Estimation of the number of irregular foreigners in Poland using non-linear count regression models

dr Maciej Beręsewicz
Department of Statistics, Poznań University of Economics and Business, Poland
Centre for the Methodology of Population Studies, Statistical Office in Poznań, Poland

BERENZ / ncn-foreigners / ojalab
/ 𝕏 @mberesewicz
mberesewicz.bsky.social

COMPAS, University of Oxford, 2025.03.28

Poznań University
of Economics
and Business

Outline

Introduction: whoami

Introduction: whoami

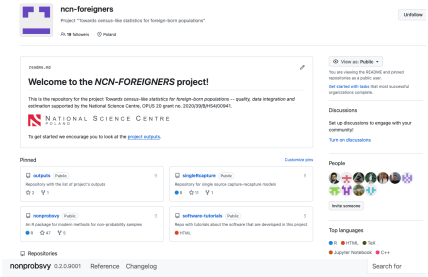
- Professor at the Department of Statistics at Poznań University of Economics and Business (PUEB) in Poland
- The head of the Centre for the Methodology of Population Studies at Statistical Office in Poznań (regional office).
- Main focus is on survey sampling and methodology, including survey sampling, non-response, calibration, non-probability samples, and population size estimation.
- The PI of The project *Towards census-like statistics for foreign-born populations – quality, data integration and estimation* (NCN 2020/39/B/HS4/00941) – https://github.com/ncn-foreigners.
- The co-PI of the project `OJALAB: Online job advertisements to study skill demand and job search patterns` (2024/53/B/HS4/01580) – https://github.com/OJALAB (e.g. we have developed a multilingual (24 languages) job ads classifier).
- I am an R enthusiast and have organised several R conferences in Poland and Europe. I have also developed several packages: `nonprobsvy`, `singleRcapture`, `jointCalib` and `blocking`.

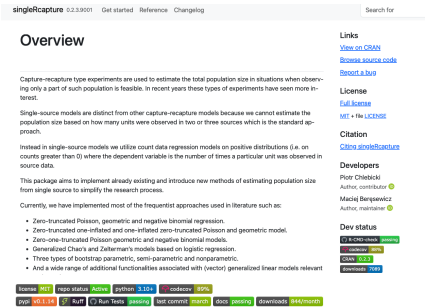
Introduction: whoami

Figure 1: Poznań University of Economics and Business

Figure 2: Statistical Office in Poznań

Introduction: whoami





Acknowledgements

This study is based on the working paper **Estimation of the number of irregular foreigners in Poland using non-linear count regression models** by Beręsewicz & Pawlukiewicz (2020) [arXiv:2008.09407]

Since this paper is being considerably revised, the model and the results may change.

Project: *Towards census-like statistics for foreign-born populations – quality, data integration and estimation* (NCN 2020/39/B/HS4/00941).

Motivation

- Irregular (undocumented) migration is hard to measure as the underlying population is hard-to-reach.
- Several approaches have been proposed in the literature, which are based for instance on residual, single-source capture-recapture or multiple estimation system methods.
- Majority of methods assumes access and integration of data at the unit-level.
- The proposed approach requires access to aggregated data and is based on a functional form and certain assumptions that can be verified using available data.

Outline

The model

- The original model is based on Prof. Li-Chun Zhang's (University of Southampton, University of Oslo, Statistics Netherlands) working paper entitled **Developing methods for determining the number of unauthorized foreigners in Norway** (2008).
- The author proposes a model that requires only three types of variables:
  1. the number of apprehended irregular foreigners (denoted as m),
  2. the number of foreigners who faced criminal charges (denoted as n),
  3. the number of foreigners registered in the central population register (denoted as N).

Assumptions

- Let M_t be the size of the population of unauthorized resident at time t (e.g. end of the year) – the random variable.
- Let N_t be the be the size of the known reference (proxy) population at the same time t – the fixed, known covariate.
- The target parameter is the theoretical size of irregular residents, which is defined as the conditional expectation of M_t given N_t with respect to $f(M_t|N_t)$ denoted by

$$\xi_t = \mathbb{E}(M_t|N_t).$$

Assumptions

- As Zhang (2008) notes, the theoretical size is defined as the conditional expectation of the random variable, which makes it possible to get rid of the spurious variation as long as the reference population size is held fixed.
- The purpose of introducing N_t is two-fold:
  1. it serves as an explanatory variable for the irregular size M_t,
  2. it provides an interpretation of the irregular size M_t in analogy to N_t.
- In this way, the theoretical size is a stable measure of the target variable as variation in M_t is linked to that of N_t.

