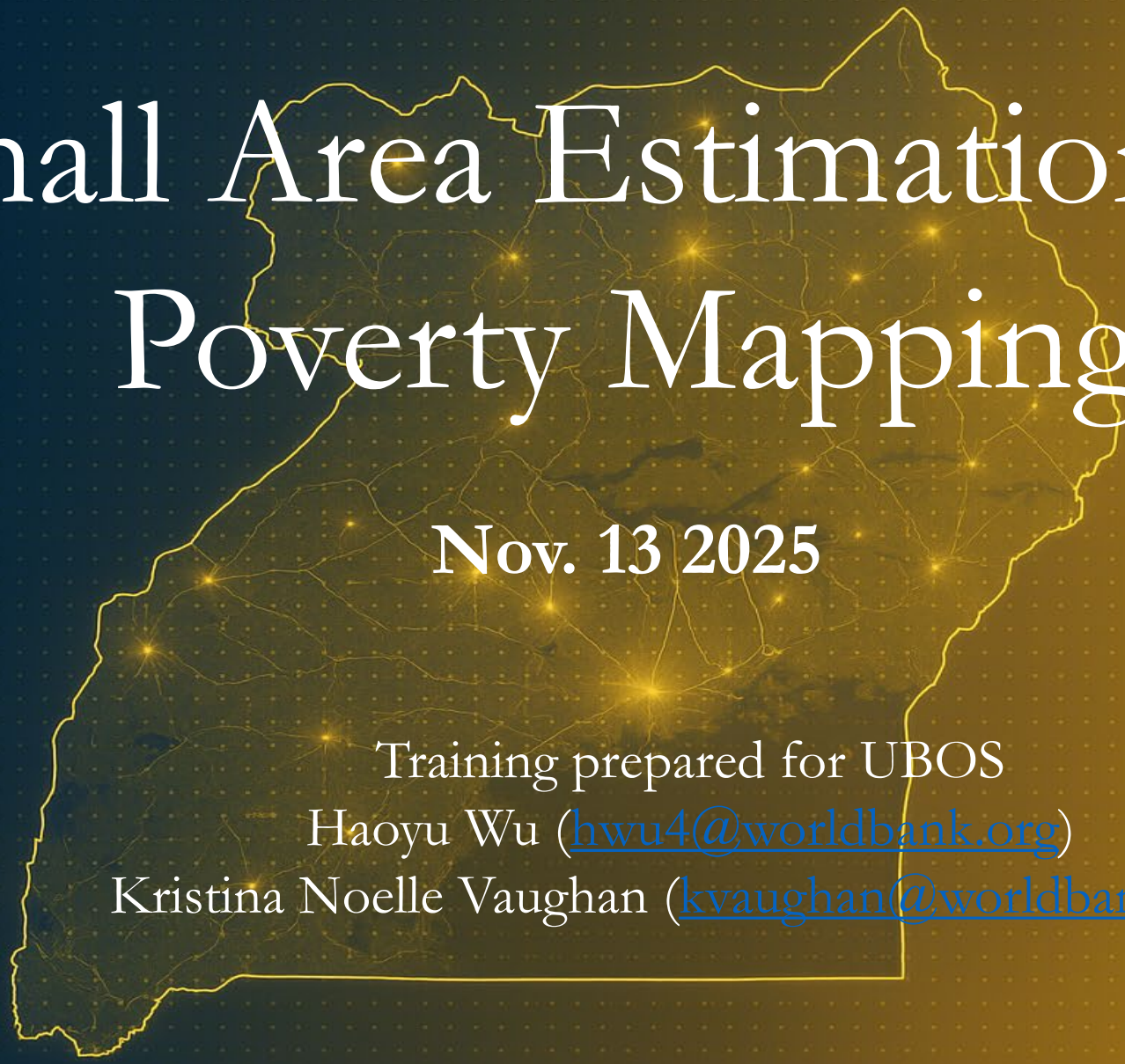
Small Area Estimation for Poverty Mapping



Nov. 13 2025

Training prepared for UBOS

Haoyu Wu (hwu4@worldbank.org)
Kristina Noelle Vaughan (kvaughan@worldbank.org)

Small Area Estimation

Area level models

Introduction

Household surveys are the main sources of indicators of living conditions, poverty, and social exclusion.

