Data-driven area-specific selection of synthetic or nested-error models

Isabel Molina

Institute of Interdisciplinary Mathematics
Department of Statistics and Operations Research
Complutense University of Madrid (Spain)

(with Agne Bikauskaite and Domingo Morales)

INTRODUCTION

TERMINOLOGY AND NOTATION

- **Population:** **Finite** set of units U, of size N.
- **Areas/domains:** Subpopulations of interest $U_1, \ldots, U_D$, of sizes $N_1, \ldots, N_D$, which form a partition of U.
- **Variable of interest:** y_{dj} for unit j within area/domain d.
- $\mathbf{y}_d = (y_{d1}, \ldots, y_{dN_d})'$ vector of outcomes for **area** d.
- **Target indicators:** General, possibly **non-linear** function of $\mathbf{y}_d$,

$$\eta_d = h_d(\mathbf{y}_d), \quad d = 1, \ldots, D.$$

$\rightsquigarrow$ Example: Means in the areas/domains,

$$\eta_d = \bar{Y}_d = \frac{1}{N_d} \sum_{j=1}^{N_d} y_{dj}, \quad d = 1, \ldots, D.$$

TERMINOLOGY AND NOTATION

- **Sample:** s sample of size n obtained from U.
- **Sub-sample:** $s_d = s \cap U_d$ sub-sample from area d, of size $n_d \leq N_d$.
- **Sample complement:** $c_d = U_d - s_d$ out-of-sample units from area d.
- **Direct estimator:** $\hat{\eta}_d$ based on the n_d sample survey observations from the area/domain, $\{y_{dj}, j \in s_d\}$.
  - $\rightsquigarrow$ No distributional assumptions about y_{dj}: **non parametric**.
  - $\rightsquigarrow$ Typically based on the sampling design, which **protects from informative sampling**.
  - $\rightsquigarrow$ Unluckily, highly **inefficient** for areas with small sample size.

INDIRECT ESTIMATORS

- **Borrow strength** from other areas by making some kind of **homogeneity** assumption across areas.
- Typically a model that uses **auxiliary information**, with **common** parameters for all the areas.

SYNTHETIC ESTIMATORS

- Hansen, Hurwitz & Madow (1953), p. 483, 1945 Radio Listening Survey.
  - $\rightsquigarrow$ Two sets of estimates of median num. of radio stations heard during day:
    x_d from mail survey (biased), available for the $D = 500$ US counties;
    y_d from intensive survey, available for $m = 85$ counties.
  - $\rightsquigarrow$ Regression $y_d = \beta_0 + \beta_1 x_d + e_d$ fitted to the $m = 85$.
    Predicted values obtained for the remaining $D = 500 - 85$ counties.
- They relate all the areas through **common** parameters, **without** allowing for area heterogeneity.
- Common model parameters estimated using data **from all the areas**. Hence design variance typically **very small**.

SYNTHETIC ESTIMATORS

- Assumptions of same behaviour for all areas often **not realistic**.
- Hence, they may be substantially **design-biased**.
- Derived basically from the model and precious survey data **waisted**.
- **Do not tend** to direct estimates as the area sample sizes grows.
- Traditionally MSE **not correctly estimated**.
  - ⇝ Design MSE: Must account for **bias**, hard to find good area-specific estimates.
  - ⇝ Model MSE: **Misleading** if synthetic assumptions do not hold.

NESTED-ERROR MODEL

NESTED-ERROR MODEL

- Nested-error linear regression model for the census:

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

- $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma_u^2, \sigma_e^2)'$ **common** parameters for all the areas.
- u_d **specific** random effect of area d, $d = 1, \ldots, D$.
- Area effects u_d and erros e_{dj} all mutually independent.

