Selection of small area estimation method for Poverty Mapping: A Conceptual Framework

Sumonkanti Das

National Institute for Applied Statistics Research Australia

University of Wollongong

The First Asian ISI Satellite Meeting on Small Area Estimation

02 September 2013

Chulalongkorn University, Bangkok, Thailand

NIASRA
NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

UNIVERSITY OF WOLLONGONG

Outline

- **Poverty, Poverty Mapping & Poverty Indicators**
- **Small Area Estimation (SAE) methods of poverty mapping**
- **SAE methods for Unit-level data**
  - **World Bank Method** (Elbers, Lanjouw and Lanjouw, 2003)
  - **Empirical Best Prediction Method** (Molina and Rao, 2010)
  - **M-Quantile Method** (Tzavidis, Salvati, Pratesi & Chambers, 2008)
- **Comparison & application of these methods on a simulated data**
- **Issues regarding selection of SAE method for Poverty Mapping**
- **Conceptual Framework for Poverty Mapping Study**

Poverty and Poverty Mapping

- **Poverty:** An economic condition where the basic needs required to comfortably live are lacking
- Common basis of poverty measurement: **Income /Consumption level**
- A person is considered poor if his/ her consumption or income level falls below the "**Poverty Line**"
- **Poverty line**
  - ♣ Minimum level of income supposed adequate in a given country
  - ♣ Total cost of all essential resources consumed by an average human adult in one year (Ravallion, Chen & Sangraula, 2008)
- **Poverty Mapping**

  A process to show the spatial distribution of poverty within a country

Poverty Indicators

Poverty Incidence (Head Count Rate):

Proportion of the population whose income or consumption level is below the poverty line

Poverty Gap (Depth of Poverty):

Expected income or consumption shortfall for people living below the poverty line relative to the poverty line

Poverty Severity (Squared Poverty Gap):

Expected squared shortfall of income or consumption for people living below the poverty line relative to the poverty line

Referred to as FGT poverty indicators (Foster, Greer and Thorbecke, 1984)

Poverty Indicators

$\boldsymbol{N_d}$: Population size of d^{th} area

$\boldsymbol{E_{dj}}$: Income or consumption for individual *j* in domain *d*

$\boldsymbol{t}$: Poverty line

FGT poverty measures for d^{th} area:

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

where $I\left(E_{dj} < t\right) = 1 \; if \; E_{dj} < t, \; otherwise \; 0$

$$F_{\alpha d} = \begin{cases} Poverty\ Incidence & if\ \alpha = 0 \\ Poverty\ Gap & if\ \alpha = 1 \\ Poverty\ Severity & if\ \alpha = 2 \end{cases}$$

Small Area Estimation (SAE) methods of poverty mapping

- Availability of Auxiliary data
- Spatial Correlation in data
- Outlier presence in data

Unit Level Model: Auxiliary variables are available for all population units

Area Level Model: Area wise auxiliary variables are available for all areas

Unit Level Model	Area Level Model
World Bank Method (ELL) **Empirical Best Prediction (EBP) Method** **M-Quantile (MQ) Method** Fast EB Method Spatial M-Quantile Method	Fay-Herriot Model Spatial Fay-Herriot Model Semi-parametric Fay-Herriot Model Spatio-Temporal Fay-Herriot Model

World Bank Method (ELL)

$$(\boldsymbol{LogIncome})_{cj} = \boldsymbol{E}\big(\boldsymbol{Log\ Income}_{cj}\big) + (\boldsymbol{ClusterEffect})_c + (\boldsymbol{Individual\ Effect})_{cj}$$

$$\left.\begin{array}{c} Y_{cj} = X_{cj}\beta + u_c + e_{cj} \\ u_c \sim iid\ N(0, \sigma_u^2)\ ; \quad e_{cj} \sim iid\ N(0, \sigma_e^2) \end{array}\right\}$$

$Y_{cj} = \log\big(E_{cj}\big)$ =Log-transformed Income or Expenditure

X_{cj} = Auxiliary variables available for whole population from Census/GIS database

Basic Procedure

- ✓ Develop the regression model using survey data at household level
- ✓ Utilize the developed model to generate B (say, B=1000) independent bootstrap populations
- ✓ Calculate poverty estimate $\left\{F_{\alpha d}^{*(b)}\right\}$ for each small area aggregating the predicted census observations
- ✓ Calculate $\hat{F}_{\alpha d}^{ELL} = B^{-1} \sum_{b=1}^{B} F_{\alpha d}^{*(b)}$

Empirical Best Prediction (EBP) Method

Random **area effect** rather than random **cluster effect**

$$(\boldsymbol{Log\ Income})_{\boldsymbol{dj}} = \boldsymbol{E}\left(\boldsymbol{LogIncome}_{\boldsymbol{dj}}\right) + (\boldsymbol{Area\ Effect})_{\boldsymbol{d}} + (\boldsymbol{Individual\ Effect})_{\boldsymbol{dj}}$$

Prediction estimator

$$\hat{F}_{\alpha d} = {N_d}^{-1}\left[\sum_{j \in s_d} F_{\alpha dj} + \sum_{j \in r_d} \hat{F}_{\alpha dj}\right]; \alpha = 0,1,2$$

- ✓ Generate L independent realisations $\{\boldsymbol{y}^{*}_{\boldsymbol{rd}}; l = 1,2, \ldots, L\}$ of $\boldsymbol{y}_{\boldsymbol{rd}}$ from the distribution of $\boldsymbol{y}_{\boldsymbol{rd}}|\boldsymbol{y}_{\boldsymbol{sd}}$ through **Monte Carlo simulation**
- ✓ Calculate $\hat{F}^{*}_{\alpha d}$ from the vectors $\boldsymbol{y}^{*}_{\boldsymbol{d}} = [\boldsymbol{y}^{\boldsymbol{t}}_{\boldsymbol{sd}}, \boldsymbol{y}^{*\boldsymbol{t}}_{\boldsymbol{rd}}]$
- ✓ Calculate $\hat{F}^{EB}_{\alpha d} \approx L^{-1} \sum_{l=1}^{L} \hat{F}^{*}_{\alpha d}$

M-Quantile (MQ) Method

- ELL and EBP are based on random effects models with
  - strong distributional assumptions
  - additive random effects
  - no easy way of doing outlier robust inference
- M-Quantile SAE
  - distribution free and allows outlier robust inference

