Calibration and Small Area Estimation Methods in Polish National Census of Population and Housing 2011 - First Results

Marcin Szymkowiak

University of Economics in Poznan

Outline

The objective of the census

The objective of the census

1. The main objective of the census was to provide the most detailed information on the numbers in the population, its territorial spread, socio-demographic and professional structures, and the socio-economic specificity of households and families, as well as their resources and dwelling conditions at all levels of the country's territorial division: national, regional, and local.
2. Considerable weight in the 2011 National Census was attached to acquiring knowledge on the changes in demographic and social processes, inter alia, due to the increased migration after Polish accession to the European Union.
3. The results of the census are directly applicable to the needs of public statistics as a basis for creating sampled frames to be employed in later sample surveys conducted on a sample of households.
4. In the census conducted in 2011 it was very important to obtain information about issues that were covered by the census in 2002. It is still necessary to conduct comparative analyses of developments over time and to describe the changes that have occurred in demographic, social and economic processes, in terms of: population, dwellings and buildings status, and households and families, in relation to housing conditions.

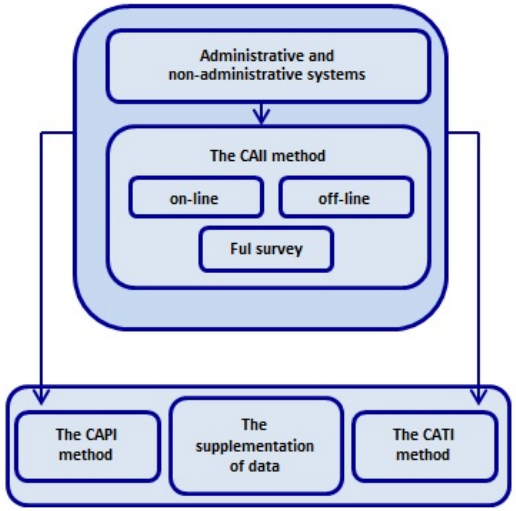
The NCPH 2011 Methodology

The NCPH 2011 Methodology

1. NCPH 2011 was carried out as a full-scale survey (administrative registers) and as a sample survey.
2. Poland used the mixed model of collecting data consisting of merging the data from administrative registers with the data obtained from direct statistical surveys.
3. Central Statistical Office in Poland decided to collect data using mixed approach because of the fact it was safer and more effective, taking into consideration the present level of development of administrative sources, their quality, and the degree of advancement of methodological work concerning the estimation and imputation of missing data in administrative sources.
4. As a result of the use of administrative registers and modern technologies for obtaining data it made possible to reduce the number of enumerators working in the field by over ten times – from approx. 170 thousand in the last census in 2002 to 18 thousand in the 2011 census. This allowed a reduction in census costs by approx. EUR 50 million.

The full-scale survey

- Data from administrative registers – the Master Record
- Data obtained with the CAII method
- Data supplemented with the CATI and CAPI methods

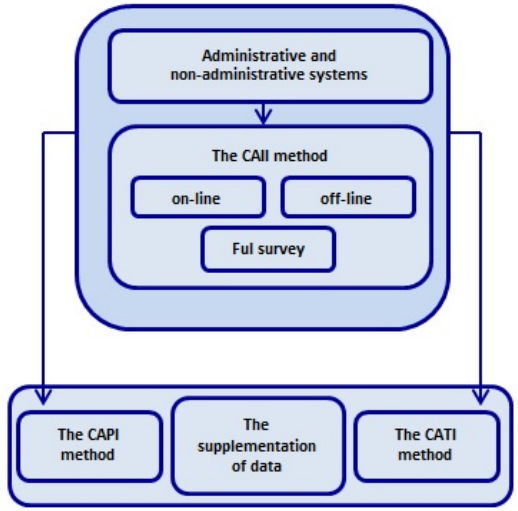


The full-scale survey

1. The full-scale survey involved population and housing, and was conducted with the use of administrative registers supplemented with a brief questionnaire to be filled in by each respondent.
2. For the first time in Poland 28 administrative sources were used in order to obtain the values of the census variables, both at the stage of creating a specification of census units (population and housing census) and for qualitative comparisons.
3. Due to a stable system of identifiers (PIN Personal Identification Number) it was possible to merge data from different registers.

