A short guide for small area estimation in household surveys

Illustration to poverty mapping in Palestine with expenditure survey and census data

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:

	Gaza	West Bank	Total
Women	2569	6550	9119
Men	2578	6666	9244
Total	5147	13216	18363

Palestinian Expenditure Comsumption 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:

	Min	Q1	Median	Mean	Q3	Max
Women	14	26	35	56.29	61.5	405
Men	13	28	36	57.06	63	464

SMALL AREA ESTIMATION

- **Areas/domains:** Subpopulations of interest (Example: Localities $\times$ gender).
- **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 has **unacceptable** sampling error.

INDIRECT ESTIMATION

- 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 to 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 can be substantial.

NOTATION

- Num. areas/domains: D
- Area popn. sizes: $N_1, \ldots, N_D$
- Area index: $d = 1, \ldots, D$
- Unit index: $j = 1, \ldots, N_d$
- **Welfare** measure for indiv. j in domain d: E_{dj}
  (Example: monthly household expenditure per adult equivalent)
- Poverty line: z

POVERTY INDICATORS

- Poverty **rate**:

$$F_{0d} = \frac{1}{N_d} \sum_{j=1}^{N_d} \underbrace{I(E_{dj} < z)}_{= \begin{cases} 1 \text{ if } E_{dj} < z; \\ 0 \text{ otherwise} \end{cases}}$$

- Poverty **gap**:

$$F_{1d} = \frac{1}{N_d} \sum_{j=1}^{N_d} \underbrace{\left(\frac{z - E_{dj}}{z} \right)}_{\text{Rel. distance to pov. line}} I(E_{dj} < z)$$

- In general, **FGT** indicators: For $\alpha \geq 0$,

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

✓ *Foster, Greer & Thornbecke (1984), Econometrika.*

EB ESTIMATORS

- The distribution of expenditures E_{dj} is highly right skewed.
- We need to take some transformation to achieve Normality: $y_{dj} = \log(E_{dj} + k)$
- We consider a model with random area effects for y_{dj}:

$$y_{dj} = \mathrm{x}'_{dj}\boldsymbol{\beta} + u_d + e_{dj}, \quad j = 1, \ldots, N_d, \; d = 1, \ldots, D$$
$$u_d \overset{iid}{\sim} N(0, \sigma_u^2), \quad e_{dj} \overset{iid}{\sim} N(0, \sigma_e^2)$$

- We obtain the **empirical best** (EB) predictor of the target indicator for each area of interest:

$$\hat{F}_{\alpha d}^{EB} = E[F_{\alpha d}|\mathrm{y}_s; \hat{\boldsymbol{\theta}}]$$

✓ *Molina & Rao (2010), Can. J. Stat.*

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.)

