Spatial Microsimulation of Apparent and Inactive Unemployment in Poland

Wojciech Roszka, PhD [1] Kamil Wilak, PhD [1,2]

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

Mannheim, 2019.02.07-08

Outline

Survey Aims

Survey Aims

- Estimation of the level of apparent and inactive unemployment in Poland by NUTS-3 spatial aggregation level.
- The quantities to be estimated are the number of apparently and inactive unemployed and the corresponding percentages of the total number of persons registered in district labour offices (DLO).

Apparent and Inactive Unemployment

Apparent and Inactive Unemployment

Conditions of registration as unemployed in DLO

A person can register as unemployed if he or she:

(1) is not engaged in paid employment,

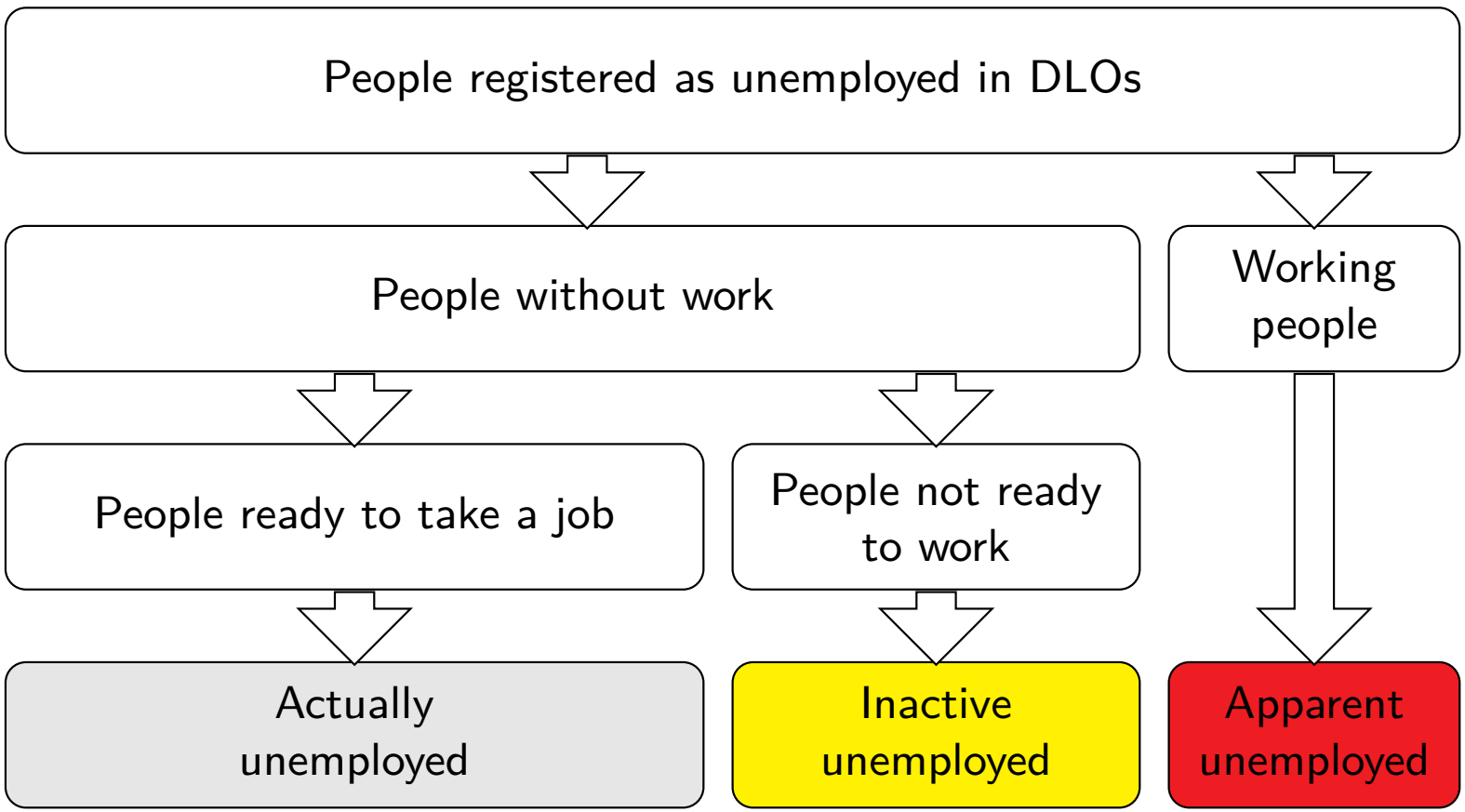
(2) is looking for a job,

(3) is ready to start work,

(4) and other.

