Measuring Welfare Impacts in Pakistan: Application of Small Area Estimation

By

Shahid Shabir
PIDE2019FPHDETS02

Supervisor
Dr. Shujaat Farooq

PIDE School of Economics

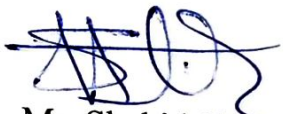
Pakistan Institute of Development Economics, Islamabad

2025

Author's Declaration

I Mr. Shahid Shabir hereby state that my PhD thesis titled **"Measuring Welfare Impacts in Pakistan: Application of Small Area Estimation"** is my own work and has not been submitted previously by me for taking any degree from **Pakistan Institute of Development Economics, Islamabad'** or anywhere else in the country/world.

At any time if my statement is found to be incorrect even after my Graduation the university has the right to withdraw my PhD degree.



Mr. Shahid Shabir

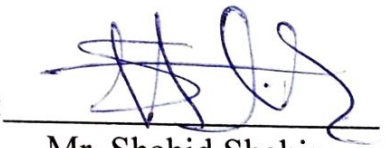
PIDE2019FPHDETS02

Plagiarism Undertaking

I solemnly declare that research work presented in the thesis titled **"Measuring Welfare Impacts in Pakistan: Application of Small Area Estimation"** is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the **HEC** and **Pakistan Institute of Development Economics, Islamabad** towards plagiarism. Therefore I as an Author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of PhD degree, the University reserves the rights to withdraw/revoke my PhD degree and that HEC and the University has the right to publish my name on the HEC/University Website on which names of students are placed who submitted plagiarized thesis.

Students/Author Signature: 

Mr. Shahid Shabir

PIDE2019FPHDETS02

Certificate of Approval

This is to certify that the research work presented in this thesis, entitled: **"Measuring Welfare Impacts in Pakistan: Application of Small Area Estimation"** was conducted by **Mr. Shahid Shabir** under the supervision of **Dr. Shujaat Farooq.** No part of this thesis has been submitted anywhere else for any other degree. This thesis is submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Econometrics from **Pakistan Institute of Development Economics, Islamabad.**

Student Name: Mr. Shahid Shabir
PIDE2019FPHDETS02

Signature: 

Examination Committee:

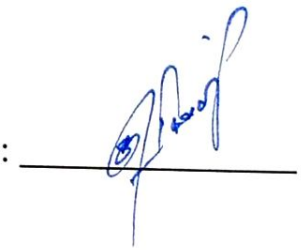
a) **External Examiner: Dr. Zahid Asghar**
Professor,
Quaid-e-Azam University
Islamabad

Signature: 

b) **Internal Examiner: Dr. Hafsa Hina**
Assistant Professor,
PIDE, Islamabad

Signature: 

Supervisor: Dr. Shujaat Farooq
Professor,
PIDE, Islamabad

Signature: 

Dr. Iftikhar Ahmad
Director PSE,
PIDE School of Economics (PSE)
PIDE, Islamabad

Signature: 

Dedication

My beloved wife, the beacon of light guiding me through the labyrinth of challenges.

My precious children, the wellsprings of joy inspiring me to greater heights.

My cherished parents—both biological and in-law—the foundations of wisdom and strength that have shaped my journey.

Your unwavering support has been the wind beneath my wings, your love the compass directing my course, and your faith the fuel propelling me forward. Your prayers have been the invisible threads weaving success into the fabric of my endeavors.

ACKNOWLEDGEMENTS

As this exhilarating chapter of my life draws to a close, I find myself overwhelmed with gratitude. First and foremost, I extend my deepest thanks to Divine, the Architect of the Universe, for being my unwavering compass throughout this journey. Your infinite mercy and grace have been the bedrock of my resilience, gifting me the fortitude, insight, and dedication needed to reach this academic pinnacle.

To my esteemed mentor and Supervisor, your guidance has been nothing short of a transformation. Your wisdom and support have been the catalyst for my intellectual growth, pushing me beyond perceived limitations.

I owe an immense debt of gratitude to my fathers - both by blood and by law. Your steadfast love and encouragement have been my anchor. To my father-in-law, your sage advice and unwavering faith in my abilities have been a wellspring of inspiration. You exemplify the virtues of perseverance and commitment, values that have been instrumental in shaping my academic pursuits.

To my mother, your selfless sacrifices and unshakeable belief in me have been the wind beneath my wings. Your nurturing spirit has been the silent force propelling me towards success.

My beloved wife, you are the unsung hero of this odyssey. Your love, patience, and understanding have been my sanctuary in tumultuous times. You've been my partner in every sense, celebrating triumphs and navigating challenges by my side. Your sacrifices have allowed me to chase my dreams while maintaining our family's equilibrium. My dear, this achievement is as much yours as it is mine.

To my precious daughters, Hareem Fatima and Anaya Fatima, and son Muhammad Ibrahim Shahid, your resilience and understanding have been a constant source of amazement. Your curiosity and zest for life have been a poignant reminder of why I embarked on this academic quest. Your ability to adapt to my demanding schedule has not gone unnoticed. You are my inspiration, my joy, and my motivation to strive for excellence.

This thesis stands as a testament to my family's collective support, love, and sacrifices. It's a culmination of my efforts and the unwavering backing of those I hold dear.

Thanks to all.

Shahid Shabir

ABSTRACT

This study pushes the boundaries of Small Area Estimation (SAE) methodologies by conducting a rigorous comparative analysis of three state-of-the-art techniques: the Elbers, Lanjouw, and Lanjouw (ELL) method, the Extended ELL method, and the Census Empirical Best (EB) method. The research addresses a critical gap in granular welfare data at the sub-population level in Pakistan, focusing on generating high-resolution district-level estimates for poverty and food insecurity indicators. This study implements a robust Monte Carlo simulation framework, involving 200 iterations of Monte Carlo simulations and bootstrap procedures for each household under the umbrellas of ELL, extended ELL, and Census EB methods. This intensive computational approach ensures the stability and reliability of the estimates, providing a solid foundation for subsequent analyses and policy recommendations. Leveraging data from the Household Integrated Economic Survey (HIES) and Pakistan Social and Living Standards Measurement (PSLM) surveys, the study employs econometric framework. This includes the application of both Ordinary Least Squares and Generalized Least Squares (GLS) regressions, augmented by the Restricted Maximum Likelihood (REML) method for parameter estimation. The research rigorously assesses the performance of these models in estimating Poverty headcount, operationalized through per-adult equivalent consumption expenditure, and food insecurity headcount, operationalized through kilocalories per capita per day.

The comparative analysis reveals the ELL method's superior efficacy over the extended ELL approach, demonstrating consistently lower Mean Squared Error (MSE) values across domains and enhanced alignment between estimates. The Census EB approach refines these estimates and shows superior performance over the ELL approach, exhibiting remarkable precision and reliability, particularly in areas with limited sample sizes. This

methodological advancement significantly outperforms direct estimation techniques, especially in data-scarce environments. Through the application of these advanced SAE techniques, the study unveils a delicate spatial distribution of welfare indicators across Pakistan. Key findings include the identification of districts experiencing extreme poverty and food insecurity conditions. Notably, the research pinpoints several districts that are epicenters of dual vulnerability, exhibiting critical levels in both poverty and food insecurity dimensions.