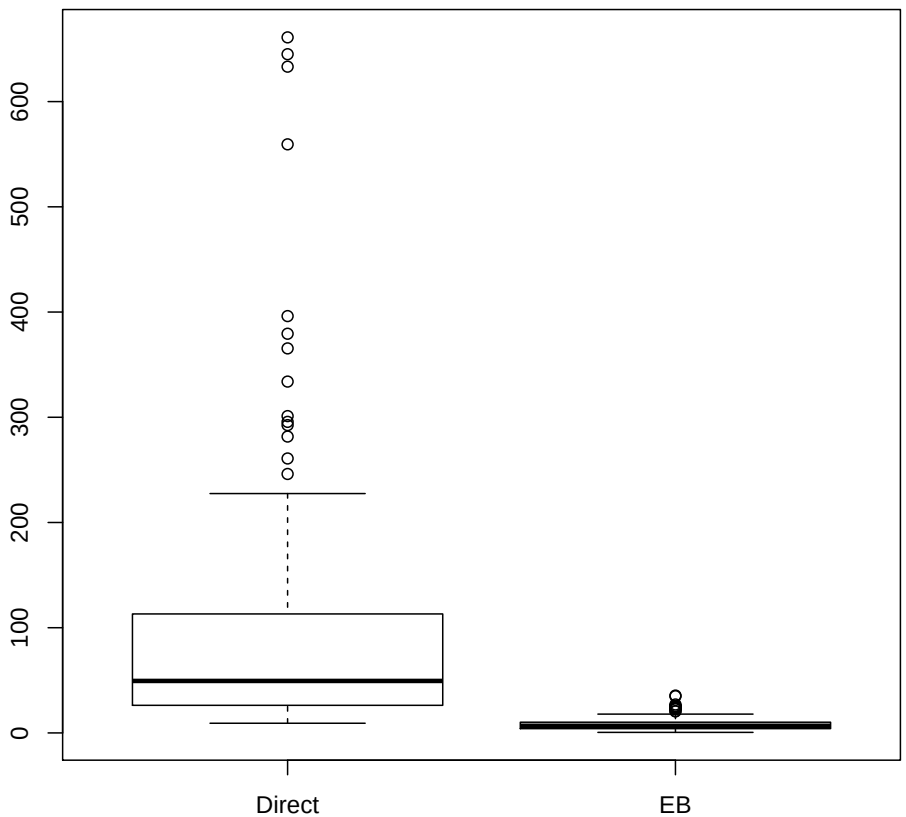
MODEL CHECKING

- 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.

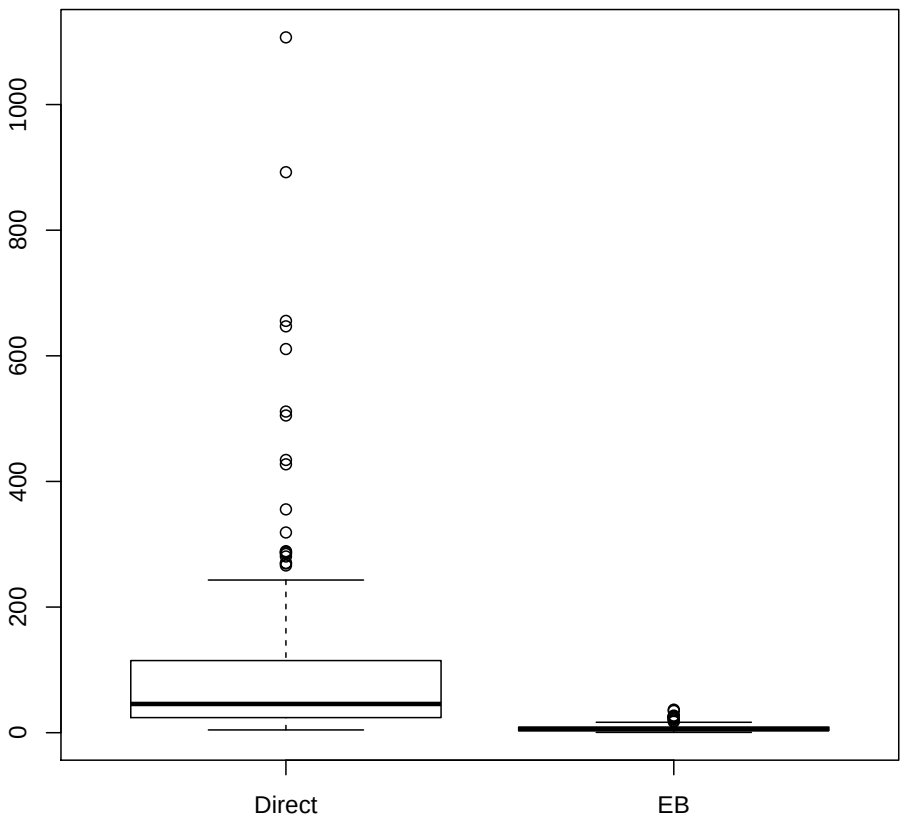
QUALITY EB vs. DIRECT: POV. RATE

- ✓ Median MSE Women: Direct **47**, EB: **6.7**
- ✓ Median MSE Men: Direct **45.8**, EB: **5.5**