The full-scale survey

- Data from administrative registers – the Master Record
- Data obtained with the CAII method
- Data supplemented with the CATI and CAPI methods

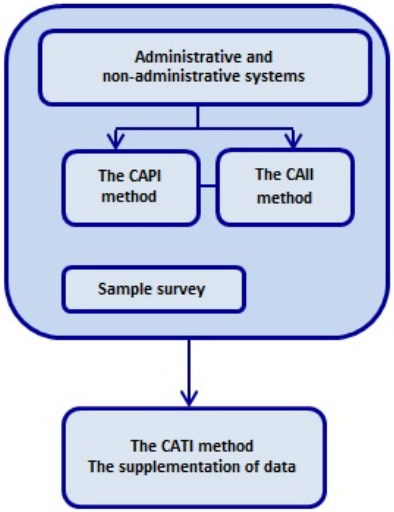


The full-scale survey

4. The supplementation of data was made using CATI (Computer Assisted Telephone Interview) and CAPI (Computer Assisted Personal Interviewing) methods.
5. They were used as supplementary channels, rather than the main channel for the acquisition of data. The basic method of obtaining data in the full-scale survey involved so called the „Master" record and the CAII method (Internet self-enumeration).
6. The Master record, being a set of variables derived from the registers, was the main channel supporting the collection of data, apart from Internet self-enumeration, phone interviews and direct interviews.

Sample survey

- Data obtained with CAII method
- Data obtained with the CAPI method
- Data supplemented with the CATI method (only if the survey requires a small supplementation)



Sample survey

1. A sample survey is carried out on persons who permanently or temporarily reside in the territory of the Republic of Poland, and whose households have been sampled.
2. A sample survey was carried out using the CAII and CAPI methods. Data were supplemented with the CATI method.
3. A sample survey was carried out on a sample of 20% of dwellings and approximately 20% of population in Poland was drawn to the sample. Design weights associated with units drawn to the sample hade to be calibrated to known demographic totals from administrative registers.

Theoretical background of calibration

Theoretical background of calibration

1. This technique was proposed by Devill and Särndal (1992) and is a method of searching for so called calibrated weights by minimizing distance measure between the sampling weights and the new weights, which satisfy certain calibration constraints.
2. As a consequence when the new weights are applied to the auxiliary variables in the sample, they reproduce the known population totals of the auxiliary variables exactly.
3. It is also important that the new weights should be as close as possible to sampling weights in sense of chosen distance measure (Särndal C-E., Lundström S. 2005, Särndal C-E. 2007).

Theoretical background of calibration

Theoretical background of calibration

- Let us assume that the whole population $U = \{1, 2, \ldots, N\}$ consists of N elements.
- From this population we draw, according to a certain sampling scheme, a sample $s \subseteq U$, which consists of n elements.
- Let π_i denote first order inclusion probability $\pi_i = P(i \in s)$ and $d_i = 1/\pi_i$ the design weight.
- Let us assume that our main goal is estimation of the total value of the variable y:

$$Y = \sum_{i=1}^{N} y_i, \tag{1}$$

where y_i denotes the value of the variable y for i-th unit, $i = 1, \ldots, N$.

Theoretical background of calibration

Theoretical background of calibration

- Let $x_1, \ldots, x_k$ denote auxiliary variables which will be used in the process of finding calibration weights and let $\mathbf{X}_j$ denote the total value for the auxiliary variable x_j, $j = 1, \ldots, k$, e.i.

$$\mathbf{X}_j = \sum_{i=1}^{N} x_{ij}, \tag{2}$$

where x_{ij} odenotes the value of j-th auxiliary variable for the i-th unit.

- In practice it occurs that:

$$\sum_{s} d_i x_{ij} \neq \mathbf{X}_j \tag{3}$$

so calibration is required.

Theoretical background of calibration

Theoretical background of calibration

- Let $\mathbf{w} = (w_1, \ldots, w_n)^T$ denote the vector of calibration weights.
- Our main goal is to look for new weights w_i which are as close as possible to the design weights d_i and which allow us to get known population totals from administrative registers exactly.
- The process of construction calibration weights depends on the properly chosen distance function.
- Let G denote function for which the second derivative exists and:
  - $G(\cdot) \geq 0$,
  - $G(1) = 0$,
  - $G'(1) = 0$,
  - $G''(1) = 1$.

