Wiadomości Statystyczne. The Polish Statistician, 2019, vol. 64, 10, 7–35 DOI: 10.5604/01.3001.0013.7589
Studia metodologiczne / Methodological studies

Estymacja liczby cudzoziemców w Polsce z wykorzystaniem metody capture-recapture[1]

Maciej Beręsewicz[a], **Grzegorz Gudaszewski**[b], **Marcin Szymkowiak**[a]

Streszczenie. Celem artykułu jest przedstawienie metody badań oraz wyników szacunku populacji cudzoziemców przebywających w Polsce. W badaniu wykorzystano administracyjne źródła danych. Zastosowano metodę capture-recapture bazującą na modelach log-liniowych. Szacuje się, że w 2015 i 2016 r. na terenie Polski mogło przebywać odpowiednio ok. 500 tys. (przyjmując 95-procentowy przedział ufności – od 369 tys. do 724 tys.) oraz ok. 744 tys. (od 601 tys. do 943 tys.) cudzoziemców. Jest to pierwsza tego typu kompleksowa analiza dotycząca próby estymacji liczby cudzoziemców w Polsce, która wpisuje się w nurt badań nad populacjami trudnymi do zbadania. Należy jednak mieć na uwadze konieczność spełnienia założeń tej metody, co również stanowił przedmiot rozważań autorów.

Słowa kluczowe: estymacja liczby cudzoziemców, populacja trudna do zbadania, metoda capture-recapture, analiza log-liniowa, rejestry administracyjne

Estimation of the number of foreigners in Poland using the capture-recapture method

Abstract. The aim of this paper is to present the methodology and the results of the estimation of the number of foreigners staying in Poland. Administrative data sources were used in the research. The authors adopted the capture-recapture method based on log-linear models. As a result, the number of foreigners staying in Poland in 2015 and 2016 has been estimated at around 500,000 persons (95% CI: 369,000–724,000) and around 744,000 persons (601,000–943,000), respectively. The study is the first comprehensive analysis of this kind which aims at estimating the number of foreigners in Poland, and thus fits in the current of research on hard-to-survey populations. It has to be remembered, however, that capture-recapture method requires close observance of its strict rules in order to be effective, which is also discussed in depth in the paper.

Keywords: estimation the number of foreigners in Poland, hard-to-survey population, capture-recapture method, log-linear analysis, administrative registers

JEL: C81, J61, J68, F22

[1] Artykuł został przygotowany na podstawie raportu podsumowującego pracę badawczą *Cudzoziemcy na krajowym rynku pracy w ujęciu regionalnym* zrealizowaną w ramach projektu *Wsparcie systemu monitorowania polityki spójności w perspektywie finansowej 2014–2020 oraz programowania i monitorowania polityki spójności po 2020 roku* współfinansowanego przez Unię Europejską ze środków Programu Operacyjnego Pomoc Techniczna 2014–2020.

[a] Uniwersytet Ekonomiczny w Poznaniu, Instytut Informatyki i Ekonomii Ilościowej.

[b] Główny Urząd Statystyczny, Departament Badań Demograficznych.

© Główny Urząd Statystyczny / Statistics Poland

Coraz częściej, zarówno na szczeblu rządowym, samorządowym, jak i lokalnym, porusza się kwestię konieczności ustalenia rzeczywistej liczby cudzoziemców przebywających w Polsce stale i czasowo oraz podejmujących tu pracę. Szczególnie istotną informację dla realizowania polityki ludnościowej, migracyjnej i gospodarczej stanowią cechy demograficzno-społeczne i ekonomiczne cudzoziemców. Ważne jest także określenie skali imigracji nierejestrowanej, tj. pozostającej poza ewidencją. Obecnie nie ma w Polsce miarodajnego i bezpośredniego źródła, które dostarczałoby wiarygodnych danych na ten temat. Należy również zwrócić uwagę, że imigracja cudzoziemców jest zjawiskiem zróżnicowanym przestrzennie, zwłaszcza w kontekście regionalnych rynków pracy.

Pozyskanie danych dotyczących liczby cudzoziemców, w tym nierejestrowanych imigrantów, to dla służb statystyki publicznej w Polsce istotne wyzwanie metodologiczne. Po pierwsze, rejestry administracyjne dostarczają informacji o populacji *de iure* (zarejestrowanej), podczas gdy statystyka zainteresowana jest populacją *de facto* (zarejestrowani i niezarejestrowani). Po drugie, cudzoziemcy są populacją trudną do zbadania przy wykorzystaniu tradycyjnych metod statystycznych. Populacja taka charakteryzuje się bowiem brakiem dostępnego (wyczerpującego) operatu losowania oraz trudnością w pozyskaniu informacji od jednostek do niej należących. O ile rozpoznanie problemów występujących w populacjach trudnych do zbadania jest możliwe na gruncie badań statystycznych (można zastosować przykładowo dobór jednostek do badania, opierając się na metodzie kuli śnieżnej i jej rozszerzeniu – metodzie RDS[2]), o tyle proces estymacji wielkości takiej zbiorowości jest z metodologicznego punktu widzenia poważnym wyzwaniem badawczym. W literaturze przedmiotu opisano jednak odpowiednie metody statystyczne, które umożliwiają estymację wielkości populacji trudnych do zbadania, bazujące na technikach capture-recapture (Böhning, van der Heijden i Bunge, 2017). Należą do nich rozwiązania, w których wykorzystuje się jedno (Godwin i Böhning, 2017; van der Heijden, Cruyff i van Houwelingen, 2003) albo co najmniej dwa źródła danych (van der Heijden, Whittaker, Cruyff, Bakker i van der Vliet, 2012; Zhang, 2008). Skuteczność tych technik w praktyce w dużej mierze zależy od dostępności danych statystycznych i jest uwarunkowana koniecznością spełnienia odpowiednich założeń leżących u podstaw poszczególnych metod.

Celem artykułu jest przedstawienie metody badań oraz wyników szacunku populacji cudzoziemców przebywających w Polsce. Podjęto w nim próbę osza-

[2] Respondent Driven Sampling – metoda doboru jednostek do próby sterowana przez respondentów. Jest to zmodyfikowana wersja metody kuli śnieżnej, w której stosuje się podwójny system zachęt polegający na wynagrodzeniu respondenta za wzięcie udziału w badaniu i zwerbowaniu kolejnych osób, które biorą w nim udział. W metodzie RDS wykorzystuje się informacje na temat sieci powiązań osób należących do danej zbiorowości.

cowania wielkości populacji cudzoziemców przebywających w Polsce pod koniec 2015 i 2016 r. i według kraju obywatelstwa[3]. W tym celu posłużono się odpowiednio zbudowanym modelem log-liniowym z szeregiem zmiennych pomocniczych. Przyjęto, zgodnie z Ustawą z dnia 12 grudnia 2013 r. o cudzoziemcach (Dz.U. 2016 poz. 1990 z późn. zm.), że cudzoziemcem jest osoba nieposiadająca obywatelstwa polskiego lub bezpaństwowiec. Należy zaznaczyć, że wybór okresu oraz źródeł danych był podyktowany ich dostępnością dla statystyki publicznej w ramach programu badań statystycznych statystyki publicznej (PBSSP). Pozyskanie innych danych jednostkowych było niemożliwe. Udostępnione dane przedstawiają stan na 31 grudnia 2016 r. Nie oznacza to, że w tym dniu wszyscy badani cudzoziemcy przebywali na terenie Polski, podobnie jak w przypadku rejestru PESEL, który nie gwarantuje, że na terenie Polski w danym dniu przebywa określona liczba obywateli polskich.

PRZEGLĄD LITERATURY

Pojęcie populacji trudnej do zbadania

W literaturze przedmiotu populacje trudne do zbadania są definiowane na różne sposoby (Tourangeau, Edwards, Johnson, Wolter i Bates, 2014). Ze względu na to, że w wielu badaniach częściowych mamy do czynienia z dużym odsetkiem odmów, terminem tym można byłoby określić każdą z badanych populacji. Jednak pojęcie populacji trudnej do zbadania ma inne znaczenie i odnosi się do zbiorowości, których badanie wiąże się ze szczególnymi wyzwaniami metodologicznymi różnego rodzaju i jest bardziej skomplikowane w porównaniu z badaniem innych populacji. Są to np. populacje rzadkie, ukryte, z których jednostkami trudno nawiązać kontakt i współpracować.

Mówiąc o populacjach trudnych do zbadania, należy rozróżnić populacje (Tourangeau i in., 2014):

- z których trudno wylosować jednostki do próby (ang. *hard-to-sample*) – bardzo rzadko zdarza się, aby istniał właściwy operat losowania, z którego można byłoby wylosować jednostki do próby, wykorzystując odpowiedni schemat jej pobierania. Z tego względu stosuje się nielosowe dobory jednostek do próby, wśród których szczególną rolę odgrywa wspomniana już metoda kuli śnieżnej czy doboru sterowana przez respondenta. Można również zastoso-

[3] W pracy badawczej, na podstawie której powstał niniejszy artykuł, rozważany był również poziom województw i podregionów. Opracowano także podstawową charakterystykę cudzoziemców uwzględniającą wybrane cechy demograficzno-społeczne, obywatelstwo czy status na rynku pracy na podstawie danych z Narodowego Spisu Powszechnego Ludności i Mieszkań 2011, Badania Aktywności Ekonomicznej Ludności oraz zezwoleń i oświadczeń Ministerstwa Rodziny, Pracy i Polityki Społecznej, które nie będą jednak omawiane w tym miejscu.

wać inne techniki doboru jednostek, zwłaszcza w odniesieniu do populacji rzadkich i trudno uchwytnych, takie jak losowanie odwrotne, lokacyjne czy schematy linia-przecięcie oraz śledzenia łączy (Jędrzejczak i Kubacki, 2014). Jednak nawet w takim przypadku dobór jednostek do próby jest problematyczny, gdyż omawiane populacje mogą być mobilne bądź nieuchwytne. Przykład: osoby bezdomne lub pracujący cudzoziemcy;

- których jednostki trudno zidentyfikować (ang. *hard-to-identify*) – w niektórych przypadkach, zwłaszcza stygmatyzowanych grup społecznych, członkowie populacji mogą nie chcieć udostępnić swoich danych z obawy przed ujawnieniem nielegalnego lub wstydliwego statusu społecznego, co uniemożliwia identyfikację jednostek należących do takich populacji. Przykład: narkomani, alkoholicy, mniejszości (LGBT, wyznawcy określonych religii czy ideologii);
- których jednostki trudno odnaleźć i z którymi trudno nawiązać kontakt (ang. *hard-to-find-and-contact*) – nawiązanie kontaktu komplikuje przede wszystkim mobilność tego typu populacji. Przykład: niezameldowani cudzoziemcy, członkowie kultur koczowniczych (Beduini z południowo-zachodniej Azji czy Tuaregowie z Afryki Północnej), mniejszości wędrowne (Romowie w Europie), osoby bezdomne;
- których jednostki trudno namówić do wzięcia udziału w badaniu (ang. *hard-to--persuade*) – niechęć do udziału w badaniu może wynikać z drażliwości poruszanej tematyki bądź z braku czasu. Przykład: aktywni zawodowo, pracujący w szarej czy czarnej strefie, cudzoziemcy;
- których jednostki można zachęcić do wzięcia udziału w badaniu, ale trudno z nimi przeprowadzić wywiad (ang. *hard-to-interview*) – przeszkodą w przeprowadzeniu wywiadu może być konieczność uzyskania zgody przełożonego, opiekuna prawnego czy rodzica na udział danej jednostki w badaniu, np. w przypadku osoby z niepełnosprawnością lub nieposługującej się językiem, w którym przygotowano kwestionariusz, a także gdy badanie należy przeprowadzić w obszarze konfliktu zbrojnego. Przykład: więźniowie, osoby z niepełnosprawnością intelektualną albo psychiczną czy cudzoziemcy nieznający języka danego kraju.

