Advance Journal of Econometrics and Finance
Vol-4, Issue-1, 2026

Advance Journal of Econometrics and Finance

Online ISSN
2959-8990
Print ISSN
2959-8982
https://ajeaf.com/index.php/Journal/About

Name of Publisher: SCHOLAR CRAFT EDUCATION & RESEARCH HUB

Review Type: Double Blind Peer Review

Journal Frequency: Quarterly Research Journal

POVERTY MAPPING AT SMALL AREA LEVEL IN PAKISTAN: A COMPARATIVE ANALYSIS OF ELL AND EXTENDED ELL METHODS

Received: 19 December 2025, Accepted: 27 February 2026

Dr. Shahid Shabir

Dr. Shahid Shabir

PhD Econometrics
Pakistan Institute of Development Economics (PIDE)
Islamabad, Pakistan
sshahid.shabir@gmail.com

Abstract

This seminal research conducts a rigorous comparative analysis for the first time to determine the optimal methodology among the ELL (Elbers et al. 2003) and extended ELL (Nguyen et al. 2018) approaches to Small Area Estimation (SAE), ultimately identifying the most deprived districts in Pakistan. The statistically robust unit-level ELL and Extended ELL Methodologies are used under the Small Area Estimation (SAE) Technique to estimate poverty headcount based on the proportion of deprived households, utilizing the 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) of Poverty estimates via the ELL and Extended ELL methods reveals that the ELL method outperforms the Extended ELL method. Apart from the methodological superiority of the traditional ELL method, both methodologies unveil the presence of extreme poverty in 13 districts, 19 districts experiencing severe poverty, and finally 04 districts with moderate poverty but on the verge of severe poverty. This granular mapping of poverty intensity across districts provides invaluable insights for the Government and policymakers regarding the priority areas for intervention. It allows for the tailoring of poverty reduction strategies to the specific needs and circumstances of each deprived district in Pakistan.

Keywords: Poverty, Small Area Estimation, ELL Method, Extended ELL Method, Monte-Carlo Simulations, Bootstrapping.

Advance Journal of Econometrics and Finance
Vol-4, Issue-1, 2026

1. Introduction

Sustainable Development Goal (SDG-1) aims to eradicate poverty in all its aspects, acknowledging that poverty is not merely a matter of economic hardship but also encompasses aspects of social exclusion, vulnerability, and lack of access to decision-making processes. One of the United Nations' Sustainable Development Goals (SDGs) is to eradicate poverty, safeguard the Planet, and ensure prosperity for all by 2030. Conversely, in the case of developing countries, data on welfare variables like poverty are not readily available at the small administrative level, or, if available, the sample size is so small that it is not representative. To overcome this problem, it is vital to generate data by utilizing some econometric modeling techniques or increasing the sample size. Increasing sample size is very costly, and sometimes not a feasible option. Due to the non-availability of the requisite data, the policymakers, researchers, and Government agencies are hampered from apportioning the development budget and resources to specific districts depending on the poverty status of those areas. This predicament shifts the focus of researchers to model-based Small Area Estimation (SAE) techniques to address the challenge of limited data granularity. This SAE-based approach works in a way that we have some variable data that is nationally representative. However, the same data is not available at some disaggregated levels. Conversely, we have some administrative records or census data at the disaggregated level that are common in both nationally representative surveys and the census. we can generate the data of the variable of interest at the district/small area level by utilizing the auxiliary or common information in both nationally representative surveys and the census. SAE methods are statistical models designed to generate or enhance the accuracy of estimates for small geographical areas or domains with limited or no sample size. The basic proposition behind these methods is to gain strength from auxiliary data sources, such as census or administrative records. SAE models commonly fall into two categories including area-level and unit-level models. Area-level models embrace the work of (Fay and Herriot 1979) and (Torabi and Rao 2014), and operate on aggregate data for each area. In contrast, unit-level models comprehend the work of (Elbers et al. 2003) and (Molina and Rao 2010), and utilize individual-level data.

The two distinct methodological approaches for small area estimation are deployed and contrasted in this study. The first approach employs the standard (Elbers et al. 2003) paradigm for error decomposition, employing a parametric approach for estimating model parameters (Shabir & Farooq 2024). In contrast, the second approach leverages an extended ELL method, as utilized by Nguyen et al. (2018), which integrates the Henderson III (H3) model for error decomposition and also implements Empirical Bayes (EB) prediction. This extended approach diverges from the original ELL by utilizing bootstrapping for parameter estimation and implementing the EB predictor to enhance the precision of area-level estimates. GLS estimates based on the ELL approach may potentially lead to an asymmetric variance-covariance matrix due to unequal weights within clusters. To mitigate this issue, the matrix is averaged with its transpose to ensure symmetry (Haslett et al. 2010). Furthermore, the integration of survey weights into the variance-covariance matrix is refined, drawing upon the advancements offered by Huang and Hidiroglou (2003) and presented by Van der Weide (2014). Consequently, the estimation of GLS parameters requires the computation of variance components. For this purpose, a modified version of Henderson's Method III Henderson (1953) is employed, specifically adapted to accommodate survey weights, as presented by Huang and Hidiroglou (2003) and further detailed by Van der Weide (2014). In light of the HIES sampling design, error decomposition is conducted using the original ELL method and the modified ELL method. In this study, two distinct model settings are utilized. Model Settings 1 refers to the ELL framework, and Model Settings 2 refers to the ELL framework with Henderson method III (H3) and EB prediction (Extended ELL method).

The ELL and Extended ELL methods generally consider nationally representative household survey data and auxiliary census information to predict individual welfare indicators. These methods involve fitting a regression model to the survey data, where the welfare indicator (per equivalent adult consumption expenditure) is the dependent variable, and numerous individual and household attributes serve as predictors. The estimated model parameters are then applied to the PSLM within the auxiliary data, generating anticipated welfare indicators for each household. These individual predictions are amassed to the desired small area level, offering measurement of poverty headcount or other pertinent welfare estimates.

2. Methodology

The study employs a two-phase approach, utilizing the ELL and its extended variant (the extended ELL method), as presented by Nguyen et al. (2018). The initial phase encompasses the estimation of core components: the Beta Model, which addresses covariate effects; the Alpha Model, which focuses on cluster-level random effects; and the Generalized Least Squares (GLS) estimation for variance components. The subsequent phase involves Monte Carlo simulations to generate small area estimates of poverty incidence, accompanied by their corresponding mean squared errors (MSEs) under the ELL and Extended ELL methods. This section delineates the empirical framework underpinning the small-area estimation techniques employed in our analysis based on both methods.

