Załącznik nr 2 Przegląd literatury w zakresie szacowania wielkości populacji trudnych do zbadania
Praca badawcza „Cudzoziemcy na krajowym rynku pracy w ujęciu regionalnym'

Spis treści

1. Idea szacowania wielkości populacji 2
   1.1. Szacowanie wielkości populacji w ujęciu historycznym 2
   1.2. Tablice będące podstawą estymacji wielkości populacji 3
   1.3. Zmienne pomocnicze w estymacji wielkości populacji 5
   1.4. Probabilistyczne łączenie rekordów 5
2. Klasyczne podejście w metodzie capture-recapture 8
   2.1. Kluczowe założenia 8
   2.1. Dwa źródła 8
   2.3. Trzy źródła 10
   2.4. Rozszerzenia klasycznych estymatorów capture-recapture 12
   2.4.1. Weryfikacja założeń dotyczących niezależności 12
   2.4.2. Probabilistyczne łączenie rekordów 13
   2.4.2.1. Uwzględnienie błędu łączenia w estymatorze 13
   2.4.2.2. Badanie wpływu błędu łączenia 14
   2.4.3 Błędy pokrycia – jednostki, które nie są elementami populacji 15
   2.4.3.1. Trimmed dual system estimation 15
   2.4.3.2. Badanie wpływu błędów pokrycia 15
   2.4.3.3. Szczególny przypadek trzech list 15
   2.4.4. Brak operatów dla badanej populacji – losowanie pośrednie 17
3. Modelowanie wielkości populacji 17
   3.1. Analiza klas ukrytych 17
   3.2. Wybrane modele regresji w estymacji wielkości populacji 20
   3.2.1. Model regresji Poissona i jego rozszerzenia 20
   3.2.2. Hierarchiczny model mieszany Poissona 21
4. Estymacja wielkości populacji trudnych do zbadania w wybranych krajach – przegląd doświadczeń 24
   4.1. Estymacja liczby niezarejestrowanych rezydentów w Holandii 25
   4.2. Estymacja liczby osób bezdomnych w Holandii metodą capture-recapture 26
   4.3. Estymacja wielkości populacji przestępców w Holandii z wykorzystaniem rejestrów policji 26
   4.4. Estymacja liczby nielegalnych cudzoziemców w Norwegii 27
   4.5. Estymacja wielkości populacji Irlandii na podstawie dwóch rejestrów 28
5. Oprogramowanie 28

Podsumowanie 29

Literatura 30

Spis tablic 32

1. Idea szacowania wielkości populacji

1.1. Szacowanie wielkości populacji w ujęciu historycznym

W typowym badaniu z dziedziny nauk biologicznych przeprowadzanym metodą wielokrotnych złowień (ang. *capture-recapture*) na badanym obszarze umieszcza się pułapki lub siatki w celu 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 składa się ono z dwóch prób: 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's z roku 1786, który wykorzystał ją do szacowania liczby ludności Francji w roku 1802, 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 artykułu Darrocha, który opracował zagadnienie od strony matematycznej.

Modele stosowane w odniesieniu do populacji zwierzęcych klasyfikuje się generalnie jako modele zamknięte i otwarte. W modelu 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 stosunkowo krótkiego czasu poza okresem godowym. W modelu otwartym, dopuszcza się przyrosty (narodziny lub imigracja) lub ubytki (śmierć lub emigracja) w populacji. Model otwarty

jest zwykle wykorzystywany 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.

W raporcie używać będziemy określenia capture-recapture (CR) w związku z nie do końca jasnym tłumaczeniem tego podejścia na język polski. Bezpośrednie tłumaczenie mogłoby brzmieć jako estymator 'wielokrotnego połowu' przy czym to określenie nie oddaje istoty tego podejścia. Oczywiście w ramach estymatorów tej klasy znajduje się estymator Petersena (lub Lincolna-Petersena), ale dotyczy on tylko jednego z przypadków, dlatego tylko w specyficznych przypadkach będzie używane to określenie zamiennie.

1.2. Tablice będące podstawą estymacji wielkości populacji

W przypadku estymatorów wykorzystywanych do szacowania wielkości populacji podstawą są dwa lub więcej niezależnych źródeł danych, które w pierwszej kolejności należy zintegrować (deterministycznie lub probabilistycznie), a następnie utworzyć odpowiednią tablicę kontyngencji. Tablica ta zawierać będzie informacje o liczebnościach w poszczególnych przekrojach. Poniżej przedstawione zostały przykłady tablic wykorzystywanych do estymacji wielkości populacji.

Najprostszy przypadek przedstawia Tablica 1, w której wykorzystywane są dwa źródła danych. Środek tablicy wypełniony jest wartościami teoretycznymi oznaczonymi przez m, a gdy dysponować będziemy wartościami obserwowanymi następuje zmiana oznaczeń na n.

Tablica 1. Przypadek dwóch źródeł – tablica kontyngencji 2x2

		Źródło B		Σ
		Tak (1)	Nie (0)	
Źródło A	Tak (1)	m_{11}	m_{10}	m_{1+}
	Nie (0)	m_{01}	m_{00}	m_{0+}
Σ		m_{+1}	m_{+0}	m_{++}

Źródło: opracowanie własne

Tablica 2 przedstawia sytuację trzech źródeł, na przykład trzech rejestrów administracyjnych, dwóch rejestrów administracyjnych i badania reprezentacyjnego czy spisu, rejestru i badania reprezentacyjnego. W obydwu przypadkach istotne jest określenie przynależności do poszczególnego źródła (oznaczone jako Tak/Nie). W obydwu przypadkach chcemy oszacować to czego nie możemy odczytać z tablic to jest odpowiednio m_{00} oraz m_{000}.

Inny układ danych, który wykorzystywany jest w przypadku modelowania z wykorzystaniem klas ukrytych (lub również w przypadku zastosowania modeli log-liniowych lub uogólnionych modeli liniowych) przedstawia Tablica 3. Kolumna Liczebność określa liczbę obserwacji spełniających określone kryteria, a przynależność do źródeł 1 do 4 określona jest przez kolumny od Z1 do Z4, gdzie 1 = TAK, 0 = NIE. W tym przypadku chcemy

dowiedzieć się ile jednostek jest poza obserwowanymi zbiorami czyli oszacować wartość teoretyczną określoną przez Z1=0, Z2=0, Z3=0 oraz Z4=0. Tablica 3 może również zawierać dane jednostkowe, tj. informacje o przynależności poszczególnych jednostek badania do określonego źródła.

Tablica 2. Przypadek trzech źródeł – tablica kontyngencji 2x2x2

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

Źródło: opracowanie własne

Tablica 3. Przypadek czterech źródeł – inny sposób zapisu

Liczebność	Z1	Z2	Z3	Z4
10	0	0	0	1
182	0	0	1	0
8	0	0	1	1
74	0	1	0	0
7	0	1	0	1
20	0	1	1	0
14	0	1	1	1
709	1	0	0	0
12	1	0	0	1
650	1	0	1	0
46	1	0	1	1
104	1	1	0	0
18	1	1	0	1
157	1	1	1	0
58	1	1	1	1

Źródło: opracowanie własne

1.3. Zmienne pomocnicze w estymacji wielkości populacji

W przypadku estymacji wielkości populacji możliwe jest wykorzystanie zmiennych pomocniczych, którymi mogą być przykładowo płeć, grupy wieku czy województwa. Celem jest z jednej strony obejście jednego z założeń metody capture-recapture (o stałej stopie pokrycia przez źródło w populacji – *enumerate rate*) i uwzględnienie faktu heterogeniczności przynależności poszczególnych jednostek do źródeł. Wykorzystanie zmiennych rozważa m.in. Gerritse (2016, Rozdział 1) czy Zwane i van der Heijden (2016).

Wyróżniamy tutaj dwa podejścia, które determinowane są dostępnością zmiennych we wszystkich, niektórych lub tylko w jednym źródle. Pierwsze podejście określa się w literaturze jako *fully observed covariates*, a drugie *partially observed covariates*. W obydwu przypadkach można wykorzystać modele log-liniowe do oszacowania poszczególnych elementów populacji.

Tablica 4. Przypadek dwóch źródeł i jednej zmiennej pomocniczej

						Σ
			Płeć (X)			
			Mężczyzna (1)		Kobieta (2)	
			Źródło B		Źródło B	
Źródło A		(1)	Tak	Nie (0)	Tak (1)	Nie (0)	
	Tak		m_{111}	m_{101}	m_{110}	m_{100}	m_{1++}
	Nie		m_{011}	m_{001}	m_{010}	m_{000}	m_{0++}
Σ			m_{+11}	m_{+01}	m_{+10}	m_{+00}	m_{+++}

Źródło: opracowanie własne

1.4. Probabilistyczne łączenie rekordów

Kluczowym aspektem estymacji wielkości populacji jest utworzenie powyżej wspomnianych tabel kontyngencji. Aby tego dokonać należy połączyć dwa lub więcej źródeł w jedno tak, aby otrzymać informację o przynależności poszczególnych jednostek do odpowiednich zbiorów. Możemy tego dokonać wykorzystując zarówno łączenie deterministyczne (po identyfikatorach), jak i probabilistyczne. Pierwszy przypadek jest najbardziej pożądany ponieważ większość metod typu capture-recapture zakłada łączenie bez błędów, jednakże istnieją również metody, które biorą tę sytuację pod uwagę. W tym miejscu skrótowo opisane zostanie, jak wygląda probabilistyczne łączenie rekordów.

Newcombe (1959) był jednym z pierwszych badaczy, który sformułował ideę probabilistycznego łączenia rekordów, a teoretyczne podstawy przedstawili Fellegi i Sunter (1969). Idea tej metody polega na znalezieniu w dwóch zbiorach danych (na przykład rejestrze i badaniu reprezentacyjnym) rekordów zawierających informacje o te samej jednostce. Decyzję polegającą na tym czy para rekordów należy do tej samej jednostki lub nie podejmuje się wykorzystując wyliczone prawdopodobieństwo opisujące stopień zgodności

przynależności pary rekordów do tej samej jednostki. Technika ta jest wykorzystywana w sytuacji gdy w obydwu parowanych zbiorach brakuje zmiennych kluczowych do konstrukcji odpowiedniego identyfikatora połączeniowego lub zmienne takie zawierają błędy. W metodzie tej dokonuje się wyboru kilku zmiennych wspólnych w obydwu zbiorach, określanych mianem zmiennych parujących, na podstawie których estymuje się prawdopodobieństwo tego, że dwa rekordy pochodzące z dwóch różnych zbiorów należą do tej samej jednostki. Więcej szczegółów na ten temat można znaleźć w pracy Winkler (1999) oraz w monografii Harrona (2015).