Basic idea of M-quantile SAE Method

- ✓ Conditional variability across the population of interest is characterized by the M-quantile coefficients (θ_i) of the population units
- ✓ Population units within an area have similar M-quantile coefficients
- ✓ Between area variation is captured by **area-specific M-quantile coefficients** (θ_d) instead of random effects

Monte Carlo simulation approach of Marchetti, Tzavidis and Pratesi (2012)

- ✓ Estimate area-specific M-quantile coefficients $\left(\hat{\theta}_d\right)$ and hence calculate the M-quantile regression coefficient $\hat{\beta}\left(\hat{\theta}_d;\psi\right)$ using IWLS algorithm
- ✓ Generate an out of sample vector of size $(N_d - n_d)$ using

$$y_i^* = \boldsymbol{x_i^t}\hat{\beta}\left(\hat{\theta}_d;\psi\right) + e_i^* \quad ; \quad i \in r_d$$

e_i^* is drawn from the empirical distribution of overall sample residuals

- ✓ Repeat the process H times and calculate H estimates of $F_{\alpha d}$ $\left(\text{say}, \boldsymbol{\widehat{F}_{\alpha d}^{(h)}}\right)$ combining sample and non-sample $y_{id}'s$ in each process.
- ✓ Calculate $\widehat{F}_{\alpha d}^{MQ} = \boldsymbol{\frac{1}{H}\sum_{h=1}^{H}\widehat{F}_{\alpha d}^{(h)}}$

Poverty Mapping Study in Bangladesh

- BBS and UNWFP (2004) conducted a poverty mapping study in Bangladesh
  - 5% of the EAs of each sub-district from Bangladesh Population & Housing Census 2001
  - Bangladesh Household Income and Expenditure Survey (HIES) 2000

Parameter	Description	Values
M	No. of total areas	507
m	No. of sample areas	295
M-m	No. of out of sample areas	212
C	No. of total clusters	12,170
c	No. of sampled cluster	442
N	No. of total household (HH) units	1,258,222
n	No. of sampled HH	7,824
$\hat{\sigma}_u^2$	Between cluster variation	0.1315
σ_e^2	Individual variation	0.6961
R^2	Coefficient of determination	0.59
P	No. of covariates	31

Construction of Simulated data

- ✓ **As explanatory variable, two correlated binary variables (X_1, X_2) are generated from bivariate Bernoulli distribution with parameters**

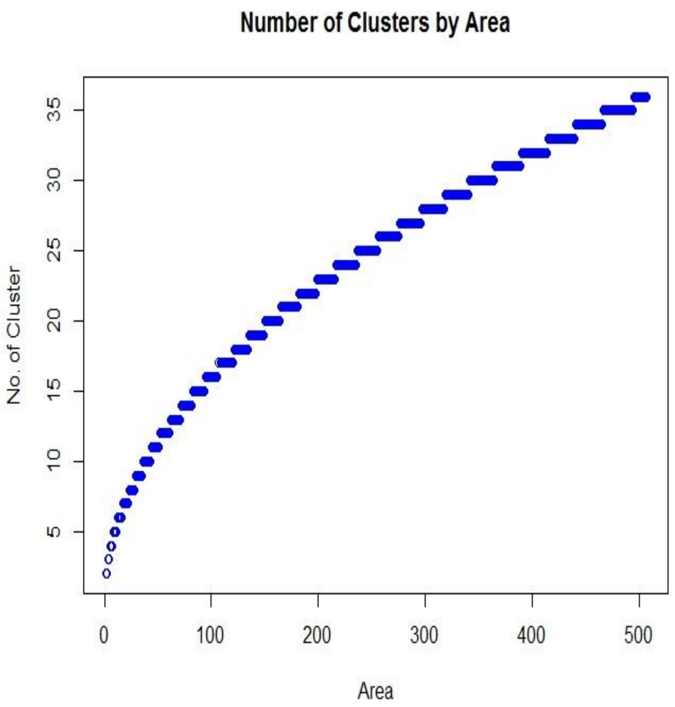
$$\begin{cases} p_{1d} = 0.3 + 0.25\dfrac{d}{507} \\ p_{2d} = 0.25 \\ \rho_d = 0.15\dfrac{d}{507} \end{cases} \qquad d = 1, 2, .., 507$$

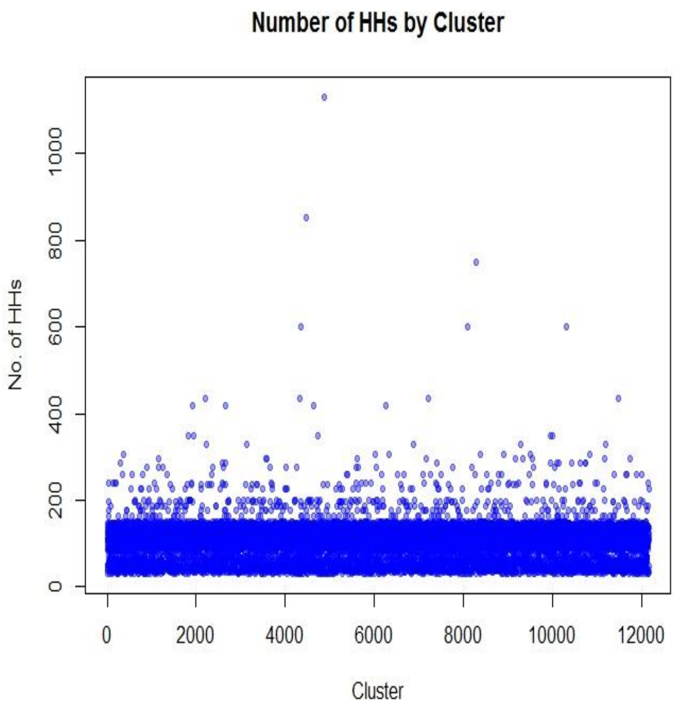
- ✓ **y_{ij}'s are generated in two (02) ways :**