Apparent and Inactive Unemployment

Definition of apparent and inactive unemployment



Apparent and Inactive Unemployment

Problem of apparent and inactive unemployment

- Apparent unemployment phenomenon is closely related with grey economy.
- In the informal sector we can distinguish two types of workers: those who are forced to work informally, because of the lack of alternatives, and those who choose to work outside the formal sector because of a higher marginal profit.
- Inactive unemployment phenomenon can by explainded by an income effect: increasing out-of-work income reduces incentives to start or continue employment.

Apparent and Inactive Unemployment

Problem of apparent and inactive unemployment

- In recent years there has been a discussion about the role of district labor offices in particular on the legitimacy of paying health insurance by these institutions, which is one of the main reasons why economically inactive people register as unemployed.
- District labor offices, instead of helping people who really want to find jobs, lose resources to serve people who are not interested in finding a job.
- Detailed knowledge on the level of apparent and inactive unemployment can be important for developing labor market policy.

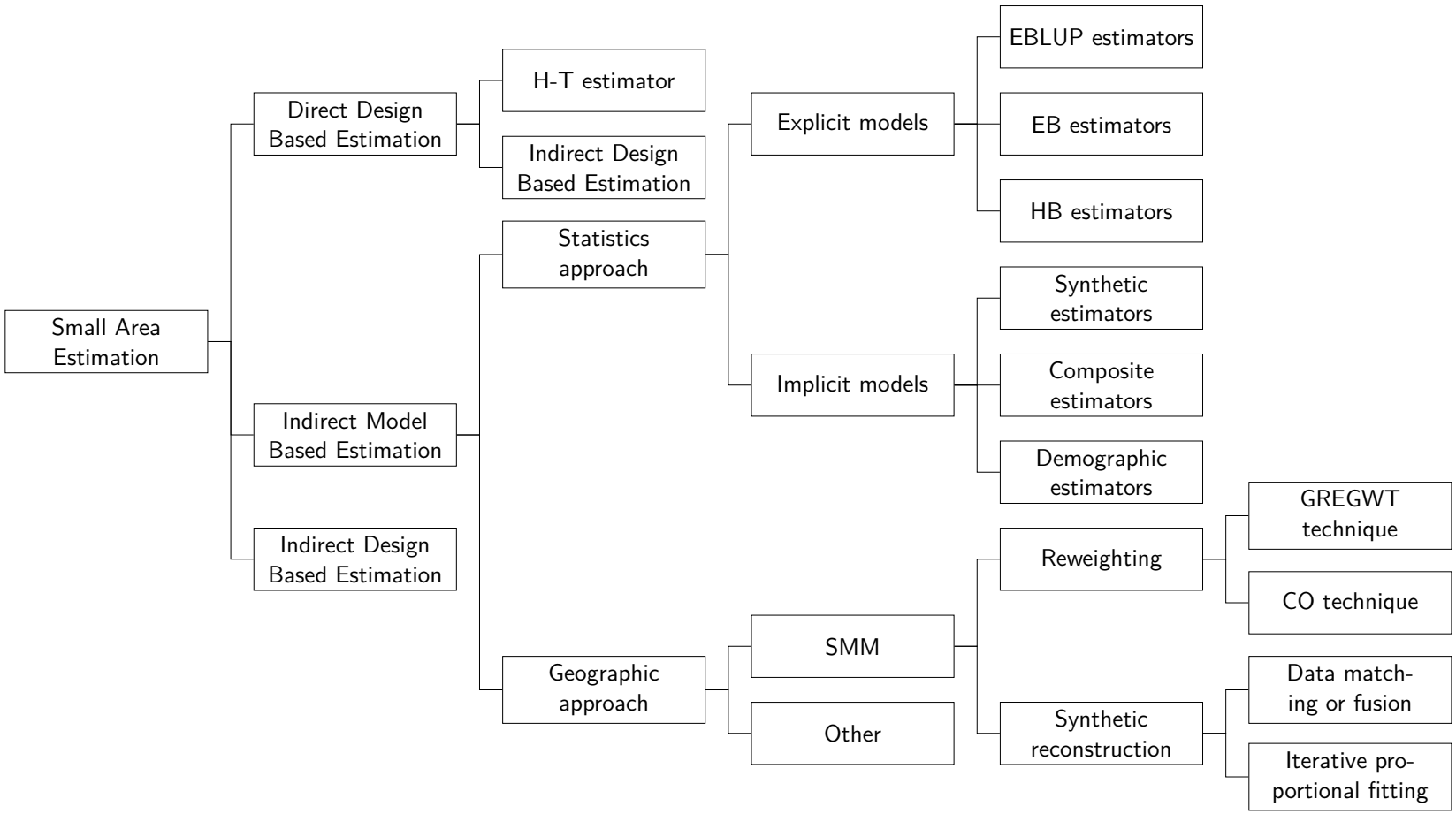
Data Sources

Data sources

- Labour Force Survey
  (microdata from 1^{st} quarter of 2011 – census reference period)
- National Census 2011
  (aggregated data, reference point – 31 March 2011):

Spatial Microsimulation Approach

Approaches in Small Area Estimation



Spatial Microsimulation

- The aim of SMM methods is a creation of dataset containing information on all units from a resulting population and a vector of many socio-economic characteristics
- The creation involves integration of sample survey microdata and small domain census constraints.
- Using different reconstruction and reweighing algorithms synthetic units are being created in such a way that the true distribution of a real population small geographical units is reflected.
- Having a multidimensional, full-coverage dataset not only small area estimation can be performed but flexible aggregation and disaggregation is possible.
- The key characteristics are then simulated in the obtained pseudo-population.

