Foreigners in the national labour market – a regional approach using the capture-recapture method

Marcin Szymkowiak\textsuperscript{1,2}, Maciej Beręsewicz\textsuperscript{1,2}, Dorota Szałtys\textsuperscript{3}

\textsuperscript{1}Poznań University of Economics and Business
\textsuperscript{2}Statistical Office in Poznań
\textsuperscript{3}Statistics Poland

Work Session on Migration Statistics
Presentation outline

1. Introduction
2. Problems of measuring the foreigner population
3. "Foreigners in the regional labour market" project
4. Summary and recommendations
5. Literature
Outline

1. Introduction
2. Problems of measuring the foreigner population
3. "Foreigners in the regional labour market” project
4. Summary and recommendations
5. Literature
A big demand for information about the real number of foreigners staying in Poland.

Data about immigration resources are an important element of various policies i.e. cohesion policy, especially given the fact that the intensity of foreign immigration varies across the country.

There is currently no reliable and direct source of information on this topic.

Administrative registers provide information about de iure (registered) population, while statistics is interested in de facto (registered + unregistered) population.
Selectivv collects and analyzes data about users of mobile phones, mobile applications and website visitors.

Using the collected information, a special study was conducted about people from Ukraine staying in Poland.

It was assumed that people with a SIM card from a Polish operator can be classified into this group if their phone has Russian or Ukrainian language settings.
Research problem – motivation

Fig. 1: Distribution of Ukrainians in Poland in 2018 based on data from smartphones. Source: Selectivv
How can foreigners be counted?
What problems are involved when counting foreigners?

- **Problems concerning the foreigner population:**
  - Definition of the foreigner population (i.e. who are we studying?),
  - Registered foreigners (e.g. registered for temporary or permanent residence, working legally),
  - Unregistered foreigners (e.g. staying in Poland temporarily, in informal / undeclared employment).

- **Problems concerning data sources:**
  - Does official statistics already use them (e.g. PESEL vs police data)?
  - Are these data from public administration or businesses (e.g. registers, Facebook, mobile phone networks)?
Goal of the presentation

The presentation aims to:

- present the problems of measuring the foreigner population,
- present results of a study conducted by a team from Statistics Poland, the Statistical Office in Poznań and Poznań University of Economics and Business,
- discuss problems and recommend solutions.
Outline

1. Introduction
2. Problems of measuring the foreigner population
3. "Foreigners in the regional labour market" project
4. Summary and recommendations
5. Literature
A hard-to-survey population is one for which no sampling frame is available and whose units of study are difficult to identify.

All populations are hard to survey, but some populations present special challenges of various sorts that make them harder to survey than the general population.

Some of these hard-to-survey populations are rare; others are hidden; some are difficult to find or contact; still others are unlikely to cooperate with survey requests.
Hard-to-sample populations

- In the ideal case, there is a complete and up-to-date frame of the target population and the sample can be drawn from this frame.
- Unfortunately, this ideal is rarely realized in practice; for most populations of interest in surveys, there is no sampling frame.
- Problems arise when the target population represents a small fraction of the population frame.
- One type of population that poses special difficulties for sample designers are mobile, foreigner or “elusive” populations.
Hard-to-identify populations

- Difficulties in identifying members of some cultural or religious minorities, such as immigrants, sexually active homosexual men or Muslims.
- Members of a highly stigmatized population, such as illicit drug users, are likely to keep this fact secret even from other household members.
Hard-to-locate populations

- Members of traditionally nomadic cultures (such as the Bedouins of Southwest Asia and the Tuareg of North Africa).
- Itinerant minorities (such as the Romani in Europe or the Travellers in Ireland).
- Persons who are temporarily mobile or displaced (recent immigrants, homeless persons, refugees).
- Persons at a mobile stage in their life cycle (college students).
Once the sample person is reached, there is still the problem of getting them to agree to take part in the survey.

