Remittances Review
August 2024,
Volume: 9, No: 4, pp.2703-2721
ISSN: 2059-6588(Print) | ISSN 2059-6596(Online)

Received: 17 July 2024, Accepted: 28 August 2024
DOI: https://doi.org/10.33282/rr.vx9i2.148

Poverty Mapping at District Level in Pakistan: An Application of ELL Method of Small Area Estimation

Shahid Shabir[1], Shujaat Farooq[2]

[1,2] Pakistan Institute of Development Economics (PIDE), Islamabad.

sshahid.shabir@gmail.com

Abstract:

This study utilizes the Elbers, Lanjouw, and Lanjouw (ELL) method of small area estimation to generate poverty data at the district level in Pakistan, addressing a critical gap in localized information. By employing the unit-level ELL model, we estimated poverty headcount ratios based on the proportion of poor households. The analysis integrates data from the Household Integrated Economic Survey (HIES) 2018-19 and the Pakistan Social and Living Standards Measurement Survey (PSLM) 2019-20. A comparison of Mean Squared Errors (MSE) between direct and model-based estimates reveals that the ELL model-based estimates are more reliable than direct estimates, particularly for districts with limited sample sizes. Our findings unveil significant spatial disparities in poverty across districts, with twelve districts exhibiting extreme levels of poverty and sixteen districts facing severe poverty conditions. This granular assessment of poverty status at the district level provides policymakers with invaluable insights for targeted interventions to alleviate poverty in these regions.

Keywords: Small Area Estimation, Poverty, District-level Analysis, ELL Method, Monte-Carlo Simulations.

Introduction:

The United Nations' Sustainable Development Goals (SDGs) represent a global call to action aimed at eradicating poverty and food insecurity, safeguarding the planet, and ensuring prosperity for all by 2030. One of the important SDGs among these is SDG-1 aims to eradicate poverty in all its manifestations, acknowledging that poverty is not solely a matter of financial hardship but also includes aspects of social exclusion, vulnerability, and marginalization from decision-making processes. However, in the case of developing countries, the data on these welfare variables (like poverty)[1] are not readily available at the

[1] Poverty relates to headcount poverty based on (monthly) per equivalent adult expenditures.

small area level, or if available but the sample size is too small that it is not representative. The solution to the problem is to increase the sample size, which is very costly and sometimes not viable and hinders the policymakers, researchers, and Government agencies from allocating the developing funds and resources to particular small areas relying on those areas' poverty status. Researchers have turned to Small Area Estimation (SAE) techniques to address the challenge of limited data granularity. When we have some variable data at the National (or Provincial) level, and the same data is not available at the district/small area level; we can generate the data of that variable at the district/small area level by using the auxiliary or common information. SAE methods are statistical models designed to enhance the accuracy of estimates for small geographical areas or domains with limited or no sample size. The fundamental principle behind these methods is to borrow strength from auxiliary data sources, such as census or administrative records, through statistical modeling. Model-based SAE techniques commonly fall into two categories: area-level models and unit-level models. Area-level models, including the work of (Fay III & Herriot, 1979) and (Torabi & Rao, 2014), operate on aggregate data for each area. In contrast, unit-level models, such as those proposed by (Elbers, Lanjouw, & Lanjouw, 2003), and (Molina & Rao, 2010) utilize individual-level data. The ELL method is a widely used unit-level small-area estimation technique that has gained prominence in poverty mapping and welfare analysis. The ELL method leverages household survey data alongside auxiliary information from the census to predict individual welfare indicators. The ELL method involves fitting a regression model to the survey data, where the welfare indicator (consumption expenditure) is the dependent variable, and various individual and household characteristics serve as predictors. The estimated model parameters are then applied to the entire population represented in the auxiliary data, generating predicted welfare indicators for each household. These individual predictions are aggregated to the desired small area level, providing estimates of poverty rates or other relevant welfare measures.

This study focuses on the application of the ELL unit-level SAE method. In this study, we generate poverty data at the district level by using HIES 2018-19 and PSLM 2019-20. In the case of Pakistan (Jamal, 2007) and (Begum, 2015) attempted to use the SAE technique to measure Poverty at the district level in Pakistan. These studies are limited in scope and use outdated methods. Our research employs the ELL method (Elbers, Lanjouw, Lanjouw, 2003), to generate comprehensive poverty data at the district level in Pakistan.

Research Methodology:

The ELL method is a revolutionary method for estimating indicators of interest, offered by (Elbers et al., 2003). Gaining renown as the "World Bank Method" due to its extensive use by the World Bank in poverty measurement and poverty map construction. The initial stage of the ELL methodology centers on modeling household welfare, represented by the dependent variable, utilizing survey data to establish its relationship with pertinent explanatory factors. This involves estimating the unconditional variance of the residual parameters, employing the ELL (2003) framework, and getting GLS parameter estimates. This first stage lays the

foundation for the subsequent small area estimation process, ensuring a robust and accurate modeling framework upon which the second stage can build. The second stage of the ELL methodology advances the analysis by incorporating simulations to generate reliable small-area estimates.

To model household per equivalent adult expenditure Y_{di}, we construct a robust empirical (beta) model by applying ordinary least squares regression to the log-transformed per equivalent adult expenditure l_nY_{di}. The OLS regression model takes the form:

$$l_nY_{di} = X_{di}\beta + u_{di}$$

Where $Y_{di} = E_{di} + c$ for $c > 0$, c is constant and E_{di} is the welfare variable (per equivalent adult expenditure) for i^{th} individual household in a cluster d and N_d is the size of the population in area d where D represents the partitioned population of size N, $i = 1, \ldots, N_d$ and $d = 1, \ldots, D$. X_{di} is the vector of explanatory variables for i^{th} household in cluster d, encompassing household and individual characteristics. β represents the vector of coefficients to be estimated. u_{di} is the error term, capturing unobserved factors influencing consumption. The error term, u_{di}, can be decomposed into two independent components, a cluster-specific effect (η_d) and a household-specific effect (e_{di}).

$$u_{di} = \eta_d + e_{di} \qquad (2)$$

The cluster effect (η_d) captures unobserved factors of households within a cluster, while the household-specific effect (e_{di}) captures idiosyncratic variations in consumption patterns at the household level. The location effect (η_d) is defined as the weighted average of the individual household errors within a specific cluster. Additionally, e_{di} is assumed to be heteroskedastic.

