

Problemy jakości danych statystycznych w przypadku badania cech rzadkich

W artykule przedstawiono szacowanie parametrów dla cech rzadkich oraz metody wyboru próby dla populacji rzadkich i trudno uchwytnych, które mogą stanowić alternatywę lub uzupełnienie badań nakierowanych na cechy powszechne. Zaprezentowano podstawowe schematy losowania dla cech rzadkich, w tym losowania: odwrotnego (*inverse sampling*), wielokrotnego połowu (losowanie typu pojmanie-uwolnienie, *capture-recapture sampling*), lokacyjnego (*location sampling*) oraz schematy linia-przecięcie (*line-intercept sampling*) i śledzenia łączy (*link-tracing design*).

Przedstawiono też koncepcję operatów wielokrotnych (*multiple frame*) oraz zagadnienie próby o zwiększonym pokryciu (*oversampling*). Trudności wynikające z zastosowania typowych schematów losowania zilustrowano przykładem symulacyjnym opartym na danych z Powszechnego Spisu Rolnego oraz danych z badania budżetów gospodarstw domowych. Wskazano na istnienie alternatywnych rozwiązań, wraz z określeniem korzyści i wad zastosowania tych metod.

Zgodnie z klasyfikacją wprowadzoną przez Purcella i Kisha¹, można wyróżnić cztery typy domen według ich wielkości w stosunku do całej badanej populacji. Według tej klasyfikacji mamy:

- dziedziny główne (*major*) — stanowiące 1/10 populacji,
- dziedziny drugorzędne (*minor*) — o liczebności od 1/10 do 1/100 populacji,
- minidziedziny (*mini*) — o wielkości od 1/100 do 1/10000 populacji,
- jednostki rzadko występujące (*rare types of individuals*) — dla których względna wielkość dziedziny jest mniejsza niż 1/10000.

W badaniach statystycznych nakierowanych na cechy powszechne liczba zaobserwowanych jednostek w zawężonej klasie (np. przez wiek czy region zamieszkania) czy też posiadających pewien rzadki atrybut (np. bezdomność) może być zbyt mała, aby osiągnąć zadowalającą precyzję szacunków. Standardowe metody, do których zalicza się estymatory Horvitz-Thompsona (HT),

¹ Purcell N. J., Kish L. (1979), s. 365—384.

można odnieść przede wszystkim do populacji stanowiących dziedziny główne. W odniesieniu do dziedzin drugorzędnych, jak i do minidziedzin konieczne jest zastosowanie innej techniki, np. estymacji pośredniej. Jeszcze trudniejsze jest szacowanie parametrów dla dziedzin lub cech rzadkich. Pojawiają się tu problemy dwojakiego rodzaju.

Po pierwsze metody estymacji pośredniej stosowane w statystyce małych domen często zawodzą w przypadku populacji rzadkich, gdyż ze względu na ich często odrębny charakter „zapożyczanie mocy” z pozostałej części próby nie znajduje uzasadnienia. **Po drugie**, jeśli jesteśmy zainteresowani nie tyle parametrami opisującymi cechy takiej podpopulacji, jak średnia czy wskaźnik struktury (np. udział osób poniżej linii ubóstwa dla rodzin posiadających dużą liczbę dzieci), ale liczebnością tej populacji lub sumą wartości cechy (np. liczba osób pobierających szczególny rodzaj świadczenia lub suma tych zasiłków), to pojawia się problem cechy rzadkiej. Chodzi tu o sytuację, gdy w próbie znajduje się jedynie znikoma lub niewielka liczba jednostek o wartościach cechy różnych od zera. Precyzja szacunków dla takich cech jest znacząco gorsza niż precyzja dla cech, które są w badaniu powszechnie obserwowane. Należy zauważyć, że problem cechy rzadkiej może dotyczyć zarówno małego, jak i dużego obszaru, przy czym w przypadku małego obszaru błędy szacunku mogą być oczywiście jeszcze większe.