Unwillingness to participate in a survey can be related to the sensitivity of the subject matter or lack of time.

Examples of this type of population: people who are professionally active, working off the books or illegally, foreigners.
These can include vulnerable populations (such as prisoners or young children), requiring explicit consent from a caretaker, parent, or guardian to be interviewed. They may suffer from cognitive or physical impairments that make it difficult or impossible to interview them, at least using standard survey protocols. They may not speak (or read) the language in which the survey questionnaire is written.
Examples of hard-to-survey populations:

- homeless people,
- drug addicts, alcoholics, Internet / smart-phone addicts, other types of addicts,
- members of various minorities,
- foreigners.
Methods of studying hard-to-survey populations

- **Sample-based methods** – snowball sampling, referral sampling or respondent-driven sampling (*which we won’t be considering as we are interested in existing data and not obtaining new data*),

- **Based on administrative sources** – *capture-recapture*, other model-based approaches (e.g. latent class models, mixed models).
Capture-recapture method

Day 1: Catch some fish, count them, tag them, and return them

First capture - $n_1$

Recaptures - $n_3$

Day 2: Catch more fish and observe the proportion tagged

Second capture - $n_2$

Total population
Capture-recapture method

To estimate the population size, it is necessary to use the Petersen / Lincoln-Petersen estimator given by (assuming the binomial distribution):

$$\hat{N}_{LP} = \frac{n_1 n_2}{n_3} = \frac{8 \times 8}{2} = 32, \quad (1)$$

where $n_1$ is the number of units caught for the 1st time, $n_2$ is the number of units caught the 2nd time and $n_3$ is the number of units in both $n_1$, and $n_2$. In the case of small samples, the Chapman estimator is recommended:

$$\hat{N}_C = \frac{(n_1 + 1)(n_2 + 1)}{n_3 + 1} - 1 = \frac{(8 + 1)(8 + 1)}{2 + 1} - 1 = 26. \quad (2)$$
Using capture-recapture to estimate the population

**Tab. 1**: Number of Polish nationals in Holland in 2009 based on the population register and police data

<table>
<thead>
<tr>
<th>Population register</th>
<th>Police register</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>374</td>
<td>39 488</td>
</tr>
<tr>
<td>No</td>
<td>1 445</td>
<td>??</td>
</tr>
<tr>
<td>∑</td>
<td>1 819</td>
<td></td>
</tr>
</tbody>
</table>

Source: based on Table 2.1 from Gerritse (2016).
Let’s use the Petersen estimator to estimate the number of Poles in the Netherlands in 2009:

\[ \hat{N}_{LC} = \frac{1819 \times 39862}{374} \approx 193874, \]  

which means that there are 193 874 - (374 + 39 488 + 1445) = 152 567 Poles not included in the two aforementioned registers.
Capture–recapture for more sources

In the case of two or more sources, Wolter (1986) formulated the following assumptions:

1. The target population is defined the same way in all sources (i.e. each unit has a positive probability of occurring in the sources),
2. The population is closed (i.e. is constant in a given period),
3. Data sources are independent,
4. There are no coverage errors or duplicates,
5. There are no linking errors (i.e. records are linked using an identifier).

It is crucial that these assumptions should be satisfied to enable the use of these methods for two or more sources.
Outline

1. Introduction
2. Problems of measuring the foreigner population
3. "Foreigners in the regional labour market" project
4. Summary and recommendations
5. Literature
Aim of the study

To develop a **method of estimating the number of foreigners staying in Poland** temporarily, with special emphasis on foreigners working in Poland at the end of 2015 and 2016 and to present results of the study at NUTS 3 level (subregions).