Algorytm integracji danych metodą probabilistycznego łączenia rekordów jest kilkustopniowy por. Roszka (2013). Obejmuje on w pierwszej kolejności proces ustalenia zmiennych parujących, harmonizację tych zmiennych oraz deduplikację. W kolejnym kroku przeprowadza się operację blokowania (grupowania), której głównym celem jest podział zbiorów podlegających procesowi integracji na podzbiory, w których znajdują się jednostki podobne do siebie ze względu na pewne kryteria (na przykład płeć, wykształcenie czy poziom agregacji przestrzennej). Jest to istotny element procedury probabilistycznego łączenia rekordów, który zmniejsza liczbę wszystkich możliwych połączeń, które należy rozpatrywać. Krok ten w istotny sposób przyczyni się do zoptymalizowania algorytmu poszukiwania odpowiednich połączeń. Następnie wyznacza się prawdopodobieństwa przynależności dwóch rekordów pochodzących z różnych zbiorów do tej samej jednostki i przeprowadza się proces ich łączenia. Ostatnim krokiem jest ocena uzyskanych połączeń.

Kluczowym problemem w metodzie probabilistycznego łączenia rekordów jest ustalenie czy dwa rekordy pochodzące z dwóch różnych źródeł należą do tej samej jednostki czy nie. W tym celu wyznacza się prawdopodobieństwo m zgodności wartości zmiennych parujących, przy przyjęciu założenia, że porównywana para rekordów należy do tej samej jednostki. Wyznacza się również tzw. prawdopodobieństwo niezgodności u wartości zmiennych parujących przy założeniu, że para jest niepołączeniem tzn. że jednostki nie należą do tej samej jednostki. Tak obliczone prawdopodobieństwa empiryczne wykorzystuje się do konstrukcji tzw. wag zgodności w_z i niezgodności w_{nz}, które ostatecznie służą do określenia stopnia przynależności porównywanych rekordów do tej samej jednostki Roszka (2013). Wagi te wyrażają się wzorami:

$$w_z = \frac{\ln\left(\frac{m}{u}\right)}{\ln 2}, \tag{1}$$

$$w_{nz} = \frac{\ln\left(\frac{1-m}{1-u}\right)}{\ln 2}. \tag{2}$$

Następnie wyznacza się tzw. wagę łączną. Dla porównywanej pary jest ona sumą wag zgodności i niezgodności dla zmiennych parujących. W przypadku gdy będzie ona dużą liczbą dodatnią, wszystkie lub większość zmiennych parujących zgadzają się co do wartości dla porównywanych par rekordów. Na odwrót, jeżeli będzie ona liczbą nieujemną to wszystkie lub większość zmiennych parujących nie zgadza się dla porównywanej pary rekordów. Za prawdopodobne połączenie uznaje się te pary rekordów, dla których wartość wagi połączeniowej jest największa - Roszka (2013). W praktyce wyznacza się najczęściej tzw. wagę progową

dla wagi łącznej, powyżej której pary uznawane są jako prawdopodobne połączenie a poniżej jako prawdopodobne niepołączenie.

W metodzie probabilistycznego integrowania danych niezwykle ważnym etapem jest ocena jakości uzyskanych połączeń. Można tutaj bowiem popełnić błędy dwojakiego rodzaju. Błąd I rodzaju polega na zaklasyfikowaniu jako niepołączenie pary rekordów, która w rzeczywistości należy do tej samej jednostki. Z kolei błąd II rodzaju polega na uznaniu dwóch rekordów jako należących do tej samej jednostki, podczas gdy w rzeczywistości nie odnoszą się one do tego samego obiektu.

Na potrzeby oceny jakości procesu łączenia rekordów pochodzących z dwóch różnych źródeł (na przykład rejestru i badania reprezentacyjnego) konstruuje się szereg wskaźników, do których można zaliczyć Roszka (2013):

• wskaźnik fałszywych połączeń:

$$fmr = \frac{n_{fp}}{n_m + n_{fp}}, \tag{3}$$

• wskaźnik wartości predykcyjnej:

$$ppv = \frac{n_m}{n_m + n_{fp}} = 1 - fmr, \tag{4}$$

• wskaźnik fałszywych niepołączeń:

$$fnmr = \frac{n_{fn}}{N_m}, \tag{5}$$

• wskaźnik czułości:

$$cz = \frac{n_m}{N_m} = 1 - fnmr, \tag{6}$$

• wskaźnik swoistości:

$$sw = \frac{n_u}{N_u}, \tag{7}$$

gdzie:

- n_m – liczba par rekordów połączonych prawidłowo (prawdziwie pozytywne),
- n_{fp} – liczba par rekordów połączonych nieprawidłowo (fałszywie pozytywne),
- n_u – liczba par rekordów niepołączonych prawidłowo (prawdziwie negatywne),
- n_{fn} –liczba par rekordów niepołączonych nieprawidłowo (fałszywie negatywne),
- N_m – ogólna liczba par rekordów odnoszących się do tych samych jednostek,

- N_u – ogólna liczba par rekordów nie odnoszących się do tej samej jednostki.

Wyznaczone powyżej wskaźniki oceny jakości połączeń różnić się będą w zależności od przyjętych wartości progowych. W praktyce poszukuje się zatem kompromisu, w którym utrzymuje się stosunkowo wysoką swoistość kosztem obniżenia czułości.

2. Klasyczne podejście w metodzie capture-recapture

2.1. Kluczowe założenia

Aby klasyczne estymatory wykorzystywane w szacowaniu wielkości populacji były nieobciążone spełnione muszą zostać następujące założenia:

- populacja jest zamknięta – dla każdej próby liczebność populacji N jest taka sama,
- prawdopodobieństwo wylosowania każdej jednostki jest takie same – może być to problematyczne w przypadku gdy korzystamy ze źródeł administracyjnych ponieważ zgodnie z regulacjami prawnymi niektóre jednostki mogą nie być uwzględnione w określonym źródle,
- próby/źródła są niezależne – pierwsze losowanie nie ma wpływu na drugie; w przypadku źródeł administracyjnych systemy powinny być niezależne (na przykład w sensie prawnym),
- jednostki możemy zidentyfikować i połączyć bez błędów (łączenie deterministyczne).

Każde złamanie założeń skutkować będzie zwiększeniem obciążenia estymatorów i w konsekwencji przedstawieniem nieprawdziwych informacji (nawet w formie przedziałów ufności). Szczegółowy opis założeń można znaleźć w pracy Wolter (1986).

2.1. Dwa źródła

W podstawowym podejściu zakładamy, że pobraliśmy próbę s_1 o wielkości n_1, w której wszystkie jednostki mogą zostać zidentyfikowane. Następnie, niezależnie pobierana jest druga próba s_2 o wielkości n_2, w której możemy zidentyfikować jednostki należące do dwóch prób $n_{1,2}$. Następnie, aby oszacować wielkość populacji, wykorzystujemy estymator Petersena (lub Lincolna-Petersena) dany wzorem (8) (por. Seber, 1982):

$$\hat{N}_{pet} = \frac{n_1 n_2}{n_{1,2}}, \tag{8}$$

którego wariancja jest równa:

$$\hat{V}(\hat{N}_{pet}) \approx \frac{n_1^2 n_2 (n_2 - n_{1,2})}{n_{1,2}^3}. \tag{9}$$

Jednakże, w przypadku populacji, w których niewielka liczba jednostek wspólnych $n_{1,2}$ znajdzie się w drugiej próbie, estymator $\hat{N}_{pet}$ może dążyć do nieskończoności. W związku z tym Chapman (1951) zaproponował estymator, który bierze pod uwagę ten problem:

$$\tilde{N}_{chap} = \frac{(n_1 + 1)(n_2 + 1)}{n_{1,2} + 1} - 1, \tag{10}$$

którego wariancja jest równa:

$$\hat{V}(\tilde{N}) = \frac{(n_1 + 1)(n_2 + 1)(n_1 - n_{1,2})(n_2 - n_{1,2})}{(n_{1,2} + 1)^2 (n_{1,2} + 2)}. \tag{11}$$

Metoda capture-recapture swoje początki ma w badaniach przyrodniczych, gdzie nie ma możliwości zbadania wszystkich organizmów, dlatego korzysta się z prób. Jednak w przypadku estymacji wielkości populacji ludzkich (na przykład liczby imigrantów, osób chorych, itp.) wykorzystuje się dane zastane (na przykład źródła administracyjne, bazy danych szpitali, itp.), które rzadko pokrywają badaną populację w całości. Problem związany z wykorzystaniem wielu źródeł administracyjnych do szacowania wielkości populacji jest powiązany z problemem wielu operatów losowania (ang. *multiple frames*).

Kontynuujmy przykład dwóch źródeł, tym razem mogą być to na przykład rejestry administracyjne o wielkościach N_1 oraz N_2, dla których także możemy ustalić ich część wspólną $N_{1,2}$ i podstawiamy dane do wzoru (8):

$$\hat{N}_{pet} = \frac{N_1 N_2}{N_{1,2}}. \tag{12}$$

Powyższy przypadek dotyczy tablicy 1, który możemy również zapisać w postaci nasyconego modelu log-liniowego (ang. *saturated loglinear model*):

$$\log m_{ij} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB}, \tag{13}$$

który musimy ograniczyć przez $\lambda_0^A = \lambda_0^B = \lambda_{00}^{AB} = \lambda_{01}^{AB} = \lambda_{10}^{AB} = 0$ aby móc oszacować. Powyższy model można zapisać symbolicznie w postaci $[AB]$. W związku z tym, że dysponujemy jedynie informacją o n_{11}, n_{01} i n_{10} należy oszacować strukturalne zero n_{00}. Dlatego, że mamy 3 wartości i szacujemy 3 parametry musimy założyć niezależność źródeł. W związku z tym model (13) musimy odpowiednio zmodyfikować:

$$\log m_{ij} = \lambda + \lambda_i^A + \lambda_j^B, \tag{14}$$

zakładając, że $\lambda_0^A = \lambda_0^B = 0$, a model możemy zapisać jako $[A][B]$ co oznacza niezależność dwóch źródeł.

Aby oszacować m_{00} możemy wykorzystać oszacowanie wyrazu wolnego, tj. $\hat{m}_{00} = \exp(\hat{\lambda})$.