Examples of G function

Examples of G function

$$G_1(x) = \frac{1}{2}(x-1)^2, \tag{4}$$

$$G_2(x) = \frac{(x-1)^2}{x}, \tag{5}$$

$$G_3(x) = x(\log x - 1) + 1, \tag{6}$$

$$G_4(x) = 2x - 4\sqrt{x} + 2, \tag{7}$$

$$G_5(x) = \frac{1}{2\alpha}\int_1^x \sinh\left[\alpha\left(t - \frac{1}{t}\right)\right] dt. \tag{8}$$

The choice of G function

The choice of G function

- The most common G function which can be used in the process of construction distance function is $G_1(x) = \frac{1}{2}(x-1)^2$. In this case we have:

$$D(\mathbf{w}, \mathbf{d}) = \sum_{i=1}^{n} d_i G\left(\frac{w_i}{d_i}\right) = \sum_{i=1}^{n} d_i \frac{1}{2}\left(\frac{w_i}{d_i} - 1\right)^2 = \frac{1}{2}\sum_{i=1}^{n} \frac{(w_i - d_i)^2}{d_i}. \quad (9)$$

The problem of finding calibration weights

The problem of finding calibration weights

(C1) Find the minimum of distance function:

$$D(\mathbf{w}, \mathbf{d}) = \frac{1}{2} \sum_{i=1}^{n} \frac{(w_i - d_i)^2}{d_i} \longrightarrow \min, \tag{10}$$

(C2) Calibration equations:

$$\sum_{i=1}^{n} w_i x_{ij} = \mathbf{X}_j, \quad j = 1, \ldots, k, \tag{11}$$

(C3) Calibration constraints:

$$L \leq \frac{w_i}{d_i} \leq U, \quad \text{where: } L < 1 \text{ i } U > 1, \ i = 1, \ldots, n. \tag{12}$$

The calibration estimator for total

The calibration estimator for total

The calibration estimator for total takes the form:

$$\hat{Y}_{cal} = \sum_{i=1}^{n} w_i y_i \,, \tag{13}$$

where the vector of calibration weights $\mathbf{w} = (w_1, w_2, \ldots, w_n)^T$ is obtained as the following minimization problem:

$$\mathbf{w} = \text{argmin}_v D(\mathbf{v}, \mathbf{d}) \,, \tag{14}$$

$$\mathbf{X} = \tilde{\mathbf{X}}, \tag{15}$$

where

$$D(\mathbf{v}, \mathbf{d}) = \frac{1}{2} \sum_{i=1}^{n} \frac{(v_i - d_i)^2}{d_i}, \tag{16}$$

$$\tilde{\mathbf{X}} = \left(\sum_{i=1}^{n} w_i x_{i1}, \sum_{i=1}^{n} w_i x_{i2}, \ldots, \sum_{i=1}^{n} w_i x_{ik} \right)^T, \quad \mathbf{X} = \left(\sum_{i=1}^{N} x_{i1}, \sum_{i=1}^{N} x_{i2}, \ldots, \sum_{i=1}^{N} x_{ik} \right)^T. \tag{17}$$

Theorem

Theorem

The solution of the minimization problem is the vector of calibration weights $\mathbf{w} = (w_1, w_2, \ldots, w_n)^T$, for which

$$w_i = d_i + d_i \left(\mathbf{X} - \hat{\mathbf{X}}\right)^T \left(\sum_{i=1}^{n} d_i \underline{\mathbf{x}}_i \underline{\mathbf{x}}_i^T\right)^{-1} \underline{\mathbf{x}}_i \tag{18}$$

where

$$\hat{\mathbf{X}} = \left(\sum_{i=1}^{n} d_i x_{i1}, \sum_{i=1}^{n} d_i x_{i2}, \ldots, \sum_{i=1}^{n} d_i x_{ik}\right)^T, \tag{19}$$

$$\underline{\mathbf{x}}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})^T. \tag{20}$$

CALMAR

CALMAR

- In many statistical packages the problem of finding calibration weights is implemented using different G functions.
- In CALMAR, which is a macro written in 4GL in SAS four distance functions were implemented: the linear method, the raking ratio metod, the logit method, the truncated linear method.
- In CALMAR 2 which is a later version of CALMAR, the distance function based on hyperbolic sinus function was also implemented.
- In the problem of finding calibration weights in NCPH 2011 G_1 function and macro CALMAR were used.