The study is an innovative approach to the way survey results are processed by official statistics.
Log-linear models

- Log-linear models are currently a very important method of analyzing data in contingency tables.
- The development of methodology devoted to this technique of data analysis was initiated in the 1960s.
- These models are particularly useful in situations where there is no precise distinction between dependent variable and independent variables, and there is a need to detect dependencies in a certain set of data.
- These models also play a special role in estimating the size of a population which is hard-to-survey.
Log-linear models

- The starting point for using log-linear models in estimating the size of a population which is hard to survey is a properly constructed contingency table.
- Such a table is created by combining information about a hard-to-survey population from two or more different sources.

**Tab. 2: The case of two sources - contingency table $2 \times 2$**

<table>
<thead>
<tr>
<th>Source A</th>
<th>Source B</th>
<th>Source A</th>
<th>Source B</th>
<th>Source A</th>
<th>Source B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes (1)</td>
<td>Yes (1)</td>
<td>$n_{11}$</td>
<td>$n_{10}$</td>
<td>$n_{1+}$</td>
<td>$n_{1+}$</td>
</tr>
<tr>
<td>No (0)</td>
<td>No (0)</td>
<td>$n_{01}$</td>
<td>$n_{00}$</td>
<td>$n_{00}$</td>
<td>$n_{0+}$</td>
</tr>
<tr>
<td>$\sum$</td>
<td>$n_{+1}$</td>
<td>$n_{+0}$</td>
<td>$n$</td>
<td>$n$</td>
<td>$n$</td>
</tr>
</tbody>
</table>
Log-linear models

- The key issue is to estimate the number of $n_{00}$ i.e. the number of units that do not exist in both A and B sources. The final estimated population size is achieved by adding all values from Table 2 after estimating the number of $n_{00}$.

- The estimation of $n_{00}$ can be obtained by fitting the log-linear model to an incomplete contingency table. For example, for a $2 \times 2$ table (Table 2) referring to data sources A and B, the saturated log-linear model $[AB]$ can be represented in the form:

$$\ln (m_{ij}) = \mu + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB}, \quad i, j = \{\text{‘Yes’, ‘No’}\}, \quad (4)$$

where $m_{ij}$ denotes the expected number in cell $i, j$. 
Log-linear models

However, since cell $m_{00} = m_{(No,No)}$ is not observed, the [AB] model has one parameter too many and cannot be estimated. In such a situation, one can consider an independence model [A][B] given by:

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

(5)

which has only three parameters to be estimated, given the lack of interaction effect $\lambda_{ij}^{AB}$.

Because we have three observed cells in Table 2 and three parameters to be estimated, we basically deal with a saturated model.
Log-linear models

After fitting this model to the data, we can use the estimated parameters to determine the number in the missing cell (‘No’, ’No’) and then determine the size of the population analyzed.

To estimate the cell size $n_{00}$ we use the following formula:

$$\hat{n}_{00} = \exp (\mu).$$  \hspace{1cm} (6)
Log-linear models

- When estimating the size of a hard-to-survey population, it is possible to use auxiliary variables, such as, for example, sex or age groups.

**Tab. 3: The case of two sources A and B and one auxiliary variable X**

<table>
<thead>
<tr>
<th>Source A</th>
<th>Auxiliary variable X</th>
<th>Source B</th>
<th>Source B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>X₁</td>
<td>X₂</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Yes (1)</td>
<td>Yes (1)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>No (0)</td>
<td>No (0)</td>
<td></td>
</tr>
<tr>
<td>Yes (1)</td>
<td>n₁₁₁</td>
<td>n₁₀₁</td>
<td>n₁₁₀</td>
</tr>
<tr>
<td>No (0)</td>
<td>n₀₁₁</td>
<td>n₀₀₁</td>
<td>n₀₁₀</td>
</tr>
<tr>
<td>∑</td>
<td>n₁⁺⁺</td>
<td>n₀⁺⁺</td>
<td></td>
</tr>
</tbody>
</table>