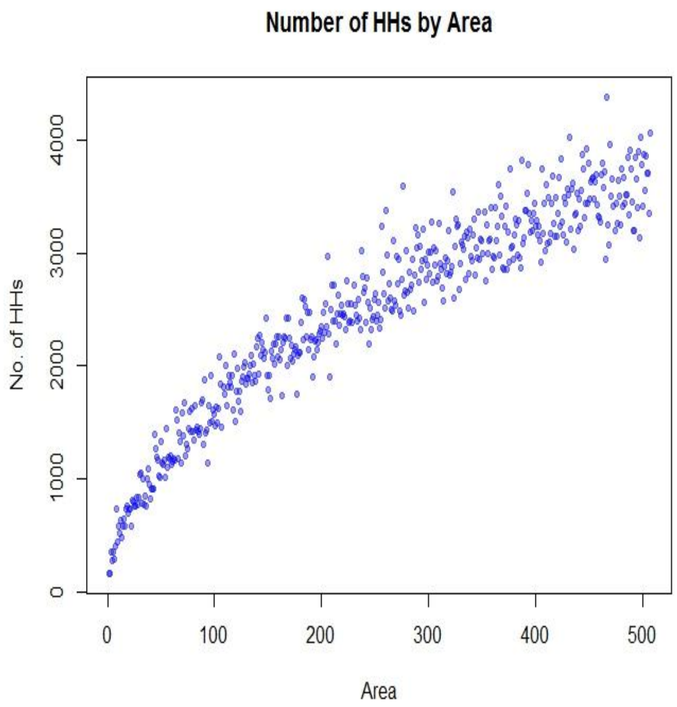
Random Cluster Effect	Random Area Effect
$y_{cj} = 6.6 + 0.16X_1 - 0.15X_2 + u_c + e_{cj}$	$y_{dj} = 6.6 + 0.16X_1 - 0.15X_2 + u_c + e_{dj}$
$u_c \sim N(0, 0.14)$ $e_{cj} \sim N(0, 0.70)$	$u_d \sim N(0, 0.14)$ $e_{dj} \sim N(0, 0.70)$
$c = 1, 2, .., 12170$; $j = 1, 2, \ldots, N_c$	$d = 1, 2, .., 507$; $j = 1, 2, \ldots, N_d$

