Estymacja długości pobytu cudzoziemców w Polsce z wykorzystaniem mobile big data

Maciej Beręsewicz
Uniwersytet Ekonomiczny w Poznaniu
Urząd Statystyczny w Poznaniu

08.11.2022

POZNAŃ UNIVERSITY
OF ECONOMICS
AND BUSINESS

Plan prezentacji

1. Wprowadzenie
2. Metodyka
3. Wybrane wyniki
4. Podsumowanie

Plan prezentacji

1. Wprowadzenie
2. Metodyka
3. Wybrane wyniki
4. Podsumowanie

Wprowadzenie

- Podstawowym źródłem informacji o długości pobytu cudzoziemców w Polsce są dane z rejestrów:
  - Urzędu ds. Cudzoziemców (m.in. pozwolenia na pobyt, karty pobytu),
  - Ministerstwa Spraw Zagranicznych (m.in. okres przyznania wizy),
  - Ministerstwa Rodziny i Polityki Społecznej (min. pozwolenia na pracę, praca sezonowa),
  - Zakładu Ubezpieczeń Społecznych (m.in. okres zgłoszenia do ubezpieczenia zdrowotnego).
- Wszystkie powyższe źródła oparte są o dokumenty i należy traktować je jako pewne przybliżenie prawdziwej długości pobytu (występuje błąd pomiaru).

Alternatywne źródła informacji

- Coraz częściej w zakresie badania migracji wykorzystuje się dane z big data (np. sieci komórkowe, systemy reklamowe).
- Z jednej strony możemy "w miarę" precyzyjnie oszacować długość pobytu na podstawie aktywności smartfonów.
- Z drugiej strony populacja smartfonów jest zmienna w czasie, a informacje społeczno-demograficzne są pozyskiwanie w wyniku algorytmów klasyfikacji.
- W niniejszej pracy wykorzystane zostaną dane zbierane przez firmę Selectivv na podstawie systemów reklamowych.

Cel prezentacji

Celem prezentacji jest estymacja długości pobytu cudzoziemców uwzględniając błąd pomiaru w zmiennych społeczno-demograficznych w okresie 2018Q1-2021Q1 w Polsce.

Badanie finansowane z grantu NCN OPUS 20 pt. *Statystyka cudzoziemców bez spisu powszechnego – jakość, integracja danych i estymacja* (2020/39/B/HS4/00941).

Plan prezentacji

1. Wprowadzenie
2. Metodyka
3. Wybrane wyniki
4. Podsumowanie

Źródła danych

- Selectivv zbiera informacje z ponad 45 mln smartfonów w Polsce (liczba aktywnych kart sim w ich bazie). Baza po deduplikacji zawiera blisko 33 mln "użytkowników smartfonów".
- Selectivv wykorzystuje systemy reklamowe w aplikacjach mobilnych i stronach internetowych, które przypisują unikalny identyfikator smartfonowi (GAID,IDFA). "Życie" takiego identyfikatora jest zdecydowanie dłuższe niż cookies.
- Następnie poszczególne smartfony przypisywane są do jednego użytkownika na podstawie współwystępowania w ciągu dnia i nocy (geolokalizacja) oraz logowań do sieci wi-fi.
- Problemy: **błąd pokrycia (duplikaty)** (*zmiana smartfonów w ciągu roku*, *resetowanie GAID / IDFA*, *blokowanie śledzenia przez aplikacje*), **błąd pomiaru** (*klasyfikacja zmiennych społeczno-demograficznych*).

Źródła danych – algorytmy klasyfikacji

- Kraj obywatelstwa – ustalany jest na podstawie języka systemowego, czasu przebywania na terenie Polski oraz zmian kart SIM na operatorów danego kraju (np. Ukrainy).
- Płeć czy wiek – ustalany na podstawie aktywności użytkowników m.in. wykorzystywanych aplikacji, stron internetowych czy geolokalizacji.
- Czas pobytu – określony na podstawie geolokalizacji smartfonów. Określono 3 grupy: *od 30 dni do 3 miesięcy*, *od 3 do 12 miesięcy* i *powyżej 12 miesięcy*.