In this analytical framework, a robust econometric model is used initially to capture household per capita expenditure Y_{di}. This model, referred to as the beta model in small area estimation literature, is developed through ordinary least squares (Elbers et al. 2002) regression. The specification of the OLS regression model can be expressed as follows:

$$l_n Y_{di} = X_{di}\beta + u_{di} \quad (1)$$

In equation 1, $Y_{di} = E_{di} + c\ for\ c > 0$, c is constant and E_{di} represents per equivalent adult expenditure for i^{th} household in an area d and N_d is the size of the population in the area d. Also, D denotes the partitioned population of size N, $i = 1, \ldots, N_d$ and $d = 1, \ldots, D$. The vector X_{di} include auxiliary variables for the i^{th} household within the cluster d, reflecting the household and its characteristics. The vector β represents the coefficients to be estimated. The error term u_{di} captures the unobserved factors affecting consumption and can be decomposed into two independent components: a cluster-specific effect (η_d) and a household-specific effect (e_{di}).

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

The cluster effect (η_d) accounts for unobserved influences affecting households within a cluster. while the household-specific effect (e_{di}) reflects idiosyncratic variations in consumption patterns at the household level.

The location effect (η_d) represents the averaged household errors within a given cluster and recognizes the interdependence of household consumption decisions within a common setting.

$$u_d \sim iid(0, \sigma_u^2) \quad (3)$$

Additionally, the household-specific effect (e_{di}) is considered heteroskedastic, indicating that variance may fluctuate across households.

$$e_{di} \sim ind(0, \sigma_e^2 k_{di}) \quad (4)$$

The unconditional variance for each location is calculated by estimating equation 1 and obtaining the residuals ($\hat{u}_{di}$). By defining $\hat{u}_{d.}$ as the average of $\hat{u}_{di}$ for a specific cluster d, the household-specific error term ($\hat{e}_{di}$) can be determined:

$$\hat{u}_{di} = \hat{u}_{d.} + (\hat{u}_{di} - \hat{u}_{d.})$$
$$\hat{u}_{di} = \hat{\eta}_d + \hat{e}_{di} \quad (5)$$

Advance Journal of Econometrics and Finance
Vol-4, Issue-1, 2026

$$\hat{e}_{di} = \hat{u}_{di} - \hat{\eta}_d \quad (6)$$

The subsequent step is estimating the variances of household and location effects, which are essential for producing reliable small-area estimates. The unconditional variance of the location effect is:

$$\hat{\sigma}_\eta^2 = \max\left(\frac{(\Sigma_d \omega_d (u_{d.} - u_{..})^2 - \Sigma_d \omega_d (1-\omega_d)\hat{\tau}_d^2)}{\Sigma_d \omega_d (1-\omega_d)} ; 0\right) \quad (7)$$

Here ω_d is the weight assigned to the cluster d, the variable $u_{d.}$ is the mean of the residuals ($\hat{u}_{di}$) within cluster d, while $u_{..}$ is the overall weighted mean of the residuals across all clusters. Finally, $\hat{\tau}_d^2$ measures the variability of household expenditures within each cluster and is estimated as follows:

$$\hat{\tau}_d^2 = \frac{\Sigma_i (e_{di} - e_{d.})^2}{n_d(n_d - 1)} \quad (8)$$

Here $e_{d.}$ denotes the average residuals for all households within a cluster d, and n_d represents the total number of households in that cluster.

Following the estimation of the unconditional variance of the location effect, the heteroskedasticity inherent in the household-specific error term is modeled. Adopting the approach proposed by (Elbers et al., 2003), a parametric form for the variance of the idiosyncratic error variances ($\sigma_{e_{di}}^2$) is utilized, employing a logistic function to ensure bounded and non-negative variance predictions.

$$\sigma_{e_{di}}^2 = \left[\frac{A exp^{z'_{di}\alpha} + B}{1 + exp^{z'_{di}\alpha}}\right] \quad (9)$$

In this equation, A is an upper bound for the variance. It ensures the variance does not get too large. B is the lower bound for the variance, corroborating that it does not become negative. The vector of household characteristics z'_{di} influence the variance of the error term. While α is a vector of parameters estimated from the data, controlling variance fluctuations based on household attributes. in z'_{di}.

The study adopts a modified approach to the logistic function, drawing inspiration from the work by Elbers et al. (2003). The function is simplified by assigning values of 1.05 and 0 to parameters A and B, respectively. The Ordinary Least Squares regression is adapted to estimate this model. In this analysis, the logarithm of squared residuals serves as the response variable, while a collection of household-specific factors (z'_{di}) acts as the predictor variable.

$$\ln\left[\frac{e_{di}^2}{A - e_{di}^2}\right] = z'_{di}\alpha + r_{di} \quad (10)$$

While drawing inspiration from Harvey's (1976) seminal work on heteroskedasticity, this methodology provides a significant advantage in contemporary data analysis by constraining the variance predictions within defined bounds.

By defining $\exp(z'_{di}\alpha)$ as D and utilizing the delta method, which is rooted in a second-order Taylor expansion for the anticipated value of the variance, the following estimator is derived:

$$\hat{\sigma}_{e_{di}}^2 \approx \left[\frac{AD}{1+D}\right] + \frac{1}{2} Var(r)\left[\frac{AD(1-D)}{(1+D)^3}\right] \quad (11)$$

In this formulation, $\hat{\sigma}_{e_{di}}^2$ denotes the estimated variance of the idiosyncratic error for the household i in cluster d. Concurrently, $Var(r)$ represents the estimated variance from the model's residuals in Equation 10.

To quantify the uncertainty associated with the estimated unconditional variance of the location effect ($\hat{\sigma}_\eta^2$), the Elbers et al. (2002) proposed two methods. The first one considers simulations, and the second method is the approximation technique, as detailed below:

$$Var(\hat{\sigma}_\eta^2) = \Sigma_d 2\left\{a_d^2\left[(\hat{\sigma}_\eta^2)^2 + (\hat{\tau}_d^2)^2 + 2\hat{\sigma}_\eta^2\hat{\tau}_d^2\right] + b_d^2\frac{(\hat{\tau}_d^2)^2}{n_d - 1}\right\} \quad (12)$$

Here $a_d = \frac{\omega_d}{\Sigma_d \omega_d(1-\omega_d)}$ and $b_d = \frac{\omega_d(1-\omega_d)}{\Sigma_d \omega_d(1-\omega_d)}$

The derived sampling variance facilitates the construction of confidence intervals for σ_η^2, enabling evaluation of the statistical significance of the observed between-cluster heterogeneity in the variable of interest.