- Provide detailed information on multiple indicators of well-being
- Samples too small to be representative for small sub-national units.
- Do not cover all areas

Population censuses

- Provide 100% coverage, permitting assessment for small areas
- Very limited information on poverty and social exclusion indicators

➔ Combine survey and census data to exploit the strengths of each information source. Requires hands-on work with national statistical institutes, use official data sources (population censuses and household surveys) to estimate risk of poverty at lowest possible sub-national level.

➔ **Small Area Estimation** is a branch of statistics focused on improving reliability of estimates and the associated measures of uncertainty for populations where samples cannot produce sufficiently reliable estimates. Poverty mapping relies on survey and census data from the same time frame.

Census microdata is not always available

Census data are infrequently collected and may not be aligned with household survey data

- A key assumption of unit-level model-based SAE is that the census and survey correspond to the same population
  - Questions may have been asked in a different manner
  - Characteristics may differ due to timing of data collection or other reasons
- Data may also be subject to access restrictions

Area-level models

- Combine survey based direct estimates of the desired indicator, at level of interest, and area-level characteristics
  - Survey-based direct estimators are noisy but unbiased
  - Indicator of interest is modeled directly

➔ Combines survey estimates and area-level data (e.g. census aggregates, geospatial data, admin data, etc.) to exploit the strengths of each information source.

➔ **Area-level SAE** will often yield estimates with better precision (less noisy) than those directly derived from the survey data, but because these require less information are often noisier than those from unit-level SAE.

FH Area Level SAE overview

- **Fay-Herriot** models are the traditional approach for cases where access to microdata is not possible or when the census and survey are not aligned
    - Fay Herriot models were introduced to estimate mean per capita income in small areas in the USA (Fay and Herriot 1979)
    - The method consists of modelling poverty rates (or other indicators) at the area level
- The **resulting estimate is a weighted average** between the direct estimates (those derived directly from the survey) and the model-based estimates
    - The weight given to each estimate in a given area depends on the sample size for that area and the quality of the model
    - For areas not in the sample we rely solely on the model-based estimates
- Because the model is only fit on sampled areas as opposed to households, the estimates obtained are often much less efficient than those obtained under unit-level models

Assumed area-level model: Fay-Harriot – set-up

- First stage assumes that the *true* district level poverty rate, $\boldsymbol{y_d}$ for all districts $\boldsymbol{d = 1, \dots, D,}$ is linearly related to a set of district level covariates, $\boldsymbol{x_d}$, through the following linking model:

$$\boldsymbol{y_d = x_d'\beta + u_d} \quad \textbf{(1)}$$

- Random errors (area effects) and represent unexplained heterogeneity between areas, assumed to have a zero mean and constant variance, $\boldsymbol{\sigma_u^2}$.
- The model presented here cannot be fit since the *true* district level poverty estimates are unobserved and instead what is observed are the survey based direct estimates of poverty at the district level, $\boldsymbol{\hat{y}_d^{dir}}$.
- Second stage models the sampling error by assuming the direct estimators are centered around the true district poverty rates:

$$\boldsymbol{\hat{y}_d^{dir} = y_d + e_d} \quad \textbf{(2)}$$

- The errors, e_d, in equation 2 are assumed to be **heteroskedastic**, where $\text{var}[e_d|y_d] = \psi_d$.

Assumed area-level model: Fay-Harriot - solution

- Combine the (1) and (2):

$$\hat{\theta}_d^{DIR} = x_d'\beta + u_d + e_d$$

This is the complete FH model.

- **What Does the Model Actually Do?**
  It merges two sources of information:
  Direct survey estimate (noisy but unbiased)
  Model-based synthetic estimate (smooth but may be biased if the model is weak)
  The final FH estimator is a **weighted average**:

$$\hat{\theta}_d^{FH} = \gamma_d\, \hat{\theta}_d^{DIR} + (1 - \gamma_d) x_d'\hat{\beta}$$

Where:

$$\gamma_d = \frac{\sigma_u^2}{\sigma_u^2 + \psi_d}$$

- Interpretation of the weight:
  If **sampling variance is large** → direct estimate is unreliable → rely more on the model
  If **sampling variance is small** → direct estimate is good → rely more on survey
  If **area has zero sample** → weight on direct estimate is zero → use model prediction only

Assumed area-level model: Fay-Harriot – Sampling Variance

Why do we smooth the sampling variances?