$$e_{di} \sim ind(0, \sigma_e^2 k_{di})$$

Following the ELL methodology, the estimation of the unconditional variance for each location is obtained by estimating equation and then using the residuals ($\hat{u}_{di}$). By defining $\hat{u}_{d.}$ as the weighted average of $\hat{u}_{di}$ for a specific cluster d, the household-specific error term ($\hat{e}_{di}$) can be derived:

$$\hat{u}_{di} = \hat{u}_{d.} + (\hat{u}_{di} - \hat{u}_{d.})$$

$$\hat{u}_{di} = \hat{\eta}_d + \hat{e}_{di}$$

The unconditional variance of the location effect ($\hat{\sigma}_\eta^2$) is estimated using the following formula:

$$\hat{\sigma}_\eta^2 = \max\left(\frac{\left(\Sigma_d \omega_d (u_{d.} - u_{..})^2 - \Sigma_d \omega_d (1 - \omega_d)\,\hat{\tau}_d^2\right)}{\Sigma_d \omega_d (1 - \omega_d)}\ ;\ 0\right)$$

Where ω_d is the weight of the cluster d, $u_{d.}$ is the weighted mean of the residuals ($\hat{u}_{di}$) within cluster d, and $u_{..}$ is the overall weighted mean of the residuals across all clusters. Finally,

$\hat{\tau}_d^2 = \frac{\Sigma_i(e_{di} - e_{d.})^2}{n_d(n_d - 1)}$ quantifies the variability of household expenditures within each cluster. A parametric form for $\sigma_{e_{di}}^2$, utilizing a logistic function to ensure variance remains non-negative.

$$\sigma_{e_{di}}^2 = \left[\frac{A\exp^{z'_{di}\alpha} + B}{1 + \exp^{z'_{di}\alpha}}\right] \qquad (4)$$

Where A is an upper bound for the variance. B is a lower bound for the variance and it ensures the variance does not become negative. z'_{di} is a vector of household characteristics that influence the variance of the error term. α is a vector of parameters that is estimated from the data. These parameters control how the variance changes depending on the household characteristics in z'_{di}.

By setting B = 0 and A = 1.05, a simplified version of this logistic function is given and is estimated using OLS regression with the log-transformed squared residuals as the dependent variable and a set of explanatory variables (z'_{di}) as the independent variables[2].

$$\ln\left[\frac{e_{di}^2}{A - e_{di}^2}\right] = z'_{di}\alpha + r_{di}$$

By defining $\exp(z'_{di}\alpha)$ as D and employing the delta method[3], the estimated variance of the idiosyncratic error for household i in cluster d is given by

$$\hat{\sigma}_{e_{di}}^2 \approx \left[\frac{AD}{1+D}\right] + \frac{1}{2}\,\mathrm{Var}(r)\left[\frac{AD(1-D)}{(1+D)^3}\right] \qquad (6)$$

Where $\mathrm{Var}(r)$ is the estimated variance from the model's residuals in Equation 5.

To quantify the uncertainty associated with $\hat{\sigma}_\eta^2$, (Elbers, Lanjouw, & Lanjouw, 2002) proposes two methods. One is a simulation-based approach[4]. Under the second method, the sampling variance is calculated using the approximation technique.

$$\mathrm{Var}(\hat{\sigma}_\eta^2) = \sum_d 2\left\{a_d^2\left[(\hat{\sigma}_\eta^2)^2 + (\hat{\tau}_d^2)^2 + 2\hat{\sigma}_\eta^2\hat{\tau}_d^2\right] + b_d^2\frac{(\hat{\tau}_d^2)^2}{n_d - 1}\right\}$$

Where $a_d = \frac{\omega_d}{\sum_d \omega_d(1-\omega_d)}$ and $b_d = \frac{\omega_d(1-\omega_d)}{\sum_d \omega_d(1-\omega_d)}$.

Having estimated the variances associated with both the location and household effects, the next step is to get the GLS estimator. The purpose of the GLS estimator is to leverage the estimated variance components to construct a variance-covariance matrix (Ω) accounting for both heteroskedasticity and cluster-level correlation. The off-diagonal elements of Ω represent

[2] This modeling approach is also known as the Alpha Model.
[3] Delta method is briefly discussed in Elber Lanjouv Lanjouv (2002).
[4] In our study, we are not utilizing this simulation-based approach to quantify the uncertainty associated with $\hat{\sigma}_\eta^2$.

the shared variance component due to location effect $\hat{\sigma}_{\eta}^{2}$, and diagonal elements represent total variance which is a combination of location and household effect variances ($\hat{\sigma}_{\eta}^{2} + \hat{\sigma}_{e_{di}}^{2}$). The variance-covariance matrix of the random effects for the complete dataset is represented as $\widehat{\Omega}$ which is structured as a block diagonal matrix. Each diagonal block associated with an individual cluster or small area allows for variations in correlation structures across different areas. The off-diagonal blocks equal to zeros highlight the assumption of independence among distinct clusters. By utilizing the ELL (2003) framework, the GLS estimates are obtained as follows:

$$\hat{\beta}_{GLS} = (X'W\Omega^{-1}X)^{-1}X'W\Omega^{-1}Y$$

Similarly, the variance estimates are obtained:

$$Var(\hat{\beta}_{GLS}) = (X'W\Omega^{-1}X)^{-1}(X'W\Omega^{-1}WX)(X'W\Omega^{-1}X)^{-1} \quad (9)$$

where W represents the diagonal matrix of sampling weights.

Building on the framework proposed by ELL (2003), Monte Carlo simulations are utilized to estimate expected welfare measures based on the initial model. This methodology applies parameter and error estimates derived from the HIES 2018-19 to the PSLM 2019-20 data. The objective is to conduct a substantial number of simulations to ensure that the welfare estimates are both robust and reliable. In Model 1, the general parameter $\delta_d = \delta_d(y_d)$ serves as the ELL estimator, derived through a bootstrap methodology. This approach yields a numerical approximation of the theoretical ELL estimator, represented as the marginal expectation $\delta_d^{ELL} = E(\delta_d)$. Additionally, the bootstrap procedure is similarly employed to estimate the mean squared error (MSE) of the ELL estimator. The following steps are involved in the bootstrap procedure:

1. By utilizing the residuals of the fitted model in , random effects η_d^*[5] and errors e_{di}^*[6] are generated.
2. Then by using the values of auxiliary variables along with the estimator $\hat{\beta}$ of the regression parameter β, bootstrap values of the welfare variable (poverty) are generated for all population units, as shown below through the generation process.

$$l_n(y_{di}^*) = x_{di}\hat{\beta} + \eta_d^* + e_{di}^*$$

3. Then we get the census of the response variable by utilizing the above model, which is used to estimate the indicator of interest based on the welfare of individual households. The process of generation is replicated for $m = 1, \ldots, M$ times to get the M full censuses. After this for each single census m, the indicator of interest $\delta_d^{*(m)} =$

[5] Random effects are generated for each area $d = 1, \ldots, D$.
[6] Errors are for each household/unit $i = 1, \ldots, N_d$.