Practical aspects of calibration in NCPH 2011

Practical aspects of calibration in NCPH 2011

1. Using data from many sources required on stage of generalization of results adjustment of initial weights assigned to all units drawn to a sample.
2. It was due to the fact that results from administrative registers and 20% sample should be consistent related to some basic demographic characteristic including gender, age and place of living.
3. In order to adjust design weights to reproduce known totals from administrative registers related to mentioned demographic characteristic calibration was used.

Practical aspects of calibration in NCPH 2011

Practical aspects of calibration in NCPH 2011

- In NCPH 2011 mixed approach of collecting data was used: administrative registers and survey sampling (20% of population).
- Some tables, especially related to demographic variables, were constructed using data from administrative registers (for example population in Poland in different cross-sections defined by sex, age and place of residence (urban areas, rural areas) in different territorial division from PESEL register.
- Many tables were created using data coming from the sample survey i.e. tables related to the level of education, labour market status etc.
- Design weights from the survey had to be calibrated because they did not reproduce known population totals from registers exactly.
- In NCPH 2011 design weights were calibrated in different cross-sections in different territorial division.

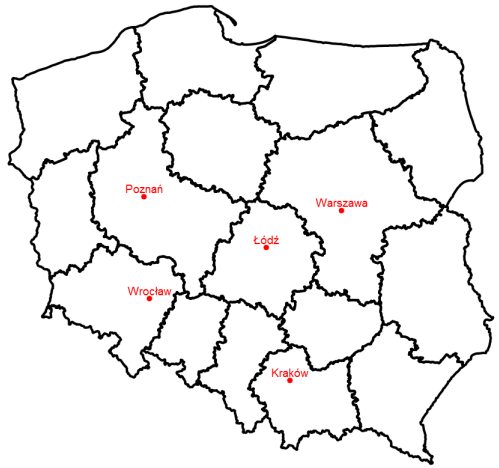
Practical aspects of calibration in NCPH 2011

- **Voivodeships: sex x place of residence x individual years of age (0,1,...,83,84,85+)**
- **Poviats: sex x place of residence x age groups (0–4,5–9,...,80–84,85+)**
- **The biggest cities: sex x individual years of age (0,1,...,83,84,85+ or 100+ for Warsaw)**

Voivodeships (16) - NUTS 2

Poviats (379) - LAU 1

The biggest cities in Poland



Practical aspects of calibration in NCPH 2011

Practical aspects of calibration in NCPH 2011

- Auxiliary variables from registers taken into account in calibration process: sex, age and place of residence

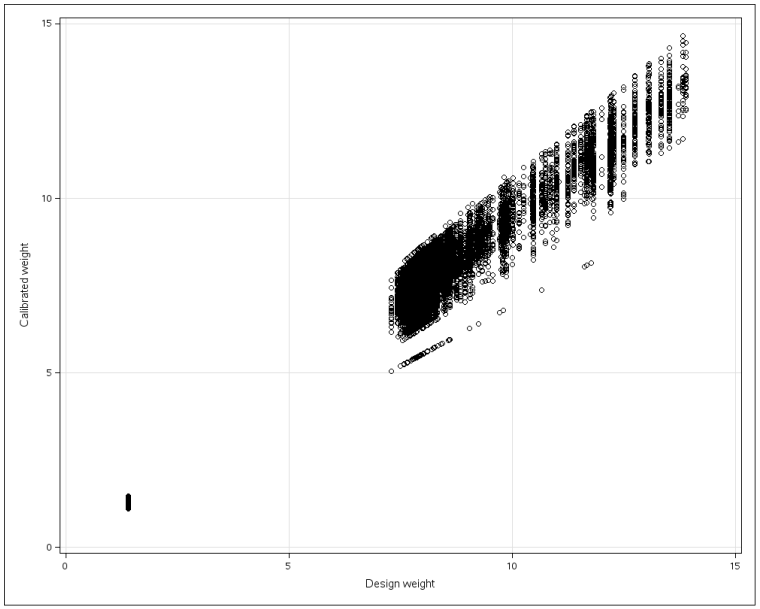
	Urban area/ Rural area	Sex	Age groups	Individual years of age	Individual years of age
	1,2	1,2	0-4, 5-9,..., 80-84, 85+	0, 1, . . . ,83, 84; 85+	0, 1, . . . ,98 99, 100+
Poland	1	1	1	1	0
Voivodeships	1	1	1	1	0
Poviats (without 5 biggest cities)	1	1	1	0	0
4 biggest cities	1	1	1	1	0
Warsaw	x	1	1	1	1
Districts of Warsaw	x	1	1	1	0
Districts of 4 biggest cities	x	1	1	1	0