The ELL methodology initially proposed a Generalized Least Squares (GLS) estimator that utilizes the estimated variance components to construct a variance-covariance matrix (Ω) accounting for heteroskedasticity and cluster-level correlation. The construction of the $\hat{\Omega}_d$ matrix occurs at the block level, employing a sophisticated spatial correlation structure. This matrix's design incorporates two key components of variance. The off-diagonal elements encapsulate the shared variance component, denoted as $\hat{\sigma}_\eta^2$. The diagonal elements of the matrix are engineered to reflect the total variance for each household, amalgamating both the aforementioned location effect and the household-specific error variance ($\hat{\sigma}_\eta^2 + \hat{\sigma}_{e_{di}}^2$). The variance-covariance matrix of the random effects for the whole dataset is denoted as $\hat{\Omega}$. The diagonal elements represent the potential heterogeneity in correlation structures across areas. The off-diagonal blocks brimming with zeros elucidate the independence between different clusters. Following the ELL (2003) framework, the GLS estimates and their variances are obtained using the following equations:

$$\hat{\beta}_{GLS} = (X'W\Omega^{-1}X)^{-1}X'W\Omega^{-1}Y \quad (13)$$

$$Var(\hat{\beta}_{GLS}) = (X'W\Omega^{-1}X)^{-1}(X'W\Omega^{-1}WX)(X'W\Omega^{-1}X)^{-1} \quad (14)$$

In Equations 13 and 14, W denotes the diagonal matrix of sampling weights. It is imperative to recognize that the product $W\Omega^{-1}$ may not always yield a symmetric matrix. Therefore, necessary adjustments are required in that case to ensure the symmetry of the resulting variance-covariance matrix. To address the challenge of potential asymmetry in the variance-covariance matrix, particularly when dealing with unequal weights within clusters, the approach recommended by Haslett and Jones (2010) can be adopted, which involves a symmetrization process, whereby the matrix is averaged with its transpose. Furthermore, the integration of survey weights into the variance-covariance matrix (Ω) is refined based on the advancements proposed by Huang and Hidiroglou (2003) and elucidated in Van der Weide (2014). Consequently, the estimation of GLS parameters necessitates a recalibration of the variance components. For this purpose, a modified version of Henderson (1953) is employed, specifically tailored to incorporate survey weights.

To facilitate the variance components estimation associated with the Generalized Least Squares (GLS) estimator, it is imperative to modify the model by centering the response variable within each cluster. This transformation can be achieved by subtracting the weighted cluster mean from individual observations. Mathematically, this can be represented as follows: Mathematically, the transformed dependent variable for cluster d, denoted as $\tilde{y}_d$, is obtained as follows:

$$\tilde{y}_d = Y_d - (\overline{Y}_d \otimes 1_T) \quad (15)$$

In Equation 15, $\tilde{y}_d$ is the transformed dependent variable for cluster d, Y_d denotes the T x 1 vector of the initial dependent variable values for all surveyed observations within the cluster d, $\overline{Y}_d$ represents the scalar weighted mean of the response variable for cluster d, 1_T denotes the $T \times 1$ vector of ones, and $\otimes$ represents the Kronecker product. The N x 1 matrix $\tilde{y}$ stacks all the cluster-specific transformed dependent variables ($\tilde{y}_d$).

The independent variables on the right-hand side of the model undergo a parallel de-meaning procedure, yielding the demeaned matrix $\tilde{x}$. This is a $D \times K$ matrix, where D represents the number of clusters or areas in the survey and K denotes the number of independent variables containing the demeaned values. Estimating variance components requires a multi-step computational approach to obtain robust Generalized Least Squares (GLS) estimates. This process integrates the transformed outcome and predictor variables, incorporating sampling and cluster-specific weights. The estimation process entails a series of calculations, including estimating the Sum of Squared Errors (SSE). In addition, three critical intermediate terms are derived: t2, t3, and t4. These terms capture distinct aspects of the data and weighting structure. The term t2 represents the interaction between the sampling weights and the transformed independent variables, providing insight into the weighted relationships between the predictors. The term t3 combines the original independent variables (X) with the sampling weights (W) and their Hadamard product, effectively accounting for the weighted covariance structure. The term t4 integrates the original and transformed independent variables, along with the cluster-specific weights (W_d) and their Hadamard product, thereby capturing the complex interplay between the predictors, weights, and clustering structure.

$$SSE = \tilde{y}'W\tilde{y} - \tilde{y}'W\tilde{x}\left(\tilde{x}'W\tilde{x}\right)^{-1}\tilde{x}'W\tilde{y} \quad (16)$$

$$t_2 = tr\left[\left(\tilde{x}'W\tilde{x}\right)^{-1}\left(\tilde{x}'(W \circ W)\tilde{x}\right)^{-1}\right] \quad (17)$$

$$t_3 = tr\left[\left(X'WX\right)^{-1}X'(W \circ W)X\right] \quad (18)$$

$$t_4 = tr\left[\left(X'WX\right)^{-1}\tilde{x}'(W_d \circ W_d)\tilde{x}\right] \quad (19)$$

Ultimately, the residual variance is computed by leveraging the previously calculated Sum of Squared Errors (SSE) and trace terms, thereby providing a quantitative measure of the unexplained variation in the data.

$$\hat{\sigma}_e^2 = \frac{SSE}{\sum_{di}\omega_{di} - \sum_d\left(\frac{\sum_i \omega_{ai}^2}{\sum_i \omega_{ai}}\right) - t2} \quad (20)$$

Concurrently, the variance of the area-level random effects is quantified by leveraging the original outcome variable (Y), both the original and transformed predictor variables, the estimated residual variance, and the trace components.

$$\hat{\sigma}_\eta^2 = \frac{Y'WY - Y'WX\left(X'WX\right)^{-1}X'WY - \left(\sum_{di}\omega_{di} - t_3\right)\hat{\sigma}_e^2}{\sum_{di}\omega_{di} - t_4} \quad (21)$$

The estimation of area-level random effects is conducted following the framework established by Van der Weide (Van der Weide, 2014). The area-level random effect is estimated by amalgamating the cluster-specific weighted average of predicted random effects integrated with the overall average weighted by the shrinkage factors, as shown below:

$$\hat{\eta}_d = \gamma_{d,\omega}\sum_i \omega_{di}\,\hat{u}_{di} - \frac{1}{D}\sum_d \gamma_{d,\omega}\left(\sum_i \omega_{di}\,\hat{u}_{di}\right) \quad (22)$$

Where $\gamma_{d,\omega} = \frac{\hat{\sigma}_\eta^2}{\hat{\sigma}_\eta^2 + \hat{\sigma}_e^2\left(\frac{\sum_i \omega_{di}}{\sum_i \omega_{di}^2}\right)}$. After the estimation of area-level random effects, the individual-level residual errors are recalibrated, which refines the error structure by accounting for the newly estimated area-level effects.