The study extends its analytical scope to explore the nexus between rural transformation (RT) and poverty alleviation through econometric modeling and typological analysis. Results indicate a statistically significant negative correlation between measures of rural transformation and rural poverty headcount, underscoring the pivotal role of structural changes in rural economies for effective poverty reduction strategies. Furthermore, the research delves into the impact of formal agricultural credit on rural income dynamics, revealing a positive yet inelastic relationship. This finding suggests the necessity for a more holistic approach to rural development that transcends mere credit access. The study also uncovers significant spatial heterogeneity in the credit-income nexus across Pakistan's districts, identifying both high-performing areas and regions where credit disbursement has translated into commensurate income gains and low-performing areas where low credit disbursement results in low-income levels.

This multifaceted research not only contributes substantially to the methodological discourse on SAE but also provides actionable, evidence-based insights for policymakers. The high-resolution district-level estimates and comprehensive analysis of socioeconomic dynamics offer a robust framework for designing targeted, context-specific interventions

to address poverty and food insecurity. Ultimately, this study lays the groundwork for more subtle, effective policy formulation aimed at fostering sustainable and inclusive economic growth in Pakistan and potentially in analogous developing economies.

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

TABLE OF CONTENTS

LIST OF FIGURES

Number *Page*

LIST OF TABLES

LIST OF ABBREVIATIONS

ADER Average Dietary Energy Requirements

BHF Battese, Harter, and Fuller

BIC Bayesian Information Criteria

BLUP Best Linear Unbiased Predictor

BP Best Predictor

CBN Cost of Basic Needs

CEBP Census Empirical Bayes Prediction

CV Control Variables

DEC Dietary Energy Consumption

DER Dietary Energy Requirement

EB Empirical Bayes/Best

EBLUP Empirical Best Linear Unbiased Predictor

EBP Empirical Bayes/Best Prediction

ELL Elbers Lanjouv and Lanjouv

FAO Food and Agriculture Organization

GLMM Generalized Linear Mixed Models

GLS Generalized Least Square

HIES Household Integrated Economic Survey

HVAP High-Value Agriculture Products

KCAL Kilocalories

Lasso Least Absolute Shrinkage and Selection Operator

MDER Minimum Dietary Energy Requirements

ML Maximum Likelihood

MSE Mean Squared Error

MSPE Mean Squared Prediction Error

NER Net Enrolment Rate

NERM Nested Error Regression Model

OLS Ordinary Least Square

PDGN Pakistan Dietary Guidelines for Better Nutrition

PIDE Pakistan Institute of Development Economics

PPCAL Price per Calorie

PSLM Pakistan Social and Living Standards Measurement

PSU Primary Sampling Unit

REML Restricted Maximum Likelihood

RRMSE Relative Root Mean Squared Error

RT Rural Transformation

SAE Small Area Estimation

SDG Sustainable Development Goals

UN United Nations

VIF Variance Inflation Factor

WLS Weighted Least Square

CHAPTER 1

INTRODUCTION

1.1 Background

Welfare impacts[1] like poverty reduction, improvement in living standards (in terms of enhanced income), and food insecurity alleviation are very important Sustainable Development Goals (SDGs) and vital for sustainable economic growth. The United Nations' SDGs represent a global call to action aimed at eradicating poverty, safeguarding the planet, and ensuring prosperity for all by 2030[2]. SDG 1, focused on poverty eradication, is central to this agenda, and SDG 2 is dedicated to reducing food insecurity. These intertwined goals are fundamental to achieving sustainable development and guarantee that every individual has access to the resources necessary for a life of dignity and well-being. SDG 1 seeks to eliminate poverty in all its forms, recognizing that poverty extends beyond mere financial deprivation to encompass social exclusion, vulnerability, and marginalization from decision-making processes. Similarly, SDG 2 strives to end hunger, achieve food security and improved nutrition, and promote sustainable agricultural practices. Despite these aspirations, developing countries are far behind in achieving these critical goals. Recent reports underscore this troubling lag. The World Bank's Poverty and Shared Prosperity 2022 report reveals that the COVID-19 pandemic has reversed years of progress in poverty reduction, with an estimated 70 million people pushed into extreme poverty in 2020 alone, primarily in

[1] "Welfare Impacts" specifically refers to measuring and analyzing two critical socioeconomic indicators: poverty and food insecurity. These variables serve as key metrics for assessing the overall well-being and quality of life within a population, providing crucial insights into the effectiveness of social policies and interventions.

[2] United Nations, "Sustainable Development Goals," https://www.un.org/sustainabledevelopment/sustainable-development-goals/.

low- and middle-income countries (World Bank, 2022). Furthermore, the State of Food Security and Nutrition in the World Bank (2022) report indicates that nearly 828 million people were affected by hunger in 2021, with the situation particularly dire in Africa, Asia, and Latin America. A significant challenge impeding progress is the scarcity of reliable, granular data at the small area level in many developing countries. This data deficit creates a paradoxical situation where the areas most in need of targeted interventions are often the least understood in terms of their specific welfare metrics. The United Nations Statistics Division (2022) notes that many developing countries still lack the capacity to collect, analyze, and disseminate high-quality, timely, and disaggregated data essential for SDG monitoring and implementation. The absence of representative data at local levels hinders the development and implementation of effective, tailored strategies. Without accurate information on poverty and food insecurity rates at a granular level, policymakers and aid organizations struggle to allocate resources efficiently and design interventions that address the unique needs of each community. To bridge this critical information gap, advanced statistical techniques such as Small Area Estimation have emerged as invaluable tools. These methods allow for the estimation of welfare indicators in areas where direct survey data is limited or non-existent, by combining available survey data with auxiliary information from other sources. By leveraging SAE and similar approaches, a more detailed picture of poverty and food insecurity can be painted at the local level. This enhanced understanding is crucial for developing targeted, evidence-based policies and interventions that can effectively address the root causes of poverty and food insecurity in the most vulnerable communities. The Small Area Technique is a tool used to

estimate a variable at the “Small Area Level” representing a subset of a larger population. 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. Small area estimation methods are statistical techniques designed to enhance the accuracy of estimates for small geographical areas or domains with limited sample sizes. 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, exemplified by the work of Fay III and Herriot (1979) and Torabi and Rao (2014), operate on aggregate data for each area. In contrast, unit-level models, such as those proposed by Elbers, Lanjouw, and Lanjouw (2003), and Molina and Rao (2010) utilize individual-level data. Elbers et al. (2003) introduced the method, which is a widely used unit-level small-area estimation technique that has gained prominence in poverty mapping and welfare analysis. Rooted in the principles of microsimulation, the ELL method leverages household survey data alongside auxiliary information from census or administrative sources to predict individual welfare indicators. The ELL method involves fitting a regression model to the survey data, where the welfare indicator (e.g., income or 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 in the auxiliary data, generating predicted welfare indicators for each household. These individual predictions are subsequently aggregated to the desired small area level,

providing estimates of poverty rates or other relevant welfare measures. Molina and Rao (2010) suggested estimating general nonlinear indicators using the Bayes/Best Predictor based on the nested error model. The Empirical Bayes Prediction (EBP) method requires identifying the survey units in the census of auxiliary variables, to identify sample and out-of-sample units. The EBP method addresses the issue of correlations of the variables of interest. It integrates a fitted area-level model for variables of interest and estimates the area location effect. It also uses different approaches to calculate uncertainty with more precision.