MSE: Women

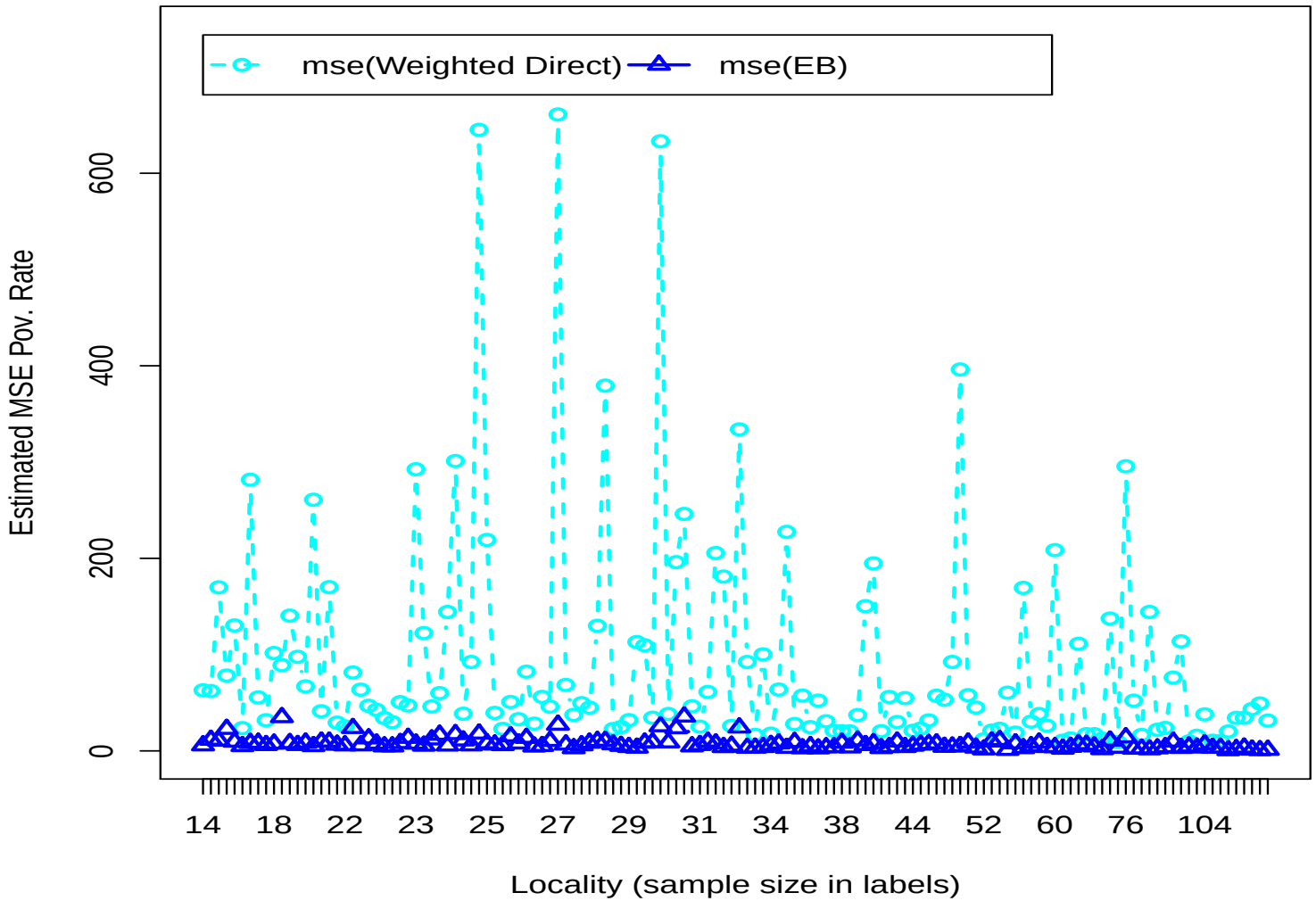


MSE: Men

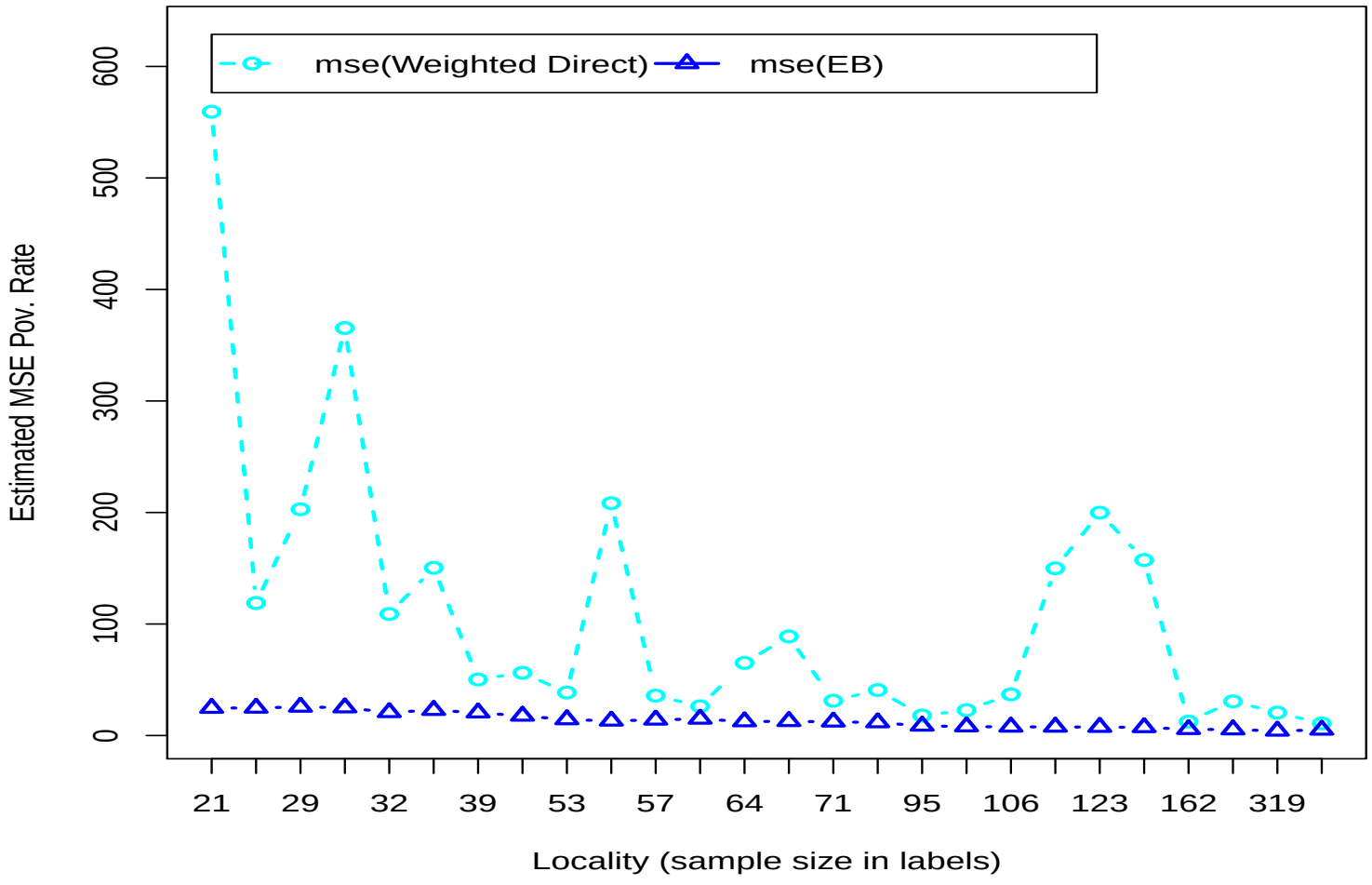


EB vs. DIRECT: WOMEN, WEST BANK

✓ Reduction in **all** but one locality, **84 %** average MSE reduction!