Assumptions

- Let m_{it} be the observed number of irregular foreigners from country i (this may also indicate more detailed populations e.g. sex-age group for a given country).
- Let n_{it} be the observed number of (legally staying) foreigners from country i.
- Let p_{it} be the probability for an irregular resident to be observed in administrative data (say Border Guards).
- Let

$$m_{it} \sim \text{Poisson}(\lambda_{it})$$

- Let $\lambda_{it} = \mu_{it} u_{it}$, where $\mu_i = \mathbb{E}\left(M_{it} p_{it} \mid n_{it}, N_{it}\right) = \mathbb{E}\left(M_{it} \mid N_{it}\right) \cdot \mathbb{E}\left(p_{it} \mid M_{it}, n_{it}, N_{it}\right)$

Assumptions

- The final model consists of the following set of equations

$$\begin{aligned} \xi_{it} &= \mathbb{E}\left(M_{it} \mid N_{it}\right) = N_{it}^{\alpha}, \\ \omega_i &= \mathbb{E}\left(p_{it} \mid M_{it}, n_{it}, N_{it}\right) = \mathbb{E}\left(p_{it} \mid n_{it}, N_{it}\right) = \left(\frac{n_{it}}{N_{it}}\right)^{\beta}, \\ u_{it} &\sim \mathrm{Gamma}(1, \phi), \end{aligned} \tag{1}$$

- From which we can derive the following relationship for μ_{it}

$$\mu_i = N_i^{\alpha} \left(\frac{n_i}{N_i}\right)^{\beta} \tag{2}$$

The target quantity

- We are interested in the target parameter describing the number of irregular residents. Given the above model, the target parameter is defined as

$$\xi = \sum_{i=1}^{C} E\left(M_i | N_i\right) = \sum_{i=1}^{C} N_i^{\alpha}, \tag{3}$$

- and its estimator is given by

$$\hat{\xi} = \sum_{i=1}^{C} N_i^{\hat{\alpha}}, \tag{4}$$

where $\hat{\alpha}$ is the estimator of α.

Estimation of the parameters and verification of assumptions

- The parameters are estimated using maximum likelihood (the loglik function, gradient and hessian are provided in the working paper).
- This model can be further extended to account for covariates.
- Assumptions of the model can be verified using the following linearized model

$$\log\left(\frac{m_i}{N_i}\right) = (\alpha - 1)\log N_i + \beta \log\left(\frac{n_i}{N_i}\right) + \epsilon_i, \tag{5}$$

- We should expect a negative relationship with $\log N_i$ and a positive one with $\log(n_i/N_i)$.

Outline

Definitions

- For administrative purposes, Polish authorities (Polish Border Guard, 2020) use the term *illegal stay*, which is defined as *a stay which does not comply with the legal provisions describing the conditions that foreigners must meet in order to enter and stay in the Republic of Poland*.
- If a foreigner is found to be staying in Poland illegally, an administrative procedure is initiated whereby the person is obliged to leave the country.

Notes

- **Note 1**: We have a population register: PESEL.
- **Note 2**: The results refer to the period before the full-scale invasion of Ukraine by Russia (but the increase in migration was observed after the annexation of Crimea).
- **Note 3**: If you are interested in research on Ukrainians after the war I would recommend works of the: The Centre of Migration Research University of Warsaw (prof. Paweł Kaczmarczyk) & The Institute for Structural Research (dr Piotr Lewandowski).

Data – Border guard data

Table 1: The number of irregular foreigners in Poland by place of apprehension and re-apprehension status in 2019

Half	Same year	Within country	Airports	Ukraine	Russia	Belarus	Total
I	No	3,190	710	6,879	106	785	11,670
I	Yes	29	1	0	0	0	30
II	No	3,437	1,016	8,492	143	1,052	14,140
II	Yes	70	0	0	0	0	70

Data – Police data

Table 2: The number of foreigners in police records by registration type and residence status (registered for temporary stay or permanent residence) in 2019

Half	Registered	Procedural	Search	Traffic	Criminal	Total
I	Yes	1,499	715	9,286	10	11,510
I	No	4,046	6,522	2,477	16	13,061
II	Yes	2,080	878	11,988	6	14,952
II	No	4,644	5,979	2,867	11	13,501

Data - registered population

Table 3: The number of foreigners in the PESEL register by registration type at quarter ends in 2019

As at	No address	Temporary	Permanent	De-registered	Expired	Outside
31.03	81,202	242,318	56,476	16,158	124,368	332,256
30.06	107,545	249,154	57,656	16,246	157,476	383,283
30.09	134,483	246,990	59,228	16,340	196,209	441,705
31.12	160,868	252,245	60,440	16,386	225,690	496,374

Note: The *Outside* means that the address of residence is outside Poland (applied for the PESEL ID but live outside Poland).