$$\hat{e}_{di} = \hat{u}_{di} - \hat{\eta}_d - \sum_{di}(\hat{u}_{di} - \hat{\eta}_d) \quad (23)$$

This equation subtracts both the estimated area-level random effect and the average deviation of predicted random effects from the original predicted random effect. Furthermore, the distribution of $\hat{e}_{di}$ is recalibrated to ensure that its variance converges to the estimated value of $\hat{\sigma}_e^2$, thereby achieving a standardized error structure. To account for potential heteroscedasticity in the individual-level residual errors (e_{di}), a variance modeling approach is employed, leveraging a linearized framework similar to Equation 10. This framework enables the estimation of parameters α, which captures the relationship between residual variance and a set of predictor variables (Z_{di}), while accounting for inherent variability through the incorporation of an additional error term (r_{di}). Subsequently, observation-specific variance estimates ($\hat{\sigma}_{e_{di}}^2$) are derived using the approximation provided by the Equation, thereby allowing for the characterization of residual variance heterogeneity. A weighted GLS estimator for β, as detailed in Van der Weide (2014), is utilized, integrating survey weights into the variance-covariance matrix. For each cluster d, the estimated variance-covariance matrix $\hat{\Omega}_d$ is constructed, representing the covariance structure of the residuals within each cluster. The diagonal elements $\left(\frac{\sum_i \omega_{di}}{\sum_i \omega_{di}^2}\right)\hat{\sigma}_\eta^2 + \frac{\hat{\sigma}_{e_{di}}^2}{\omega_{di}}$ denotes the total variance for each observation within the cluster and the off-diagonal elements $\left(\frac{\sum_i \omega_{di}}{\sum_i \omega_{di}^2}\right)\hat{\sigma}_\eta^2$ capture the covariance between observations within the same cluster. The overall variance-covariance matrix $\hat{\Omega}$ for the entire dataset is constructed, containing each block along the diagonal corresponding to the $\hat{\Omega}_d$ matrix for a specific cluster d which reflects the assumption of independence between different areas while allowing for heterogeneity in the correlation structures within each area. Analogously, a cluster-specific variance-covariance matrix $\hat{V}_d$ is derived for each cluster d, following a similar conceptual framework to the original ELL methodology. Furthermore, the aggregate variance-covariance matrix $\hat{V}$ matrix for the entire dataset is assembled as a block diagonal structure, wherein each block corresponds to the cluster-specific variance-covariance matrix $\hat{V}_d$ for a particular cluster d, thereby capturing the heterogeneous covariance patterns across clusters. Ultimately, the GLS estimator (based on the Henderson (H3) method for β, along with its associated variance-covariance matrix, is derived using the subsequent equations:

$$\widehat{\beta}_{GLS} = \left(X'\widehat{\Omega}^{-1}X\right)^{-1}X'\widehat{\Omega}^{-1}Y \quad (24)$$

$$Var\left(\widehat{\beta}_{GLS}\right) = \left(X'\widehat{\Omega}^{-1}X\right)^{-1}\left(X'\widehat{\Omega}^{-1}\widehat{V}\widehat{\Omega}^{-1}X\right)\left(X'\widehat{\Omega}^{-1}X\right)^{-1} \quad (25)$$

The first stage of the analysis focuses on developing statistically robust parameter estimates using the ELL and extended ELL methods, enabling the derivation of reliable and efficient parameter estimates. The second stage involves a simulation-based inference approach, wherein the ELL and Extended ELL methods are employed to generate a large number of bootstrap samples from the census data (PSLM 2019-20), using the parameter and error estimates derived from the survey data (HIES 2018-19). The Monte Carlo simulations are utilized within the ELL methodological framework to derive expected welfare measures conditional on the first-stage model. In the case of the extended ELL method, Empirical Bayes (EB) prediction is employed as a precursor to the simulations, enabling the refinement of area-level estimates by leveraging both prior information from the model and observed data. This approach, as outlined in Van der Weide (2014), yields posterior estimates of area-level parameters, assuming the normality of the random effects. By integrating survey data, the EB prediction technique enhances the accuracy of location-specific effect estimates, ultimately improving the reliability of forecasts for areas included in the survey. Furthermore, under the assumption of normality for both area-level random effects and individual-level residuals, the conditional distribution of η_d given e_d is also normally distributed, as demonstrated in Van der Weide (2014). This normality supposition is crucial in deriving the distribution of the random location effects. Within the Empirical Bayes (EB) prediction framework, the posterior mean of the location-specific effect, conditional on the residuals of the households within the location, is derived as:

$$E[(\eta_d|e_d)] = \hat{\eta}_d = \gamma_{d,\omega}\left(\frac{\sum_i\left(\frac{\omega_{d,i}}{\sigma^2_{e_{di}}}\right)e_{di}}{\sum_i\frac{\omega_{d,i}}{\sigma^2_{e_{di}}}}\right) \quad (26)$$

Where $\gamma_{d,\omega} = \frac{\sigma^2_\eta}{\sigma^2_\eta+\sum_i\omega^2_{di}\left(\sum_i\omega_{di}\sum_i\frac{\omega_{di}}{\sigma^2_{e,di}}\right)^{-1}}$ is the shrinkage factor serves as a balancing mechanism, integrating information from the cluster-level predicted random effects with the overall average. Meanwhile, accounting for the estimated variances and survey weights. When the within-area sample size is substantial, and the individual-level residuals exhibit low variances, the shrinkage factor approaches 1, effectively assigning more weight to the cluster-specific information and reducing the influence of the overall average and vice versa.

The simulation protocol entails integrating parameter estimates and error estimates derived from survey (HIES 2018-19) data into census (PSLM 2019-20) data, with the ultimate goal of generating a substantial number of simulations that yield robust and reliable welfare estimates. This multi-step simulation process comprises the following key steps:

Step 1: This step entails generating simulated values of regression coefficients via random draws from a multivariate normal distribution, thereby introducing stochastic variability into the modeling framework.

$$\widetilde{\beta}_{GLS} \sim N\left(\widehat{\beta}_{GLS}, Var\left(\widehat{\beta}_{GLS}\right)\right)$$

The mean and variance-covariance structure utilized in this step are obtained from the survey (HIES 2018-19) in stage 1 estimations.