Synthetic dataset

Definition

Synthetic dataset is an anonymous microdata with appropriate variables and marginal and multivariate distributions that are at least *quasi*-identical to reality.

Components

1. Survey microdata file - provides comprehensive information on different characteristics of individual and/or households.
2. Small area constraints - aggregated census (or other reliable source) tables for small geographical areas.

Synthetic dataset

Creating rules

- The distribution of the synthetic population by region and stratum must be *quasi*-identical to the distribution of the true population.
- Marginal and joint distributions between variables – the correlation structure of the true population – must be accurately represented.
- Heterogeneities between subgroups, especially regional aspects, must be allowed.
- The records in the synthetic population should not be created by pure replication of units from the underlying sample, as this will usually lead to unrealistically small variability of units within smaller subgroups.
- Data confidentiality must be ensured.

Synthetic reconstruction

The approach consists of combining information from two types of sources of data.

Sources

1. Aggregated data in the form of **census (or administrative) tables** – provides the marginal distributions of relevant categorical socio- demographic variables covering the whole population of interest.
2. **Survey micro-dataset** representative of the population of interest – contains information on at least the same variables for a sample of individuals (referred as *seed*).

Synthetic reconstruction

The synthetic population dataset is generated using a two-step procedure:

Steps

1. **Estimation** – a joint distribution is estimated using both sources of data. The correlation structure of the seed should be preserved.
2. **Selection** – individuals are randomly selected from the seed dataset and added to the synthetic population so that the joint probabilities calculated in the previous step are respected.

The **iterative proportional fitting** (IPF) technique is commonly used in the estimation process.

Iterative proportional fitting

The IPF algorithm is an iterative procedure that fits an n-dimensional table with unknown entries to a set of known and fixed marginal distributions – **sample weights are calibrated according to known marginal population totals**.

Problem of estimating the population total $Y = \sum_{i=1}^{N} y_i$ for a finite population of size N.

- The weighted estimator is unbiased H-T estimator: $\hat{Y}_d = d_i y_i$, where $d_i = \frac{1}{\pi_i}$.
- If an auxiliary variable x is available from the sample with the condition that the population total $X = \sum_{i=1}^{N} x_i$ is known, usually $\hat{X} = \sum_{i=1}^{n} d_i x_i \neq X$
- Each individual in the sample dataset is given a probability of selection according to the original sampling weights and the expected number of similar individuals that need to be added to the synthetic population.
- Individuals are randomly selected from the sample until the expected number of persons in each group is reached.
- For each individual added to the synthetic population, all attributes – not only those controlled for in the first step – are automatically selected.

Model approach

Modelling the categorical target variable

- Due to need of preservation of variability, units cannot be simply replicated using weights.
- Categorical variables are conditionally drawn using multinomial regression estimates:

1. Simulated variable is selected from sample S. Independent variables must be present in both sample S and population U,

$$S = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,j} & x_{1,p+1} \\ x_{2,1} & x_{2,2} & \dots & x_{2,j} & x_{2,p+1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,j} & x_{n,p+1} \end{bmatrix}$$

   where $i = 1, \dots, n$ are sample units and $k = 1, \dots, j$ is the number of variable. X_1 to X_j is an independent variables vector and X_{p+1} is the target (dependent) variable.
2. The model is estimated in every small area using sample S units. In result β coefficients are obtained.

Model approach

Modelling the categorical target variable

3 The model is estimated in every small area using sample S units. In result β coefficients are obtained.

4 For every $i = 1, \ldots, N$ unit of the selected variable new outcome category is predicted. The conditional probability of selecting r-th category for each i-th $\hat{x}^*_{i,j+1}$ is:

$$\hat{p}_{i1} = \frac{1}{1+\sum_{r=2}^{R} exp(\hat{\beta}_{0r}+\hat{\beta}_{1r}\hat{x}_{i,1}+\cdots+\hat{\beta}_{jr}\hat{x}_{i,1})},$$

$$\hat{p}_{ir} = \frac{exp(\hat{\beta}_{0r}+\hat{\beta}_{1r}\hat{x}_{i,1}+\cdots+\hat{\beta}_{jr}\hat{x}_{i,1})}{1+\sum_{r=2}^{R} exp(\hat{\beta}_{0r}+\hat{\beta}_{1r}\hat{x}_{i,1}+\cdots+\hat{\beta}_{jr}\hat{x}_{i,1})},$$

where $r = 2, \ldots, R$ and $\hat{\beta}_{0r}, \ldots, \hat{\beta}_{jr}$ are the estimates of multinomial logistic regression model. The new $\hat{x}^*_{i,j+1}$ values are computed.