This study focuses on the application and comparison of ELL, extended ELL, and EBP unit-level SAE methods. Model-based methods, as described by (Tzavidis, Marchetti, & Chambers, 2010) assume a data-generating process that aligns with the model's assumptions, treating the target parameters as random. In our study, we generate poverty and food insecurity data at the district level by using HIES and PSLM surveys of Pakistan[3]. The estimation methods we consider include the ELL method (Elbers et al., 2003), the Extended ELL method (Nguyen, Corral Rodas, Azevedo, & Zhao, 2018), and the Census EB method which is extended over the EBP (Molina & Rao, 2010). Notably, instead of the original EBP method, we apply the Census Empirical Bayes Prediction (CEBP) variant described by (Molina, Rao, & Datta, 2015). We generate and compare the poverty and food insecurity data by following the updated guidelines based on the ELL method offered by (Elbers et al., 2003), the extended ELL method

[3] HIES 2018-19 and PSLM 2019-20 is used in Chapter 4 and 5 to ascertain the current state of poverty and food insecurity. Similarly, for panel analysis, the data at two points in time is considered and PSLM rounds 2008-09 and 2019-20 are utilized.

by (Nguyen et al., 2018),[4] and the Census EB method in updated form by (Corral, Himelein, McGee, & Molina, 2021)[5].

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. However these studies are specific and for a certain period, the methodology used in these studies is obsolete now. In contrast, this Research study incorporates three state-of-the-art techniques to ascertain the state of poverty and food insecurity at the district level in Pakistan. We also study the relationship between rural poverty and rural transformation and rural income with agriculture credit (lending). A panel study is conducted at the district level, wherein 02 rounds of PSLM (2008 and 2019) are considered.

1.2 Statement of Problem

The major problem faced by researchers is the non-availability of the data at the sub-population level which impedes the researchers from inferring about the current welfare situation at the small area level which further leads to a lack of concrete policies through which welfare goals (like poverty and food insecurity reduction) can be attained comprehensively, so that sustainable economic growth can be achieved.

1.3 Research Problems

The research problems stem from the broader issues outlined in the Statement of Problem and can be articulated as follows:

- The core research problem is rooted in a critical absence of granular, disaggregated data at the sub-population or small-area level. This data

[4] For unit-level ELL method application.
[5] For the comparison of unit level ELL and EBP methods of SAE.

deficiency creates a cascade of challenges, as it severely impedes the ability to make reliable inferences about localized welfare conditions. Consequently, the development of targeted and effective policies for poverty and food insecurity is hindered.

- Methodological gaps: Consequently, there is a clear methodological imperative to develop or adapt statistical techniques capable of overcoming these data limitations to produce reliable small-area estimates and foster a path toward equitable and sustainable economic growth.

1.4 Research Questions

- How can Small Area Estimation techniques be effectively employed to generate reliable poverty and food insecurity estimates at the district level in Pakistan?
- Which SAE method (ELL, extended ELL, or Census EBP) provides the most accurate and precise estimates of poverty and food insecurity at the district level, as measured by area-specific MSE?
- How do the generated district-level estimates of poverty and food insecurity contribute to a more nuanced understanding of welfare disparities across Pakistan? Which districts in Pakistan can be identified as most vulnerable in terms of poverty and food insecurity
- What is the nature and strength of the relationship between rural poverty and rural transformation at the district level in Pakistan? How do the speeds of rural transformation and rural poverty vary across districts, and what implications does this have for balanced regional development?

- How does the availability of agricultural credit correlate with rural income levels across districts in Pakistan? How do the speeds of rural income growth and agricultural credit expansion vary across districts, and what implications does this have for balanced regional development?

1.5 Objectives of the Research

- The objective of this study is to generate the Poverty and food insecurity data at the district level in Pakistan by using the special unit-level Technique of Small Area Estimation, to identify the most deprived districts in Pakistan, which further leads:
- To evaluate the best method among the ELL, Extended ELL, and EBP methods based on the MSE of the indicator of interest and area-specific MSE, and to generate the Poverty and food insecurity data to identify vulnerability across districts in Pakistan based on the best methods.
- To explore the relationship between rural poverty and rural transformation, and rural income and agricultural credit (lending), at the district level in Pakistan. The analysis also involves examining the speed of poverty reduction and conducting a Typology analysis to identify the most vulnerable districts, enabling better monitoring of poor districts in the future in terms of poverty reduction. Similarly, through typology analysis, the speed of rural income and agricultural credit is also monitored.

1.6 Contribution of the Study

This research conduct a comprehensive comparative analysis of three advanced unit-level SAE methods. The research advances the current understanding of SAE techniques by implementing the latest variants of these SAE methods. The study contributes to the refinement and application of cutting-edge statistical techniques in poverty and food insecurity estimation. Through a rigorous comparison of the ELL, Extended ELL, and Census EB methods, the research provides valuable insights into the relative performance and suitability of these approaches within the Pakistani context, along with the identification of more deprived districts that require urgent interventions and policy implications, addressing a gap in the literature on method selection for developing countries.

1.7 Significance of Study

By using the small area estimation technique, we can generate the poverty and food insecurity data at the district level in Pakistan to identify the state of poverty and food insecurity at the district level by utilizing the HIES and PSLM rounds. The study enables us to identify the deprivation status of districts by choosing the best methods among the ELL (or extended ELL) and EBP methods for data generation at the district level in Pakistan. The study also permits ascertaining the speed of rural poverty along with how rural transformation and credit can play a role in poverty reduction at the district level in Pakistan. The study will help Government policymakers to establish well-designed policies by keeping in mind the status of households at district levels in Pakistan. Granular poverty and food insecurity data at the district level serve as crucial inputs for evidence-based policy formulation and targeted resource allocation. By

identifying highly vulnerable districts, governments can implement spatially-tailored interventions and channel development funds more effectively. This precision in resource distribution facilitates the initiation of context-specific projects designed to stimulate local economies and enhance employment opportunities. Such targeted approaches have the potential to catalyze socioeconomic mobility and elevate the quality of life for residents in these high-priority areas.

1.8 Study Plan

Chapter 1 discusses the introduction of this research study, including the introduction to the SAE technique, a statement of the problem, research questions, objectives, contribution, and significance of the study. Chapter 2 discusses the literature review where different theories related to the SAE are discussed. Chapter 3 is based on research methodology where we discuss ELL and EB methods of SAE along with extensions. In Chapter 4, by utilizing the ELL and extended ELL method of SAE, the data on poverty and food insecurity is generated using HIES 2018-19 and PSLM 2019-20 to check the poverty status at the district level in Pakistan, and we compare both methods ELL and extended ELL to decide the best method for further comparison with EBP method. In Chapter 05, we generate the poverty and food insecurity data at the district level by using the EBP (Empirical Bayes prediction) method[6] of (Molina & Rao, 2010) in its updated form (Census EB method), and we compare both methods ELL and extended ELL whichever is best (from Chapter 4) with Census EB method to decide the best method for data generation context. In Chapter 6, the relationship of rural poverty is studied with rural transformation, and formal agriculture credit with

[6] In this study, the EBP refers to the Census EBP method due to the structure of HIES and PSLM.

rural households' income at the district level in Pakistan. Finally, section 7 is the overall conclusion, along with policy recommendations and limitations of the study.