Structure of the Simulated Data Set

Distribution of Clusters & HHs by Area

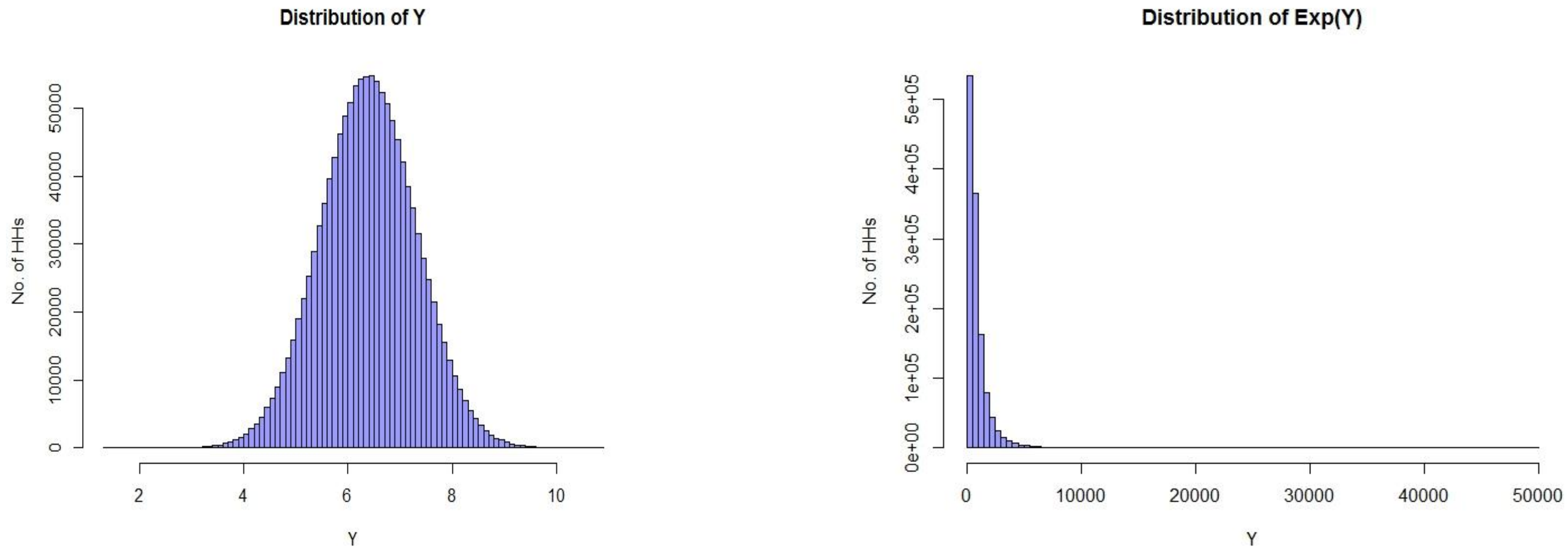




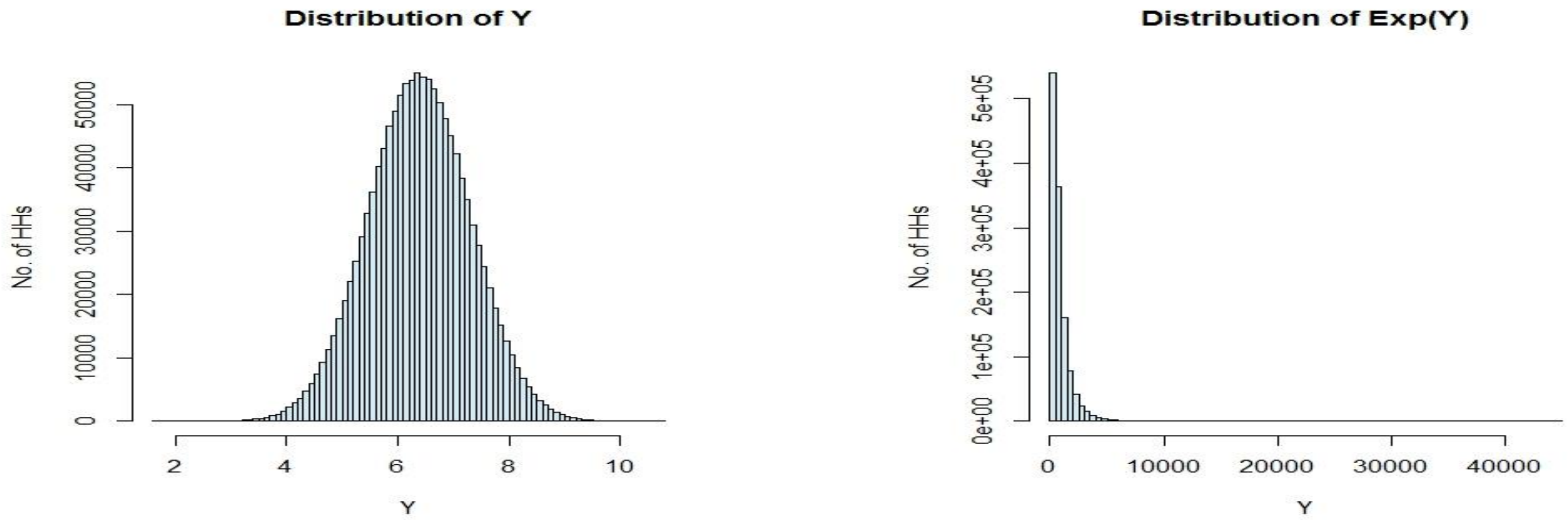


Distribution of Y: log(Income) and Exp(Y): Income

Cluster Effect



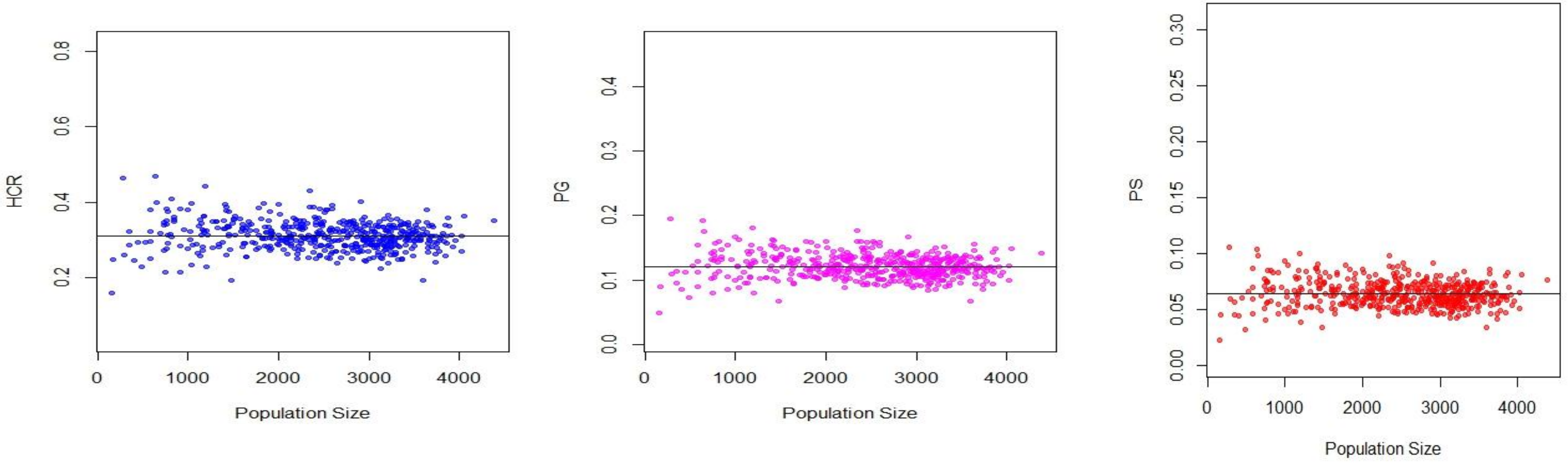
Area Effect



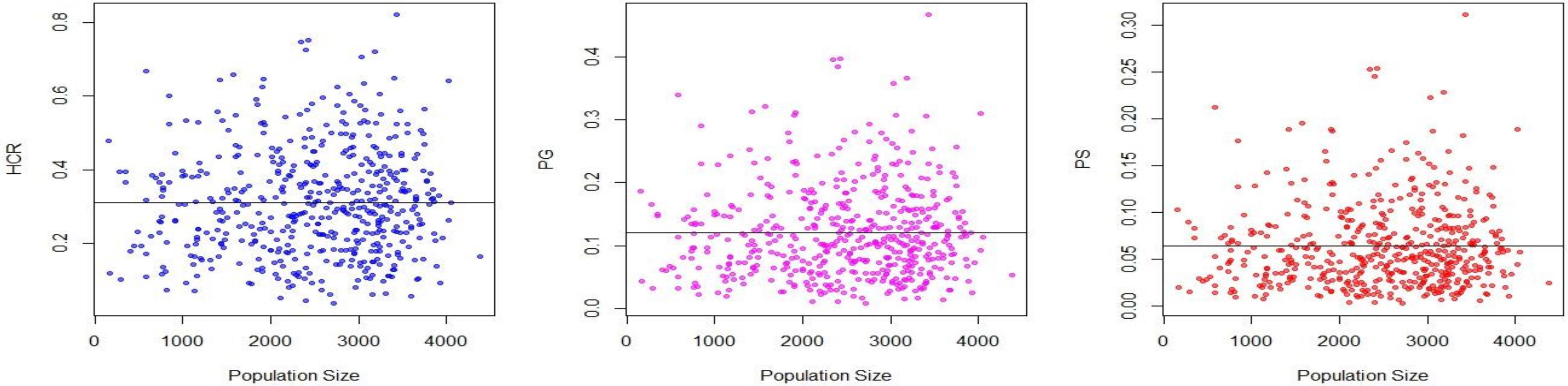
Distribution of FGT 0, FGT 1 & FGT 2 by Area Size

Head Count Rate (HCR): FGT 0	Poverty Gap (PG): FGT 1	Poverty Severity (PS): FGT 2

