POZNAŃ UNIVERSITY OF ECONOMICS AND BUSINESS

Statistical Office in Poznań

Statistics Poland

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

Marcin Szymkowiak[1,2], Maciej Beręsewicz[1,2], Dorota Szałtys[3]

[1]Poznań University of Economics and Business
[2]Statistical Office in Poznań
[3] Statistics Poland

Work Session on Migration Statistics
29.10.2019–31.10.2019, Geneva

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

Research problem – motivation

- 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.

Research problem – motivation

- 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

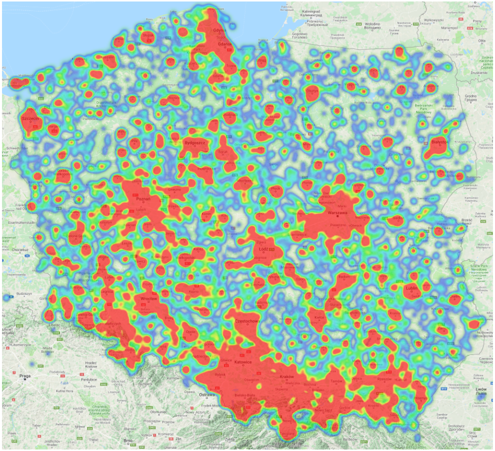


Fig. 1: Distribution of Ukrainians in Poland in 2018 based on data from smartphones. Source: Selectivv

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

Hard-to-survey population

- 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).

Hard-to-persuade populations

- 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.

Hard-to-interview populations

- 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.

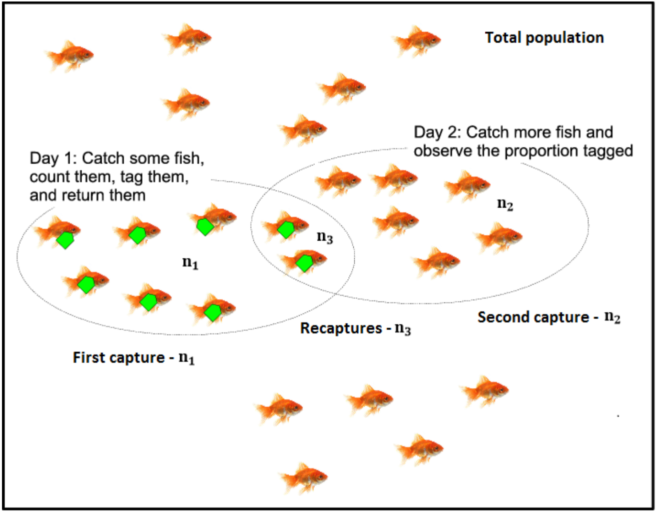
Examples of hard-to-survey populations:

- 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


Total population
Day 2: Catch more fish and observe the proportion tagged
Day 1: Catch some fish, count them, tag them, and return them
n_2
n_3
n_1
Second capture - n_2
Recaptures - n_3
First capture - n_1

Capture-recapture method

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

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

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:

$$\hat{N}_C = \frac{(n_1+1)(n_2+1)}{n_3+1} - 1 = \frac{(8+1)(8+1)}{2+1} - 1 = 26. \tag{2}$$

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

	Police register		
Population register	Yes	No	$\sum$
Yes	374	39 488	39 862
No	1 445	??	
$\sum$	1 819		

Source: based on Table 2.1 from Gerritse (2016).

Poles in the Netherlands

Let's use the Petersen estimator to estimate the number of Poles in the Netherlands in 2009:

$$\hat{N}_{LC} = \frac{1819 \times 39862}{374} \approx 193874, \tag{3}$$

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 2×2

		Source B		
		Yes (1)	No (0)	$\sum$
Source A	Yes (1)	n_{11}	n_{10}	n_{1+}
	No (0)	n_{01}	n_{00}	n_{0+}
$\sum$		n_{+1}	n_{+0}	n

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 2×2 table (Table 2) referring to data sources A and B, the saturated log-linear model [AB] can be represented in the form:

$$\ln\left(m_{ij}\right) = \mu + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB}, \quad i,j = \left\{'\mathrm{Yes}', '\mathrm{No}'\right\}, \tag{4}$$

where m_{ij} denotes the expected number in cell i, j.