Step 2: Within the ELL variance estimation framework, the cluster-specific random effect is drawn from a normal distribution, characterized by a mean of zero and a variance of $\widehat{\sigma}^2_\eta$, thereby capturing cluster-level variability.

$$\tilde{\eta}_d \sim N\left(0, \widehat{\sigma}^2_\eta\right)$$

To incorporate uncertainty in the estimation of $\widehat{\sigma}^2_\eta$, its variance is modeled using a Gamma distribution, allowing for the quantification of uncertainty in the cluster-level variance component. The variance of $\widehat{\sigma}^2_\eta$, denoted as $Var\left(\widehat{\sigma}^2_\eta\right)$, is subsequently estimated through the application of Equation 12.

$$\widehat{\sigma}^2_\eta \sim Gamma\left(\widehat{\sigma}^2_\eta, Var\left(\widehat{\sigma}^2_\eta\right)\right)$$

In the context of Henderson's Method III (H3), the distributional form of $\widehat{\sigma}^2_\eta$ is not explicitly specified. To address this, a bootstrapping approach is employed to estimate $\widehat{\sigma}^2_\eta$, in conjunction with other pertinent parameters, at each simulation iteration. This resampling-based methodology is particularly crucial when utilizing Empirical Bayes (EB) methods. Under EB, the bootstrap procedure generates cluster-specific vectors of η_s values and estimates of $\widehat{\sigma}^2_{\eta_d}$, thereby facilitating the incorporation of cluster-level variability. Consequently, for each simulated scenario, $\tilde{\eta}_d$ values are sampled from a normal distribution with a mean η_d and variance $\widehat{\sigma}^2_{\eta d}$ for clusters represented in the survey data. In contrast, for non-surveyed clusters, $\tilde{\eta}_d$ values are sampled from a normal distribution with mean 0 and variance $\widehat{\sigma}^2_\eta$, thus accounting for between-cluster heterogeneity.

Step 3: The individual-level residual errors $\tilde{e}_{di}$ are simulated from a normal distribution, characterized by a mean of zero and a variance of $\hat{e}^2_{di}$, thereby modeling the stochastic variability in individual-level responses. Where $\hat{e}^2_{di}$ is estimated by utilizing Equation 11 in stage 1 of the analysis. Ultimately, by integrating all the estimated parameters into the census (PSLM 2019-20) data, a synthetic welfare vector is generated, thereby facilitating the creation of a simulated dataset that combines the strengths of both survey and census data sources.

$$\widetilde{Y} = X\widetilde{\beta}_{GLS} + \tilde{\eta}_d + \tilde{e}_{di} \quad (27)$$

The equation yields simulated values of the outcome variable $\widetilde{Y}$ for each unit in the census (PSLM-2019-20) data, effectively combining the fixed effects $X\widetilde{\beta}_{GLS}$, the simulated area-level random component $\tilde{\eta}_d$, and the simulated individual-level error term $\tilde{e}_{di}$. The resulting simulated welfare values $\widetilde{Y}$ form the basis for generating M simulated replicates of the target variable. Subsequently, the mean indicator and its associated standard error are computed for each domain of interest, enabling the assessment of small-area estimates and their corresponding measures of uncertainty.

3. Data description and analysis for small area unit-level models

The data description involves the identification of auxiliary variables between the HIES 2018-19 and the PSLM 2019-20. These linking variables are defined across both datasets to generate poverty data. The HIES incorporates a detailed consumption module, which is instrumental in deriving poverty metrics; however, its sampling framework ensures representativeness only at provincial and national levels, limiting its granularity for district-level analysis. Conversely, while the PSLM does not include a consumption module, it has a large sample representative at the district level. Poverty estimation is derived from the Household Integrated Economic Survey (HIES) 2018-19 using monthly per-adult

equivalent consumption expenditure as the key welfare indicator. The Cost of Basic Needs (CBN) approach is employed to establish a spatially adjusted poverty line, expressed in monetary terms (minimum monthly per-adult equivalent expenditure). Households falling below this threshold are classified as poor, enabling a standardized poverty assessment. To ensure comparability between HIES 2018-19 and PSLM 2019-20, auxiliary variables are constructed with consistent definitions, question phrasing, and categorical classifications. These variables span four key domains: household head characteristics (age, education, gender, etc.); household composition (size, dependency ratios, etc.); dwelling attributes (construction materials, drinking water, sanitation, etc.); and Asset ownership (durable goods, etc.). A rigorous matching process ensures distributional consistency of auxiliary variables across both datasets. Variables with missing values exceeding 1% of observations are excluded to preserve data integrity. For retention, auxiliary variables must exhibit weighted mean and standard deviation ratios between 0.95 and 1.05 when comparing HIES and PSLM. Finally, a seven-digit Primary Sampling Unit (PSU) codes are harmonized at the cluster level using a hierarchical structure starting from province, division, district, rural or urban, and PSU respectively.

Following data preparation, the next phase involves modeling household consumption expenditure using the HIES 2018-19 dataset. The process begins with the estimation of a Beta Model via Ordinary Least Squares, subject to the preliminary steps, including the mitigation of multicollinearity. The Variance Inflation Factor (VIF) is employed to diagnose and address multicollinearity among predictors. Variables exhibiting a VIF exceeding 7 are systematically excluded from the model to ensure robust parameter estimation. Similarly, a stepwise selection procedure (forward and backward induction) is applied, retaining variables with p-values ≤ 0.05 to optimize statistical significance. The Least Absolute Shrinkage and Selection Operator (Lasso) is also utilized to penalize model complexity, further refining the predictor set by shrinking irrelevant coefficients to zero to get the most relevant auxiliary variables. Finally, the candidate models derived from stepwise and Lasso regression are subjected to an additional stepwise selection process, evaluated based on adjusted R-squared, Akaike, and Bayesian Information Criteria. Following the selection of predictors for the Beta model, the analysis proceeds through a multi-stage estimation process. A vector of residuals is generated by using the finalized Beta model. These residuals are subsequently modeled by utilizing the auxiliary variables excluded from the Beta model, interaction terms between these auxiliary variables and the residuals, and their squared form. This specification forms the basis of the Alpha model, which is systematically evaluated for multicollinearity using Variance Inflation Factors (VIFs). To optimize the Alpha model's predictive performance, a stepwise selection procedure is implemented, retaining only statistically significant predictors. This process simultaneously calculates the unconditional variance, providing crucial information about the error structure. The final estimation employs Generalized Least Squares (GLS), which explicitly accounts for heteroskedasticity in the error terms. Compared to OLS, GLS delivers enhanced efficiency in parameter estimation, robust point estimates of regression coefficients, and proper consideration of the underlying distributions for both coefficients and error terms. This approach yields statistically efficient estimators that are consistent with the model's theoretical foundations. Additionally, to ensure the robustness of our Beta model, we conducted rigorous diagnostic checks to identify and address influential observations that could disproportionately affect model estimates. Observations with Cook's distance greater than $4/N$, absolute studentized residuals exceeding 2, and leverage values more than $(2K+2)/N$ (where N is the total number of observations and K is the number of predictors) are excluded to improve the model's robustness and reliability. The core estimation framework remains consistent between the traditional ELL method and its extended variant, with both approaches following the fundamental poverty mapping workflow. However, the Extended ELL method incorporates two key enhancements, including Henderson Method III (H3) Integration and Empirical Bayes Prediction. H3 method introduces a more sophisticated variance components estimation technique, provides improved decomposition of error structures at different hierarchical levels, and enables better separation of household-specific and area-level variation. Similarly, EB prediction implements shrinkage estimation for small area predictions, reduces mean squared error by borrowing strength across similar areas, and produces more stable estimates for domains with limited sample sizes. The current analysis employs a dual simulation framework for estimating welfare measures in small areas, utilizing parametric and bootstrap techniques within a Monte Carlo paradigm. The Parametric Approach (ELL Framework) utilizes theoretically derived normal distributions for regression coefficients and error terms. Bootstrap Approach based on Extended ELL employs empirical resampling with replacement from observed survey data.