Jak pokazują powyższe rozważania, problemy związane z badaniem określonych populacji mogą wynikać z wielu czynników. Tak jest w przypadku cudzoziemców w Polsce, spośród których trudno wylosować jednostki do próby (brak pełnego operatu losowania oraz kompleksowych źródeł danych statystycznych, z których można czerpać wiedzę na temat cudzoziemców), z którymi trudno nawiązać kontakt (mobilność cudzoziemców na rynku pracy oraz brak stałego miejsca zamieszkania) i przeprowadzić wywiad ze względu na barierę językową. Czynniki te powodują również, że estymacja liczebności tego typu populacji, zwłaszcza z uwzględnieniem dodatkowych przekrojów, jest niezwykle złożonym zadaniem. W literaturze przedmiotu proponuje się jednak pewne rozwiązania,

które mogą być remedium na problemy związane z określeniem rzeczywistych rozmiarów populacji trudnych do zbadania. Należą one do grupy technik określanych terminem capture-recapture[4].

Szacowanie wielkości populacji trudnej do zbadania z wykorzystaniem metody capture-recapture

Metoda capture-recapture wywodzi się z nauk przyrodniczych. Pierwotnie użyto jej do oszacowania liczby ryb w jeziorze (Goudie i Goudie, 2007). Idea tego podejścia polega na tym, że w typowym badaniu z obszaru nauk przyrodniczych przeprowadzanym metodą capture-recapture na analizowanym terytorium umieszcza się pułapki lub siatki w celu wielokrotnego wyłapywania osobników danej populacji. W pierwszej próbie zostaje złowiona pewna liczba osobników, które po oznakowaniu są wypuszczane na wolność. W każdej kolejnej próbie zapisuje się i znakuje każde nieoznaczone zwierzę, notuje się każde zwierzę, które zostało wcześniej oznakowane, i ponownie wypuszcza się je na wolność. Po zakończeniu badania uzyskuje się pełną historię złowień dla każdego osobnika. Badania tego typu określane są mianem mark-recapture, tag-recapture czy multiple-record system.

W najprostszej wersji metoda capture-recapture składa się z dwóch prób lub źródeł[5]. Pierwsza to próba zawierająca osobniki złowione za pierwszym razem, druga – zwierzęta złowione za drugim razem. Ten szczególny przypadek złożony z dwóch prób w kontekście szacowania błędu niedostatecznego pokrycia określany jest jako system podwójny (ang. *dual system*) lub system podwójnego zapisu (ang. *dual-system record*). Od wielu lat metodę wielokrotnych złowień stosuje się do szacowania parametrów demograficznych w populacjach zwierzęcych. Biolodzy już dawno zauważyli, że nie jest konieczne ani nawet możliwe zliczenie wszystkich zwierząt w celu dokładnego oszacowania wielkości populacji. Informacja na temat liczby ponownych złowień (lub proporcji ponownych złowień) uzyskiwana poprzez znakowanie odgrywa tu istotną rolę, ponieważ można ją wykorzystać do oszacowania liczby osobników nieujętych w próbach, przyjmując odpowiednie założenia. W najprostszym ujęciu można założyć, że w przypadku gdy liczba ponownie złowionych osobników w kolejnych próbach jest niewielka, rozmiar populacji jest większy niż liczba unikatowych osobników, które zostały złowione. Natomiast jeśli wskaźnik ponownych złowień jest stosunkowo wysoki, można przypuszczać, że złowiono większość zwierząt

[4] Autorzy posługują się angielskim terminem *capture-recapture*, który w dosłownym tłumaczeniu na język polski mógłby brzmieć „metoda wielokrotnego połowu". Określenie to nie oddaje jednak istoty tego podejścia, w szczególności w odniesieniu do rejestrów administracyjnych, w których nie dokonano losowań czy „połowów" jednostek.

[5] Możliwe jest również zastosowanie metody w przypadku jednego źródła, o czym będzie mowa później.

z danej populacji. Pomysł zastosowania techniki złożonej z dwóch prób można odnaleźć w pracach Pierre'a Simona de Laplace'a z 1786 r., który wykorzystał ją do szacowania liczby ludności Francji w 1802 r., a nawet wcześniej, w pracach Johna Graunta, który za jej pomocą oszacował skutki zarazy wśród ludności Anglii ok. 1600 r. W dziedzinie ekologii techniki tej najwcześniej użyto w badaniach Petersena i Dahla dotyczących populacji ryb odpowiednio w 1896 i 1907 r. oraz w przeprowadzonym przez Lincolna w 1930 r. badaniu powrotów zaobrączkowanych ptaków wodnych. Modele oparte na dwóch próbach zostały rozszerzone na przypadki zawierające większą liczbę prób przez Schnabela w 1938 r., stąd też metoda wielokrotnego połowu nazywana jest również spisem Schnabela. Bardziej zaawansowana teoria statystyczna i procedury wnioskowania w kontekście populacji trudnych do zbadania pojawiły się po publikacji prac Darrocha, który opracował zagadnienie pod względem matematycznym (Böhning, van der Heijden i Bunge, 2018).

Założenia stosowane w odniesieniu do populacji zwierzęcych klasyfikuje się generalnie jako modele zamknięte i otwarte. W modelu zamkniętym zakłada się, że wielkość populacji będącej przedmiotem badania jest stała w okresie prowadzenia badania. Założenie to jest zwykle zachowane w przypadku danych zbieranych w stosunkowo krótkim czasie poza okresem godowym. W modelu otwartym dopuszcza się przyrosty (narodziny lub imigracja) lub ubytki (śmierć lub emigracja) w populacji. Założenie otwartej populacji zwykle wykorzystuje się w długoterminowych badaniach zwierząt lub ptaków wędrownych. Poza wielkością populacji w momencie poszczególnych prób badane parametry obejmują również wskaźnik przeżywalności oraz liczbę narodzin pomiędzy próbami.

Warto zaznaczyć, że współcześnie pojęcie capture-recapture jest szerokie i odnosi się do wielu metod mających na celu oszacowanie wielkości nieznanej populacji. Zwykle wykorzystuje się różnego rodzaju narzędzia statystyczne, np. modele log-liniowe, modele klas ukrytych czy uogólnione modele liniowe. W prezentowanym badaniu zastosowano metodę capture-recapture wykorzystującą analizę log-liniową. Natomiast warto pamiętać, że wybór odpowiedniej techniki w estymacji liczebności populacji trudnych do zbadania podyktowany jest w dużej mierze liczbą dostępnych źródeł, którą można podzielić na przypadki wyłącznie jednego albo dwóch lub więcej źródeł.

Kluczowym aspektem metody capture-recapture są założenia, których niespełnienie skutkuje obciążonymi szacunkami wielkości populacji. W przypadku jednego źródła zakładamy: (1) możliwość zidentyfikowania jednostek, (2) wielokrotną obserwację jednostek (np. dana osoba popełniła więcej niż jedno przestępstwo), (3) stałość populacji w czasie, (4) określony rozkład prawdopodobieństwa wielokrotnego wystąpienia w zbiorze danych (np. ucięty rozkład Poissona) oraz (5) niezależność kolejnych obserwacji (van der Heijden i in., 2003). Założenie o niezależności zdarzeń jest bardzo restrykcyjne i w praktyce rzadko możliwe do spełnienia (Zhang, 2008). Dlatego Godwin i Böhning (2017) zapro-

ponowali wykorzystanie dodatniego rozkładu Poissona z podwyższoną liczbą jedynek do opisu liczby wystąpień w jednym źródle, będącego wynikiem: (1) nauczenia się przez badane jednostki, jak być nierozpoznanym/uniknąć złapania lub (2) nieprzyjemności związanych z pierwszym zdarzeniem i niechęci do powtórzenia sytuacji.

W kontekście dwóch lub więcej źródeł Wolter (1986) zdefiniował następujące założenia: (1) definicje populacji we wszystkich źródłach są takie same (tj. każda jednostka z populacji ma dodatnie prawdopodobieństwo pojawienia się w wybranych źródłach), (2) populacja jest zamknięta (tj. stała w danym czasie), (3) źródła danych są niezależne, (4) brak błędów pokrycia i duplikatów, (5) brak błędów łączenia między źródłami (tj. łączenie następuje na podstawie identyfikatora) oraz (6) prawdopodobieństwo włączenia do co najmniej jednego z rejestrów powinno być jednorodne. Spełnienie tych założeń jest kluczowe w kontekście możliwości stosowania omawianych metod zarówno w przypadku dwóch, jak i wielu źródeł. Wrażliwość estymatorów wielkości populacji na złamanie powyższych założeń jest obecnie poddawana dyskusji w literaturze poświęconej statystyce publicznej (Di Cecco, Di Zio, Filipponi i Rocchetti, 2018; Di Consiglio i Tuoto, 2015; Gerritse, 2016; Gerritse, van der Heijden i Bakker, 2015; Griffin, 2014; Zhang, 2015; Zhang i Dunne, 2018). Z powodu ograniczonego miejsca nie podjęto w artykule próby oceny wrażliwości na niespełnienie powyższych założeń. Planowane jest to w przyszłych pracach autorów.

W kontekście statystyki publicznej metodę capture-recapture wykorzystuje się do oceny jakości spisów w ramach badań pospisowych czy spisów kontrolnych (ang. *post-enumeration surveys* – PES albo *Census Coverage Survey* – CSS). Ocena ta polega na przeprowadzeniu niezależnego badania reprezentacyjnego w celu określenia pokrycia spisu (Gołata, 2018). Przykładowo w przypadku Narodowego Spisu Powszechnego 2002 oraz 2011 wykorzystano spisy kontrolne, jednakże ich wyniki nie zostały opublikowane (Gołata, 2012).

Metodę capture-recapture zaadaptowano także do określenia wielkości populacji wyłącznie na podstawie rejestrów administracyjnych. W takim wypadku można ją znaleźć pod pojęciem dualnej metody estymacji (ang. *dual-system estimation* – DSE), jeżeli wykorzystuje się dwa źródła danych, czy potrójnej metody estymacji (ang. *triple-system estimation* – TSE) w przypadku trzech źródeł. Na przykład Zhang i Dunne (2018) rozważali zastosowanie metody capture-recapture do estymacji populacji Irlandii na podstawie rejestru aktywności osób (Person Activity Register) będącego wynikiem łączenia 10 rejestrów administracyjnych według podejścia opartego na znakach życia (ang. *signs-of-life*) oraz ewidencji praw jazdy. Bakker, van der Heijden i Gerritse (2017) podjęli próbę estymacji liczby niezarejestrowanych rezydentów w Holandii, posługując się trzema źródłami danych: rejestrem ludności, rejestrem zatrudnionych oraz rejestrem podejrzanych o przestępstwa prowadzonym przez policję. W celu oszacowania wielkości populacji niezarejestrowanych rezydentów wykorzystano od-

powiednio zbudowany model log-liniowy, uwzględniając zmienne pomocnicze w postaci czasu pobytu, płci oraz wieku, wcześniej dokonawszy deterministycznego i probabilistycznego łączenia rekordów z trzech wspomnianych źródeł danych. Przeprowadzono również analizę wrażliwości uzyskanych wyników na przypadek występowania błędów połączenia oraz poprawności procesu parowania jednostek.

