UNIVERSITÀ DI PISA

Freie Universität Berlin

A Unit-level Quantile Nested Error Regression Model for Domain Prediction

Timo Schmid, Nikos Tzavidis, Nicola Salvati and Beate Weidenhammer

Workshop

Pisa
July 2016

UNIVERSITÀ DI PISA

Freie Universität Berlin

Overview

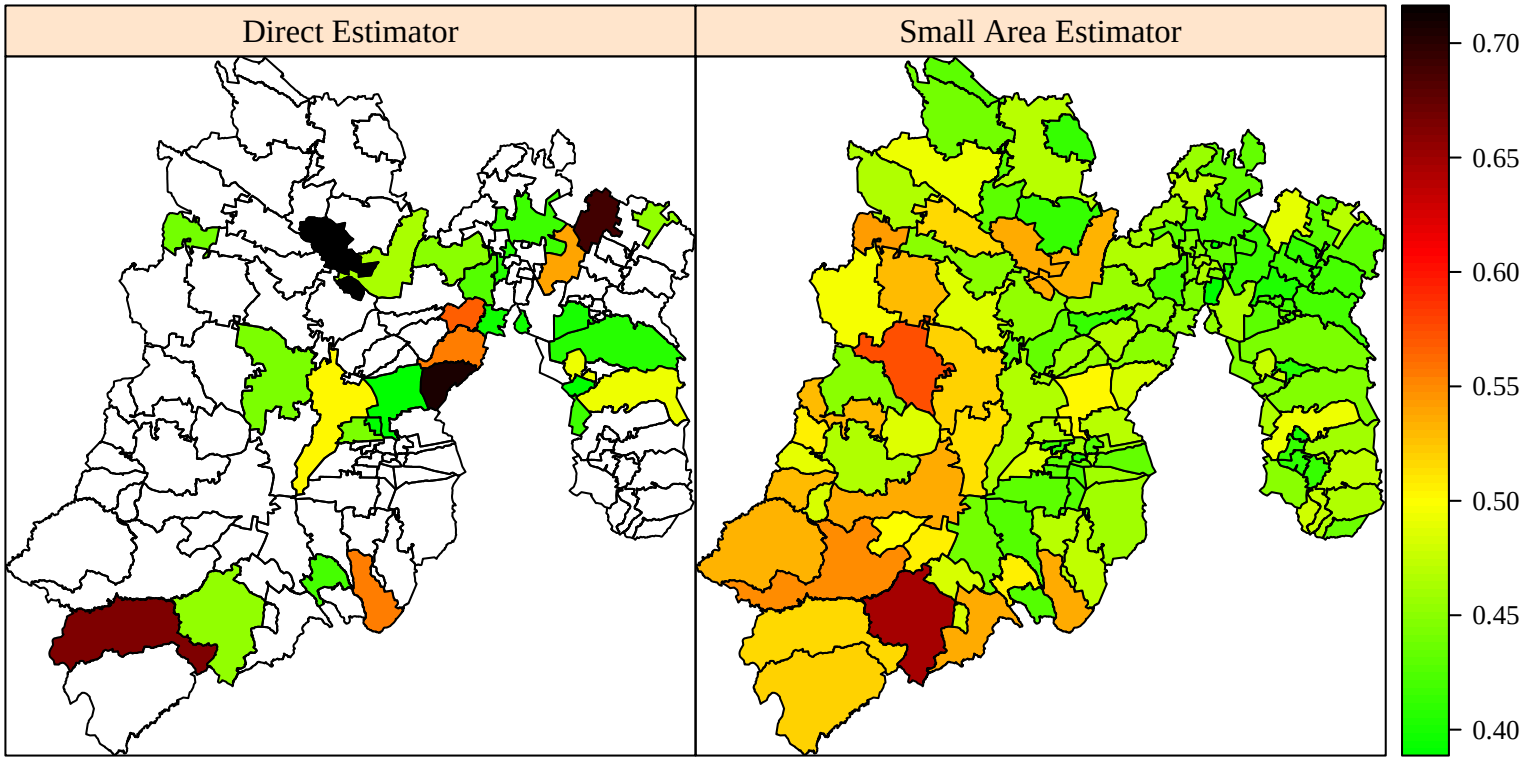
UNIVERSITÀ DI PISA

Freie Universität Berlin

What is the problem?

- **Surveys** are designed to produce **reliable estimates at the National level** and also at pre-specified sub-National levels
- What happens if after collecting the data we are interested in **producing estimates for groups/areas** consisting of very few observations?
- Using only the sample data will result in estimates that have **high sampling variance**.
- Small area estimation is concerned with the development of statistical procedures for producing **efficient** estimates for domains (**planned** or **unplanned**) with **small** or **zero** sample sizes.