$\delta_d Y_d^{*(m)}$, is calculated. Where $Y_d^{*(m)} = Y_{d1}^{*(m)}, \ldots, Y_{dN_d}^{*(m)}$ are the values of the Welfare variable variable in the area d that is obtained via the bootstrap census.

4. To obtain the ELL estimator, we simply average over the M censuses.

$$\hat{\delta}_d^{ELL} = \frac{1}{M}\sum_{m=1}^{M} \delta_d^{*(m)}$$

5. To estimate MSE under ELL

$$\text{mse}_{ELL}\left(\hat{\delta}_d^{ELL}\right) = \frac{1}{M}\sum_{m=1}^{M}\left(\delta_d^{*(m)} - \hat{\delta}_d^{ELL}\right)^2$$

Data Description and Application

The data description starts with the data preparation stage which is concerned with identifying the common variables that exist between the HIES 2018-19 and the PSLM 2019-20. These variables serve as the bridge for predicting consumption levels within the PSLM dataset. To generate poverty data, these linking variables are defined across both datasets. The HIES includes a consumption module that facilitates the estimation of poverty while providing comprehensive information on various well-being indicators. However, the HIES sample is representative only at the provincial and national levels, not at the district level. In contrast, while the PSLM lacks a consumption module, it has a large sample that is representative at the district level. The data preparation phase involves:

1. Welfare Variables Creation: involves creating a variable of interest (poverty) by utilizing HIES 2018-19. To measure poverty, "monthly per-adult equivalent consumption expenditure" at the household level is utilized. The Cost of Basic Needs (CBN) method is used to measure poverty. The poverty line, quantified in monetary terms as (monthly) per equivalent adult expenditure, acts as a crucial threshold[7]. Households with spatially adjusted monthly per-adult equivalent consumption expenditures below this threshold are categorized as poor.
2. Definition Matching: involves creating common variables in HIES 2018-19 and PSLM 2019-20 with nearly the same questions, definitions of variables and categories, etc. The common variables between HIES 2018-19 and PSLM 2019-20 are known as auxiliary variables. These auxiliary variables are based on household head characteristics, household characteristics, dwelling characteristics, and household asset possession.

[7] Based on HIES 2018-19, the poverty line is set at 3776 by the Planning Commission of Pakistan.

3. Statistical Matching: involves the distribution of the common variables being the same between HIES 2018-19 and PSLM 2019-20. Any variable having missing values for more than 1% of the observations is discarded to maintain data quality. If the ratios of the weighted means and the standard deviations of auxiliary variables between HIES 2018-19 and PSLM 2019-20 fall within the range of 0.95 to 1.05, the variable is deemed suitable for inclusion in the regression model. After, removing the irrelevant variables, a final list of 42 variables is considered for the model selection procedure out of 130 total variables.
4. Location Matching: entails the aligning of PSU variables at the cluster level between HIES 2018-19 and PSLM 2019-20. The first digit represents province, the second digit represents division (if no division then it is equal to zero), the third digit represents district (if no district then it is equal to zero), fourth digit represents region (rural or urban), last three-digit represents the code of PSU.

After data preparation, the next step is the modeling of consumption utilizing the HIES 2018-19 data. The process starts with estimating the Beta Model using OLS. Before the estimation of the Beta model, there are some perquisites:

1. Adjust the Beta model for multicollinearity by using the Variance Inflation Factor (VIF). Those variables excluded from the Beta model have VIF values surpassing 7.
2. The Beta model is adjusted for stepwise backward and forward induction using p-values. The general threshold of 0.05 is used for statistical significance to select the best model. Similarly, model selection via Lasso (Least Absolute Shrinkage and Selection Operator) is utilized. Having established a preliminary set of predictors through stepwise and lasso regression, an additional stepwise selection procedure is utilized to further refine the model based on both the adjusted R-squared and the Bayesian Information Criterion (BIC).

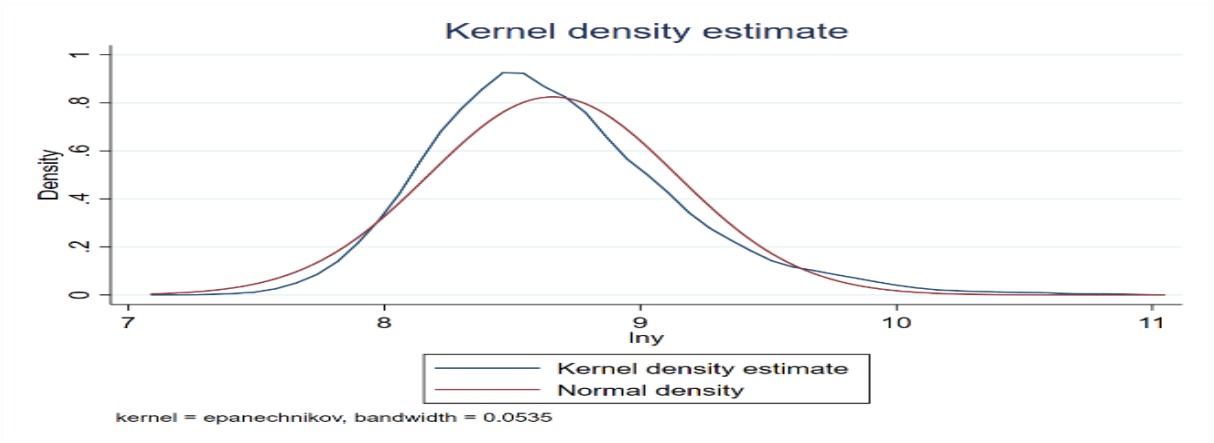
After finalizing the set of predictors for the Beta model, an Ordinary Least Squares regression is estimated to obtain the residuals. These residuals are then modeled with regressors including the auxiliary variable that is not used in the Beta model, and interaction of these auxiliary variables with the residuals and their squared counterparts. The resulting Alpha model is then scrutinized for multicollinearity using VIF. To enhance the predictive accuracy of the Alpha model, a stepwise selection procedure is implemented. The whole procedure is designed to calculate the unconditional variance[8]. Then Generalized Least Squares (GLS) estimation is employed[9], which explicitly accommodates the varying error structures within the model. This method produces more efficient estimates than OLS, offering point estimates for the regression coefficients while also considering the underlying distributions of both the coefficients and the errors. In the current analysis, a parametric approach to Monte Carlo

[8] The procedure of Alpha model for the calculation of unconditional variance is also explained in (Ngyen et al., 2018).
[9] GLS estimation is described in Research methodology part.

simulation is employed within the ELL framework, assuming normality for the distributions of coefficients and error terms. This method utilizes theoretically derived distributions to introduce variability. Although computationally efficient, its effectiveness relies on the validity of these distributional assumptions.