✓ *Battese, Harter & Fuller (1988), JASA*

NESTED-ERROR MODEL

- In matrix notation:

$$\mathbf{y}_d \overset{ind.}{\sim} N(\mathrm{X}_d\boldsymbol{\beta}, \mathrm{V}_d(\sigma_u^2, \sigma_e^2)), \quad d = 1, \ldots, D.$$

- Covariance matrix:

$$\mathrm{V}_d(\sigma_u^2, \sigma_e^2) = \begin{pmatrix} \sigma_u^2 + \sigma_e^2 & \sigma_u^2 & \cdots & \sigma_u^2 \\ \sigma_u^2 & \sigma_u^2 + \sigma_e^2 & \cdots & \sigma_u^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_u^2 & \sigma_u^2 & \cdots & \sigma_u^2 + \sigma_e^2 \end{pmatrix} = \sigma_u^2 \mathbf{1}_{N_d} \mathbf{1}'_{N_d} + \sigma_e^2 \mathbf{I}_{N_d}.$$

- Observations from **different** areas are **independent**:

$$\mathrm{cov}(y_{dj}, y_{\ell j}) = 0, \text{ for all } d \neq \ell.$$

✓ *Battese, Harter & Fuller (1988), JASA*

TRADITIONAL ELL METHOD

- Removing area effects (i.e. setting $\sigma_u^2 = 0$) in nested-error model $\Rightarrow$ Synthetic regression.
- If auxiliary information explain most of area heterogeneity ($\sigma_u^2 \approx 0$), synthetic regression **OK**.
- Elbers, Lanjouw & Lanjouw (2003, Econometrica) proposed a method for **general** indicators.
- ELL assume the nested error model with random **cluster** effects.
- Even if **clusters** were **equal** to the **areas**, Molina & Rao (2010, CJS) showed that ELL method is basically **synthetic**.

TRADITIONAL ELL METHOD

- In areas with zero sample size (not recommended!), synthetic regression model is used to predict.
- ELL argue that:
  - ⇝ ELL is usually applied in countries where only a small portion of the areas are sampled.
  - ⇝ ELL uses rich census auxiliary information that may explain the area heterogeneity.
  - ⇝ Contextual (area-level) covariates reduce area heterogeneity.

✓ *Elbers, Lanjouw & Lanjouw (2003), Econometrica*

MACHINE-LEARNING TECHNIQUES

- Flexible modeling might also reduce area heterogeneity: Machine learning techniques (random forests, boosting, neural networks,...).
- Corral et al. (2022) evaluated synthetic gradient boosting techniques:
  - $\rightsquigarrow$ With rich auxiliary information (census): great performance.
  - $\rightsquigarrow$ With weak auxiliary information (georeferenced area-level data): very poor performance.
- Conclusion: Need to account for **area effects**.

SELECTION OF AREA EFFECTS

- Synthetic regression ignores area effects **without distinction** between areas, wasting the survey data even for areas with very large n_d.
- In absence of area effects, synthetic model leads to **efficient** estimators.
- In area-level models:
  - ⇝ Datta, Hall & Mandal (2011, JASA) used a testing procedure to decide between synthetic or area-effects model.
  - ⇝ Datta & Mandal (2015, JASA) considered a "spike and slab" distribution for the area effects (mixture giving mass p to the usual normal and $1 - p$ to zero), allowing for area-specific selection, in the Bayesian context.
- We deal with **unit-level** models in the **frequentist** setup.

EMPIRICAL BEST PREDICTION

BEST PREDICTOR UNDER NESTED-ERROR MODEL

- Assumes the nested-error model with random effects for all the areas.
- Best predictor of $\eta_d = h_d(\mathbf{y}_d)$: Predictor $\tilde{\eta}_d$ minimizing the model MSE,

$$\text{MSE}_{\mathbf{y}}(\tilde{\eta}_d) = E_{\mathbf{y}}\left[(\tilde{\eta}_d - \eta_d)^2\right].$$

- Separate $\mathbf{y}_d = (y_{d1}, \ldots, y_{dN_d})'$ into sample and sample-complement:

$$\mathbf{y}_d = (\mathbf{y}'_{ds}, \mathbf{y}'_{dc})', \quad d = 1, \ldots, D.$$

- Best predictor of η_d:

$$\tilde{\eta}_d^B(\boldsymbol{\theta}) = E_{\mathbf{y}_{dc}}[\eta_d | \mathbf{y}_{ds}].$$

✓ *Molina & Rao (2018), CJS*

EMPIRICAL BEST PREDICTOR

- The best predictor depends on $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma_u^2, \sigma_e^2)'$.
- $\mathbf{y}_s = (\mathbf{y}_{1s}', \ldots, \mathbf{y}_{Ds}')'$ overall sample data, of size $n = \sum_{d=1}^{D} n_d$.
- Obtain a consistent estimator (as $D \to \infty$) $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ based on $f(\mathbf{y}_s; \boldsymbol{\theta})$.
- **Empirical** best (EB) predictor:

$$\hat{\eta}_d^{EB} = \tilde{\eta}_d^B(\hat{\boldsymbol{\theta}}).$$

- For certain non-linear indicators, EB has no closed form.
- In those cases, **Monte Carlo simulation** used to approximate EB.

