – comparing the confidence ellipsoid volumes. Since their distribution is unknown – permutation tests were applied. A Monte-Carlo simulation let to compare power of these tests with T^2 Hotelling tests. Proposed methods can be used, when the assumptions for parametric tests couldn't be verified.