Estymacja liczby cudzoziemców w Polsce w latach 2015 i 2016 z wykorzystaniem rejestrów administracyjnych i metody capture-recapture

Maciej Beręsewicz, Grzegorz Gudaszewski, Marcin Szymkowiak

Streszczenie. W pracy zaproponowano metody szacunku wielkości populacji cudzoziemców przebywających w Polsce na koniec 2015 i 2016 roku, ze szczególnym uwzględnieniem cudzoziemców pracujących na terytorium Polski. W tym celu wykorzystano administracyjne źródła danych oraz techniki bazujące na metodzie *capture-recapture*, ze szczególnym uwzględnieniem modeli log-liniowych. Szacuje się, że w 2015 i 2016 roku na terenie Polski mogło przebywać odpowiednio około 500 tys. (95% przedział ufności 369–724 tys.) oraz około 744 tys. (601–943 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 stanowiło obszar rozważań autorów w niniejszym artykule.

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

Estimation the number of foreigners in Poland in 2015 and 2016 using administrative registers and the capture-recapture method

Summmary. The article describes methods of estimating the size of the foreigner population in Poland at the turn of 2015/2016, with special emphasis on foreigners engaged in paid employment. To achieve this goal the authors used administrative data sources and techniques of the capture-recapture method as well as log-linear models. The number of foreigners staying in Poland in 2015 and 2016 was estimated to be at about 500,000 (95% CI: 369–724,000) and about 744,000 (601-943,000), respectively. The study is the first comprehensive attempt to estimate the number of foreigners in Poland, which is an example of research into hard-to-survey populations. However, one should be aware of restrictive assumptions of the capture-recapte methods which will be also discussed in depth in the paper.

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

JEL: C81, J61, J68, F22

WSTĘP

Coraz częściej, zarówno na szczeblu rządowym, samorządowym jak i lokalnym, podnoszona jest kwestia konieczności dysponowania rzeczywistą skalą liczby cudzoziemców przebywających w Polsce stale i czasowo oraz podejmujących pracę. Szczególnie istotna dla realizowania polityk ludnościowych, migracyjnych i gospodarczych jest informacja o cechach demograficzno-społecznych i ekonomicznych cudzoziemców. Kolejna ważna kwestia dotyczy skali imigracji nierejestrowanej, tj. pozostającej poza ewidencją. Obecnie nie ma w Polsce miarodajnego i bezpośredniego źródła danych, które dostarczałoby wiarygodnych informacji w tym zakresie. Zwrócić należy również uwagę na fakt, iż imigracja cudzoziemców nie jest zjawiskiem dotykającym obszaru całego kraju w równym stopniu i podlega zróżnicowaniu przestrzennemu, zwłaszcza w kontekście regionalnych rynków pracy.

Posiadanie informacji na temat liczby cudzoziemców, uwzględniając w tym nierejestrowanych imigrantów, stanowi 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* (zarejestrowanej i niezarejestrowanej). Po drugie, cudzoziemcy stanowią populację 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 w oparciu o metodę kuli śnieżnej i jej rozszerzenie – metodę RDS[1]), o tyle proces estymacji wielkości takiej zbiorowości jest z metodologicznego punktu widzenia poważnym wyzwaniem badawczym. W literaturze przedmiotu istnieją jednak odpowiednie metody statystyczne, które umożliwiają estymację wielkości populacji trudnych do zbadania bazujące na technikach capture-recapture (por. Böhning i in., 2017). Zaliczyć tutaj można rozwiązania, w których wykorzystuje się jedno (por. Van Der Heijden i in., 2003a; Godwin i Böhning, 2017) albo co najmniej dwa źródła danych (por. Van der Heijden i in., 2012; Zhang, 2008). Skuteczne wykorzystanie tych technik w praktyce w dużej mierze zależy od dostępności danych statystycznych i jest uwarunkowane koniecznością spełnienia odpowiednich założeń leżących u podstaw poszczególnych metod.

[1]ang. *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 jak i zwerbowaniu kolejnych osób, które biorą w nim udział. W metodzie RDS wykorzystuje się informacje na temat sieci powiązań osób należących do danej zbiorowości.

W artykule podjęto próbę oszacowania wielkości populacji cudzoziemców przebywających w Polsce w końcu 2015 i 2016 roku na terenie kraju i według kraju obywatelstwa[2]. Przyjęto przy tym następującą definicję cudzoziemca – osoba nieposiadająca obywatelstwa polskiego lub bezpaństwowiec (podstawa prawna – ustawa z dnia 12 grudnia 2013 r. o cudzoziemcach, Dz.U. z 2016 r. poz. 1990 z późn. zm.). W tym celu posłużono się odpowiednio zbudowanym modelem log-liniowym z szeregiem zmiennych pomocniczych. Przyjęto przy tym definicję cudzoziemca jako osoby nieposiadającej obywatelstwa polskiego lub bezpaństwowca[3].

W kolejnych częściach artykułu przedstawiono przegląd literatury w zakresie estymacji populacji trudnych do zbadania. Opisano przy tym ideę metody capture-recapture oraz jej zastosowań w kontekście szacowania liczby cudzoziemców. Następnie omówiono wykorzystane źródła danych, a także opisano wybrane aspekty modeli log-liniowych, które stanowiły podstawową metodę estymacji wykorzystaną w procesie szacowania liczby cudzoziemców w Polsce w latach 2015–2016. Należy zaznaczyć, że wybór okresu oraz źródeł danych był podyktowany ich aktualną dostępnością dla statystyki publicznej w ramach Program badań statystycznych statystyki publicznej (PBSSP) i niemożliwe było pozyskanie innych danych jednostkowych. Udostępnione dane były aktualne na 31.12.2016 r. Należy podkreślić, że nie oznacza to, że w tym dniu wszyscy badani cudzoziemcy przebywali na terenie Polski, podobnie jak w rejestrze PESEL, który nie gwarantuje, że na terenie Polski w danym dniu przebywa określona liczba obywateli polskich. W dalszej części artykułu przedstawiono także w jaki sposób podjęto próbę spełnienia założenia metody *capture-recapture*, wyniki estymacji z uwzględnieniem wybranych zmiennych demograficznych i porównanie do liczebności z rejestrów administracyjnych. Całość artykułu stanowi podsumowanie, w którym sformułowano dalsze kroki badawcze.

ESTYMACJA LICZEBNOŚCI POPULACJI TRUDNEJ DO ZBADANIA – PRZEGLĄD LITERATURY

Populacje trudne do zbadania

W literaturze przedmiotu populacje trudne do zbadania można rozumieć na wiele różnych sposobów (Tourangeau i in., 2014). W gruncie rzeczy, ze względu na fakt, że

[2]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 niniejszym artykule.

[3]Podstawa prawna – ustawa z dnia 12 grudnia 2013 r. o cudzoziemcach, Dz.U. z 2016 r. poz. 1990 z późn. zm.

w wielu badaniach częściowych mamy do czynienia z dużą frakcją odmów, terminem tym można byłoby określić każdą z badanych populacji. Populacja trudna do zbadania ma jednak inne znaczenie i odnosi się do zbiorowości, które przedstawiają szczególne wyzwania metodologiczne różnego rodzaju oraz sprawiają, że są trudniejsze do zbadania w porównaniu z innymi populacjami. Niektóre z trudności związane mogą być z tym, że są to populacje rzadkie, ukryte, z jednostkami których trudno nawiązać kontakt czy ciężko współpracować.