Nasze uwagi dotyczą sytuacji, kiedy analizowane badanie statystyczne, w którym interesujące nas cechy mają charakter cech rzadkich, było nakierowane na cechy powszechne. Nie były więc one uwzględniane w schemacie losowania próby, np. jako warstwy o zwiększonym pokryciu. Można sobie jednak wyobrazić sytuację, kiedy ze względu na niezwykle istotny z punktu widzenia polityki społecznej czy badań naukowych charakter danego zjawiska, celowe jest przeprowadzenie odrębnego badania, w którym cechy te miałyby charakter powszechne. Chodzić tu może np. o badania mniejszości etnicznych i religijnych, osób bezdomnych czy też rzadkich chorób. W badaniach takich jednostki posiadające rzadki atrybut mogłyby stanowić, w określonych warunkach, populację celu. Istotnym problemem jest jednak dotarcie do tych populacji, gdyż często brak jest odpowiednich operatów losowania.

Dodatkowo pojawić się może problem, że badane jednostki są także „trudno uchwytnie”, co jest częste np. w badaniach marketingowych, gdy chcemy zbadać liczbę oraz opinie klientów sieci handlowej. Często uniemożliwia to zastosowanie klasycznej losowej techniki doboru próby. W celu zminimalizowania błędów konieczne jest zastosowanie specjalnej techniki doboru próby oraz estymacji.

W pierwszej części artykułu przedstawiono problemy estymacji dla cech rzadkich, wynikające z zastosowania typowych schematów losowania, które zilustrowano przykładem symulacyjnym opartym na danych z Powszechnego Spisu Rolnego 2010 (PSR 2010) oraz danych z badania budżetów gospodarstw domowych.

PRZYKŁADY OGRANICZEŃ W STOSOWANIU TECHNIKI W BADANIACH REPREZENTACYJNYCH W PRZYPADKU CECH RZADKICH

Ilustracja przedstawia wynik badania symulacyjnego wykonanego podczas przygotowań do PSR 2010. W eksperymencie użyto danych z powszechnego spisu rolnego w 2002 r. wraz z kryteriami doboru próby, przyjętymi w PSR 2010 dla badania małych gospodarstw². Jako punkt wyjścia wybrano populację małych gospodarstw woj. opolskiego (o powierzchni użytków rolnych do 1 ha) i niespełniających progów przyjętych w PSR 2010. Województwo to wybrano ze względu na jego wielkość i konieczność przeprowadzenia stosownych obliczeń odpowiednio szybko. Zdecydowano się — podobnie jak w badaniu struktury gospodarstw rolnych — na trzy warstwy, wraz z warstwą górną. Schemat zaprojektowano tak, aby granice warstw i ich liczebność wyznaczone były na podstawie optymalnego schematu warstwowego, przyjmując jako zmienną warstwową całkowitą powierzchnię gospodarstwa. Zastosowano tu m.in. schemat losowania zaproponowany przez Lednickiego i Wieczorkowskiego³.

Wyniki badania symulacyjnego gospodarstw rolnych dla jednej z gmin woj. opolskiego przedstawiono na wyk. 1—4. Wykr. 1 ilustruje rozkład dla cechy powszechnej (powierzchnia gospodarstwa ogółem). Widoczna jest tutaj normalność rozważanego rozkładu. Wyniki obliczeń wskazują też na brak obciążenia dla tak dobranej metody estymacji. Z kolei dla cechy rzadkiej (powierzchnia sadów), która została przedstawiona na wyk. 2, obserwuje się inną postać rozkładu obserwowanych wartości. Może to skutkować wystąpieniem obciążenia *in minus*. W takim bowiem przypadku wylosowanie jednostki o wartości wyższej od średniej jest zjawiskiem rzadszym niż wylosowanie jednostki o wartości cechy poniżej średniej. Wskutek tego, gdy wykonuje się losowania prostego w warstwie, częściej otrzymamy próby o wartości zaniżonej w stosunku do wartości parametru w populacji niż próby o wartości zawyżonej.