✓ *Molina & Rao (2010), CJS*

AREA-SPECIFIC SELECTION OF AREA EFFECTS

MIXTURE SYNTHETIC AND NESTED-ERROR MODELS

- Multivariate mixture between synthetic regression and nested-error models:

$$\mathbf{y}_d \overset{ind.}{\sim} \pi N(\mathrm{X}_d\boldsymbol{\beta}_1, \mathrm{V}_{1d}) + (1-\pi)N(\mathrm{X}_d\boldsymbol{\beta}_0, \mathrm{V}_{0d}),$$

  where $\pi \in [0,1]$ and

$$\mathrm{V}_{1d} = \mathrm{V}_d(\tau^2, \sigma_1^2), \quad \mathrm{V}_{0d} = \sigma_0^2 \mathbf{I}_{N_d}, \ d = 1, \ldots, D.$$

- Latent variables: $Z_d \overset{iid}{\sim} \text{Bern}(\pi)$, $d = 1, \ldots, D$.
- The mixture model is then:

$$\mathbf{y}_d | Z_d = 1 \sim N(\mathrm{X}_d\boldsymbol{\beta}_1, \mathrm{V}_{1d}) \text{ and } \mathbf{y}_d | Z_d = 0 \sim N(\mathrm{X}_d\boldsymbol{\beta}_0, \mathrm{V}_{0d}).$$

✓ *Ph.D. thesis of A. Bikauskaite (2024)*

MIXTURE SYNTHETIC AND NESTED-ERROR MODELS

- Two different mixture model specification:
  - ⇝ Equal model parameters:

$$\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0 = \boldsymbol{\beta},\ \sigma_1^2 = \sigma_0^2 = \sigma^2 \Rightarrow \boldsymbol{\theta} = (\pi, \boldsymbol{\beta}', \tau^2, \sigma^2)'$$

  - ⇝ Different model parameters:

$$\boldsymbol{\theta} = (\pi, \boldsymbol{\beta}_1', \tau^2, \sigma_1^2, \boldsymbol{\beta}_0', \sigma_0^2)'$$

✓ *Ph.D. thesis of A. Bikauskaite (2024)*

MIXED BEST PREDICTORS

- Mixed best predictor of η_d (if Z_d was observed):

$$\tilde{\eta}_d^{MB}(Z_d) = \begin{cases} E_{\mathbf{y}_{dc}}(\eta_d|\mathbf{y}_{ds}, Z_d = 1) \text{ if } Z_d = 1; \\ E_{\mathbf{y}_{dc}}(\eta_d|\mathbf{y}_{ds}, Z_d = 0) \text{ if } Z_d = 0. \end{cases}$$

- Z_d unobservable, but we use posterior probabilities:

$$p_{1d} = P(Z_d = 1|\mathbf{y}_{ds}), \; p_{0d} = 1 - p_{1d}.$$

✓ *Ph.D. thesis of A. Bikauskaite (2024)*

MIXED BEST PREDICTORS

- Mixed best predictor 1 (MB1):

$$\tilde{\eta}_d^{MB1} = \begin{cases} E_{\mathbf{y}_{dc}}(\eta_d|\mathbf{y}_{ds}, Z_d = 1) \text{ if } p_{1d} \geq 0.5, \\ E_{\mathbf{y}_{dc}}(\eta_d|\mathbf{y}_{ds}, Z_d = 0) \text{ if } p_{1d} < 0.5. \end{cases}$$

- Mixed best predictor 2 (MB2):

$$\tilde{\eta}_d^{MB2} = p_{1d} E_{\mathbf{y}_{dc}}(\eta_d|\mathbf{y}_{ds}, Z_d = 1) + p_{0d} E_{\mathbf{y}_{dc}}(\eta_d|\mathbf{y}_{ds}, Z_d = 0).$$

✓ *Ph.D. thesis of A. Bikauskaite (2024)*

EMPIRICAL MIXED BEST PREDICTORS

- For each of the two model specifications:
  - ⇝ Model fitting: ML through **E-M** algorithm.
  - ⇝ **Empirical** MB predictors: Replace unknown model parameters in MB1 and MB2 with ML estimators.
  - ⇝ **Closed-form** expressions for indicators of special interest.
  - ⇝ **MC simulation** algorithm for general indicators.
  - ⇝ **Bootstrap** for MSE estimation.