- Direct survey variances $\hat{V}_d$ at district level can be unstable: small samples, singleton PSUs, collapsed strata.
- Unstable variances → unstable FH estimates, convergence issues.
- Variance smoothing improves model performance and stabilizes shrinkage.

Step 1. Log-Linear Variance Model

We model the sampling variances as a function of an area-level size measure:

$$\log(\hat{V}_d) = \alpha + \beta \log(n_d/N) + u_d$$

$\hat{V}_d$ is sampling variance of the direct estimate; n_d =sample size, N is the area population and $u_d$is the model error term

Step 2. Predicted Variance (Exponentiated)

$$\tilde{V}_d = \exp(\alpha + \beta \log(share))$$

This gives a smooth, model-based approximation to the noisy variances.

Step 3. Normalization (Rescaling)

Ensure the total smoothed variance matches the total raw variance: $\tilde{V}_d^{\star} = \tilde{V}_d \cdot \frac{\sum_d \hat{V}_d}{\sum_d \tilde{V}_d}$

This keeps the overall variability constant while smoothing area-level noise.

Why normalize explanatory variables in the Fay–Herriot Model?

- **To compare predictors fairly**
  Different variables are measured on very different scales; standardization makes them comparable.

- **To improve numerical stability**
  Large scale differences can cause unstable estimation or convergence problems.

- **To enhance interpretability**
  Coefficients reflect the impact of a one-standard-deviation change, making comparisons meaningful.

- **To reduce multicollinearity**
  Standardization stabilizes the covariance matrix and lowers inflated correlations.

- **Standard practice in SAE**
  Normalization is widely recommended in the FH literature and World Bank SAE Guidelines.

Introducing the `fhsae` Command in Stata

`fhsae` fits the Fay–Herriot (FH) area-level small area estimation model in Stata. It is translated from the SAE R package by Molina and Marhuenda.

- **What the FH model does**
  Combines direct survey estimates (e.g., district poverty) with sampling variances and auxiliary predictors from census/administrative sources to produce EBLUP estimates for each area

- **When to use `fhsae`**
  When your survey sample is too small for reliable district-level estimates
  When you have census or admin variables to strengthen predictions
  When you want transparent, reproducible FH results in Stata

- **What `fhsae` solves**
  Noisy direct estimates
  Highly variable sampling variances
  Inconsistent shrinkage across areas
  Need for model-based small-area estimates

Fhsae Stata command for Area-level models:
https://github.com/jpazvd/fhsae

Application of Fay-Herriot Model for Ghana

1. Objective
2. Data requirements
3. FH model
4. Direct estimates
5. Model selection
6. Check assumptions
7. Evaluate estimates

Application of Fay-Herriot Model for Ghana

Objective

Producing an updated poverty map for Ghana at the district level using area level models.

- Presents the poverty headcount for all 216 districts in Ghana in 2017

Target indicators: poverty headcount (FGT0) for all 216 districts in Ghana as of 2016/17

The poverty map presented here takes advantage of **district-level** aggregate characteristics derived from the 10 percent sample of the 2010 Population and Housing Census and direct estimates of poverty at the district level obtained from the Ghana Living Standards Survey corresponding to 2016/17. We use a FH model to estimate poverty.

Application of Fay-Herriot Model for Ghana - Data requirements

1. **Direct estimates** of indicators of interest and its sampling variance for the areas considered (from the survey).
   - GLSS7: 14,009 households, nationally and regionally representative
   - We will need poverty estimates at the district level derived from GLSS-7
   - National poverty rate in GLSS7: 23.4%
   - Only missing a sample in a few districts

2. **Aggregate data** at district level of all necessary covariates for the model
   - 2010 Population and Housing Census, 10% sample at the district level.
   - The data provides information on population and housing characteristics for the entire country, the 10 administrative regions and the 170 districts that existed at the time of the 2010 census.
   - District level aggregates from the 2010 census. (we will only require census aggregates, so no need for access to micro data)

3. Additionally, we need a **location variable** to link the census (or any other auxiliar data) and survey at that level.