CHAPTER 2

REVIEW OF LITERATURE

2.1 Overview of Small Area Estimation

The evolution of Small Area Estimation represents a significant paradigm shift in statistical methodology, addressing the limitations inherent in traditional survey-based approaches. This progression was catalyzed by the increasing demand for granular, localized data in an era where policy decisions and resource allocations require delicate, area-specific insights. Initially, researchers relied on direct survey estimates to gauge various socioeconomic indicators at sub-national levels. (Horvitz & Thompson, 1952) introduced one of the direct estimators known as the HT estimator which is used to estimate the total and mean of population in a stratified sample. (Hájek, 1971) also introduced HA estimator. The HA Estimator is a weighted average of the sample observations from the area, and it uses weights as the sampling weights. These direct estimators are unbiased based on the assumption that the possible samples that are drawn from the population are to the corresponding sample design. However, these methods often fell short when confronted with insufficient sample sizes in smaller geographic units or specific demographic subgroups. This inadequacy led to imprecise estimates, potentially misguiding policy interventions and resource distribution. To overcome these challenges, statisticians and researchers developed model-based SAE techniques. These methods leverage the power of statistical modeling to produce more reliable estimates for areas with limited sample data. The evolution of SAE methodologies has been marked by several seminal contributions. Fay and Herriot (1979) introduced an area-level model that combines direct estimates with area-specific

auxiliary information, laying the groundwork for many subsequent developments in SAE. The Elbers, Lanjouw, and Lanjouw (2003) revolutionized poverty mapping by utilizing unit-level models that integrate survey and census data, and Nguyen et al. (2018) have further refined these methodologiesMolina and Rao (2010) proposed the EBP approach, which offers improved efficiency in estimating complex, non-linear indicators such as poverty measures. These model-based approaches share a common thread that they are designed-biased estimators that leverage auxiliary information and account for location effects to compensate for the deficiencies in direct estimates stemming from limited sample sizes. By borrowing strength from related areas and utilizing correlations with auxiliary variables, these methods can produce more stable and reliable estimates for small areas.

Small Area Estimation has become an indispensable tool within official statistics, enabling the production of reliable estimates for sub-national regions or domains where sample sizes are often insufficient for direct estimation. This framework outlines a structured approach to SAE, consisting of three primary stages: specification, analysis and adaptation, and evaluation (Tzavidis, Zhang, Luna, Schmid, & Rojas-Perilla, 2018).

Figure 2.1 illustrates the comprehensive methodological framework for generating small-area estimates, encompassing a tripartite process from conceptualization to validation. This schematic representation delineates the integral stages of the estimation procedure, beginning with the initial specification phase, progressing through the analytical core, and culminating in a rigorous evaluation protocol. The

diagram encapsulates the systematic approach employed in this study, providing a visual roadmap of the methodological continuum that underpins the production of granular, spatially disaggregated estimates.

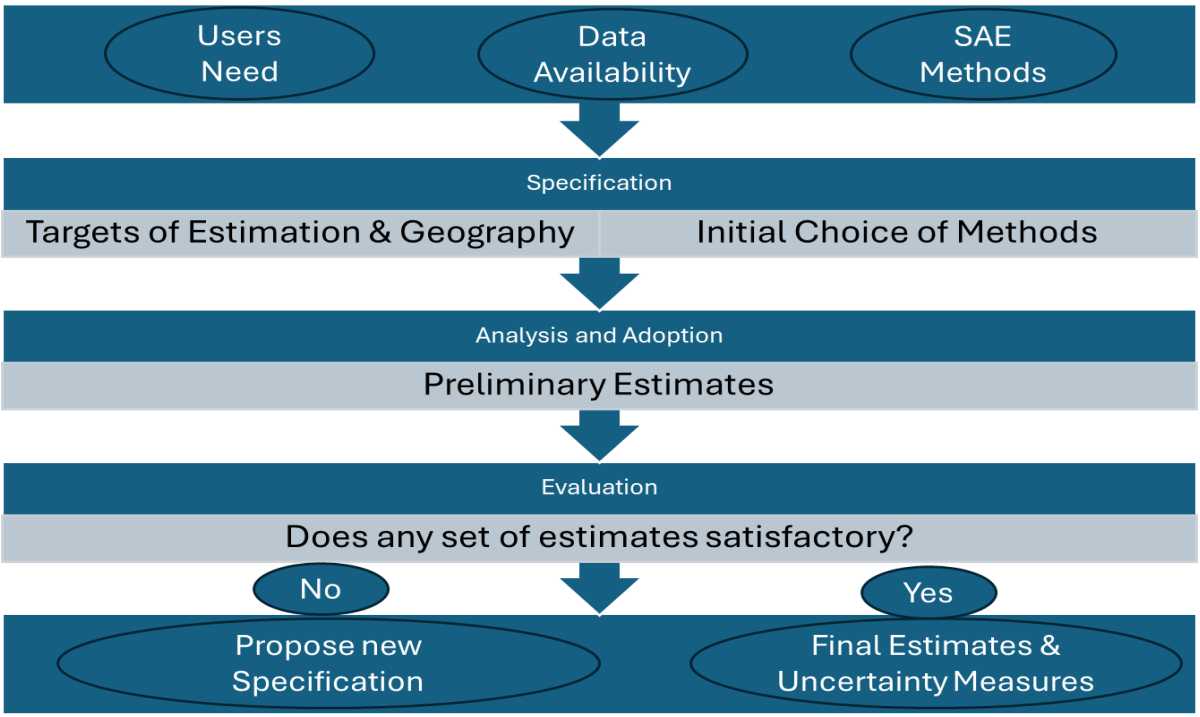


Figure 2.1 **Framework of Producing Small Area Estimates**

Source: Tzavidis et al. (2018)

2.2 Background of SAE

Sub-administrative units for which data is required are referred to as areas or domains. To estimate a particular variable of interest (like poverty or food insecurity) for these areas/domains, we rely on a direct estimator. These direct estimators are unbiased based on the assumption that the possible samples that are drawn from the population are to the corresponding sample design. However, if the survey is not designed to a small area level or if considered but the sample size is not sufficient, then the direct estimator

produces biased results in terms of large sampling errors which further leads to poor quality estimates (Molina et al., 2015). To cope with this problem, indirect estimation methods are generally deemed suitable to consider sample data related to a particular area of interest as well as from other areas by using information from auxiliary variables that are related to the variable of interest. The relationship is supposed to be the same for all areas/domains and is represented through a model that connects all the areas/domains employing conventional/common parameters. These parameters are then measured using the sampled information from all the areas/domains that result in more efficient parameters (particularly in terms of mean squared error) as compared to the direct method. The simplest indirect estimators are often biased due to the reason that they are based on unrealistic hypotheses. (Purcell & Kish, 1980) explained the synthetic estimation procedure wherein the variable of interest is estimated at some higher level for different sub-classes of the population then it scales it down in proportion to the small areas of interest. However synthetic estimators are biased when the same do not account for heterogeneity that usually exists between different areas. The most famous synthetic estimators are post-stratified and regression-synthetic (model-based) estimators. To overcome the bias of the synthetic estimator, one of the classical estimators is the composite. Albacea (1999) ascertained that composite estimators generate more reliable results compared to design-based direct estimators. The drawback of these composite estimators is the weight assigned to these estimators that does not depend on the model's goodness of fit which is a basic assumption for synthetic estimators. Similarly, weights assigned to the direct estimator are close to one that indicates some information is borrowed from other areas resulting in the in-

efficiency of the direct estimate. The most famous synthetic estimators are post-stratified and regression-synthetic estimators. The most refined indirect estimators that can better capture the between-area heterogeneity are based on mixed regression models. The two most famous models are the area-level models and the unit-level models.

2.2.1 Area Level Models

Small area estimation methods are essential statistical tools for deriving reliable estimates for regions with limited sample sizes. These methods "borrow strength" from auxiliary data sources such as census or administrative records, thereby enhancing the efficiency of estimates compared to those obtained using direct survey data alone. Model-based SAE techniques can be categorized into area-level models (Fay III & Herriot, 1979; Torabi & Rao, 2008) and unit-level models (Molina et al., 2015). Model-based methods operate under the assumption of a data-generating process that aligns with the model's assumptions, treating target parameters as random whereas design-based simulation, on the other hand, involves applying the selected methods under realistic conditions (Tzavidis et al., 2018). Small area estimation is any of several statistical techniques involving the estimation of parameters in small ‘sub-populations’ of interest included in a larger “survey”. The term “small area” in this context generally refers to a small geographical area such as a county, census tract, or district (Ghosh, 2020).