W literaturze można znaleźć również wiele innych przykładów wykorzystania omawianej metody do szacunku liczebności specyficznych subpopulacji, np. liczby bezdomnych (Coumans, Cruyff, van der Heijden, Wolf i Schmeets, 2017; Hudson, 1998; Schepers i Nicaise, 2017), narkomanów (Bouchard, 2007, 2008; Bouchard i Tremblay, 2005; van der Heijden, Cruts i Cruyff, 2013; Rossi i Mascioli, 2008), nietrzeźwych kierowców (van der Heijden i in., 2003), ofiar konfliktów (Chen, Shrivastava i Steorts, 2018) czy liczby cudzoziemców. Ciekawy przegląd zastosowań metody capture-recapture w estymacji liczebności populacji trudnych do zbadania przedstawili Godwin i Böhning (2017).

Metoda capture-recapture w estymacji liczby cudzoziemców

Van der Heijden i współpracownicy (2003) rozważali wykorzystanie jednego źródła danych do estymacji liczby cudzoziemców nielegalnie przebywających w 1995 r. w Amsterdamie, Rotterdamie, Hadze oraz Utrechcie, którzy nie zostali skutecznie wydaleni z Holandii. Cudzoziemcy ci byli wielokrotnie obserwowani w zbiorach danych policji. Do estymacji wielkości tak zdefiniowanej populacji badacze zastosowali rozkład Poissona ucięty w zerze oraz odpowiadający mu uogólniony model liniowy (ang. *zero-truncated Poisson regression model*), wykorzystując następujące zmienne pomocnicze: wiek (do 40, powyżej 40 lat), płeć, narodowość (Turcja, Afryka Północna, pozostała część Afryki, Surinam, Azja, Ameryka i Australia) oraz powód wydalenia (nielegalne przebywanie, pozostałe).

Godwin i Böhning (2017) ponownie przeanalizowali zbiór danych wykorzystany przez van der Heijdena i współpracowników (2003), ale zakładając, że zdarzenia są zależne, tj. cudzoziemcy raz złapani przez policję mogą nauczyć się, w jaki sposób unikać kolejnego spotkania, lub postanowili zalegalizować swój pobyt. W tym celu zaproponowali wykorzystanie dodatniego rozkładu Poissona z podwyższoną liczbą jedynek (pierwszych złapań) oraz uogólnionego modelu liniowego zakładającego ten rozkład dla badanej cechy. Zastosowanie tego podejścia znacząco obniżyło szacunki wielkości populacji w porównaniu z podejściem van der Heijdena i współpracowników (2003) – z 7080 do 3455. Natomiast wykorzystanie zmiennych pomocniczych zwiększyło estymowaną liczbę cudzoziemców nielegalnie przebywających na terenie wyżej wymienionych miast w 1995 r. do odpowiednio 6272 oraz 12690. Wydaje się, że w wypadku wykorzystania jednego źródła danych podejście zaproponowane przez Godwina i Böhninga (2017) jest właściwe.

W kontekście dwóch i większej liczby źródeł van der Heijden i współpracownicy (2012) przedstawili z kolei interesującą technikę estymacji osób urodzonych na Bliskim Wschodzie (Afganistan, Irak oraz Iran), ale przebywających w Holandii. W tym celu wykorzystali modele log-liniowe uwzględniające tzw. pasywne i aktywne zmienne pomocnicze. W procesie szacowania tak zdefiniowanej populacji posłużyli się dwoma rejestrami: rejestrem osób, którym wydano zezwolenie na pobyt w Holandii, oraz rejestrem policyjnym, zawierającym informacje o osobach podejrzanych o popełnienie przestępstwa.

Z kolei Gerritse i współpracownicy (2015) rozważali problem estymacji liczby Polaków oraz osób urodzonych na Bliskim Wschodzie, a przebywających w Holandii odpowiednio w 2011 i 2009 r. W tym celu sięgnęli po dane, podobnie jak van der Heijden i współpracownicy (2012), z dwóch rejestrów administracyjnych: rejestru osób zameldowanych w Holandii oraz rejestru policyjnego, skupiając się jednak na wrażliwości estymatora wielkości populacji opartego na modelach log-liniowych na złamanie założenia o niezależności tych dwóch źródeł. W przypadku osób urodzonych na Bliskim Wschodzie wpływ złamania założeń capture-recapture jest niewielki, podczas gdy dla obywateli Polski różnice w wielkości populacji są bardzo znaczące. W swojej rozprawie doktorskiej Gerritse (2016) analizowała problemy niespełnienia założeń metody capture-recapture (zależności źródeł oraz błędów w łączeniu rekordów) oraz wpływu imputacji danych na estymację liczby rezydentów według czasu przebywania. Oprócz rejestru ludności i policji autorka wykorzystała rejestr osób zatrudnionych.

Alternatywę dla danych jednostkowych pochodzących z wielu źródeł zaproponował Zhang (2008) w kontekście estymacji subpopulacji cudzoziemców. Na potrzeby estymacji wielkości populacji odnoszącej się do nielegalnie przebywających w Norwegii cudzoziemców[6] sięgnął do trzech źródeł danych: Centralnego Rejestru Osób (Central Personel Register), z którego wykorzystał informacje na temat liczby zameldowanych osób urodzonych poza Norwegią według kraju urodzenia i w wieku 18 lat i więcej, zbioru danych z Krajowego Urzędu Statystycznego w Norwegii na temat liczby obcokrajowców według kraju obywatelstwa oskarżonych o popełnienie przestępstwa oraz rejestru DUF (Datasystemet for Utlendings og Flyktningsaker), w którym znajdują się wszystkie osoby ubiegające się o zamieszkanie w Norwegii (jest to baza obejmująca imigrantów i uchodźców, którym przyznawany jest 12-cyfrowy numer w momencie ubiegania się przez nich o możliwość zamieszkania w Norwegii). Z tego źródła uzyskano informację o liczbie wniosków o wydalenie z Norwegii z uwzględnieniem osób wnioskujących o azyl. Na potrzeby estymacji wielkości populacji nielegalnie przebywających w Norwegii cudzoziemców zastosowano hierarchiczny model gamma Poissona, który należy do rodziny modeli mieszanych z efektami losowymi. Jako efekt losowy wykorzystano kraj pochodzenia cudzoziemców.

[6] Autor w swojej pracy używał zamiennie pojęć *unauthorized foreigners* oraz *irregular foreigners* w kontekście rezydentów, którzy przebywali na terenie Norwegii bez wymaganych dokumentów.

Na podstawie powyższych rozważań należy zauważyć pewną powtarzalność w kontekście doboru źródeł danych. Wszystkie estymacje opierały się na rejestrze osób (populacji *de iure*) oraz danych policji. Główną przesłanką takiego wyboru jest spełnienie założenia o niezależności źródeł danych. Dlatego aby poprawnie oszacować wielkość populacji, kluczowe jest dobranie odpowiednich zbiorów administracyjnych. Od spełnienia tego założenia zależy zasadność stosowania metody capture-recapture.

Przytoczone przykłady wskazywały na praktyczne wykorzystanie metody capture-recapture bazującej na modelach log-liniowych czy Poissona w estymacji liczby cudzoziemców w innych krajach. W przypadku Polski brak jest kompleksowych opracowań skupiających się na estymacji faktycznej liczby cudzoziemców. Częściowo może wynikać to z faktu, że dopiero w ostatnich latach tematyka cudzoziemców w Polsce (zwłaszcza osób pochodzących z Ukrainy) nabrała dużego znaczenia, szczególnie w kontekście rynku pracy. Warto jednak podkreślić, że pracownicy Narodowego Banku Polskiego dokonują prób estymacji na podstawie danych zagregowanych na potrzeby modelu NECMOD (ekonometrycznego modelu polskiej gospodarki) oraz szacunków przekazów pieniężnych. Także w mediach pojawiają się różne szacunki, jednak w żaden sposób nie są weryfikowalne. Rejestr PESEL zawiera bowiem wyłącznie osoby zameldowane na pobyt czasowy lub stały, Zakład Ubezpieczeń Społecznych (ZUS) dysponuje liczbą cudzoziemców zgłoszonych do ubezpieczenia, Ministerstwo Rodziny, Pracy i Polityki Społecznej – danymi na temat oświadczeń o chęci zatrudnienia cudzoziemców, Urząd ds. Cudzoziemców (UdSC) – danymi dotyczącymi ubiegania się o wizy czy karty pobytu, straż graniczna – statystykami m.in. ruchu granicznego czy liczby cudzoziemców nielegalnie przebywających na terenie Polski, a policja – Krajowym Systemem Informacji, który zawiera dane o popełnionych przestępstwach. Mimo to wydaje się, że polska statystyka publiczna, wspierana zasobami informacyjnymi pochodzącymi od innych organów, dysponuje wszelkimi zbiorami umożliwiającymi podjęcie rzetelnej próby szacunku liczby cudzoziemców. Do tej pory, zgodnie z aktualną wiedzą autorów, w Polsce nie podejmowano prób estymacji liczby cudzoziemców z wykorzystaniem wyżej rozważanych metod. Niniejszy artykuł oraz projekt badawczy, na podstawie którego powstał, wychodzą naprzeciw oczekiwaniom wielu odbiorców w tym zakresie.

MODELE LOG-LINIOWE W SZACOWANIU WIELKOŚCI POPULACJI TRUDNYCH DO ZBADANIA

Na potrzeby estymacji liczby cudzoziemców w Polsce z uwzględnieniem dodatkowych przekrojów zdecydowano się na zastosowanie metody capture-recapture bazującej na modelach log-liniowych. Wynikało to przede wszystkim z dostępności odpowiednich źródeł danych (które można wykorzystać w tego typu szacunkach), odpowiednich pakietów programu R (w których zaimplemen-

towane są funkcje na potrzeby estymacji parametrów modeli log-liniowych oraz kodów na procedurę bootstrap umożliwiającą znalezienie właściwych przedziałów ufności), a także z faktu, że w literaturze przedmiotu właśnie te modele są z powodzeniem wykorzystywane w estymacji liczebności populacji trudnych do zbadania. Przykładem mogą być wspomniane już prace Coumans i współpracowników (2017) oraz van der Heijdena i współpracowników (2012). W pierwszej z nich wykorzystano modele log-liniowe do oszacowania liczby bezdomnych w Holandii, w drugiej – modele log-liniowe oraz koncepcje pasywnych i aktywnych zmiennych pomocniczych do oszacowania liczby osób urodzonych na Bliskim Wschodzie, a przebywających w Holandii.

Modele log-liniowe stanowią obecnie bardzo ważną metodę analizy danych zawartych w tablicach kontyngencji. Rozwój metodologii opartej na tej technice analizy danych jakościowych zapoczątkowany został w latach 60. XX w. Goodman (1964, 1968, 1969) był jednym z pierwszych badaczy, którzy spopularyzowali modele log-liniowe w naukach społecznych. Modele te są szczególnie przydatne w sytuacji, gdy nie ma precyzyjnego rozróżnienia między zmienną objaśnianą a zmiennymi objaśniającymi, a zachodzi potrzeba wykrycia zależności w pewnym zbiorze danych.

