Statystyczna integracja rejestrów administracyjnych na potrzeby estymacji charakterystyk populacji cudzoziemców w Polsce

Urząd Statystyczny w Poznaniu
Uniwersytet Ekonomiczny w Poznaniu
Maciej Beręsewicz

Główny Urząd Statystyczny
Dorota Szałtys

MET2023

2023-07-03 Warszawa

Plan prezentacji

- Wprowadzenie do problemu
- Probabilistyczna integracja rejestrów administracyjnych
- Wyniki
- Wnioski i rekomendacje
- Literatura

Wprowadzenie

- Praca badawcza pt. Cudzoziemcy na krajowym rynku pracy w ujęciu regionalnym realizowana w ramach projektu “Wsparcie systemu monitorowania polityki spójność w perspektywie finansowej 2014-2020 oraz programowania i monitorowania polityki spójności po 2020 roku”.
- Praca metodologiczna nr 3.254 pt. Opracowanie metody szacunku zasobów cudzoziemców na krajowym rynku pracy w ujęciu regionalnym (NUTS3)
- Grant NCN OPUS 20 Statystyka cudzoziemców bez spisu powszechnego – jakość, integracja danych i estymacja (2020/39/B/HS4/00941).

Wprowadzenie

Na potrzeby projektu wykorzystano następujące bazy danych (numery zgodne z systemem ISODS):

- **PESEL** (2785) – Dane jednostkowe z rejestru PESEL, o osobach żyjących, gromadzone na podstawie ustawy z dnia 24 września 2010 r. o ewidencji ludności (Dz. U. z 2019 r. poz 1397),
- **ZUS** (2834) — Dane jednostkowe z konta ubezpieczonego i konta płatnika składek, o cudzoziemcach, którzy zostali zgłoszeni do ubezpieczenia społecznego lub zdrowotnego, oraz o płatnikach składek, którzy ich zgłosili do ubezpieczenia, gromadzone na podstawie art. 40 i art. 45 ustawy z dnia 13 października 1998 r. o systemie ubezpieczeń społecznych,
- **NFZ** (2808) – Dane jednostkowe z Centralnego Wykazu Ubezpieczonych, o osobach zarejestrowanych w Centralnym Wykazie Ubezpieczonych prowadzonym na podstawie ustawy z dnia 27 sierpnia 2004 r. o świadczeniach opieki zdrowotnej finansowanych ze środków publicznych (Dz. U. z 2019 r. poz. 1373, 1394, 1590, 1694 i 1726),
- **KRUS** (2793) – Dane dotyczące ubezpieczonych i płatników składek,
- **MF** (2869) – Dane jednostkowe o osobach fizycznych prowadzących samodzielnie działalność gospodarczą i nieprowadzących jej, gromadzone na podstawie ustawy z dnia 13 października 1995 r. o zasadach ewidencji i identyfikacji podatników i płatników.
- **UdSC** (2873) – Dane jednostkowe z krajowego zbioru rejestrów, ewidencji i wykazu w sprawach cudzoziemców, o cudzoziemcach, którzy posiadali ważny dokument, wydany w Rzeczypospolitej Polskiej, uprawniający do pobytu na jej terytorium,

Źródła danych

Tablica 1 – Liczba cudzoziemców według rejestrów zgodnie ze stanem na 31.12.2021

Rejestr	Liczba cudzoziemców
PESEL	2 004 765
ZUS	957 539
MF	1 513 129
KRUS	67 932
NFZ	2 034 434
UdSC	545 873

Łączenie danych

Proces łączenia danych przedstawiał się następująco:

- Ustalono, że głównym rejestrem jest PESEL i do niego dołączano kolejne rejestry,
- Deduplikacja rekordów w ramach rejestru oraz przypisanie nr PESEL gdzie to jest możliwe,
- Łączenie rejestrów po nr PESEL,
- Łączenie rejestrów na podstawie dokładnej zgodności: imienia, nazwiska, płci, daty urodzenia oraz kodu obywatelstwa,
- Wstępne łączenie danych na podstawie probabilistycznego łączenia rekordów (PLR) celem utworzenia próby uczącej i testowej do PLR z wykorzystaniem nadzorowanego uczenia maszynowego (m.in. metoda wektorów nośnych oraz sztuczne sieci neuronowe),
- Połączenie rejestrów z wykorzystaniem nadzorowanego uczenia maszynowego.