Kolejne dwa wykresy przedstawiają rozkład względnych błędów szacunku otrzymanych dla cechy powszechnej i cechy rzadkiej. Na podstawie wyk. 3 można zauważyć, że rozkład względnego błędu szacunku dla powierzchni gruntów ogółem jest w przybliżeniu normalny, co jest typowe dla cech powszechnych. W przypadku cech rzadkich rozkład ten ma wyraźnie odmienny charakter, najczęściej jest asymetryczny, a niekiedy wręcz nieregularny, co jest szczególnie widoczne dla takich cech i małych gmin. Prawdopodobnie tę uwagę zauważono także w przeprowadzonym badaniu symulacyjnym, gdzie rozkład względnego błędu szacunku dla powierzchni sadów jest silnie prawostronnie asymetryczny.

² Kubacki J., Piasecki T. (2009), s. 17—59.

³ Lednicki B., Wieczorkowski R. (2003), s. 287—305.

Co jednak najistotniejsze, na podstawie przedstawionych przykładów zauważyć można znaczne różnice w precyzji obu estymatorów, bowiem dla cechy powszechnej względne błędy szacunków wynoszą najczęściej ok. 5% szacowanych parametrów, podczas gdy dla cechy rzadkiej typowe wartości błędów to 60—70%. Dodatkowo należy podkreślić, że w wielu przypadkach estymacja precyzji była niemożliwa w wyniku zerowania się estymatora parametru lub estymatora jego wariancji.

Kolejne wykresy pokazują, z jakimi trudnościami można się zetknąć wykonując obliczenia w przypadku rzadziej występującej cechy, gdy dodatkowo potrzebne są oszacowania na poziomie małych domen, a więc podpopulacji, które nie były uwzględnione w schemacie losowania. Są one w próbie reprezentowane przez liczbę jednostek zbyt małą do osiągnięcia zadowalającej precyzji, nawet dla cech powszechnych. Przedstawione przykłady dotyczą rozkładu precyzji (względnego błędu szacunku) dla dochodu rozporządzalnego i zasiłku dla bezrobotnych w powiatach na podstawie badania budżetów gospodarstw domowych, otrzymanego z użyciem bezpośredniej techniki estymacji (wykr. 5 i 6). Wykr. 5 pokazuje, że względny błąd szacunku dla cechy powszechnej, jaką był w badaniu dochód rozporządzalny, jest zwykle poniżej 20%. Przy zastosowaniu metod szacunku dla małych obszarów może to pozwolić na oszacowanie wartości takiej zmiennej z zadowalającą precyzją.

Z kolei dla cechy rzadszej, jaką był w badaniu zasiłek dla bezrobotnych, szacunki dla takiej zmiennej mają zbyt małą precyzję — błąd względny jest zwykle większy niż 20%, a nierzadko przekracza nawet 100% szacowanych parametrów (wykr. 6). Tak duży błąd nie gwarantuje, że użycie nawet zaawansowanej pośredniej techniki estymacji dla małych obszarów pozwoli na uzyskanie wystarczająco precyzyjnych ocen.

Podejście modelowe jest często stosowane w statystyce małych obszarów wtedy, kiedy dysponujemy danymi dodatkowymi pozwalającymi lepiej wyznaczyć oceny szacowanych zmiennych w przypadku, gdy mamy wiarygodne dane pochodzące z innych źródeł, np. z rejestrów administracyjnych. W takim przypadku model statystyczny może pozwolić nie tylko na lepsze oszacowanie konkretnej zmiennej (charakteryzujące się np. mniejszym błędem szacunku), ale również pozwolić na określenie wartości takiej zmiennej dla jednostek niewystępujących w badaniu.

Do poprawy precyzji szacunków dochodu rozporządzalnego oraz zasiłku dla bezrobotnych na osobę uzyskanych na podstawie próby pochodzącej z badania budżetów gospodarstw domowych wykorzystano dane dodatkowe pochodzące z rejestru Poltax. W szczególności z systemu Poltax otrzymano następujące cechy dla powiatów dotyczące przychodu z: wynagrodzenia, emerytury i renty,

działalności wykonywanej osobiście, praw majątkowych, najmu, składki na ubezpieczenia społeczne i składki na ubezpieczenie zdrowotne. Wykorzystując agregatowe modele dla małych obszarów, uwzględniające te cechy, jako zmienne objaśniające oszacowano oba parametry dla powiatów, przy czym estymację przeprowadzono metodą EBLUP, EB i HB. Względne błędy szacunku dla cech powszechnej i rzadkiej w przypadku metod bezpośrednich i opartych na modelach przedstawiają wykry. 7 i 8.