5 The population U is:

$$U = \begin{bmatrix} \hat{x}_{1,1} & \hat{x}_{1,2} & \cdots & \hat{x}_{1,j} & \hat{x}^*_{1,j+1} \\ \hat{x}_{2,1} & \hat{x}_{2,2} & \cdots & \hat{x}_{2,j} & \hat{x}^*_{2,j+1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \hat{x}_{N,1} & \hat{x}_{N,2} & \cdots & \hat{x}_{N,j} & \hat{x}^*_{N,j+1} \end{bmatrix}.$$

Such an approach minimizes appearance of so-called *random zeroes* (domains that exist in the population but didn't occur in the sample).

Model approach

Modelling the continuous target variable

For continuous variables one of suggested approaches (Templ *et al.* 2017) involves as follows:

1. Dependent x_{j+1} in discretized in y_{j+1} by creating R cutoff values $c_1 \leq \ldots \leq c_R$:

$$y = \begin{cases} 1 & \text{if } c_1 \leq x_{ij} < c_2, \\ 2 & \text{if } c_2 \leq x_{ij} < c_3, \\ \vdots & \vdots \\ , R & \text{if } c_{R-1} \leq x_{ij} \leq c_R. \end{cases}$$

2. Multinomial logistic regression model (the same as for categorical variables) is estimated with the dependent variable y_{j+1} and the independent variables vector $x_1, x_2, \ldots, x_j$ for each k-th domain (small area) separately.
3. Within each r-th class estimates $\hat{x}$ are drawn from a uniform distribution with boundaries of classes as parameters. The exception is the last class where due to outliers values are drawn using generalized Pareto distribution:

$$\hat{x}^*_{i,j+1} \approx \begin{cases} U(c_r, c_{r+1}) & \text{if } \hat{y}_i = r \text{ and } 1 \leq r \leq R-1, \\ GPD(\mu, \sigma, \xi, x) & \text{if } \hat{y}_i = R. \end{cases}$$