   → **We make sure both data sets are linkable at the district level.** **This is a must for the method.**

Application of Fay-Herriot Model for Ghana - Direct estimates

- Direct estimators are those that come directly from the survey
  - Horvitz-Thompson (1952) – design unbiased but large variance
  - Hajek (1971) – slightly biased but smaller variance
- Indirect estimators for an area's indicators are those that make use of information of other areas; it borrows strength from other areas.
- We calculate direct estimates of poverty at the **district level** from the 2016/17 GLSS7
- See Corral et al. (2022, Ch2) for a discussion on direct estimates.

Open following do-file: **~01.dofiles/FayHerriot/1.SVY_prep.do**

Application of Fay-Herriot Model for Ghana - Model selection

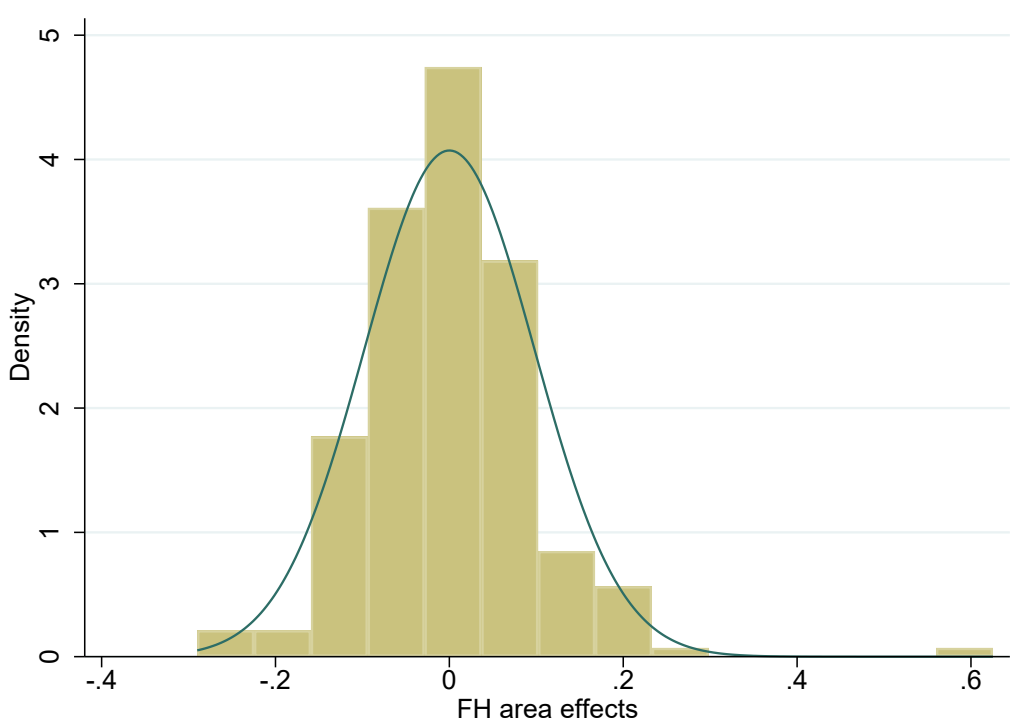
- Covariates are standardized (mean=0, std. dev.=1) before model selection.
- Model selection follows a procedure similar to Corral et al. (2022), gradually removing non-significant covariates.
- The model selection stage uses the FH (Fay Herriot's moments method) due to lower computational requirements.
- Covariates with a VIF above 5 are excluded from the final model.
- The final model employs 8 covariates and an intercept to explain poverty across 210 districts.
- The adjusted R^2 of the model is 0.73, indicating substantial explanatory power.
- Assumptions of the model need to be verified, as mentioned in the introduction.