W przypadku gdy obserwujemy również zmienną pomocniczą, tak jak w Tablicy 4, wielkość populacji możemy oszacować według następującego wzoru (nazywanego również stratyfikowanym estymatorem Petersena):

$$\hat{N}_{pet,X} = \sum_{x=1}^{2} \frac{N_{1,x} N_{2,x}}{N_{1,2,x}} \tag{15}$$

lub wykorzystując modele log-liniowe:

$$\log m_{ijx} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_x^X + \lambda_{ix}^{AX} + \lambda_{jx}^{BX}, \tag{16}$$

przy założeniu $\lambda_0^A = \lambda_0^B = \lambda_2^X = \lambda_{01}^{AX} = \lambda_{02}^{AX} = \lambda_{12}^{AX} = \lambda_{02}^{BX} = \lambda_{12}^{BX} = 0$. Gdy zakładamy niezależność A i B względem X to $\lambda_{ij}^{AB} = \lambda ijx^{ABX} = 0$. Model dany wzorem (16) zapisujemy jako $[AX][BX]$, natomiast wielkość populacji przy założeniu niezależności A i B dana jest wzorem:

$$\hat{N} = \sum_{x=1}^{2} n_x + \sum_{x=1}^{2} \hat{m}_{00x}. \tag{17}$$

Więcej o modelach log-liniowych można znaleźć w Knoke i Burke (1980); Agresti (2013) lub w języku polskim Brzezińska (2015). Niemniej, w celu ułatwienia czytelności dokumentu, N zarezerwujemy dla wielkości badanej populacji, a n określać będzie próbę (w końcu rejestr administracyjny też jest pewną próbą z superpopulacji o określonej wielkości).

2.3. Trzy źródła

Rozważamy teraz sytuację przedstawioną w Tablicy 2. Nasycony model log-liniowy dany jest wzorem:

$$\log m_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC}, \tag{18}$$

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 oznacza się $[AB][BC][AC]$. Oszacowanie brakującej liczby jednostek populacji otrzymujemy poprzez $\hat{m}_{000} = exp(\lambda)$.

Inny sposób estymacji wielkości populacji dla trzech i więcej źródeł zaproponowali Chao i Tsay (1998); Chao i inni (2001), które określili mianem podejścia pokrycia prób (ang. *sample coverage approach*). Poniżej przedstawiony zostanie przykład trzech źródeł, który jak wskazują Chao i inni (2001, s. 3137) można uogólnić na wiele prób. Główne założenia tej metody przedstawiają się następująco:

- populacja jest zamknięta,

- każda jednostka ma szansę być zapisana w każdym ze źródeł przy czym prawdopodobieństwo nie musi być jednakowe,
- we wszystkich źródłach jednostki są zidentyfikowane i połączone bez błędów.

Podejście to nie zakłada niezależności źródeł oraz homogeniczności prawdopodobieństwa wystąpienia w poszczególnych źródłach. Chao i Tsay (1998) rozważają trzy przypadki biorąc pod uwagę sytuację, w której źródła między sobą nie są niezależne.

Przypadek 1.

Próby są niezależne od siebie. W takim wypadku estymator wielkości populacji oznaczony $\hat{N}_0$ dany jest wzorem:

$$\hat{N}_0 = D/\hat{C}, \tag{19}$$

gdzie

$$\hat{C} = 1 - \frac{1}{3}\left(\frac{Z_{100}}{n_1} + \frac{Z_{010}}{n_2} + \frac{Z_{001}}{n_3}\right) = \frac{1}{3}\left[\left(1 + \frac{Z_{100}}{n_1}\right) + \left(1 + \frac{Z_{010}}{n_2}\right) + \left(1 + \frac{Z_{001}}{n_3}\right)\right] \tag{20}$$

oraz

$$D = \frac{1}{3}\left[(M - Z_{100}) + (M - Z_{010}) + (M - Z_{001})\right] = M - \frac{1}{3}\left(Z_{100} + Z_{010} + Z_{001}\right), \tag{21}$$

gdzie $Z_{100}, Z_{010}, Z_{001}$ określa liczbę jednostek obserwowanych jedynie w źródle 1, 2 i 3, a $M = n_1 + n_2 + n_3$ oznacza sumę wszystkich obserwacji.

Przypadek 2.

Źródła są zależne i pokrycie jest duże (przynajmniej 55% według badań Chao). Wtedy estymator wielkości populacji dany jest wzorem:

$$\hat{N} = \frac{\dfrac{Z_{+11} + Z_{1+1} + Z_{11+}}{3\hat{C}}}{1 - \dfrac{1}{\hat{C}}\left[\dfrac{(Z_{1+0} + Z_{+10})Z_{11+}}{n_1 n_2} + \dfrac{(Z_{10+} + Z_{+01})Z_{1+1}}{n_1 n_3} + \dfrac{(Z_{0+1} + Z_{01+})Z_{+11}}{n_2 n_3}\right]}, \tag{22}$$

gdzie oznaczenia są takie same jak we wcześniejszym wzorze.

Przypadek 3.

Źródła są zależne, a pokrycie jest niewielkie (mniejsze od 55%), wtedy estymator wielkości populacji dany jest wzorem:

$$\hat{N}_1 = \frac{D}{\hat{C}} + \frac{1}{3\hat{C}}\left[(Z_{1+0} + Z_{+10})\hat{\gamma}_{12} + (Z_{10+} + Z_{+01})\hat{\gamma}_{13} + (Z_{01+} + Z_{0+1})\hat{\gamma}_{23}\right], \tag{23}$$

gdzie

$$\hat{\gamma}_{12} = \hat{N}' \frac{Z_{11+}}{n_1 n_2},$$
$$\hat{\gamma}_{13} = \hat{N}' \frac{Z_{1+1}}{n_1 n_3}, \quad (24)$$
$$\hat{\gamma}_{23} = \hat{N}' \frac{Z_{+11}}{n_2 n_3}$$

oraz

$$\hat{N}' = \frac{D}{\hat{C}} + \frac{1}{3\hat{C}} \quad [(Z_{1+0} + Z_{+01})\left(\frac{D}{\hat{C}}\frac{Z_{11+}}{n_1 n_2} - 1\right) + (Z_{10+} + Z_{+01})\left(\frac{D}{\hat{C}}\frac{Z_{1+1}}{n_1 n_3} - 1\right) + (Z_{01+} + Z_{0+1})\left(\frac{D}{\hat{C}}\frac{Z_{+11}}{n_2 n_3} - 1\right)] \quad (25)$$

Powyższe estymatory nazywa się w literaturze pojęciem *sample coverage population size estimators*. Wariancja powyższych estymatorów obliczana jest na podstawie procedury parametrycznego bootstrapu i jest oprogramowana w pakiecie CARE-1 (Hsieh, 2012). Rozszerzenia powyższych estymatorów o wykorzystanie zmiennych pomocniczych oprogramowane są w CARE-2.

W kolejnym podrozdziale przedstawione zostaną przypadki uchylenia niektórych z powyższych założeń.

2.4. Rozszerzenia klasycznych estymatorów capture-recapture

2.4.1. Weryfikacja założeń dotyczących niezależności

Gerritse (2016) rozważa przypadki złamania założenia warunkowej niezależności, nieprecyzyjnego połączenia rekordów (ang. *imperfect linkage*) czy błędnego zliczenia jednostek do badanej populacji (ang. *erroneous captures*).

W pierwszym przypadku zaproponowano miarę implikowanego pokrycia (ang. *implied coverage*, IC), która określa prawdopodobieństwo wpływu jednego źródła na drugie. Aby wyliczyć miarę IC należy wyznaczyć prawdopodobieństwa warunkowe, gdzie przez 0 oznaczamy pierwsze źródło, a przez 1 drugie:

$$\hat{p}(0|1) = \frac{n_{01}}{n_{+1}},$$
$$\hat{p}(1|1) = \frac{n_{11}}{n_{+1}}, \quad (26)$$

gdzie oznaczenia są takie same jak wcześniej. $\hat{p}(0|1)$ określa oszacowane prawdopodobieństwo nowych danych uzyskanych dzięki drugiemu źródle, a $\hat{p}(1|1)$ to oszacowanie prawdopodobieństwa, że dane ze źródła drugiego obserwowane są w źródle pierwszym. Mając te informacje można oszacować liczbę brakujących elementów populacji:

$$\hat{m}_{00} = \frac{n_{10}\hat{p}(0|1)}{\hat{p}(1|1)} = \frac{n_{10}\hat{p}(0|1)}{1-\hat{p}(0|1)}, \tag{27}$$

która jest funkcją prawdopodobieństwa uwzględnienia nowych obserwacji dzięki drugiemu źródłu.

2.4.2. Probabilistyczne łączenie rekordów

2.4.2.1. Uwzględnienie błędu łączenia w estymatorze

W przypadku występowania błędów wynikających z probabilistycznego łączenia rekordów estymator Petersena może być obciążony. Dlatego należy wprowadzić pewne poprawki, które umożliwiają nieobciążone (lub bliskie nieobciążonemu) oszacowanie pokrycia danych źródeł lub wielkości populacji. W niniejszym paragrafie zaprezentowane zostanie podejście opisane w pracach Ding i Fienberg (1994); Di Consiglio i Tuoto (2015). Obydwie prace znoszą założenie dokładnego łączenia rekordów.