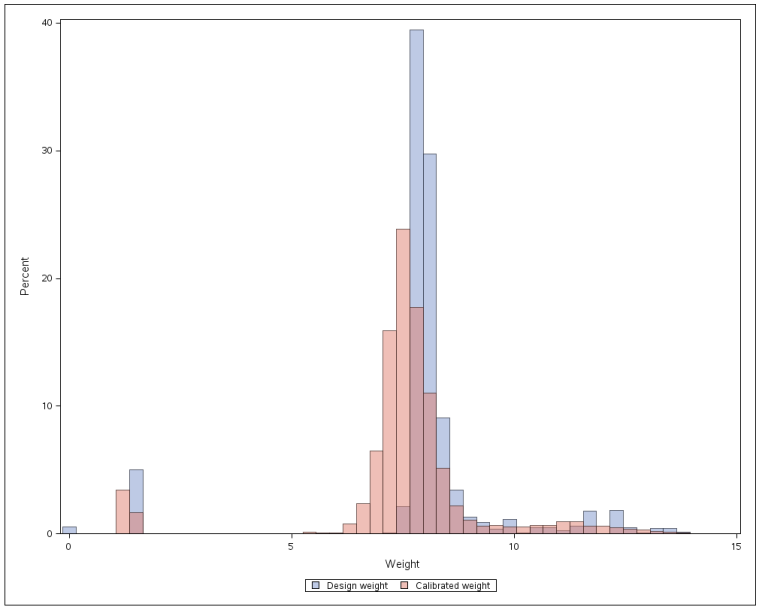
- **Legend:** 1–calibration possible, 0–calibration impossible, x–cross-section inadequate

Practical aspects of calibration in NCPH 2011

Poznanski poviat

Descriptive statistics					
Variable	Minimum	Maximum	Sum	Median	Std Dev
Design weights	1.3919308	13.8937500	350920.53	7.9896301	1.8675295
Calibrated weight	1.0884322	14.4946168	331525.00	7.5480397	1.8096110





Small Area Estimation in NCPH 2011

Small Area Estimation in NCPH 2011

1. **The main goal:** estimation of unemployment people in Poland at LAU 2 level of aggregation.
2. **Sources of data:** administartive registers and survey sample e.i. data from NCPH 2011 collected in so called „Golden Record".
3. **Estimators:** direct estimator, synthetic estimators, composite estimators (Rao 2003).

Gminas (2511) - LAU 2

Estimators

Estimators

- **Horvitz-Thompson estimator:**

$$Y_d^{HT} = \sum_{i \in s_d} y_i d_i \tag{21}$$

- **Simple synthetic estimator BARE – Broad Area Ratio Estimator:**

$$BARE_d^{no} = \frac{Y^{HT}}{N} \cdot N_d \tag{22}$$

- **Post-stratified synthetic estimator:**

$$BARE_d^{with} = \sum_g N_{d,g} \frac{Y_g^{HT}}{N_{.,g}} \tag{23}$$

- **Synthetic regression estimator:**

$$Y_d^{SYNT\_REG} = \beta_d X_d \tag{24}$$

Estimators

Estimators

- **Composite estimators:**

$$Y_d^{COMP\_i} = {}_i\gamma_d \cdot Y_d^{HT} + (1 - {}_i\gamma_d) \cdot SYNT_d \tag{25}$$

where:

- $i = 1$ then $\gamma_d = 0.5$
- $i = 2$ then $\gamma_d = \frac{n_d}{N_d}$
- $i = 3$ then $\gamma_d = \begin{cases} 1 & \text{for} \quad \hat{N}_d \geq \delta N_d \\ \frac{\hat{N}_d}{\delta N_d} & \text{for} \quad \hat{N}_d < \delta N_d \end{cases}$
- $i = 4$ then $\gamma_d = \begin{cases} 1 & \text{for} \quad \hat{N}_d \geq N_d \\ \left(\frac{\hat{N}_d}{N_d}\right)^{h-1} & \text{for} \quad \hat{N}_d < N_d \end{cases}$

and $SYNT_d$ could be equal to one of described above synthetic estimators e.i. $BARE_d^{no}$, $BARE_d^{with}$ or $Y_d^{SYNT\_REG}$.