Próba walidacyjna

- Aby ocenić błąd klasyfikacji zastosowano próbę walidacyjną.
- Jest to podejście często stosowane w literaturze do oceny jakości rejestrów administracyjnych (ang. *two-phase studies*).
- Przygotowano próbę walidacyjną opartą o stratyfikowaną próbę losową realizowaną przez systemy reklamowe. Do stratyfikacji wykorzystano obywatelstwo określone przez Selectivv.
- Realizacja blisko 1%, finalna próba 501 respondentów.
- W kwestionariuszu pytano o obywatelstwo, płeć, wiek, czas pobytu oraz informacje o liczbie smartfonów.
- Deklaracje zestawiono ze zmiennymi określonymi przez algorytmy Seletivv.

Problemy badawcze

- Jak oszacować błąd klasyfikatora na podstawie próby walidacyjnej?
- Jak skorygować błąd pomiaru?
- Jak oszacować długość pobytu cudzoziemców w Polsce?

Oznaczenia

- n to wielkość próby walidacyjnej, a m to wielkość zrealizowanej próby walidacyjnej,
- $h^* = 1, ..., H^*$ oznacza warstwę określoną przez kombinację obywatelstwa, płci oraz wieku wyznaczoną na podstawie algorytmów klasyfikacyjnych (występuje błąd pomiaru; $H^* = 24$),
- $h = 1, ..., H$ oznacza prawdziwą warstwę określoną przez te same zmienne co h^* uzyskaną na podstawie próby walidacyjnej ($H = 24$),
- x_h i x_{h^*} oznaczają wartości cechy x w warstwie h oraz h^*,
- y_h oraz y_{h^*} oznaczają czas przebywania (30 dni do 3 miesięcy, 3-12 miesięcy oraz 12+ miesięcy),

Oznaczenia cd.

- $\mathbf{P}_{x,h^*,h}$ to macierz prawdopodobieństw utworzoną przez zestawienie x_h i x_{h^*}. Elementy macierzy $\mathbf{P}$ określone są przez $\Pr(x_h|x_{h^*})$.
- $\Gamma_{y,x,h}$ to macierz prawdopodobieństw związanych z czasem pobytu (y). Elementy macierzy Γ są określone przez $\Pr(y_h|x_h)$.
- Estymację elementów macierzy $\mathbf{P}$ oraz Γ dokonano metodą bootstrap na podstawie próby walidacyjnej.

Korekta błędu pomiaru

Błąd pomiaru wynikający z klasyfikacji korygujemy jest według dwustopniowej procedury:

1. $m_{x,h} \sim \text{Multinomial}(m_{x,h*}, \mathbf{P}_{x,h^*,h})$ – korygujemy liczebność warstwy h^* uwzględniając macierz prawdopodobieństwa $\mathbf{P}$,
2. $m_{y,h} \sim \text{Multinomial}(m_{x,h}, \Gamma_{y,x,h})$ – estymujemy czas pobytu uwzględniając skorygowane zmienne społeczno-demograficzne.

Do estymacji wariancji wykorzystano metodę bootstrap.

Plan prezentacji

1. Wprowadzenie
2. Metodyka
3. Wybrane wyniki
4. Podsumowanie

Próba walidacyjna – wyniki

Table 1: Poprawność klasyfikacji danych Selectivv na podstawie próby walidacyjnej

Zmienna	Poziomy	Trafność	Wielkość próby
Kraj	Białoruś	96.0	101
	Polska	96.8	247
	Ukraina	93.5	153
Płeć	Kobieta	87.3	221
	Mężczyzna	89.3	280
Wiek	18-24	88.1	236
	25-29	84.8	151
	30-39	92.2	64
	40+	96.0	50
Pobyt	3m	61.4	44
	3m-12m	78.6	112
	12m+	97.9	97

Próba walidacyjna – długość pobytu

Table 2: Długość pobytu według Selectivv (wiersze), a deklaracje w próbie audytowej (kolumny)

	3m	3m-12m	12m+
3m	27 (61.4)	16 (36.4)	1 (2.3)
3m-12m	0 (0.0)	88 (78.6)	24 (21.4)
12m+	1 (1.0)	1 (1.0)	95 (98.0)

Próba walidacyjna – wyniki

Table 3: Zależność między zmiennymi demograficznymi, a czasem pobytu

Zmienna	χ^2	df	p-value	V Cramer'a
Przed korektą błąd klasyfikacji				
Kraj	0.32	2.00	0.85	0.04
Wiek	4.13	6.00	0.66	0.09
Płeć	10.17	2.00	0.01	0.21
Po korekcje błędu klasyfikacji				
Kraj	1.00	2.00	0.61	0.06
Wiek	3.78	6.00	0.71	0.09
Płeć	4.95	2.00	0.08	0.14

