NR 9 (496)

WIADOMOŚCI STATYSTYCZNE

CZASOPISMO GŁÓWNEGO URZĘDU STATYSTYCZNEGO
I POLSKIEGO TOWARZYSTWA STATYSTYCZNEGO WRZESIEŃ 2002

STUDIA METODOLOGICZNE

Wojciech NIEMIRO, Waldemar POPIŃSKI, Jacek WESOŁOWSKI, Robert WIECZORKOWSKI

Optymalna alokacja próby w warunkach niekompletnej realizacji badania

Badania reprezentacyjne prowadzone są na próbkach losowych, wybieranych zgodnie z planem, będącym odzwierciedleniem badanej populacji. Jednym z powszechnych mankamentów tych badań jest różnica między próbką wylosowaną i próbką zbadaną, opisywana przez parametr zwany kompletnością badania. Jest wiele czynników wpływających na obniżenie kompletności badania. Należą do nich np.: niska jakość operatu i odmowy odpowiedzi. W wielu przypadkach wpływ odmów można zmniejszyć przez intensywniejszą pracę badających — to wiąże się z większym obciążeniem jednostek prowadzących badanie. Dlatego wydaje się rozsądne przyjęcie degresywnego modelu realizacji badania, tzn. takiego, w którym liczebność próbki zbadanej nie może przekroczyć pewnego progu oraz gdy jej stosunek do liczebności próbki wylosowanej maleje wraz ze wzrostem liczebności próby wylosowanej.

W naszej pracy przyjęto wykładniczy model zależności liczebności próbki zbadanej y od liczebności próbki wylosowanej x:

$$y = f(x) = b[1 - exp(-x/a)] \tag{1}$$

przy czym parametry a i b są liczbami dodatnimi i mają spełniać warunek $b/a \leq 1$, który można zapisać inaczej jako ograniczenie na wartość pochodnej funkcji f w zerze

$f'(0) \leq 1$, a interpretować jako prawdopodobieństwo udzielenia odpowiedzi przez pojedynczą jednostkę populacji.

Oznaczmy funkcję odwrotną do f przez g:

$$x = g(y) = -a\, ln(1 - y/b) \tag{2}$$

Rozważymy populację podzieloną na domeny, czyli takie podzbiory, dla których chcemy estymować wartość globalną pewnej ustalonej cechy. Dodatkowo populacja podzielona będzie na podpopulacje, w których zależności między liczebnością próbki zbadanej i wylosowanej mogą się różnić w ramach przyjętego modelu wykładniczej zależności funkcyjnej, tzn. wartości parametrów a i b mogą zmieniać się z podpopulacji na podpopulację. Każdy element populacji jest ponadto zakwalifikowany do jednej z klas (w szczególności, gdy żadne klasy nie są wyróżnione — to odpowiada sytuacji, gdy liczba klas wynosi jeden). Przyjmujemy warstwowy plan próbkowania wewnątrz każdej domeny z warstwami będącymi przekrojami podpopulacji i klas.

Celem przedstawianej pracy jest podanie algorytmu takiej alokacji próby (o zadanej całkowitej liczebności) w przekrojach (domena × podpopulacja × klasa), aby precyzja warstwowego estymatora wartości globalnej w domenach była jednakowa i możliwie najmniejsza. Nasza praca mieści się więc w nurcie badań dotyczących optymalnej alokacji stałoprecyzyjnej — zob. np. Lednicki i Wesołowski (1994), Niemiro i Wesołowski (2001). Z matematycznego punktu widzenia problem taki sprowadza się do minimalizacji pewnej funkcji z ograniczeniami. Problem ten udaje się rozwiązać jedynie numerycznie, co jest zadowalające z punktu widzenia aplikacji. Natomiast brak rozwiązania analitycznego utrudnia badanie uniwersalnych własności otrzymywanych rozwiązań.