Zgodnie z oznaczeniami Ding i Fienberg (1994), niech $L1, L2$ oznaczają dwa źródła danych, β będzie prawdopodobieństwem błędnego połączenia (ang. *false link probability*), a $1-\alpha$ prawdopodobieństwem błędu niepołączenia (ang. *false nonlink probability*). Biorąc pod uwagę poniższe założenia:

- prawdopodobieństwo poprawnego połączenia rekordów między L1, a L2 równe jest α,
- błędne połączenie rekordów, które składają się na część wspólną ($N_{1,2}$) jest pomijane,
- prawdopodobieństwo wystąpienia błędnego połączenia rekordów dla wszystkich jednostek jest takie samo i wynosi β,
- do zbioru $L1$ dołączany jest zbiór $L2$,

estymator zaproponowany przez Ding i Fienberg (1994) dany jest wzorem:

$$\tilde{\tilde{N}}_{DF} = \frac{M}{\hat{\tau}_{1,DF} + \hat{\tau}_{2,DF} - (\alpha-\beta)\hat{\tau}_{1,DF}\hat{\tau}_{2,DF} - \beta\hat{\tau}_{1,DF}}, \tag{28}$$

gdzie $M = m_{11} + m_{12} + m_{21} = n_{11} + n_{12} + n_{21}$ oznacza liczbę rekordów występujących w sumie dwóch zbiorów, gdzie m_{ij} oznacza teoretyczną liczbę rekordów w poszczególnych zbiorach, podczas gdy n_{ij} oznaczają wartości obserwowane w wyniku procedury łączenia rekordów. Oszacowania $\hat{\tau}_{1,DF}$ oraz $\hat{\tau}_{1,DF}$ określają pokrycie pierwszego i drugiego źródła i wyrażają się jako:

$$\hat{\tau}_{1,DF} = \frac{-n_{11} + \beta(n_{11}+n_{12})}{(\beta-\alpha)(n_{11}+n_{21})}, \tag{29}$$

oraz

$$\hat{\tau}_{2,DF} = \frac{-n_{11} + \beta(n_{11}+n_{12})}{(\beta-\alpha)(n_{11}+n_{12})}. \tag{30}$$

Di Consiglio i Tuoto (2015) zmodyfikowali model Ding and Fienberg (1994), który znosi założenie (d) dotyczące kierunku łączenia. Zmodyfikowany estymator Ding i Fienberga dany jest wzorem:

$$\tilde{N}_{MDF} = \frac{M}{\hat{\tau}_{1,MDF} + \hat{\tau}_{2,MDF} - (\alpha\hat{\tau}_{1,MDF}\hat{\tau}_{2,MDF} + \beta(\hat{\tau}_{1,MDF} + \hat{\tau}_{2,MDF} - 2\hat{\tau}_{1,MDF}\hat{\tau}_{2,MDF}))}, \tag{31}$$

gdzie

$$\hat{\tau}_{1,MDF} = \frac{2\beta n_{11} + \beta n_{12} - n_{11}}{(2\beta - \alpha)(n_{11} + n_{21})}, \tag{32}$$

oraz

$$\hat{\tau}_{2,MDF} = \frac{2\beta n_{11} + \beta n_{12} - n_{11}}{(2\beta - \alpha)(n_{11} + n_{12})}. \tag{33}$$

W obydwu przypadkach zakładamy, że błędy połączenia są stałe. W takim wypadku należałoby rozważyć wykorzystanie tych estymatorów w warstwach, na przykład określonych przez płeć, województwo czy grupę wiekową.

2.4.2.2. Badanie wpływu błędu łączenia

Gerritse (2016) symulacyjnie sprawdziła wpływ błędów łączenia, które określone są przez poniższy wzór:

$$\beta = \frac{\tilde{n}_{01}}{n_{01}} = \frac{n_{01} - b}{n_{01}}, \tag{34}$$

gdzie $\beta = 1$ oznacza, że nie ma problemów z łączeniem, a b to liczba jednostek, które zostały błędnie połączone między pierwszym, a drugim źródłem. W takim przypadku estymator nieobserwowanej części populacji wyraża się wzorem:

$$\hat{m}_{00,\beta} = \frac{(n_{10} - b)(n_{01} - b)}{n_{11} - b}, \tag{35}$$

natomiast estymator wielkości populacji równy jest:

$$\hat{N} = \hat{m}_{00,\beta} + (n_{11} + b) + (n_{10} - b) + (n_{01} - b) = \hat{m}_{00,\beta} + n_{11} + n_{10} + n_{01} - b. \tag{36}$$

Należy jednak zaznaczyć, że podejście Gerritse (2016) służy raczej do badania wpływu tegoż błędu na wielkość populacji. W praktyce bardzo trudne jest precyzyjne oszacowanie b.

Podsumowując, powyższe rozwiązanie przedstawione jest jedynie dla tabel 2x2 w związku z tym wymagane są pewne rozszerzenia.

2.4.3 Błędy pokrycia – jednostki, które nie są elementami populacji

2.4.3.1. Trimmed dual system estimation

Trimmed dual system estimation (TDSE) został zaproponowany przez Zhang i Dunne (2017), który bierze pod uwagę sytuację błędnie zakwalifikowanych rekordów w jednym ze źródeł oraz części wspólnej. Estymator ma następującą postać:

$$\hat{N}_{TDSE} = \frac{n_1(n_2 - k)}{n_{1,2} - k_{1,2}}, \tag{37}$$

o wariancji danej wzorem:

$$\hat{V}(\hat{N}_{TDSE}) = \frac{n_1(n_1 - n_{1,2} + k_{1,2})(n_2 - k)(n_2 - k - n_{1,2} + k_{1,2})}{(n_{1,2} - k_{1,2})^3}, \tag{38}$$

gdzie n_1 to liczebność pierwszego źródła, n_2 to liczebność drugiego źródła, $n_{1,2}$ to liczba jednostek wspólnych dla obydwu źródeł, k to liczba jednostek błędnie zakwalifikowanych w drugim zbiorze, a $k_{1,2}$ to liczba jednostek błędnie zakwalifikowanych, które znajdują się w obydwu zbiorach.
Estymator ten ma dwa założenia: brak błędów związanych z łączeniem rekordów (łączenie dokładne, deterministyczne) oraz stała liczba błędnie zakwalifikowanych jednostek. Zaproponowane podejście uwzględnia jedynie dwa źródła.

2.4.3.2. Badanie wpływu błędów pokrycia

Gerritse (2016) symulacyjnie sprawdziła wpływ błędnego zliczania poszczególnych jednostek do badanej populacji zakładając, że wystąpił w drugim zbiorze. Błąd ten oznaczony jest przez γ i zawiera się w przedziale $[0,1]$. W takim przypadku część wspólna dana jest wzorem:

$$\hat{m}_{00,\gamma} = \frac{n_{10}(n_{01}\gamma)}{n_{11}} = \gamma \hat{m}_{00}. \tag{39}$$

2.4.3.3. Szczególny przypadek trzech list

Zhang (2015) rozważał przypadek trzech źródeł, w których źródło A i B charakteryzowało się błędami pokrycia (nadreprezentacji i niedoreprezentacji), a trzecie źródło S charakteryzuje się jedynie błędem pokrycia oraz stałą frakcją jednostek wspólnych (tj. taka sama liczba jednostek wspólnych według płci,

województwa i innych zmiennych). Przykładowym źródłem S może być badanie reprezentacyjne wykorzystywane do badania pokrycia (ang. *coverage survey*), a pozostałymi źródłami mogą być rejestry oraz spis. Zhang (2015) wziął pod uwagę dwie sytuacje nadreprezentacji:

1. Model 1 – założenie, że jednostki, które nie należą do populacji znajdują się jedynie w dwóch źródłach, a nie części wspólnej:

$$P(i \notin U \mid i \in A \cup B) = P(i \notin U \mid i \in A \setminus B)P(i \notin \mid i \in B \setminus A). \tag{40}$$

Wtedy błąd pokrycia dla poszczególnych list jest równy:

$$\hat{\theta}_{10} = \frac{x_{01}}{n_{01}}\left(\frac{n_{11}}{x_{11}} - \frac{n_{10}}{x_{10}}\right),$$

$$\hat{\theta}_{01} = \frac{x_{10}}{n_{10}}\left(\frac{n_{11}}{x_{11}} - \frac{n_{01}}{x_{01}}\right),$$

a pokrycie (ang. *capture rate*) zbioru S jest równe:

$$\hat{\pi} = \frac{n_{10}}{x_{10}(1-\hat{\theta}_{10})} = \frac{n_{01}}{x_{01}(1-\hat{\theta}_{01})}.$$

2. Model 2 – założenie, że jednostki, które nie należą do populacji znajdują się jedynie w dwóch źródłach, a w części wspólnej jest ich bardzo mało (pomijalne):

$$P(i \notin U \mid i \in A \cup B) = P(i \notin U \mid i \in A)P(i \notin \mid i \in B). \tag{41}$$

Wtedy błąd pokrycia dla poszczególnych źródeł jest równy:

$$\hat{\theta}_{1+} = \frac{x_{+1}}{n_{+1}}\left(\frac{n_{11}}{x_{11}} - \frac{n_{1+}}{x_{1+}}\right),$$

$$\hat{\theta}_{+1} = \frac{x_{1+}}{n_{1+}}\left(\frac{n_{11}}{x_{11}} - \frac{n_{+1}}{x_{+1}}\right),$$

a pokrycie (ang. *capture rate*) zbioru S jest równe:

$$\hat{\pi} = \frac{n_{1+}}{x_{1+}(1-\hat{\theta}_{1+})} = \frac{n_{+1}}{x_{+1}(1-\hat{\theta}_{+1})}.$$

gdzie $x_{ab} = \#(A \cap B)$, a $n_{ab} = \#(A \cap B \cap S)$ to liczba jednostek, które znajdują się również w zbiorze S, gdzie $\#$ oznacza liczbę jednostek danego zbioru.

Ostatecznie, estymator wielkości populacji niezależnie czy rozważamy model 1 czy 2 dany jest wzorem:

$$\hat{N} = x_{11}(1-\hat{\theta}_{11}) + x_{10}(1-\hat{\theta}_{10}) + x_{01}(1-\hat{\theta}_{01}) + \frac{n_{00}}{\hat{\pi}}. \tag{42}$$

2.4.4. Brak operatów dla badanej populacji – losowanie pośrednie

Lavallée i Rivest (2012) rozważali sytuację, w której nie jest możliwe pełne wykorzystanie dostępnych źródeł, a jedynie prób z nich pobranych. Dodatkowo, rozszerzają podstawowy estymator Petersena na sytuację, w której wykorzystane operaty losowania (i próby) nie dotyczą bezpośrednio populacji (na przykład nie istnieje taki operat), którą badają, a innych populacji, które są z nimi związane (na przykład populacja osób -> gospodarstwa domowe). Wykorzystali w tym celu losowanie pośrednie – *indirect sampling* (Deville i Lavallée, 2006). Lavallée i Rivest (2012) rozważali przypadek estymacji braku pokrycia gospodarstw domowych przez spis (ang. *census undercount*) oraz sprzedawców herbaty (w tym osób, które robią to nielegalnie).

3. Modelowanie wielkości populacji

3.1. Analiza klas ukrytych

Zmienne ukryte stanowią podstawę modeli ze zmiennymi ukrytymi, które składają się na szerzej rozumiane metody struktur ukrytych (ang. *latent structure methods*). Podziału tych metod dokonuje się ze względu na charakter zmiennej obserwowalnej oraz zmiennej ukrytej. Gdy zarówno zmienna obserwowalna, jak i zmienna ukryta mają charakter dyskretny, metoda ta nazywana jest analizą klas ukrytych (ang. *latent class analysis*).