One of the final calls before the application of ELL methodology is to check certain diagnostics essential for detecting outliers and effectively predicting poverty estimates. While progressing toward diagnostics, the introductory step is to transform the dependent variable to some appropriate scale for analysis. The analysis in this study used the natural logarithm transmutation based on the positively skewed nature of the dependent variable. The figure appended below demonstrates the shape of the log-transformed variable against the normal transformation.

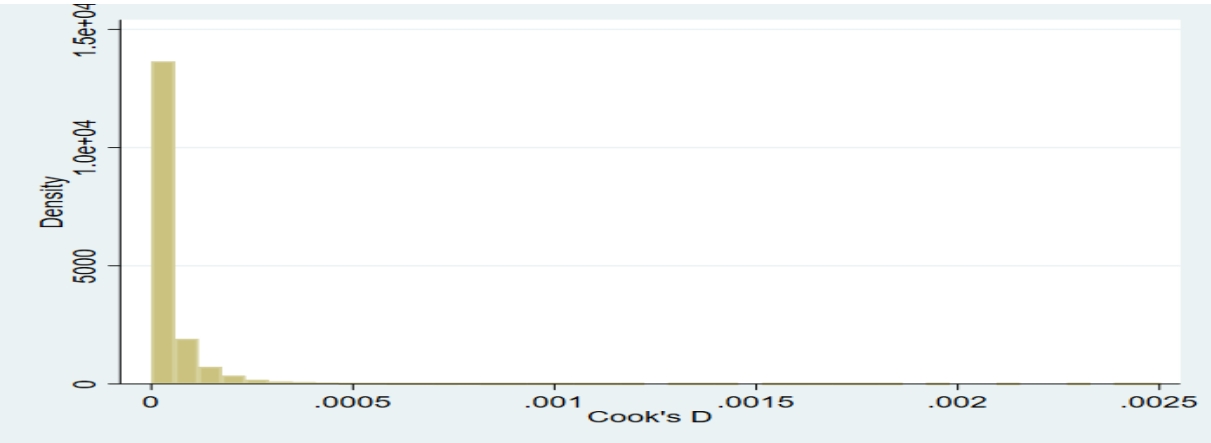
Fig 01: Kernel Density Plot based on logarithm of per equivalent adult expenditure



The kernel density plot (KDE) demonstrates a distribution that approximates a normal curve. As transformation is utilized to induce normality. However, attaining a perfect fit is rarely met in practice (Marhuenda, Molina, Morales, & Rao, 2017). Figure 01 illustrates that the data exhibiting a near-normal distribution, allows us to move further.

The initial step in model diagnostics involves identifying the influential observations and outliers that can distort model fit and stability. For this purpose, Cook's distance plot, studentized residual, and leverage plot are observed. Cook's distance a metric that relates leverage and residual information, is a valuable tool for this purpose. As detailed by Cook (1977) and (Molina, Rao, & Datta, 2015), Cook's distance measures the impact of excluding a specific observation on the model's parameter estimates. Observations with an absolute Cook's distance exceeding 4/N, where N is the total number of observations, are assumed potentially influential observations.

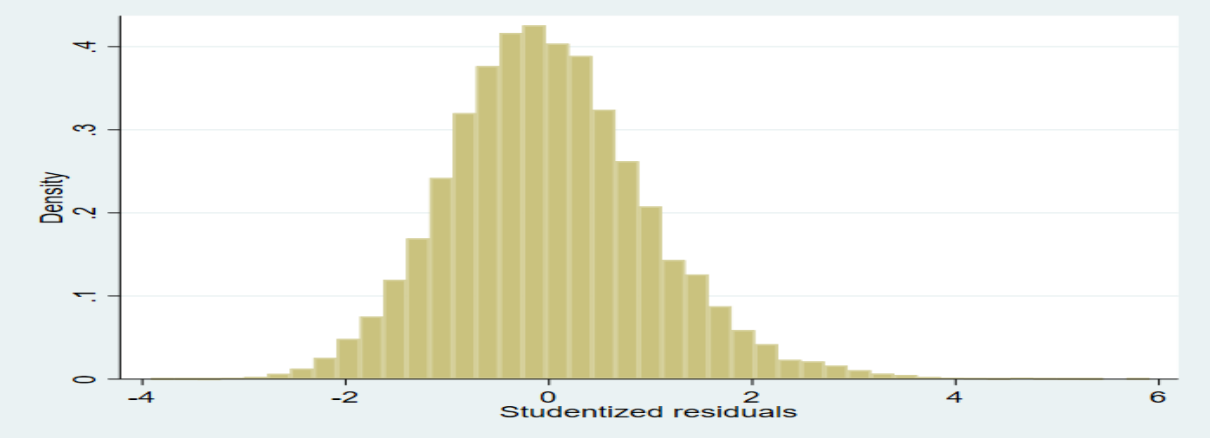
Fig 02 : Cook's Distance Plot



Cook's distance plot reveals that most observations have a trivial influence on the model's fit. However, a few observations with elevated Cook's distance values may warrant further investigation to assess their impact on model parameter estimates.

Studentized residuals help to identify observations that deviate significantly from the model's predictions. Large absolute values of studentized residuals indicate potential outliers or influential points. Observations with studentized residuals exceeding a threshold of 2 or 3 in absolute terms may warrant further investigation.