Punktem wyjścia do zastosowania modeli log-liniowych w estymacji liczebności populacji trudnych do zbadania jest odpowiednio skonstruowana tablica kontyngencji[7], w której wykorzystuje się informacje z dwóch lub większej liczby źródeł danych. Tablica 1 odnosi się do przypadku, gdy dysponujemy dwoma niezależnymi źródłami danych (A i B).

TABL. 1. PRZYPADEK DWÓCH ŹRÓDEŁ – TABLICA KONTYNGENCJI 2 x 2

A	B		Suma
	Tak (1)	Nie (0)	
Tak (1)	n_{11}	n_{10}	n_{+1}
Nie (0)	n_{01}	n_{00}	n_{0+}
Suma	n_{+1}	n_{+0}	n

U w a g a. Tak – jednostka występuje w danym źródle, Nie – jednostka nie występuje w danym źródle.

Ź r ó d ł o: opracowanie własne.

W wypadku dwóch źródeł danych A i B może mieć miejsce sytuacja, w której po połączeniu jednostek[8] występują one tylko w źródle A, a nie występują w

[7] Na potrzeby szacunku liczby cudzoziemców rozpatrywane były złożone tablice wielowymiarowe. W artykule ograniczono się do tablic typu 2 x 2 oraz 2 x 2 x 2 celem przedstawienia idei modeli log-liniowych w tym zagadnieniu.

[8] W tym celu można zastosować łączenie deterministyczne z wykorzystaniem odpowiedniego identyfikatora lub probabilistyczne łączenie rekordów.

źródle B, występują w źródle B i nie występują w źródle A lub występują jednocześnie w źródle A i B. Przykładowo w tabl. 1 n_{01} oznacza liczbę jednostek, które nie występują w źródle A, a występują w źródle B. Kluczową kwestią jest zatem oszacowanie liczebności n_{00}, tj. liczby jednostek, które nie występują ani w źródle A, ani B. Ostatecznie oszacowaną liczebność populacji uzyskuje się bowiem poprzez dodanie wszystkich wartości z tabl. 1 po wcześniejszej estymacji liczebności n_{00}.

Oszacowanie liczebności n_{00} można uzyskać poprzez dopasowanie modelu log-liniowego do niekompletnej tablicy kontyngencji. Przykładowo dla tabl. 1 typu 2 x 2 odnoszących się do źródeł danych A i B pełny model log-liniowy [AB][9] może być przedstawiony w postaci modelu nasyconego (ang. *saturated model*):

$$\ln(m_{ij}) = \mu + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB} \quad i,j = \{'\text{Tak}', '\text{Nie}'\} \tag{1}$$

gdzie m_{ij} oznacza oczekiwaną liczebność w komórce i, j. Ponieważ jednak komórka $m_{00} = m_{(\text{Nie, Nie})}$ nie jest obserwowana, model [AB] ma jeden parametr za dużo i nie może być estymowany. W takiej sytuacji można rozważyć model niezależności [A][B] postaci:

$$\ln(m_{ij}) = \mu + \lambda_i^A + \lambda_j^B \tag{2}$$

w którym są tylko trzy parametry do oszacowania w związku z brakiem efektu interakcji λ_{ij}^{AB}. Trzy obserwowane komórki w tabl. 1 oraz trzy parametry do oszacowania w zasadzie czynią z tego modelu model nasycony. Po dopasowaniu go do danych możemy użyć oszacowanych parametrów do wyznaczenia liczebności brakującej komórki (Nie, Nie), a następnie wyznaczyć liczebność populacji poddanej analizie. Oszacowanie liczebności komórki n_{00} otrzymujemy ze wzoru:

$$\hat{n}_{00} = \exp(\mu) \tag{3}$$

Podobne rozumowanie można przeprowadzić w odniesieniu do tablic trójdzielczych typu 2 x 2 x 2, tj. w sytuacji gdy dysponujemy trzema źródłami danych.

Tablica 2 ilustruje przypadek wykorzystania trzech źródeł (A, B, C), np. trzech rejestrów administracyjnych lub dwóch rejestrów administracyjnych i badania reprezentacyjnego czy spisu. Podobnie jak w tablicy 2 x 2 x 2 istotne jest określenie przynależności do danego źródła, a celem – oszacowanie tego, czego nie możemy odczytać z tablicy, tj. n_{000}. Na potrzeby estymacji liczebności n_{000} można również wykorzystać koncepcję modeli log-liniowych.

[9] Notacja nawiasowa, często stosowana w przypadku modeli log-liniowych.

TABL. 2. PRZYPADEK TRZECH ŹRÓDEŁ – TABLICA KONTYNGENCJI 2 x 2 x 2

A	C				Suma
	Tak (1)		Nie (0)		
	B				
	Tak (1)	Nie (0)	Tak (1)	Nie (0)	
Tak (1)	n_{111}	n_{101}	n_{110}	n_{100}	n_{1++}
Nie (0)	n_{011}	n_{001}	n_{010}	n_{000}	n_{0++}
Suma	n_{+11}	n_{+01}	n_{+10}	n_{+00}	n

U w a g a. Jak przy tabl. 1.
Ź r ó d ł o: jak przy tabl. 1.

W tym celu budujemy model log-liniowy postaci (bez efektu głównego λ_{ijk}^{ABC}):

$$\ln(m_{ij}) = \mu + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} \tag{4}$$

który musimy ograniczyć przez: $\lambda_0^A = \lambda_0^B = \lambda_0^C = \lambda_{00}^{AB} = \lambda_{10}^{AB} = \lambda_{01}^{AB} = \lambda_{00}^{AC} = \lambda_{10}^{AC} = = \lambda_{01}^{AC} = \lambda_{00}^{BC} = \lambda_{10}^{BC} = \lambda_{01}^{BC} = 0$, aby móc oszacować parametry. Dodatkowym założeniem jest to, że nie występuje interakcja między A, B i C, tj. $\lambda_{ijk}^{ABC} = 0$. Model ten w notacji nawiasowej oznacza się jako [AB][BC][AC]. Oszacowanie brakującej liczby jednostek populacji otrzymujemy ze wzoru:

$$\hat{n}_{000} = \exp(\mu) \tag{5}$$

po uprzednim wyestymowaniu wszystkich parametrów.

W przypadku estymacji wielkości populacji możliwe jest wykorzystanie zmiennych pomocniczych, którymi mogą być przykładowo płeć czy grupy wieku. Ma to na celu obejście jednego z założeń metody capture-recapture (o stałej stopie pokrycia przez źródło w populacji) i uwzględnienie heterogeniczności przynależności poszczególnych jednostek do źródeł. Wykorzystanie zmiennych pomocniczych w kontekście modeli log-liniowych rozważają m.in. Coumans i współpracownicy (2017), Gerritse (2016), van der Heijden i współpracownicy (2012) czy Zwane i van der Heijden (2005). Wyróżniamy przy tym dwa podejścia, które determinowane są dostępnością zmiennych we wszystkich lub niektórych źródłach, lub tylko w jednym źródle. Pierwsze określa się w literaturze jako podejście z całkowicie obserwowalnymi zmiennymi (ang. *fully observed covariates*), a drugie – z częściowo obserwowalnymi zmiennymi (ang. *partially observed covariates*). W obydwu przypadkach można wykorzystać modele log-liniowe do oszacowania poszczególnych elementów populacji. Tego typu podejście zostało również zastosowane na potrzeby tego artykułu. Przykładowo w przypadku dwuwymiarowej tablicy kontyngencji 2 x 2 oprócz przynależności do dwóch źródeł A i B można rozpatrywać dodatkową cechę *X* (np. płeć), przez co należy rozszerzyć

tablicę do trójdzielczej (tabl. 3) oraz dopasować model log-liniowy [AX][BX] postaci:

$$\ln\left(m_{ijx}\right) = \mu + \lambda_i^A + \lambda_j^B + \lambda_x^X + \lambda_{ix}^{AX} + \lambda_{jx}^{BX} \tag{6}$$

gdzie λ_{ix}^{AX} oraz λ_{jx}^{BX} oznaczają efekty interakcji pomiędzy zmienną pomocniczą *X* i źródłami danych A oraz B.

TABL. 3. PRZYPADEK DWÓCH ŹRÓDEŁ I ZMIENNEJ POMOCNICZEJ

A	*X*				Suma
	X_1		X_2		
	B				
	Tak (1)	Nie (0)	Tak (1)	Nie (0)	
Tak (1)	n_{111}	n_{101}	n_{110}	n_{100}	n_{1++}
Nie (0)	n_{011}	n_{001}	n_{010}	n_{000}	n_{0++}
Suma	n_{+11}	n_{+01}	n_{+10}	n_{+00}	n

U w a g a. Jak przy tabl. 1.
Ź r ó d ł o: jak przy tabl. 1.

W przypadku dwóch źródeł A i B oraz jednej zmiennej pomocniczej *X*, przyjmującej przykładowo dwa warianty X_1 oraz X_2 (np. mężczyzna i kobieta), mamy do czynienia z trójdzielczą tablicą kontyngencji 2 x 2 x 2, w której brakujące liczebności podlegające estymacji to n_{001} oraz n_{000}. Dla sześciu komórek zatem znane są liczebności obserwowane w tabl. 3, w związku z czym model (6) zawiera sześć parametrów, które należy oszacować (nasycony model log-liniowy). Po dopasowaniu modelu do danych brakujące liczebności komórek ustala się ze wzorów: $\hat{n}_{000} = \exp(\mu)$ oraz $\hat{n}_{001} = \exp\left(\mu + \lambda_{X_1}^X\right)$. Powyższe postępowanie można rozszerzyć na większą liczbę zmiennych pomocniczych oraz źródeł. Zwiększa się przez to w oczywisty sposób złożoność analizowanych modeli log-liniowych, jednak wykorzystanie odpowiednich pakietów (np. *stats* i *parallel*) języka R (The R Development Core Team, 2018) znacznie skraca proces estymacji wszystkich możliwych do zbudowania modeli.

METODY OCENY JAKOŚCI MODELI LOG-LINIOWYCH

W analizie log-liniowej głównym celem jest wybór modelu o możliwie najprostszej postaci, który jednocześnie byłby najlepiej dopasowany do danych. W literaturze przedmiotu (Brzezińska, 2015; Goodman, 1964, 1968, 1969) proponuje się różnego rodzaju kryteria oceny modeli. Zostały one również wykorzystane na potrzeby niniejszego artykułu w procesie wyboru i oceny finalnego modelu. Do najważniejszych kryteriów zaliczamy iloraz wiarygodności, dewiancję, *AIC* (Akaike Information Criterion) oraz *BIC* (Bayerian Information Criterion).