Log-linear models

- However, since cell $m_{00} = m_{(\mathrm{No,No})}$ 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:

  $$\ln\left(m_{ij}\right) = \mu + \lambda_i^A + \lambda_j^B, \tag{5}$$

  which has only three parameters to be estimated, given the lack of interaction effect λ_{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:

$$\hat{n}_{00} = \exp\left(\mu\right). \tag{6}$$

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

		Auxiliary variable X				
		X_1		X_2		
		Source B		Source B		
		Yes (1)	No (0)	Yes (1)	No (0)	$\sum$
Source A	Yes (1)	n_{111}	n_{101}	n_{110}	n_{100}	n_{1++}
	No (0)	n_{011}	n_{001}	n_{010}	n_{000}	n_{0++}
$\sum$		n_{+11}	n_{+01}	n_{+10}	n_{+00}	n

Log-linear models

- For example, in the case of a two-dimensional 2×2 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 3 and a log-linear model [AX][BX] can be fitted to the data:

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

where λ_{ix}^{AX} and λ_{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: $\hat{n}_{000} = \exp\left(\mu\right)$ and $\hat{n}_{001} = \exp\left(\mu + \lambda^{X}_{X_1}\right)$.

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 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

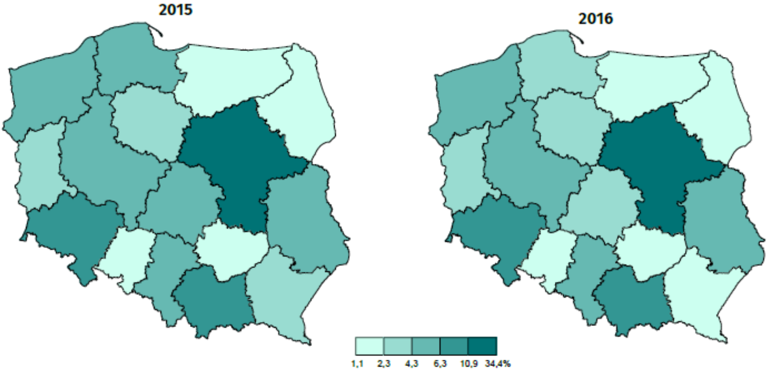
Year	$\hat{N}$	Lower bound	Upper bound	Precision (in %)
2015	507 693	369 135	724 407	17.64
2016	743 665	600 796	943 124	11.70

Estimation results – the foreigner population

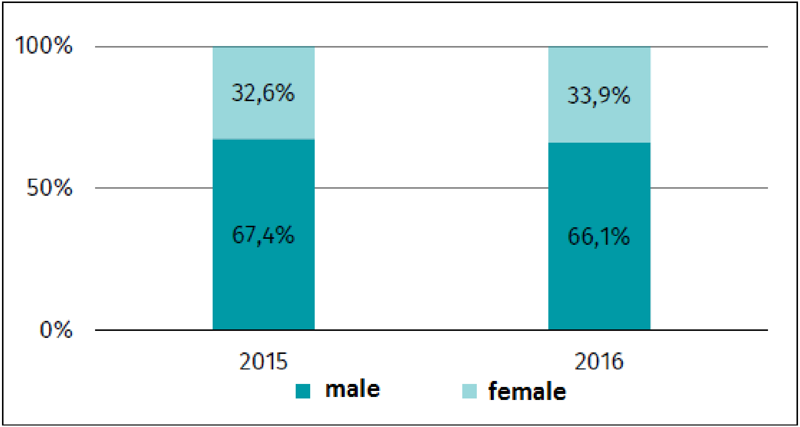
Tab. 5: Estimated number of foreigners in Poland by citizenship