✓ *Ph.D. thesis of A. Bikauskaite (2024)*

SIMULATION EXPERIMENTS

SIM. EXPERIMENT: EQUAL MODEL PARAMETERS

- Population and sample sizes:

$$N = 20000, \quad D = 80,$$
$$N_d = 250, \quad n_d = 20, 50, \; d = 1, \ldots, D$$

- Area effects and errors:

$$u_d \sim N(0, \tau^2), \quad e_{dj} \sim N(0, \sigma^2), \quad \tau^2 = 0.15^2, \; \sigma^2 = 0.5^2.$$

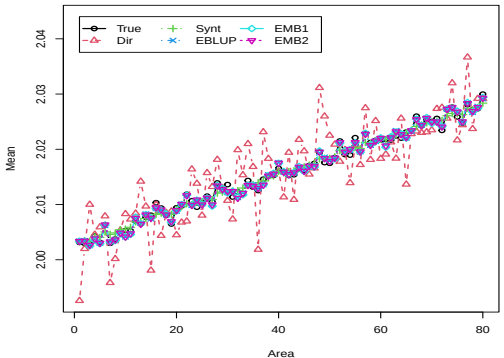
- Latent variables: For $\pi = 3/4$, $Z_d \overset{iid}{\sim} \text{Bern}(\pi)$, $d = 1, \ldots, D$.
- Target variables: For $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)' = (2, 0.05, -0.06)'$,

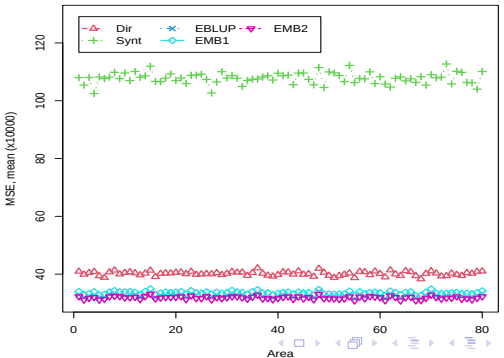
$$y_{dj} = \begin{cases} \beta_0 + \beta_1 x_{1dj} + \beta_2 x_{2dj} + u_d + e_{dj} & \text{if } Z_d = 1, \\ \beta_0 + \beta_1 x_{1dj} + \beta_2 x_{2dj} + e_{dj} & \text{if } Z_d = 0. \end{cases}$$

✓ *Ph.D. thesis of A. Bikauskaite (2024)*

AREA MEAN: $n_d = 50$

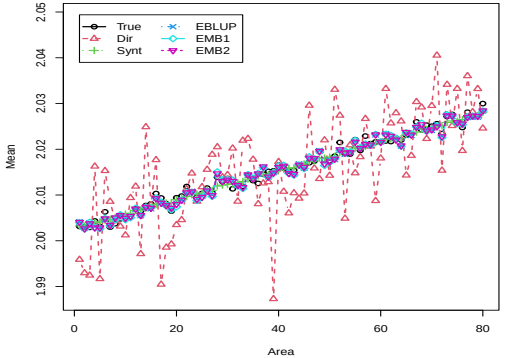
- Synthetic **performing much worse** than Direct!
- EB=EBLUP **performing well** for all the areas.

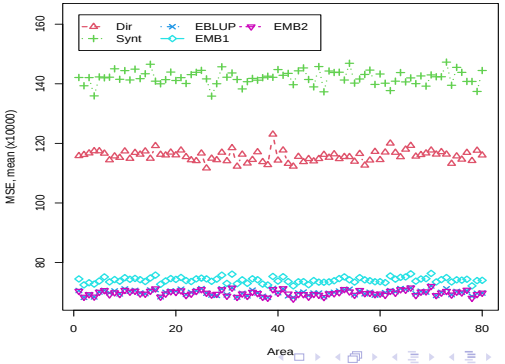




AREA MEAN: $n_d = 20$

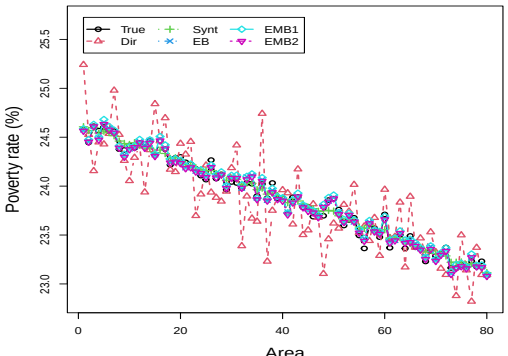
- EB=EBLUP **very close to** EMB2 for all the areas.