Illustation using Mexican data - Gini coefficient



(white - municipalities with zero sample sizes)

Domain Prediction of Non-linear Indicators

- Dominated recent SAE literature
- Motivated by growing needs of statistics agencies
- Examples
  - Estimate the income distribution
  - Estimate poverty and inequality indicators
- SAE Data Requirements:
  - **Survey Data:** Available for y and for x related to y
  - **Census/Administrative Data:** Available for x but not for y

Auxiliary information available for every unit in the population is needed.

Non-linear Indicators

- FGT measures (Foster et al.,1984)

$$FGT(\alpha, t) = \sum_{i=1}^{N} \left(\frac{t - y_i}{t} \right)^{\alpha} 1(y_i \leq t)$$

$\alpha = 0$ - Head Count Ratio; $\alpha = 1$ - Poverty Gap

- Gini coefficient
- Quintile Share Ratio

$$QSR_{80/20} = \frac{\sum_{i=1}^{N} [y_i 1(y_i > q_{0.8})]}{\sum_{i=1}^{N} [y_i 1(y_i \leq q_{0.2})]}$$

Recent Methodologies

- The World Bank method
  (Elbers et al., 2003, Econometrica)
- The Empirical Best Predictor (EBP) method
  (Molina & Rao, 2010, CJS)
- EBP based on normal mixtures
  (Lahiri and Gershunskaya, 2011; Elbers & Van der Weide, 2014)
- Methods based on M-Quantiles
  (Marchetti et al., 2012, CSDA)

The EBP Method

- Point of departure: Nested error regression model Battese, Harter & Fuller (1988, JASA)

  Notation: (k =domain, i =individual)

$$y_{ik} = \boldsymbol{x}_{ik}^T \boldsymbol{\beta} + \boldsymbol{z}_{ik}^T u_k + \epsilon_{ik}, i = 1, ..., n_k, k = 1, ..., D$$

- Use sample data to estimate β, σ_u, σ_ϵ, u_k
- Generate $u_k^* \sim N(0, \hat{\sigma}_u^2 * (1 - \gamma_k))$ & $\epsilon_{ik}^* \sim N(0, \hat{\sigma}_\epsilon^2)$

Micro-simulating a synthetic population:

- Generate a synthetic population under the model a large number of times each time estimating the target parameter
- Linear and non-linear indicators can be computed

UNIVERSITÀ DI PISA

Freie Universität Berlin

Motivating Alternative Methods

- EBP relies on assumptions about the distribution of the data
- What if these fail?
- Option 1: Explore the use of transformations
  - Deciding on appropriate transformations requires some work
- Option 2: EBP formulation under an alternative distribution
- Option 3: Estimate the quantile function of the target empirical distribution
  - Do this by using a nested error regression type model

UNIVERSITÀ DI PISA

Freie Universität Berlin

Overview

Alternative View

- Let $\boldsymbol{x}$ denote a set of covariates and k a domain
- Let $Q_{y|\boldsymbol{x},k}(q|\boldsymbol{x},k)$ denote the quantile function of an unknown $F(y|\boldsymbol{x},k)$
- Interested in **estimating this quantile function**
- Simplest case: Assume a linear random effects model for the quantiles

$$Q_{y|\boldsymbol{x},k}(q|\boldsymbol{x},k) = \boldsymbol{x}_{ik}^T \boldsymbol{\beta}_q + v_k$$

- v_k domain random effect **capturing unobserved heterogeneity**

Model for the Quantiles

- $Q_{y|\boldsymbol{x},k}(q|\boldsymbol{x},k)$ is estimated by using the link between quantile regression and the $ALD(\mu_q, \sigma, q)$ likelihood (Geraci & Bottai, 2014) with $\mu_q = \boldsymbol{x}_{ik}^T \boldsymbol{\beta}_q + v_k$
- $v_k \sim N(0, \sigma_v^2)$; Other specifications are possible (Non-parametric)
- The approach for fitting is done separately at each q
- Estimation by ML, obtain $\hat{\boldsymbol{\beta}}_q$, $\hat{\sigma}_q$, $\hat{\sigma}_{v,q}^2$
- Plug-in predictor v is q-specific
- The predictor for the q-quantile of y given $\boldsymbol{x}$ in domain k is

$$\widehat{Q}_{y|\boldsymbol{x},k}(q|\boldsymbol{x},k) = \boldsymbol{x}_{ik}^T \hat{\boldsymbol{\beta}}_q + \hat{v}_{qk}$$