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

- z których jednostki trudno wylosować do próby (ang. *hard-to-sample*) – w przypadku populacji trudnych do zbadania bardzo rzadko zdarza się, aby istniał właściwy operat losowania, z którego jednostki można byłoby wylosować do próby wykorzystując odpowiedni schemat jej pobierania. Z tego względu, w odniesieniu do takich populacji, stosuje się nielosowe dobory jednostek do próby, wśród których szczególną rolę odgrywają wspomniana już metoda kuli śnieżnej czy metoda doboru sterowana przez respondenta. Można również zastosować inne techniki doboru jednostek, zwłaszcza w odniesieniu dla populacji rzadkich i trudno uchwytnych, takie jak losowanie odwrotne, lokacyjne czy schematy linia-przecięcie oraz śledzenia łączy (Jędrzejczak i Kubacki, 2014). Istnieją jednak nawet i w takim przypadku trudności z doborem jednostek do próby gdyż populacje takie mogą być mobilne bądź nieuchwytne. Przykładem tego typu populacji mogą być osoby bezdomne lub pracujący cudzoziemcy;

- których jednostki trudno jest zidentyfikować (ang. *hard-to-identify*) – w niektórych przypadkach, szczególnie w odniesieniu do stygmatyzowanych grup społecznych, członkowie populacji mogą nie chcieć udostępnić swoich cech, co wiązać się może z lękiem przed ujawnieniem nielegalnego lub wstydliwego statusu społecznego. W takim przypadku utrudniona jest identyfikacja jednostek należących do takich populacji. Przykład tego typu populacji stanowić mogą narkomani, alkoholicy czy różne mniejszości (na przykład osoby LGBT, wyznawcy określonych religii czy ideologii);

- których jednostki trudno znaleźć i nawiązać z nimi kontakt (ang. *hard-to-find-and-contact*) – trudność w nawiązaniu kontaktu związana jest przede wszystkim z mobilnością tego typu populacji. Przykładem tego typu populacji mogą być niezameldowani cudzoziemcy, członkowie koczowniczych kultur (Beduini z południowo-zachodniej Azji czy Tuaregowie z Afryki Północnej), mniejszości wędrowne (Romowie w Europie), osoby bezdomne;

- której jednostki trudno namówić do wzięcia udziału w badaniu (ang. *hard-to-persuade*) – niechęć do wzięcia udziału w badaniu związana może być z drażliwością

poruszanej tematyki bądź z brakiem czasu. Przykładem tego typu populacji mogą być aktywni zawodowo, pracujący w szarej czy czarnej strefie, cudzoziemcy;

- której jednostki można zachęcić do wzięcia udziału w badaniu, ale ciężko przeprowadzić wywiad (ang. *hard-to-interview*) – trudność w przeprowadzeniu wywiadu może wynikać z tego, że należy uzyskać na udział w badaniu danej jednostki zgody przełożonego, opiekuna prawnego czy rodzica. Trudność ta może być również konsekwencją występowania niepełnosprawności czy bariery językowej, jeśli osoba ankietowana nie mówi w języku, w którym przygotowany został odpowiedni kwestionariusz. Wreszcie może być ona pochodną tego, że badanie należy przeprowadzić w obszarze konfliktu zbrojnego. Przykład tego typu populacji stanowić mogą więźniowie, osoby niepełnosprawne psychicznie czy cudzoziemcy nie znający języka danego kraju.

Jak pokazują powyższe rozważania trudność w zbadaniu określonych populacji może być pochodną wielu czynników. Tak jest w przypadku populacji cudzoziemców w Polsce, której jednostki trudno wylosować do próby (brak pełnego operatu losowania oraz kompleksowych źródeł danych statystycznych, z których można czerpać wiedzę na temat cudzoziemców), z którą trudno nawiązać kontakt (mobilność cudzoziemców na rynku pracy oraz brak stałego miejsca zamieszkania) czy też przeszkodą może być bariera językowa. 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 stanowić mogą swego rodzaju remedium na problemy związane z określeniem rzeczywistych rozmiarów populacji trudnych do zbadania. Należą one do grupy technik określanych wspólnym terminem *capture-recapture*[4]. W dalszej części artykułu wskażemy na kilka praktycznych zastosowań technik statystycznych wchodzących w skład metod typu *capture-recapture* w szacowaniu liczebności takich populacji, również z uwzględnieniem populacji cudzoziemców.

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

Metody *capture-recapture*, które wykorzystane zostały na potrzeby oszacowania liczby cudzoziemców w Polsce, wywodzą się z nauk przyrodniczych. Pierwotnie użyto ich 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

[4]W artykule używać będziemy angielskiego terminu capture-recapture (CR) w związku z nie do końca jasnym tłumaczeniem tego podejścia na język polski. Bezpośrednie tłumaczenie mogłoby brzmieć jako 'metodę wielokrotnego połowu' przy czym to określenie nie oddaje istoty tego podejścia. W szczególności w odniesieniu do rejestrów administracyjnych, w których nie dokonujemy losowań czy 'połowów' jednostek.

wielokrotnego wyłapywania osobników danej populacji. W pierwszej próbie złowiona jest 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 jako badania *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 i druga zawierająca 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 nie ujętych 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, jakie zostały złowione. Natomiast jeśli wskaźnik ponownych złowień jest stosunkowo wysoki, można przypuszczać, że złowiona została większość zwierząt z danej populacji. Pomysł zastosowania techniki złożonej z dwóch prób można odnaleźć w pracach Pierre'a Simona Laplace'a z 1786 roku, który wykorzystał ją do szacowania liczby ludności Francji w 1802 roku, a nawet wcześniej, w pracach Johna Graunta, który zastosował tę technikę do szacowania skutków zarazy wśród ludności Anglii około roku 1600. W dziedzinie ekologii technika ta najwcześniej użyta została w badaniach Petersena i Dahla dotyczących populacji ryb odpowiednio w roku 1896 i 1907 oraz w przeprowadzonym przez Lincolna badaniu powrotów zaobrączkowanych ptaków wodnych z roku 1930. Modele oparte na dwóch próbach zostały rozszerzone na przypadki zawierające większą liczbę prób przez Schnabela w roku 1938. Stąd też metoda wielokrotnych złowień nazywana jest również spisem Schnabela. Bardziej zaawansowana teoria statystyczna i procedury wnioskowania pojawiły się po publikacji prac Darrocha, który opracował zagadnienie od strony matematycznej (Böhning i in., 2018, Rozdział 1).

Założenia stosowane w odniesieniu do populacji zwierzęcych klasyfikuje się generalnie jako modele zamknięte i otwarte. W przypadku zamkniętym zakłada się, że wielkość populacji, która jest przedmiotem badania, jest stała w czasie prowadzonego badania. Założenie to jest zwykle zachowane w przypadku danych zbieranych na przestrzeni sto-

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

sunkowo krótkiego czasu poza okresem godowym. W modelu otwartym, dopuszcza się przyrosty (narodziny lub imigracja) lub ubytki (śmierć lub emigracja) w populacji. Założenie otwartej populacji jest zwykle wykorzystywane 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. W dalszej części uwaga skupiona zostanie na modelach zamkniętych w odniesieniu do populacji ludzi.

Warto również zaznaczyć, że współcześnie pojęcie *capture-recapture* jest szerokie i odnosi się do szeregu metod mających na celu oszacowanie wielkości nieznanej populacji. Zwykle wykorzystuje się różnego rodzaju narzędzia statystyczne, na przykład modele log-liniowe, modele klas ukrytych czy uogólnione modele liniowe. W prezentowanym artykule, celem oszacowania liczby cudzoziemców wykorzystane zostały metody *capture-recapture* wykorzystujące 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 przypadek wyłącznie jednego albo dwóch lub więcej źródeł.