Fig 02: Studentized Residuals

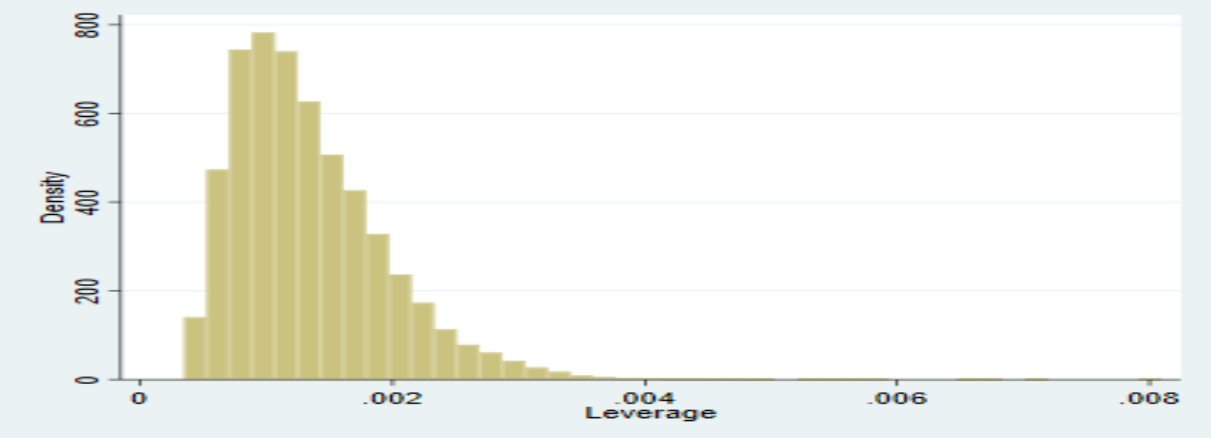


The studentized residual plot highlights an approximately symmetric distribution, centered around zero, suggesting that the model's assumptions of homoscedasticity and normality of residuals are reasonably met. However, a few outliers visible in the tails of the distribution may warrant further investigation.

Leverage is a measure of the extent to which an independent variable varies from its mean. An observation with an unusual value on a predictor variable is considered a high-leverage point. These points can influence the calculated regression coefficients, leading to biased estimates.

Fig 03: Leverage

Remittances Review
August 2024,
Volume: 9, No: 4, pp.2703-2721
ISSN: 2059-6588(Print) | ISSN 2059-6596(Online)



The leverage plot displays a right-skewed distribution, with the majority of the observations having low leverage. However, a few observations with high leverage values may influence the model's fit, leading to biased parameter estimates that require further investigation.

To verify the robustness of our model, we conducted diagnostics and set aside observations with high leverage, which are data points with extreme values on predictor variables that can disproportionately influence the model's parameter estimates. Observations with Cook's distance exceeding $\frac{4}{N}$, absolute studentized residuals exceeding 2, or leverage values exceeding $\frac{2k+2}{N}$ (where N is the total number of observations and K is the number of predictors) were excluded to improve the model's robustness and reliability.

Results and Discussion

After addressing influential observations and outliers via diagnostic checks, the regression model is re-estimated. The resulting parameter estimates such as Beta, Alpha, and GLS model estimates along with simulation-based estimates for the variable of interest (poverty) are presented in the following section.

Table 01: Model Selection - Beta Model Estimates

Dependent Variable: Per-adult equivalent Consumption Expenditure in logarithm					
Variable	**Definition/Coding**	**Coefficient**	**Standard Error**	**95% Confidence Interval**	
cleanwater	1=improved drinking water source; 0=otherwise	0.0986***	0.0137	0.0718	0.1253
cooking	1=improved cooking source; 0=otherwise	0.0439***	0.0063	0.0315	0.0562
dryer	1 = household owns item; 0 otherwise	0.0720***	0.0081	0.0561	0.0879
edu	Educational attainment	0.0654***	0.0026	0.0602	0.0705
edu_1	1=No Schooling; 0=otherwise	0.0571***	0.0071	0.0432	0.0710
edu_4	1= Lower secondary (grade 9-10); 0=otherwise	-0.0163**	0.0073	-0.0305	-0.0020
fan	1 = household owns item; 0 otherwise	0.0337***	0.0086	0.0169	0.0505
floor	1=pacca floor; 0=otherwise	0.0617***	0.0064	0.0491	0.0743
geaser	1 = household owns item; 0 otherwise	0.1980***	0.0111	0.1763	0.2198
highestedu	Highest level of educational attainment in the HH	0.0017	0.0022	-0.0026	0.0060
highestedu_4	1= Lower secondary (grade 9-10), 0=otherwise	-0.0167***	0.0061	-0.0286	-0.0048
highestedu_6	Undergraduate	0.0447***	0.0087	0.0278	0.0617
internet	1=HH has an internet connection; 0=otherwise	0.0909***	0.0057	0.0798	0.1020
iron	1 = household owns item; 0 otherwise	0.0366***	0.0067	0.0235	0.0498
language_new_4	1=pashto; 0=otherwise	-0.0982***	0.0118	-0.1213	-0.0752
lnage	Age in natural logarithm	-0.0231***	0.0079	-0.0385	-0.0076
marital_2	1=married; 0=otherwise	-0.0713***	0.0081	-0.0871	-0.0554
microwave	1 = household owns item; 0 otherwise	0.2388***	0.0110	0.2174	0.2603
ownership_2	1=rented house; 0=otherwise	-0.0605***	0.0078	-0.0758	-0.0452
prov_2	Punjab	-0.0172	0.0114	-0.0396	0.0051
prov_3	Sindh	-0.0445***	0.0122	-0.0684	-0.0207
prov_4	Balochistan	-0.0710***	0.0132	-0.0968	-0.0452
roof	1=pacca roof; 0=otherwise	0.0453***	0.0065	0.0325	0.0580
Sexratio	No. of men (15-65 yrs old)/HH size	-0.0453***	0.0135	-0.0718	-0.0189

Source: Author's calculation

Table 01 depicts the model selection via OLS indicating that most included predictors are statistically significant at the 5% significance level. While the prov_2 variable did not reach statistical significance, it is retained in the model as a control variable.

Table 02: Model Selection Alpha Model Estimates

Dependent Variable: Residual			
Variable	**ELL Method**		
	Definition/Coding	**Coefficient**	**p-value**
employ_1	1=employed; 0=otherwise	-0.1801	0.0000
room	Number of Rooms in Dwelling	0.0477	0.0000
toilet_detail_2	1=flush connected to public sewerage, 0=otherwise	-0.0763	0.0690
toilet_detail_7	1=dry pit latrine, 0=otherwise	-0.2027	0.0050
_cons	Constant	-3.8204	0.0000

Source: Author's calculation

As shown in Table 02, the included predictors are statistically significant, indicating a strong association with the outcome variable. At this stage of analysis, the requirement is that the regressors used in the model should be significant.