- ∑: Total
Log-linear models

For example, in the case of a two-dimensional \(2 \times 2\) contingency table, apart from the fact if belonging to two sources A and B, an additional variable X (for example, sex) can be considered, so the table should be extended to a three-way table and a log-linear model \([AX][BX]\) can be fitted to the data:

\[
\ln(m_{ijx}) = \mu + \lambda_i^A + \lambda_j^B + \lambda_x^X + \lambda_{ix}^{AX} + \lambda_{jx}^{BX},
\] (7)

where \(\lambda_{ix}^{AX}\) and \(\lambda_{jx}^{BX}\) denote the interaction effects between the auxiliary variable X and the data sources A and B respectively.
Log-linear models

In the case of two sources A and B and one auxiliary variable X, with, say, two levels \( X_1 \) and \( X_2 \) (e.g. male and female), we deal with a three-way \( 2 \times 2 \times 2 \) contingency table, in which the missing numbers to be estimated are \( n_{001} \) and \( n_{000} \).

There are now six cells in Table 3 for which the observed numbers are known. In this case model (7) contains six parameters to be estimated (a saturated log-linear model).

After fitting the model to the data, the missing cell numbers are estimated using the formulas: \( \hat{n}_{000} = \exp (\mu) \) and \( \hat{n}_{001} = \exp (\mu + \lambda^X_{X_1}) \).
Data sources – estimation of the number of foreigners

The number of foreigners (total, by sex, age, place of residence) in Poland was estimated using the following sources (as at 31 Dec 2016):

- **The Office for Foreigners - the „Pobyt” system** – a set of registers concerning foreigners’ affairs,
- **Ministry of Digital Affairs** – the PESEL register concerning foreigners registered for permanent or temporary residence,
- **Social Insurance Institution** – the Central Register of Insured Persons concerning insured foreigners and their family members.
Additional data sources – characteristics of foreigners

The target variables (incl. NACE section, length of stay, level of education, labour market status), assuming the relationships are correct, were estimated using the following sources:

- **Ministry of Family, Labour and Social Policy** – concerning work permits and employers’ declarations of willingness to employ a foreigner,

- **Census 2011** – the survey-based component.

- **Labour Force Survey 2015 and 2016.**

**Note:** in the case of these variables, it was assumed that population structures obtained from a given source were correct for the target population!
Estimation results – the foreigner population

Tab. 4: Estimated number of foreigners in Poland in 2015 and 2016

<table>
<thead>
<tr>
<th>Year</th>
<th>ÎN</th>
<th>Lower bound</th>
<th>Upper bound</th>
<th>Precision (in %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015</td>
<td>507 693</td>
<td>369 135</td>
<td>724 407</td>
<td>17.64</td>
</tr>
<tr>
<td>2016</td>
<td>743 665</td>
<td>600 796</td>
<td>943 124</td>
<td>11.70</td>
</tr>
</tbody>
</table>