Na podstawie wykry. 7 można zauważyć, że dla cechy powszechnej, jaką był w badaniu dochód rozporządzalny, użycie modeli dla małych obszarów okazało się skuteczne. Względny błąd szacunku, nawet w przypadku zastosowania estymacji bezpośredniej, nie przekraczał na ogół 15% i jedynie sporadycznie był większy niż 20%, natomiast przy użyciu estymatorów opartych na modelach uległ znacznej redukcji — przeciętnie o 25%.

Jednak dla cechy rzadszej przeprowadzane badanie potwierdziło obawy, że użycie techniki dla małych obszarów może okazać się niewystarczające w wielu przypadkach (wykr. 8). Mimo większej redukcji błędu szacunku, wynoszącej nawet 50%, tylko część szacunków ma akceptowalny poziom precyzji.

SCHEMATY LOSOWANIA DLA RZADKICH POPULACJI

Losowanie, którego celem jest dotarcie do populacji rzadkich bądź trudno uchwytnych może być żmudne oraz kosztowne. W literaturze występuje kilka schematów losowania nadających się do wykorzystania dla tego typu populacji. Są to schematy:

- losowania odwrotnego;
- losowania wielokrotnego połowu;
- losowania lokacyjnego;
- linia-przecięcie;
- śledzenia łączy, który obejmuje losowanie:
 - sieciowe,
 - metody „kuli śnieżnej”,
 - metody przypadkowego błędzenia,
 - adaptacyjne.

Losowanie odwrotne

Koncepcję losowania odwrotnego, które miało być techniką pozwalającą na losowanie dla cech rzadkich w populacji, wprowadził Haldane⁴. W przypadku badania zjawisk rzadkich, próba o ustalonej wielkości może nie pozwolić na wylosowanie dostatecznej liczby jednostek mających wymagany atrybut. Losowanie odwrotne polega na losowaniu jednostka po jednostce z użyciem schematu losowania prostego, dopóki nie wybierze się określonej liczby elementów posiadających daną cechę. Jest więc odwrotnie niż w klasycznych schematach losowania — to liczebność próby, a nie liczba wyróżnionych elementów jest zmienną losową. W pracy Haldane rozważano przypadek nieograniczonej populacji. Określono rozmiar próby oraz rozkład prawdopodobieństwa dla takiego przypadku. Przedstawiono też formuły na nieobciążony estymator dla proporcji oraz wariancję tego estymatora.

Losowanie metodą wielokrotnego połowu

Metoda wielokrotnego połowu była po raz pierwszy zaproponowana przez Sebera⁵ do szacowania populacji rzadkich gatunków zwierząt. W celu oszacowania całkowitej liczby jednostek w populacji posiadających dany rzadki atrybut pobiera się wstępną próbkę tych jednostek. Są one oznaczane i powracają do populacji. Następnie, niezależnie, wybiera się drugą losową próbkę i oblicza ponownie oznakowane jednostki. Jeśli druga próbka jest reprezentatywna dla całej populacji, to otrzymany udział oznakowanych jednostek powinien być taki sam, jak udział tych jednostek w całej populacji. Korzystając z tego założenia można w prosty sposób określić całkowitą liczbę jednostek w populacji.

Losowanie lokacyjne

Technika ta opisana jest w pracy Kaltona⁶ jako metoda losowania jednostek z populacji przemieszczających się, jak turyści lub odwiedzający centra handlowe, chorzy na HIV czy bezdomni. Zwykle stosuje się losowanie dwustopniowe, przy czym jednostki te konstruowane są jako kombinacje miejsc obserwacji i okresów. Dodatkowo na pierwszym stopniu stosuje się warstwowanie względem np. godzin otwarcia i rodzaju obiektu, a jednostki wybierane są proporcjonalnie do ich wielkości. Jednostki badania losowane są zazwyczaj systematycznie w ramach wylosowanych okresów.