Random Cluster Effect



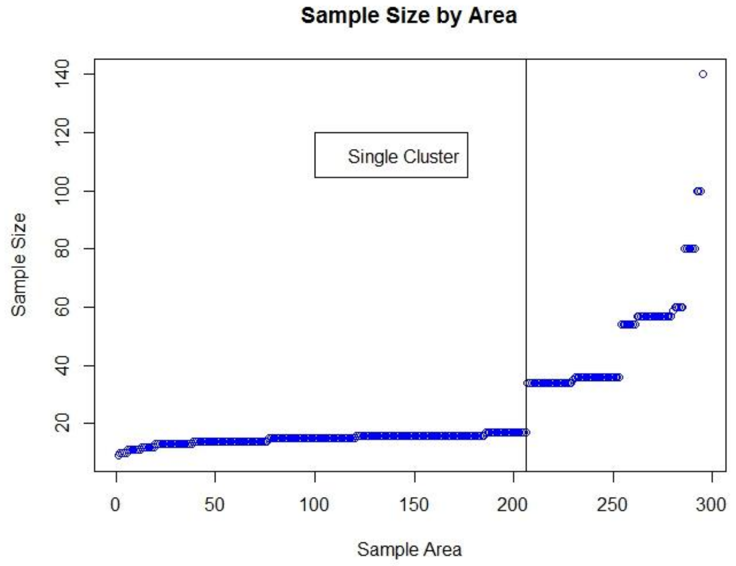
Random Area Effect



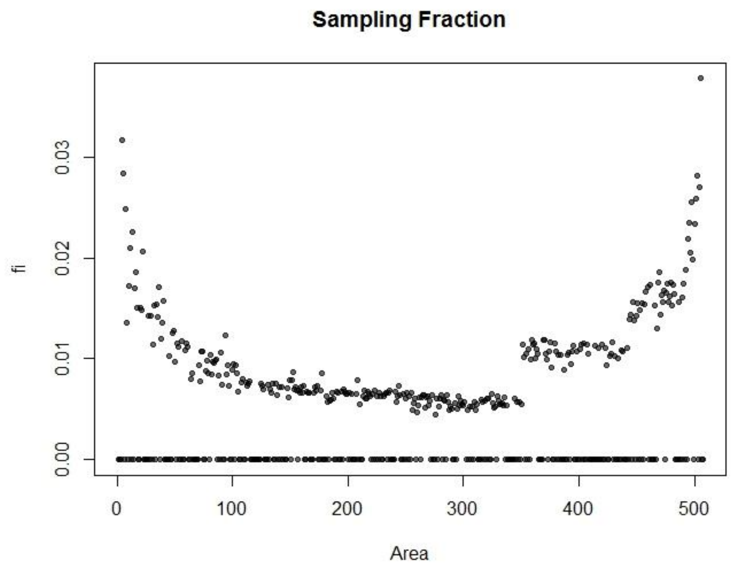
Sampling and Sampling Fraction

Description of Sample

- ✓ **7428 HHs are selected following three-stage random sampling**
- ✓ **9-20 HHs are selected from the selected clusters (442) belong to the selected areas (295)**
- ✓ **About 70% selected areas (206) have only single cluster.**



$$f_i = 0.00 \ : 42\%$$
$$f_i < 0.01 \ : 33\%$$
$$0.01 \le f_i < 0.04 \ : 25\%$$



Design-Based Monte-Carlo Simulation Study

Correlations among Estimates of FGT 0: Sample Areas

Random Cluster Effect

	ELL	EBP	MQ
True	0.0921	0.8951	0.8874
ELL		0.2301	0.2639
EBP			0.9859

Random Area Effect

	ELL	EBP	MQ
True	-0.058	0.9938	0.9969
ELL		-0.047	-0.028
EBP			0.9938

Correlations among Estimates of FGT 0: Non-Sample Areas

Random Cluster Effect

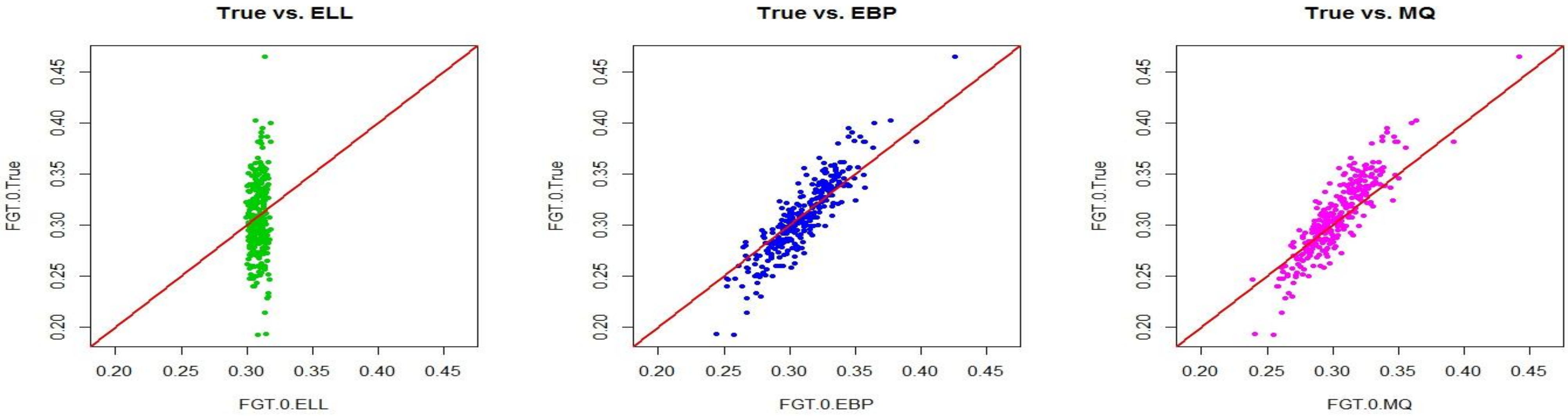
	ELL	EBP	MQ
True	0.1659	0.1694	0.1657
ELL		0.9921	0.9994
EBP			0.9936