Kluczowym aspektem wszystkich metod *capture-recapture* są założenia, których niespełnienie skutkuje obciążonymi szacunkami wielkości populacji. W przypadku jednego źródła zakładamy, że: (1) jednostki możemy zidentyfikować, (2) jednostki obserwujemy wielokrotnie (na przykład dana osoba popełniła więcej niż jedno przestępstwo), (3) populacja jest stała w czasie, (4) zakładamy określony rozkład prawdopodobieństwa wielokrotnego wystąpienia w zbiorze danych (na przykład ucięty rozkład Poissona) oraz (5) niezależność kolejnych obserwacji (por. Van Der Heijden i in., 2003a,b). Założenie o niezależności zdarzeń jest bardzo restrykcyjne i w praktyce rzadko możliwe do spełnienia (por. Zhang, 2008). Dlatego Godwin i Böhning (2017) zaproponowali wykorzystanie dodatniego rozkładu Poissona z podwyższoną liczbą jedynek do opisu liczby wystąpień w jednym źródle łagodząc w ten sposób założenie o niezależności zdarzeń 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 po identyfikatorze) oraz (6) prawdopodobieństwa włączenia do co najmniej jednego z rejestrów powinny 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 aktualnie poddawane dyskusji w literaturze poświęconej statystyce publicznej (por. Zhang, 2015; Gerritse i in., 2015; Gerritse, 2016; Di Consiglio i Tuoto, 2015; Di Cecco i in., 2018; Griffin, 2014; Zhang i Dunne, 2018). W artykule z racji ograniczonego miejsca nie podjęto próby oceny wrażliwości na niespełnienie powyższych założeń. Planowane jest to w przyszłych pracach autorów.

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

Metodę *capture-recapture* zaadaptowano także do określenia wielkości populacji wyłącznie na podstawie rejestrów administracyjnych. W takim wypadku, metodę *capture-recapture* można 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 wykorzystanie metody *capture-recapture* do estymacji populacji Irlandii na podstawie rejestru aktywności osób (ang. *Person Activity Register*) będącego wynikiem łączenia 10 rejestrów administracyjnych według podejścia opartego na znakach życia (ang. *signs-of-life*) oraz ewidencji praw jazdy. Bakker i in. (2017) podjęli próbę estymacji liczby niezarejestrowanych rezydentów w Holandii wykorzystując trzy źródła danych: rejestr ludności, rejestr zatrudnionych oraz rejestr podejrzanych o przestępstwa prowadzony przez policję. Wykorzystano, celem oszacowania wielkości populacji niezarejestrowanych rezydentów, odpowiednio zbudowany model log-liniowy uwzględniając zmienne pomocnicze w postaci czasu pobytu, płci oraz wieku, wcześniej dokonując deterministycznego i probabilistycznego łączenia rekordów z trzech wspomnianych źródeł danych. Została również przeprowadzona analiza wrażliwości uzyskanych wyników na przypadek występowania błędów połączenia jak i poprawności procesu parowania jednostek.

W literaturze, można znaleźć również wiele innych przykładów wykorzystania omawianej metody do szacunku liczebności specyficznych subpopulacji, na przykład liczby bezdomnych (Hudson, 1998; Coumans i in., 2017; Schepers i Nicaise, 2017), narkomanów (Van der Heijden i in., 2013; Bouchard, 2007, 2008; Bouchard i Tremblay, 2005; Rossi i Mascioli, 2008), nietrzeźwych kierowców (Van Der Heijden i in., 2003b), ofiar konfliktów (Chen i in., 2018) czy liczby cudzoziemców, na której skupimy się w kolejnej części artykułu. Ciekawy przegląd zastosowań metod *capture-recapture* w estymacji liczebności populacji trudnych do zbadania można również znaleźć w pracy Godwin i Böhning (2017).

Metody capture-recapture w estymacji liczby cudzoziemców

Van Der Heijden i in. (2003a) rozważał wykorzystanie jednego źródła danych do estymacji liczby cudzoziemców nielegalnie przebywających w 1995 roku w Amsterdamie, Rotterdamie, Hadze oraz Utrechcie, którzy nie zostali skutecznie wydaleni z Holandii. Cudzoziemcy ci byli wielokrotnie obserwowani w zbiorach danych policji. Van Der Heijden i in. (2003a) do estymacji wielkości tak zdefiniowanej populacji 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, Północna Afryka, reszta Afryki, Surinam, Azja, Ameryka i Australia) oraz powód wydalenia (nielegalne przebywanie, pozostałe).

Godwin i Böhning (2017) ponownie rozważyli zbiór danych wykorzystany przez Van Der Heijden i in. (2003a) 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 autorzy zaproponowali wykorzystanie dodatniego rozkładu Poissona z podwyższoną liczbą jedynek (pierwszych złapań) oraz uogólnionego modelu liniowego zakładającego ten rozkład dla badanej cechy. Wykorzystanie tego podejścia znacząco obniżyło szacunki wielkości populacji (3 455) w porównaniu z podejściem Van Der Heijden i in. (2003a) (7 080). Natomiast wykorzystanie zmiennych pomocniczych zwiększyło estymowaną liczbę cudzoziemców nielegalnie przebywających na terenie wyżej wymienionych miast w 1995 roku do odpowiednio 6 272 oraz 12 690. Wydaje się, że w przypadku wykorzystania jednego źródła danych podejście zaproponowane przez Godwin i Böhning (2017) jest właściwe.

W kontekście dwóch i większej liczby źródeł Van der Heijden i in. (2012) przedstawili z kolei interesującą technikę estymacji osób urodzonych na Środkowym 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 użyte zostały dwa rejestry: rejestr osób, którym wydano zezwolenie na pobyt w Holandii oraz rejestr policyjny zawierający informacje o osobach, które podejrzane są o popełnienie przestępstw.

Z kolei Gerritse i in. (2015) rozważali problem estymacji liczby Polaków oraz osób urodzonych na Środkowym Wschodzie, a przebywających w Holandii odpowiednio w 2011 i 2009 roku. W tym celu autorzy wykorzystali, podobnie jak Van der Heijden i in. (2012), dwa rejestry administracyjne – rejestr osób zameldowanych w Holandii oraz rejestr policyjny – 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. W tym celu, oprócz rejestru ludności i policji, wykorzystała rejestr osób zatrudnionych.

Ciekawą alternatywę dla wykorzystania danych jednostkowych 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], Zhang wykorzystał trzy źródła danych. Pierwsze źródło stanowił Centralny Rejestr Osób (Central Personel Register), z którego wykorzystano informacje na temat liczby zameldowanych osób urodzonych poza Norwegią według kraju urodzenia i w wieku 18+. Drugim z wykorzystanych źródeł były dane na temat liczby obcokrajowców według kraju obywatelstwa, którzy zostali oskarżeni o popełnienie przestępstwa. Tego typu informacje dostarcza Krajowy Urząd Statystyczny w Norwegii. Ostatnim źródłem danych był rejestr DUF (nor. *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 wykorzystano informację o liczbie wniosków o wydalenie z Norwegii uwzględniając podział czy dane osoby wnioskowały o azyl. Na potrzeby estymacji wielkości populacji nielegalnie przebywających w Norwegii cudzoziemców wykorzystano hierarchiczny model gamma Poissona, który należy do rodziny modeli mieszanych z efektami losowymi. W charakterze efektu losowego wykorzystano kraj pochodzenia cudzoziemców.

Na podstawie powyższych rozważań należy zauważyć pewną powtarzalność w kontekście doboru źródeł danych. Podstawą wszystkich estymacji było wykorzystanie rejestru osób (populacji *de iure*) oraz danych pochodzących z policji. Główną przesłanką takiego wyboru jest spełnienie założenia o niezależności źródeł danych. Dlatego, aby poprawnie oszacować wielkość populacji, kluczowe jest dobranie odpowiednich zbiorów administracyjnych celem spełnienia tego kluczowego założenia, od którego zależy zasadność stosowania metod *capture-recapture*.

Powyżej przytoczone przykłady wskazywały na praktyczne wykorzystanie metod *capture-recapture* bazujących 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. Po części może wynikać to z faktu, że dopiero w ostatnich latach tematyka cudzoziemców w Polsce (zwłaszcza osób pochodzących z Ukrainy) nabrała dużego znaczenia, zwłaszcza w kontekście rynku pracy. Warto jednak podkreślić, że pewne próby estymacji dokonują pracownicy Narodowego Banku Polskiego na podstawie danych zagregowanych na potrzeby modelu NECMOD (ekonometrycznego modelu polskiej gospodarki) oraz szacunków przekazów pieniężnych.