**Estimation results – the foreigner population**

**Tab. 5: Estimated number of foreigners in Poland by citizenship**

<table>
<thead>
<tr>
<th>Year</th>
<th>Country</th>
<th>( \hat{N} )</th>
<th>Lower bound</th>
<th>Upper bound</th>
<th>Precision (in %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015</td>
<td>Armenia</td>
<td>3 168</td>
<td>2 263</td>
<td>4 505</td>
<td>18.33</td>
</tr>
<tr>
<td>2016</td>
<td>Armenia</td>
<td>4 773</td>
<td>3 897</td>
<td>6 032</td>
<td>11.35</td>
</tr>
<tr>
<td>2015</td>
<td>Belarus</td>
<td>19 868</td>
<td>14 429</td>
<td>27 951</td>
<td>17.38</td>
</tr>
<tr>
<td>2016</td>
<td>Belarus</td>
<td>25 813</td>
<td>20 832</td>
<td>32 569</td>
<td>11.81</td>
</tr>
<tr>
<td>2015</td>
<td>Moldova</td>
<td>2 693</td>
<td>1 613</td>
<td>4 227</td>
<td>25.59</td>
</tr>
<tr>
<td>2016</td>
<td>Moldova</td>
<td>7 580</td>
<td>5 355</td>
<td>10 617</td>
<td>17.99</td>
</tr>
<tr>
<td>2015</td>
<td>Russia</td>
<td>22 611</td>
<td>16 040</td>
<td>32 237</td>
<td>18.62</td>
</tr>
<tr>
<td>2016</td>
<td>Russia</td>
<td>25 534</td>
<td>20 685</td>
<td>32 344</td>
<td>12.07</td>
</tr>
<tr>
<td>2015</td>
<td>Vietnam</td>
<td>7 408</td>
<td>5 554</td>
<td>9 942</td>
<td>15.45</td>
</tr>
<tr>
<td>2016</td>
<td>Vietnam</td>
<td>11 728</td>
<td>10 008</td>
<td>14 170</td>
<td>9.10</td>
</tr>
<tr>
<td>2015</td>
<td>Ukraine</td>
<td><strong>283 714</strong></td>
<td><strong>203 946</strong></td>
<td><strong>415 732</strong></td>
<td><strong>18.55</strong></td>
</tr>
<tr>
<td>2016</td>
<td>Ukraine</td>
<td><strong>454 974</strong></td>
<td><strong>361 512</strong></td>
<td><strong>584 696</strong></td>
<td><strong>12.27</strong></td>
</tr>
<tr>
<td>2015</td>
<td>EU countries</td>
<td>70 901</td>
<td>53 579</td>
<td>97 126</td>
<td>15.63</td>
</tr>
<tr>
<td>2016</td>
<td>EU countries</td>
<td>59 571</td>
<td>50 914</td>
<td>71 169</td>
<td>8.77</td>
</tr>
<tr>
<td>2015</td>
<td>others</td>
<td>97 329</td>
<td>70 037</td>
<td>138 339</td>
<td>17.86</td>
</tr>
<tr>
<td>2016</td>
<td>others</td>
<td>153 692</td>
<td>124 170</td>
<td>196 140</td>
<td>12.06</td>
</tr>
</tbody>
</table>
Spatial distribution of foreigners in Poland
Distribution of foreigners in Poland by sex

- 2015:
  - Male: 67.4%
  - Female: 32.6%

- 2016:
  - Male: 66.1%
  - Female: 33.9%
### Distribution of foreigners in Poland by sex and province

<table>
<thead>
<tr>
<th>Province</th>
<th>Male 2015</th>
<th>Female 2015</th>
<th>Male 2016</th>
<th>Female 2016</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dolnośląskie</td>
<td>67.5</td>
<td>32.5</td>
<td>64.5</td>
<td>35.5</td>
</tr>
<tr>
<td>Kujawsko-pomorskie</td>
<td>76.1</td>
<td>23.9</td>
<td>74.1</td>
<td>25.9</td>
</tr>
<tr>
<td>Lubelskie</td>
<td>65.6</td>
<td>34.4</td>
<td>64.5</td>
<td>35.5</td>
</tr>
<tr>
<td>Lubuskie</td>
<td>78.2</td>
<td>21.8</td>
<td>76.5</td>
<td>23.5</td>
</tr>
<tr>
<td>Łódzkie</td>
<td>71.8</td>
<td>28.2</td>
<td>70.4</td>
<td>29.6</td>
</tr>
<tr>
<td>Małopolskie</td>
<td>64.9</td>
<td>35.1</td>
<td>63.2</td>
<td>36.8</td>
</tr>
<tr>
<td>Mazowieckie</td>
<td>62.9</td>
<td>37.1</td>
<td>62.1</td>
<td>37.9</td>
</tr>
<tr>
<td>Opolskie</td>
<td>73.5</td>
<td>26.5</td>
<td>71.2</td>
<td>28.8</td>
</tr>
<tr>
<td>Podkarpackie</td>
<td>64.2</td>
<td>35.8</td>
<td>63.9</td>
<td>36.1</td>
</tr>
<tr>
<td>Podlaskie</td>
<td>62.8</td>
<td>37.2</td>
<td>60.8</td>
<td>39.2</td>
</tr>
<tr>
<td>Pomorskie</td>
<td>68.5</td>
<td>31.5</td>
<td>67.1</td>
<td>32.9</td>
</tr>
<tr>
<td>Śląskie</td>
<td>71.7</td>
<td>28.3</td>
<td>71.5</td>
<td>28.5</td>
</tr>
<tr>
<td>Świętokrzyskie</td>
<td>68.7</td>
<td>31.3</td>
<td>68.4</td>
<td>31.6</td>
</tr>
<tr>
<td>Warmińsko-Mazurskie</td>
<td>66.7</td>
<td>33.3</td>
<td>66.4</td>
<td>33.6</td>
</tr>
<tr>
<td>Wielkopolskie</td>
<td>71.4</td>
<td>28.6</td>
<td>71.8</td>
<td>28.2</td>
</tr>
<tr>
<td>Zachodniopomorskie</td>
<td>77.0</td>
<td>23.0</td>
<td>73.8</td>
<td>26.2</td>
</tr>
</tbody>
</table>
Distribution of foreigners in Poland by age