Data – comparison

Table 4: The number of foreigners and countries by data source, sex and period before applying the condition for the model

Classification		Number of foreigners		Number of countries	
Source	Sex	1^{st} period	2^{st} period	1^{st} period	2^{st} period
PESEL	Total	232,468	234,194	151	147
	Women	137,424	137,880	145	140
	Men	95,044	96,314	127	130
Border Guard	Total	3,187	3,435	77	68
	Women	762	776	40	39
	Men	2,425	2,659	72	67
Police (all)	Total	20,138	23,330	100	98
	Women	3,017	3,079	58	57
	Men	17,121	20,251	94	94

Data – data for the model

- In our study we used Polish data from two halves of 2019 for the foreign population aged 18+.
- In addition, we derived data broken down by sex and economic age group (18-59 and 60+ for women; 18-64 and 65+ for men).
- The PESEL register contained people from 151 and 147 countries in the first and second half of the year, respectively, police data – around 100, and Border Guard records – around 70.
- The model requires that the following conditions hold: $m_{tij} > 0$, $n_{tij} > 0$ and $n_{tij}/N_{tij} < 1$, so we created a new dataset that meets these requirements.
- After applying this condition, we received a total of 73 countries (including category *other*), of which 50 were observed in both periods and 23 only in one (65 in the first and 58 in the second half of 2019).

Outline

Assumptions

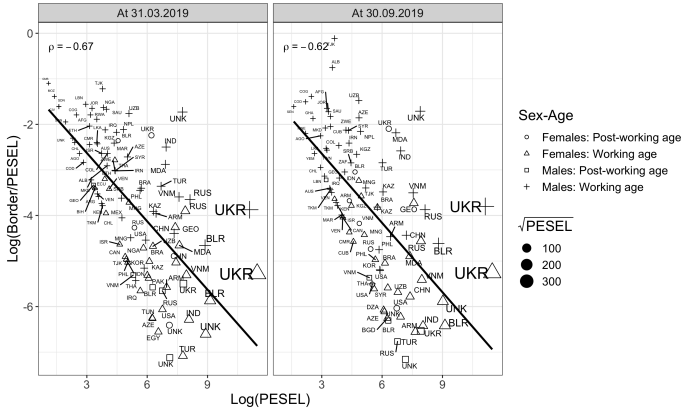


Figure 3: The relationship between the log of the PESEL population and the log of the BG-to-PESEL counts at the end of first and third quarter of 2019

Assumptions

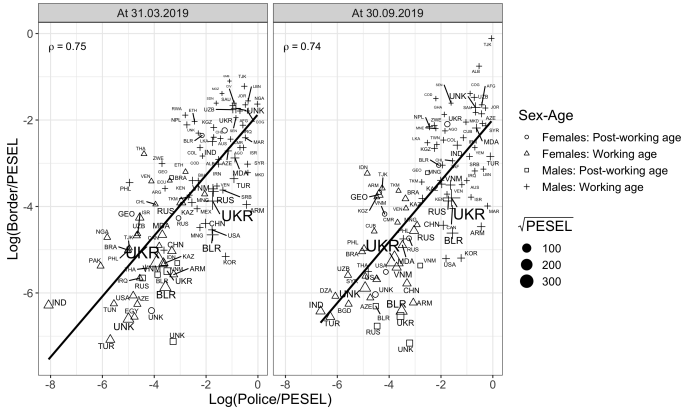


Figure 4: The relationship between the log of police-to-PESEL counts and the log of BG-to-PESEL counts at the end of first and third quarter of 2019

Population size estimation results

Table 5: Quality of models used in the study and the estimated population $\hat{\xi}$

Covariates for α	LogLik	AIC	BIC	$\hat{\xi}$
At the end of 1^{st} quarter 2019				
No covariates	-733.1	1,470.3	1,475.5	24,119.9
Ukraine	-648.7	1,303.5	1,311.3	20,835.8
Sex	-682.5	1,371.0	1,378.8	51,982.8
Ukraine & Sex	-630.1	1,268.1	1,278.6	34,870.1
At the end of 3^{rd} quarter of 2019				
No covariate	-822.2	1,648.3	1,653.4	23,582.6
Ukraine	-735.7	1,477.5	1,485.1	21,139.0
Sex	-742.2	1,490.3	1,497.9	65,011.0
Ukraine & Sex	-689.8	1,387.6	1,397.8	49,080.1

Outline

Discussion

- In the paper we propose a different approach to estimating the hard-to-reach population of irregular foreigners based on a flexible non-linear count regression model.
- The approach is an alternative to classic capture-recapture methods, which rely on one or multiple sources, and the interpretation of results is more intuitive because the irregular population is conditionally dependent on the regular population.
- The approach only requires administrative data and, as a result, the quality of our estimates depends on the availability of high-quality register-based statistics.
- Selection of data for the model should be strictly connected with the definition of the irregular population used in the study.

Outline

Literature (selected)

- Beręsewicz, M., Gudaszewski, G., and Szymkowiak, M. (2019). Estymacja liczby cudzoziemców w Polsce z wykorzystaniem metody capture-recapture. Wiadomości Statystyczne. The Polish Statistician, 64(10), 7-35.
- Beręsewicz, M., & Pawlukiewicz, K. (2020). Estimation of the number of irregular foreigners in Poland using non-linear count regression models. arXiv preprint arXiv:2008.09407.
- Polish Border Guard (2020). Consequences of illegal stay
- Zhang, L.-C. (2008). Developing methods for determining the number of unauthorized foreigners in Norway. Statistics Norway (SSB), Division for Statistical Methods and Standards. www. ssb. no.(accessed July 28, 2008)