Iloraz wiarygodności jest miarą pozwalającą ocenić dopasowanie modelu do danych. Przykładowo dla tablic 2 x 2 wyraża się on wzorem:

$$G^2 = 2\sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij} \ln\left(\frac{n_{ij}}{\widehat{m}_{ij}}\right) \tag{7}$$

gdzie $\widehat{m}_{ij} = \frac{n_{i+}n_{+j}}{n}$ stanowią oszacowania liczebności teoretycznych wyznaczonych dla danego modelu log-liniowego. W sytuacji gdy wartość ilorazu wiarygodności G^2 jest duża, model taki powinien być odrzucony jako ten, który w nieprawidłowy sposób odwzorowuje zależności między badanymi zmiennymi. Współczynnik G^2 może być także wykorzystywany do porównania oceny różnych modeli. Jeżeli porównuje się dwa modele, współczynnik G^2 może zostać przedstawiony w postaci (dla tablic 2 x 2):

$$G^2 = 2\sum_{i=1}^{2}\sum_{j=1}^{2} \widehat{m}_{ij}^{0} \ln\left(\frac{\widehat{m}_{ij}^{0}}{\widehat{m}_{ij}^{1}}\right) \tag{8}$$

gdzie 0 odnosi się do liczebności teoretycznych modelu ogólniejszego, tj. zawierającego wszystkie możliwe parametry, natomiast 1 dotyczy liczebności teoretycznych modelu zagnieżdżonego o uproszczonej postaci i zawierającego się w modelu 0.

Współczynnik ten może być również przedstawiony w postaci:

$$G^2(M_0|M_1) = G^2(M_0) - G^2(M_1) \tag{9}$$

Powyższa statystyka ma rozkład chi-kwadrat o liczbie stopni swobody $df = df(M_0) - df(M_1)$, gdzie M_0 jest modelem zagnieżdżonym, a M_1 modelem ogólnym z większą liczbą parametrów, i nazywana jest dewiancją. Dewiancja pozwala ocenić, czy parametr występujący w modelu M_1, a niewystępujący w modelu M_0 jest statystycznie istotny.

Statystyką służącą do porównywania ze sobą większej liczby modeli jest kryterium informacyjne Akaike oraz Schwarza (bayesowskie). Kryterium informacyjne Akaike wyraża się wzorem:

$$AIC = G^2 - df \tag{10}$$

gdzie G^2 to iloraz wiarygodności badanego modelu, a *df* to liczba odpowiadających mu stopni swobody. Z kolei bayesowskie kryterium informacyjne wyraża się wzorem:

$$BIC = G^2 - df \cdot \ln(n) \tag{11}$$

gdzie *n* to liczebność w tablicy kontyngencji. Preferowane są przy tym modele, dla których miary *AIC* i *BIC* przyjmują mniejsze wartości. W pracy wykorzystano kryterium *BIC* do określenia najlepszego modelu.

PRECYZJA OSZACOWAŃ LICZEBNOŚCI POPULACJI TRUDNEJ DO ZBADANIA

Kluczową kwestią w zagadnieniu estymacji liczebności populacji trudnej do zbadania jest jakość uzyskanych wyników. Oceną precyzji oszacowań uzyskanych za pomocą technik capture-recapture zajmowali się liczni badacze oraz wiele instytucji. Przykładowo Międzynarodowa Grupa Robocza ds. Monitorowania i Prognozowania Chorób (International Working Group for Disease Monitoring and Forecasting) prowadzi prace nad konstrukcją niesymetrycznych przedziałów ufności dla liczebności populacji trudnej do zbadania. Z kolei Chao (1989) podjął próbę konstrukcji symetrycznych przedziałów ufności polegającą na odpowiedniej transformacji oszacowanej liczebności populacji, głównie z wykorzystaniem transformacji logarytmicznej. Wreszcie ostatnio stosowane techniki w konstrukcji estymatorów wariancji liczebności populacji trudnej do zbadania bazują na metodzie bootstrap, zarówno nieparametrycznej, jak i parametrycznej (Buckland i Garthwaite, 1991; Gemmell, Millar i Hay, 2004).

W artykule na potrzeby oceny jakości oszacowań liczby cudzoziemców w odpowiednich przekrojach skonstruowano 95-procentowe przedziały ufności oraz względne błędy szacunku. W tym celu wykorzystano parametryczny bootstrap, który jest szeroko stosowany w badaniach poświęconych estymacji populacji trudnych do zbadania (Zwane i van der Heijden, 2003). Decyzja o konstrukcji odpowiednich przedziałów ufności oraz względnych błędów szacunku bazujących na parametrycznej metodzie bootstrap wynikała również z tego, że jest to stosunkowo łatwa w implementacji technika w kontekście tablic kontyngencji, które nie są w pełni obserwowalne (nieznajomość liczebności niektórych komórek). Ogólnie w celu utworzenia przedziałów ufności oraz wyznaczenia względnych błędów szacunku w pierwszej kolejności dokonuje się oszacowania liczebności populacji trudnej do zbadania z wykorzystaniem odpowiedniego modelu log-liniowego. Estymację parametrów modelu log-liniowego przeprowadza się na obserwowalnych komórkach tablicy kontyngencji. Na podstawie oszacowanych parametrów modelu oraz liczebności brakujących komórek można wyznaczyć prawdopodobieństwo teoretyczne przynależności dla wszystkich komórek w tablicy kontyngencji. Następnie losuje się próbę z rozkładu wielomianowego przy uwzględnieniu oszacowanych prawdopodobieństw, która na dalszym etapie jest korygowana, tak aby odpowiadała strukturze obserwowanych danych. Wówczas dopasowuje się odpowiedni model log-liniowy do kompletnej tablicy kontyngencji i uzyskuje pierwsze oszacowanie bootstrapowe liczebności populacji trudnej do zbadania. Procedurę tę przeprowadza się wielokrotnie, wyznaczając wariancję, a następnie przedział ufności dla liczebności populacji.

Ujmując zagadnienie bardziej formalnie, sposób wyznaczania względnych błędów szacunku oraz przedziałów ufności liczebności populacji trudnej do zbadania w parametrycznej metodzie bootstrap można zapisać w następujących krokach:

1. Oszacowanie parametrów odpowiedniego modelu log-liniowego dla zadanej tablicy kontyngencji i komórek, dla których istnieją wartości empiryczne.
2. Oszacowanie wielkości populacji we wszystkich założonych przekrojach z wykorzystaniem parametrów wyznaczonego modelu log-liniowego.
3. Wyznaczenie całkowitej liczebności populacji trudnej do zbadania $\widehat{N} = \widehat{N}_1 + + \cdots + \widehat{N}_P$, gdzie *P* to liczba komórek w tablicy kontyngencji, a $\widehat{N}_p$ to oszacowana wielkość populacji w komórce p, przy czym $p = 1, \ldots, P$.
4. Wyznaczenie wektora długości *P* złożonego z prawdopodobieństw $\widehat{\boldsymbol{\pi}}_P = = (\widehat{N}_1/\widehat{N}, \ldots, \widehat{N}_P/\widehat{N})^T$.
5. Generowanie z rozkładu wielomianowego wektora $\boldsymbol{n}^* = (N_1^*, \ldots, N_P^*)^T$ długości *P* odpowiadającego populacji o liczebności $\widehat{N}$ z prawdopodobieństwami $\widehat{\boldsymbol{\pi}}_P$. Jest to wektor złożony z pseudoliczebności populacji we wszystkich przyjętych *P* przekrojach.
6. Utworzenie tablicy kontyngencji na bazie uzyskanej pseudoliczebności, układem odpowiadającej tablicy wyjściowej. Oszacowane są parametry modelu log-liniowego dla tych samych komórek co w punkcie 1.
7. Znajdowanie oszacowania liczebności dla przekrojów nieobserwowanych w tablicy kontyngencji.
8. Powtórzenie *B* razy[10] kroków 5–7.
9. Oszacowanie liczebności populacji $\widehat{N}^b$, dla $b = 1, \ldots, B$ oraz liczebności we wszystkich przekrojach rozważanej tablicy kontyngencji.
10. Wyznaczenie na podstawie otrzymanych oszacowań wartości oczekiwanej, wariancji, względnego błędu szacunku oraz 95-procentowego przedziału ufności liczebności populacji trudnej do zbadania[11]:
    - wartość oczekiwana:

$$\widehat{N} = \frac{\sum_{b=1}^{B} \widehat{N}^b}{B} \tag{12}$$

    - wariancja empiryczna:

$$\hat{V}(\widehat{N}) = \frac{1}{B-1} \sum_{b=1}^{B} \left(\widehat{N}^b - \widehat{N}\right)^2 \tag{13}$$

[10] Na potrzeby artykułu przyjęto $B = 500$.

[11] W podobny sposób można wyznaczyć te miary dla liczebności populacji w odpowiednich przekrojach.

- względny błąd szacunku (precyzja):

$$REE\left(\widehat{N}\right) = \frac{\sqrt{\widehat{V}\left(\widehat{N}\right)}}{\widehat{N}} \tag{14}$$

- 95-procentowy przedział ufności[12]:

$$\left[\widehat{N}_{2,5\%}, \widehat{N}_{97,5\%}\right]$$

ŹRÓDŁA DANYCH

Wybór zbiorów

Postawione w artykule cele realizowane były z wykorzystaniem danych pochodzących z zasobów informacyjnych statystyki publicznej za lata 2015 i 2016 (PBSSP), w szczególności danych administracyjnych i statystycznych gromadzonych w ramach badań: Zasoby migracyjne w Polsce, Cudzoziemcy w Polsce. Legalizacja pobytu cudzoziemców na terytorium RP, Operat do Badań Społecznych, Charakterystyka demograficzno-społeczna i ekonomiczna gospodarstw domowych i rodzin oraz Badanie Aktywności Ekonomicznej Ludności.

Po dokonaniu analizy i oceny zasobów danych jako główne źródła administracyjne w modelach log-liniowych posłużyły:

- System Pobyt (UdSC) – zbiór rejestrów, ewidencji i wykazu w sprawach cudzoziemców w zakresie wydanych zezwoleń na pobyt;
- rejestr PESEL (Ministerstwo Cyfryzacji) – w przypadku cudzoziemców zameldowanych wyłącznie na pobyt stały;
- Centralny Rejestr Ubezpieczonych (ZUS) – w przypadku ubezpieczonych cudzoziemców oraz członków ich rodzin (udostępniony zbiór nie obejmował wszystkich ubezpieczonych)[13].

Przygotowanie zbiorów danych do badania

Zbiory wejściowe wykorzystane w projekcie badawczym poddano przetwarzaniu umożliwiającemu porównywanie, łączenie i analizę danych z różnych źródeł oraz oszacowanie wyników. Można wyróżnić kilka, wzajemnie się przenikających i uzupełniających, grup działań:

[12] Przedział ten wyznaczany jest metodą percentylową, np. 95-procentowy percentylowy przedział ufności ma dolną i górną granicę wyznaczoną przez 2,5 i 97,5 percentyla wartości bootstrapowych $\widehat{N}^b$.

[13] W projekcie, który jest podstawą tego artykułu, wykorzystano więcej źródeł, nieuwzględnionych tutaj z racji innego zastosowania oraz ograniczonego miejsca.