Micro-simulation & Domain Prediction

- Interested in domain-specific finite population parameters
- Some target indicators heavily depend on the tails of the distribution
- Capture the shape of the target distribution
- Define a fine grid of quantiles $q_g = \{0.01, 0.02, \ldots, 0.99\}$
- Fit the quantile model for $q \in q_g$
- Estimate $\widehat{Q}_{y|\boldsymbol{x},k}(q|\boldsymbol{x}, k)$
- One additional step: Draw Monte-Carlo samples from by using the estimated Quantile function

Micro-simulation & Domain Prediction

- Use the inverse transform method
- Draw $h = 1, \ldots, MC$, Uniform random variables $\sim (0, 1)$
- Invert the corresponding quantile function, obtain $\tilde{y}_{ik}$

$$\tilde{y}_{ik} = (\tilde{y}_{ik}^{(1)}, \tilde{y}_{ik}^{(2)}, \ldots, \tilde{y}_{ik}^{(MC)})$$

- Leads to microsimulation in domain k

$$\tilde{y}_k = (\tilde{y}_{1k}^{(1)}, \ldots, \tilde{y}_{1k}^{(MC)}, \tilde{y}_{2k}^{(1)}, \ldots, \tilde{y}_{2k}^{(MC)}, \ldots, \tilde{y}_{N_k k}^{(1)}, \ldots, \tilde{y}_{N_k k}^{(MC)})$$

UNIVERSITÀ DI PISA

Freie Universität Berlin

Micro-simulation & Domain Prediction

- Estimate domain target parameter from

$$\tilde{y}_k = (\tilde{y}_{1k}^{(1)}, \ldots, \tilde{y}_{1k}^{(MC)}, \tilde{y}_{2k}^{(1)}, \ldots, \tilde{y}_{2k}^{(MC)}, \ldots, \tilde{y}_{N_k k}^{(1)}, \ldots, \tilde{y}_{N_k k}^{(MC)})$$

- Domain average

$$\widehat{mean}_k = mean(\tilde{y}_k) = \frac{1}{N_k \cdot MC} \sum_{i=1}^{N_k} \sum_{mc=1}^{MC} \tilde{y}_{ik}^{(mc)}$$

- Domain median

$$\widehat{median}_k = median(\tilde{y}_k)$$

- Head Count Ratio

$$\widehat{HCR}_k = \frac{1}{N_k \cdot MC} \sum_{i=1}^{N_k} \sum_{mc=1}^{MC} I(\tilde{y}_{ik}^{(mc)} \leq t)$$

Constrained Fitting

- Currently the fitting process is done separately for each q
- Follows Lum & Gelfand (Bayesian Anal., 2012) ; Geraci & Bottai (Stats & Comp, 2014)
- Joint quantile specification is complex
- Alternative: Constrain the fitting process
- Allows for one common random effect across q
- Allows implicitly for correlation
- Can impose monotonicity at the same time

UNIVERSITÀ DI PISA

Freie Universität Berlin

MvQ: MSE Estimation

Bootstrap:

- Select q from U $\sim (0, 1)$
- Generate v_k^*
  - Using the assumed distribution
  - Non-parametrically with rescaling to adjust for shrinkage
- Generate ϵ_{ik}^*
  - Option 1: Re-sample from the empirical distribution of residuals appropriately rescaled
  - Option 2: Investigate the use of wild bootstrap to accommodate the non-id case (Feng et al., Biometrika, 2011)

$$y_{ik}^* = \mathbf{x}_{ik}^T \hat{\boldsymbol{\beta}}_q + v_k^* + \epsilon_{qik}^*$$

MvQ: MSE Estimation

- Construct B bootstrap populations
- For each population b compute the population indicators, θ_k^{*b}
- From each bootstrap population select a bootstrap sample
- Implement the MvQ with the bootstrap sample, get $\hat{\theta}^{*b}$

$$\widehat{MSE}(\hat{\theta}_k) = B^{-1} \sum_{b=1}^{B} (\hat{\theta}_k^{*b} - \theta_k^{*b})^2$$

UNIVERSITÀ DI PISA

Freie Universität Berlin

Discrete Outcomes - Counts

- Goal: Estimate the distribution function of y
- BUT: Quantiles of discrete y are not continuous
- Jittering: Machado & Santos Silva (JASA, 2005)
  - Impose smoothness by adding to y random noise from a distribution F with support in $(0, 1)$
  - Example: $F \sim Uniform(0, 1)$
- $1 - 1$ relationship between the quantiles of the count and those of the jittered outcome
- Nested error model for quantiles of jittered outcome
- Estimates of the quantiles of discrete y obtained via appropriate backtransformation

Theory - Consistency (Weidenhammer, 2016)

For a fixed $q \in (0,1)$ the quantile estimator

$$Q_{y|\boldsymbol{x},k}(q|\boldsymbol{x},k) = \boldsymbol{x}_{ik}^T \boldsymbol{\beta}_q + v_k \qquad k = 1, ..., D, \text{ and } i = 1, ..., n_k$$

is consistent for $D \to \infty$ and $n_k \to \infty$.