The term small area is described as a domain for which the survey data alone cannot provide reliable direct estimates (Molina et al., 2015) either because the sample size is too small or because there is no sample at all for the domain. Small area estimation

refers to techniques to “borrow strength” from auxiliary variables, typically through modeling. Brackstone (1987) discussed the potential uses of administrative records in the production of a wide range of official statistics and pointed out their merits and demerits. There are various ways administrative records can be used to produce small-area statistics. Some methods are purely based on administrative registers, commonly referred to as register-based methods, and use no survey data at all. Such use of administrative records in small area estimation can be traced back to the eleventh century in England and the seventeenth century in Canada (Brackstone, 1987). L.-C. Zhang and Fosen (2012) constructed register-based small-area employment rates and evaluated the progressive measurement errors in the small-area estimates, using historical data from the Norwegian Employer Employee Register (NEER). L. C. Zhang and Giusti (2016) discussed register-based small-area methods. Demographers have been using administrative records such as birth, death, migration, housing records, etc., in conjunction with the population census to estimate the population for small areas for a long time. Ghosh and Rao (1994) have done remarkable work in terms of demographic methods. A recent comprehensive review of various techniques in small-area estimation can be found in (Rao & Molina, 2015).

Fay III and Herriot (1979) proposed area-level linear mixed models that use only aggregated data/information to estimate areas. The Area level model has further two-level formation. In the first level, the variable of interest (income) is considered to be linearly associated with the auxiliary variables that is the decrease/increase in area mean income is the same in all areas by keeping another factor constant. In other words, the variability due to the modeling process is first ascertained using ordinary least

square (Elbers, Lanjouw, & Lanjouw, 2002). In the second level, in the presence of true values of area variables (or indicators) of interest, the values of the corresponding direct estimator are centered on these true values with known variances that represent the sampling errors of the direct estimators and depend upon the area sample size. In other words, the weighted least square (WLS) procedure is applied together with sampling variance, the predicted values then serve as the regression synthetic estimates. The FH-based models are successful in the sense that estimators for the areas are the result of weighted averages between direct estimators and synthetic regression estimators with weights depending on the area sample size. So, when the synthetic model fails to explain between-area heterogeneity, or if the sample size of an area is huge, then the estimator assigns more weight to the direct estimator which further enhances accuracy. Similarly, if the synthetic model fits well, or the area sample size is not large, the estimator assigns more weight to the synthetic regression estimator which leads to increased efficiency, as the synthetic estimator has a common regression coefficient for all the areas that are then measured by using data from all the areas. One problem with FH models is to determine the values of variances of direct estimators. In the case of small sample sizes in some of the areas, the estimates of these variances lack precision. To solve this problem generalized variance function or nonparametric estimation of variances can be used for smoothening (González-Manteiga, Lombarda, Molina, Morales, & Santamaría, 2010). But still, the problem is to incorporate the error of estimation of these variances into the error of the final estimator.

Area-level models improve upon direct survey estimators by assuming a relationship between the small area parameters of interest and covariates via hierarchical models.

They have the advantage that only area-level aggregate summaries of the covariates from administrative records are needed. Some early applications of area-level models in small-area estimation can be found in (G. M. Carter & Rolph, 1974; Efron & Morris, 1973, 1975). These authors used area-level covariates implicitly in forming groups of similar small areas. However, they did not explicitly use any area-level covariate in modeling and did not use complex survey data. Fay III and Herriot (1979) considered the generalization of the (Efron & Morris, 1973, 1975) and (G. M. Carter & Rolph, 1974) Bayesian models. The Fay-Herriot model is a special case of a multilevel or hierarchical model where the first level captures the errors of the direct survey estimates of the target variable due to sampling and the second level, often called the linking model, links the true small area parameters to a set of auxiliary variables. In practice, the sampling variance must be estimated, typically using survey data. The Fay-Herriot model is a type of general linear mixed model where the model or linking errors and sampling errors are independent (Rao & Molina, 2015). Over the years, other approaches to MSE estimation have appeared, some quite appealing as well as elegant. The two most prominent ones appear to be the ones due to Jackknife and Bootstrap. K. Das, Jiang, and Rao (2004); Jiang and Lahiri (2001); Jiang, Lahiri, and Wan (2002); D. Lahiri et al. (2002) all considered the Jackknife estimation of the MSE that avoids the detailed Taylor series expansion of the MSE. A detailed discussion paper covering many aspects of related methods appears in (Jiang & Lahiri, 2006). Butar and Lahiri (2003); Pfeffermann and Tiller (2005) considered bootstrap estimation of the MSE. Booth and Hobert (1998) introduced a conditional approach for estimating the MSE. In a different vein, P. Lahiri and Rao (1995) dispensed with the normality assumption

of the random effects, assuming its eighth moment in the Fay-Herriot model. Pfeffermann and Correa (2012) proposed an approach that they showed to perform much better than the "classical" jackknife and bootstrap methods. Pfeffermann and Ben-Hur (2019) used a similar approach for estimating the design-based MSE of model-based predictors. The Best Predictor (BP) is obtained by minimizing the mean squared prediction error (MSPE) given that parameters are known (Li & Lahiri, 2010). Casas-Cordero Valencia, Encina, and Lahiri (2016) using the FH model demonstrates how to use the empirical best/Bayes method to obtain poverty estimates for Chilean communas. The extension of the FH model and how to obtain small sample variance estimation in the Bayesian framework of SAE are also discussed. Li and Lahiri (2010) used the ML method and showed that the method can produce strictly positive estimates of variance components and at the same time achieve the properties of the ML method. They also demonstrated how to apply the parametric bootstrap method to obtain highly efficient confidence intervals for small-area means. Hirose and Lahiri (2018) showed that by using the maximum likelihood method how to estimate shrinkage parameters so that the bias is small. They also showed that MSE estimation can be done without bias correction, if an adjusted ML method is used. Hirose and Lahiri (2018) showed how to obtain prior, or Bayesian estimates such that the Bayesian method is similar to the classical prediction method in the asymptotic sense.

2.2.2 Unit-Level Models

The unit-level models typically relate the values of the variable of interest at the unit level to the unit-specific auxiliary variables. Battese, Harter, and Fuller (1988) proposed this type of model, also known as the nested error model. The model is linear

and incorporates random effects associated with areas, in addition to the individual model errors. The area effects are particularly used to capture heterogeneity among the areas that are not being explained by the auxiliary variables. These models are superior in the sense that they incorporate more information (by considering the whole sample size "n") than area-level models.

Elbers et al. (2003) built a method known as the ELL method (based on unit level) that is used for the estimation of indicators of interest. The ELL method is the first of its type that is designed to measure the more complex indicators of interest than the average or total if they are a function of the welfare variable. ELL estimator can be biased, if the considered auxiliary variables cannot fully explain the between-area heterogeneity of the response variable. In addition to this, in the ELL method, under the bootstrap procedure, the model cannot be fitted again, and indicators cannot be re-estimated with each bootstrap sample that should be drawn from the bootstrap censuses to replicate the real-world phenomenon (Molina & Rao, 2010). As a result, the real-world process cannot be replicated in the ELL bootstrap procedure. Therefore, the MSE estimated through this method cannot accurately reproduce the error incurred in the real world. Also, in the original ELL method, the random effects included in the model are meant for clusters (first stage/ primary sample units) rather than for areas of interest which may lead to the understatement of the MSE estimate (S. Das & Chambers, 2017), if available auxiliary variables cannot fully explain the between-area heterogeneity.