Błąd pomiaru i jego korekta

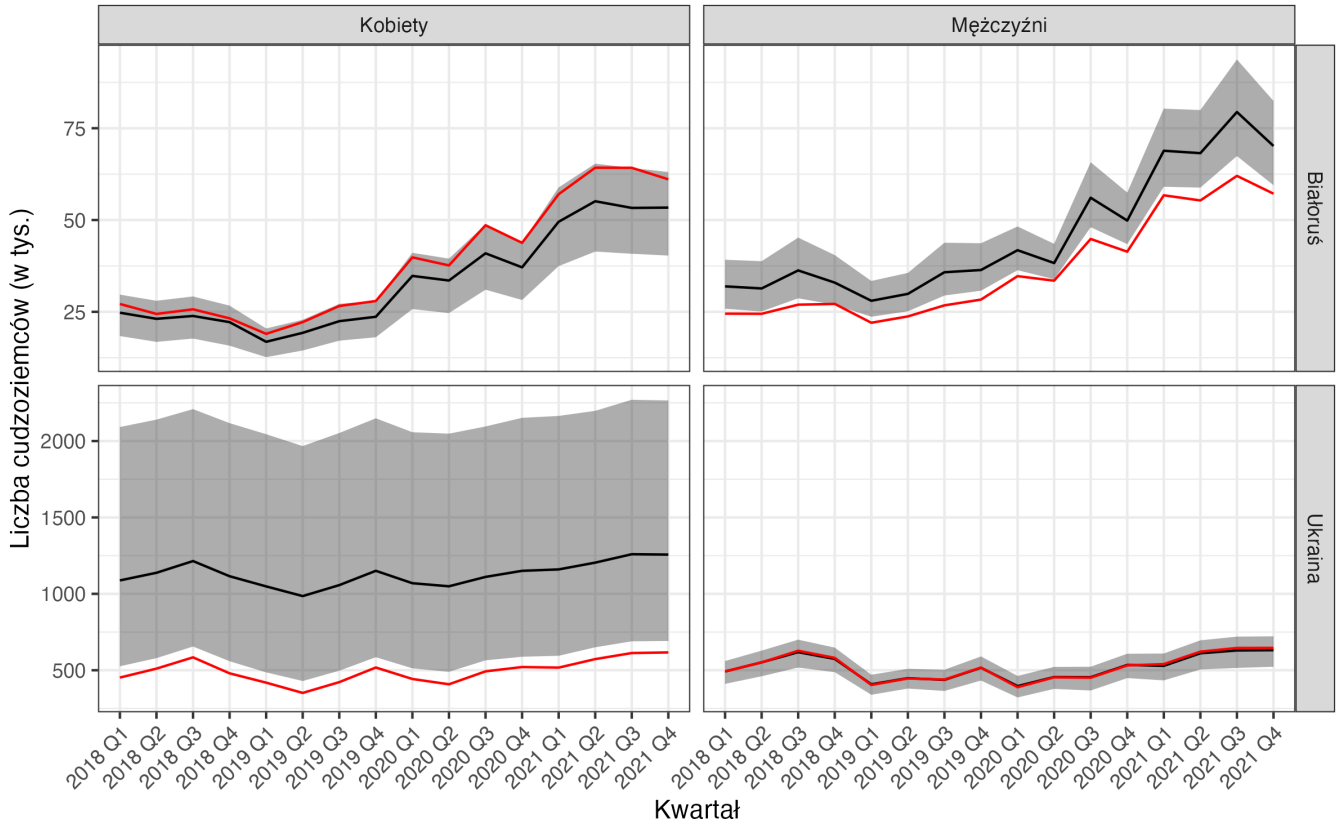

Figure 1: Korekta błędu pomiaru według kraju i płci. Kolor czerwony oznacza liczbę cudzoziemców przed korektą.