Random Area Effect

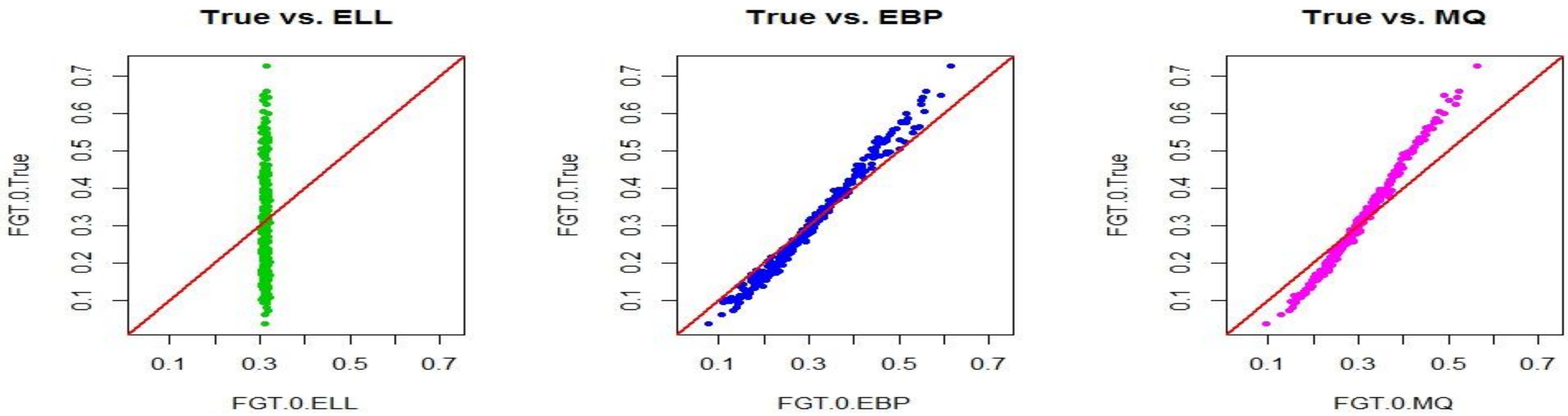
	ELL	EBP	MQ
True	-0.044	-0.055	-0.045
ELL		0.9864	0.9987
EBP			0.9884