Proof (Steps):

1. Consistency of the parameter estimates $\boldsymbol{\theta} = (\boldsymbol{\beta}_q, \sigma_q, \sigma^2_{v,q})$
   - Define $\hat{\boldsymbol{\theta}}$ to be the maximum likelihood estimator of $\boldsymbol{\theta}$
   - Showing consistency of $\hat{\boldsymbol{\theta}}$ relies on Weiss (1971, 1974)
   - The construction of the proof follows Miller (1977) & Pinheiro (1994)
   - Two conditions for showing asymptotic normality $\hat{\boldsymbol{\theta}}$
2. Consistency of the random effect v_k

Theory - Consistency (Step 1)

Let $\ell(\theta|y)$ be the likelihood of the observed data y and $\boldsymbol{\theta}^0$ is the unknown value of $\boldsymbol{\theta}$

- Assumption 1: ✓

$$-\frac{1}{K(n)}\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ell(\theta|y)\Big|_{\theta^0} \xrightarrow{P} B(\theta^0)$$

  where $B(\theta^0)$ is a positive definite matrix

- Assumption 2: ✓
  Convergence in 1 has to be at a certain rate for all $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}^0$

UNIVERSITÀ DI PISA

Freie Universität Berlin

Overview

Model-based simulations

Population data is generated for $D = 50$ areas with $N = 200$ via

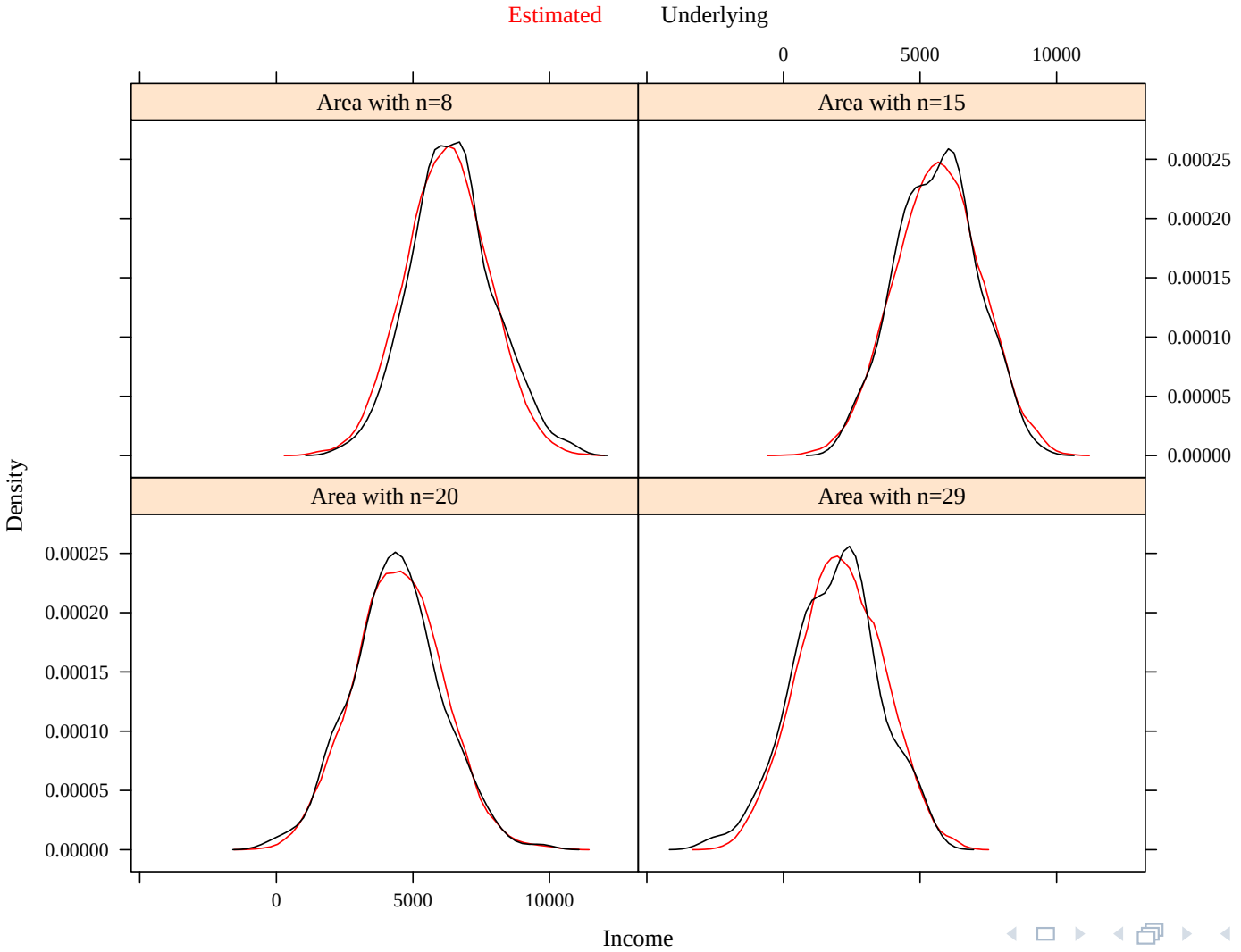
$$y_{ik} = 4500 - 400x_{ik} + u_k + \epsilon_{ik}$$