Błąd pomiaru i jego korekta

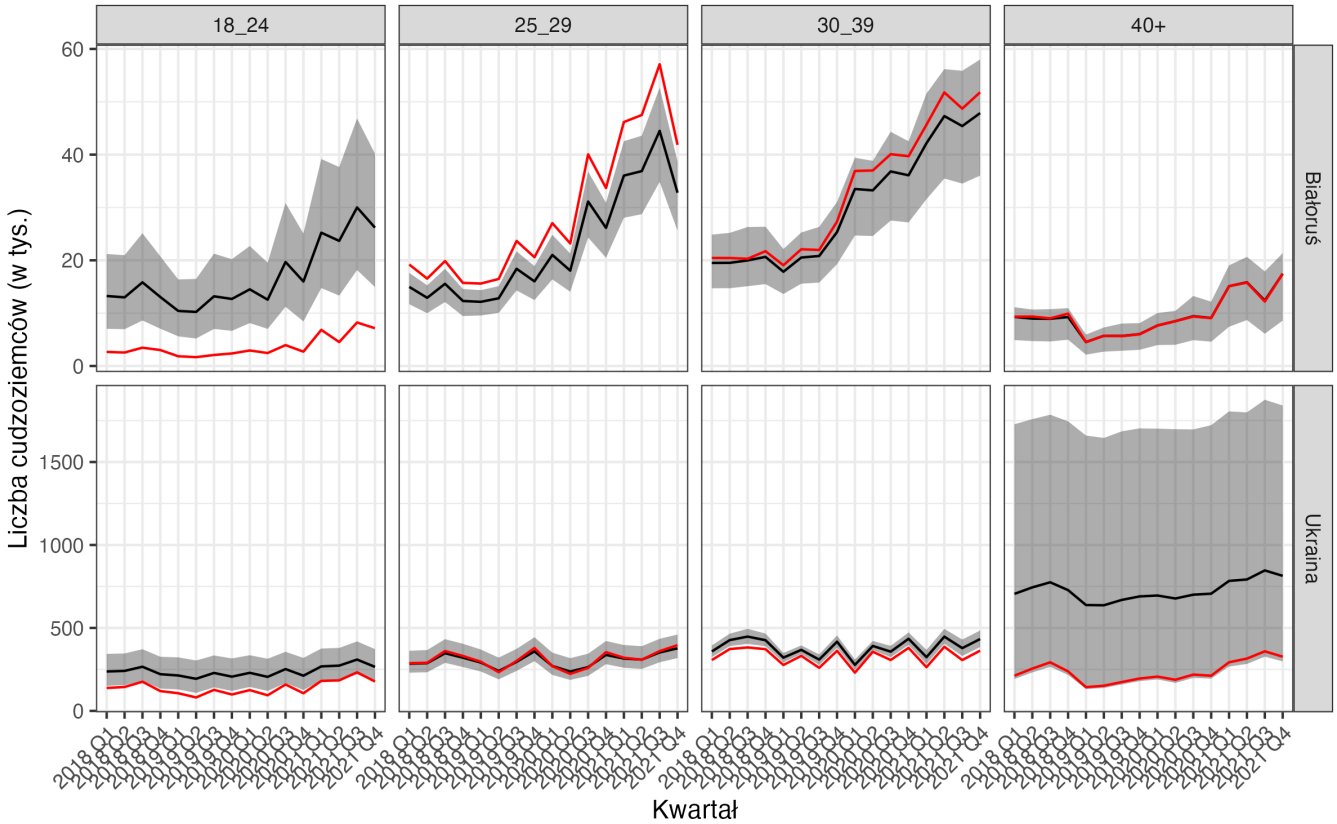

Figure 2: Korekta błędu pomiaru według kraju i wieku. Kolor czerwony oznacza liczbę cudzoziemców przed korektą.