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

Także w mediach pojawiają się różne szacunki, które w żaden sposób nie są weryfikowalne. Rejestr PESEL zawiera bowiem wyłącznie osoby, które są zameldowane na pobyt czasowy lub stały, Zakład Ubezpieczeń Społecznych dysponuje liczbą cudzoziemców zgłoszonych do ubezpieczenia, Ministerstwo Rodziny, Pracy i Polityki Społecznej z kolei dysponuje danymi o chęci zatrudnienia cudzoziemców, Urząd ds. Cudzoziemców posiada dane dotyczące ubiegania się o wizy czy karty pobytu, Straż Graniczna dostarcza statystyk dotyczących m.in. ruchu granicznego czy liczby cudzoziemców nielegalnie przebywających na terenie Polski, a Policja dysponuje Krajowym Systemem Informacji, który zawiera dane o popełnionych przestępstwach. Wydaje się jednak, 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 było jednak podejmowanych prób estymacji liczby cudzoziemców z wykorzystaniem wyżej rozważanych metod. Niniejszy artykuł oraz wspomniany na wstępie projekt badawczy wychodzą naprzeciw oczekiwaniom wielu odbiorców odnośnie informacji o liczbie cudzoziemców w Polsce.

MODELE LOG-LINIOWE W SZACOWANIU WIELKOŚCI POPULACJI TRUDNYCH DO ZBADANIA

Na potrzeby estymacji liczby cudzoziemców w Polsce z uwzględnieniem dodatkowych przekrojów zdecydowano się wykorzystać metodę *capture-recapture* bazującą 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 zaimplementowane 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. Przykład stanowić mogą wspomniane już prace Coumans i in. (2017) oraz Van der Heijden i in. (2012). W pierwszej z prac wykorzystano modele log-liniowe do oszacowania liczby bezdomnych osób w Holandii. W drugim z przytoczonych artykułów zastosowanie modeli log-liniowych oraz koncepcji pasywnych i aktywnych zmiennych pomocniczych umożliwiło oszacowanie 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 dedykowanej tej technice analizy danych jakościowych zapoczątkowany został w latach 60-tych XX wieku. Goodman (1964, 1968, 1969) był jednym z pierwszych badaczy, którzy spopularyzowali modele log-liniowe w naukach społecznych. Modele te są szczególnie przydatne w sytuacjach, gdy brak jest 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. W Tabeli 1 przedstawiono przypadek, gdy dysponujemy dwoma niezależnymi źródłami danych (powiedzmy A i B). Tabela taka powstaje poprzez połączenie informacji o populacji trudnej do zbadania z dwóch różnych źródeł.

Tablica 1: Przypadek dwóch źródeł - tablica kontyngencji 2×2

		Źródło B		
		Tak (1)	Nie (0)	$\sum$
Źródło A	Tak (1)	n_{11}	n_{10}	n_{1+}
	Nie (0)	n_{01}	n_{00}	n_{0+}
$\sum$		n_{+1}	n_{+0}	n

Źródło: opracowanie własne

W przypadku dwóch źródeł danych A i B może mieć miejsce sytuacja, w której po połączeniu jednostek[8] występują jednostki tylko w źródle A, a nie występują w źródle B, występują w źródle B i nie występują w źródle A oraz występują jednocześnie w źródle A i B. W powyższej tabeli Tak (1) oznacza, że dana jednostka występuje w odpowiednim źródle, a Nie (0), że jednostka w tym źródle nie występuje. Przykładowo, 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 zarówno w źródle A jak i B. Ostatecznie oszacowaną liczebność populacji uzyskuje się bowiem poprzez dodanie wszystkich wartości z Tabeli 1 po wcześniejszej estymacji liczebności n_{00}.

Oszacowanie liczebności n_{00} może być uzyskane poprzez dopasowanie modelu log-liniowego do niekompletnej tablicy kontyngencji. Przykładowo, dla Tabeli 1 wymiarów 2×2 odnoszących się do źródeł danych A i B pełen model log-liniowy [AB][9] może być przedstawiony w postaci (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

[7]Na potrzeby szacunku liczby cudzoziemców rozpatrywane były złożone tablice wielowymiarowe. Celem przedstawienia idei modeli log-liniowych w tym zagadnieniu, w artykule ograniczymy się do tablic typu 2×2 oraz $2 \times 2 \times 2$.

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

[9]Jest to tzw. notacja nawiasowa, która w przypadku modeli log-liniowych jest często stosowana.

zatem estymowany. W takiej sytuacji można rozważyć model niezależności [A][B] postaci:

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

który ma tylko trzy parametry do oszacowania w związku z brakiem efektu interakcji λ_{ij}^{AB}. Ponieważ mamy trzy obserwowane komórki w Tabeli 1 oraz trzy parametry do oszacowania mamy w zasadzie do czynienia z modelem nasyconym. Po dopasowaniu tego modelu 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} znajdujemy przy tym ze wzoru:

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

Podobne rozumowanie można przeprowadzić w odniesieniu do tablic trójdzielczych typu $2 \times 2 \times 2$, tj. w sytuacji, gdy dysponujemy trzema źródłami danych A,B i C.

Tablica 2: Przypadek trzech źródeł - tablica kontyngencji $2 \times 2 \times 2$

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

Źródło: opracowanie własne

Tabela 2 może przedstawiać sytuację trzech źródeł, na przykład trzech rejestrów administracyjnych, dwóch rejestrów administracyjnych i badania reprezentacyjnego czy spisu. Podobnie jak w przypadku tabeli $2\times2\times2$ istotne jest określenie przynależności do poszczególnego źródła (oznaczone jako Tak/Nie). Również i w tym przypadku chcemy oszacować to czego nie możemy odczytać z tabeli, tj. n_{000}. Na potrzeby estymacji liczebności n_{000} można również wykorzystać koncepcję modeli log-liniowych. 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 przez [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. Celem jest z jednej strony obejście jednego z założeń metody *capture-recapture* (o stałej stopie pokrycia przez źródło w populacji) i uwzględnienie faktu heterogeniczności przynależności poszczególnych jednostek do źródeł. Wykorzystanie zmiennych pomocniczych w kontekście modeli log-liniowych rozważa m.in. Gerritse (2016), Coumans i in. (2017), Van der Heijden i in. (2012) czy Zwane i van der Heijden (2005). Wyróżniamy przy tym dwa podejścia, które determinowane są dostępnością zmiennych we wszystkich, niektórych lub tylko w jednym źródle. Pierwsze określa się w literaturze jako podejście z pełni 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 tabeli kontyngencji 2×2 oprócz przynależności do dwóch źródeł A i B można rozpatrywać dodatkową cechę X (na przykład płeć) przez co należy rozszerzyć tabelę do trójdzielczej Tabeli 3 oraz dopasować model log-liniowy [AX][BX] postaci:

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

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

Tablica 3: Przypadek dwóch źródeł A i B oraz jednej zmiennej pomocniczej X

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

Źródło: opracowanie własne

W przypadku dwóch źródeł A i B oraz jednej zmiennej pomocniczej X, przyjmującej przykładowo dwa warianty X_1 oraz X_2 (na przykład mężczyzna i kobieta), mamy do czynienia z trójdzielczą tablicą kontyngencji $2 \times 2 \times 2$, w której brakujące liczebności

podlegające estymacji to n_{001} oraz n_{000}. Mamy zatem sześć komórek, dla których znane są obserwowane liczebności w Tabeli 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 znajdujemy ze wzorów: $\hat{n}_{000} = \exp(\mu)$ oraz $\hat{n}_{001} = \exp\left(\mu + \lambda_{X_1}^{X}\right)$. Powyższe rozumowanie w naturalny sposób można rozszerzyć na większą liczbę zmiennych pomocniczych oraz liczbę analizowanych źródeł. Zwiększa się przez to w oczywisty sposób złożoność analizowanych modeli log-liniowych, jednak wykorzystanie odpowiednich pakietów (np. `stats` i `parallel`) języka R (R 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 (Goodman, 1964, 1968, 1969; Brzezińska, 2015) proponuje się na potrzeby oceny modeli różnego rodzaju kryteria. Zostały one również wykorzystane na potrzeby artykułu w procesie wyboru i oceny finalnego modelu. Do najważniejszych kryteriów zaliczamy iloraz wiarogodności, dewiancję, AIC oraz BIC.