Table 03: Model Selection - GLS Model estimates

Dependent Variable: Per-adult equivalent Consumption Expenditure in logarithm		
Variable	**ELL**	
	Coefficient	**p-value**
cleanwater	0.0788***	0.0000
cooking	0.0545***	0.0000
dryer	0.0735***	0.0000
edu	0.0623***	0.0000
edu_1	0.0526***	0.0000
edu_4	-0.0170**	0.0170
fan	0.0291***	0.0010
floor	0.0695***	0.0000
geaser	0.1866***	0.0000
highestedu	0.0017	0.4340
highestedu_4	-0.0128***	0.0330
highestedu_6	0.0403***	0.0000
internet	0.0949***	0.0000
iron	0.0353***	0.0000
language_new_4	-0.0713***	0.0000
lnage	-0.0258***	0.0010
marital_2	-0.0640***	0.0000
microwave	0.2238***	0.0000
ownership_2	-0.0705***	0.0000
prov_2	0.0049	0.7470
prov_3	-0.0230	0.1560
prov_4	-0.0587***	0.0010
roof	0.0486***	0.0000
sexratio	-0.0586***	0.0000
table	0.0800***	0.0000
toilet_1	0.0481***	0.0000
ups	0.1737***	0.0000
urban	0.0189**	0.0250
wall	0.0479***	0.0000
washingmachine	0.0462***	0.0000
water	-0.0159***	0.0000
water_2	-0.0367***	0.0000
water_5	0.1465***	0.0000
workratio_adult	0.3619***	0.0000
_cons	8.0786***	0.0000

Source: Author's calculation

Table 03 presents the estimated coefficients for both model specifications. All coefficients, except for the highest_edu variable and the prov_3, are statistically significant.

Table 04: Comparison of OLS vs GLS Estimates

Dependent Variable: Per-adult equivalent Consumption Expenditure in logarithm		
Variable	**ELL Model Coefficients**	
	OLS Estimates	**ELL GLS Estimates**
cleanwater	0.0986	0.0788
cooking	0.0439	0.0545
dryer	0.0720	0.0735
edu	0.0654	0.0623
edu_1	0.0571	0.0526
edu_4	-0.0163	-0.0170
fan	0.0337	0.0291
floor	0.0617	0.0695
geaser	0.1980	0.1866
highestedu_4	-0.0167	-0.0128
highestedu_6	0.0447	0.0403
internet	0.0909	0.0949
iron	0.0366	0.0353
language_n~4	-0.0982	-0.0713
lnage	-0.0231	-0.0258
marital_2	-0.0713	-0.0640
microwave	0.2388	0.2238
ownership_2	-0.0605	-0.0705
prov_2	-0.0172	0.0049
prov_3	-0.0445	-0.0230
prov_4	-0.0710	-0.0587
roof	0.0453	0.0486
sexratio	-0.0453	-0.0586
table	0.0897	0.0800
toilet_1	0.0457	0.0481
ups	0.1846	0.1737
urban	0.0225	0.0189
wall	0.0612	0.0479
washingmac~e	0.0344	0.0462
water	-0.0182	-0.0159
water_2	-0.0285	-0.0367
water_5	0.1606	0.1465
workratio_~t	0.3557	0.3619
_cons	8.0644	8.0786

Source: Author's calculation

A comparison of the coefficient estimates from OLS and GLS regressions, as presented in Table 04, reveals minimal differences across most coefficients related to auxiliary variables. However, the province-specific variables exhibit more substantial discrepancies between the two models which is included as a control factor in the model.

Table 05: Summary Statistics

Model setting	
Error decomposition	ELL
Beta drawing	Parametric
Eta drawing method	normal
Epsilon drawing method	normal
Empirical best method	No
Beta-model diagnostics	
Number of observations	22677
Adjusted R-squared	0.57
R-squared	0.5706
Root MSE	0.2802
F-stat	884.974
Alpha-model diagnostics	
Number of observations	22677
Adjusted R-squared	0.0026
R-squared	0.0028
Root MSE	2.2726
F-stat	15.8203
Model Parameters	
Sigma ETA sq.	0.0098
Ratio of sigma eta sq over MSE	0.1251
Variance of epsilon	0.0687
Sampling variance of Sigma eta sq.	0.0000003

Source: Author's calculation

As depicted in Table 05, this study employs the ELL framework for error decomposition, utilizing a parametric method to estimate model parameters. The analysis assumes that both area-specific and household-specific errors adhere to a normal distribution. The model selection is done via the Beta model. The adjusted R-squared value of 0.57 indicates a reasonable fit for the model. Additionally, the F-statistic of 885 reinforces the overall statistical significance of the model. For estimating unconditional variance and model parameters, the analysis incorporates the alpha model and employs a GLS approach. It is noted that the alpha model typically exhibits a low adjusted R-squared value[10]. Concerning model parameters, a critical guideline is that the ratio of the variance of location effects (sigma eta squared) to the mean squared error (MSE) should not be very high. This ratio of 0.1251 is reasonable and it illustrates the extent of the residual variance that can be attributed to location effects.

[10] The Alpha model is characterized by typically low R-squared. In most applications, the adjusted R-squared of the alpha model is rarely exceeds 0.05 (Corral, Molina, Cojocaru, & Segovia, 2022).

The findings from the simulations are presented systematically, facilitating a detailed comparison between direct and model-based estimates.

Table 06: Comparison of Poverty Estimates at the Provincial and National Level

Provincial / National	Poverty Model-Based Estimates	Poverty Estimates-HIES 18-19	Absolute Difference	Squared Difference
KPK	0.25	0.28	0.0277053	0.0007676
Punjab	0.15	0.16	0.0107123	0.0001148
Sindh	0.24	0.24	0.0017839	0.00000318
Balochistan	0.37	0.42	0.0448443	0.002011
Pakistan	0.20	0.21	0.0127932	0.0001637

Note: Direct Estimates of poverty are the Official estimates reported by the Pakistan Planning Commission based on HIES 2018-19.

Table 06 demonstrates that the ELL model-based district-level estimates, when aggregated nationally (0.20), are closely aligned with the national-level poverty estimates (0.21) derived from the HIES 2018-19 data, thereby validating the benchmarking criterion. The estimates for Punjab and Sindh show a particularly strong alignment. In contrast, we observe differences in Balochistan, followed by Khyber Pakhtunkhwa (KPK), between the survey and model-based estimates. Additionally, the sample size for Balochistan is relatively small, resulting in higher Mean Squared Errors (MSEs).