- Covariates $x_{ik} \sim N(\mu_k, 3^2)$ with $\mu_k \sim U(-3, 3)$
- Random effects $u_k \sim N(0, 500^2)$
- Unbalanced design leading to a sample size of $n = 921$ ($min = 8$, $mean = 18.4$, $max = 29$)
- 100 Monte Carlo replicates with B=100 bootstraps

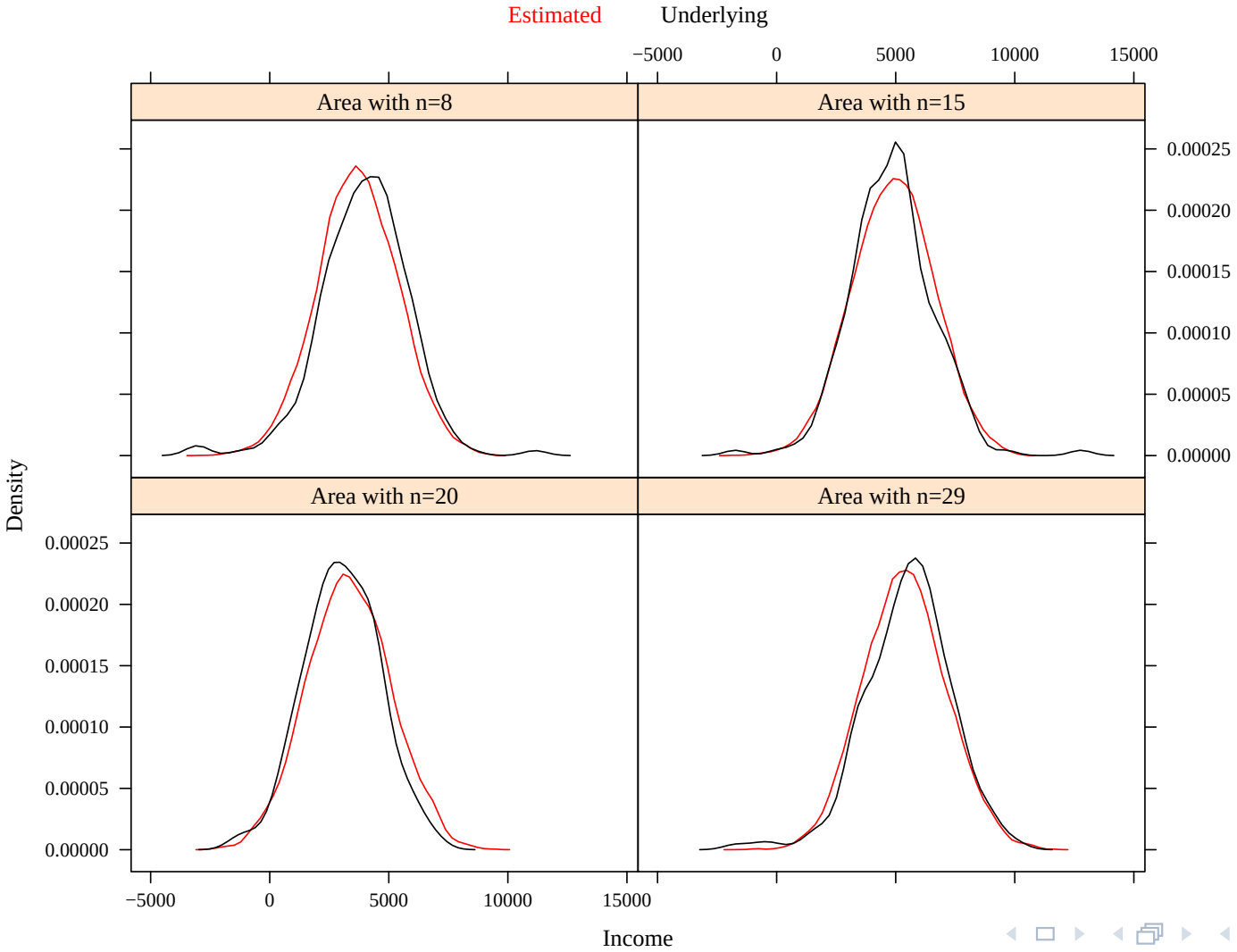
Scenarios:

- Normality: $\epsilon_{ik} \sim N(0, 1000^2)$
- Contamination: $\epsilon_{ik} \sim 0.98N(0, 1000^2) + 0.02N(0, 6000^2)$
- Heteroscedasticity: $\epsilon_{ik} = (1 + 0.1x_{ik})e$ with $e \sim N(0, 1000^2)$

Estimating the Domain Distribution - Normality



Estimating the Domain Distribution - Contamination



Quality Measures

Root mean square error:

$$RMSE_k = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}_{k,r} - \theta_{k,r}\right)^2}$$

Relative bias [%]:

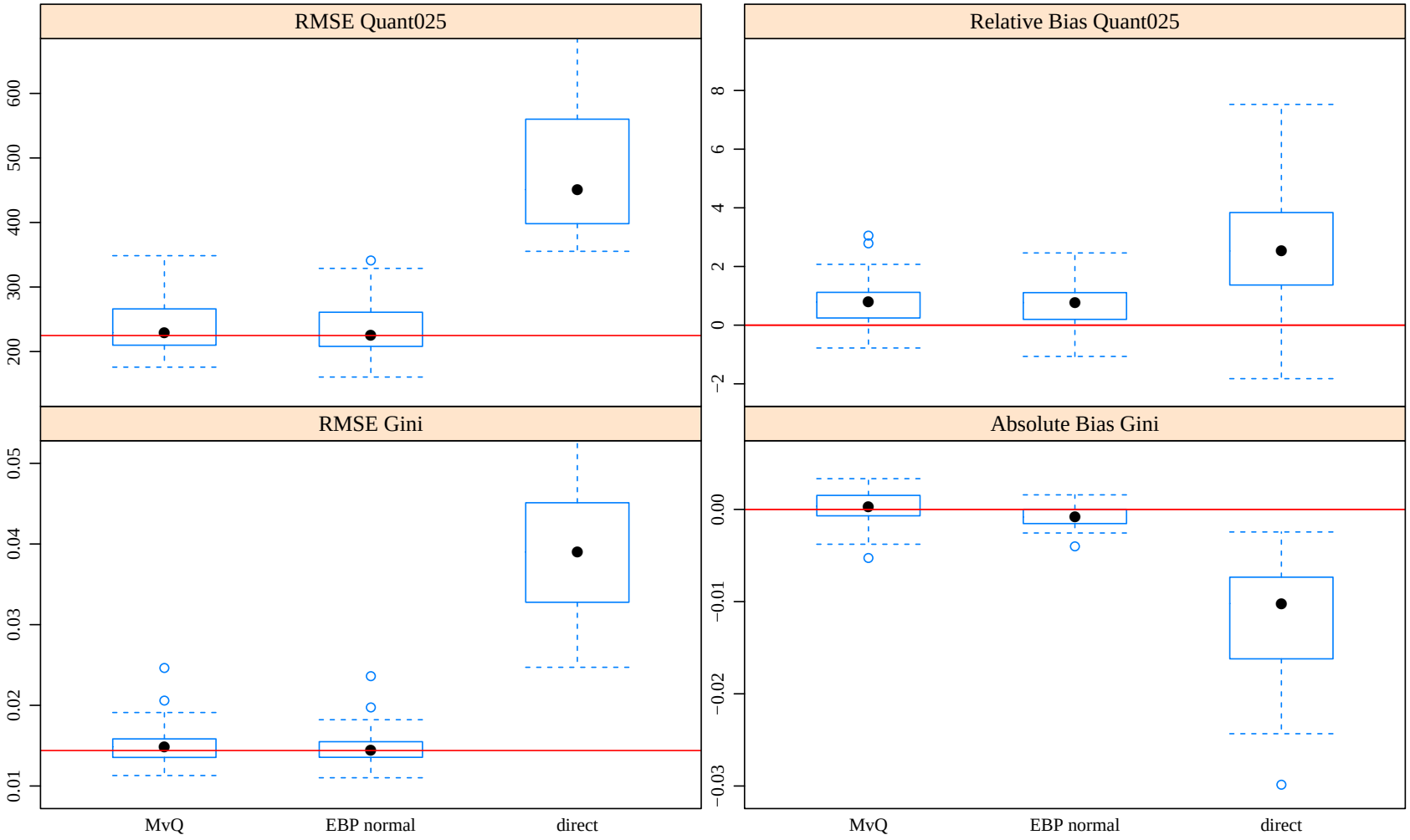
$$RB_k = \frac{1}{R}\sum_{r=1}^{R}\frac{\hat{\theta}_{k,r} - \theta_{k,r}}{\theta_{k,r}} \cdot 100$$