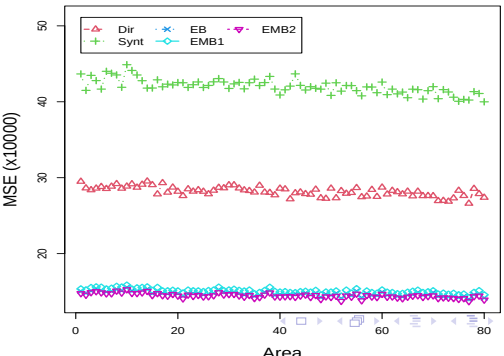




POVERTY RATE: $n_d = 50$

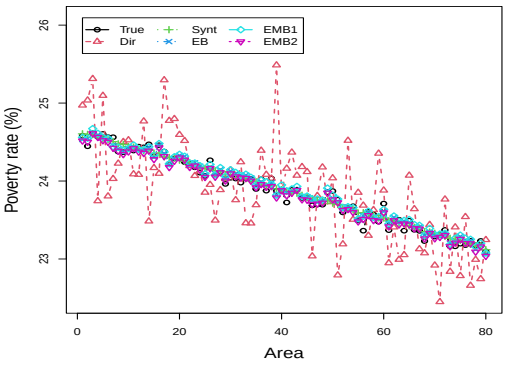
- Synthetic **performing even worse** than Direct!
- EB **performing well** for all the areas.

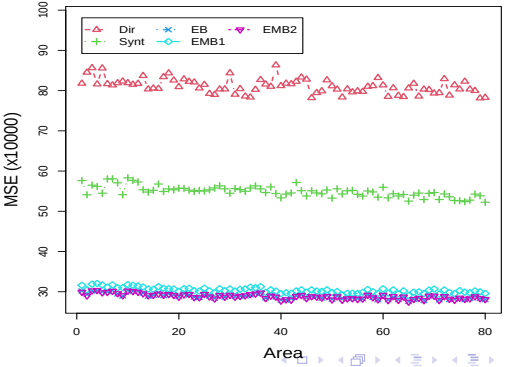




POVERTY RATE: $n_d = 20$

- EB **very close to** EMB2 for all the areas.





SIM. EXPERIMENT: DIFFERENT MODEL PARAMETERS

- Area effects and errors:

$$u_{1d} \sim N(0, \tau^2),\ e_{kdj} \sim N(0, \sigma_k^2),\ \tau^2 = 0.15^2,\ \sigma_1^2 = 0.5^2,\ \sigma_0^2 = 0.75^2.$$

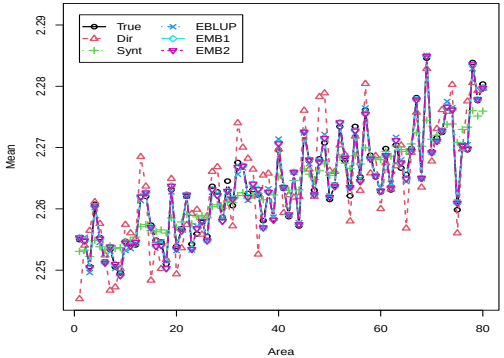
- Latent variables: For $\pi = 3/4$, $Z_d \overset{iid}{\sim} \text{Bern}(\pi)$, $d = 1, \ldots, D$.
- Target variables: For $\boldsymbol{\beta}_0 = (\beta_{00}, \beta_{01}, \beta_{02}) = (3, 0.03, -0.04)$ and $\boldsymbol{\beta}_1 = (\beta_{10}, \beta_{11}, \beta_{12}) = (2, 0.05, -0.06)$,