To further investigate the Benchmarking properties at the district level, we compared the (rural) district-level estimates calculated from HIES 2018-19 with the ELL model-based poverty estimates. The direct estimates of poverty are calculated via (monthly) per equivalent adult expenditure with weighting factors. Furthermore, we established a predetermined threshold to refine our estimates[11]. For variance estimation, both the weighting scheme and the inherent clustering within the dataset are considered to generate a domain-specific variance to account for the variability present in our sample. The direct estimates were computed for 93 rural districts based on the HIES 2018-19 data due to the survey's structural characteristics[12]. These direct estimates are compared with model-based estimates generated under the ELL model framework.

[11] The poverty line of 3776 as officially reported by the Pakistan Planning Commission serves as a threshold.
[12] HIES 2018-19 collects data at the district level for rural strata except for Balochistan province. Similarly, it collects data at the divisional level for urban strata including Balochistan where both rural and urban data are attributable to divisions.

Fig 04: MSE of Poverty Estimates at District level

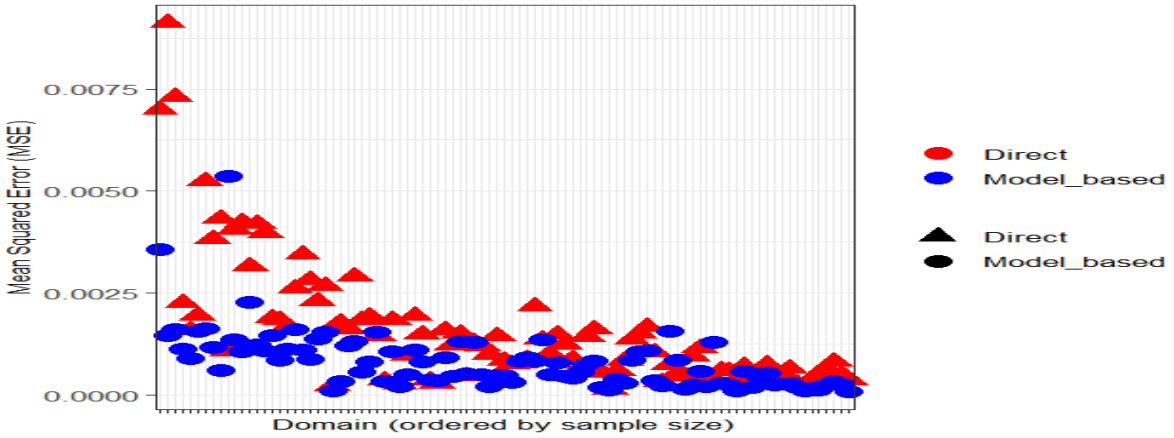
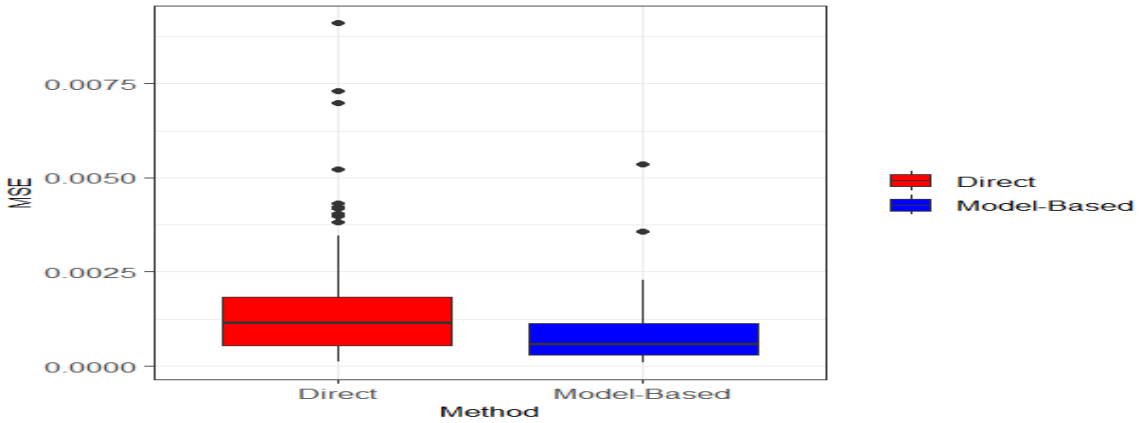


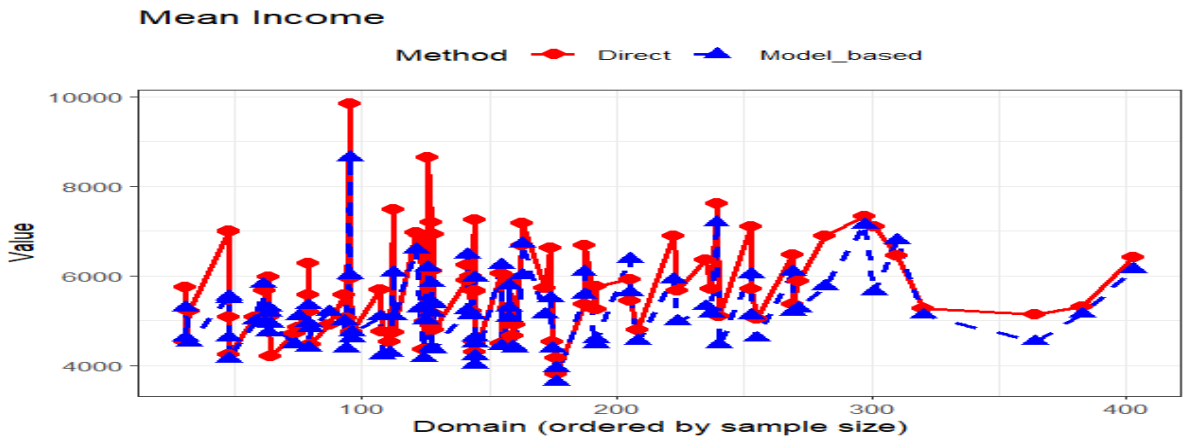
Figure 04 illustrates the results of the Mean Squared Error (MSE) of poverty estimates at the district level in Pakistan. In instances of limited sample sizes, the MSE for direct estimates is higher than model-based estimates. The MSE values indicate that model-based estimates are more reliable, particularly when the sample size is small.

Fig 05: Box Plot of MSE related to Poverty Estimates at the District level



In the box plot depicted in Figure 05, the ELL model-based method exhibits superior performance. The model-based estimates present a significantly lower median MSE and the interquartile range of MSE values for the model-based technique is notably narrower, indicating greater precision and consistency in its estimates. In contrast, the direct estimation method shows a wider spread of MSE values, followed by a higher median and extended whiskers in the box plot, suggesting potentially less reliable direct estimates.