Design-Based Monte-Carlo Simulation Study

Estimated Values against True Values: Random Cluster Effect for Sample Areas

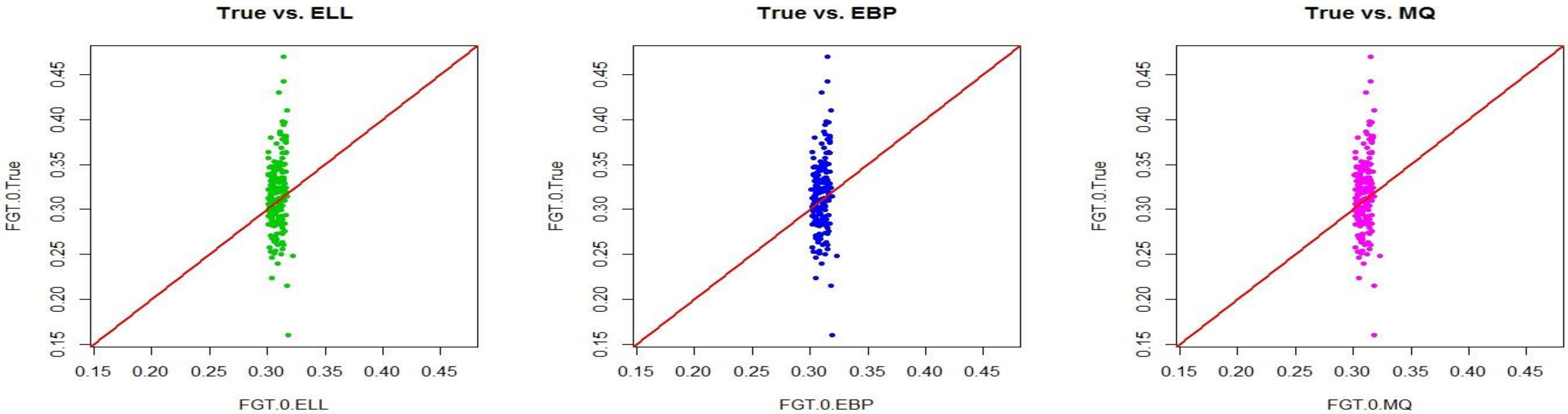


Estimated Values against True Values: Random Area Effect for Sample Areas

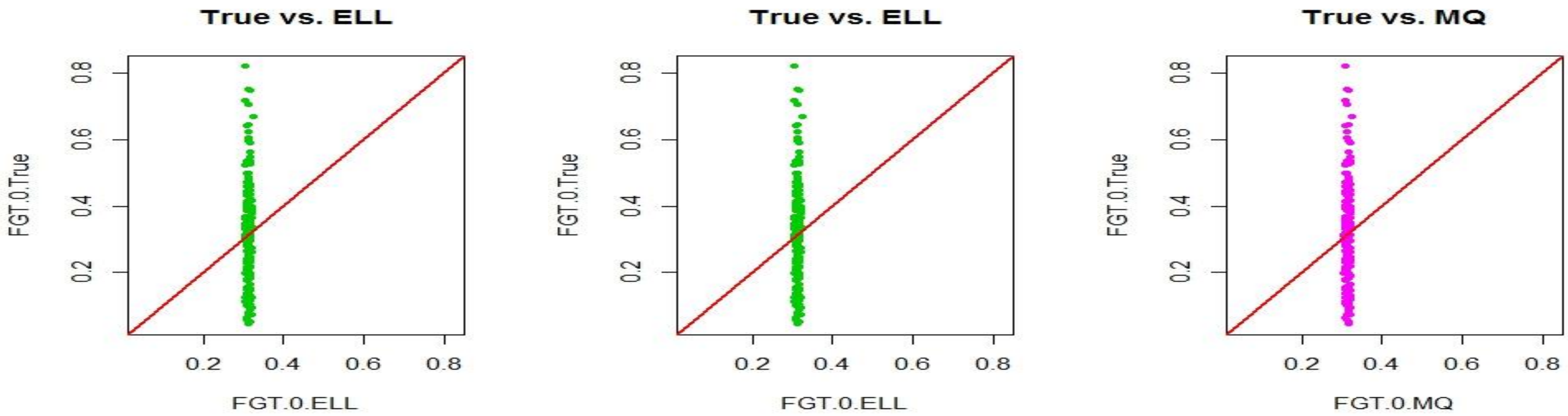


Design-Based Monte-Carlo Simulation Study

Estimated Values against True Values: Random Cluster Effect for Non-Sample Areas



Estimated Values against True Values: Random Area Effect for Non-Sample Areas



Model-Based Monte-Carlo Simulation Study

Correlations among Estimates of FGT 0: Sample Areas

Random Cluster Effect

	ELL	EBP	MQ
True	0.9333	0.7763	0.8160
ELL		0.7834	0.8315
EBP			0.9805

Random Area Effect

	ELL	EBP	MQ
True	0.6018	0.8959	0.8996
ELL		0.6173	0.7745
EBP			0.9418

Correlations among Estimates of FGT 0: Non-Sample Areas

Random Cluster Effect

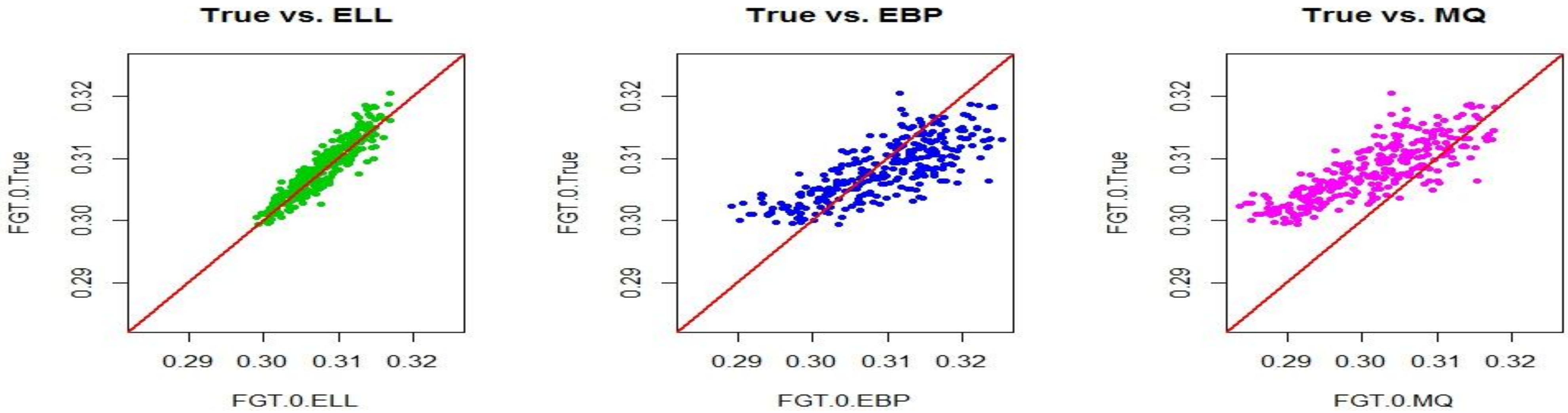
	ELL	EBP	MQ
True	0.9464	0.9411	0.9488
ELL		0.9908	0.9993
EBP			0.9911

Random Area Effect

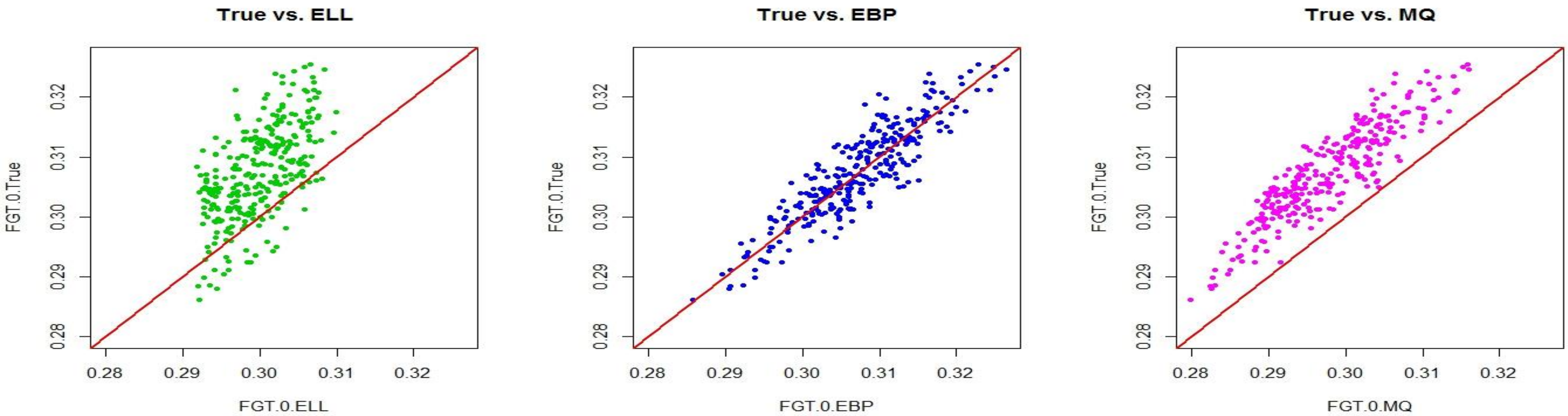
	ELL	EBP	MQ
True	0.5893	0.5810	0.5915
ELL		0.9864	0.9979
EBP			0.9892