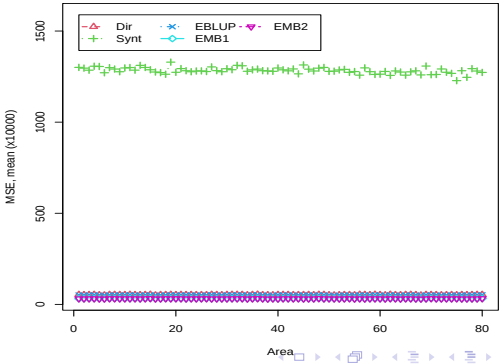
$$y_{dj} = \begin{cases} \beta_{10} + \beta_{11} x_{1dj} + \beta_{12} x_{2dj} + u_{1d} + e_{1dj} & \text{if } Z_d = 1, \\ \beta_{00} + \beta_{01} x_{1dj} + \beta_{02} x_{2dj} + e_{0dj} & \text{if } Z_d = 0. \end{cases}$$

✓ *Ph.D. thesis of A. Bikauskaite (2024)*

AREA MEAN: $n_d = 50$

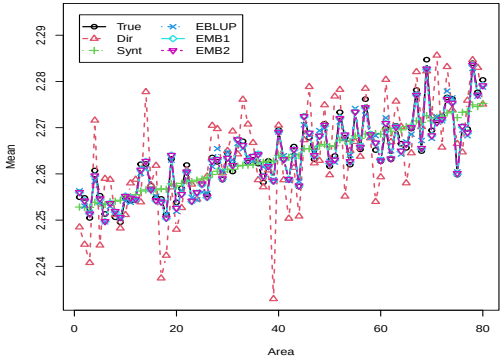
- Synthetic **performing much worse** than all other!

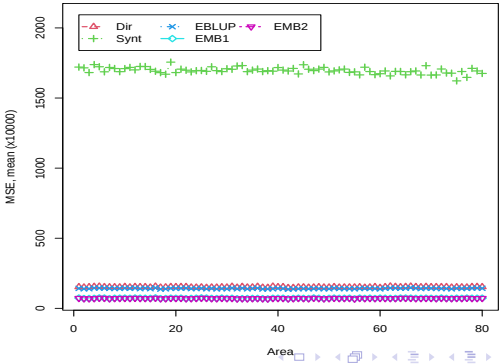




AREA MEAN: $n_d = 20$

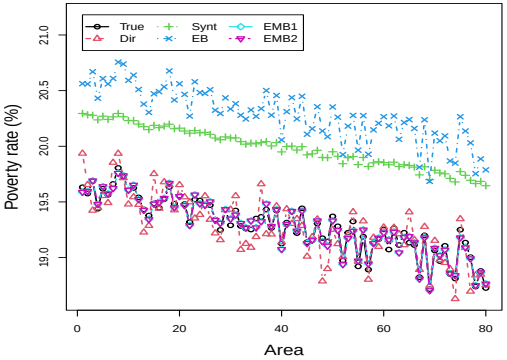
- EB=EBLUP now slightly worse than EMB2.

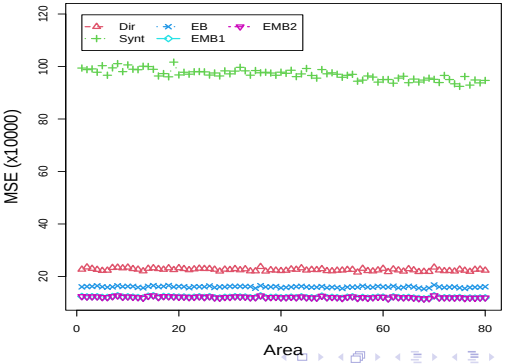




POVERTY RATE: $n_d = 50$

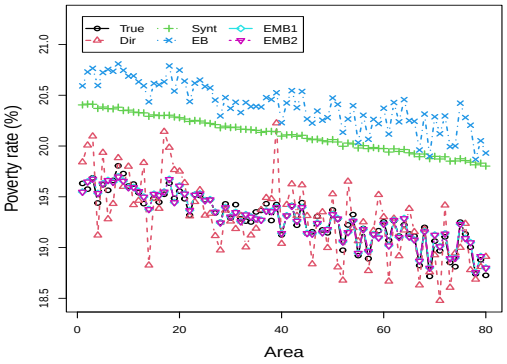
- Synthetic **performing much worse** than Direct!