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

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

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

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

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 tzw. kryterium informacyjne Akaike oraz Schwarza (bayesowskie). Kryterium informacyjne Akaike wyraża się wzorem:

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

gdzie G^2 to iloraz wiarogodności 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 TRUDNYCH DO ZBADANIA

Kluczową kwestią w zagadnieniu estymacji liczebności populacji trudnej do zbadania jest jakość uzyskanych wyników. Prace nad oceną precyzji oszacowań uzyskanych w oparciu o techniki *capture-recapture* prowadzone były przez wielu badaczy oraz instytucji. Przykładowo, Międzynarodowa Grupa Robocza ds. Monitorowania i Prognozowania Chorób, podjęła prace nad konstrukcją niesymetrycznych przedziałów ufności dla liczebności populacji trudnych do zbadania (International Working Group for Disease Monitoring and Forecasting, 1995). Chao (1989) podjął z kolei 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 trudnych do zbadania bazują na metodzie bootstrap, zarówno nieparametrycznej jak i parametrycznej (Buckland i Garthwaite, 1991; Gemmell i in., 2004).

W artykule na potrzeby oceny jakości oszacowań liczby cudzoziemców w odpowiednich przekrojach dokonano konstrukcji 95% przedziałów ufności oraz względnych błędów 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 faktu, że jest to stosunkowo łatwa w implementacji technika w kontekście tablic kontyngencji, które nie są w pełni obserwowalne (nieznajomość liczebności niektórych komórek).

Ogólnie, celem 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 przy tym na obserwowalnych komórkach tablicy kontyngencji. Mając oszacowane parametry modelu oraz liczebności brakujących komórek wyznaczane są prawdopodobieństwa teoretyczne przynależności dla wszystkich komórek w tablicy kontyngencji. Następnie losowana jest próba z rozkładu wielomianowego przy uwzględnieniu oszacowanych prawdopodobieństw, która w dalszym etapie jest korygowana, tak aby odpowiadała strukturze obserwowanych danych. Wówczas dokonuje się dopasowania odpowiedniego modelu log-liniowego do kompletnej tablicy kontyngencji i uzyskuje pierwsze oszacowanie bootstrapowe liczebności populacji trudnej do zbadania. Procedurę tą 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 boostrap można zapisać w następujących krokach:

1. Dla zadanej tablicy kontyngencji i komórek, dla których istnieją wartości empiryczne, dokonaj oszacowania parametrów odpowiedniego modelu log-liniowego.

2. Wykorzystując parametry wyznaczonego modelu log-liniowego dokonaj oszacowania wielkości populacji we wszystkich założonych przekrojach.

3. Wyznacz całkowitą liczebność populacji trudnej do zbadania $\hat{N} = \hat{N}_1 + \ldots + \hat{N}_P$, gdzie P to liczba komórek w tablicy kontyngencji, a $\hat{N}_p$ to oszacowana wielkość populacji w komórce p, przy czym $p = 1, \ldots, P$.

4. Wyznacz wektor długości P złożony z prawdopodobieństw $\hat{\boldsymbol{\pi}}_P = (\hat{N}_1/\hat{N}, \ldots, \hat{N}_P/\hat{N})^T$.

5. Wygeneruj z rozkładu wielomianowego wektor $\boldsymbol{N}^* = (N_1^*, \ldots, N_P^*)^T$ długości P odpowiadający populacji o liczebności $\hat{N}$ z prawdopodobieństwami $\hat{\boldsymbol{\pi}}_P$. Jest to wektor złożony z pseudo-liczebności populacji we wszystkich założonych P przekrojach.

6. Na bazie uzyskanych pseudo-liczebności stwórz tablicę kontyngencji układem odpowiadającą wyjściowej tablicy. Oszacuj parametry modelu log-liniowego dla tych samych komórek co w punkcie 1.

7. Dokonaj oszacowania dla przekrojów nieobserwowanych w tablicy kontyngencji.

8. Powtórz kroki 5–7 B razy[10].

9. Oszacuj liczebność populacji $\hat{N}^b$, dla $b = 1, \ldots, B$ oraz liczebności we wszystkich przekrojach rozważanej tablicy kontyngencji.

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

10. Na podstawie otrzymanych oszacowań wyznacz wartość oczekiwaną, wariancję, względny błąd szacunku oraz 95% przedział ufności liczebności populacji trudnej do zbadania[11]:

    - wartość oczekiwana:
    $$\hat{N} = \frac{\sum_{b=1}^{B} \hat{N}^B}{B}, \tag{12}$$
    - wariancja empiryczna:
    $$\hat{V}\left(\hat{N}\right) = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{N}^B - \hat{N}\right)^2, \tag{13}$$
    - względny błąd szacunku (precyzja):
    $$REE\left(\hat{N}\right) = \frac{\sqrt{\hat{V}\left(\hat{N}\right)}}{\hat{N}}, \tag{14}$$
    - 95% przedział ufności[12]:
    $$\left[\hat{N}_{2,5\%}, \hat{N}_{97,5\%}\right]. \tag{15}$$

ŹRÓDŁA DANYCH

Wybór zbiorów

Postawione w artykule cele realizowane były z wykorzystaniem informacji pochodzących z zasobów informacyjnych statystyki publicznej za lata 2015 i 2016 (PBSSP), w szczególności z wykorzystaniem 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", „Badanie Aktywności Ekonomicznej Ludności".

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

- System „Pobyt" (Urząd do Spraw Cudzoziemców – UdSC) – zbiór rejestrów, ewidencji i wykazu w sprawach cudzoziemców w zakresie wydanych zezwoleń na pobyt,
- Rejestr PESEL (Ministerstwo Cyfryzacji) – w zakresie cudzoziemców zameldowanych wyłącznie na pobyt stały,

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

[12] Przedział ten wyznaczany jest metodą percentylową. Na przykład, 95% percentylowy przedział ufności ma dolną i górną granicę wyznaczoną przez 2,5 i 97,5 percentyl wartości bootstrapowych $\hat{N}^B$.

- Centralny Rejestr Ubezpieczonych (Zakład Ubezpieczeń Społecznych – ZUS) – w zakresie ubezpieczonych cudzoziemców oraz członków ich rodzin (udostępniony zbiór niestety nie pokrywał wszystkich ubezpieczonych).

W projekcie, który jest podstawą tego artykułu, wykorzystano więcej źródeł, które z racji innego zastosowania oraz ograniczonego miejsca nie zostały tutaj uwzględnione.

Przygotowanie zbiorów danych na potrzeby badania

Wejściowe zbiory wykorzystane w projekcie badawczym poddano przetwarzaniu umożliwiającemu porównywanie, łączenie i analizę danych z różnych źródeł oraz oszacowanie wyników. W tym obszarze prac można wyróżnić kilka, wzajemnie przenikających i uzupełniających się, grup działań.

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 dobory rekordów, tzn. tak, aby dotyczyły one cudzoziemców (np. w przypadku zbiorów z badań obejmujących szersze kategorie ludności) w odpowiednich do założonych w badaniu momentów obserwacji (31.12.2015 r. i 31.12.2016 r.) pod względem okresu przebywania w Polsce (w przypadku zbiorów rejestrowych odnotowujących fakty i daty dotyczące pobytu).