ANALITYCZNE SFORMUŁOWANIE I NUMERYCZNE ROZWIĄZANIE PROBLEMU

Niech i oznacza numer domeny, j numer podpopulacji, a h numer klasy. Oznaczmy przez x_{ijh} oraz y_{ijh}, odpowiednio, liczebność próbki wylosowanej i zbadanej w przekroju (i, j, h), a ponadto niech:

$$x_j = \sum_i \sum_h x_{ijh} \text{ oraz } y_j = \sum_i \sum_h y_{ijh} \tag{3}$$

Zakładamy, że

$$\begin{aligned} y_j = f_j(x_j) &= b_j[1 - exp(-x_j / a_j)], \text{ czyli} \\ x_j = g_j(y_j) &= -a_j\, ln(1 - y_j / b_j) \end{aligned} \tag{4}$$

(Wydaje się, że dopasowanie tego modelu do konkretnej sytuacji musi polegać na wykorzystaniu danych dotyczących kompletności wcześniejszych badań do estymacji parametrów a_j i b_j).

Dodatkowo przyjmujemy założenie, że w każdej podpopulacji frakcja próbki zbadanej (w stosunku do wylosowanej) nie zmienia się dla różnych przekrojów (domena × klasa), tzn. dla dowolnych (i, j, h) mamy:

$$\frac{y_{ijh}}{y_j} = \frac{x_{ijh}}{x_j} \tag{5}$$

Dla przedstawionego modelu wydaje się naturalne sformułowanie problemu optymalnej alokacji stałoprecyzyjnej. Rozważamy estymator warstwowy pewnej cechy, przy tym interesują nas wartości globalne tej cechy w *poszczególnych domenach*. Przy założonej maksymalnej wartości procentowej p, dla precyzji nieobciążonego estymatora $\hat{t}$ wartości globalnej t, która to precyzja dana jest wzorem $196\sqrt{V(\hat{t})}/t$, gdzie $V(\hat{t})$ oznacza wariancję estymatora $\hat{t}$, poszukujemy takich wielkości losowanej próbki x_{ijh} we wszystkich przekrojach (i, j, h), aby spełnione były następujące warunki:

1. Dla każdej domeny i estymator $\hat{t}_i$ ma precyzję nie gorszą niż p. Uwzględniając wzór na wariancję estymatora wartości globalnej w losowaniu warstwowym (np. Bracha 1996), warunek ten prowadzi do nierówności:

$$196^2 \sum_j \sum_h \left(\frac{S_{ijh}^2 N_{ijh}^2}{t_i^2 y_{ijh}} - \frac{S_{ijh}^2 N_{ijh}}{t_i^2} \right) \leq p^2 \tag{6}$$

   gdzie t_i jest wartością globalną badanej cechy w i-tej domenie, S_{ijh}^2 jest wariancją tej cechy w przekroju (i, j, h), a N_{ijh} jest liczbą jednostek w tym przekroju.

2. Sumaryczna liczebność wylosowanej próbki $x = \sum_i \sum_j \sum_h x_{ijh}$ jest minimalna.

Zauważmy, że jest to sformułowanie dualne do minimalizacji stałej precyzji p, przy założonej długości próbki losowanej x.

Opisane zadanie optymalizacji można sformułować jako zagadnienie minimalizacji nieliniowej funkcji celu przy liniowych ograniczeniach. Wykorzystując ideę opisaną w pracy Grenia (1966) wprowadzamy nowe zmienne $z_{ijh} = 1/y_{ijh}$, co prowadzi do zadania minimalizacji funkcji:

$$F[(z_{ijh})_{ijh}] = \sum_j -a_j ln\left(1 - b_j^{-1} \sum_i \sum_h z_{ijh}^{-1}\right) \to min \tag{7}$$

po układach liczb $(z_{ijh})_{ijh}$, gdzie (i, j, h) przebiega wszystkie przekroje, przy ograniczeniach dla każdej domeny i postaci:

$$\sum_j \sum_h \left(\frac{S_{ijh}^2 N_{ijh}^2}{t_i^2} z_{ijh} - \frac{S_{ijh}^2 N_{ijh}}{t_i^2} \right) \leq (p/196)^2 \qquad (8)$$

Zauważmy, że funkcja celu (7) jest wypukła oraz ograniczenia (liniowe) (8) również tworzą zbiór wypukły, zatem z klasycznej teorii programowania wypukłego wynika istnienie jednoznacznego rozwiązania problemu.