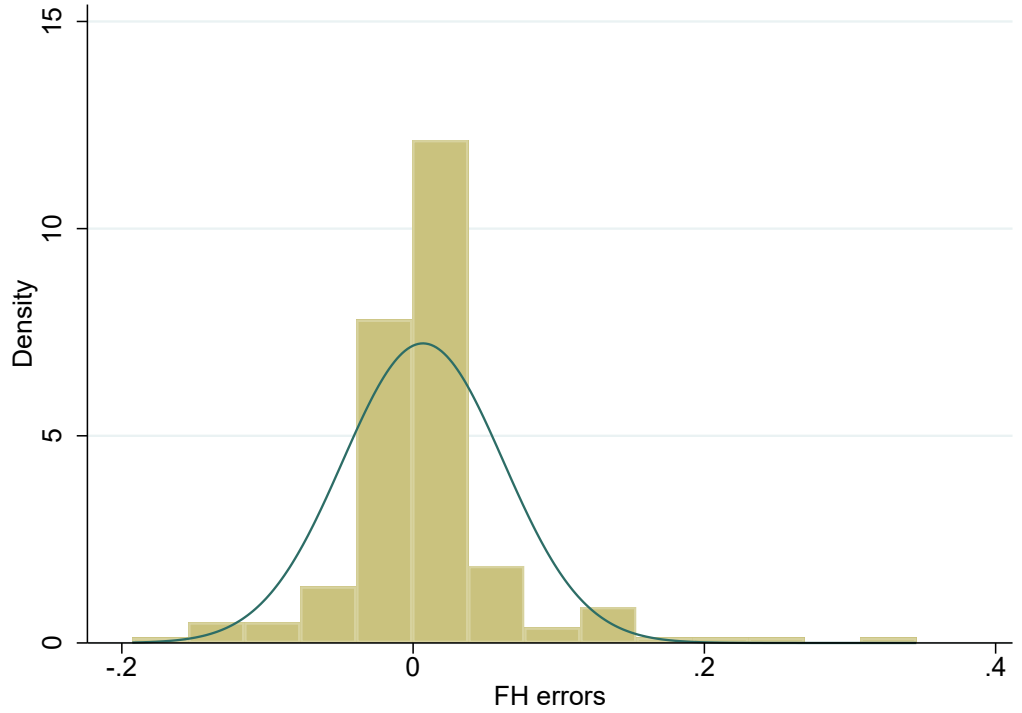
Open following do-file: **~01.dofiles/FayHerriot/2.FH_model_select.do**