Wyliczanie cech pochodnych na podstawie przekształceń surowych danych – w ramach tej grupy działań wykonano szereg wyliczeń i przekształceń surowych danych, mających przede wszystkim na celu: (1) utworzenie (wyprowadzenie) cech potrzebnych do opisu badanej populacji, czyli tego typu operacje, jak 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.

Redukcja nadmiarowych danych – ta faza składała się z dwóch kroków: (1) *deduplikacji w obrębie pojedynczych zbiorów danych*, polegającej na wykrywaniu i usuwaniu zwykłych (ewidentnych) dubli – powielonych rekordów danych, czyli takich, które pomimo różnych technicznych (bazodanowych) identyfikatorów rekordu, zawierały dokładne powtórzenie wszystkich wartości; (2) *niwelowaniu redundancji podmiotowej danych w kilku zbiorach jednego rejestru wartości/danych*. Łączenie (parowanie) poszczególnych zbiorów w ramach danego rejestru i wykrywanie rekordów dotyczących tych samych podmiotów (osób), a następnie — w oparciu o przesłanki merytoryczne i utworzone na ich podstawie hierarchie adekwatności – dokonano wyboru najbardziej odpowiedniego 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.

Wyodrębnianie podstawowych jednostek/podmiotów badania – w wielu zbiorach rejestrowych dedykowanych cudzoziemcom podstawowe jednostki danych (rekordy) nie odnoszą się bezpośrednio do pojedynczych osób, lecz do różnego rodzaju faktów dotyczących osób. Stąd niezbędne były przekształcenia i transformacje zbiorów wejściowych, w wyniku których 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 różnorodne wartości dotyczące jednej osoby zarejestrowane w kilku rekordach (różne warianty cechy) zapisywano w kolumnach jednego rekordu odnoszącego się do osoby.

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[13].

WYNIKI BADANIA

Spełnianie założeń metody capture-recapture

W związku z wykorzystaniem w niniejszym artykule metody *capture-recapture* bazującej na wielu źródłach, do oszacowania liczby cudzoziemców poza dostępnymi statystycznymi źródłami danych, należy w pierwszej kolejności określić przyjęte założenia metodologiczne. Poniżej przedstawiono listę kluczowych założeń, których spełnienie jest niezbędne z punktu widzenia przyjętych rozwiązań modelowych. Wskazano również działania, które miały na celu ich spełnienie.

Definicje populacji we wszystkich rozważanych źródłach są takie same – określono populację cudzoziemców jako osoby posiadające obywatelstwo inne niż polskie w wieku 18+, które przebywały w Polsce w końcu 2015 i 2016 roku. Każde z wykorzystanych źródeł zostało ograniczone do tej populacji.

[13]W łączeniu zbiorów wykorzystanych w niniejszym artykule zastosowano również parowanie według kluczy alternatywnych wobec numeru PESEL – głównie w odniesieniu do łączenia zbioru UDSC, w którym znaczna część rekordów nie miała numeru PESEL. Wykorzystano w nich, jako klucza podstawowego, zestawienia uwzględniającego 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 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. 31.12.2016 r., dlatego podjęto następujące kroki przy wyodrębnianiu populacji na lata 2015 i 2016:

- populacja na dzień 31.12.2015 r.:
  - na podstawie PESEL, UdSC i ZUS wybrano tylko osoby urodzone przed 31 grudnia 1997 r.,
  - na podstawie UdSC wybrano tylko te osoby, które miały decyzję umożliwiającą pobyt w Polsce wydaną między 01.01.2015 r. oraz 31.12.2015 r.
- populacja na dzień 31.12.2016 r.:
  - na podstawie PESEL, UdSC i ZUS wybrano tylko osoby urodzone przed 31 grudnia 1998 r.,
  - na podstawie UdSC wybrano tylko tych, którzy mieli decyzję umożliwiającą pobyt w Polsce wydaną między 01.01.2016 r. oraz 31.12.2016 r.

W przypadku rejestru UdSC nie wyłączono z analizy cudzoziemców, którym upłynęła data ważności dokumentu wydanego przez Urząd do Spraw Cudzoziemców w ciągu 2016 r., tj. przed 31.12.2016 r. ponieważ mogły przebywać w Polsce nielegalnie.

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 pokrycie tej samej populacji.

Brak błędów nadreprezentacji i duplikatów – zakłada się, że źródła są pozbawione błędów nadreprezentacji, tj. źródła zawierają wyłącznie jednostki z badanej populacji oraz zostały zdeduplikowane. Podstawowym źródłem był zintegrowany zbiór danych powstały na podstawie łączenia kilku rejestrów administracyjnych i zdeduplikowany. Rejestr ten zawierał zmienną dotyczącą jakości danego rekordu, który jest przybliżeniem błędu nadreprezentacji. Dla rekordów występujących w PESEL, ZUS lub UDSC wyodrębniono te, dla których określono kod jakości 1 oznaczający *sytuacja referencyjna (istnienie osoby potwierdzone)*, 3 wskazujący *osoby w wieku 90+* oraz kod 6 oznaczający *osobę*

zidentyfikowaną tylko w jednym rejestrze, który był wyznaczony przed dołączeniem rejestru 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.

Każdą jednostkę będzie można zidentyfikować i połączyć między źródłami bez błędów – w tym celu zastosowano integrację danych za pomocą identyfikatora PESEL lub kombinacji zmiennych jednoznacznie wskazujących daną osobę (łączenie deterministyczne). Nie dokonywano łączenia probabilistycznego.

Prawdopodobieństwa włączenia do co najmniej jednego z rejestrów powinny być jednorodne – aby spełnić to założenie w procesie estymacji wykorzystano modele zawierające następujące zmienne: 1) kraj obywatelstwa, 2) płeć, 3) wiek (2 grupy) i 4) województwo (16 oraz nieustalone). Wybór zmiennych podyktowany był z jednej strony ich dostępnością, z drugiej zaś koniecznością spełnienia warunku, aby w odpowiednio utworzonych grupach prawdopodobieństwa włączenia cudzoziemca do danego źródła były jednakowe. Jest to jeden ze sposobów spełnienia założenia dotyczącego homogeniczności prawdopodobieństw, który rekomenduje się w literaturze poświęconej metodom capture-recapture (por. Van der Heijden i in., 2012, s. 2).

Opis danych

W Tablicy 4 przedstawiono liczbę cudzoziemców według występowania w trzech źródłach według badanych lat.

Wartość „Tak" oznacza, że cudzoziemiec został zidentyfikowany, a wartość „Nie" oznacza, że nie został zidentyfikowany w danym źródle. W przypadku kombinacji PESEL, UdSC, ZUS w 2015 roku wykorzystano informacje o ponad 68,5 tys., a w 2016 roku dla blisko 154 tys. cudzoziemców.

W odniesieniu do 2015 roku jedynie 9 871 cudzoziemców było zidentyfikowanych jednocześnie w PESEL, UdSC i ZUS. W 2016 roku 18 951 cudzoziemców występowało jednocześnie w PESEL, UdSC i ZUS. W tabelach pojawia się również wartość "–", która oznacza nieznaną liczbę cudzoziemców będących poza wymienionymi rejestrami.