Głównym celem analizy klas ukrytych jest redukcja liczby zmiennych przy jak najmniejszej utracie informacji o badanym zjawisku, a także odkrycie nieobserwowalnej heterogeniczności w populacji. Klasy ukryte pełnią wówczas funkcję nieobserwowalnych czynników, które wpływają na zależność pomiędzy jednostkami w danej klasie.

Analiza klas ukrytych oparta jest na dwóch założeniach. Pierwsze założenie mówi o tym, że populacja składa się z rozłącznych (ang. *mutually exclusive*) i spójnych (ang. *exhaustive*) jednorodnych podpopulacji, które łącznie tworzą klasę ukrytą. Oznacza to, że jeden obiekt może należeć tylko do jednej klasy ukrytej. Drugie założenie nazywane jest warunkiem lub aksjomatem lokalnej niezależności (ang. *local independence assumption*), zgodnie z którym związek między zmiennymi obserwowalnymi zależy od relacji pomiędzy zmiennymi obserwowalnymi a zmiennymi ukrytymi. Oznacza to, że jeśli zmienna ukryta jest stała, zmienne obserwowalne powinny być statystycznie niezależne. Warunek ten spełniony jest we wszystkich rodzajach modeli struktur ukrytych.

Analiza klas ukrytych ma na celu znalezienie oraz zidentyfikowanie odpowiedniej liczby klas, w których zmienne obserwowalne są od siebie niezależne. Inaczej mówiąc, metoda ta umożliwia rozwarstwienie tablicy kontyngencji zawierającej zmienne obserwowalne przez zmienną ukrytą, przy czym poszczególne klasy stanowią kategorie zmiennej ukrytej o charakterze dyskretnym. Model taki w efekcie przydziela obserwacje do klas ukrytych, a w dalszym etapie pozwala na przypuszczenie, jak zmienne obserwowalne zachowają się pod wpływem zmiennych ukrytych.

W modelach klas ukrytych wyróżnia się następujące rodzaje zmiennych, w zależności od rodzaju skali pomiaru:

- zmienne ukryte (ang. latent variables), które mogą być mierzone na skalach nominalnych lub porządkowych,
- zmienne obserwowalne (ang. *manifest vaviables, response variaables*) lub zmienne objaśniane (ang. *dependent variables*), które mogą być mierzone na różnych skalach pomiaru,
- zmienne towarzyszące (ang. *concomitant variables, covariates*) i zmienne objaśniające (ang. *predictor variables*), które mogą być mierzone na różnych skalach pomiaru.

Model musi zawierać przynajmniej jedną zmienną ukrytą i jedną zmienną obserwowalną; może on także zawierać zmienne towarzyszące. Zmienna ukryta jest zatem statyczna i dzieli populację na podpopulacje, zwane klasami ukrytymi.

W analizie klas ukrytych modele różnią się jedynie liczbą klas ukrytych. Modele zawierające większą liczbę parametrów (większą liczbę klas ukrytych) zapewniają lepsze dopasowanie do danych niż te, które opisane są przez mniejszą liczbę klas por. Brzezińska (2015). Podstawą analizy klas ukrytych w przypadku wykorzystania zmiennych pomocniczych jest przykładowo Tablica 5.

Modele klas ukrytych można interpretować w odniesieniu do prawdopodobieństwa, że dana jednostka została zakwalifikowana do danego źródła danych oraz skłonności do przynależności do określonej klasy ukrytej. Modele klas ukrytych były wykorzystywane do szacowania wielkości populacji m.in. przez:

- Agresti (1994) – wykorzystał klasy ukryte bez zmiennych towarzyszących do szacowania wielkości populacji zwierząt,
- Dorazio i Andrew Royle (2003); Böhning i inni (2005) – wykorzystali klasy ukryte bez zmiennych towarzyszących do szacowania wielkości populacji ludzkich,
- Bartolucci i Forcina (2006); Stanghellini i van der Heijden (2004); Van Der Heijden i inni (2003a); Cruyff i van der Heijden (2008); Van Der Heijden i inni (2003b); Thandrayen i Wang (2009) – wykorzystali klasy ukryte ze zmiennymi towarzyszącymi do szacowania wielkości populacji.

Tablica 5. Przypadek czterech źródeł – inny sposób zapisu

Liczebność	Z1	Z2	Z3	Płeć
1	0	0	0	1
2	0	1	1	1
3	1	0	1	1
4	0	1	0	2
5	1	1	0	2
6	0	1	1	2
7	0	1	1	2
8	1	0	0	1
9	1	1	0	2
10	1	1	1	1

…	…	…	…	..
n	1	0	1	2

Źródło: Opracowanie własne

Niech N będzie nieznaną wielkością populacji, a $i = 1,\mathrm{K}\,,N$ określa identyfikator poszczególnych jednostek. Niech $\mathbf{Y}_i = (Y_{i1},...,Y_{iJ})^T$ będzie wektorem składającym się z wartości 0,1 (rozkład Bernoulliego) o wielkości J, która określa liczbę dostępnych źródeł danych. Wartości $\{0,1\}$ określają czy dana jednostka była obserwowana w danym źródle. Niech $\mathbf{Y}$ o rozmiarze $N \times J$ określa macierz przynależności poszczególnych jednostek do określonej listy.

Załóżmy teraz, że populację możemy podzielić na L klas i niech S_i określa przynależność danej i-tej jednostki do danej klasy. Niech $p(S_i = l) = q_i, l = 1,\mathrm{K}\,,L$ określa prawdopodobieństwo przynależności do klasy ukrytej. Niech $\mathbf{P}$ określa macierz prawdopodobieństw pojawienia się w danym źródle o wymiarze $J \times L$, gdzie każdy element określony jest jako $p_{jl}, j = 1,..,J, l = 1,...,L$.

W przypadku braku zmiennych pomocniczych, prawdopodobieństwo przynależności do danej klasy dane jest wzorem:

$$p(\mathbf{Y}_i = \mathbf{y}_i) = \sum_{l=1}^{L} q_l \prod_{j=1}^{J} p_{jl}^{y_{ij}} (1 - p_{jl})^{1-y_{ij}}, \tag{43}$$

natomiast jeżeli założymy, że dysponujemy informacjami pomocniczymi $\mathbf{x}_i = (x_{i1},\mathrm{K}\,,x_{1H})^T$, gdzie H określa liczbę zmiennych pomocniczych, wtedy model (43) określony jest wzorem:

$$p(\mathbf{Y}_i = \mathbf{y}_i \mid \mathbf{x}_i) = \sum_{l=1}^{L} q_l(\mathbf{x}_i) \prod_{j=1}^{J} p_{jl}^{y_{ij}} (1 - p_{jl})^{1-y_{ij}}, \tag{44}$$

w którym $q_l(\mathbf{x}_i)$ i p_{jl} nie jest znane i musi być estymowane. Procedura estymacji opisana jest szczegółowo w pracy Thandrayen i Wang (2009). Po oszacowaniu $\hat{p}_{ij}$ oraz $q_l(\mathbf{x}'_i \hat{\boldsymbol{\beta}})$ możemy obliczyć wielkość populacji zgodnie z następującym wzorem:

$$\hat{N} = \sum_{i=1}^{N} \frac{I_i}{\hat{\pi}_i}, \tag{45}$$

gdzie $I_i = \{0,1\}$ określa czy dana jednostka była obserwowana w próbie czy nie, a $\hat{\pi}_i = 1 - \sum_{l=1}^{L} q_l(\mathbf{x}'_i \hat{\boldsymbol{\beta}}) \prod_{j=1}^{J} (1 - \hat{p}_{ij})$.

Istotnym wkładem pracy Baffour-Awuah (2009) jest zaproponowanie modelu klas ukrytych do oszacowania wielkości populacji z uwzględnieniem błędów pokrycia w badanych źródłach.

3.2. Wybrane modele regresji w estymacji wielkości populacji

3.2.1. Model regresji Poissona i jego rozszerzenia

W przypadku regresji Poissona i jej rozszerzeń (na przykład tzw. Zero-inflated Poisson) przy wykorzystaniu zmiennych pomocniczych estymacja wielkości populacji opiera się na Tablicy 6, gdzie kolumna „Liczba zliczeń" odnosi się do liczby wystąpień danej jednostki w badanych źródłach (przykładowo liczba 3 oznacza, że jednostka pojawiła się w trzech źródłach) i przebiega w następujący sposób:

- szacuje się prawdopodobieństwo p wystąpienia jednostki w jednym, dwóch i więcej źródeł (zakładając rozkład Poissona dla zmiennej 'Liczba zliczeń') – podstawą jest tabela z charakterystykami osób $\mathbf{x}_i$ (na przykład płeć, grupa wieku, obywatelstwo) oraz liczba wystąpień w poszczególnych źródłach ('Liczba zliczeń'),
- następnie dla każdej osoby sumuje się odwrotność tego prawdopodobieństwa i otrzymujemy wielkość populacji.

Tablica 6. Przykładowe dane wykorzystywane w procesie modelowania

ID	Liczba zliczeń	Płeć	Grupa wieku	Obywatelstwo polskie
1	3	0	1	1
2	1	0	2	0
3	2	0	3	1
4	5	0	1	0
5	1	1	3	1
6	1	1	4	0
7	2	1	1	1
...	...	...	...	...
n	1	1	2	1

Źródło: Opracowanie własne

W takim przypadku estymator wielkości populacji dany jest wzorem:

$$\hat{N} = \sum_{i=1}^{N} \frac{I_i}{p(\mathbf{x}_i, \boldsymbol{\beta})} = \sum_{i=1}^{n} \frac{1}{p(\mathbf{x}_i, \boldsymbol{\beta})}, \tag{46}$$

gdzie $I_i = 0$ gdy jednostka nie była obserwowana w żadnym ze źródeł, a $I_i = 1$ gdy była obserwowana przynajmniej w jednym źródle, $p(\mathbf{x}_i, \boldsymbol{\beta})$ to prawdopodobieństwo bycia obecnym w jednym, dwóch lub więcej źródeł uwzględniając charakterystyki $\mathbf{x}_i$ jednostek, które są obserwowane. Warto zauważyć, że

estymator ten jest w swojej formie podobny do estymatora Horvitza-Thompsona i takiej nazwy używa się do określenia tego estymatora (na przykład estymator wielkości populacji Horvitza-Thompsona). Wariancja estymatora (46) dana jest wzorem:

$$var(\hat{N}) = E[var(\hat{N} \mid I_i)] + var(E[\hat{N} \mid I_i]). \tag{47}$$

Należy zauważyć, że aby oszacować wielkość populacji wystarczy informacja pochodząca z próby. Szczegóły estymacji i rozszerzeń powyższych modeli można znaleźć w Van Der Heijden i inni (2003a,b).