Łączenie deterministyczne

Tablica 2 – Liczba cudzoziemców dołączona do rejestru PESEL według typu łączenia: po etapie 1 (po identyfikatorze PESEL i po etapie 2 (po dokładnej zgodności imienia, nazwiska, płci, daty urodzenia i kodu obywatelstwa) oraz liczba rekordów niepołączonych (bez identyfikatorów)

Rejestr	Etap 1	Etap 2	Bez identyfikatorów
KRUS	4 674	18 317	42 693
MF	1 043 769	1 132 840	352 088
NFZ	1 987 884	1 987 891	42 524
ZUS	624 113	760 765	117 926
Wszystkie	1 988 650	1 989 390	--

Łączenie probabilistyczne

- Różne rejestry, różne zapisy

Tablica 3 – Przykładowe dane z rejestrów

Imie	Imie2	Nazwisko
Miguel	Luis	Pereira-Tinoco
Miguel Luis		Pereira Tinoco
Miguel		Pereira-Tinoco
Thi	Huyen Trang	Dao
Dao Thi Huyen Trang		
Thi		Dao

Procedura probabilistycznego łączenia rekordów

- Ujednolicenie danych między źródłami
- Redukcja liczby porównań z zastosowaniem blokowania
- Porównanie par rekordów w ramach bloków
- Uczenie nadzorowane lub nienadzorowane

Ujednolicanie

- Przykład:
- [„THI”, „HUYEN TRANG”, „DAO”] => [„THI”, „HUYEN”, „TRANG”, „DAO”] => [„DAO”, „HUYEN”, „THI”, „TRANG”] => [„DAO HUYEN THI TRANG”]

imie	imie2	nazwisko	dane_imiona2	do_qgrams
MIGUEL	LUIS	PEREIRA-TINOCO	LUIS MIGUEL PEREIRA TINOCO	LUISMIGUELPEREIRATINOCO198911031620
MIGUEL LUÍS		PEREIRA TINOCO	LUIS MIGUEL PEREIRA TINOCO	LUISMIGUELPEREIRATINOCO198911031620
MIGUEL		PEREIRA-TINOCO	MIGUEL PEREIRA TINOCO	MIGUELPEREIRATINOCO198911031620

imie	imie2	nazwisko	dane_imiona2	do_qgrams
THI	HUYEN TRANG	DAO	DAO HUYEN THI TRANG	DAOHUYENTHITRANG199412262704
DAO THI HUYEN TRANG			DAO HUYEN THI TRANG	DAOHUYENTHITRANG199412262704
THI		DAO	DAO THI	DAOTHI199412262704

Blokowanie rekordów

- HNSW – Hierarchical Navigable Small Words.
- Wykorzystano pakiet *RcppHNSW* w R, który zawiera bibliotekę GitHub - nmslib/hnswlib: Header-only C++/python library for fast approximate nearest neighbors
- Bardzo szybka, duże możliwości.
- Jeżeli mamy *n* jednostek to liczba porównań wynosi *n(n-1)/2*.

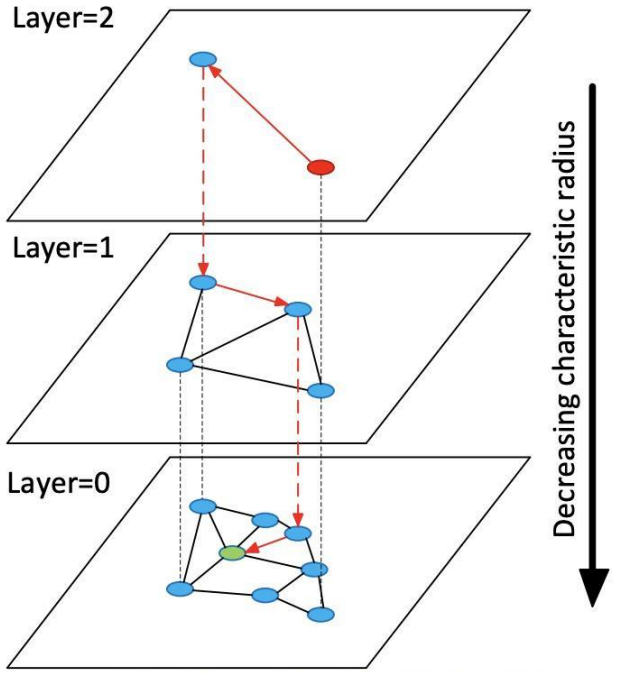

Fig. 1. Illustration of the Hierarchical NSW idea. The search starts from an element from the top layer (shown red). Red arrows show direction of the greedy algorithm from the entry point to the query (shown green).