Głównym celem niniejszego artykułu jest 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 wspomniane już 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:

Tablica 4: Zestawienie liczby cudzoziemców w wieku 18+ według PESEL, UdSC i ZUS zgodnie ze stanem na 31.12.2015 r. i 31.12.2016.r

Stan na 31.12.2015 r.						
				UdSC		
				Nie	Tak	$\sum$
PESEL	Nie	ZUS	Nie	–	30 090	30 090
			Tak	3 821	5 583	9 404
	Tak	ZUS	Nie	7 042	7 476	14 518
			Tak	4 620	9 871	14 491
$\sum$				15 483	53 020	68 503
Stan na 31.12.2016 r.						
				UdSC		
				Nie	Tak	$\sum$
PESEL	Nie	ZUS	Nie	–	92 106	92 106
			Tak	3 821	11 224	15 045
	Tak	ZUS	Nie	7 115	16 549	23 664
			Tak	4 641	18 951	23 592
$\sum$				15 577	138 830	154 407

Źródło: Opracowanie własne na podstawie podstawie danych PESEL, ZUS i UdSC.

- Płeć – 1 = Mężczyzna, 2 = Kobieta,
- Wiek – Produkcyjny (18–59 dla kobiet, 18–64 dla mężczyzn), Poprodukcyjny – 60+ dla kobiet, 65+ 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. Poniższy opis dotyczy wyników uzyskanych dla modelu wykorzystującego województwa. Przyjęto zatem następujące oznaczenia: PESEL = P, UdSC = U, ZUS = Z, Płeć = S (od angielskiego słowa sex), Wiek = A (od angielskiego słowa age), Kraj obywatelstwa = C (od angielskiego słowa citizenship), Województwo = V (od angielskiego słowa voivodeship).

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

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

Tablica 5: Wybrane miary jakości modeli log-liniowych według roku

Rok	Model	Dewiancja M0	df M0	G2	AIC	BIC	Dewiancja	df r
2015	1	216 672	2 218	-24 972	50 003	50 168	41 580	2 190
	2	216 672	2 218	-8 909	18 349	19 860	9 453	1 954
	2s	216 672	2 218	-8 909	18 349	19 860	9 453	1 954
2016	1	651 403	2 399	-50 619	101 295	101 463	9 1543	2 371
	2	651 403	2 399	-12 260	25 051	26 583	14 827	2 135
	2s	651 403	2 399	-12 260	25 049	26 575	14 827	2 136

Źródło: opracowanie własne na podstawie danych PESEL, ZUS i UdSC. Wyjaśnienia: 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 step, które odnosi się do pojęcia regresji krokowej, ang. *stepwise selection; stepwise regression*), df = stopnie swobody, M0 —model jedynie z wyrazem wolnym (inaczej model pusty), df r — różnica między liczbą stopni swobody modelu pustego, a modelu w danym wierszu.

Estymacja punktowa i przedziałowa

Tablica 6 przedstawia postać finalnego modelu wraz z oszacowaną wielkością populacji cudzoziemców w Polsce w latach 2015 i 2016 oraz 95% bootstrapowym przedziałem ufności. Model dla 2015 r. nieznacznie różni się od modelu dla 2016 r., ponieważ posiada jeden dodatkowy element — interakcję między płcią, a wiekiem (oznaczone w tabeli jako [SA]). Wynik modelowania sugeruje, że prawdopodobieństwa 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 w końcu 2015 r. wynosiła 507,7 tys. (95% 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 publikacjami ZUS, liczba ubezpieczonych

Tablica 6: Postać ostatecznego modelu log-liniowego (notacja nawiasowa) wraz z oszacowaną wielkością populacji cudzoziemców w Polsce, 95% przedziałem ufności i precyzją oszacowań (w %)

Rok	Model	$\hat{N}$	Przedział ufności	Precyzja
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*]	507 693	(369 135, 724 407)	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]	743 665	(600 796, 943 124)	11,70

Źródło: opracowanie własne na podstawie danych PESEL, ZUS i UdSC. Wyjaśnienia: Notacja nawiasowa oznacza efekty główne oraz interakcje. Litery oznaczają odpowiednio PESEL = P, UdSC = U, ZUS = Z, Płeć = S, Wiek = A, Kraj obywatelstwa = C, Województwo = V.

cudzoziemców zgłoszonych do ubezpieczeń emerytalnych i rentowych wynosiła 184 188 w końcu 2015 roku.

Analogiczne oszacowano liczbę cudzoziemców w wieku 18 lat i więcej przebywających w Polsce w końcu 2016 r. na 743,7 tys. (95% poziom ufności 600,8–943,1 tys). Liczba ta – oprócz cudzoziemców zameldowanych na pobyt czasowy – obejmowała również cudzoziemców zameldowanych na pobyt stały (takich osób wg rejestru PESEL było 41,4 tys.). W 2016 r. odnotowano wyraźny wzrost liczby cudzoziemców w stosunku do roku poprzedniego. Wzrosła liczba obywateli Ukrainy, Białorusi, Rosji, Wietnamu i innych krajów spoza UE — liczba obywateli UE nieznacznie zmniejszyła się. Według statystyk ZUS, liczba osób fizycznych o obywatelstwie innym niż polskie zgłoszonych do ubezpieczeń społecznych i rentowych w końcu 2016 roku wynosiła 293 188, natomiast liczba cudzoziemców pracowników zgłoszonych do tego samego ubezpieczenia wynosiła 169 350. Tablica 7 przedstawia szczegółowe zestawienie wyników w podziale na kraj obywatelstwa.

Wśród cudzoziemców przebywających w Polsce zdecydowanie przeważają obywatele krajów trzecich (co oznacza każdą osobę, która nie jest obywatelem Unii Europejskiej w rozumieniu art. 17 ust. 1 Traktatu, w tym bezpaństwowców). Wynika to z faktu, że polski rynek pracy jest atrakcyjny dla cudzoziemców zza naszej wschodniej granicy, z jednej strony z powodu bliskości geograficznej, sieci migracyjnych, które pozwalają zminimalizować koszty pobytu przynajmniej w pierwszych tygodniach, zdecydowanie wyższych zarobków niż w krajach rodzimych, z drugiej – liberalizacji zasad dostępu obywateli do

Tablica 7: Szacunek wielkości populacji cudzoziemców w Polsce według kraju obywatelstwa

Rok	Kraj	$\hat{N}$	95% Przedział ufności		Precyzja
Armenia	2015	3 168	2 263	4 505	18,33
	2016	4 773	3 897	6 032	11,35
Białoruś	2015	19 868	14 429	27 951	17,38
	2016	25 813	20 832	32 569	11,81
Mołdawia	2015	2 693	1 613	4 227	25,59
	2016	7 580	5 355	10 617	17,99
Rosja	2015	22 611	16 040	32 237	18,62
	2016	25 534	20 685	32 344	12,07
Ukraina	2015	283 714	203 946	415 732	18,55
	2016	454 974	361 512	584 696	12,27
Wietnam	2015	7 408	5 554	9 942	15,45
	2016	11 728	10 008	14 170	9,10
kraje EU	2015	70 901	53 579	97 126	15,63
	2016	59 571	50 914	71 169	8,77
pozostałe	2015	97 329	70 037	138 339	17,86
	2016	153 692	124 170	196 140	12,06

Źródło: Opracowanie własne na podstawie podstawie danych PESEL, ZUS i UdSC.

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 krajów trzecich to właśnie obywatele Ukrainy stanowią największą zbiorowość. Szacuje się, że w 2015 r. przebywało 283,7 tys. (95% przedział ufności: 203,9 tys. - 415,7 tys.), a w 2016 r. 455,0 tys. (95% przedział ufności: 361,5 tys. - 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 holenderskich i norweskich badaczy.