4. Results

This study presents a rigorous methodological comparison between two established approaches for variance decomposition in small area estimation: ELL and the Extended ELL approach. The analysis systematically evaluates their respective performance in quantifying and partitioning error components, particularly their implications for precision estimation in poverty mapping applications.

Table 1 Summary Statistics

Model setting	**1**	**2**
Error decomposition	ELL	H3
Beta drawing	Parametric	Bootstrapped
Eta drawing method	normal	normal
Epsilon drawing method	normal	normal
Empirical best method	No	Yes
Beta-model diagnostics		
Number of observations	22677	22677
Adjusted R-squared	0.5700	0.5700
R-squared	0.5706	0.5706
Root MSE	0.2802	0.2802
F-stat	884.9740	884.9740
Alpha-model diagnostics		
Number of observations	22677	22677
Adjusted R-squared	0.0026	0.0023
R-squared	0.0028	0.0025
Root MSE	2.2726	2.2792
F-stat	15.8203	13.9915
Model Parameters		
Sigma ETA sq.	0.0098	0.0102
Ratio of sigma eta sq over MSE	0.1251	0.1294

Advance Journal of Econometrics and Finance
Vol-4, Issue-1, 2026

Variance of epsilon	0.0687	0.0685
Sampling variance of Sigma eta sq.	0.0000003	N/A

Source: Author's Calculations

As illustrated in Table 1, this study evaluates and contrasts two small area estimation techniques: the conventional ELL and the enhanced ELL approach that integrates Henderson's Method III with Empirical Bayes prediction. The key findings that provide meaningful insights are discussed in detail. Both methods share a common first-stage OLS regression (adjusted $R^2 = 0.57$) for predictor selection but diverge in their treatment of residual variation. The higher alpha model's adjusted R^2 is higher for the traditional ELL approach (0.0026 vs 0.0023), indicating a somewhat better explanation of variations under the ELL method. The original ELL approach demonstrates marginally better performance with a lower σ_η^2/MSE ratio (0.1251 vs 0.1294), as the lower this ratio, the lower the portion of residual errors attributable to location effect, suggesting that auxiliary information can convincingly explain the area-level heterogeneity.

Table 2 GLS Model Estimates

Variables	ELL Coefficient	Extended ELL Coefficient
Clean Water	0.0788***	0.0781***
Cooking	0.0545***	0.0552***
Dryer	0.0735***	0.0736***
Educational Attainment	0.0623***	0.0620***
Education 1	0.0526***	0.0524***
Education 4	-0.0170**	-0.0169**
Fan	0.0291***	0.0291***
Floor	0.0695***	0.0699***
Geyser	0.1866***	0.1861***
Highest Education 4	-0.0128***	-0.0128**
Highest Education 6	0.0403***	0.0399***
Internet	0.0949***	0.0950***
Iron	0.0353***	0.0351***
Language New 4	-0.0713***	-0.0686***
Log (Age)	-0.0258***	-0.0263***
Marital 2	-0.0640***	-0.0640***
Microwave	0.2238***	0.2229***
Ownership 2	-0.0705***	-0.0710***
Province 2	0.0049	0.0079
Province 3	-0.023	-0.0199
Province 4	-0.0587***	-0.0566***
Roof	0.0486***	0.0486***
Sex Ratio	-0.0586***	-0.0586***
Table	0.0800***	0.0798***
Toilet 1	0.0481***	0.0485***
UPS	0.1737***	0.1732***
Urban	0.0189**	0.0185**
Wall	0.0479***	0.0475***
Washing Machine	0.0462***	0.0465***
Water	-0.0159***	-0.0156***
Water 2	-0.0367***	-0.0369***
Water 5	0.1465***	0.1450***
Work Ratio adult	0.3619***	0.3590***
Constant	8.0786***	8.0788***

Note: If p < 0.01 (***), p < 0.05 (**), p < 0.10 (*). Source: Author's Calculations

The results presented in Table 2 demonstrate that all estimated coefficients in both the ELL and Extended ELL model specifications are statistically significant at the standard 5% significance level, except for the provincial control variables (Province 2 and Province 3, where the reference category is Province 1, which represents KPK) that failed to reach significance. This finding is particularly important as Small Area Estimation methodologies - including both the standard ELL framework and its extended variant - require that the Generalized Least Squares (GLS) parameters used for census imputation be statistically significant. The consistent significance of coefficients in both approaches validates the robustness of our predictor variables and supports their use in subsequent estimation stages. The insignificant provincial control variables do not substantially affect performance, as control variables primarily serve to account for potential confounding effects rather than directly contribute to the predictive capacity.

The second stage of analysis implements simulation-based techniques to produce district-level welfare estimates in both methodological frameworks. Building upon the first-stage parameter estimates (including regression coefficients, variance components, and location effects), this phase systematically combines information from the Household Integrated Economic Survey (HIES 2018-19) with auxiliary variables from the Pakistan Social and Living Standards Measurement (PSLM 2019-20). The methodology employs Monte Carlo simulation techniques with 200 independent replications to generate comprehensive welfare distributions. Each simulation iteration incorporates key components, including the deterministic predictions based on fixed-effect coefficients, stochastic area-level random effects, and household-specific residuals. Within the Extended ELL framework, the implementation of Empirical Bayes prediction serves to optimize the estimation of location-specific random effects, thereby improving the precision of small-area estimates. Through repeated simulation and subsequent aggregation across all iterations, the analysis produces robust poverty metrics that account for both model-based and sampling uncertainties. The simulation outcomes are analyzed through a dual-model analytical framework that enables comprehensive methodological comparison. This

approach facilitates rigorous evaluation of each model's predictive performance while assessing its respective sensitivities to different parametric specifications. Our comparative examination focuses on two distinct estimation paradigms, including the conventional Elbers-Lanjouw-Lanjouw (ELL) methodology, and its enhanced counterpart integrating Henderson III (H3) variance decomposition with Empirical Bayes (EB) prediction, hereafter denoted as ELL_H3EB or Extended ELL Method.