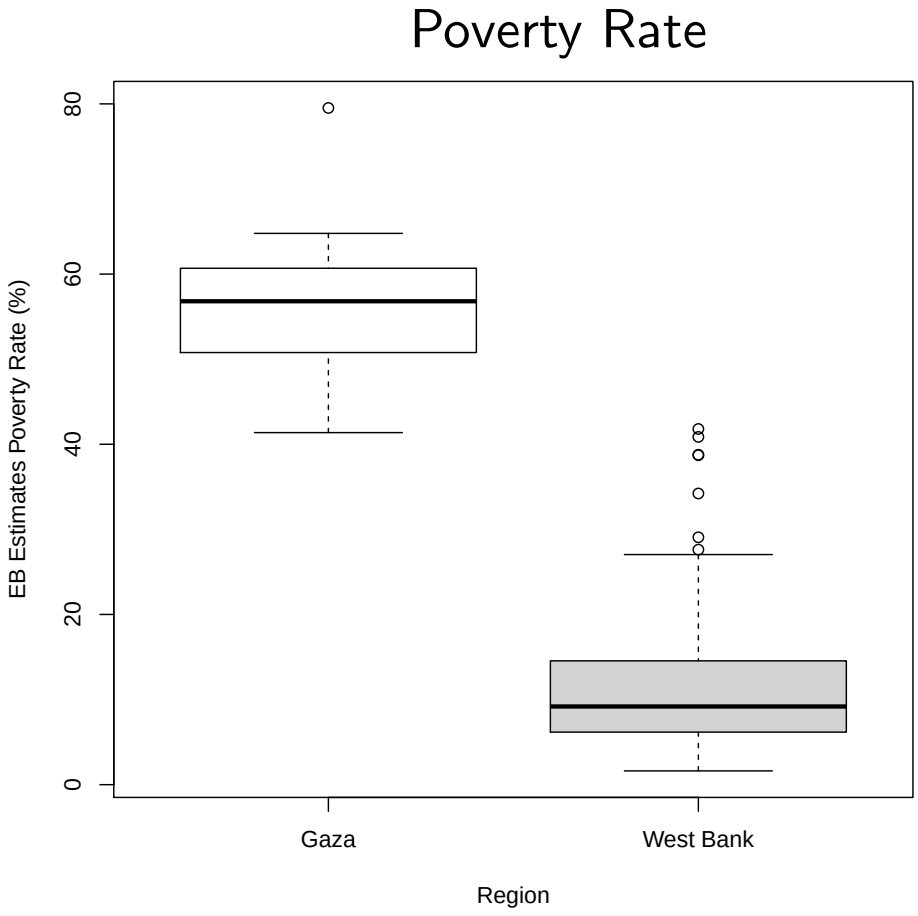
EB vs. DIRECT: WOMEN, GAZA

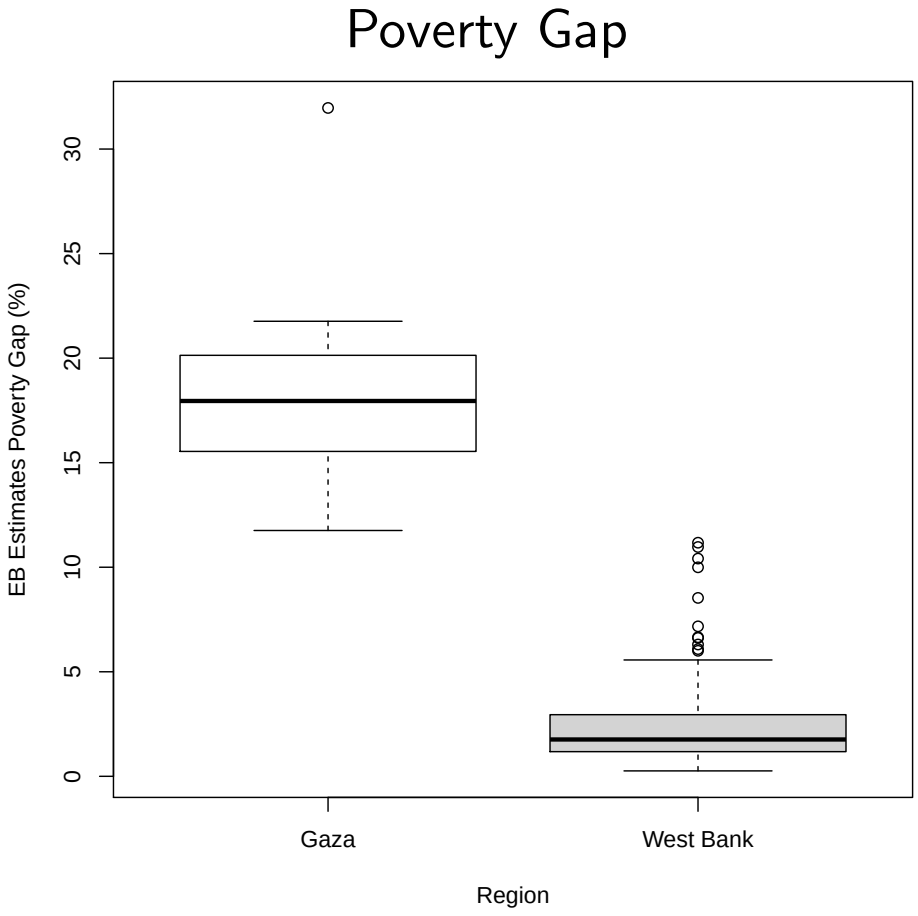


✓ **Great gains** also for Pov. **Gap** (not shown)!

