Limits of data disaggregation in household surveys for population subgroups and geographical areas and the requirements to overcome them

Application to poverty mapping in Palestine

Isabel Molina and Eduardo García
Dept. of Statistics, Univ. Carlos III de Madrid
Palestinian Expenditure Consumption Survey (PECS) 2016/17

- Sample size: $n = 18,363$ persons out of $N = 4,266,953$ (43 out of 10,000)

- Sample sizes of regions by gender are fine:

<table>
<thead>
<tr>
<th></th>
<th>Gaza</th>
<th>West Bank</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Women</td>
<td>2569</td>
<td>6550</td>
<td>9119</td>
</tr>
<tr>
<td>Men</td>
<td>2578</td>
<td>6666</td>
<td>9244</td>
</tr>
<tr>
<td>Total</td>
<td>5147</td>
<td>13216</td>
<td>18363</td>
</tr>
</tbody>
</table>
Palestinian Expenditure Consumption Survey (PECS) 2016/17

- What if we wish to estimate at local level?
- 315 localities in census: 162 in PECS, 157 unsampled.
- Sample sizes localities by gender:

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>Q1</th>
<th>Median</th>
<th>Mean</th>
<th>Q3</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Women</td>
<td>14</td>
<td>26</td>
<td>35</td>
<td>56.29</td>
<td>61.5</td>
<td>405</td>
</tr>
<tr>
<td>Men</td>
<td>13</td>
<td>28</td>
<td>36</td>
<td>57.06</td>
<td>63</td>
<td>464</td>
</tr>
</tbody>
</table>
BASIC TERMINOLOGY

• **Areas/domains:** Subpopulations of interest.

• **Direct estimator:** Based *only* on the survey data *from* the target area/domain.

• **Small area:** Area/domain for which the considered *direct* estimator of the target indicator is *inefficient* (too large sampling error).
The idea is *borrowing strength* from the other areas.

Use auxiliary data sources (census or other, ideally a census) that contains some variables related to our target variable and observed also in the survey.

Consider that the target variable is related with the auxiliary variables *similarly* for all the areas (regression model).

Include (random) area effects to account for unexplained between-area heterogeneity (mixed regression model).
MODEL-BASED ESTIMATION

- Fit the model with the survey data from all the areas.
- Total survey sample size is typically large, so borrowing a lot of strength.
- Use the fitted model to estimate in the small areas.
- Efficiency gains are typically great.
POVERTY INDICATORS

- $D$ areas/domains ($d = 1, \ldots, D$) of sizes $N_1, \ldots, N_D$.
- $E_{dj}$ welfare measure for indiv. $j$ in domain $d$.
- $z =$ poverty line.
- **FGT poverty indicator of order $\alpha$ for domain $d$:**

$$F_{\alpha d} = \frac{1}{N_d} \sum_{j=1}^{N_d} \left( \frac{z - E_{dj}}{z} \right)^{\alpha} I(E_{dj} < z), \quad \alpha \geq 0.$$ 

- When $\alpha = 0 \Rightarrow \text{Poverty rate}$ (or at-risk-of-poverty rate)
- When $\alpha = 1 \Rightarrow \text{Poverty gap}$

✓ Foster, Greer & Thornbecke (1984), Econom.
NESTED ERROR MODEL

• In our application, as welfare measure of individuals we use household expenditure per adult equivalent.
• The distribution of expenditures $E_{dj}$ is highly right skewed.
• We need to transform expenditures so that the distribution is approximately Normal: $y_{dj} = \log(E_{dj} + k)$
• We consider a nested error model for $y_{dj}$:

$$y_{dj} = x'_{dj} \beta + u_d + e_{dj}, \quad j = 1, \ldots, N_d, \quad d = 1, \ldots, D$$

$$u_d \overset{iid}{\sim} N(0, \sigma_u^2), \quad e_{dj} \overset{iid}{\sim} N(0, \sigma_e^2)$$

• We obtain the empirical best (EB) predictor of the target indicator for each area of interest.

(✓ Molina & Rao (2010), CJS)
DATA DESCRIPTION

• **Data:** Palestinian Expenditure Consumption Survey (PECS) from 2016/2017 and Population Census from 2017.

• **Target:** Estimate poverty rates and gaps for Palestinian localities by gender.

• **Areas:** In census, 319 localities $\rightarrow D = 162$ in survey. We compute estimates for each sampled locality by gender.

• **Welfare measure:** $E_{dj}$ monthly expenditure per adult equivalent (ILS).

• **Poverty line:** $z = 10,027$ ILS $\rightarrow$ approx. 26% popn. below pov. line.
**FITTED MODEL**

- We fit a separate model for each gender.
- **Explanatory variables:**
  - Indicators of region (Gaza, West Bank), type of locality (rural/urban, camp).
  - Household characteristics (size, prop. females, employed ratio).
  - Household head characteristics (unemployed, employisrasett, employnatgov, refugstat, diff, neverschool, secondabove).
  - Dwelling characteristics (type, tenure, num. rooms).
  - Supplies (water, waste, heating systems, freezer, etc.)
MODEL CHECKING

- Model coefficients take reasonable signs.
- All covariates with significant categories for both genders.
- **Explanatory power**: $R^2 = 53.6\%$, both genders.
- Data indicates nothing against normality of model residuals, linearity, heteroscedasticity. Model seems to fit well.
COMPARISON BY REGION

✓ Median Pov. Rate: Gaza 55%, West Bank: 8.3%
✓ Median Pov. Gap: Gaza 17.4%, West Bank: 1.5%

Poverty Rate

Poverty Gap
QUALITY EB vs. DIRECT: POV. RATE

✓ Median MSE Women: Direct 47, EB: 6.7
✓ Median MSE Men: Direct 45.8, EB: 5.5
MSE EB vs. DIRECT POV. RATE: WEST BANK

✓ Reduction in all but one locality, 84% average MSE reduction!
MSE EB vs. DIRECT POV. RATE: GAZA

✓ Great gains also for Pov. Gap (not shown)!
ESTIMATED POV. RATE: WOMEN

West Bank

Gaza

Locality (sample size in labels)

Locality (samp. size in labels)
EB POV. RATE BY GENDER: WEST BANK
EB POV. RATE BY GENDER: GAZA
EB POV. GAP BY GENDER: WEST BANK
EB POV. GAP BY GENDER: GAZA

![Graph showing the estimated poverty gap by gender in Gaza, with data points for both women and men across different localities.](image-url)
CONCLUSIONS

- The use of census data in a model allows us to obtain much more efficient estimates.
- The considered model-based methodology allows to disaggregate at any desired level.
- We can estimate whatever indicator that is function of expenditure.
- The considered model fits rather well these data.
- The efficiency gains of model-based estimators with respect to direct estimators are notorious (over 82% reduction in MSE for pov. rates and gaps).
CONCLUSIONS

• **Direct** estimates equal to **zero** for many localities (32 for Men, 29 for Women) and **highly unstable**.

• **EB** estimates **never zero** and much more stable. Perhaps some underestimation in few localities, model variations can be further explored.

• Gaza has **much larger** pov. rates and gaps. Perhaps using a different pov. line.

• No great differences between men and women, although women with slightly greater estimates for about 70% of localities in West Bank.
✓ MANY THANKS TO UN-ESCWA AND PCBS FOR GREAT DATA PREPARATION!

✓ THANK YOU ALL FOR YOUR ATTENTION!