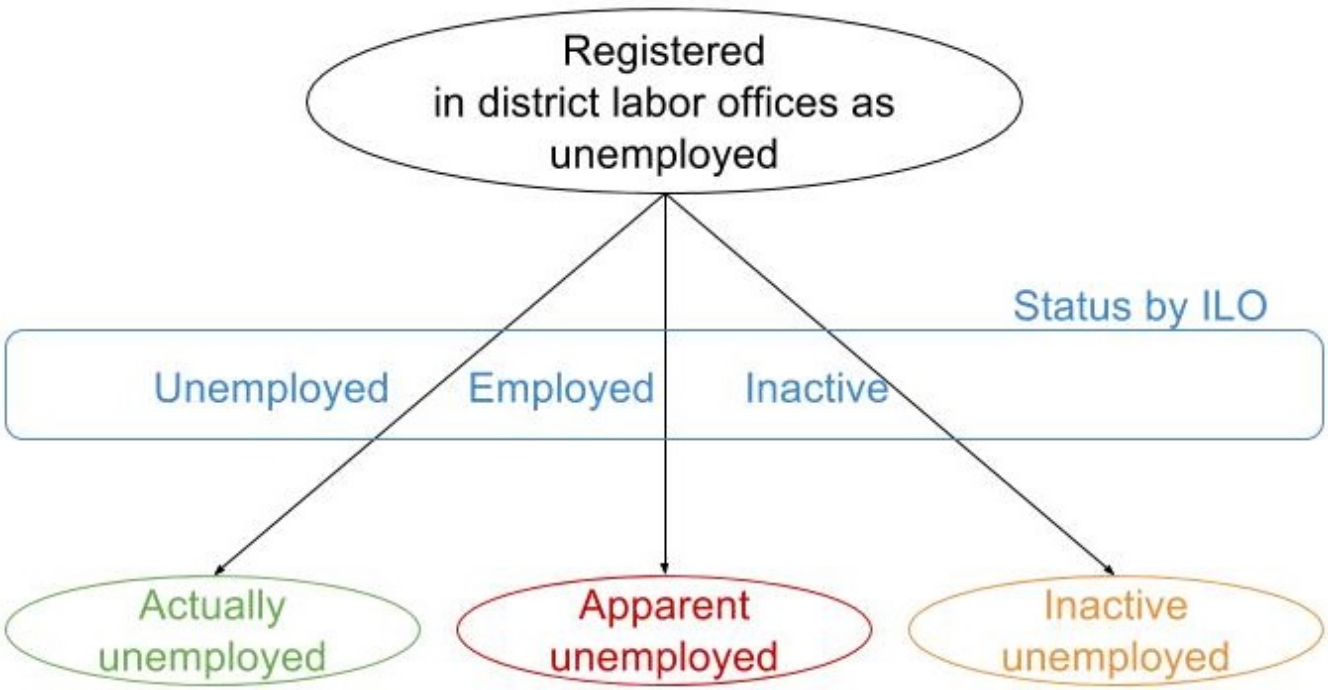
Research Procedure

Research procedure

1. Apparent and inactive unemployed were identicated in LFS microdata.
2. Original survey weights were calibrated using population distribution for NUTS3 $\times$ sex $\times$ five-year age groups.
3. The population was allocated to small areas using IPF algorithm - **a synthetic dataset of 32 mln units was created**.
4. The multinomial logistic regression model was estimated.
5. Labour force status categories were conditionally drawn using multinomial regression estimates.
6. Estimates were assessed in terms of:
   - Estimation effectiveness.
     - 6.1 Estimation process was conducted 1000 times.
     - 6.2 For each iteration the percentage of apparent and inactive unemployment in each sub-region was estimated.
     - 6.3 Means and standard deviations were calculated and $CV_j = \frac{\hat{\sigma}_j}{\hat{\mu}_j}$ for each j-th sub-region was calcutated.
   - Estimation bias - obtained estimates were compared with direct estimators at NUTS3 level.

Apparent and Inactive Unemployment

Identification of apparently and inactive unemployed



Auxiliary variables

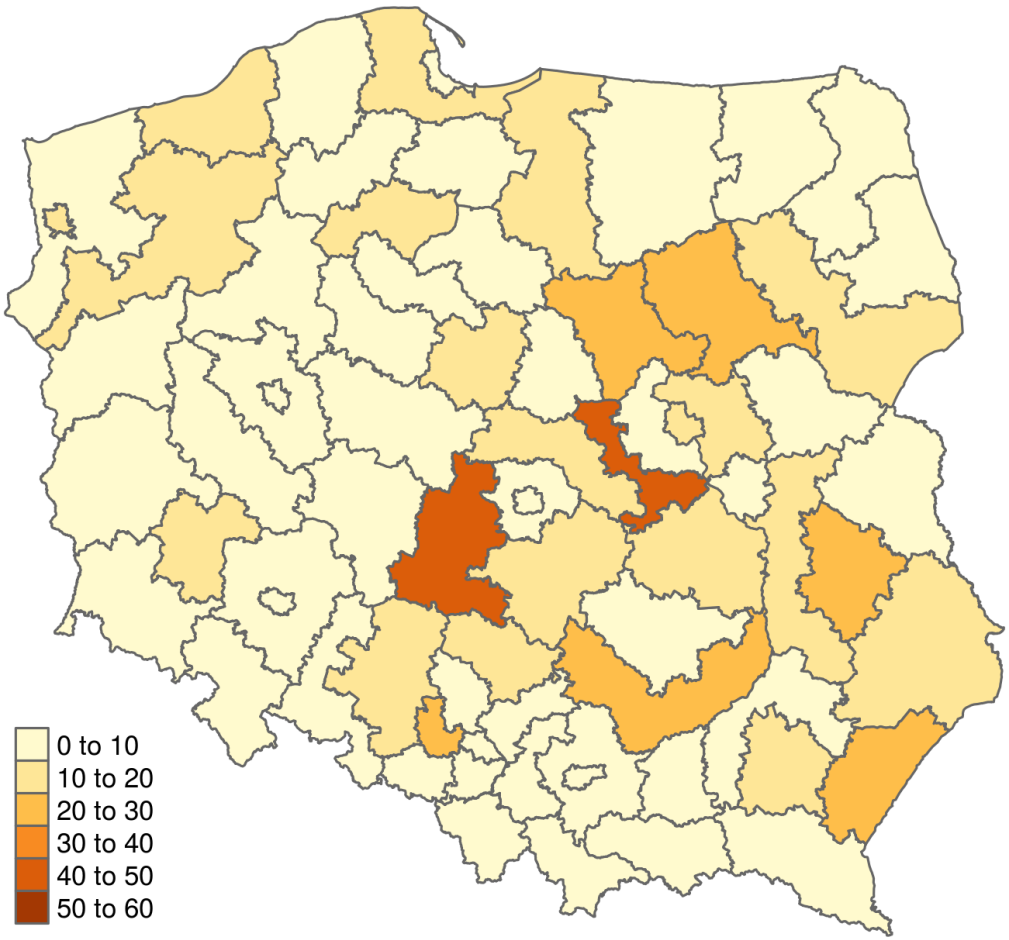
Table: Independent vars in model

Variable	**Categories**
Sex	male; female
Age	5 years groups, 65+
Marital status	bachelor, miss; married; widower, widow; divorced, separated
Class of place of residence	cities 100k+; 50k - 100k;20k - 50k; 10k - 20k; 5k - 10k; 2k - 5k; >2k; village
Disability level	a severe degree of disability; a moderate degree; a slight degree; no disability
Source of income	employees; farmers; workers using a farm; self-employed; pensioners; others
Education level	ISCED levels
Household size	number of household's members