Oczywiście, znajomość rozwiązania powyższego problemu minimalizacji wyznacza optymalną alokację zgodnie ze wzorami:

$$x_{ijh} = y_j^{-1} z_{ijh}^{-1} \cdot x_j \quad \text{oraz} \quad x_j = -a_j \, ln\left(1 - b_j^{-1} \sum_i \sum_h z_{ijh}^{-1} \right) \qquad (9)$$

W celu numerycznego rozwiązania zadania wykorzystano język AMPL oraz możliwość rozwiązywania problemów optymalizacji za pośrednictwem Internetu, w środowisku serwera NEOS (http:/www.neos.mcs.anl.gov.). Język AMPL (ang. *A Mathematical Programming Language*) jest językiem programowania przeznaczonym specjalnie do opisu złożonych problemów optymalizacji — zob. Fourer, Gay i Keringhan (1990). Zapis zadania w tym języku jest bardzo zbliżony do odpowiedniego sformułowania w zapisie matematycznym. Ponadto AMPL zakłada oddzielenie samego opisu problemu od części wprowadzającej konkretne dane, stanowiące parametry modelu. AMPL pozwala użytkownikowi skupić się na samym modelu matematycznym, unikając wielu szczegółów implementacji algorytmów obliczeniowych. Konkretne realizacje AMPL działają na zasadzie transformacji odpowiedniego opisu do postaci wymaganych przez wyspecjalizowane programy rozwiązujące różne klasy zagadnień optymalizacyjnych (tzw. *solvery*). Przykładem takiej realizacji jest właśnie udostępnione w Internecie środowisko serwera NEOS. Pozwala ono użytkownikowi na zdalne, za pomocą układu pośredniczącego przeglądarki WWW, rozwiązywanie zadań optymalizacji zapisanych uprzednio w języku AMPL jako pliki tekstowe na lokalnym komputerze. Dany model można wykorzystywać wielokrotnie, zmieniając jego parametry wejściowe, a także uruchamiając procesy obliczeniowe dla różnych *solverów* dostępnych na serwerze NEOS dla danej kategorii rozważanego problemu. Przykładowo, dla interesującej nas minimalizacji możemy wybrać w grupie programów optymalizacji nieliniowej z ograniczeniami m.in. takie *solvery,* jak: MOSEK, MINOS, LANCELOT, LOQO.

ZASTOSOWANIE DO MODYFIKACJI ALOKACJI W BADANIU MAŁYCH PRZEDSIĘBIORSTW

Bezpośrednią przyczyną rozważania opisanej problematyki są trudności występujące w badaniu małych jednostek gospodarczych prowadzonym w GUS. W warunkach tego badania domenami są rodzaje działalności gospodarczej (77 rodzajów), podpopulacjami są województwa (16), a klasy (4) zdefiniowane są przez przekroje formy prawnej (2) i wielkości przedsiębiorstwa (2).

Kompletność badania w ostatnich trzech latach w podziale na podpopulacje (województwa) przedstawiona jest w tabl. 1.