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

Spatial distribution of foreigners in Poland



Distribution of foreigners in Poland by sex

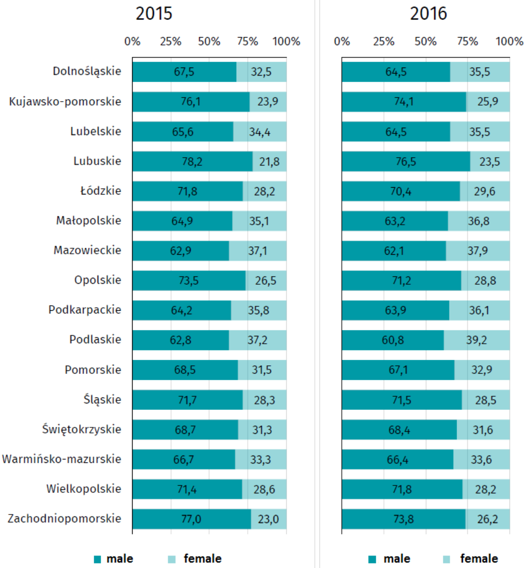


POZNAŃ UNIVERSITY OF ECONOMICS AND BUSINESS

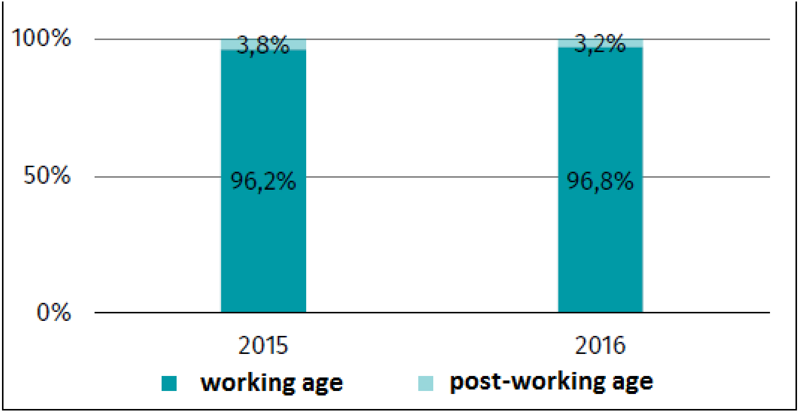
Statistical Office in Poznań

Statistics Poland

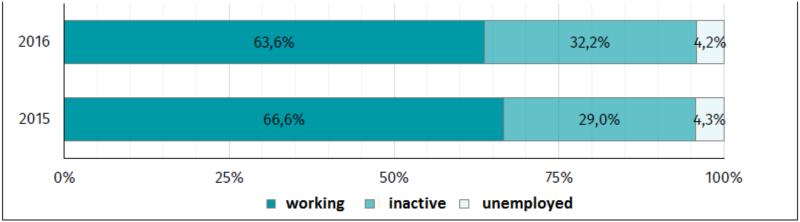
Distribution of foreigners in Poland by sex and province



Distribution of foreigners in Poland by age



Distribution of foreigners in Poland by labor market status



POZNAŃ UNIVERSITY OF ECONOMICS AND BUSINESS

Statistical Office in Poznań

Statistics Poland

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

		PESEL		Σ
		Yes (1)	No (0)	
Police data	Yes (1)	178	2 083	2 261
	No (0)	33 581	?	
Σ		33 759		

$$\hat{n}_{00} = \frac{\hat{n}_{1+}\hat{n}_{+1}}{\hat{n}_{11}} = \frac{33759 \times 2261}{178} = 428815.2 \tag{8}$$

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).

Outline

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

Literature I

- Agresti A. (2013), *Categorical Data Analysis*. Wiley.
- 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.
- 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.
- 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.
- Rivest L.P., Baillargeon S. (2014), *Rcapture: Loglinear Models for Capture-Recapture Experiments*. R package version 1.4-2.

Literature II

- Wolter K.M. (1986), *Some coverage error models for census data*. Journal of the American Statistical Association, 81(394):337–346.
- Van der Heijden, P. G., Whittaker, J., Cruyff, M., Bakker, B., & Van der Vliet, R. (2012). People born in the Middle East but residing in the Netherlands: Invariant population size estimates and the role of active and passive covariates. The Annals of Applied Statistics, 6(3), 831-852.
- 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., 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.

Thank you for your attention!

POZNAŃ UNIVERSITY
OF ECONOMICS
AND BUSINESS

Statistical Office
in Poznań

Statistics Poland