1. Dobór podmiotowy i przedmiotowy. W tej fazie prac – na podstawie wstępnej analizy zawartości zbiorów wejściowych oraz stosownie do przyjętego zakresu przedmiotowego badania i przesłanek metodologicznych – dokonano selekcji potencjalnie przydatnych zmiennych ze zbiorów. Z kolei stosownie do zakresu podmiotowego badania zastosowano dobór rekordów w taki sposób, aby dotyczyły one cudzoziemców (np. w przypadku zbiorów z badań obejmujących szersze kategorie ludności) w odpowiednich do przyjętych w badaniu momentów obserwacji (31 grudnia 2015 r. i 2016 r.) pod względem okresu przebywania w Polsce (w przypadku zbiorów rejestrowych odnotowujących fakty i daty dotyczące pobytu).
2. Wyliczanie cech pochodnych na podstawie przekształceń surowych danych. W ramach tej grupy działań wykonano szereg wyliczeń i przekształceń surowych danych, mających na celu przede wszystkim: (1) utworzenie (wyprowadzenie) cech potrzebnych do opisu badanej populacji, czyli np. wyliczanie okresu pobytu cudzoziemca na podstawie dat zarejestrowanych w dokumentach; (2) zapewnienie zgodności definicyjnej i zakresowej cech pochodzących z różnych źródeł, np. dostosowanie różnorakich konwencji zapisów kraju obywatelstwa do ujednoliconego słownika kodów krajów.
3. Redukcja nadmiarowych danych. Ta faza składała się z dwóch kroków: (1) deduplikacji w obrębie pojedynczych zbiorów danych, polegającej na wykrywaniu i usuwaniu ewidentnych dubli – powielonych rekordów danych, czyli takich, które pomimo różnych technicznych (bazodanowych) identyfikatorów rekordu zawierały dokładne powtórzenie wszystkich wartości; (2) niwelowania redundancji podmiotowej danych w kilku zbiorach jednego rejestru wartości/danych. Dokonano łączenia (parowania) poszczególnych zbiorów w ramach danego rejestru i wykrywania rekordów dotyczących tych samych podmiotów (osób), a następnie – na podstawie przesłanek merytorycznych i utworzonych na ich podstawie hierarchii adekwatności – wyboru najodpowiedniejszego rekordu reprezentującego danego cudzoziemca w rejestrze. W konsekwencji w odniesieniu do określonego rejestru powstawał – w zależności od potrzeb – jeden zbiór zawierający dane dotyczące unikalnych jednostek lub kilka zbiorów, ale podmiotowo rozłącznych.
4. Wyodrębnianie podstawowych jednostek/podmiotów badania. W wielu zbiorach rejestrowych obejmujących cudzoziemców podstawowe jednostki danych (rekordy) nie odnoszą się bezpośrednio do pojedynczych osób, lecz do różnego rodzaju faktów ich dotyczących. Stąd niezbędne były przekształcenia zbiorów wejściowych, w wyniku czego otrzymywano rekordy danych odnoszące się do osób. W szczególności były to działania oparte na: (1) grupowaniu (agregowaniu) rekordów danych, w ramach którego utworzono rekordy dla osób oraz wyprowadzano za pomocą operacji i funkcji agregujących przewidziane w badaniu cechy charakteryzujące cudzoziemców lub cechy pomocnicze; (2) restrukturyzacji (transpozycji) danych, w wyniku których pewne różno-

rodne wartości dotyczące jednej osoby zarejestrowane w kilku rekordach (różne warianty cechy) zapisywano w kolumnach jednego rekordu odnoszącego się do osoby.

5. Łączenie (parowanie) rekordów z różnych zbiorów. Operacje łączenia przeprowadzane były zarówno w obrębie zbiorów pochodzących z jednego rejestru – zazwyczaj na podstawie przygotowanego przez gestora sztucznego identyfikatora rekordów/osób – jak i kojarzenia zbiorów z różnych rejestrów czy badań, w tym wypadku na ogół za pomocą uniwersalnego identyfikatora (numer PESEL) lub na podstawie kombinacji wartości kilku cech[14].

WYNIKI BADANIA

Spełnianie założeń metody capture-recapture

W związku z wykorzystaniem w badaniu metody capture-recapture bazującej na wielu źródłach, w celu oszacowania liczby cudzoziemców poza dostępnymi statystycznymi źródłami danych, w pierwszej kolejności określono założenia metodologiczne. Kluczowe założenia, których spełnienie jest niezbędne z punktu widzenia przyjętych rozwiązań modelowych, oraz podjęte działania są następujące:

1. Definicje populacji we wszystkich rozważanych źródłach są takie same – określono populację cudzoziemców jako osoby w wieku 18 lat i więcej posiadające obywatelstwo inne niż polskie, które przebywały w Polsce pod koniec 2015 i 2016 r. Każde z wykorzystanych źródeł zostało ograniczone do tej populacji.
2. Populacja jest zamknięta – zakłada się, że w badanym okresie wielkość populacji jest stała. Ponadto należy podkreślić, że wszystkie rejestry były aktualne na ten sam dzień, tj. 31 grudnia 2016 r., dlatego podjęto następujące kroki przy wyodrębnianiu populacji:
   - populacja na 31 grudnia 2015 r.:
     – na podstawie rejestrów PESEL i ZUS oraz zbioru UdSC wybrano tylko osoby urodzone przed 31 grudnia 1997 r.,
     – na podstawie zbioru UdSC wybrano tylko te osoby, które miały decyzję umożliwiającą pobyt w Polsce wydaną między 1 stycznia a 31 grudnia 2015 r.;

[14] W łączeniu zbiorów wykorzystanych w niniejszym artykule zastosowano również parowanie według kluczy alternatywnych wobec numeru PESEL – głównie w odniesieniu do zbioru UdSC, w którym znaczna część rekordów nie miała numeru PESEL. Wykorzystano w nich, jako klucz podstawowy, zestawienia uwzględniające datę urodzenia, płeć i kraj obywatelstwa oraz – w zależności od rodzaju podejścia i dostępności zapisów w kolumnach – różne kombinacje spośród takich cech, jak: kod gminy, nazwa miejscowości i numer budynku. W tym wypadku liczba połączeń niejednoznacznych była stosunkowo niewielka i ostatecznie zrezygnowano z łączenia stochastycznego.

- populacja na 31 grudnia 2016 r.:
  - na podstawie rejestrów PESEL i ZUS oraz zbioru UdSC wybrano tylko osoby urodzone przed 31 grudnia 1998 r.,
  - na podstawie zbioru UdSC wybrano tylko te osoby, które miały decyzję umożliwiającą pobyt w Polsce wydaną między 1 stycznia a 31 grudnia 2016 r.

  W wypadku zbioru UdSC nie wyłączono z analizy cudzoziemców, którym data ważności wydanego dokumentu upłynęła w ciągu 2016 r., tj. przed 31 grudnia 2016 r., ponieważ mogli przebywać w Polsce nielegalnie.

3. Źródła danych są niezależne – w przypadku źródeł administracyjnych systemy powinny być niezależne (w sensie statystycznym), aby możliwe było zastosowanie metody capture-recapture wykorzystującej modele log-liniowe. Niezależność w kontekście źródeł administracyjnych oznacza, że prawdopodobieństwo znalezienia się jednostki w jednym źródle nie zależy od przynależności tej jednostki do drugiego źródła. Ostatecznie na potrzeby artykułu wykorzystano kombinacje trzech źródeł danych, które umożliwiają spełnienie tego założenia (tj. PESEL, UdSC i ZUS). Głównym uzasadnieniem wyboru tych źródeł danych było ich bieżące wykorzystywanie w statystyce publicznej na potrzeby innych badań (nie wymagało to pozyskania danych spoza PBSSP) oraz objęcie tej samej populacji.
4. Brak błędów nadreprezentacji i duplikatów – zakłada się, że źródła są pozbawione błędów nadreprezentacji, tj. zawierają wyłącznie jednostki z badanej populacji oraz zostały zdeduplikowane. Podstawowym źródłem był zintegrowany zbiór danych powstały w wyniku łączenia kilku rejestrów administracyjnych i zdeduplikowany. Zawierał on zmienną dotyczącą jakości danego rekordu, który jest przybliżeniem błędu nadreprezentacji. Z rekordów występujących w rejestrach PESEL i ZUS lub zbiorze UdSC wyodrębniono te, dla których określono kody jakości: 1 – oznaczający sytuację referencyjną (potwierdzenie istnienia osoby), 3 – wskazujący osoby w wieku 90 lat i więcej oraz 6 – oznaczający osobę zidentyfikowaną tylko w jednym rejestrze, który był wyznaczony przed dołączeniem zbioru UdSC. Dodatkowo przyjęto przy tym założenie, że cudzoziemcy będący w rejestrach przebywają na terenie Polski, niezależnie od tego, czy mają ustalone miejsce pobytu. Jest to kluczowe zwłaszcza w przypadku Systemu Pobyt, którego gestorem jest UdSC.
5. Każdą jednostkę będzie można zidentyfikować i połączyć między źródłami bez błędów – w tym celu zintegrowano dane za pomocą identyfikatora PESEL lub kombinacji zmiennych jednoznacznie wskazujących daną osobę (łączenie deterministyczne). Nie dokonywano łączenia probabilistycznego.
6. Prawdopodobieństwo włączenia do co najmniej jednego z rejestrów powinno być jednorodne – aby spełnić to założenie, w procesie estymacji wykorzystano modele zawierające następujące zmienne: kraj obywatelstwa, płeć, wiek

(2 grupy) i województwo (16 oraz nieustalone). Wybór zmiennych podyktowany był z jednej strony ich dostępnością, z drugiej zaś koniecznością spełnienia warunku, aby w odpowiednio utworzonych grupach prawdopodobieństwo włączenia cudzoziemca do danego źródła było jednakowe. Jest to jeden ze sposobów spełnienia założenia dotyczącego homogeniczności prawdopodobieństw, który rekomenduje się w literaturze poświęconej metodzie capture-recapture (van der Heijden i in., 2012).

Opis danych

W tabl. 4 przedstawiono liczbę cudzoziemców według występowania w trzech źródłach dla analizowanych lat.

TABL. 4. LICZBA CUDZOZIEMCÓW W WIEKU 18 LAT I WIĘCEJ (stan na 31 grudnia)

Wyszczególnienie				UdSC		Suma
				Nie	Tak	
2015						
PESEL	Nie	ZUS	Nie	x	30090	30090
			Tak	3821	5583	9404
	Tak		Nie	7042	7476	14518
			Tak	4620	9871	14491
Suma				15483	53020	68503
2016						
PESEL	Nie	ZUS	Nie	x	92106	92106
			Tak	3821	11224	15045
	Tak		Nie	7115	16549	23664
			Tak	4641	18951	23592
Suma				15577	138830	154407

U w a g a. Jak przy tabl. 1, x – nieznana liczba cudzoziemców poza rejestrami.
Ź r ó d ł o: opracowanie własne na podstawie rejestrów PESEL i ZUS oraz zbioru UdSC.

W przypadku kombinacji PESEL, UdSC i ZUS w 2015 r. wykorzystano informacje o ponad 68,5 tys., a w 2016 r. – blisko 154 tys. cudzoziemców.

W odniesieniu do 2015 r. jedynie 9871 cudzoziemców było zidentyfikowanych jednocześnie w rejestrach PESEL i ZUS i zbiorze UdSC, a w 2016 r. – 18951. Głównym celem jest zatem oszacowanie liczby cudzoziemców będących poza tymi rejestrami, tj. nieznanej wartości liczbowej na przecięciu pól: PESEL = Nie, UdSC = Nie i ZUS = Nie. Na potrzeby estymacji tej liczebności wykorzystano modele log-liniowe.