TABL. 1. LICZEBNOŚĆ JEDNOSTEK WYLOSOWANYCH I ZBADANYCH W POSZCZEGÓLNYCH WOJEWÓDZTWACH (podpopulacjach) W BADANIU MAŁYCH PRZEDSIĘBIORSTW

Symbole województw	1998			1999			2000		
	jednostki wylosowane	jednostki zbadane	w % jednostek zbadanych	jednostki wylosowane	jednostki zbadane	w % jednostek zbadanych	jednostki wylosowane	jednostki zbadane	w % jednostek zbadanych
02	7858	4153	52,9	9233	4675	50,6	10419	4560	43,8
04	5281	2739	51,9	5651	2777	49,1	6758	2752	40,7
06	6107	3981	65,2	3210	1983	61,8	3608	1821	50,5
08	2442	1451	59,4	2798	1654	59,1	3050	1487	48,8
10	7730	3458	44,7	7566	3807	50,3	7209	3276	45,4
12	8349	3979	47,7	10694	5777	54,0	8416	4017	47,7
14	21813	8877	40,7	25478	8869	34,8	26252	8838	33,7
16	1927	1259	65,3	2588	1603	61,9	2042	1127	55,2
18	5778	3490	60,4	3002	1817	60,5	3318	1667	50,2
20	3116	1883	60,4	2257	1416	62,7	2244	1198	53,4
22	5395	2935	54,4	7107	3705	52,1	6448	2992	46,4
24	12962	7297	56,3	12546	6460	51,5	10532	4354	41,3
26	2741	1745	63,7	2768	1712	61,8	3194	1704	53,4
28	3360	1705	50,7	3478	1815	52,2	4427	2079	47,0
30	10072	5740	57,0	8764	4771	54,4	11301	5030	44,5
32	5057	2731	54,0	6106	3135	51,3	4965	1980	39,9

Źródło: GUS.

Z zestawienia widać, że kompletność badania jest niska i zasadniczo mieści się w przedziale 40—60%. Co więcej, zróżnicowanie pomiędzy województwami oraz większa kompletność w przypadku mniejszej liczebności próbki wylosowanej pozwalają sądzić, iż przedstawiony degresywny model wykładniczy jest adekwatnym, przybliżonym opisem obserwowanej sytuacji.

W dalszej części przedstawiono zastosowanie modelu do alokacji próbki w badaniu małych przedsiębiorstw. Do obliczeń i symulacji wykorzystano dane z roku 2000, zarówno dane operatowe jak i wyniki tego badania.

Parametry a_j i b_j funkcji f_j były dla 16 województw w Polsce wyznaczone za pomocą procedury o nazwie *model,* dostępnej w systemie SAS (wersja 8.12). Wyznaczane są one iteracyjnie metodą najmniejszych kwadratów (Gaussa-Newtona) z uwzględnieniem warunku $b_j / a_j \leq 1$. Jako dane wejściowe została wprowadzona liczebność próbki wylosowanej (wartości zmiennej niezależnej x) i zrealizowanej (wartości zmiennej zależnej y) w badaniu małych przedsiębiorstw (SP3) z lat 1998—2000, w poszczególnych województwach, podane w tabl. 1. Wyniki estymacji parametrów a_j i b_j przedstawione są zaś w tabl. 2.

Oczywiście należy pamiętać, że do obliczenia wartości estymatorów można było wykorzystać jedynie dane pochodzące z trzech ostatnich lat. Dlatego konkretne wartości mogą być obarczone znacznym błędem losowym, przy czym błąd ten mniej znacząco wpływać będzie na alokację, natomiast szczególnie wyraźnie uwidacznia się w ostatniej kolumnie tabl. 2. Przypominamy, że iloraz $b_j / a_j \leq 1$ można traktować jako ocenę prawdopodobieństwa odpowiedzi dla pojedynczej jednostki w danym województwie i w zasadzie nie jest możliwe, aby jego realna wartość w badaniu małych jednostek gospodarczych była równa jeden.