3.2.2. Hierarchiczny model mieszany Poissona

Ciekawą koncepcję, w kontekście szacowania wielkości populacji, przedstawił Zhang (2008), który na potrzeby estymacji liczby nielegalnie przebywających w Norwegii cudzoziemców wykorzystał modele mieszane. Bardziej precyzyjnie w celu oszacowania tej wielkości wykorzystał on hierarchiczny model gamma Poissona (ang. *A hierarchical Poisson gamma model*), w którym wprowadził efekt losowy w postaci kraju nielegalnie przebywającego w Norwegii cudzoziemca.

Z formalnego punktu widzenia w procesie estymacji przyjęto, że liczba nielegalnie przebywających w Norwegii cudzoziemców podlegała rozkładowi Poissona postaci:

$$m_i \sim Poisson(\lambda_i), \tag{48}$$

gdzie $i = 1, \mathrm{K}, t$ jest indeksem odnoszącym się do podpopulacji cudzoziemców według kraju ich pochodzenia. Kluczową kwestią jest zatem estymacja jedynego w rozkładzie Poissona parametru – λ_i. W badaniu przyjęto założenie, że parametr λ_i zależy od dwóch wielkości: całkowitej liczby nielegalnie przebywających cudzoziemców z kraju i, oznaczanej jako M_i oraz prawdopodobieństwa p_i, że nielegalnie przebywający cudzoziemiec znajduje się w tzw. rejestrze DUF odnoszącym się do imigrantów i uchodźców, którzy chcą zamieszkać w Norwegii. W takim przypadku można przyjąć, że:

$$\lambda_i = M_i p_i. \tag{49}$$

W modelu tym przyjęto, że:

$$u_i = \frac{M_i p_i}{E(M_i p_i \mid n_i, N_i)}, \tag{50}$$

gdzie $E(M_i p_i \mid n_i, N_i)$ oznacza warunkową wartość oczekiwaną $M_i p_i$ przy danym n_i oraz N_i, przy czym n_i oznacza liczbę cudzoziemców z kraju i, którzy dopuścili się przestępstwa a N_i liczbę legalnych cudzoziemców z kraju i. Z kolei u_i oznacza efekt losowy, który odpowiada za zmienność szacowanego parametru w obrębie różnych krajów, z których pochodzić mogą nielegalni cudzoziemcy. Przy przyjętych oznaczeniach mamy:

$$\lambda_i = \mu_i u_i, \tag{51}$$

gdzie:

$$\mu_i = E(M_i p_i \mid n_i, N_i) = E(M_i \mid N_i) \cdot E(p_i \mid M_i, n_i, N_i). \tag{52}$$

Ostatecznie model wykorzystywany w estymacji liczby nielegalnie przebywających w Norwegii cudzoziemców można zapisać w następującej postaci:

$$\xi_i = E(M_i \mid N_i) = N_i^{\alpha_i}, \tag{53}$$

$$E(p_i \mid M_i, n_i, N_i) = E(p_i \mid n_i, N_i) = \left(\frac{n_i}{N_i}\right)^{\beta}, \tag{54}$$

$$u_i \sim Gamma(1, \phi), \tag{55}$$

gdzie $Gamma(1,\phi)$ oznacza rozkład Gamma z wartością oczekiwaną $E(u_i) = 1$ i wariancją $V(u_i) = 1/\phi$. Formuły (48)–(54) definiują tzw. hierarchiczny model gamma Poisonna. Konsekwencją tak przyjętego modelu jest następujące równanie:

$$E\left(\frac{m_i}{M_i} \mid M_i, n_i, N_i\right) = \left(\frac{n_i}{N_i}\right)^{\beta}. \tag{56}$$

Hierarchiczność modelu odnosi się do zmienności, która jest obserwowana na dwóch poziomach. Po pierwsze, na poziomie populacji λ_i, która jest parametrem w modelu Poissona, zależy od u_i, który jest zmienną losową o rozkładzie Gamma wśród wszystkich krajów. Z kolei μ_i zawiera efekty stałe α i β, które są takie same dla wszystkich krajów. Po drugie przy danym u_i, m_i obarczone jest błędem losowym podlegającym rozkładowi Poissona.

Z równań (52)–(54) wynika, że:

$$\mu_i = N_i^{\alpha_i}\left(\frac{n_i}{N_i}\right)^{\beta}. \tag{57}$$

Kluczową kwestią jest oszacowanie parametrów α, β i ϕ. W praktyce robi się to z wykorzystaniem metody największej wiarogodności – MNW (ang. Maximum Likelihood Estimation – MLE). Przyjmując, że:

$$\xi = \sum_{i=1}^{t} E(M_i \mid N_i) = \sum_i N_i^{\alpha}, \tag{58}$$

jego estymator możemy wyrazić jako $\hat{\xi} = \sum_i N_i^{\hat{\alpha}}$ gdzie $\hat{\alpha}$ jest oszacowaniem parametru α. Oznaczając przez $L(\eta; \mathbf{m})$ funkcję wiarogodności $\eta = (\alpha, \beta, \phi)$ dla danego m_i, $i = 1, \mathrm{K}, t$ przy założeniu hierarchicznego modelu gamma Poissona funckję gęstości można wyrazić następującym wzorem:

$$f(m_i, u_i; \eta) = \frac{e^{-\mu_i u_i}(\mu_i u_i)^{m_i}}{m_i!} \cdot \frac{\phi^{\phi} u_i^{\phi-1} e^{-\phi u_i}}{\Gamma(\phi)} = \frac{\mu_i^{m_i}\phi^{\phi}}{m_i!\Gamma(\phi)} e^{-u_i(\mu_i+\phi)} u_i^{m_i+\phi-1}, \tag{59}$$

gdzie $\Gamma(\ )$ jest tzw. funkcją gamma. Stąd:

$$f(m_i;\eta)=\int_0^\infty f(m_i,u_i;\eta)d(u_i)$$

$$=\frac{\mu_i^{m_i}\phi^\phi}{m_i!\Gamma(\phi)}\int_0^\infty e^{-\left(\sqrt{u_i}\right)^2(\mu_i+\phi)}\left(\sqrt{u_i}\right)^{2(m_i+\phi-1)}2\sqrt{u_i}\,d\left(\sqrt{u_i}\right)$$

$$=\frac{\mu_i^{m_i}\phi^\phi}{m_i!\Gamma(\phi)}(\mu_i+\phi)^{-(m_i+\phi)\Gamma(m_i+\phi)} \tag{60}$$

co wynika z warunku:

$$\int_0^\infty e^{-\gamma z^2}z^k dz=\frac{1}{2}\gamma^{-\frac{k+1}{2}}\Gamma\left(\frac{k+1}{2}\right), \tag{61}$$

przy czym $z=\sqrt{u_i}$ oraz $k=2(m_i+\phi)-1$. Wspomnianą powyżej funkcję wiarogodności $L(\eta;\mathbf{m})$ można zatem wyrazić wzorem:

$$L(\eta;\mathbf{m})=\prod_{i=1}^{t}f(m_i;\eta). \tag{62}$$

Logarytm funkcji wiarogodności, pomijając stałą, wyraża się wzorem:

$$l(\eta;\mathbf{m})=\sum_{i=1}^{t}l_i(\eta), \tag{63}$$

gdzie:

$$l_i(\eta)=m_i\log\mu_i-(m_i+\phi)\log(\mu_i+\phi)+\log\Gamma(m_i+\phi)+\phi\log\phi-\log\Gamma(\phi)$$

$$=m_i\log\mu_i-(m_i+\phi)\log(\mu_i+\phi)+\phi\log\phi$$

$$+(m_i+\phi-0.5)\log(m_i+\phi)-(m_i+\phi)-(\phi-0.5)\log\phi+\phi$$

$$=m_i\log\mu_i-(m_i+\phi)\log(\mu_i+\phi)+(m_i+\phi-0.5)\log(m_i+\phi)+0.5\log\phi,$$

przy czym skorzystano z aproksymacji wzoru Stirlinga:

$$\log\Gamma(z)=(z-0.5)\log z+0.5\log(2\pi)-z. \tag{64}$$

Zważywszy na fakt, że $\log\mu_i$ jest liniową kombinacją wektora zmiennych pomocniczych x_i i parametru γ tj. $\log\mu_i=x_i^T\gamma$ mamy:

$$\frac{\partial l_i(\eta)}{\partial\gamma}=\frac{\partial l_i(\eta)}{\partial\mu_i}\frac{\partial\mu_i}{\partial\log\mu_i}\frac{\partial\log\mu_i}{\partial\gamma}=\frac{\partial l_i(\eta)}{\partial\mu_i}\mu_i x_i=\frac{m_i-\mu_i}{\mu_i+\phi}\phi x_i, \tag{65}$$

przy czym:

$$\frac{\partial l_i(\eta)}{\partial\mu_i}=\frac{m_i}{\mu_i}-\frac{m_i+\phi}{\mu_i+\phi} \tag{66}$$

oraz:

$$\frac{\partial l_i(\eta)}{\partial \phi} = -\log(\mu_i + \phi) - \frac{m_i + \phi}{\mu_i + \phi} + \log(m_i + \phi) + \frac{m_i + \phi - 0.5}{m_i + \phi} + \frac{1}{2\phi}. \tag{67}$$

Co więcej, pochodne drugiego rzędu można zapisać jako:

$$\frac{\partial^2 l_i(\eta)}{\partial \gamma \partial \gamma^T} = \frac{\partial^2 l_i(\eta)}{\partial \mu_i^2} \mu_i x_i \frac{\partial \mu_i}{\partial \gamma^T} + \frac{\partial^2 l_i(\eta)}{\partial \mu_i} x_i \frac{\partial \mu_i}{\partial \gamma^T} = -\left(\frac{m_i + \phi}{\mu_i + \phi} \mu_i \phi \right) x_i x_i^T, \tag{68}$$

$$\frac{\partial^2 l_i(\eta)}{\partial \phi^2} = -\frac{2\mu_i + \phi - m_i}{(\mu_i + \phi)^2} + \frac{m_i + \phi + 0.5}{(m_i + \phi)^2} - \frac{1}{2\phi^2}, \tag{69}$$

$$\frac{\partial^2 l_i(\eta)}{\partial \gamma \partial \phi} = \left(\partial \left(\frac{\partial l_i(\eta)}{\partial \mu_i} \right) / \partial \phi \right) \mu_i x_i = -\frac{\mu_i - m_i}{(\mu_i + \phi)^2} \mu_i x_i = \left(\frac{\partial^2 l_i(\eta)}{\partial \phi \partial \gamma^T} \right)^T. \tag{70}$$