Otrzymana próba jest próbą probabilistyczną z punktu widzenia cechy badanych jednostek (np. satysfakcji z odwiedzania danego sklepu czy muzeum).

⁴ Haldane J. B. S. (1945), s. 222—225.

⁵ Seber G. A. F. (1973).

⁶ Kalton G. (1991), s. 183—194.

Jednak, gdy celem badania jest poznanie liczebności danej rzadkiej populacji powstają problemy z określeniem prawdopodobieństwa inkluzji, związane z powtarzalnością obserwacji tych samych obiektów.

Losowanie metodą linia-przecięcie

Losowanie metodą linia-przecięcie⁷ odnosi się do losowania dla rzadkich populacji w celu określenia całkowitej liczebności takiej populacji. Jest to schemat losowania, w którym wyznacza się n linii przecięcia przez wybór n pozycji tych linii wzdłuż pewnej linii stanowiącej podstawę, która poprowadzona jest wzdłuż szerokości badanego obszaru, natomiast linie przecięcia są poprowadzone przez badany obszar prostopadle do linii będącej podstawą dla każdej z wybranych pozycji. Za każdym razem, gdy obiekt w populacji jest przecięty przez jedną lub więcej wylosowanych linii, wykonuje się zapis rozpatrywanej zmiennej. We wnioskowaniu o wielkości populacji można użyć estymatora HT, w którym prawdopodobieństwo inkluzji obiektu wynika z rzutu szerokości obiektu na linię stanowiącą podstawę odniesioną do długości tej linii⁸.

Losowanie metodą śledzenia łączy

Losowanie tą metodą polega na stosowaniu rozmaitej techniki, która w celu otrzymania próby wykorzystuje powiązania między jednostkami populacji. Dotyczą one często więzi rodzinnych lub społecznych pomocnych w identyfikowaniu przypadków rzadkich chorób lub dotarciu do trudno dostępnych populacji, takich jak np. bezdomni.

Schemat śledzenia łączy obejmuje następujące techniki pobierania próby, tj. losowanie:

- 1) sieciowe,
- 2) metodą „kuli śnieżnej”,
- 3) metodą przypadkowego błędzenia,
- 4) adaptacyjne.

Losowanie sieciowe

Schemat losowania sieciowego opisał po raz pierwszy socjolog James Coleman⁹ w kontekście badań struktur społecznych opartych na powiązaniach nieformalnych, a więc trudno uchwytnych dla badacza. Sieć w takim przypadku jest zbiorem obserwowanych jednostek wraz z określonym wspólnym wzorcem połączeń, przy czym jednostka może być połączona z więcej niż jedną siecią.

⁷ Lucas H. A., Seber G. A. F. (1977), s. 618—622.

⁸ Thompson S. K. (2012), s. 275—282.

⁹ Coleman J. S. (1958), s. 28—36.

Początkowo pobierana jest próba prosta lub warstwowa, a następnie dołączane są wszystkie jednostki powiązane z początkowo wylosowanymi. Nieobciążony estymator dla losowania sieciowego zaproponowano w pracy Birnbauma i Sirkena¹⁰. Autorzy ci podali także alternatywną formułę opartą na estymatorze HT.

Losowanie metodą „kuli śnieżnej”

Termin ten odnosi się do dwóch podobnych schematów pobierania próby, opartych na tzw. efekcie „kuli śnieżnej”. Pierwszy z nich był zaproponowany przez Goodmana¹¹ do badania struktur społecznych i polega na wylosowaniu pewnej liczby jednostek, które w następnym etapie wskazują określoną liczbę następnych itd., przy czym liczba etapów jest także określona z góry.

Druga wersja metody (stosowana w przypadku badania populacji rzadkich) była przedstawiona w pracy Kaltona i Andersona¹². Początkowo identyfikowana jest niewielka grupa jednostek, które potem wprowadzają do próby następne itd. Otrzymana w efekcie próba nieprobabilistyczna może być wykorzystana bezpośrednio do estymacji lub też służy jako operat do pobrania próby losowej.