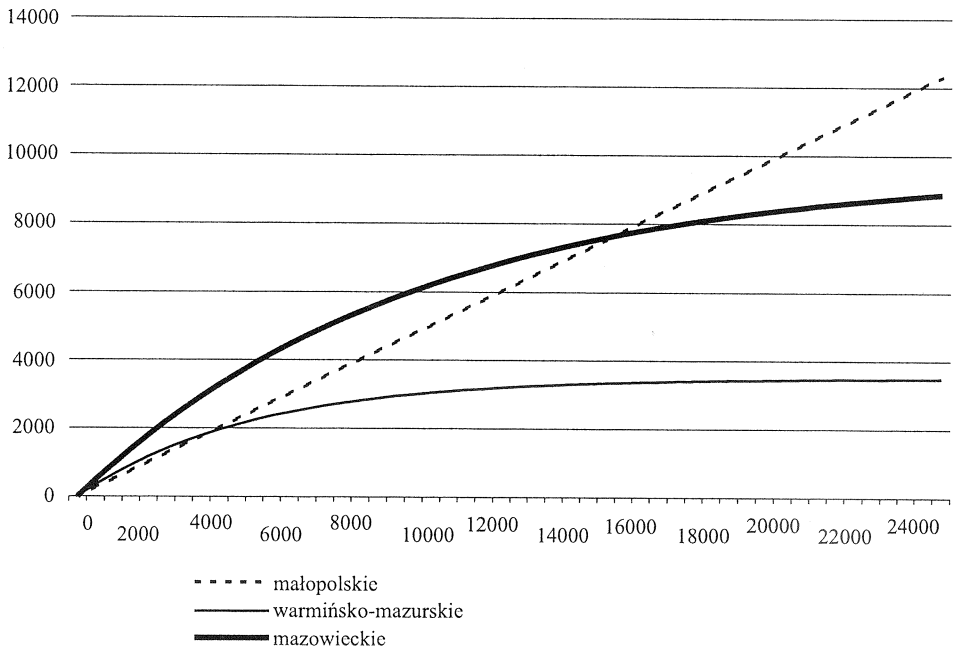
TABL. 2. WARTOŚĆ PARAMETRÓW FUNKCJI WYKŁADNICZEJ DLA POSZCZEGÓLNYCH WOJEWÓDZTW (podpopulacji)

Symbole województw j	a_j	b_j	b_j / a_j
02	5,5335E+03	5,5335E+03	1,00
04	3,3148E+03	3,3148E+03	1,00
06	4,6898E+10	2,8817E+10	0,61
08	2,0821E+03	2,0821E+03	1,00
10	3,9805E+09	1,8646E+09	0,47
12	2,0275E+13	1,0246E+13	0,51
14	9,6194E+03	9,6194E+03	1,00
16	7,5200E+10	4,5780E+10	0,61
18	8,3215E+11	4,8538E+11	0,58
20	8,0182E+10	4,7496E+10	0,59
22	1,8528E+04	1,1129E+04	0,60
24	2,4201E+12	1,2284E+12	0,51
26	2,4981E+03	2,4981E+03	1,00
28	4,9029E+03	3,5022E+03	0,71
30	6,6576E+03	6,6576E+03	1,00
32	8,6131E+12	4,2112E+12	0,49

Źródło: obliczenia własne.

Z kolei na wykresie przedstawiamy funkcję f_j, z wartościami parametrów podanymi w tabl. 2, dla trzech województw: małopolskiego (12), mazowieckiego (14) i warmińsko-mazurskiego (28).

Wykres 1. FUNKCJA f_j DLA j=12, 14 I 28 UZYSKANE DLA WARTOŚCI PARAMETRÓW PODANYCH W TABL. 2



Do obliczeń, oprócz parametrów a_j i b_j, konieczne było wyznaczenie oszacowań wariancji, wartości globalnych (w tym celu wykorzystano wyniki badania SP-3 za rok 2000, w zakresie cechy *liczba pracujących*) oraz liczebność w rozważanych przekrojach (w tym celu na potrzeby prowadzonych symulacji wykorzystano wartości z operatu, który służył do wylosowania próbki na rok 2000). W naszych eksperymentach, związanych z rozwiązywaniem opisanego wcześniej zadania optymalizacji, najefektywniejszy okazał się *solver* MOSEK. Należy podkreślić dużą złożoność numeryczną rozważanego problemu (ok. 3000 zmiennych i ponad 70 ograniczeń). Problem tego typu nie mógłby być rozwiązany za pomocą narzędzi dostępnych w popularnych pakietach statystycznych.