Absolute bias [%]:

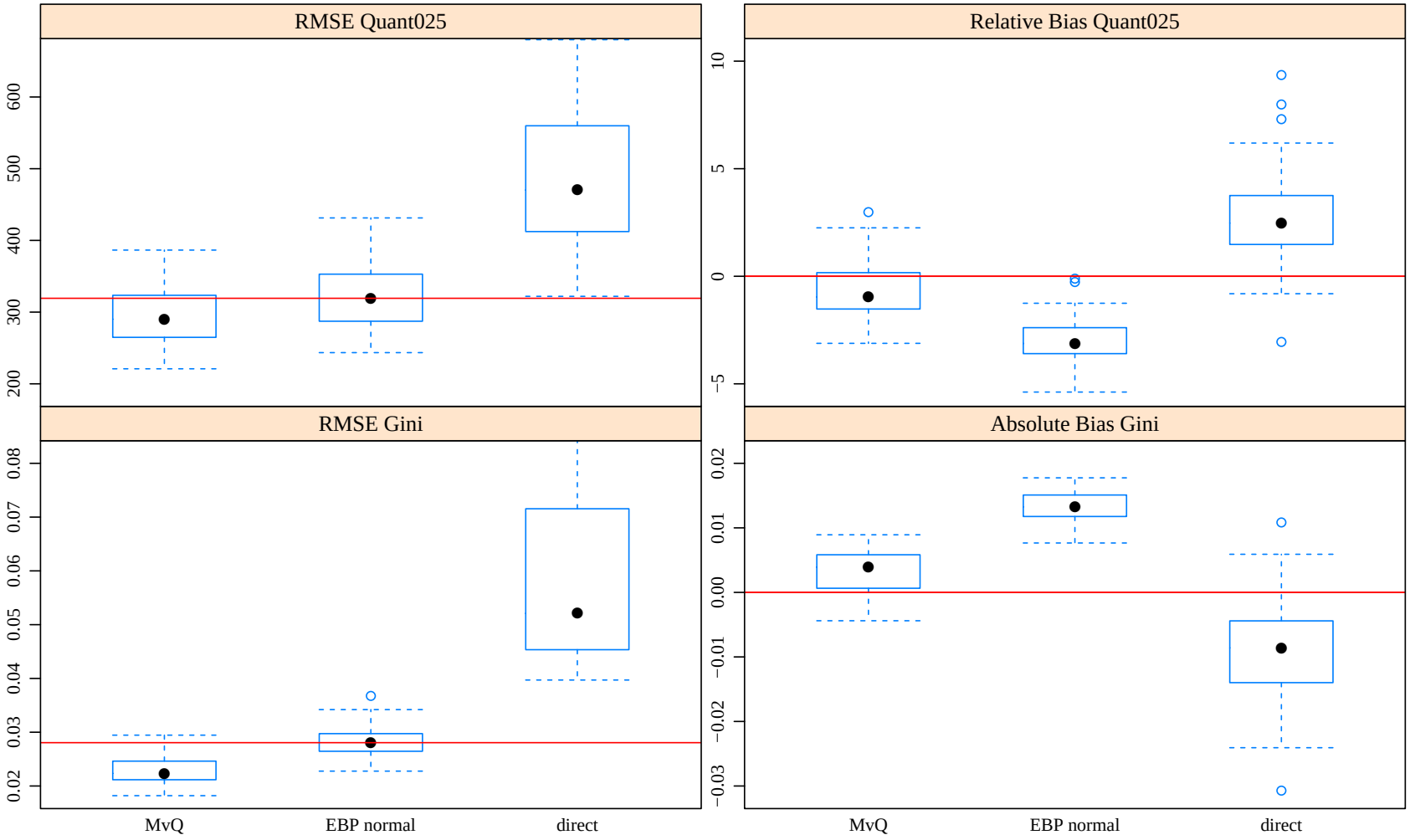
$$Bias_k = \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}_{k,r} - \theta_{k,r}$$