Losowanie metodą przypadkowego śledzenia

Schemat błędzenia przypadkowego jest schematem losowania pozwalającym na otrzymanie próby probabilistycznej dla dużych sieci społecznych¹³. Jednostki losowane są do próby w kilku etapach zwanych falami. Jeśli na pewnym etapie wylosowano daną jednostkę, to w następnym losuje się jedną z jednostek wskazanych przez jej łącze. Procedurę powtarza się tak długo aż osiągnie się żadaną liczebność próby. W ten sposób otrzymujemy próbę losową o liczebności n . Do oszacowania średniej na podstawie tej próby można wykorzystać asymptotycznie nieobciążony estymator ilorazowy oparty na formule Hansena-Hurwitza.

Losowanie adaptacyjne

Adaptacyjny schemat losowania sieciowego to schemat, w którym decyzja o włączeniu do próby powiązanych z danym węzłem jednostek zależy od wartości zmiennej obserwowanej w węźle. Procedurę adaptacyjną zaproponowali Thompson i Ramsey w 1983 r.¹⁴. Thompson w 2006 r.¹⁵ przedstawił nowy schemat losowania nazwany adaptacyjnym losowaniem sieciowym (*adaptive*

¹⁰ Birnbaum Z. W., Sirken M. G. (1965), s. 1—8.

¹¹ Goodman L. A. (1961), s. 148—170.

¹² Kalton G., Anderson D. W. (1986), s. 65—82.

¹³ Klov Dahl A. S. i in. (1977), s. 169—172.

¹⁴ Thompson S. K., Ramsey F. L. (1983).

¹⁵ Thompson S. K. (2006), s. 1—24.

web sampling — AWS), który można zastosować w przypadku sieci oraz schematów przestrzennych. Wybór jednostek odbywa się sekwencyjnie z zastosowaniem mieszaniny rozkładów wykorzystujących zależności sieciowe lub przestrzenne, a także wartości otrzymane z próby. Schemat ten ma przewagę nad opracowanymi wcześniej schematami śledzenia łączy, gdyż pozwala m.in. na sterowanie rozmiarem próby.

Operat wielokrotny

W przypadku gdy rozważa się badanie cech rzadkich czasami skuteczne może być zastosowanie wielu wykazów jednostek będących przedmiotem badania. Często sytuacja taka zachodzi, gdy badane są osoby mające określone schorzenie względnie badane są specyficzne gospodarstwa rolne. W takim przypadku podczas estymacji uwzględnić trzeba liczebność wszystkich wykorzystywanych operatów, jak również występowanie ich części wspólnych. Istnienie informacji dodatkowych w próbie może też być pomocne przy stworzeniu oddzielnych warstw, których wyodrębnienie może poprawić jakość otrzymanych szacunków.

Próba o zwiększonym pokryciu (nadreprezentacja)

W celu zwiększenia frakcji losowania w pewnych warstwach populacji można wykorzystać zarówno informację dotyczącą lokalizacji geograficznej (np. w przypadku mniejszości etnicznych), jak również inne informacje dotyczące badanych jednostek (np. typ i wielkość gospodarstwa rolnego lub rodzaj działalności wykonywanej przez podmiot ekonomiczny oraz jego wielkość).

Przykłady:

- nadreprezentacja obszarów wiejskich w Badaniu Aktywności Ekonomicznej Ludności;
- badania rolnicze (w tym badanie struktury gospodarstw rolnych) — wyodrębnienie warstw dla gospodarstw hodowlanych i sadowniczych czy ekologicznych;
- schemat warstwowania w badaniu działalności gospodarczej przedsiębiorstw (SP-3) uwzględniający liczebność poszczególnych grup przedsiębiorstw oraz wyodrębnienie warstwy górnej ze względu na skalę przychodów.