A potential challenge in implementing the original ELL method often arises from the asymmetry of the variance-covariance matrix, caused by unequal weights assigned to clusters. To address this, Stephen Haslett, Jones, Noble, and Ballas (2010) recommend

averaging the matrix with its transpose. This procedure effectively symmetrizes the matrix, enhancing the robustness and reliability of the resulting estimates. Furthermore, the integration of survey weights within the variance-covariance matrix is refined by drawing upon the advancements put forth by (R. Huang & Hidiroglou, 2003) and elaborated upon by (Van der Weide, 2014). This refinement acknowledges that survey weights are not merely scaling factors but integral components of the estimation process, particularly when dealing with complex survey designs involving stratification, clustering, and unequal selection probabilities. By explicitly incorporating survey weights into the variance-covariance structure, the estimation procedure more accurately reflects the sampling design and yields more precise estimates of the model parameters and their associated standard errors. In addition to incorporating GLS and Henderson's Method III into the ELL framework, Empirical Bayes (EB) prediction has been introduced to enhance small-area estimation. EB prediction leverages survey data to refine predictions of location effects, thereby improving the accuracy of estimates for areas represented in the survey. This approach capitalizes on the availability of both survey and census data by using survey information to inform and adjust predictions derived from the census. Under the assumption of normality for both the location effects and the unit-level errors, the conditional distribution of location effects given the unit-level errors, also follows a normal distribution (Van der Weide, 2014).

Molina and Rao (2010) expanded upon the ELL method by introducing the EBP method. While ELL focuses on predicting individual-level outcomes, EBP directly estimates small-area means, incorporating both fixed and random effects to account for

variability within and between small areas. This approach has proven to be more flexible, accommodating complex survey designs and offering direct estimates of policy-relevant parameters. SJ Haslett and Jones (2010) demonstrated the practical utility of the ELL method by applying it to poverty mapping in the UK. Their study highlighted the potential of this method to inform policy decisions at the local level by providing reliable poverty estimates for small areas. The versatility of the ELL method was further showcased by (Tarozzi & Deaton, 2009), who employed it to evaluate the impact of a cash transfer program on poverty in South Africa. By utilizing household survey data and administrative records, they were able to estimate the counterfactual consumption of households that did not receive the transfer, demonstrating the ELL method's effectiveness in impact evaluation. Recognizing the challenges posed by small sample sizes, Nguyen et al. (2018) proposed a bootstrap-based extension of the ELL method. This extension involves resampling the data to generate multiple estimates, enhancing the stability and reliability of the estimates, especially in situations with limited data.

While the ELL method has been widely adopted, researchers have also critically examined its strengths and weaknesses. Corral et al. (2021) conducted a comprehensive comparative study of the ELL and EBP methods, providing valuable insights into their performance under different conditions. Their work helps guide researchers in selecting the most appropriate approach for their specific research questions and data constraints. Furthermore, Rao and Molina (2015) offered a comprehensive review of SAE methods, including the ELL method, discussing its theoretical underpinnings and practical applications. This review serves as a valuable resource for understanding the broader

landscape of SAE techniques and their relevance in various fields. They delved into the limitations of the ELL method regarding data disaggregation. They also highlighted the challenges of estimating poverty and inequality for subpopulations not explicitly defined in the survey data, raising important questions about the applicability of the method in complex settings and stimulating further research in this area.

Christiaensen, Lanjouw, Luoto, and Stifel (2012) conducted a comprehensive comparison of various SAE methods, including ELL, for estimating poverty in Mozambique. Utilizing both survey and census data, they assessed the performance of various methods in terms of accuracy and precision. Their findings revealed that the ELL method demonstrated strong performance, particularly in resource-constrained settings, making it a valuable tool for poverty estimation in developing countries. A comparative analysis of the ELL and EBP methods is conducted by Souza (2015) using the 2010 Census data of Minas Gerais, Brazil, with a focus on the target variable of per capita household income. A simulation study was undertaken, wherein 400 samples were randomly selected from the known population, based on the 2008-2009 Consumer Expenditure Survey (González-Manteiga et al., 2010). The ELL and EBP methods are then applied to estimate poverty incidence and poverty gap at the municipal level. Notably, only 20% of municipalities (195 out of 853) were represented in the survey dataset. The results indicated that the ELL estimator outperformed the EBP estimator in terms of both relative bias and relative root mean squared error (RRMSE). Although both estimators provided synthetic estimates for out-of-sample areas, the EBP estimator exhibited higher overestimation compared to the ELL estimator. A study focused on poverty mapping in the Philippines, Fujii (2014) employed the ELL method

to generate poverty estimates for small administrative units (municipalities). These estimates were then utilized to create visual poverty maps, highlighting the spatial distribution of poverty within the country. This approach provides valuable insights for policymakers and program designers to identify high-poverty areas and tailor interventions accordingly. Tarozzi, Desai, and Johnson (2015) utilized the ELL method within a broader methodological framework to evaluate the impact of microcredit programs on poverty reduction in Ethiopia. By estimating the counterfactual consumption of households not participating in microcredit, Tarozzi assessed the causal impact of the program, showcasing the versatility of the ELL method for impact evaluation studies.

The ELL method's widespread adoption has led to the development of various tools and extensions, such as the World Bank's PovMap software (Nguyen et al., 2018), which has been extensively used for poverty mapping in developing countries. Additionally, researchers have explored robust extensions of the ELL method to address model misspecification and outliers, as well as spatial extensions to incorporate geographic heterogeneity. Van der Weide (2014) work is astonishing in this respect. Molina and Rao's groundbreaking 2010 paper introduced the EBP method as a novel approach to address the challenges of small area estimation. They leveraged the nested error regression model, a common framework in small area estimation, to derive EBP estimators. Their methodology ingeniously combined survey data with auxiliary information to enhance the accuracy of estimates for areas with limited sample sizes. The authors demonstrated the effectiveness of EBP through both theoretical analysis and empirical applications, particularly focusing on poverty mapping. This seminal

work laid the foundation for subsequent research on EBP. Rao and Molina (2015) offered a comprehensive and authoritative treatment of the subject. They not only provided a detailed exposition of the EBP method but also discussed a wide array of other small-area estimation techniques. The book delved into theoretical underpinnings, practical considerations, and various applications across different domains. Marhuenda, Molina, and Morales (2013) recognized the potential limitations of assuming normality in the nested error regression model. They proposed a nonparametric EBP approach where they relaxed the normality assumption, thereby extending the applicability of EBP to situations where the error distribution might deviate from the normal distribution. This nonparametric EBP variant enhanced the flexibility and robustness of the methodology, making it suitable for a wider range of data scenarios. Molina, Nandram, and Rao (2014) enriched the EBP framework by introducing a hierarchical Bayes extension. By incorporating prior information and adopting a Bayesian perspective, their approach allowed for the simultaneous estimation of model parameters and small-area means. This hierarchical Bayes EBP method not only improved the precision of estimates but also facilitated the quantification of uncertainty through posterior distributions. Correa, Molina, and Rao (2012) extended the conventional Empirical Best (EB) predictor to develop the "census EB" predictor, addressing a key limitation in practical applications. While the standard EB predictor assumes that the survey data is a subset of the census and individual survey units can be matched to their census counterparts, this linkage is not always feasible. The census EB predictor overcomes this constraint by assuming that the survey and census data are independent but jointly follow the nested error model. This

allows for the estimation of small-area parameters even when direct linkage between the survey and census datasets is not possible, significantly broadening the applicability of the EB prediction approach.