Results

Results

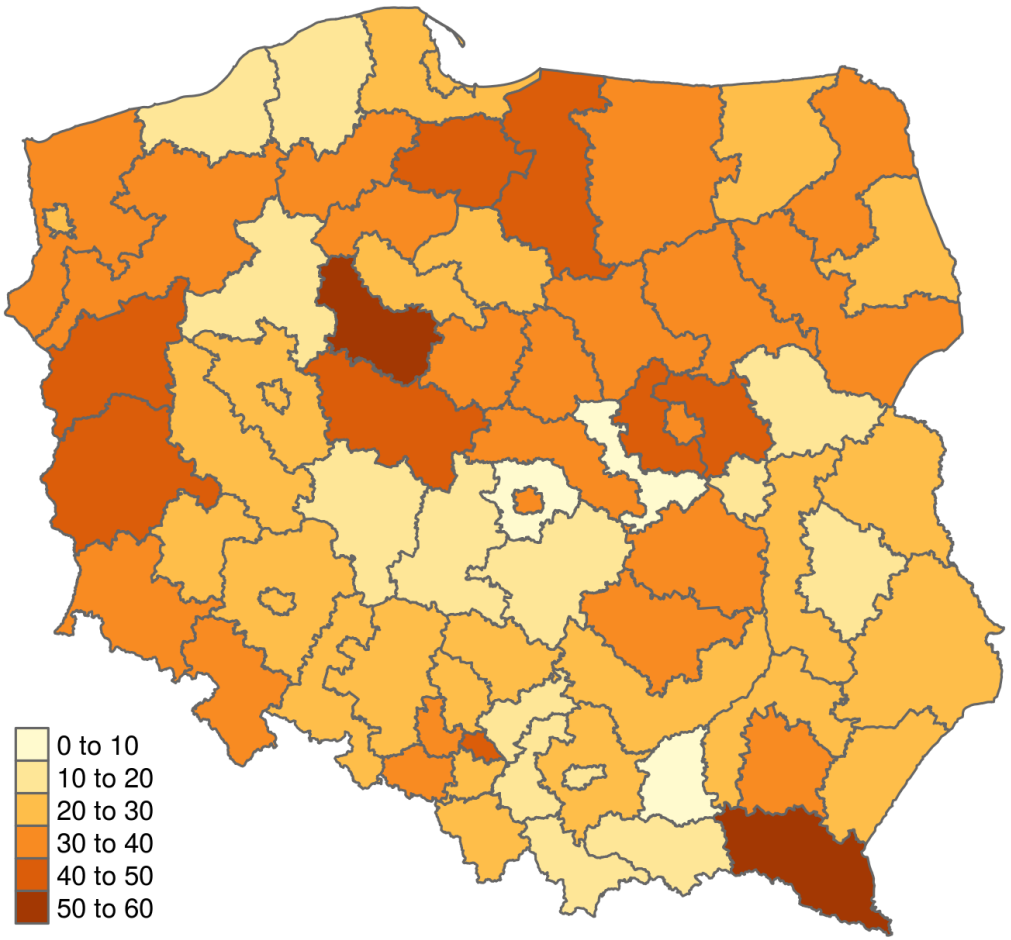
Figure: Fraction of apparently unemployed among registered as unemplyed in DLOs by NUTS-3 level, 1st quarter 2011.



Source: Own study

Results

Figure: Fraction of inactive unemployed among registered as unemplyed in DLOs by NUTS-3 level, 1st quarter 2011.



Source: Own study

Results

Table: Distribution characteristics of fractions of apparently and inactive unemployed among persons registered in DLOs by NUTS-3 level, 2nd quarter 2011

Unempl. type	Min	Q1	Q2	Q3	Max	Mean	Sd
apparent	0.9	5.3	8.7	12.4	46.9	10.8	8.9
inactive	6.7	21.3	27.4	36.3	53.5	28.9	10.5