Application of Fay-Herriot Model for Ghana - Checking assumptions

Open following do-file: **~01.dofiles/FayHerriot/3.results_check.do**

Figure 1: Fay-Herriot Residual Plots

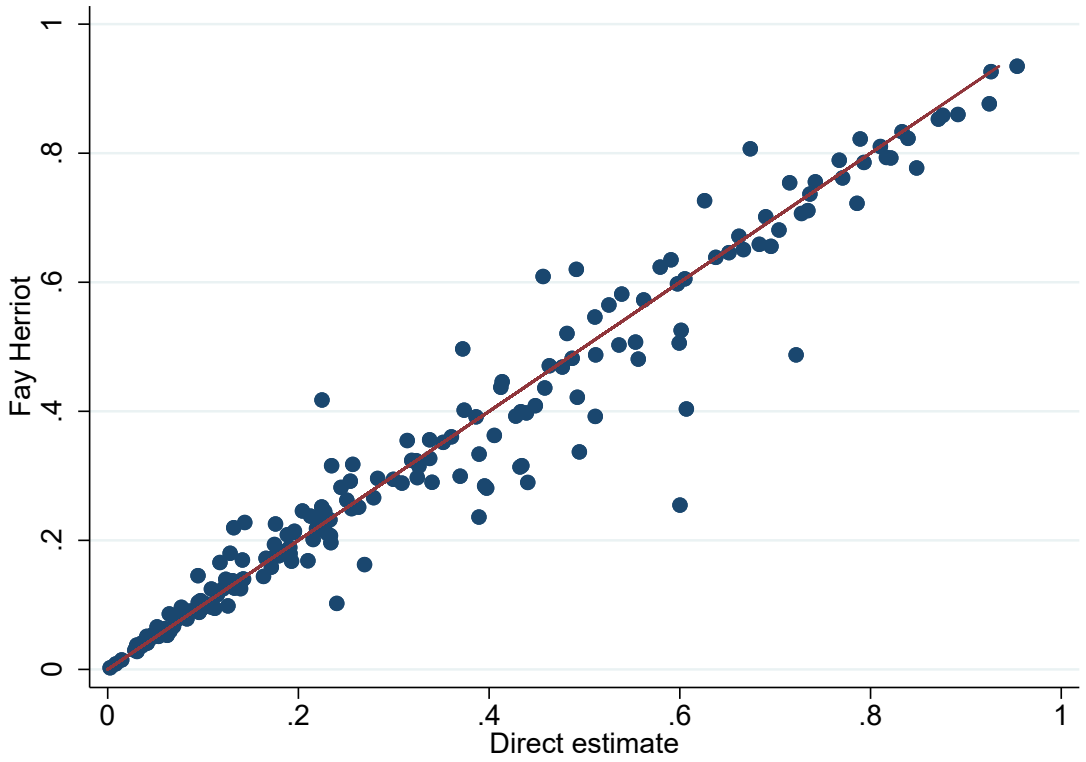


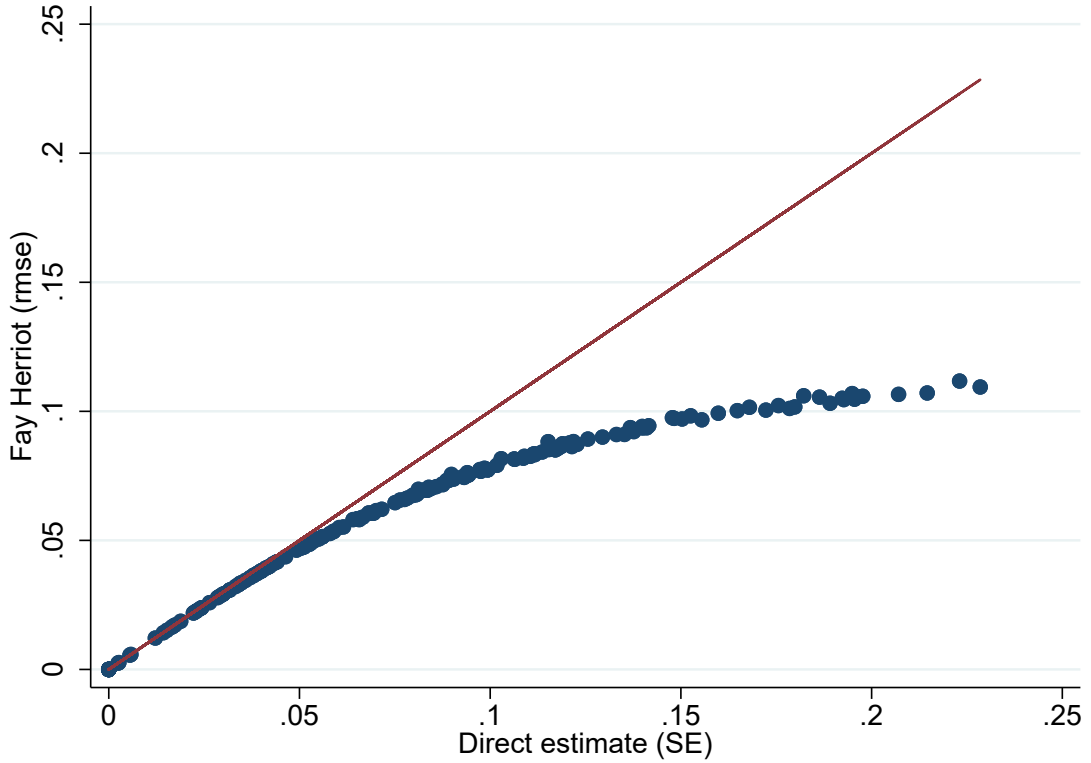


WORLD BANK GROUP

Application of Fay-Herriot Model for Ghana - Estimates' evaluation

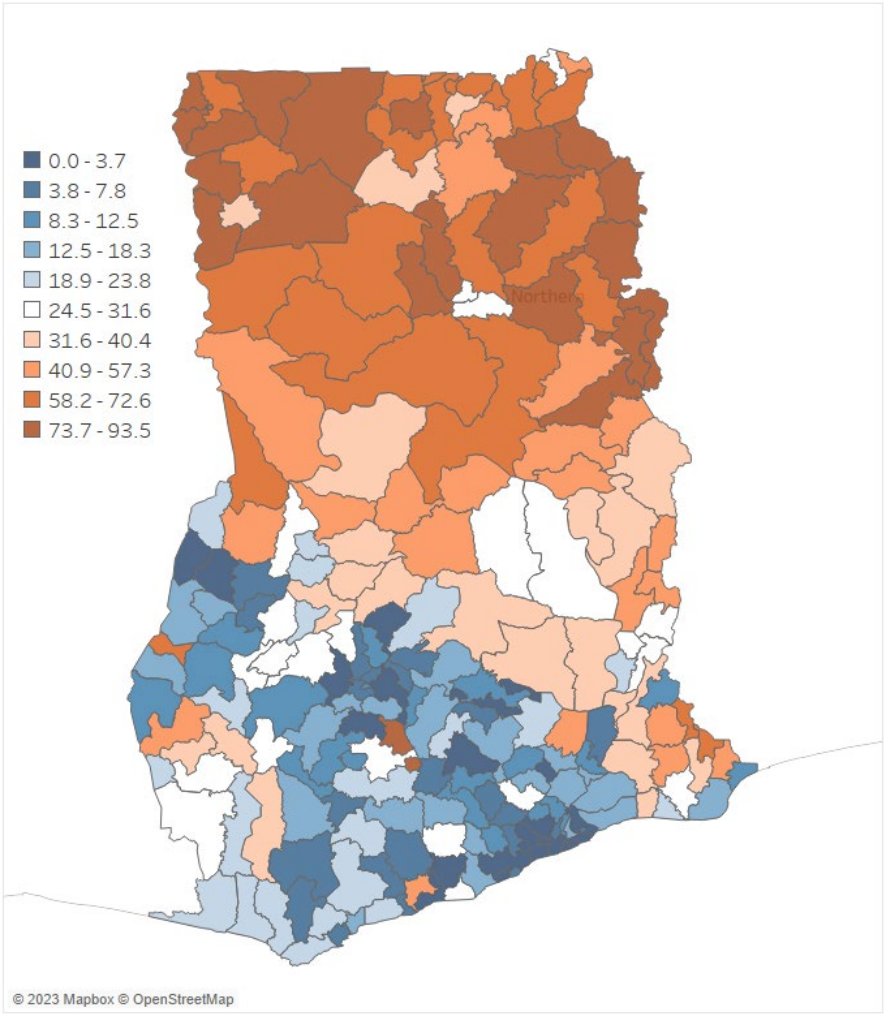
Figure 2: Direct vs Small Area Estimates (Left, point) (Right, SE)





WORLD BANK GROUP

Poverty map - Fay Herriot Small Area Estimates of Poverty (FGT0 deciles)



The poverty map corresponding to 2016/17 for Ghana, at the district level.

Note: Area-level small area estimates of poverty obtained from a Fay Herriot model.

WORLD BANK GROUP

Main references for area-level small area estimation

- Fay, Robert E and Roger A Herriot (1979). "Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data". In: Journal of the American Statistical Association 74.366a,pp. 269–277.
- Molina, Isabel (2019). Desagregación De Datos En Encuestas De Hogares: Metodologías De Estimación En áreas Pequeñas. CEPAL. url: https://repositorio.cepal.org/handle/11362/44214.
- Rao, JNK and Isabel Molina (2015). Small Area Estimation. 2nd. John Wiley & Sons.
- Corral, Paul, Isabel Molina, Alexandru Cojocaru, and Sandra Segovia, 2022. Guidelines to Small Area Estimation for Poverty Mapping. Washington, DC: World Bank.