Estymator MNW parametru η, który oznaczamy przez $\hat{\eta}$ poszukuje się następnie jako rozwiązanie następujących równań:

$$\frac{\partial l(\eta; \mathbf{m})}{\partial \eta} = \sum_{i=1}^{t} \frac{\partial l_i(\eta)}{\partial \eta} = 0. \tag{71}$$

Estymator MNW parametru η można znaleźć wykorzystując metodę Newtona-Raphsona. Jako wartości początkowe w tej metodzie przyjmujemy wstępne oszacowania parametrów uzyskanych klasyczną metodą najmniejszych kwadratów poprzez dopasowanie modelu postaci:

$$\log \frac{m_i}{N_i} = (\alpha - 1) \log N_i + \beta \log \frac{n_i}{N_i} + \varepsilon_i. \tag{72}$$

Po oszacowaniu parametrów α i β modelu (72) traktujemy je jako wartości początkowe tych samych parametrów w hierarchicznym modelu gamma Poissona a odwrotność oszacowanej macierzy wariancji-kowariancji $V(\varepsilon_i)$ jako wartość początkową dla ϕ. W ten sposób można uzyskać oszacowania wszystkich parametrów rozważanego w tej części opracowania hierarchicznego modelu gamma Poisona, który następnie można wykorzystać do oszacowania liczby nielegalnie przebywających imigrantów.

4. Estymacja wielkości populacji trudnych do zbadania w wybranych krajach – przegląd doświadczeń

W rozdziale tym przedstawiono doświadczenia wybranych państw w obszarze estymacji wielkości populacji. Opis ma charakter syntetyczny i składa się z informacji o analizowanej populacji, wykorzystanych źródłach danych, sposobie łączenia danych oraz zastosowanej metodzie estymacji wielkości populacji.

4.1. Estymacja liczby niezarejestrowanych rezydentów w Holandii

Populacja[1]: osoby zamieszkałe w Holandii w wieku 15–65 w 2010 roku.

Źródła: wykorzystano trzy źródła – rejestr ludności (Population Register – PR), rejestr zatrudnionych (Employment Register – ER) oraz rejestr podejrzanych o przestępstwa prowadzony przez policję (Crime Suspects Register – CSR). Tablica 7 przedstawia informacje o liczebnościach wspólnych dla tych rejestrów. Zastosowano imputację zmiennej dotyczącej czasu pobytu tzw. metodą Predictive Mean Matching.

Sposób łączenia: wykorzystano łączenie deterministyczne (identyfikator osoby) oraz probabilistyczne. Tablica 8 przedstawia udział poszczególnych typów łączenia w zależności od źródła. Największy odsetek trafnych połączeń w przypadku łączenia probabilistycznego występował w odniesieniu do łączenia PR i CSR oraz ER i CSR[2]. W przypadku łączenia probabilistycznego zastosowano blokowanie (według kodu pocztowego i daty urodzenia) a procesu połączenia dokonywano następująco:

- PR-ER oraz PR-CSR – data urodzenia, płeć, kod pocztowy, numer doku oraz suffix (blokowanie według kodu pocztowego, daty urodzenia),
- ER-CSR – data urodzenia, miejsce zamieszkania, adres, numer mieszkania (blokowanie według daty urodzenia lub miejsca zamieszkania, miesiąca lub dnia urodzenia).

Tablica 7. Pokrycie między źródłami – PR, ER i CSR (w tysiącach)

		CSR		
PR	ER	1 tak	0 nie	Razem
1 tak	1 tak	2.1	259.8	261.9
1 tak	0 nie	4.9	350.6	355.4
0 nie	1 tak	0.4	112.5	112.9
0 nie	0 nie	5.1	.-	5.1
	Razem	12.4	722.9	735.3

Źródło: Opracowanie własne na podstawie Bakker i inni (2017)

Tablica 8. Informacje o łączeniu rejestrów

		Niepołączony	Połączony		Razem
Połączenie	Źródło				
			Deterministyczne	Probabilistyczne	
		%	%	%	x1000

[1] Opracowano na podstawie Bakker i inni (2017).

[2] Uwaga. W tabeli poszczególne wartości należy interpretować jako: 57.6% wszystkich jednostek w PR jest połączonych z ER, 42.4% wszystkich jednostek zostało połączonych deterministycznie a żadna probabilistycznie. Daje to łącznie 617.3 tys. jednostek.

PR ↔ ER	PR	57.6	42.4	0.0	617.3
	ER	30.1	69.9	0.0	374.8
PR ↔ CSR	PR	98.9	1.1	0.0	617.3
	CSR	43.8	54.3	1.9	12.4
ER ↔ CSR	ER	99.3	0.6	0.1	374.8
	CSR	80.2	17.8	2.0	12.4

Źródło: Opracowanie własne na podstawie Bakker i inni (2017)

Model: wykorzystano model log-liniowy z i bez zmiennych pomocniczych (czas pobytu – poniżej 1 roku i powyżej, płeć oraz grupa wieku). Przeprowadzono analizę wrażliwości wyników na błędy braku połączenia oraz łączenia.

4.2. Estymacja liczby osób bezdomnych w Holandii metodą capture-recapture

Populacja[3]: osoby bezdomne w wieku 18–64.

Źródła: zarejestrowani w schroniskach (Basic Municipal Administration System), osoby które dostały wsparcie ale nie miały stałego miejsca zamieszkania (Received Income Support), zarejestrowani jako osoby bezdomne w rejestrze osób nadużywających alkoholu i narkotyków (National Alcohol and Drugs Information System).

Sposób łączenia: identyfikator osoby (łączenie deterministyczne).

Model: wykorzystano model log-liniowy ze zmiennymi określającymi charakterystyki osób bezdomnych i ich interakcją ze źródłami danych (płeć, grupa wieku, miejsce oraz pochodzenie). Modele liczone były oddzielnie dla poszczególnych lat 2009–2013.

4.3. Estymacja wielkości populacji przestępców w Holandii z wykorzystaniem rejestrów policji

Populacja[4]: liczba osób w Holandii, która prowadzi pojazdy w stanie nietrzeźwym oraz liczba osób, która posiada broń, a jej nie zarejestrowała. Populacja ograniczona jest do osób, które można złapać (na przykład wyłączone są osoby, które mają broń zakopaną i nigdy jej nie użyją).

Źródła: Dane policyjne za lata 1996–2001 ograniczono jedynie do pięciu regionów, dla których dane charakteryzowały się największą jakością. Tablica 9 zawiera informację o liczbie osób, które w badanych latach zostały złapane na przestępstwach. Lewa część tabeli odnosi się do danych obserwowanych i szacowanych na podstawie modelu (nielegalne posiadanie broni palnej). Z kolei prawa strona tabeli odnosi się do wartości

[3] Opracowano na podstawie Coumans i inni (2017).
[4] Opracowano na podstawie Van Der Heijden i inni (2003b).

obserwowanych i modelowych liczby osób prowadzących samochód w stanie nietrzeźwym.

Sposób łączenia: brak łączenia, ponieważ wykorzystano jedną bazę za lata od 1996 do 2001 roku.

Model: zastosowano regresję Poissona z wyłączeniem zera (ang. *Zero-Truncated Poisson Model*). Wykorzystano następujące zmienne: typ przestępstwa (6 poziomów), wiek, płeć, wiek w którym pierwszy raz popełniono przestępstwo, liczbę popełnionych przestępstw oraz region zdefiniowany przez policję.

Tablica 9. Dane będące podstawą analizy wraz z oszacowaniami pochodzącymi z modelu

k	Rzeczywiste	Oszacowane	Reszty	Rzeczywiste	Oszacowane	Reszty
	0	60,084.0	-	0	104,352.0	-
1	2,561	2,558.9	0.04	8,877	8,847.2	0.32
2	72	76.4	-0.50	481	534.4	-2.31
3	5	2.6	1.48	52	34.0	3.08
4	-	-	-	8	2.9	2.98
5	-	-	-	1	0.4	1.06

Źródło: Opracowanie własne na podstawie Van Der Heijden i inni (2003b)

4.4. Estymacja liczby nielegalnych cudzoziemców w Norwegii

Populacja[5]: nielegalnie przebywający w Norwegii cudzoziemcy na dzień pierwszego stycznia 2006 r.

Źródła: na potrzeby estymacji wielkości populacji odnoszącej się do nielegalnie przebywających w Norwegii cudzoziemców wykorzystane zostały trzy źródła danych. Pierwsze źródło stanowił Centralny Rejestr Osób (Central Personel Register) z którego wykorzystano informacje na temat liczby 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 dopuścili się przestępstw i byli sądzeni. Tego typu informacje dostarcza Krajowy Urząd Statystyczny w Norwegii. Ostatnim źródłem danych był rejestr 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.

Łączenie: nie dokonywano łączenia danych, które wykorzystano w procesie modelowania w zagregowanej formie.

Model: na potrzeby estymacji wielkości populacji zdefiniowanej jako liczba nielegalnie przebywających w

[5] Opracowano na podstawie Zhang (2008).

Norwegii cudzoziemców wykorzystano hierarchiczny model gamma Poisona, który należy do rodziny modeli mieszanych z efektami losowymi. W charakterze efektu losowego wykorzystano kraj pochodzenia cudzoziemców.

4.5. Estymacja wielkości populacji Irlandii na podstawie dwóch rejestrów

Populacja[6]: mieszkańcy Irlandii.

Źródła: Rejestr Aktywności Ludzkiej (Person Activity Register – PAR), który utworzony jest na podstawie wielu rejestrów odnoszących się m.in. do edukacji, opieki społecznej, emerytur, podatków oraz Rejestr Praw Jazdy (Driving Licence Dataset – DLD).

Łączenie: deterministyczne po zakodowanym identyfikatorze osoby.