As depicted in Figure 1, the comprehensive visual analytics is used to examine poverty patterns across 126 districts, comparing both headcount ratios and their associated estimation errors. Both estimation methodologies reveal a coherent spatial pattern of deprivation levels across the geographical continuum. A rigorous comparative analysis of mean squared error distributions is conducted to assess and contrast the predictive performance of each modeling framework. The findings demonstrate an inverse relationship between estimation precision and methodological divergence - as MSE values escalate, the discrepancy between the two approaches becomes increasingly pronounced. The empirical results establish that the conventional ELL estimator maintains consistently superior precision metrics throughout the domain distribution, with particularly notable advantages in high-variance regions. The spatial-error analysis further reveals that while both methods capture the fundamental poverty gradient, the traditional approach achieves greater stability in uncertainty quantification, especially for peripheral districts with limited sample representation.

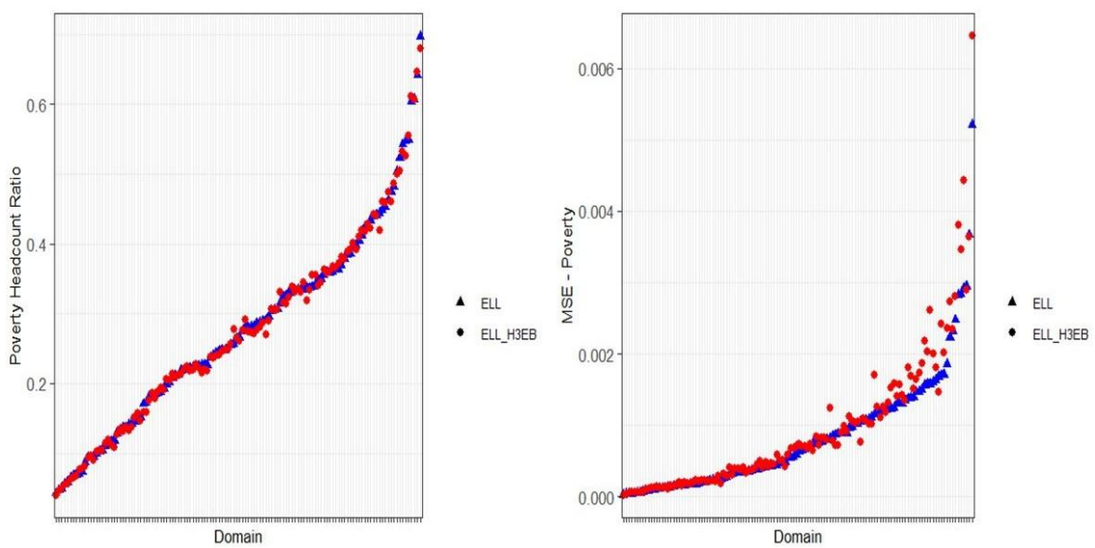

Fig.1 Poverty Estimates (left) and MSE of Poverty (right) at the district level in Pakistan. The estimates presented herein are derived from the author's analytical computations, leveraging small-area estimation (SAE) methods. These calculations integrate harmonized microdata from the HIES 2018–19 and the PSLM 2019–20

To substantiate our comparative findings, Figure 2 presents box-plot visualizations of the Mean Squared Error (MSE) of poverty estimates at both district and primary sampling unit (PSU) levels. The ELL methodology demonstrates systematically lower median MSE values across all geographical scales compared to the ELL_H3EB approach. This persistent advantage in central tendency suggests superior predictive accuracy in capturing true poverty headcount. The interquartile ranges reveal that ELL maintains tighter error distributions at the district and PSU levels. These findings collectively demonstrate that the parametric ELL framework provides more precise and consistent poverty estimates across Pakistan's administrative hierarchy. The observed performance differentials may stem from ELL's more efficient handling of the underlying data structure's heteroskedasticity, whereas the extended method's additional complexity appears to introduce greater variability without commensurate gains in accuracy.

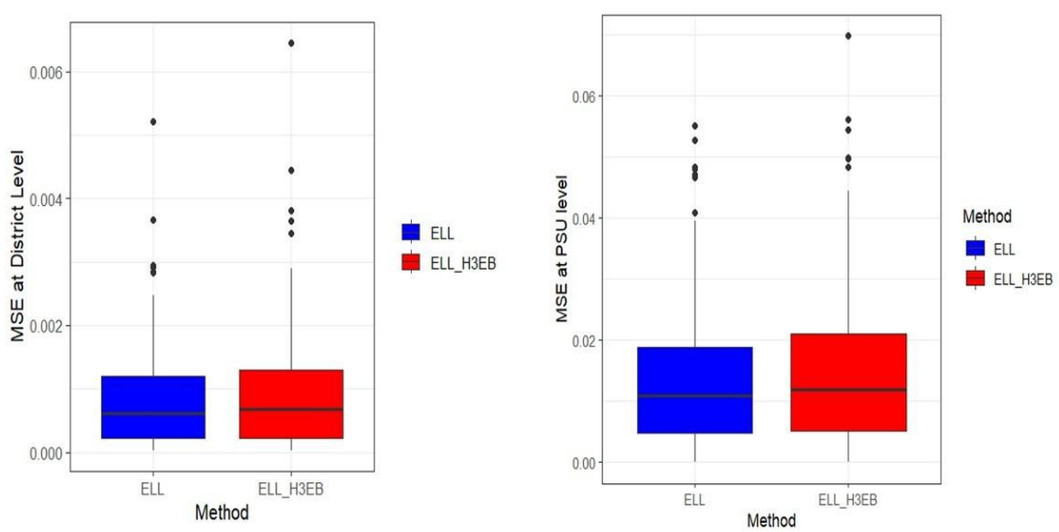

Fig.2 MSE of Poverty Estimates at the district (left) and PSU (right) level in Pakistan. The estimates presented herein are derived from the author's analytical computations, leveraging small-area estimation (SAE) methods. These calculations integrate harmonized microdata from the HIES 2018–19 and the PSLM 2019–20

The analytical process culminates with the critical task of determining household poverty status through district-level headcount estimation across Pakistan. Building upon our comparative methodological assessment, we implement the empirically validated ELL framework to generate precise poverty classifications as depicted in Figure 3.