Wyniki zebrane są w tabl. 3, w której jednocześnie porównano je z rzeczywistą liczebnością próbki wylosowanej i zbadanej w badaniu za rok 2000.

TABL. 3. LICZEBNOŚĆ PRÓBEK WYLOSOWANYCH I ZBADANYCH JEDNOSTEK W POSZCZEGÓLNYCH WOJEWÓDZTWACH (podpopulacjach) W BADANIU MAŁYCH PRZEDSIĘBIORSTW W ROKU 2000 ORAZ WEDŁUG TESTOWANEGO MODELU

Województwa	Badanie	Testowany model	Różnica (a–b)	Badanie	Testowany model	Różnica (d–e)
	jednostki wylosowane			jednostki zbadane		
	a	b	c	d	e	f
02	10419	7536	2883	4560	4316	244
04	6758	5832	926	2752	2909	–157
06	3608	4844	–1236	1821	3133	–1312
08	3050	6759	–3709	1487	2147	–660
10	7209	9739	–2530	3276	4763	–1487
12	8416	9349	–933	4017	4884	–867
14	26252	12430	13822	8838	7183	1655
16	2042	3715	–1673	1127	2411	–1284
18	3318	5160	–1842	1667	3140	–1473
20	2244	3705	–1461	1198	2295	–1097
22	6448	8387	–1939	2992	4226	–1234
24	10532	13100	–2568	4354	6826	–2472
26	3194	3976	–782	1704	2108	–404
28	4427	4782	–355	2079	2316	–237
30	11301	7902	3399	5030	4829	201
32	4965	7256	–2291	1980	3757	–1777

Źródło: obliczenia własne.

W obu przypadkach próbki są mniej więcej tej samej wielkości: 114183 dla wylosowanej realnie oraz 114472 dla proponowanej. Liczebność próbki zrealizowanej wyniosła 48920, co stanowi ok. 43% próbki wylosowanej. Dla próbki proponowanej wielkość ta odpowiednio wynosi: 61243 oraz ok. 54%. Warto zwrócić uwagę na fakt, że zgodnie z oczekiwaniami nastąpiło przesunięcie wylosowanej próbki w kierunku województw, w których kompletność badania jest większa — szczególnie widoczne jest to dla woj. mazowieckiego oraz w nieco mniejszym stopniu dla województw: dolnośląskiego oraz wielkopolskiego. W wyniku takiej zmiany alokacji próbki następuje też zmiana precyzji dla badanej cechy w domenach. Zmiany te przedstawiono w kolejnej tablicy.

TABL. 4. PORÓWNANIE OCEN PRECYZJI PROCENTOWEJ DLA LICZBY PRACUJĄCYCH W BADANIU MAŁYCH PRZEDSIĘBIORSTW, UZYSKANYCH W DOMENACH W ROKU 2000 ORAZ DLA TESTOWANEGO MODELU