Corral et al. (2021) introduced model fit with GLS to account for weights and heteroskedasticity in the census EB estimates. They also conducted rigorous simulation studies, comparing EBP with the well-established Elbers, Lanjouw, and Lanjouw (ELL) method for small area estimation. Their findings consistently favored EBP in terms of mean squared error, highlighting its superior performance. The researchers further solidified the empirical evidence for the effectiveness of EBP by applying both methods to real-world poverty data. These comparative studies underscored the practical relevance and advantages of EBP over alternative techniques.

Organizations like the World Bank and various national statistical offices refine poverty estimates in developing countries with limited survey data. Bui and Nguyen (2017), for instance, leveraged EBP to estimate poverty rates within Vietnam's provinces, contributing to more targeted poverty reduction strategies. EBP has enabled researchers to delve into disparities at a granular level. They utilized EBP to assess geographic variations in infant mortality rates across the United States, identifying areas requiring focused interventions. The EBP method has proven instrumental in monitoring environmental indicators such as air and water quality. Holan and Wikle (2016) employed a spatial-temporal EBP model to estimate fine particulate matter concentrations throughout the continental United States, informing air quality management strategies. EBP has found applications in assessing educational outcomes in small areas like school districts or neighborhoods.

2.3 Miscellaneous Contributions and Conclusion

This framework offers a structured approach to navigating the complexities of small-area estimation. By emphasizing user needs, data availability, and methodological considerations, it ensures that the resulting estimates are tailored to the specific requirements of the stakeholders and are produced transparently and rigorously.

Beyond the widely recognized ELL and EBP methods, the landscape of unit-level SAE encompasses a rich array of alternative approaches that contributed a lot to the unit-level model literature. One such method is the unit-level Empirical Best Linear Unbiased Predictor (EBLUP), introduced by (Battese et al., 1988). The Battese, Harter, and Fuller (BHF) model extends the classical Best Linear Unbiased Predictor (BLUP) by incorporating unit-level auxiliary information, thereby enhancing the precision of small-area mean predictions. The BHF model is the unit-level Nested Error Regression Model (NERM) . The NERM's ability to model both between-area and within-area variability makes it particularly well-suited for scenarios where heterogeneity within small areas is substantial. Similarly, Fay III and Herriot (1979), initially introduced an area-level model but they also proposed that the same model can be adopted for unit-level analysis by incorporating unit-level covariates. This adoption combines the simplicity of the FHM with the advantage of utilizing individual-level information for improved small-area mean estimations. For scenarios involving non-normal response variables, such as binary or count data, Jiang and Lahiri (2006) provided a comprehensive review of the unit-level Generalized Linear Mixed Models (GLMM) framework, which offers a flexible solution in SAE. This approach accommodates various types of outcomes and allows for the modeling of complex relationships. In

cases where small areas exhibit spatial dependence, unit-level spatial models, Cressie (1993) in his seminal work on spatial statistics, incorporate spatial correlation structures into the model, and methods to improve prediction accuracy when spatial patterns are present. Opsomer, Claeskens, Ranalli, Kauermann, and Breidt (2008) introduced unit-level nonparametric models, providing a flexible and data-driven alternative to traditional parametric models. These models relax assumptions of linearity and normality, making them robust to model misspecification and capable of capturing complex, non-linear relationships. Chambers and Tzavidis (2006) introduced the M-quantile method, a robust alternative to traditional mean-based SAE methods, proving valuable when dealing with data contaminated by outliers. This method minimizes a specific loss function less sensitive to outliers, proving particularly valuable when dealing with data contamination. There are also other methods used to measure poverty at a small area level including the M-quantile (Chambers & Tzavidis, 2006), and the Hierarchical Bayes (HB) method (Molina et al., 2014). However, this study focuses on the ELL, extended ELL, and (Census) EB methods that are particularly used by the World Bank for the measurement of welfare or poverty at the small area level.

2.4 Research Gap

The literature review reveals several critical gaps in the current body of knowledge regarding SAE techniques and their application to welfare indicators in developing countries, particularly in Pakistan. The literature lacks studies that bridge the gap between advanced statistical methodologies and practical policy formulation, particularly in leveraging granular poverty and food insecurity estimates for targeted

interventions in developing countries. While various SAE methods have been developed and applied individually, there is a dearth of comprehensive studies comparing the effectiveness of multiple contemporary techniques, specifically the ELL, Extended ELL, and Census EBP methods, in estimating poverty and food insecurity at granular levels. The application of advanced SAE techniques to Pakistan's socioeconomic landscape remains limited. Previous studies by Jamal (2007) and Begum (2015) employed SAE for district-level poverty estimation in Pakistan, but these efforts are now outdated and utilize less sophisticated methodologies. There is a notable absence of research simultaneously addressing both poverty and food insecurity estimation at the district level in Pakistan using state-of-the-art SAE techniques. This study aims to address these gaps by providing a comprehensive, methodologically rigorous analysis of poverty and food insecurity at the district level in Pakistan. The empirical results and associated confidence levels of this study are a direct outcome of a significantly more robust methodological framework than those employed in earlier works. While previous research in Pakistan relied on simpler OLS regression models, this study incorporates a rigorous suite of diagnostic and model selection criteria. The process of generating more reliable welfare estimates includes the use of advanced criteria such as stepwise forward and backward selection, Lasso, and information criteria (BIC/AIC) to identify the optimal model.

The application of sophisticated statistical models, including the Beta Model, the Alpha Model, the GLS Model, and the REML Model, which are designed to produce more accurate and stable results. A detailed diagnostic analysis is also conducted to ensure that all models meet the necessary statistical assumptions, thereby validating the

integrity of the welfare estimates. Furthermore, Bootstrapping and Monte Carlo simulations are utilized to generate robust and more reliable welfare estimates by providing a deeper understanding of the sampling distribution and prediction uncertainty. Ultimately, the robustness of these estimates is quantified through a comparative analysis where MSE serves as the final and conclusive metric for assessing model performance. This comparative approach of multiple SAE methodologies is a core component of this study and provides valuable insights that were not available in earlier works. By comparing advanced SAE techniques, exploring their policy implications, and linking the findings to rural development indicators, this research seeks to contribute significantly to both the theoretical understanding of SAE methodologies and their practical application in developing country contexts.

CHAPTER 3

RESEARCH METHODOLOGY

The research methodology undertakes a rigorous exploration of welfare impact assessment, focusing on poverty and food insecurity, by applying advanced small-area estimation (SAE) techniques. Informed by a comprehensive review of extant literature, we highlight the prominent unit-level models—ELL method (Elbers et al., 2003), extended ELL method (Nguyen et al., 2018) and the Census Empirical Best (EB) predictor method (Molina et al., 2015)—extensively utilized by the World Bank for granular poverty and income estimation. This study not only delves into the theoretical underpinnings of these models in terms of methodological procedures but also examines their contemporary advancements and refinements. By meticulously evaluating these unit-level models, including their most recent upgrades, we aim to illuminate their strengths, limitations, and applicability in the milieu of measuring welfare impacts, thereby contributing to a more delicate and up-to-date understanding of poverty and food insecurity dynamics within the target population.