Advance Journal of Econometrics and Finance
Vol-4, Issue-1, 2026

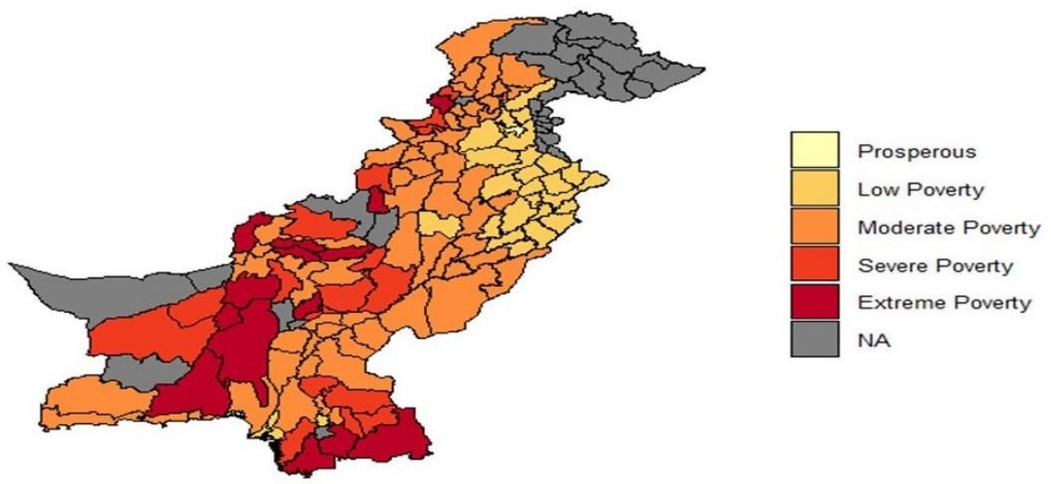

Fig.3 Poverty Mapping at the district level in Pakistan. NA denotes districts with missing data in PSLM 2019–20. Estimates derived via the ELL method of small-area estimation (HIES 2018–19 & PSLM 2019–20)

The geospatial visualization demonstrates pronounced heterogeneity in poverty concentration across Pakistan's administrative divisions. Our empirical analysis identifies thirteen districts exhibiting extreme deprivation: Sheerani, Awaran, Tharparkar, Khuzdar, Shaheed Sikandarabad, Ziarat, Killa Abdullah, Harnai, Mohmand, Kalat, Sujawal, Duki, and Nasirabad. These territories represent the most economically vulnerable hotspots, displaying poverty indicators significantly above national thresholds, thereby warranting immediate, intensive policy interventions. The analytical framework further delineates nineteen districts experiencing severe economic distress: Badin, Bajur, Tando Muhammad Khan, Barkhan, Killa Saifullah, Kachhi, Kharan, Dera Bugti, Nuski, Washuk, Thatta, Mirpur Khas, Jaffarabad, Umerkot, Shaheed Benazirabad, Khyber, Orakzai, South Waziristan, and Sohbatpur. While marginally less deprived than the extreme poverty cohort, these regions still manifest critical development deficits that demand prioritized allocation of poverty mitigation resources. Notably, our cluster analysis reveals a transitional group of districts, Rajanpur, Sanghar, Khairpur, and Shikarpur, which, while currently classified as moderately impoverished, exhibit socioeconomic indicators perilously proximate to severe poverty benchmarks. This borderline cohort represents a critical intervention zone where preventive policy measures could potentially avert further economic deterioration.

5. Conclusion

This study conducts a rigorous comparative assessment of the established small area estimation (SAE) methodologies, including the conventional Elbers, Lanjouw, and Lanjouw (ELL) framework (2003), and its enhanced variant, the Extended ELL approach (Nguyen et al., 2018). Empirical results indicate that while both methods yield broadly consistent spatial patterns in poverty distribution, the traditional ELL estimator exhibits systematically lower Mean Squared Error (MSE) values across geographical domains. This enhanced precision suggests that the original ELL framework provides greater estimation accuracy and reliability for district-level poverty measurement within the context of Pakistan than its extended counterpart. Beyond the demonstrable superiority of the traditional ELL method in estimation accuracy over its extended counterpart, both methodologies consistently identify the same geographic hotspots of deprivation across Pakistan. The analysis reveals thirteen districts suffering from extreme poverty, nineteen experiencing severe economic distress, and an additional four classified as moderately poor but perilously close to crossing into severe poverty thresholds. These findings underscore the urgent need for targeted policy interventions to alleviate poverty in these high-priority districts. The convergence of results across both methodological approaches strengthens the reliability of these poverty mappings, providing a robust evidence base for policymakers to design and implement precision-targeted poverty reduction strategies. The persistent identification of these disadvantaged regions across independent estimation frameworks reinforces their status as key intervention zones requiring immediate and sustained developmental focus. This study recognizes the valuable opportunity for methodological expansion through the incorporation of the Empirical Best Predictor (EBP) framework (Molina and Rao, 2010) and its subsequent methodological developments in future investigations. The integration of these advanced Bayesian prediction techniques would allow for comprehensive triangulation of results across multiple estimation paradigms, potentially yielding nuanced validation of current poverty mapping outcomes.

References:

Corral P, Molina I, Cojocaru A, Segovia S (2022) Guidelines to small area estimation for poverty mapping. World Bank, Washington
Elbers C, Lanjouw JO, Lanjouw P (2002) Micro-level estimation of welfare. World Bank Publications
Elbers C, Lanjouw JO, Lanjouw P (2003) Micro-level estimation of poverty and inequality. Econometrica 71**:**355-364.
Fay R, Herriot R (1979) Estimates of income for small places: an application of James-Stein procedures to census data. J Am Stat Assoc 74**:**269-277.
Haslett S, Jones G (2010) Small-area estimation of poverty: the aid industry standard and its alternatives. Aus & N Zea J of Stat 52**:**341-362.
Haslett S, Jones G, Noble A, Ballas D (2010) More for Less? Comparing small area estimation, spatial microsimulation, and mass imputation. JSM 1584-1598.
Henderson CR (1953) Estimation of variance and covariance components. Biometrics*,* 9**:**226-252.
Huang R, Hidiroglou M (2003) Design consistent estimators for a mixed linear model on survey data: Proceedings of the Survey Research Methods Section. J Am Stat Assoc 1897-1904.
Molina I, RAO JN (2010) Small area estimation of poverty indicators. Canadian Journal of Statistics 38**:**369-385.
Nguyen M, Corral R, Azevedo JP, Zhao Q (2018) SAE: A Stata package for unit-level small area estimation. World Bank Policy Research Working Paper.
Torabi M, Rao J (2014) On small area estimation under a sub-area level model. J Mul Anal, 127**:**36-55.

Van Der Weide R (2014) GLS estimation and empirical bayes prediction for linear mixed models with heteroskedasticity and sampling weights: a background study for the povmap project. World Bank Policy Research Working Paper.

Shabir S, Farooq S (2024) Poverty Mapping at District Level in Pakistan: Application of the ELL Method of SAE. Rem Rev 9:2703-2721.