Mapping poverty for the agglomeration of Poznań using methods of indirect estimation

Grzegorz Grygiel¹  Tomasz Józefowski¹  Tomasz Klimanek¹  Łukasz Wawrowski²

¹Centre for Small Area Estimation
Statistical Office in Poznan

²Department of Statistics
Poznan University of Economics and Business

6 June 2018
Presentation outline

Motivation

Poverty indicators

Data sources

Poznan agglomeration

The approach

Results

Measures of precision

Conclusions
Motivation

- demand for information concerning the quality of life, especially poverty, for low-level unplanned domains, such as the Poznan Agglomeration
- estimates at lower levels of the territorial division than regions (NUTS 1) or provinces (NUTS 2) have not been published so far
- such estimates can be calculated by using indirect estimation methods, which rely on out-of-sample data, and tend to improve estimation precision
- opportunity to draw on CSAE’s experiences of poverty mapping:
  - for subregions - a project with the World Bank (2013)
  - for districts - an EU-funded POPT project (2015)
Poverty indicators

- **at-risk-of-poverty rate (head count ratio (HCR)), poverty incidence** - the share of people with an equivalised disposable income (after social transfers) below the poverty threshold, (60 % of the national median EDI)

- **relative median at-risk-of-poverty gap** - the difference between the median EDI of people below the poverty threshold

- **Gini coefficient** - measure of income distribution inequality: ranges from 0 (maximal equality) to 1 (maximal inequality)

- **inequality of income distribution, income quintile share ratio (QSR), S80/S20 ratio** - the ratio of total income received by the richest top quintile of the population to that received by the bottom poorest quintile
### Poverty indicators

**Table 1:** Poverty indicators for NTS 1 regions and Poland based on EU-SILC 2015

<table>
<thead>
<tr>
<th>NTS1 region</th>
<th>Gini</th>
<th>HCR</th>
<th>QSR</th>
<th>Pov-Gap</th>
</tr>
</thead>
<tbody>
<tr>
<td>central</td>
<td>34.0</td>
<td>16.5</td>
<td>5.7</td>
<td>-</td>
</tr>
<tr>
<td>southern</td>
<td>29.0</td>
<td>14.5</td>
<td>4.6</td>
<td>-</td>
</tr>
<tr>
<td>eastern</td>
<td>28.5</td>
<td>22.0</td>
<td>4.4</td>
<td>-</td>
</tr>
<tr>
<td>north-western</td>
<td>29.2</td>
<td>17.6</td>
<td>4.6</td>
<td>-</td>
</tr>
<tr>
<td>south-western</td>
<td>30.1</td>
<td>14.4</td>
<td>4.8</td>
<td>-</td>
</tr>
<tr>
<td>northern</td>
<td>29.7</td>
<td>20.5</td>
<td>4.7</td>
<td>-</td>
</tr>
<tr>
<td>Poland</td>
<td>30.6</td>
<td>17.6</td>
<td>4.9</td>
<td>22.3</td>
</tr>
</tbody>
</table>
The European Union Statistics on Income and Living Conditions (EU-SILC)

- two-stage stratified sampling with proportional allocation: PSU - census output areas (stratified), SSU - households
- first sample in 2005 - 24,000 HHs, divided into 4 subsamples, each year one subsample replaced with a new one
- results published for domains defined by place of residence (urban/rural), locality size class, NTS1 regions, socio-demographic variables
The National Census of Population and Housing (2011)

- mixed mode census - registers + 20% sample
- **information available at NUTS2 level (provinces):**
  - economic activity by age group
  - unemployed people by duration of job search
  - economically inactive by cause of inactivity
  - employed people by employment status
- **information available at NUTS4 level (districts):**
  - economic activity by sex - level of education
  - marital status
  - sources of income
  - disability
**Table 2:** Number of households in the two datasets Wielkopolskie province and Poznań agglomeration

<table>
<thead>
<tr>
<th>Dataset</th>
<th>province</th>
<th>agglomeration</th>
</tr>
</thead>
<tbody>
<tr>
<td>EU-SILC 2015</td>
<td>821</td>
<td>173</td>
</tr>
<tr>
<td>Census 2011</td>
<td>228 632</td>
<td>40 840</td>
</tr>
</tbody>
</table>
Datasets and variables (2)