Model-Based Monte-Carlo Simulation Study

Estimated Values against True Values: Random Cluster Effect for Sample Areas

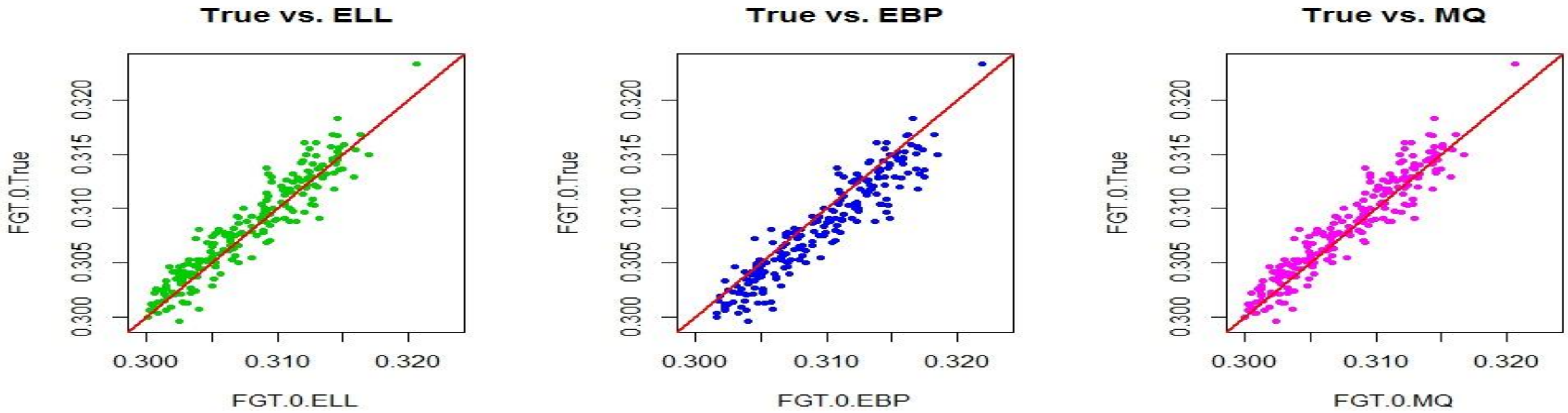


Estimated Values against True Values: Random Area Effect for Sample Areas

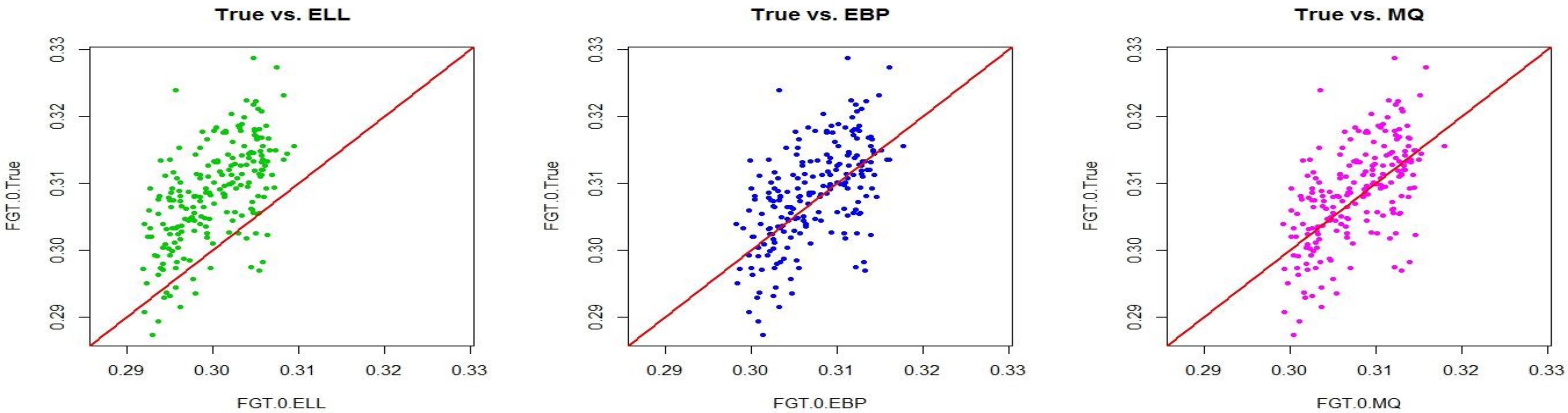


Model-Based Monte-Carlo Simulation Study

Estimated Values against True Values: Random Cluster Effect for Non-Sample Areas



Estimated Values against True Values: Random Area Effect for Non-Sample Areas



Outstanding issues regarding selection of SAE method for Poverty Mapping

Design-Based Study

- ELL method provides synthetic estimate with insufficient between area variability and consequently fails to picture the true poverty situation in case of either random cluster or area effect
- EBP and MQ provide a better result than ELL for the sample areas even when the situation is favourable to ELL, but behaves like ELL for the non-sample areas
- Estimation of accurate FGT indicators for out-of-sample areas is a big problem for all the methods
- Including area effect may improve the ELL estimates beside cluster effect

Model-Based Study

Random cluster Effect

- ELL is doing the best in its favourable condition (random cluster effect) for both sample and non-sample areas
- EBP and MQ behave almost similar to ELL for non-sample areas but fails to track the exact trend for sample areas.

Random Area Effect

- For sample areas, EBP is doing the best. Unfortunately, MQ tracks the trend but underestimates the true values.
- For non-sample areas all the three methods fail to track the trend
- Estimation of accurate FGT indicators for out-of-sample areas is also a big problem for all the methods here

Conceptual Framework for Poverty Mapping Study

1. Selection of poverty indicators and its measurement
2. Detailed study on the sample survey data and the census data
3. Selection of the auxiliary variables
4. **Selection of an appropriate Small Area Estimation (SAE) method**
   - Aggregation level (Area/cluster) where variation is higher
   - Number of areas & sampling fraction
   - Outlier existence in the data
   - Others characteristics like spatial correlation between areas

5. Estimation of the Small area parameter of interest following the considered SAE method in step 4.
6. Diagnostic checking of the estimated parameters
7. Drawing the Poverty Map using the estimates of poverty indicator

This conceptual framework is not only for poverty indicator but also for income/expenditure distribution.

References

BBS and UNWFP (2004) “Local Estimation of Poverty and Malnutrition in Bangladesh”, Bangladesh Bureau of Statistics and the United Nations World Food Program.

Elbers, C., Lanjouw, J. O. and Lanjouw, P. (2003) “Micro-Level Estimation of Poverty and Inequality”, *Econometrica*, Vol. 71, No. 1, pp. 355–364.

Foster J., Greer J., Thorbeck, E. (1984) A class of decomposable poverty measures. *Econometrica*, 52(3):761–66.

Molina I. and Rao, J.N.K. (2010) “Small area estimation of poverty indicators”, *Canad. J. Statistics.*, 38, 369-385.

Ravallion, M., Chen, S. and Sangraula, P. (2008) “Dollar a Day Revisited”, *Policy Research Working Paper*, World Bank, Washington DC.

Tzavidis, N.,Salvati, N.,Pratesi, M., and Chambers, R. (2008), M-quantile models with application to poverty mapping, *Stat Meth Appl* (2008) 17:393–411.