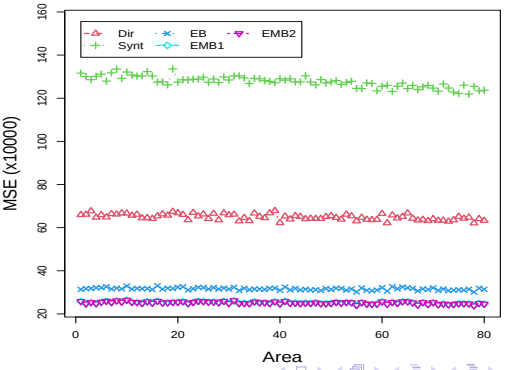




POVERTY RATE: $n_d = 20$

- EB based on common model **biased**.





CONCLUSIONS

- When a fraction of areas is generated from synthetic model, EB obtained including area effects for all the areas performs OK when model parameters in synthetic and nested-error models are **equal**.
- When model parameters are **different**, EB **biased** for non-linear indicators.
- Synthetic estimators perform poorly as soon as some area effects area needed.
- Mixture model **worth** when model parameters are **different**.

ILLUSTRATION: POVERTY IN PALESTINE 2016/17

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

- **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.)

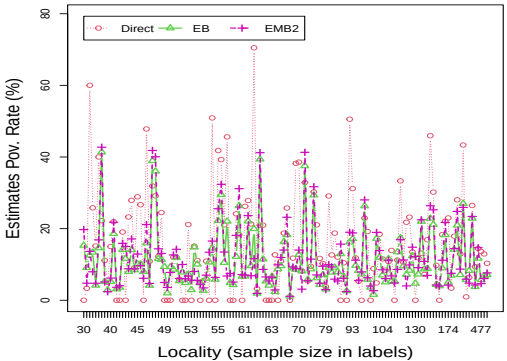
✓ *García-Portugués & Molina (2020), ESCWA.*

ESTIMATES OF MODEL PARAMETERS AND 95% CIs

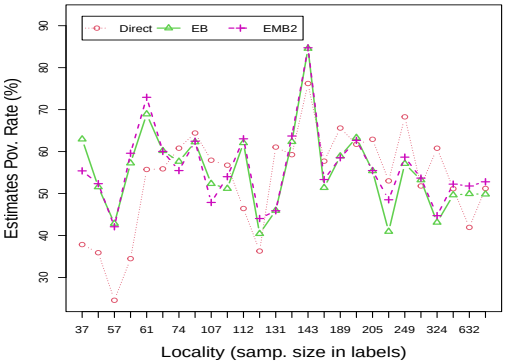
	NE			Mixture($k=1$)			Mixture($k=2$)			
Parameter	LL	Est	UL	LL	Est	UL	LL	Est	UL	
τ^2, τ_1^2	0.012	0.0156	0.019	0.013	0.0178	0.022				
σ^2, σ_k^2	0.102	0.1055	0.109	0.090	0.0918	0.094	0.118	0.1218	0.126	*
π_k				0.635	0.7004	0.766	0.234	0.2996	0.365	