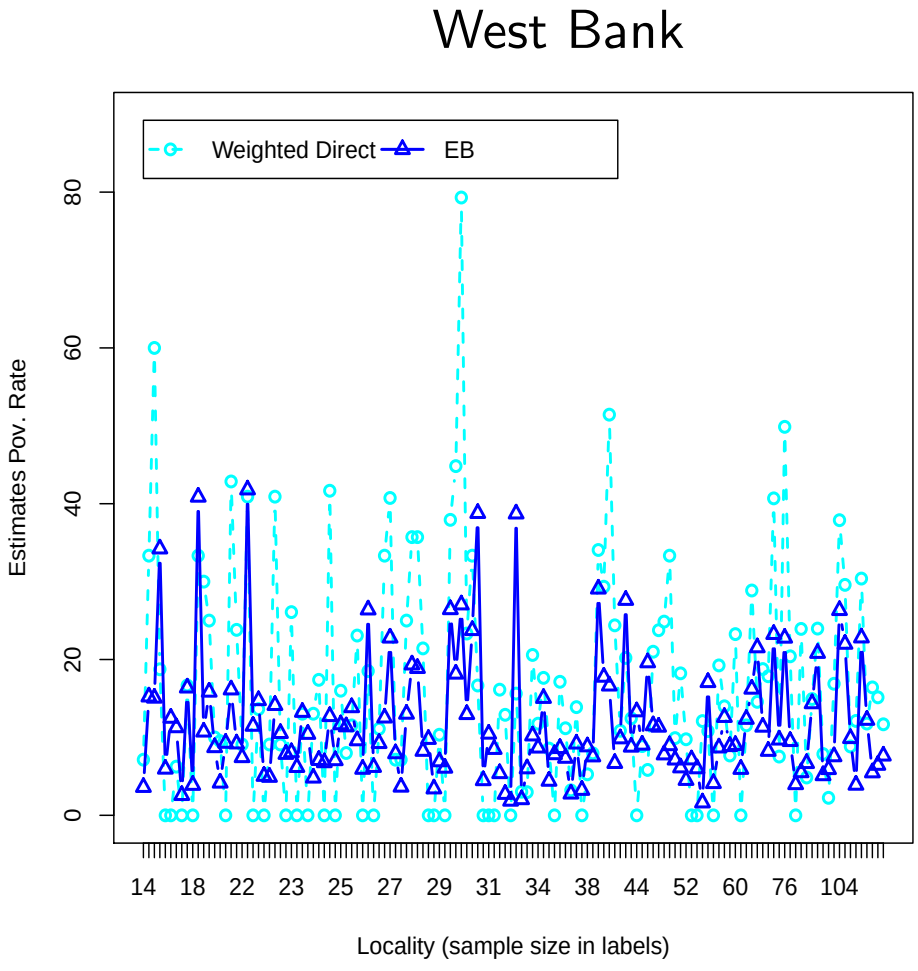
ESTIMATES BY REGION

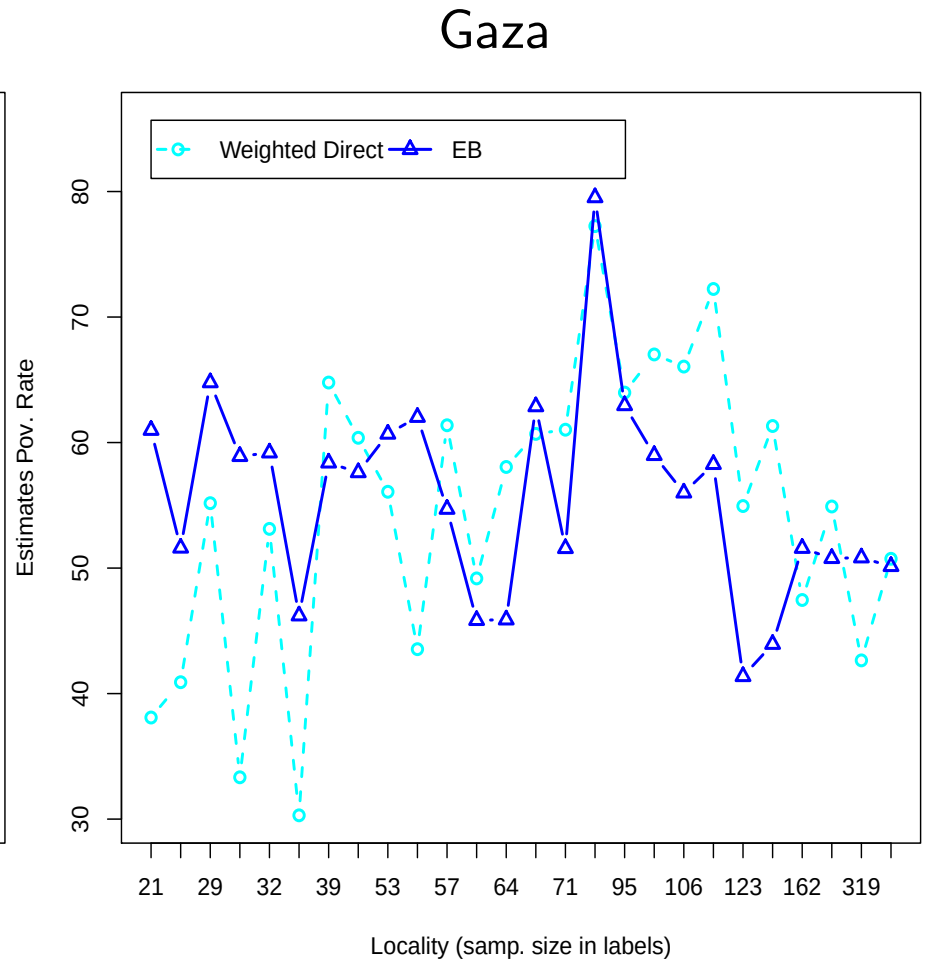
- ✓ Median Pov. Rate: Gaza **55 %**, West Bank: **8.3 %**
- ✓ Median Pov. Gap: Gaza **17.4 %**, West Bank: **1.5 %**





ESTIMATED POV. RATE: WOMEN





CONCLUSIONS

- The use of **census data** in a model allows us to obtain disaggregated estimates of **much better quality**.
- **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 (smooth), without visible systematic design bias.
- **The reduction of error** of model-based estimators with respect to direct estimators is notorious (over 82 %).

CONCLUSIONS

- The considered unit level 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. Still, model variations are being 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!