Model: zastosowano metodę typu DSE (ang. Dual-System Estimation) oraz jej odmianę TDSE (ang. Trimmed Dual-System Estimation). Zastosowano ograniczenie zbiorów (ang. *trimming*).

5. Oprogramowanie

W rozdziale tym wyszczególniono najważniejsze oprogramowanie, które może być przydatne na potrzeby estymacji wielkości populacji z wykorzystaniem wielu źródeł danych:

- RecordLinkage (Borg i Sariyar, 2016), fastLink (Enamorado i inni, 2017), Relais – oprogramowanie przydatne w procesie probabilistycznego łączenia rekordów,
- SMERED: Split and MErge REcord linkage and De-duplication toolbox for Java – (Steorts i inni, 2016), https://bitbucket.org/resteorts/smered/,
- Pakiety R: CARE1 (Hsieh, 2012), Rcapture (Rivest i Baillargeon, 2014), unmarked (Fiske i Chandler, 2011), multimark (McClintock, 2015)– pakiety R do oszacowania wielkości populacji,
- Skrypty dostępne w rozprawach doktorskich (Baffour-Awuah, 2009; Gerritse, 2016) oraz innych pracach Zwane i van der Heijden (2005),
- Programy oraz skrypty dostępne na stronach internetowych następujących autorów *Anna Chao* (http://chao.stat.nthu.edu.tw/wordpress/), *Francesco Bartolucci* (https://sites.google.com/site/bartstatistics/software),
- Procedury języka SAS – PROC GLM lub R – glm,
- STAN (Carpenter i inni, 2016) – język programowania służący głównie do szacowania modeli metodami bayesowskimi.

Spis pozostałego oprogramowania do estymacji wielkości populacji można znaleźć na stronie internetowej http://www.capturerecapture.co.uk/software.html.

[6] Opracowano na podstawie Zhang and Dunne (2017).

Podsumowanie

W raporcie dokonano przeglądu metod statystycznych, które można wykorzystać w procesie estymacji wielkości populacji trudnych do zbadania. Przykładem takiej zbiorowości mogą być cudzoziemcy czasowo przebywający w Polsce z uwzględnieniem nierejestrowanych imigrantów. Jak to zostało wzmiankowane we wstępie raportu 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. W podsumowaniu raportu sformułowane zostały zatem pewne wytyczne, których uwzględnienie w dalszych pracach stanowić będzie warunek niezbędny w procesie estymacji liczby cudzoziemców czasowo przebywających w Polsce wraz z niezarejestrowanymi imigrantami.

Założenia, które należy wziąć pod uwagę, aby poprawnie oszacować wielkość populacji zdefiniowanej na potrzeby pracy badawczej (część z nich jest uchylana w omawianych tu metodach) odnoszą się przede wszystkim do:

- sprawdzenia czy definicje populacji we wszystkich źródłach są takie same,
- określeniu czy populacja jest zamknięta – zakłada się bowiem, że w badanym okresie wielkość populacji jest stała,
- sprawdzeniu czy każda jednostka ma szansę być zapisana w każdym ze źródeł (przy czym prawdopodobieństwo nie musi być jednakowe),
- określeniu czy źródła danych są niezależne – w przypadku źródeł administracyjnych systemy powinny być niezależne (na przykład w sensie prawnym),
- określeniu czy każdą jednostkę będzie można zidentyfikować i połączyć między źródłami bez błędów (łączenie deterministyczne) – istnieją estymatory uwzględniające błędy łączenia wynikające z probabilistycznego łączenia rekordów.

Aby ponadto dokonać estymacji wielkości populacji trudnych do zbadania należy posiadać minimum dwa źródła danych, które pokrywają największą część populacji i spełniają założenie o niezależności. Każde z wybranych źródeł bowiem jedynie częściowo pokrywa badaną populację. Na potrzeby dalszych prac analitycznych należy podczas szacowania wielkości populacji cudzoziemców poszukać odpowiedzi na następujące pytania:

- czy spełnione są założenia metod typu capture-recapture – metody te są bowiem wrażliwe na niespełnienie założeń,
- czy możemy łączyć źródła danych po identyfikatorze (deterministyczne łączenie) czy z wykorzystaniem zmiennych pomocniczych (probabilistyczne łączenie rekordów),
- czy rozważane źródła zawierają błędy nadreprezentacji,
- czy dostępne są zmienne pomocnicze (na przykład wiek, płeć, województwo), które można wykorzystać w procesie łączenia jednostek pochodzących z różnych zbiorów jak i w procesie estymacji.

Wybór odpowiednich estymatorów i metod na potrzeby oszacowania wielkości populacji cudzoziemców czasowo przebywających w Polsce wraz z nierejestrowanymi imigrantami zależy zatem od spełnienia odpowiednich założeń i od znalezienia odpowiedzi na powyżej sformułowane pytania. Warunkowane będzie to również dostępnością odpowiednich danych statystycznych oraz ich jakością. Stanowić to będzie przedmiot rozważań w dalszych pracach, w których podjęta zostanie próba estymacji założonej do oszacowania w projekcie populacji cudzoziemców w Polsce.

Literatura

Agresti A. (2013), Categorical Data Analysis. Wiley.

Agresti A. (1994), Simple capture-recapture models permitting unequal catchability and variable sampling effort. Biometrics, 494–500.

Baffour-Awuah B. Estimation of population totals from imperfect census, survey and administrative records. PhD thesis, University of Southampton, 2009.

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

Bartolucci F., Forcina A. (2006), A class of latent marginal models for capture–recapture data with continuous covariates. Journal of the American Statistical Association, 101(474):786–794.

Böhning D., Ekkehart D., Ronny K., Schön D (2005), Mixture models for capture-recapture count data. Statistical Methods & Applications, 14(1):29–43.

Borg A., Sariyar M. (2016), RecordLinkage: Record Linkage in R, R package version 0.4-10.

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

Carpenter B., Gelman A., Hoffman M., Lee D., Goodrich B., Betancourt M., Brubaker M., Guo J., Li P., Riddell A. (2016), Stan: A probabilistic programming language. Journal of Statistical Software, 20:1–37.

Chao A., Tsay PK. (1998), A sample coverage approach to multiple-system estimation with application to census undercount. Journal of the American Statistical Association, 93(441):283–293.

Chao A., Tsay PK., Lin S.H., Shau W.Y., Chao D.Y. (2001), The applications of capture-recapture models to epidemiological data. Statistics in medicine, 20(20):3123–3157.

Chapman CJ. (1951), Some properties of the hypergeometric distribution with applications to zoological censuses. University of California Public Static, 1:131–160.

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

Cruyff M., van der Heijden P.G.M. (2008), Point and Interval Estimation of the Population Size Using a Zero-Truncated Negative Binomial Regression Model. Biometrical Journal, 50(6):1035–1050.

Deville J.D., Lavallée P. (2006), Indirect Sampling: The Foundations of the Generalized Weight Share Method. Survey Methodology, 32(2):165–176.

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

Ding Y., Fienberg S.E. (1994), Dual system estimation of Census undercount in the presence of matching error. Survey Methodology, 20(2):149–158.

Dorazio R.M., Royle A.J. (2003), Mixture models for estimating the size of a closed population when capture rates vary among individuals. Biometrics, 59(2):351–364.

Enamorado T., Fifield B., Imai K. (2017), fastLink: Fast Probabilistic Record Linkage with Missing Data. R package version 0.1.1.

Fellegi I.P., Sunter A.B. (1969), A theory for record linkage. Journal of the American Statistical Association, 64(328):1183–1210.

Fiske I., Chandler R. (2011), unmarked: An R Package for Fitting Hierarchical Models of Wildlife Occurrence and Abundance. Journal of Statistical Software, 43(10):1–23.

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

Harron K. Goldstein H., Dibben Ch. (2015), Methodological developments in data linkage. John Wiley & Sons.

Hsieh T.C. (2012), CARE1:Statistical package for population size estimation in capture-recapture models. R package version 1.1.0.

Knoke D., Burke P. J. (1980), Log-linear models, volume 20. Sage.

Lavallée P., Rivest L.p. (2012), Capture-Recapture Sampling and Indirect Sampling. Journal of official statistics, 28(1):1–27.

McClintock B. T. (2015), multimark: an R package for analysis of capture-recapture data consisting of multiple “noninvasive” marks. Ecology and Evolution.

Newcombe H.B., Kennedy J.,Axford S., James A. (1959), Automatic linkage of vital records. Science, 130(3381):954–959.

Rivest L.P., Baillargeon S. (2014), Rcapture: Loglinear Models for Capture-Recapture Experiments. R package version 1.4-2.

Roszka W. (2013), Statystyczna integracja danych w badaniach społeczno-ekonomicznych. PhD Thesis, Poznań University of Economics.

Stanghellini E., van der Heijden P.G.M. (2004), A Multiple-Record Systems Estimation Method that Takes Observed and Unobserved Heterogeneity into Account. Biometrics, 60(2):510–516.

Steorts R.C., Hall R., Fienberg S.E. (2016), A Bayesian approach to graphical record linkage and deduplication. Journal of the American Statistical Association, 111(516):1660–1672.

Thandrayen J., Wang Y. (2009), A latent variable regression model for capture–recapture data. Computational Statistics & Data Analysis, 53(7):2740–2746.

Van Der Heijden P.G.M., and Bustami R., Cruyff M.,Engbersen G., Van Houwelingen H.C. (2003), 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.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.

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, 2008.

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

Zhang L.C., Dunne J. (2017), Trimmed dual system estimation. In Böhning, Dankmar and Bunge, John and Heijden, P. G. M. van der, editors, Capture-recapture methods for the social and medical sciences, chapter 17, pages 237–257. CRC Press.

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.

Spis tablic

Tablica 1. Przypadek dwóch źródeł – tablica kontyngencji 2x2 3

Tablica 2. Przypadek trzech źródeł – tablica kontyngencji 2x2x2 4

Tablica 3. Przypadek czterech źródeł – inny sposób zapisu 4

Tablica 4. Przypadek dwóch źródeł i jednej zmiennej pomocniczej 5

Tablica 5. Przypadek czterech źródeł – inny sposób zapisu 18

Tablica 6. Przykładowe dane wykorzystywane w procesie modelowania 20

Tablica 7. Pokrycie między źródłami – PR, ER i CSR (w tysiącach) 25

Tablica 8. Informacje o łączeniu rejestrów 25

Tablica 9. Dane będące podstawą analizy wraz z oszacowaniami pochodzącymi z modelu 27