Blokowanie

- Wstępne blokowanie rekordów wg roku urodzenia (do 1945, 1946, ...) oraz płci.
- W ramach subpopulacji tworzona jest macierz bigramów na podstawie sklejonych imion, nazwiska, daty urodzenia i kodu obywatelstwa.

```
> stringdist::qgrams("DAOHUYENTHITRANG199412262704", q=2)
   DA OH UY YE TH TR RA HU NT IT NG EN HI G1 AO AN 19 99 94 41 62 70 22 26 27 12 04
V1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
```

- Następnie dokonano wyszukiwania 2 najbliższych sąsiadów.

```
> hnsw_test <-  hnsw_knn(X = qgrams, k = 2, verbose = T, distance = "cosine", n_threads = 2, ef = 200, M = 25,
  ef_construction = 200)
```

- Na końcu wykorzystuję paczkę *igraph* do grupowania najbliższych sąsiadów.

Blokowanie - wynik

- Na podstawie rejestrów utworzono zbiór zawierający ponad 6 mln rekordów, które mogły zawierać 0 lub więcej błędów w zmiennych takich jak imiona, nazwiska, daty urodzeń czy kody obywatelstw.
- Blokowanie było bardzo udane:
  - Jednostki w tym samym bloku – ponad 2 mln jednostek
  - Jednostki w dwóch blokach – 22 jednostki
  - Jednostki w trzech blokach – 1182 jednostki (ale uwaga! W ten sposób znaleziono błędy w ponad 300 numerach PESEL)

Uczenie maszynowe

- Dane do nadzorowanego uczenia maszynowego przygotowano z pakietem reclin2 w R
- Zastosowano 10-krotną walidację krzyżową z wykorzystaniem nnet::nnet i e1071::tune

Tablica 4 – Macierz klasyfikacji dla wybranego algorytmu klasyfikacji

Znane / Przewidywane	Różne	Te same
Zbiór treningowy		
Różne	137 778 (99,89)	147 (0,11)
Te same	120 (0,09)	137 755 (99,91)
Zbiór testowy		
Różne	58 943 (99,88)	70 (0,12)
Te same	57 (0,10)	58 956 (99,90)
Zbiór testowy (przynajmniej jedna różnica)		
Różne	58 943 (99,94)	37 (0,06)
Te same	57 (2,33)	2 395 (97,67)

Łączenie deterministyczne

Tablica 5 – Liczba cudzoziemców dołączona do rejestru PESEL według typu łączenia: po etapie 1 (po identyfikatorze PESEL, po etapie 2 (po dokładnej zgodności imienia, nazwiska, płci, daty urodzenia i kodu obywatelstwa) oraz po etapie 3 (probabilistyczna integracja danych)

Rejestr	Etap 1	Etap 2	Etap 3
KRUS	4 674	18 317	19 148
MF	1 043 769	1 132 840	1 149 721
NFZ	1 987 884	1 987 891	1 959 426
ZUS	624 113	760 765	763 548
Wszystkie	1 988 650	1 989 390	1 989 803

Wyniki

Tablica 6 – Liczba cudzoziemców w wieku 18+ według liczby źródeł zgodnie ze stanem na 31.12.2021

Liczba źródeł	Liczba	Odsetek
1	9 699	0,5
2	501 581	26,8
3	795 365	42,5
4	563 776	30,1
5	1 505	0,1

Tablica 7 – Liczba cudzoziemców w wieku 18+ według regionu obywatelstwa zgodnie ze stanem na 31.12.2021

Kontytent	Liczba	Odsetek
Europa	1 598 080	85,8
Azja	201 708	10,8
Afryka	22 288	1,2
Ameryka Płn i Płd	21 587	1,2
Oceania	1 146	0,1
Nieustalone	21 587	0,9

Wyniki

Tablica 8 – Liczba cudzoziemców w wieku 18+ według kraju obywatelstwa zgodnie ze stanem na 31.12.2021

Kraj	Liczba	Odsetek
Ukraina	1 223 476	65,7
Białoruś	124 556	6,7
Niemcy	43 249	2,3
Gruzja	39 895	2,1
Mołdawia	37 145	2,0
Rosja	36 045	1,9
Turcja	24 756	1,3
Indie	24 487	1,3
Chiny	17 779	1,0
Wietnam	16 511	0,9

Obywatele Ukrainy

Tablica 9 – Liczba obywateli Ukrainy w wieku 18+ według grup wieku i płci zgodnie ze stanem na 31.12.2021

Grupa wieku	Mężczyźni [N]	Kobiety [N]	Mężczyźni [%]	Kobiety [%]
[18, 25)	108 428	86 851	14,87	17,57
[25, 35)	245 706	132 215	33,69	26,75
[35, 45)	199 616	121 369	27,37	24,56
[45, 55)	131 181	106 853	17,99	21,62
[55, 65)	40 228	39 869	5,52	8,07
65+	4 108	7 052	0,56	1,43