Dobór modelu

Zgodnie z literaturą poświęconą szacowaniu wielkości nieznanej populacji założono, że prawdopodobieństwo pokrycia przez określone źródła danych nie jest jednakowe. Dlatego na potrzeby procesu modelowania wykorzystano następujące zmienne:

- płeć – 1 = mężczyzna, 2 = kobieta;
- wiek – produkcyjny (18–59 lat dla kobiet, 18–64 lat dla mężczyzn), poprodukcyjny – 60 lat i więcej dla kobiet, 65 lat i więcej dla mężczyzn;
- kraj obywatelstwa – UE, Armenia, Mołdawia, Białoruś, Rosja, Ukraina, Wietnam, pozostałe;
- województwo: 16 województw kodowanych 1, …, 16, nieustalone (jeżeli nie zostało określone miejsce pobytu).

Na potrzeby wyboru końcowego modelu log-liniowego, który wykorzystano w procesie estymacji liczby cudzoziemców w Polsce w odpowiednich przekrojach, w pierwszej kolejności dokonano zakodowania zmiennych, zgodnie z symbolicznym zapisem (notacja nawiasowa) charakterystycznym dla modeli log-liniowych.

Procedurę modelowania przeprowadzono oddzielnie dla lat 2015 i 2016 oraz dla kombinacji źródeł. Oznacza to, że ostatecznie przeprowadzono dwie niezależne procedury szacunku wielkości populacji (tabl. 5).

TABL. 5. WYBRANE MIARY JAKOŚCI MODELI LOG-LINIOWYCH WEDŁUG ROKU

Modele	Dewiancja M_0	*df* M_0	G^2	*AIC*	*BIC*	Dewiancja	*df r*
			2015				
1	216672	2218	–24972	50003	50168	41580	2190
2	216672	2218	–8909	18349	19860	9453	1954
2s	216672	2218	–8909	18349	19860	9453	1954
			2016				
1	651403	2399	–50619	101295	101463	91543	2371
2	651403	2399	–12260	25051	26583	14827	2135
2s	651403	2399	–12260	25049	26575	14827	2136

U w a g a. 1 – model wyłącznie z efektami głównymi, 2 – model z efektami głównymi i interakcjami pierwszego rzędu, 2s – model 2 z zastosowaną procedurą krokową (s pochodzi od ang. *step*, które odnosi się do pojęcia regresji krokowej, ang. *stepwise selection*; *stepwise regression*), *df* – stopnie swobody, M_0 – model jedynie z wyrazem wolnym (pusty model), *df r* – różnica między liczbą stopni swobody modelu pustego a liczbą stopni swobody modelu w danym wierszu.

Ź r ó d ł o: jak przy tabl. 4.

Tablica 5 zawiera zestawienie wybranych miar jakości dla zastosowanych modeli log-liniowych według kombinacji źródeł oraz roku. W przypadku modelu opartego na kombinacji PESEL, UdSC i ZUS modele 2 i 2s okazały się iden-

tyczne w 2015 r. (co potwierdzają kryteria informacyjne), a w przypadku 2016 r. model 2s był nieznacznie lepszy od modelu 2, ponieważ zarówno kryteria informacyjne *AIC*, jak i *BIC* są niższe.

Estymacja punktowa i przedziałowa

Tablica 6 przedstawia finalny model wraz z oszacowaną wielkością populacji cudzoziemców w Polsce w latach 2015 i 2016 oraz 95-procentowym bootstrapowym przedziałem ufności. Model dla 2015 r. różni się od modelu dla 2016 r. dodatkowym elementem – interakcją między płcią a wiekiem. Wynik modelowania sugeruje, że prawdopodobieństwo pokrycia przez badane źródła danych jest stałe, co skutkuje stabilnością modelu w czasie.

Według szacunków liczba cudzoziemców w wieku 18 lat i więcej przebywających w Polsce na koniec 2015 r. wynosiła 507,7 tys. (95-procentowy przedział ufności – od 369,1 tys. do 724,4 tys.). Liczba ta – oprócz cudzoziemców zameldowanych na pobyt czasowy – obejmowała również cudzoziemców zameldowanych na pobyt stały (takich osób według rejestru PESEL było 39,1 tys.). Dla porównania, zgodnie z danymi ZUS, liczba ubezpieczonych cudzoziemców zgłoszonych do ubezpieczeń emerytalnych i rentowych wynosiła 184188 na koniec 2015 r.

TABL. 6. OSTATECZNY MODEL LOG-LINIOWY

Model	$\hat{N}$	Przedział ufności	Precyzja w %
2015			
[P][Z][U][V][S][A][C][PZ][PU][PV][PS][PA][PC] [ZU][ZV][ZA][ZC][UV][UC][VS][VA][VC][SA][AC] [UA][US][ZS][SC]	507693	(369135, 724407)	17,64
2016			
[P][Z][U][V][S][A][C][PZ][PU][PV][PS][PA][PC] [ZU][ZV][ZA][ZC][UV][UC][VS][VA][VC][SA][AC] [UA][US][SC]	743665	(600796, 943124)	11,70

U w a g a. P – PESEL, U – UdSC, Z – ZUS, S – płeć, A – wiek, SA – interakcja między płcią a wiekiem, C – kraj obywatelstwa, V – województwo. Notacja nawiasowa oznacza efekty główne oraz interakcje.

Ź r ó d ł o: jak przy tabl. 4.

Analogicznie oszacowano liczbę cudzoziemców w wieku 18 lat i więcej przebywających w Polsce pod koniec 2016 r. na 743,7 tys. (95-procentowy poziom ufności – 600,8–943,1 tys.). Liczba ta oprócz cudzoziemców zameldowanych na pobyt czasowy obejmowała również cudzoziemców zameldowanych na pobyt

stały (takich osób według rejestru PESEL było 41,4 tys.). W 2016 r. odnotowano wyraźny wzrost liczby cudzoziemców w stosunku do roku poprzedniego. Zwiększyła się liczba obywateli Ukrainy, Białorusi, Rosji, Wietnamu i innych krajów spoza UE, natomiast liczba obywateli UE nieznacznie spadła. Według statystyk ZUS liczba osób fizycznych mających inne obywatelstwo niż polskie zgłoszonych do ubezpieczeń społecznych i rentowych pod koniec 2016 r. wynosiła 293188, natomiast liczba pracowników cudzoziemców zgłoszonych do tego samego ubezpieczenia – 169350. Tablica 7 przedstawia szczegółowe zestawienie wyników w podziale na kraj obywatelstwa.

TABL. 7. SZACUNEK WIELKOŚCI POPULACJI CUDZOZIEMCÓW W POLSCE WEDŁUG KRAJU OBYWATELSTWA

Kraje		$\hat{N}$	Przedział ufności 95%		Precyzja w %
			dolna granica	górna granica	
Armenia	2015	3168	2263	4505	18,33
	2016	4773	3897	6032	11,35
Białoruś	2015	19868	14429	27951	17,38
	2016	25813	20832	32569	11,81
Mołdawia	2015	2693	1613	4227	25,59
	2016	7580	5355	10617	17,99
Rosja	2015	22611	16040	32237	18,62
	2016	25534	20685	32344	12,07
Ukraina	2015	283714	203946	415732	18,55
	2016	454974	361512	584696	12,27
Wietnam	2015	7408	5554	9942	15,45
	2016	11728	10008	14170	9,10
UE	2015	70901	53579	97126	15,63
	2016	59571	50914	71169	8,77
Pozostałe	2015	97329	70037	138339	17,86
	2016	153692	124170	196140	12,06

Źródło: jak przy tabl. 4.

Wśród cudzoziemców przebywających w Polsce zdecydowanie przeważają obywatele krajów trzecich (co oznacza każdą osobę, która nie jest obywatelem UE w rozumieniu art. 17 ust. 1 Traktatu o Unii Europejskiej, w tym bezpaństwowców). Polski rynek pracy jest atrakcyjny dla cudzoziemców zza wschodniej granicy ze względu na bliskość geograficzną, sieci migracyjne, które pozwalają zminimalizować koszty pobytu przynajmniej w pierwszych tygodniach, zdecydowanie wyższe zarobki niż w krajach rodzimych, a także z powodu liberalizacji zasad dostępu obywateli do polskiego rynku pracy. Uregulowania prawne wprowadzające uproszczoną procedurę zezwalają na podejmowanie pracy (oświadczenia pracodawców o powierzeniu pracy cudzoziemcowi) przez obywateli sześciu krajów trzecich: Armenii, Białorusi, Gruzji, Mołdawii, Rosji i Ukrainy. Spośród nich obywatele Ukrainy stanowią największą zbiorowość. Szacuje się, że w 2015 r. w Polsce przebywało 283,7 tys. (95-procentowy przedział ufności –

od 203,9 tys. do 415,7 tys.), a w 2016 r. – 455,0 tys. (95-procentowy przedział ufności – od 361,5 tys. do 584,7 tys.) obywateli tego kraju.

PODSUMOWANIE

Wybrana do oszacowania liczby cudzoziemców na krajowym rynku pracy metoda capture-recapture, bazująca na modelach log-liniowych, jak dotąd nie była stosowana w badaniach statystycznych w Polsce. Jedynymi doświadczeniami, z których można było skorzystać, są empiryczne badania zrealizowane przez badaczy holenderskich i norweskich.

Przedstawione w artykule wyniki estymacji wielkości populacji cudzoziemców, będące pochodną rezultatów otrzymanych w projekcie badawczym, mogą stanowić dobrą podstawę do wyprowadzenia (wtórnie) różnego rodzaju wskaźników dla wyodrębnionych jednostek terytorialnych, takich jak np. bilans migracyjny netto czy wskaźnik aktywności zawodowej cudzoziemców. Te ostatnie zaś mogą być wykorzystywane przez władze samorządowe, m.in. do monitorowania wielkości zatrudnienia, wysokości stopy bezrobocia czy popytu na pracę cudzoziemców o wysokich kwalifikacjach oraz do oceny wpływu powierzania pracy cudzoziemcom na wysokość płac pracowników rodzimych. Dodatkowo przedstawione w artykule szacunki mogą zostać użyte do monitorowania grup narażonych na wykluczenie zawodowe i społeczne poprzez zapobieganie substytucji rodzimych zasobów pracy przez cudzoziemców. Ponadto będą mogły stanowić podstawę do prowadzenia analiz statystycznych dotyczących sytuacji społeczno-gospodarczej poszczególnych regionów kraju oraz prognoz ich rozwoju.

Przedstawione wyniki mają innowacyjny charakter ze względu na zastosowaną metodę opracowania szacunku oraz wykorzystane źródła danych. Należy jednak zaznaczyć, że w trakcie badań napotkano poważne trudności. Wynikały one głównie z tego, że metoda szacunku opierała się na źródłach danych administracyjnych pozyskiwanych w ramach PBSSP do innych badań, a zatem o zakresie informacyjnym zdefiniowanym przez określoną jednostkę realizującą własne badanie. Jako przykład można wskazać rejestry ZUS pozyskiwane na potrzeby realizacji badań z zakresu rynku pracy, które nie zawierały informacji o wszystkich ubezpieczonych cudzoziemcach. Bardzo cenne zbiory dotyczące zezwoleń na pracę i oświadczeń pracodawców o zamiarze powierzenia pracy cudzoziemcowi nie uwzględniały cech identyfikacyjnych, w związku z czym nie można było ich połączyć deterministycznie z innymi zbiorami. Rejestr PESEL z kolei nie obejmował cudzoziemców przebywających czasowo i zameldowanych w gminach, którzy nie posiadali numeru PESEL. Co więcej, wszystkie wykorzystane rejestry były aktualne na 31 grudnia 2016 r., co mogło wpłynąć na wyniki uzyskane na 31 grudnia 2015 r. Tym samym wtórne wykorzystanie rejestrów i zawartych w nich zmiennych miało istotny wpływ zarówno na sam wybór źródeł danych oraz metodę, jak i w konsekwencji na konstrukcję wskaźników.