Fig 06: Per equivalent Adult Expenditure at the District level



A review of the mean income distribution across districts reveals the direct estimation approach results in a significantly wider dispersion of mean income specifically, in areas with limited sample sizes. Conversely, the model-based estimation technique shows a more compact distribution of mean income estimates particularly, in those areas with small sample sizes. Our comprehensive analysis strongly supports the model-based method for poverty estimation, demonstrating its superior robustness and reliability compared to alternative techniques.

Fig 07: Poverty Mapping at the District level in Pakistan – ELL Approach

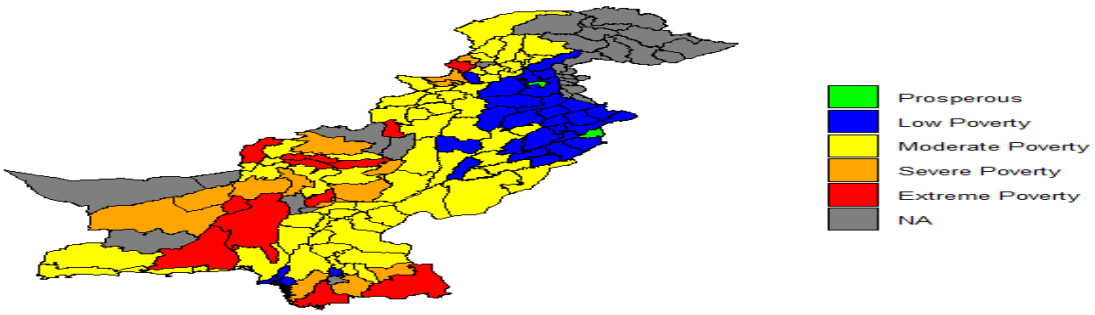


Figure 07 highlights poverty at the district level in Pakistan[13]. The results indicate that 12 districts are facing extreme poverty: Sheerani, Awaran, Tharparkar, Khuzdar, Shaheed Sikandarabad, Ziarat, Killa Abdullah, Harnai, Mohmand, Sujawal, Duki, and Nasirabad, marking them as the most economically disadvantaged areas in the country that require urgent targeted interventions. Additionally, the study identified 16 districts suffering from severe

[13] Poverty Mapping is done by considering the following thresholds: If less than 5th percentile=Prosperous, 5th to less than 25th percentile=Low Poverty, 25th to less than 75th percentile= Moderate Poverty, 75th to 90th percentile= Severe Poverty and over 90th percentile= Extreme Poverty.

poverty: Kalat, Badin, Bajur, Tando Allah Yar, Barkhan, Killa Saifullah, Kachhi, Kharan, Dera Bugti, Washuk, Thatta, Jaffarabad, Umerkot, Khyber, Orakzai and Sohbatpur. Although these areas do not fall into the extreme category, they still experience significant economic challenges and warrant focused attention in poverty alleviation efforts. The analysis also uncovered districts classified as experiencing Moderate Poverty. However, they are very close to the Severe Poverty threshold: Shaheed Benazirabad, Mirpur Khas, South Waziristan, Nuski, Rajanpur, Sanghar, Khairpur, and Shikarpur. This detailed poverty mapping offers crucial insights for policymakers, enabling targeted interventions and customized strategies to address the unique economic challenges of each district effectively.

Conclusion:

This study utilized the (Elbers et al., 2003) method to generate district-level poverty data[14] for Pakistan, using HIES 2018-19 and PSLM 2019-20 surveys. To generate robust poverty estimates along with MSE, the Monte Carlo simulation approach is employed, conducting 200 iterations for each household represented in the PSLM 2019-20 dataset. A comparison of Mean Squared Errors (MSE) between direct and model-based estimates reveals that the ELL model-based estimates are more reliable than direct estimates, particularly for districts with limited sample sizes. The poverty mapping based on the ELL model-based estimates revealed 12 districts experiencing extreme poverty and 16 facing severe poverty, requiring immediate intervention. Several districts under Moderate Poverty are identified as being close to the Severe Poverty threshold, highlighting the need for proactive measures. This granular mapping provides invaluable insights for policymakers, enabling precise identification of priority areas and tailored poverty reduction strategies. The study acknowledges the potential for further validation using the extended ELL method (Nguyen, Corral Rodas, Azevedo, & Zhao, 2018) and suggests exploring alternative approaches like the Empirical Best/Bayes Predictor method (Molina & Rao, 2010) in future research.

Acknowledgement:

The author wishes to express sincere gratitude to Mr. Paul Corral, Senior Economist at the World Bank for sharing relevant STATA files/materials that significantly enhanced the quality of our analysis. The author is deeply appreciative of Mr. Corral's professional courtesy and collaborative spirit, which exemplify the best practices in academic and institutional cooperation.

References:

Begum, S. (2015). The livelihood and poverty mapping analysis at regional level in Pakistan: [Netherlands]:[publisher not identified].

Corral, P., Molina, I., Cojocaru, A., & Segovia, S. (2022). Guidelines to small area estimation for poverty mapping: World Bank Washington.

[14] For generation of poverty estimates, Monte carlo simulations 200 times for each household in the PSLM 2019-20 is conducted.

Elbers, C., Lanjouw, J. O., & Lanjouw, P. (2002). Micro-level estimation of welfare (Vol. 2911): World Bank Publications.

Elbers, C., Lanjouw, J. O., & Lanjouw, P. (2003). Micro-level estimation of poverty and inequality. Econometrica, 71(1), 355-364.

Fay III, R. E., & Herriot, R. A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. Journal of the American Statistical Association, 74(366a), 269-277.

Jamal, H. (2007). Income poverty at district level: An application of small area estimation technique.

Marhuenda, Y., Molina, I., Morales, D., & Rao, J. (2017). Poverty mapping in small areas under a twofold nested error regression model. Journal of the Royal Statistical Society Series A: Statistics in Society, 180(4), 1111-1136.

Molina, I., Rao, J., & Datta, G. S. (2015). Small area estimation under a Fay-Herriot model with preliminary testing for the presence of random area effects. Survey Methodology, 41(1), 1-20.

molimolina, I., & Rao, J. N. (2010). Small area estimation of poverty indicators. Canadian Journal of statistics, 38(3), 369-385.

Nguyen, M., Corral Rodas, P. A., Azevedo, J. P., & Zhao, Q. (2018). sae: A stata package for unit level small area estimation. World Bank Policy Research Working Paper(8630).

Torabi, M., & Rao, J. (2014). On small area estimation under a sub-area level model. Journal times