Raking and variance estimation

Raking and variance estimation

- For composite and synthetic regression estimator, the estimates of unemployment for small areas do not add up to the direct estimate in poviat.
- A simple adjustment was needed in order to ensure coherence of estimates at different levels (Rao 2003).

$$\hat{Y}_d^{raking} = \frac{\hat{Y}_d}{\sum_d \hat{Y}_d} \cdot Y^{HT} \tag{26}$$

where $\hat{Y}_d^{raking}$ is adjusted estimator of total of unemployment people in d area (gmina) and $\hat{Y}_d$ is composite or synthetic regression estimator in d area (gmina).

- In order to estimate the variance of synthetic regression and composite estimators bootstrap was used (500 replications).

Chosen results

Chosen results

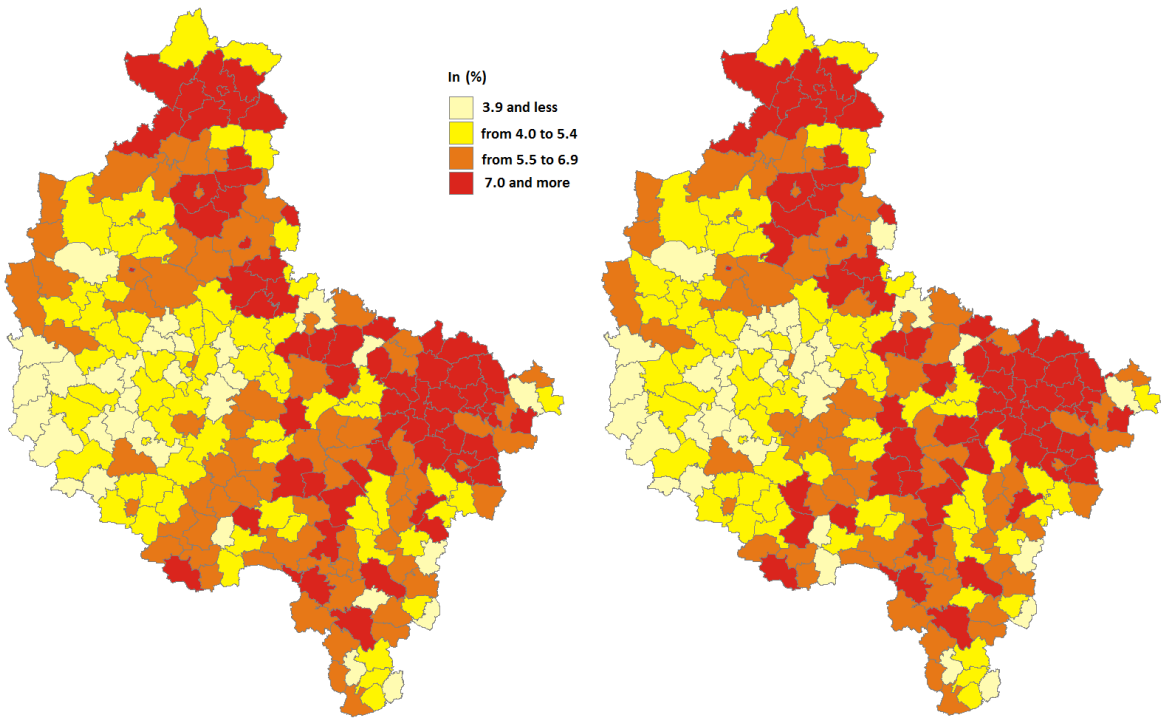
- Descriptive statistics of CV for gminas in wielkopolskie voivodeship

Estimator	N	Min	Max	Mean	S_x	Q_1	Q_2	Q_3
DIR_cv	230	2,89	22,73	9,40	3,38	7,46	9,10	11,23
SYNT_reg_cv	230	2,31	5,91	3,62	0,79	2,97	3,51	4,18
COMP_1_reg_cv	230	1,44	13,03	4,93	1,84	3,83	4,72	5,92
COMP_2_reg_cv	230	0,41	7,01	2,53	1,14	1,71	2,48	3,13
COMP_3_reg_cv	230	2,89	26,05	9,87	3,68	7,66	9,44	11,85
COMP_4_reg_cv	230	2,89	26,05	9,74	3,62	7,61	9,37	11,65