Household variables used in the model

- equivalised disposable income
- share of men
- share of people aged 65 or older
- share of unemployed persons
- share of disabled persons
- share of people with primary education
- share of people with higher education
- ratio of children (aged <15) to persons aged 16-65
- binary – HH in rural area or in a municipality < 20000
- binary – HH with 1 room
- binary – HH with 3 or more rooms
Poznan agglomeration
Poznan agglomeration

- created in 2007 as a way of fostering cooperation between the city of Poznan and 17 surrounding municipalities (the Poznan district)
- later on extended by including 4 municipalities from outside the Poznan district
- 11% of the province’s area
- 30% of the total population (appr. 1 million)
- 40% of all companies registered in the province
- provide employment to 40% of all employees
The approach

1. estimate a unit-level model for the province of wielkopolskie (unit - household) with random effect $u_d$ at district level (Poznan agglomeration is treated as one district)

2. using the model parameters, apply the EB method to predict values of the dependent variable $Y_{dj}$: equivalised disposable income in the household for out-of-sample units
The linear mixed model

\[ Y_{dj} = x_{dj}^T \beta + u_d + e_{dj}, \quad j = 1, \ldots, N_d, \quad d = 1, \ldots, D, \]

where:

- \( Y_{dj} \) - Box-Cox transformed income of \( j \)-th HH in district \( d \)
- \( x_{dj}^T \) - vector of independent variables for \( j \)-th HH in district \( d \)
- \( \beta \) - vector of regression parameters
- \( u_d \) - random area effect with \( u_d \sim iid N(0, \sigma_u^2) \)
- \( e_{dj} \) - residual errors \( e_{dj} \sim iid N(0, \sigma_e^2) \)

Calculations made using the \texttt{lmer} function from the \texttt{lme4} package in R.
1. fit a linear mixed model describing the household income based on (transformed) data from a representative survey - $y_s$

2. draw $L$ out-of-sample vectors $y_r$ but with the unknown parameters replaced by the estimators obtained in (1) - model parameters and random effects variances

3. With the $L$ generated vectors of theoretical values and using the sample data $y_s$, compute EBPs of the poverty measures for unit-level auxiliary data from the census using the Monte Carlo approximation (minimizing MSE under the model).

4. estimate MSE using parametric bootstrap

Calculations made using the `ebp` function from the `emdi` package in R (key parameters: $L = 500$, $B = 200$)
Results: Gini coefficient & Quintile share ratio
Results: HCR (poverty rate) & Poverty gap

HCR
16%
24%
33%

Pov-Gap
4.0
5.4
6.8
8.2
9.6
11.0

16.4%
4.5%
Measures of precision

Parametric bootstrap for MSE estimation - approximated through a Monte Carlo procedure repeated B times and then taking the mean over the B replicates.

\[
\text{mse}(\hat{Y}_d^{EB}) = \frac{1}{B} \sum_{b=1}^{B} \left( \hat{Y}_d^{EB*}(b) - Y_d^{(b)*} \right)^2
\]

Relative root mean square error (CV) (%):

\[
RRMSE(\hat{y}) = \frac{\text{RMSE}(\hat{y})}{\hat{y}} \times 100
\]

\(\hat{y}\) - estimate of a poverty measure
Root mean square error

- Gini
- QSR
- HCR
- Pov_Gap

<table>
<thead>
<tr>
<th>estimator</th>
<th>direct</th>
<th>indirect</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Relative root mean square error (CV) (%)

- Gini
- QSR
- HCR
- Pov_Gap

estimator: direct, indirect

RRMSE (CV) (%)
Conclusions

- methods of indirect estimation make it possible to estimate poverty for unplanned domains (when sample size is small or even zero)
- further work is required to select the most optimal explanatory variables for the model (especially variables from administrative registers, including the tax register)
- estimation should be conducted for subsequent years to verify the model’s stability over time
GUS, (2017), *Income and living conditions of the population of Poland (report from the EU-SILC survey if 2015)*, Warszawa


Wawrowski, Ł. (2017), Indirect estimation of poverty at the regional and local level in Poland, (doctoral dissertation)
Thank you for your attention.