Osoby / rekordy niepołączone

Tablica 10 - Liczba rekordów niepołączonych, po deduplikacji

Rejestr	Początkowa	Po integracji z PESELem	Po deduplikacji
KRUS	42 693	40 724	28 904
MF	352 088	297 067	272 350
NFZ	42 524	42 524	29 876
ZUS	117 926	85 962	74 142

Podsumowanie

- Dzięki algorytmowi HNSW możliwe było zredukowanie liczby porównań i przygotowanie zbioru do probabilistycznej deduplikacji rekordów.
- Probabilistyczna integracja danych umożliwia łączenie danych bez identyfikatorów, co może rozwiązać pewne problemy w zakresie badania populacji cudzoziemców w Polsce.
- Wypracowany w pracy metodologicznej algorytm klasyfikacji jest stosowany do deduplikacji danych w ramach prac zespołu ds. UKR.

Literatura

- Beręsewicz, M., Gudaszewski, G., & Szymkowiak, M. (2019). Estymacja liczby cudzoziemców w Polsce z wykorzystaniem metody capture-recapture. *Wiadomości Statystyczne. The Polish Statistician*, *64*(10), 7–35.
- Brittain, S., & Böhning, D. (2009). Estimators in capture–recapture studies with two sources. *AStA Advances in Statistical Analysis*, *93*(1), 23–47.
- Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, 783–791.
- Chao, A. (1989). Estimating population size for sparse data in capture-recapture experiments. *Biometrics*, 427–438.
- Chao, A., Chu, W., & Hsu, C.-H. (2000). Capture–recapture when time and behavioral response affect capture probabilities. *Biometrics*, *56*(2), 427–433.
- Chatterjee, K., & Bhuyan, P. (2017). On the estimation of population size from a post-stratified two-sample capture–recapture data under dependence. *Journal of Statistical Computation and Simulation*, *90*, 819–838.
- Chatterjee, K., & Bhuyan, P. (2019). On the estimation of population size from a dependent triple-record system. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *182*(4), 1487–1501.
- Chatterjee, K., & Mukherjee, D. (2021). On the estimation of population size under dependent dual-record system: an adjusted profile-likelihood approach. *Journal of Statistical Computation and Simulation*, *91*(13), 2740–2763. https://doi.org/10.1080/00949655.2021.1908284
- Chatterjee, K., & Mukherjee, D. (2020). Identifying the direction of behavioral dependence in two-sample capture-recapture study. *Journal of Official Statistics*, *36*(1), 25–48.
- Di Cecco, D., Di Zio, M., Filipponi, D., & Rocchetti, I. (2018). Population size estimation using multiple incomplete lists with overcoverage. *Journal of Official Statistics*, *34*(2), 557–572.
- Di Consiglio, L., & Tuoto, T. (2015). Coverage evaluation on probabilistically linked data. *Journal of Official Statistics*, *31*(3), 415429.

Literatura

- Di Consiglio, L., Tuoto, T., & Zhang, L.-C. (2019). Capture-recapture methods in the presence of linkage errors. In *Analysis of integrated data* (pp. 39–71). Chapman; Hall/CRC.
- Ding, Y., & Fienberg, S. (1994). Dual system estimation of census undercount in the presence of matching error. *Survey Methodology*, *20*(2), 149158.
- 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., Heijden, P. G. van der, & 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.
- 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.
- Nour, E.-S. (1982). On the estimation of the total number of vital events with data from dual collection systems. *Journal of the Royal Statistical Society: Series A (General)*, *145*(1), 106–116. https://doi.org/https://doi.org/10.2307/2981424
- Sariyar, M., & Borg, A. (2010). The RecordLinkage Package: Detecting Errors in Data. *The R Journal*, *2*(2), 61–67. https://doi.org/10.32614/RJ-2010-017
- Wolter, K. M. (1986). Some coverage error models for census data. *Journal of the American Statistical Association*, *81*(394), 337–346.
- Yee, T. W., Stoklosa, J., & Huggins, R. M. (2015). The **VGAM** Package for Capture-Recapture Data Using the Conditional Likelihood. *Journal of Statistical Software*, *65*(5). https://doi.org/10.18637/jss.v065.i05
- 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. *Capture-Recapture Methods for the Social and Medical Sciences*, 237–257.

Główny Urząd Statystyczny
Dorota Szałtys

Urząd Statystyczny w Poznaniu

Uniwersytet Ekonomiczny w Poznaniu
Maciej Beręsewicz

stat.gov.pl

2023-07-03 Warszawa