Chosen results

Percentage share of the unemployed in the total number of population in age 15 and more – gminas in wielkopolskie voivodeship

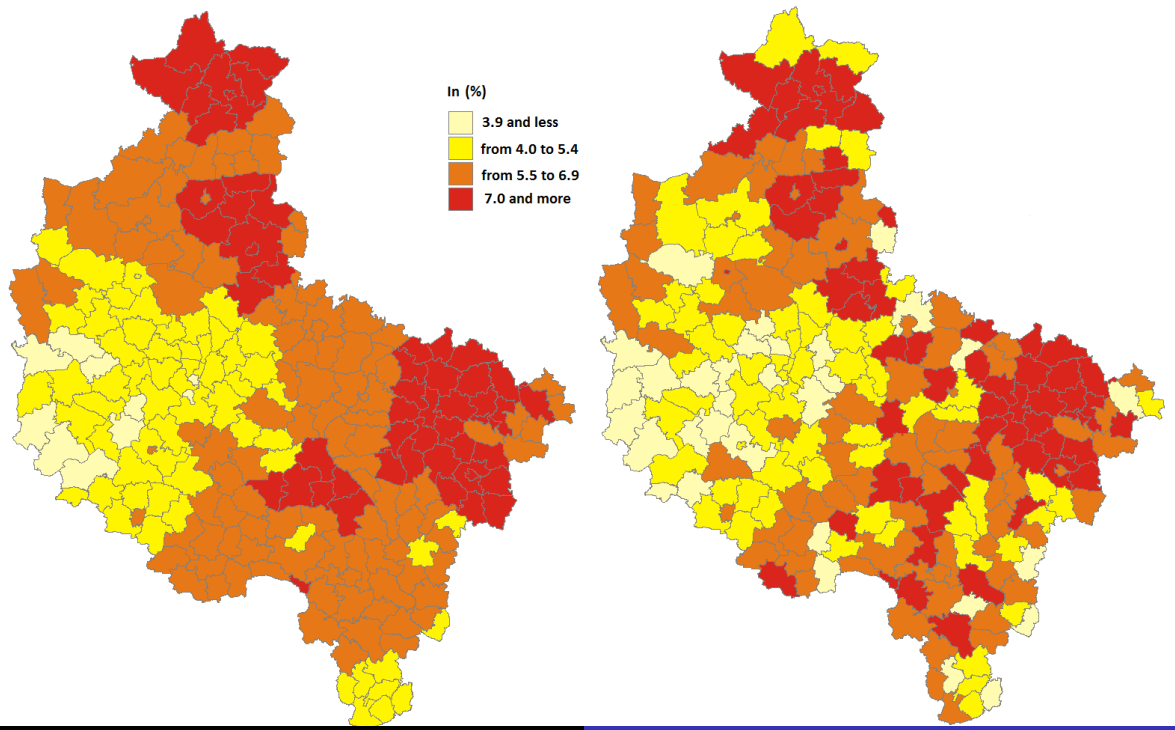
- **Direct estimator**
- **Synthetic ratio estimator with sex as a strata**



Chosen results

Percentage share of the unemployed in the total number of population in age 15 and more – gminas in wielkopolskie voivodeship

- **Synthetic regression estimator**
- **Composite estimator (synt_reg, γ_d for $i = 4$)**



Literature

Literature


Rao J.N.K (2003), „*Small Area Estimation*", Wiles Series in Survey Methodology, A John Wiley & Sons, INC., Publication.


Särndal C-E., Lundström S. (2005), „*Estimation in Surveys with Nonresponse*", John Wiley & Sons, Ltd.


Deville J-C., Särndal C-E. (1992), „*Calibration Estimators in Survey Sampling*", Journal of the American Statistical Association, Vol. 87, 376–382.


Särndal C-E. (2007), „*The Calibration Approach in Survey Theory and Practice*", Survey Methodology, Vol. 33, No. 2, 99–119.

Thank you very much for your attention!

Acknowledgments: Many thanks for support to all my colleagues (Ewa, Łukasz, Tomasz and Tomasz) from the Center for Small Area Estimation!