POTENCJALNE MOŻLIWOŚCI ZASTOSOWANIA OMAWIANYCH METOD W PRAKTYCE

Analizując literaturę przedmiotu, wskazać można na przykłady zastosowań omawianych schematów losowania dla zjawisk rzadkich, które mogłyby być wykorzystane także w badaniach społecznych. Wspomnieć tutaj można o zastosowaniu metody losowania:

- typu wielokrotnego połówu¹⁶ oraz operatorów wielokrotnych¹⁷ w badaniu bezdomności;
- sieciowego w badaniu wiktylizacji¹⁸;
- sieciowego w badaniach zdrowia i niepełnosprawności¹⁹.

Z kolei koncepcja operatu wielokrotnego może znaleźć zastosowanie w badaniach społecznych oraz badaniach rolniczych. Wspomnieć tutaj można o rozpoczętych już pracach, prowadzonych przez GUS, w tym metodologicznych — budowa operatu statystycznego do badań społecznych oraz analiza możliwości zastępowania i łączenia danych administracyjnych oraz statystycznych z zakresu rolnictwa, a także opracowaniu FAO pt. *Multiple Frame Agricultural Surveys*²⁰.

Podsumowanie

Przedstawione przykłady wskazują na znacznie gorszą precyzję estymatorów parametrów dla cech rzadkich w porównaniu z powszechnymi. Częstym problemem jest niemożność estymacji precyzji szacunku w wyniku zerowania się estymatora parametru i estymatora jego precyzji, co jest związane z egzotycznością cechy. Z przedstawionego przeglądu metod losowania dla cech rzadkich wynika, że możliwe jest zastosowanie tych metod (często wykorzystujących metody biometryczne) także w przypadku badań społecznych. Zastosowanie metod, takich jak operaty wielokrotne oraz nadreprezentacja potwierdza skuteczność stosowania obu tych metod w praktyce. Wydaje się, że celowe jest doskonalenie metodologii statystycznej pozwalającej na pełniejszą ocenę zjawisk rzadkich. Wykorzystanie takich metod musi uwzględniać zarówno korzyści, jak i trudności (w tym dotyczące kosztów badania) związane z zastosowaniem takich metod.

dr hab. Alina Jędrzejczak — Uniwersytet Łódzki, **dr Jan Kubacki** — Urząd Statystyczny w Łodzi

LITERATURA

- Birnbaum Z. W., Sirken M. G. (1965), *Design of Sampling Surveys to Estimate the Prevalence of Rare Diseases: Three Unbiased Estimates*, Vital and Health Statistics, Washington D.C.: U.S. Government Printing Office
- Coleman J. S. (1958), *Relational Analysis: the Study of Social Organizations with Survey Methods*. *Human Organization*, No. 17, <http://sfaa.metapress.com/content/q5604m676260q8n7/>

¹⁶ Fattorini L., Ghellini G. (2009).

¹⁷ Lohr S. (2007), s. 3257—3264.

¹⁸ Czaja R., Blair J. (1988), s. 38—43.

¹⁹ Sirken M. G. (1998), s. 1—6.

²⁰ *Multiple...* (1996), s. 105—109.