Niezbędne jest podjęcie prac nad rozpoznaniem również innych źródeł, które ze względu na swój charakter i zakres mogą być bardzo przydatne do weryfikacji charakterystyki cudzoziemców, np. rejestr policji dotyczący cudzoziemców podejrzanych o popełnienie przestępstw czy zbiory: Państwowej Inspekcji Pracy w zakresie kontroli legalności zatrudniania cudzoziemców, Komendy Głównej Straży Granicznej w zakresie legalności pobytów lub Ministerstwa Spraw Zagranicznych w zakresie wiz. W tym celu konieczne będzie nawiązanie bądź zintensyfikowanie współpracy z gestorami poszczególnych rejestrów i baz danych. Jednocześnie należy podkreślić, że w dalszych pracach planuje się wykorzystanie zarówno metod uwzględniających łączenie deterministyczne i probabilistyczne oraz analizy wrażliwości na złamanie założeń metody capture-recapture, jak i stosowanych modeli.

PODZIĘKOWANIA

Autorzy składają podziękowania wszystkim osobom, które przyczyniły się do powstania raportu[15] podsumowującego pracę badawczą, na podstawie której powstał niniejszy artykuł, w szczególności: kierownikowi projektu dyrektor Departamentu Badań Demograficznych GUS Dorocie Szałtys oraz członkom zespołu badawczego: Michałowi Adamskiemu, Mariuszowi Chmielewskiemu, Piotrowi Filipowi, Danielowi Godlewskiemu, Tomaszowi Józefowskiemu, Pawłowi Kaczorowskiemu, Zofii Kostrzewie, Jackowi Kowalewskiemu, Arlecie Olbrot-Brzezińskiej, Arturowi Owczarkowskiemu, Joannie Stańczak, Karinie Stelmach oraz Annie Wysockiej.

BIBLIOGRAFIA

Bakker, B. F. M., van der Heijden, P. G. M., Gerritse, S. C. (2017). Estimation of non-registered usual residents in the Netherlands. W: D. Böhning, P. G. M. van der Heijden, J. Bunge (red.), *Capture-Recapture Methods for the Social and Medical Sciences* (s. 259–273). Boca Raton, Florida: CRC Press.

Bouchard, M. (2007). A Capture-Recapture Model to Estimate the Size of Criminal Populations and the Risks of Detection in a Marijuana Cultivation Industry. *Journal of Quantitative Criminology*, *23*(3), 221–241. DOI: 10.1007/s10940-007-9027-1.

Bouchard, M. (2008). Towards a Realistic Method to Estimate Cannabis Production in Industrialized Countries. *Contemporary Drug Problems*, *35*(2–3), 291–320. DOI: 10.1177/009145090803500206.

Bouchard, M., Tremblay, P. (2005). Risks of Arrest across Drug Markets: A Capture-Recapture Analysis of ”Hidden” Dealer and User Populations. *Journal of Drug Issues*, *35*(4), 733–754. DOI: 10.1177/002204260503500404.

[15] Raport wraz z załącznikami dostępny jest na stronie http://stat.gov.pl/statystyka-regionalna/statystyka-dla-polityki-spojnosci/statystyka-dla-polityki-spojnosci-2016-2018/badania/rynek-pracy-ubostwo-i-wykluczenie-spoleczne/.

Brzezińska, J. (2015). *Analiza logarytmiczno-liniowa: teoria i zastosowania z wykorzystaniem programu R*. Warszawa: Wydawnictwo C.H. Beck.

Buckland, S. T., Garthwaite, P. H. (1991). Quantifying Precision of Mark-Recapture Estimates Using the Bootstrap and Related Methods. *Biometrics*, *47*(1), 255–268. DOI: 10.2307/2532510.

Böhning, D., van der Heijden, P. G. M., Bunge, J. (2017). *Capture-Recapture Methods for the Social and Medical Sciences*. Boca Raton, Florida: CRC Press.

Böhning, D., van der Heijden, P. G. M., Bunge, J. (2018). Basic concepts of capture-recapture. W: D. Böhning, P. G. M. van der Heijden, J. Bunge (red.), *Capture-Recapture Methods for the Social and Medical Sciences* (s. 237–257). Boca Raton, Florida: CRC Press.

Chao, A. (1989). Estimating Population Size for Sparse Data in Capture-Recapture Experiments. *Biometrics*, *45*(2), 427–438. DOI: 10.2307/2531487.

Chen, B., Shrivastava, A., Steorts, R. C. (2018). Unique entity estimation with application to the Syrian conflict. *The Annals of Applied Statistics*, *12*(2), 1039–1067. DOI: 10.1214/18-AOAS1163.

Coumans, A. M., Cruyff, M., van der Heijden, P. G. M., Wolf, J. R. L. M., Schmeets, H. (2017). Estimating Homelessness in the Netherlands Using a Capture-Recapture Approach. *Social Indicators Research*, *130*(1), 189–212. DOI: 10.1007/s11205-015-1171-7.

Di Cecco, D., Di Zio, M., Filipponi, D., Rocchetti, I. (2018). Population Size Estimation Using Multiple Incomplete Lists with Overcoverage. *Journal of Official Statistics*, *34*(2), 557–572. DOI: 10.2478/jos-2018-0026.

Di Consiglio, L., Tuoto, T. (2015). Coverage Evaluation on Probabilistically Linked Data. *Journal of Official Statistics*, *31*(3), 415–429. DOI: 10.1515/jos-2015-0025.

Gemmell, I., Millar, T., Hay, G. (2004). Capture-recapture estimates of problem drug use and the use of simulation based confidence intervals in a stratified analysis. *Journal of Epidemiology & Community Health*, *58*(9), 758–765. DOI: 10.1136/jech.2003.008755.

Gerritse, S. C. (2016). *An application of population size estimation to official statistics: sensitivity of model assumptions and the effect of implied coverage*. Pobrane z: https://dspace.library.uu.nl/handle/1874/337476.

Gerritse, S. C., van der Heijden, P. G. M., Bakker, B. F. M. (2015). Sensitivity of Population Size Estimation for Violating Parametric Assumptions in Log-linear Models. *Journal of Official Statistics*, *31*(3), 357–379. DOI: 10.1515/jos-2015-0022.

Godwin, R. T., Böhning, D. (2017). Estimation of the population size by using the one-inflated positive Poisson model. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *66*(2), 425–448. DOI: 10.1111/rssc.12192.

Gołata, E. (2012). Spis ludności i prawda. *Studia Demograficzne*, *161*(1), 23–55. DOI: 10.2478/v10274-012-0002-y.

Gołata, E. (2018). *Koniec ery tradycyjnych spisów ludności*. Poznań: Wydawnictwo Uniwersytetu Ekonomicznego w Poznaniu.

Goodman, L. A. (1964). Simple Methods for Analyzing Three-Factor Interaction in Contingency Tables. *Journal of the American Statistical Association*, *59*(306), 319–352. DOI: 10.1080/01621459.1964.10482163.

Goodman, L. A. (1968). The Analysis of Cross-Classified Data: Independence, Quasi-Independence, and Interactions in Contingency Tables with or without Missing Entries: R. A. Fisher memorial lecture. *Journal of the American Statistical Association*, *63*(324), 1091–1131. DOI: 10.1080/01621459.1968.10480916.

Goodman, L. A. (1969). On Partitioning χ^2 and Detecting Partial Association in Three-Way Contingency Tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *31*(3), 486–498. DOI: 10.1111/j.2517-6161.1969.tb00808.x.

Goudie, I. B. J., Goudie, M. (2007). Who captures the marks for the Petersen estimator? *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *170*(3), 825–839. DOI: 10.1111/j.1467-985X.2007.00479.x.

Griffin, R. A. (2014). Potential uses of administrative records for triple system modeling for estimation of census coverage error in 2020. *Journal of Official Statistics*, *30*(2), 177–189. DOI: 10.2478 /jos-2014-0012.

van der Heijden, P. G. M., Cruts, G., Cruyff, M. (2013). Methods for population size estimation of problem drug users using a single registration. *International Journal of Drug Policy*, *24*(6), 614–618. DOI: 10.1016/j.drugpo.2013.04.002.

van der Heijden, P. G. M., Cruyff, M., van Houwelingen, H. C. (2003). Estimating the Size of a Criminal Population from Police Records Using the Truncated Poisson Regression Model. *Statistica Neerlandica*, *57*(3), 289–304. DOI: 10.1111/1467-9574.00232.

van der Heijden, P. G. M., Whittaker, J., Cruyff, M., Bakker, B., van der Vliet, R. (2012). People born in the Middle East but residing in the Netherlands: invariant population size estimates and the role of active and passive covariates. *The Annals of Applied Statistics*, *6*(3), 831–852.

Hudson, C. G. (1998). Estimating Homeless Populations through Structural Equation Modeling. *The Journal of Sociology & Social Welfare*, *25*(2), 136–154.

Jędrzejczak, A., Kubacki, J. (2014). Problemy jakości danych statystycznych w przypadku badania cech rzadkich. *Wiadomości Statystyczne*, (6), 11–26.

Rossi, C., Mascioli, F. (2008). Capture-recapture methods to estimate prevalence indicators for evaluating drug policies. *Bulletin on Narcotics*, *15*(1–2), 5–25.

Schepers, W., Nicaise, I. (2017). Estimating the homeless population. Sampling strategies. (HIVA Working Paper Series No. 20). Pobrane z: https://www.belspo.be/belspo/brain-be/projects /FinalReports/MEHOBEL_Final%20report_Annex%20I.pdf.

The R Development Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Tourangeau, R., Edwards, B., Johnson, T. P., Wolter, K. M., Bates, N. (2014). *Hard-to-Survey Populations*. Cambridge: Cambridge University Press.

Wolter, K. M. (1986). Some Coverage Error Models for Census Data. *Journal of the American Statistical Association*, *81*(394), 337–346. DOI: 10.1080/01621459.1986.10478277.

Zhang, L.-C. (2008). *Developing methods for determining the number of unauthorized foreigners in Norway*. Oslo: Statistisk Sentralbyrå. Pobrane z: https://www.ssb.no/a/english/publikasjoner/pdf /doc_200811_en/doc_200811_en.pdf.

Zhang, L.-C. (2015). On Modelling Register Coverage Errors. *Journal of Official Statistics*, *31*(3), 381–396. DOI: 10.1515/jos-2015-0023.

Zhang, L.-C., Dunne, J. (2018). Trimmed dual system estimation. W: D. Böhning, P. G. M. van der Heijden, J. Bunge (red.), *Capture-recapture methods for the social and medical sciences* (s. 237–257). Boca Raton, Florida: CRC Press.

Zwane, E. N., van der Heijden, P. G. M. (2003). Implementing the parametric bootstrap in capture--recapture models with continuous covariates. *Statistics & Probability Letters*, *65*(2), 121–125. DOI: 10.1016/j.spl.2003.07.010.

Zwane, E., van der Heijden, P. G. M. (2005). Population estimation using the multiple system estimator in the presence of continuous covariates. *Statistical Modelling*, *5*(1), 39–52. DOI: 10.1191 /1471082X05st086oa.