Symbole rodzajów działalności gospodarczej[a]	Badanie	Testowany model
	w %	
01	14,19	5,59
02	10,37	5,84
03	14,57	6,33
04	7,35	5,61
05	10,13	4,90
06	8,25	4,78
07	17,91	5,06
08	10,04	5,73
09	19,22	6,35
10	8,60	5,55
11	16,69	6,11
12	11,48	6,36
13	9,57	6,15
14	7,58	6,07
15	13,06	5,02
16	15,41	4,80
17	18,25	6,92
18	12,10	5,20
19	13,59	6,81
20	11,79	5,63
21	9,55	5,30
22	7,98	5,20
23	13,96	5,18
24	10,14	6,80
25	5,32	5,05
26	6,73	5,73
27	6,27	5,13
28	26,83	6,81
29	7,49	4,72
30	10,35	6,01
31	4,99	4,63
32	5,66	2,48
33	16,69	5,44
34	8,03	4,85
35	8,03	5,03
36	13,50	5,22
37	4,04	5,65
38	9,33	5,45
39	2,66	4,25
40	6,94	4,86
41	8,29	5,72
42	4,12	5,99
43	5,13	5,96
44	5,75	5,59
45	6,26	6,98
46	6,01	6,12
47	6,59	6,27
48	3,95	6,25
49	6,84	4,97
50	8,68	0,00
51	2,69	4,10
52	6,40	5,81
53	6,30	5,76
54	8,46	5,33
55	6,64	5,81

a Według klasyfikacji przyjętej w badaniu małych jednostek gospodarczych.

TABL. 4. PORÓWNANIE OCEN PRECYZJI PROCENTOWEJ DLA LICZBY PRACUJĄCYCH W BADANIU MAŁYCH PRZEDSIĘBIORSTW, UZYSKANYCH W DOMENACH W ROKU 2000 ORAZ DLA TESTOWANEGO MODELU (dok.)

Symbole rodzajów działalności gospodarczej[a]	Badanie	Testowany model
	w %	
56	8,16	5,83
57	1,91	5,03
58	4,23	5,84
59	24,92	3,91
60	8,86	4,89
61	14,30	5,23
62	10,81	6,51
63	4,36	4,68
64	7,16	5,39
65	12,86	4,52
66	6,57	5,67
67	11,77	3,59
68	2,86	5,09
69	5,95	6,24
70	8,31	5,41
71	9,58	6,65
72	8,73	7,10
73	11,88	7,09
74	49,66	6,24
75	10,34	6,76
76	6,61	4,65
77	6,42	6,17

a Według klasyfikacji przyjętej w badaniu małych jednostek gospodarczych.

Ź r ó d ł o: obliczenia własne.

Warto zwrócić uwagę, że zgodnie z przyjętym modelem do obliczeń przyjęto precyzję *p* wartości globalnej dla cechy *liczba pracujących* równą 8,8%. Tymczasem precyzja przedstawiona w tabl. 4 jest znacznie lepsza. Wynika to z faktu, że populacja badana różni się od populacji operatowej — liczebność populacji badanej przyjęto na podstawie uogólnień uzyskanych w badaniu za rok 2000. Tę samą liczebność zastosowano przy liczeniu precyzji wyników badania za rok 2000. Poza efektem przesunięcia próby i w rezultacie zwiększenia kompletności badania, dodatkowo wpływ na znaczące polepszenie precyzji miał fakt alokacji próbki, biorący pod uwagę zmienność cechy *liczba pracujących*, a nie cechy *przychody*, jak to miało miejsce przy alokacji próbki do badania za rok 2000.

dr hab. Wojciech Niemiro — *Uniwersytet Warszawski*, **dr Waldemar Popiński** — *GUS*, **dr hab. inż. Jacek Wesołowski** — *Politechnika Warszawska*, **dr Robert Wieczorkowski** — *GUS*

LITERATURA

1. Bracha Cz. (1996), *Teoretyczne podstawy metody reprezentacyjnej*. PWN, Warszawa
2. Greń J. (1966), *O pewnym zastosowaniu programowania nieliniowego do metody reprezentacyjnej*. „Przegląd Statystyczny", R. III
3. Fourer R., Gay D.M., Kernighan B.W. (1990), *A modelling language for mathematical programming*. „Management Science" 36, 519—554
4. Lednicki B., Wesołowski J. (1994), *Lokalizacja próby pomiędzy subpopulacje*. „Wiadomości Statystyczne" 9 (400), 2—4
5. Niemiro W., Wesołowski J. (2001), *Fixed precision optimal allocation in two-stage sampling*. „Applicationes Mathematicae" 28, 1, 73—82