Target parameters: Gini coefficient (Gini) and 25% quantile of the small areas

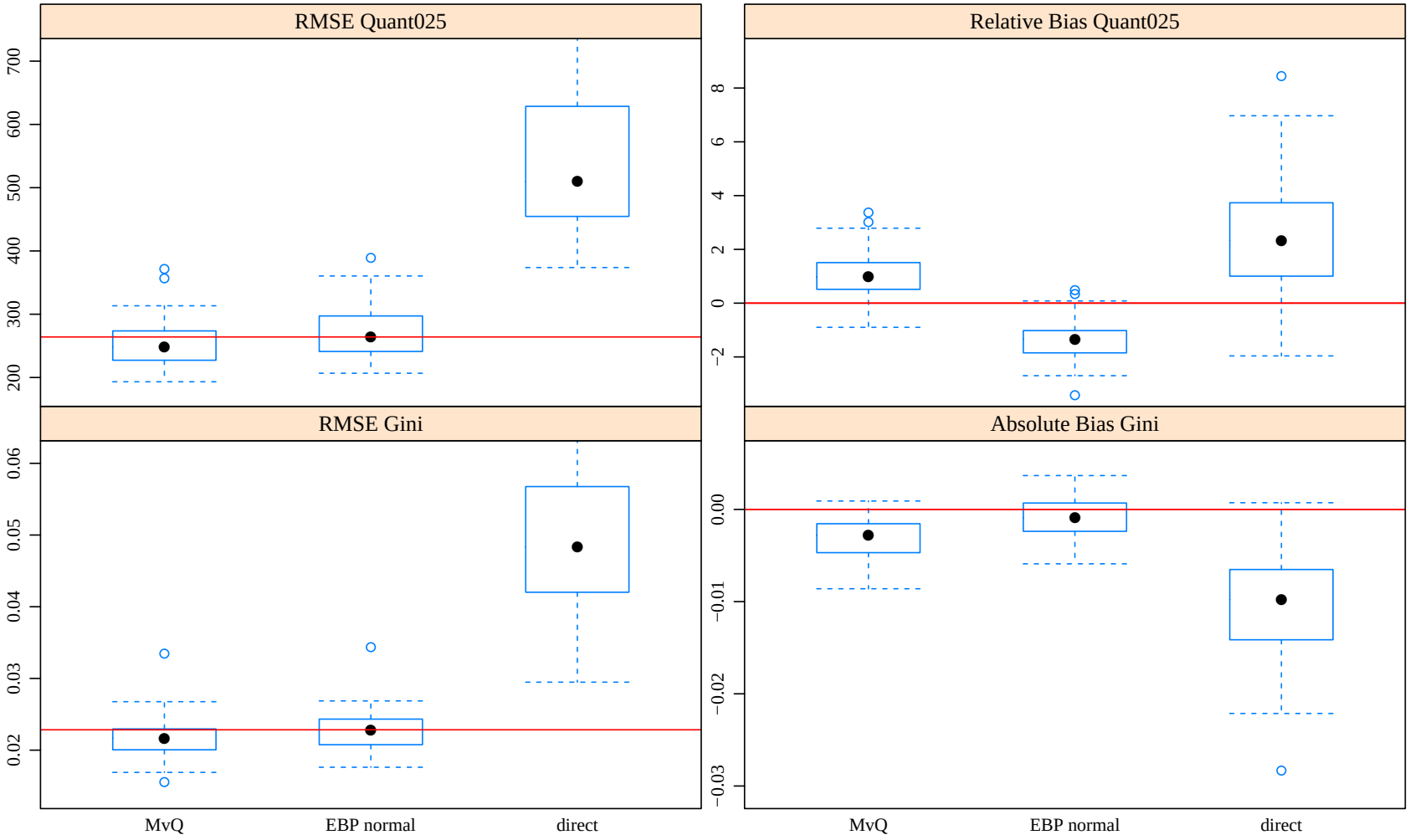
Performance under Normality



Performance under Contamination



Performance under Heteroscedasticity



Remarks and next steps

Remarks:

- Methodology can be used for domain estimation with count outcomes.
- Proposed bootstrap approach (count and continuous case) reveal promising results.

Next steps:

- Constrained quantile regression to avoid quantile crossing.
- Relax parametric assumptions regarding the random effects.

Thank you very much for your attention.

Timo Schmid
Nikos Tzavidis
Beate Weidenhammer
Nicola Salvati