POVERTY RATE

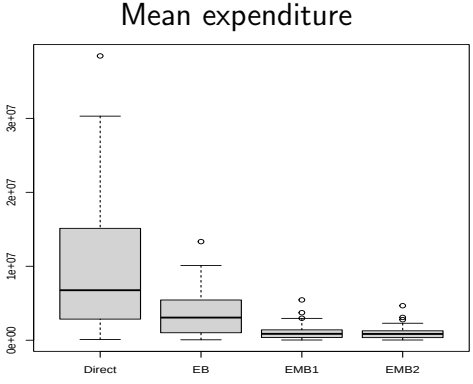
West Bank

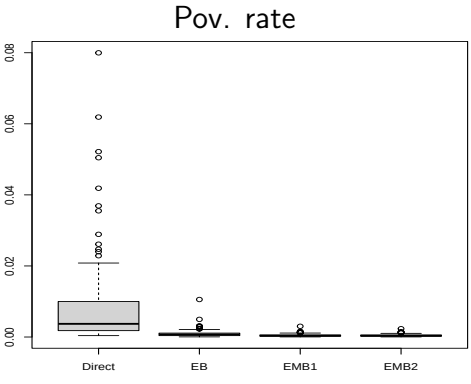


Gaza



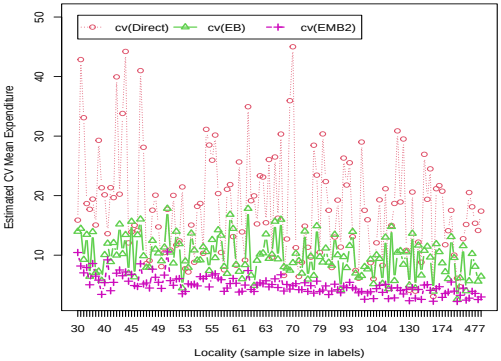
ESTIMATED MSE



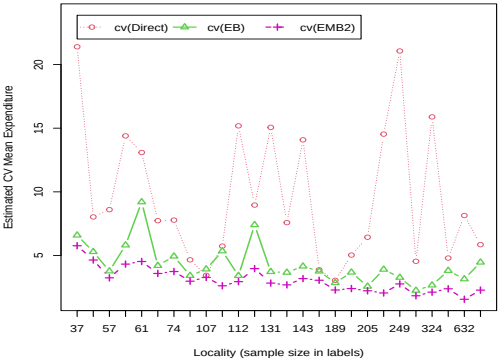


CV: MEAN EXPENDITURE

West Bank

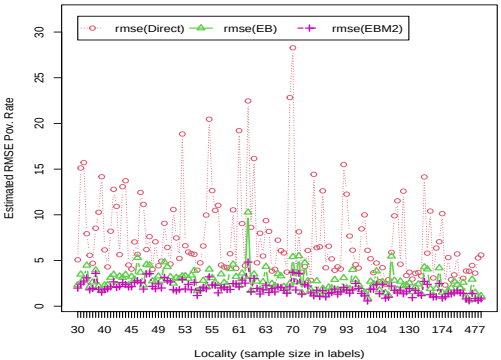


Gaza

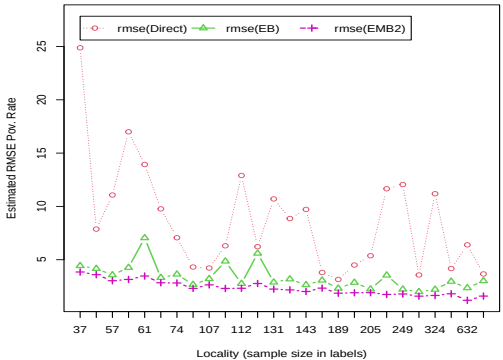


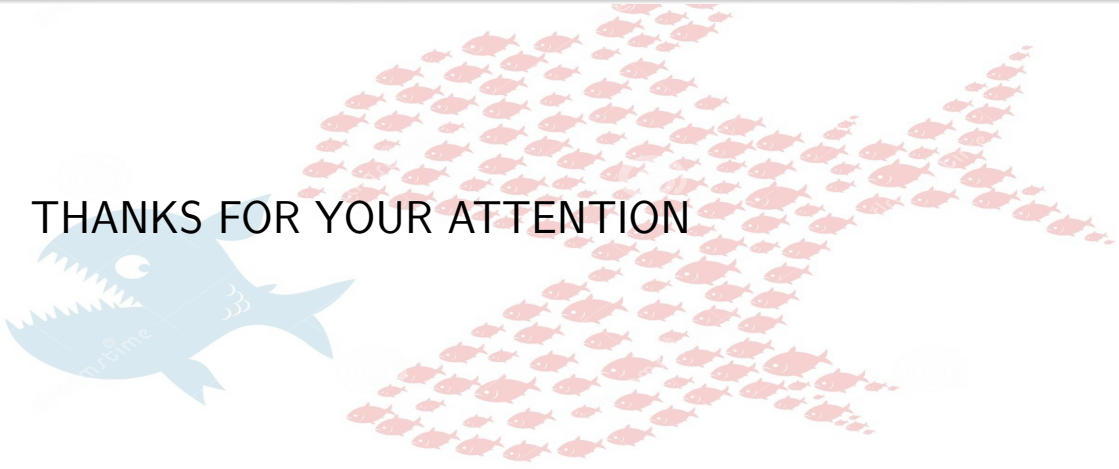
RMSE: POV. RATE

West Bank



Gaza





THANKS FOR YOUR ATTENTION