Source: Own study

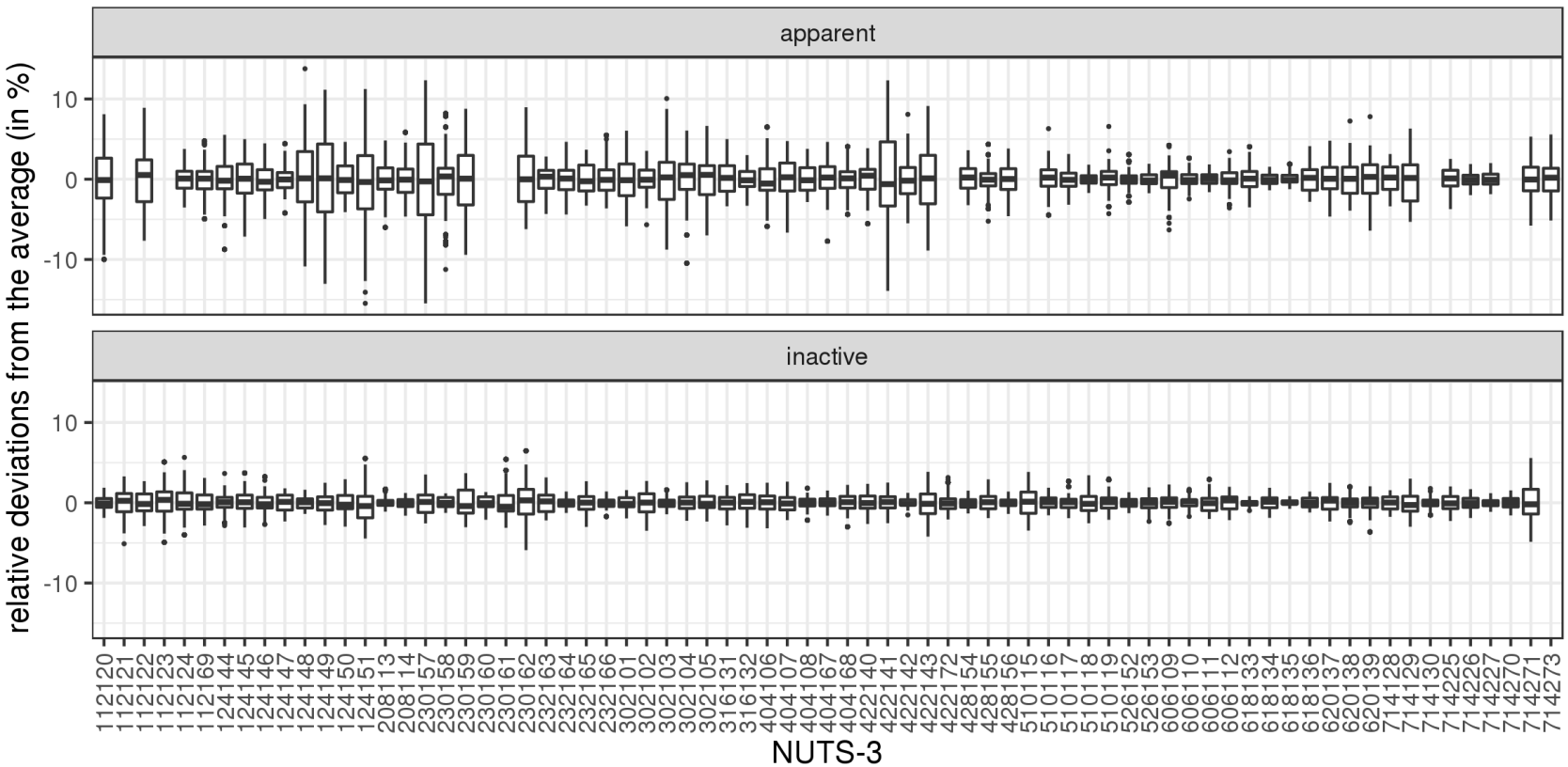
Table: Distribution characteristics of variation coefficients (in %) of fraction estimates

Unempl. type	Min	Q1	Q2	Q3	Max	Mean	Sd
apparent	0.6	1.5	2.0	2.6	6.2	2.3	1.2
inactive	0.3	0.7	1.0	1.3	2.3	1.0	0.4