- Czaja R., Blair J. (1988), *Using Network Sampling for Rare Populations: An Application to Local Crime Victimization Surveys*, Proceedings of the Survey Research Methods Section, American Statistical Association, http://www.amstat.org/sections/SRMS/Proceedings/papers/1988_007.pdf
- Fattorini L., Ghellini G. (2009), *Estimating the size of street-dwelling populations by means of mark-resighting counts: theoretical considerations and empirical results*, Proceedings of the XXVI IUSSP International Population Conference, <http://iussp2009.princeton.edu/abstracts/90984>
- Goodman L. A. (1961), *Snowball Sampling*, „Annals of Mathematical Statistics”, No. 32, <http://projecteuclid.org/DPubS?service=UI&version=1.0&verb=Display&handle=euclid.aoms/1177705148>
- Haldane J. B. S. (1945), *On a Method of Estimating Frequencies*, „Biometrika”, No. 33 (3), <http://www.jstor.org/stable/2332299>
- Kalton G. (1991), *Sampling flows of mobile human populations*, „Survey Methodology”, No. 17
- Kalton G., Anderson D. W. (1986), *Sampling Rare Populations*, „Journal of the Royal Statistical Society”, No. 149
- Klov Dahl A. S. et al. (1977), *Social Networks in an Urban area: First Canberra Study*, „Australian and New Zealand Journal of Sociology”, No. 3(2)
- Kubacki J., Jędrzejczak A., Piasecki T. (2011), *Wykorzystanie metod statystyki małych obszarów do opracowania wyników badań statystycznych*, raport z pracy metodologicznej 3.065, Ośrodek Statystyki Matematycznej, Urząd Statystyczny w Łodzi
- Kubacki J., Piasecki T. (2009), *Opracowanie modelu/estymacji danych w PSR 2010 z wykorzystaniem danych z systemów informacyjnych i badań statystycznych*, Ośrodek Realizacji Badań, Urząd Statystyczny w Łodzi, raport powstały w ramach prac podgrupy ds. statystyczno-matematycznych w ramach prac przygotowawczych do PSR 2010
- Lednicki B., Wiecek R. (2003), *Optimal Stratification and Sample Allocation Between Subpopulations and Strata*, „Statistics in Transition”, October 2003, Vol. 6, No. 2
- Lohr S. (2007), *Recent Developments in Multiple Frame Surveys*, Proceedings of the Survey Research Methods Section, American Statistical Association, <http://www.amstat.org/sections/SRMS/Proceedings/y2007/Files/JSM2007-000580.pdf>
- Lucas H. A., Seber G. A. F. (1977), *Estimating Coverage and Particle Density Using the Line Intercept Method*, „Biometrics”, No. 64, <http://www.jstor.org/stable/2345342>
- Multiple Frame Agricultural Surveys* (1996), FAO, Vol. 1
- Purcell N. J., Kish L. (1979), *Estimation for Small Domains*, „Biometrics”, No. 35, <http://www.jstor.org/stable/2530340>
- Seber G. A. F. (1973), *The Estimation of Animal Abundance*, London, Griffin
- Sirken M. G. (1998), *A Short History of Network Sampling*, Proceedings of the Survey Research Methods Section, American Statistical Association, http://www.amstat.org/sections/SRMS/Proceedings/papers/1998_001.pdf
- Thompson S. K. (2006), *Adaptive web sampling*, „Biometrics”, No. 62, <http://www.jstor.org/stable/4124545>
- Thompson S. K. (2012), *Sampling*, 3rd ed., Hoboken, Wiley
- Thompson S. K., Ramsey F. L. (1983), *Adaptive sampling of animal populations*, „Technical Report”, No. 82, Corvallis, OR, Department of Statistics, Oregon State University

SUMMARY

The authors presented estimation of parameters for rare characteristics as well as methods of the sample selection to the rare and hardly perceptible populations, which may provide an alternative or complement to research aimed at

the common characteristics. The basic sampling schemes for rare characteristics, including the inverse sampling, capture-recapture sampling, location sampling, as well as line-intercept sampling and link-tracing design were described.

Also the multiple frame concept and the trials of increased coverage are presented in the article. Difficulties arising from the use of conventional sampling schemes were exemplified with a simulation based on data from the Agricultural Census and household budget surveys. The existence of alternatives for such cases, along with the advantages and disadvantages of those solutions are pointed, too.

РЕЗЮМЕ

Авторы статьи представили оценку параметров для редких признаков, а также методы определения выборки для редких и неуловимых популяций, которые могут быть альтернативой или дополнением обследований направленных на общие признаки. Статья характеризует основные схемы выборки для редких признаков, в том числе для обратной выборки, схему выборки многократной ловли (типа захват-освобождение), расположительную выборку, а также схемы линия пересечения и отслеживания ретранслятора.

Была представлена также концепция многократных инструментариев выборки, а также вопрос выборки с увеличенным покрытием. Трудности, возникающие из использования типичных схем выборки были изображены симуляционным примером, опирающимся на данных из всеобщей сельскохозяйственной переписи и обследования бюджетов домашних хозяйств. Было указано на существование альтернативных решений для таких случаев вместе с определением преимуществ и недостатков этих решений.