Błąd pomiaru i jego korekta

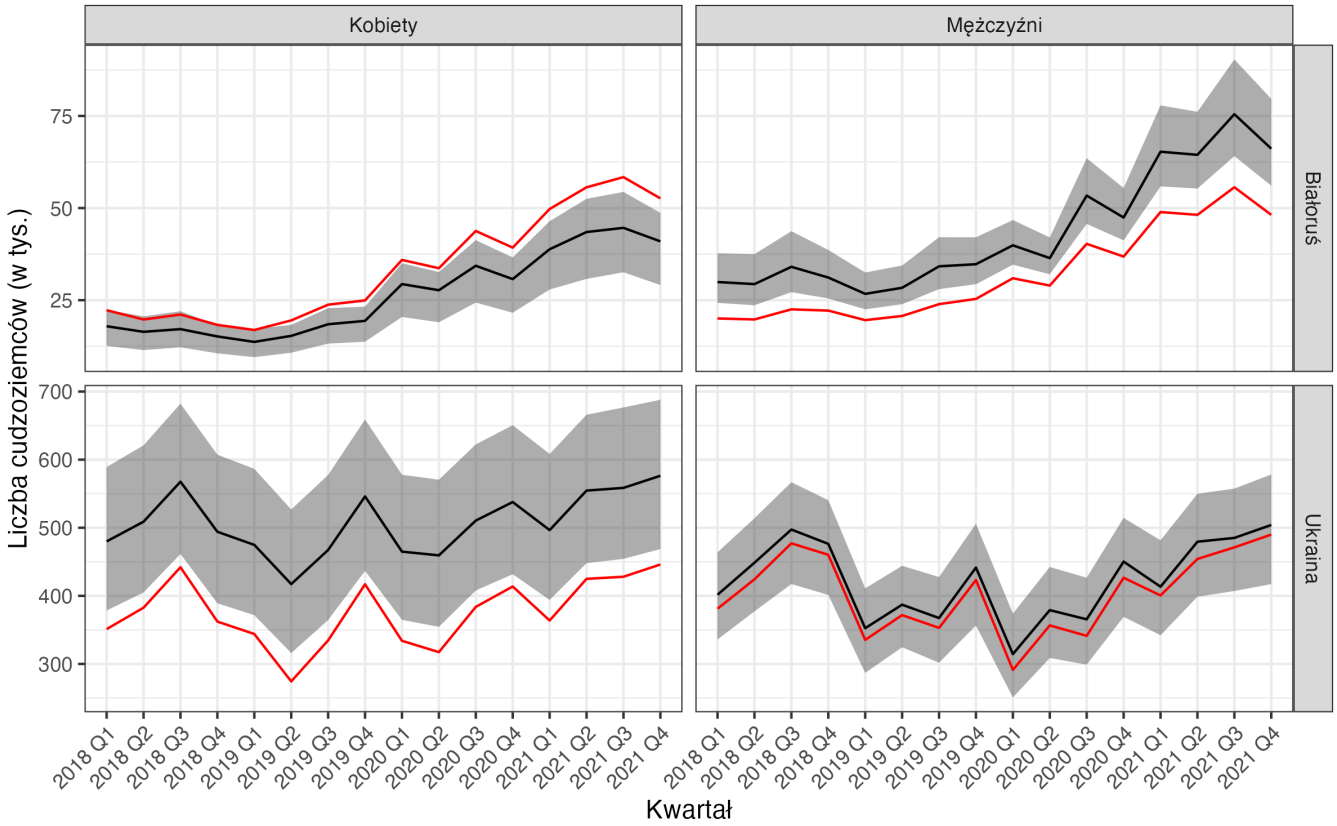

Figure 3: Korekta błędu pomiaru według kraju i płci (bez 40+). Kolor czerwony oznacza liczbę cudzoziemców przed korektą.