Przedstawione w artykule wyniki estymacji wielkości populacji cudzoziemców, będące pochodną rezultatów otrzymanych we wspomnianym 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 oceny wpływu powierzania pracy cudzoziemcom na wysokość płac pracowników rodzimych. Dodatkowo przedstawione w artykule szacunki mogą być wykorzystywane do monitorowania grup narażonych na wykluczenie zawodowe i społeczne poprzez zapobieganie substytucji rodzimych zasobów pracy przez cudzoziemców. Wreszcie dane te będą mogły stanowić podstawę do prowadzenia analiz statystycznych na temat 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. Najważniejszą ich przyczyną był fakt, że metoda szacunku oparta była na źródłach danych administracyjnych, pozyskiwanych w ramach PBSSP dla innych badań, zatem o zakresie informacyjnych zdefiniowanym przez określoną jednostkę/departament realizującą własne badanie. Jako przykład można wskazać rejestry ZUS pozyskiwane na potrzeby realizacji badań z zakresu rynku pracy, które nie zawierały wszystkich ubezpieczonych cudzoziemców. Bardzo cenne zbiory dotyczące zezwoleń na pracę i oświadczeń pracodawców o zamiarze powierzenia pracy cudzoziemcowi nie zawierały cech identyfikacyjnych, w związku z czym nie można było ich połączyć deterministycznie z innymi zbiorami. Zbiór 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 dzień 31.12.2016 r., co mogło wpłynąć na wyniki uzyskane na dzień 31.12.2015 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 również podjęcie prac nad rozpoznaniem innych źródeł, które ze względu na swój charakter i zakres mogą być niezwykle cenne do weryfikacji charakterystyki cudzoziemców, np. rejestr policji dotyczący cudzoziemców podejrzanych o popełnienie przestępstw, zbiory Państwowej Inspekcji Pracy w zakresie kontroli legalności zatrudniania cudzoziemców, Komendy Głównej Straży Granicznej w zakresie legalności pobytów lub zbiory 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 metod uwzględniających łączenie deterministyczne i probabilistyczne oraz analizy wrażliwości na złamania założeń metody capture-recapture, jak i stosowanych modeli.

PODZIĘKOWANIA

Niniejszy artykuł został przygotowany na podstawie raportu podsumowującego pracę badawczą pt.*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.

Autorzy artykułu składają podziękowania wszystkim osobom, które przyczyniły się do powstania raportu. W szczególności są to: kierownik projektu - dyrektor Departamentu Badań Demograficznych Dorota Szałtys oraz członkowie zespołu badawczego: Michał Adamski, Mariusz Chmielewski, Piotr Filip, Daniel Godlewski, Tomasz Józefowski, Paweł Kaczorowski, Zofia Kostrzewa, Jacek Kowalewski, Arleta Olbrot-Brzezińska, Artur Owczarkowski, Joanna Stańczak, Karina Stelmach oraz Anna Wysocka[14].

Maciej Beręsewicz, Marcin Szymkowiak – *Katedra Statystyki Uniwersytet Ekonomiczny w Poznaniu; Ośrodek Statystyki Małych Obszarów Urzędu Statystycznego w Poznaniu*

Grzegorz Gudaszewski – *Departament Badań Demograficznych, Główny Urząd Statystyczny*

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

Literatura

Bakker, B. F. M., van der Heijden, P. G. M., i Gerritse, S. G. (2017). Estimation of non-registered usual residents in the Netherlands. In Böhning, D., Bunge, J., i van der Heijden, P. G. M., editors, *Capture-recapture methods for the social and medical sciences*, chapter 18, pages 259–273. CRC Press.

Böhning, D., Bunge, J., i van der Heijden, P. G. M. (2018). Basic concepts of capture-recapture. In Böhning, D., Bunge, J., i van der Heijden, P. G. M., editors, *Capture-recapture methods for the social and medical sciences*, chapter 17, pages 237–257. 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*, 23:221–241.

Bouchard, M. (2008). Towards a realistic method to estimate cannabis production in industrialized countries. *Contemporary Drug Problems*, 35(2-3):291–320.

Bouchard, M. i 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.

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

Buckland, S. T. i Garthwaite, P. H. (1991). Quantifying precision of mark-recapture estimates using the bootstrap and related methods. *Biometrics*, pages 255–268.

Böhning, D., Bunge, J., i van der Heijden, P. G. M., editors (2017). *Capture-recapture methods for the social and medical sciences*. CRC Press, Boca Raton, Florida.

Chao, A. (1989). Estimating population size for sparse data in capture-recapture experiments. *Biometrics*, pages 427–438.

Chen, B., Shrivastava, A., Steorts, R. C., i in. (2018). Unique entity estimation with application to the syrian conflict. *The Annals of Applied Statistics*, 12(2):1039–1067.

Coumans, A., Cruyff, M., Van der Heijden, P. G., Wolf, J., i Schmeets, H. (2017). Estimating homelessness in the Netherlands using a capture-recapture approach. *Social Indicators Research*, 130(1):189–212.

Di Cecco, D., Di Zio, M., Filipponi, D., i Rocchetti, I. (2018). Population size estimation using multiple incomplete lists with overcoverage. *Journal of Official Statistics*, 34(2):557–572.

Di Consiglio, L. i Tuoto, T. (2015). Coverage evaluation on probabilistically linked data. *Journal of Official Statistics*, 31(3):415–429.

Gemmell, I., Millar, T., i 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.

Gerritse, S. C. (2016). *An application of population size estimation to official statistics: Sensitivity of model assumptions and the effect of implied coverage*. PhD thesis, Utrecht University.

Gerritse, S. C., van der Heijden, P. G., i Bakker, B. F. (2015). Sensitivity of population size estimation for violating parametric assumptions in log-linear models. *Journal of official statistics*, 31(3):357–379.

Godwin, R. T. i 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.

Gołata, E. (2012). Spis ludności i prawda. *Studia Demograficzne*, 161(1):23–55.

Goodman, L. A. (1964). Simple methods for analyzing three-factor interaction in contingency tables. *Journal of the American Statistical Association*, 59(306):319–352.

Goodman, L. A. (1968). The analysis of cross-classified data: Independence, quasi-independence, and interactions in contingency tables with or without missing entries: Ra fisher memorial lecture. *Journal of the American Statistical Association*, 63(324):1091–1131.

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)*, pages 486–498.

Goudie, I. B. i 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.

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.

Hudson, C. G. (1998). Estimating homeless populations through structural equation modeling. *J. Soc. & Soc. Welfare*, 25:136.

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

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

Rossi, C. i Mascioli, F. (2008). Capture-recapture methods to estimate prevalence indicators for evaluating drug policies. *Bulletin on Narcotics*, 60.

Schepers, W. i Nicaise, I. (2017). Estimating the homeless population. sampling strategies. *Working Paper*, 20:1–28.

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

Van Der Heijden, P. G., Bustami, R., Cruyff, M. J., Engbersen, G., i Van Houwelingen, H. C. (2003a). Point and interval estimation of the population size using the truncated Poisson regression model. *Statistical Modelling*, 3(4):305–322.

Van der Heijden, P. G., Cruts, G., i 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.

Van Der Heijden, P. G., Cruyff, M., i Van Houwelingen, H. C. (2003b). Estimating the size of a criminal population from police records using the truncated Poisson regression model. *Statistica Neerlandica*, 57(3):289–304.

Van der Heijden, P. G., Whittaker, J., Cruyff, M., Bakker, B., Van der Vliet, R., i in. (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.

Wolter, K. M. (1986). Some coverage error models for census data. *Journal of the American Statistical Association*, 81(394):337–346.

Zhang, L.-C. (2008). Developing methods for determining the number of unauthorized foreigners in Norway. *Statistisk Sentralbyrå/Utlendingsdirektoratet, Oslo Garcia, Jose Miguel Morales*.

Zhang, L.-C. (2015). On modelling register coverage errors. *Journal of Official Statistics*, 31(3):381–396.

Zhang, L.-C. i Dunne, J. (2018). Trimmed dual system estimation. In Böhning, D., Bunge, J., i Heijden, P. G. M. v. d., editors, *Capture-recapture methods for the social and medical sciences*, chapter 17, pages 237–257. CRC Press.

Zwane, E. i Van der Heijden, P. (2003). Implementing the parametric bootstrap in capture–recapture models with continuous covariates. *Statistics & probability letters*, 65(2):121–125.

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