Source: Own study

Results

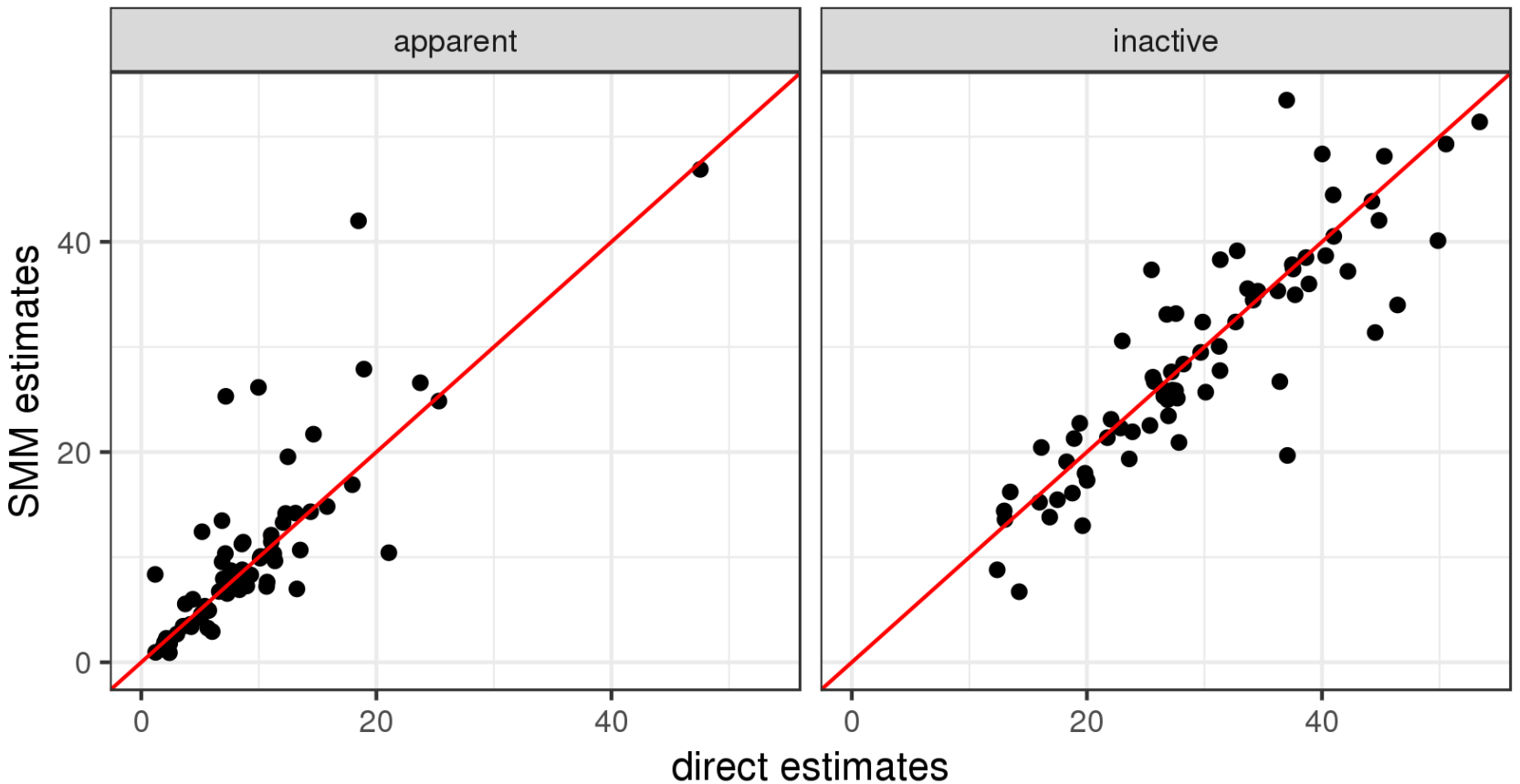
Figure: Relative deviations from the mean, 1000 iteration of SMM.



Source: Own study

Results

Figure: SMM estimates vs direct estimates of fraction apparently and inactive unemployed among registered as unemplyed in DLOs by NUTS-3 level, 1st quarter 2011.



Source: Own study

Conclusions and Future Directions of Research

Conclusions and Future Directions of Research

- The applied microsimulation method enabled to estimate apparent and inactive unemployment at a low level of spatial aggregation.
- Significant share of people registered in district labor offices work in the informal economy, or do not want / are not ready to start employment.
- Apparent and inactive unemployment are characterized by a spatial variation.
- There is a need to develop a method for assessing estimates obtained by spatial microsimulations.
- Data from registered unemployment will be used in the estimation process.

Selected Literature

Selected Literature

- Czapiński J. Panek T. (2014). "Wykluczenie społeczne. In Diagnoza Społeczna 2013 - Warunki i jakość życia Polaków", Warszawa
- Rahman A. (2009), "Small Area Estimation Through Spatial Microsimulation Models: Some Methodological Issues", University of Canberra, 2nd International Microsimulation Association Conference, Ottawa, Canada
- Rahman A., Harding A., Tanton R., Liu S. (2010), "Methodological Issues in Spatial Microsimulation Modelling for Small Area Estimation", International Journal of Microsimulation (2010) 3(2)
- Rahman A., Harding A. (2017), "Small Area Estimation and Microsimulation Modelling", CRC Press
- Tanton R., Edwards K. L. ed. (2013), "Spatial Microsimulation: A Reference Guide for Users", Springer
- Tanton R. (2014), "A Review of Spatial Microsimulation Methods", International Journal of Microsimulation (2014) 7(1)
- Wilak K. (2018), "Estymacja bezrobocia pozornego i biernego z wykorzystaniem strukturalnych modeli szeregów czasowych. PhD thesis", Uniwersytet Ekonomiczny w Poznaniu, Poznań