![Bar chart showing distribution of foreigners in Poland by age in 2015 and 2016, with percentages for working age and post-working age categories.](image-url)
Distribution of foreigners in Poland by labor market status

- **2016**
  - Working: 63.6%
  - Inactive: 32.2%
  - Unemployed: 4.2%

- **2015**
  - Working: 66.6%
  - Inactive: 29.0%
  - Unemployed: 4.3%
Estimation results - use the police data

After we finished the project, we were granted access to (aggregated) data from the police register about persons with Ukrainian citizenship. Based on this information and data from the PESEL register, we used the Lincoln-Petersen estimator assuming that all assumptions of the capture-recapture method were met.

Tab. 6: The number of citizens of Ukraine in the PESEL register and police data

<table>
<thead>
<tr>
<th>Police data</th>
<th>PESEL</th>
<th>∑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes (1)</td>
<td>178</td>
<td>2 083</td>
</tr>
<tr>
<td>No (0)</td>
<td>33 581</td>
<td>2 261</td>
</tr>
<tr>
<td>∑</td>
<td>33 759</td>
<td></td>
</tr>
</tbody>
</table>

\[
\hat{n}_{00} = \frac{\hat{n}_1 + \hat{n}_1}{\hat{n}_{11}} = \frac{33759 \times 2261}{178} = 428815.2
\]
Outline

1. Introduction
2. Problems of measuring the foreigner population
3. "Foreigners in the regional labour market” project
4. Summary and recommendations
5. Literature
Summary and recommendations

- Assumptions adopted in the study were largely satisfied.
- Further research is required to improve the estimation method – requirement: the scope of administrative registers needs to be extended to include, among other things, identification variables.
- Further work is required to identify other potential sources:
  - police registers on foreigners suspected of committing crimes (the basic source mentioned in the literature),
  - data of the National Labour Inspectorate from inspections of compliance with regulations on the employment of foreigners,
  - data of the Polish Border Guard concerning the legality of stays,
  - data of the Ministry of Foreign Affairs concerning visas.
Summary and recommendations

An attempt to obtain more reliable estimates for subregions and two regions (the capital city of Warsaw and mazowiecki region) and the distribution of variables:

- length of stay,
- labour market status,
- NACE section,
- place of residence by province.

An attempt to identify the unregistered part of the estimated population (not included in the registers).
Further steps

- Estimate the number of foreigners staying illegally (without valid or with expired documents).
- Estimate the number of foreigners working illegally – e.g. on the basis of data from the Border Guard and the Ministry of Family, Labour and Social Policy (e.g. revoked permits).
Literature I

Literature II


Thank you for your attention!