Czas pobytu – wyniki estymacji

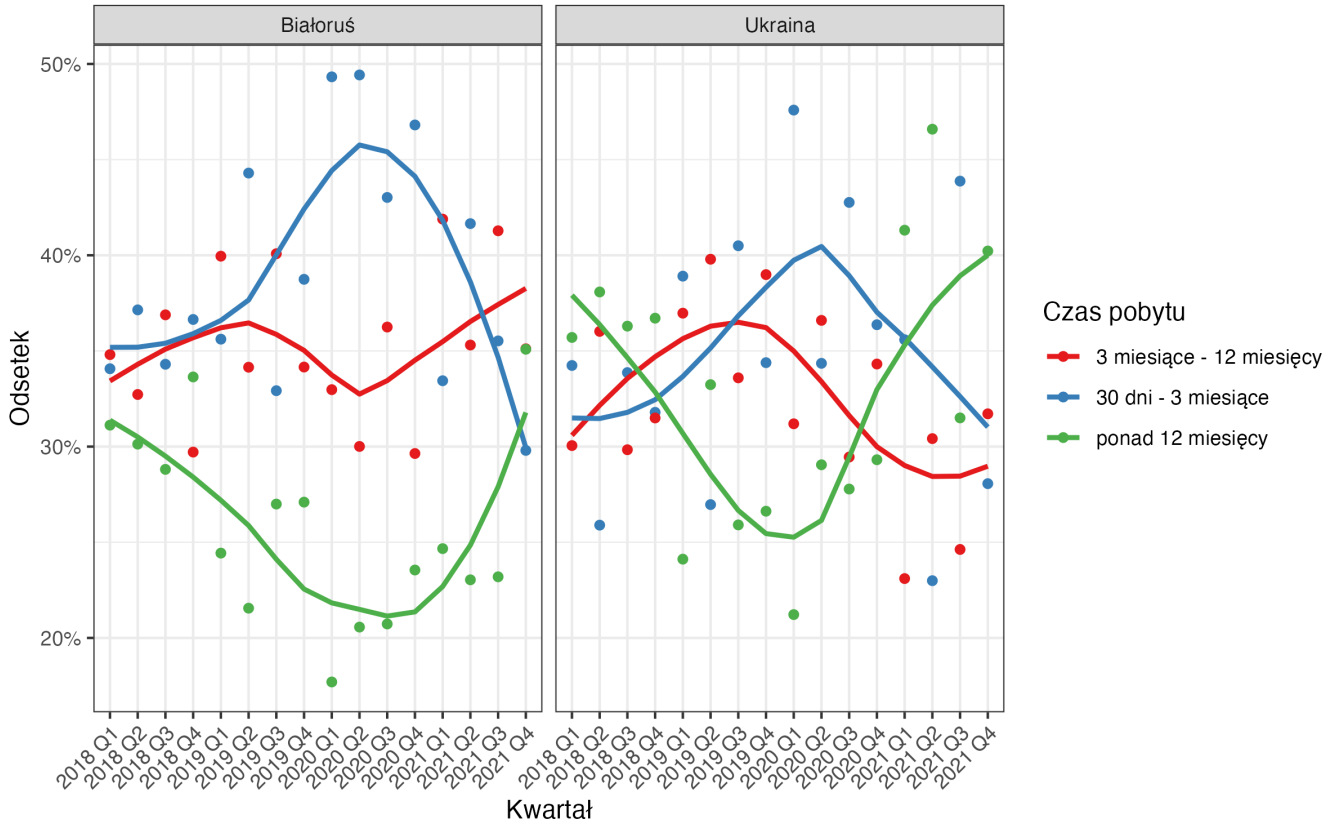

Figure 4: Czas pobytu według kraju pochodzenia wyznaczony bez grupy wiekowej 40+

Czas pobytu – porównanie

Table 4: Porównanie czasu pobytu po uwzględnieniu błędu pomiaru w danych Selectivv za 2021Q4 oraz danych z ZUS za 2022Q1

Kraj	Okres	Selectivv	ZUS
Białoruś	do 3 miesięcy	29.8	24.9
	od 3 miesięcy do 1 roku	35.1	29.5
	powyżej 1 roku	35.1	45.6
Ukraina	do 3 miesięcy	28.1	26.5
	od 3 miesięcy do 1 roku	31.7	29.4
	powyżej 1 roku	40.2	44.1

Plan prezentacji

1. Wprowadzenie
2. Metodyka
3. Wybrane wyniki
4. Podsumowanie

Podsumowanie

- W prezentacji skupiono się na problemie estymacji czasu pobytu z wykorzystaniem źródeł big data.
- Podjęto próbę skorygowania błędu pomiaru w zmiennych społeczno-demograficznych z wykorzystaniem próby walidacyjnej.
- W związku z brakiem informacji o wielkości populacji cudzoziemców nie można było skorygować błędu pokrycia.

Literatura (wybrana)

- 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.
- Schenkel, J. F., & Zhang, L. C. (2022). Adjusting misclassification using a second classifier with an external validation sample. Journal of the Royal Statistical Society: Series A (Statistics in Society).
- Biemer, P. P., & Bushery, J. M. (2000). On the validity of markov latent class analysis for estimating classification error in labor force data. Survey Methodology, 26(2), 139-152.
- Chen, T. T. (1989). A review of methods for misclassified categorical data in epidemiology. Statistics in Medicine, 8(9), 1095-1106.
- Pankowska, P., Bakker, B., Oberski, D. L., & Pavlopoulos, D. (2018). Reconciliation of inconsistent data sources by correction for measurement error: The feasibility of parameter re-use. Statistical Journal of the IAOS, 34(3), 317-329.
- Pavlopoulos, D., & Vermunt, J. K. (2015). Measuring temporary employment. Do survey or register data tell the truth. Survey Methodology, 41(1), 197-214.