3.1 ELL Method (Elbers et al., 2003)

The ELL pioneered a revolutionary method in 2003 for estimating indicators of interest, aptly named the ELL method. 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 both the ELL (2003) framework and Henderson's Method III (H3). The culmination of this stage is the derivation of GLS estimates for the regression coefficients (β), accounting for

the inherent correlation structure within the data. 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. This stage encompasses two key components, Empirical Bayes (EB) Prediction by assuming normality of the random effects[7], and EB prediction is employed to obtain posterior estimates of the area-level parameters. This approach leverages both the prior information from the model and the observed data to derive more precise and stable estimates, particularly for areas with limited sample sizes. Secondly, Monte Carlo Simulations are used to quantify the uncertainty associated with the small area estimates. The Monte Carlo simulations are conducted by repeatedly drawing samples from the estimated distributions of the model parameters, and a range of plausible values for the small area estimates is generated. This enables the construction of confidence intervals and facilitates a more comprehensive understanding of the estimation's precision. These simulation-based techniques in the second stage augment the ELL methodology's capability to produce robust and informative small-area estimates, even in the presence of data limitations at the local level.

3.1.1 First Stage – Modeling:

To model household per capita expenditure Y_{di}, we initially construct a robust empirical model also known as the beta model by applying ordinary least squares

[7] EB estimates are not employed in stage 2 when GLS estimation is done via the ELL (2003) Method in stage 1.

regression to the log-transformed per capita expenditure l_nY_{di}. The OLS regression model takes the form:

$$l_nY_{di} = X_{di}\beta + u_{di} \quad (3.1)$$

Where $Y_{di} = E_{di} + c\ for\ c > 0$, c is constant and E_{di} is the welfare variable (income/expenditure) for i^{th} individual household in a cluster/area d and N_d is the size of the population in area d where D represents the partitioned of population size N, $i = 1, \ldots., N_d$ and $d = 1, \ldots., D$. X_{di} is the vector of explanatory variables for household i in cluster d encompassing household and individual characteristics such as demographics, household head characteristics, household characteristics, dwelling characteristics, and possession of household assets. 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}, within the household consumption model can be decomposed into two independent components, as elucidated by (Elbers et al., 2002; Elbers et al., 2003): a cluster-specific effect (η_d) and a household-specific effect (e_{di}).

$$u_{di} = \eta_d + e_{di} \quad (3.2)$$

This decomposition represents the inherent correlation among households residing within the same cluster, a common characteristic of household survey designs. The cluster effect (η_d) captures unobserved factors shared by households within a particular geographical or social cluster, while the household-specific effect (e_{di}) encapsulates idiosyncratic variations in consumption patterns at the household level.

The cluster or location-specific effect (η_d) is the weighted average of the individual household errors within a specific cluster. This formulation acknowledges the interdependence of household consumption decisions within a shared environment and accounts for the potential impact of unobserved cluster-level factors on individual household expenditure.

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

Additionally, the household-specific effect (e_{di}) is assumed to be heteroskedastic, meaning its variance can differ across households. This allows for the possibility that some households may exhibit greater variability in their consumption patterns than others, due to individual preferences, income fluctuations, or other unobserved idiosyncratic factors.

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

This decomposition of the error term, as delineated in equation (3.2), aligns with the methodological framework proposed by (SJ Haslett & Jones, 2010), enabling the incorporation of clustering effects and heteroskedasticity into the estimation of household consumption models. Consequently, the resulting model to be estimated is a linear mixed model, as (Van der Weide, 2014) outlined. Although Molina and Rao (2010) refer to this as a nested error linear regression model, the underlying structure remains consistent.

Following the ELL (2002 and 2003) methodology, the estimation of the unconditional variance for each location is obtained by estimating the 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:

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

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

This decomposition of the residual into a cluster-level component ($\hat{\eta}_d$) and a household-specific component ($\hat{e}_{di}$) aligns with the nested error structure of the model, acknowledging the inherent correlation among households within the same cluster. The subsequent steps involve estimating the variances of these two components, which are crucial for obtaining reliable small-area estimates of household consumption.

The unconditional variance of the location effect ($\widehat{\sigma}_{\eta}^2$), representing the between-cluster heterogeneity, is estimated using the following formula:

$$\widehat{\sigma}_{\eta}^2 = \max\left(\frac{\left(\Sigma_d \omega_d (u_{d.} - u_{..})^2 - \Sigma_d \omega_d (1 - \omega_d)\, \hat{\tau}_d^2\right)}{\Sigma_d \omega_d (1 - \omega_d)} \; ; \; 0\right) \qquad (3.3)$$

Where $\widehat{\sigma}_{\eta}^2$ is the unconditional variance of the location effect $\hat{\eta}_d$, Σ_d is the summation of overall clusters d, ω_d is the weight of the cluster d (reflecting its relative size or importance), $u_{d.}$ 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, $\hat{\tau}_d^2$ quantifies the variability of household expenditures within each cluster.

Following the estimation of the unconditional variance of the location effect, we proceeded to model the heteroskedasticity inherent in the household-specific error term. Adopting the approach proposed by (Elbers et al., 2003), we specified a

parametric form for the variance of the idiosyncratic error ($\sigma^2_{e_{di}}$), utilizing a logistic function to ensure that the predicted variance remains bounded and non-negative.

$$\sigma^2_{e_{di}} = \left[\frac{Aexp^{z'_{di}\alpha}+B}{1+ exp^{z'_{di}\alpha}}\right] \qquad (3.4)$$

Where A is an upper bound for the variance. It ensures the variance does not get too large. 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 will be estimated from the data. These parameters control how the variance changes depending on the household characteristics in z'_{di}.

In line with ELL (2003), we employed a simplified version of this logistic function, setting $B = 0$, and $A = 1.05\, max(e^2_{di})$. This simplified form is then estimated using Ordinary Least Squares regression, with the log-transformed squared residuals as the dependent variable and a set of household-level explanatory variables (z'_{di}) as the independent variables[8].

$$\ln\left[\frac{e^2_{di}}{A - e^2_{di}}\right] = z'_{di}\alpha + r_{di} \qquad (3.5)$$

This approach, while reminiscent of Harvey's (1976) work on heteroskedasticity, offers the advantage of ensuring bounded variance predictions, thereby enhancing the robustness of the subsequent small-area estimations.

[8] This modeling approach is also known as the Alpha Model.

To obtain a household-specific conditional variance estimator for the idiosyncratic error term (e_{di}), we follow the approach outlined in (Elbers et al. 2003). By defining $\exp(z'_{di}\alpha)$ as D and employing the delta method, which stems from a second-order Taylor expansion for the expected value of the variance, the following estimator is derived:

$$\hat{\sigma}^2_{e_{di}} \approx \left[\frac{AD}{1+D}\right] + \frac{1}{2} Var(r) \left[\frac{AD(1-D)}{(1+D)^3}\right] \qquad (3.6)$$

Where $\hat{\sigma}^2_{e_{di}}$ represents the estimated variance of the idiosyncratic error for household i in cluster d and $Var(r)$ is the estimated variance from the model's residuals in equation (3.5). This estimator effectively combines the logistic function (represented by the terms involving A and D) with an adjustment term involving $Var(r)$. The logistic component ensures that the estimated variance remains bounded within the interval $[0, A]$, while the adjustment term refines the estimate based on the variability observed in the residuals of the alpha model.

To quantify the uncertainty associated with the estimated unconditional variance of the location effect ($\widehat{\sigma}^2_{\eta}$), the ELL (2002) proposed two methods. The first one considers simulations that involve:

Using a simulation-based approach ELL (2002), the following steps are undertaken:

(a) The variance component σ^2_{η} is estimated using equation (3.3), yielding an estimate of $\widehat{\sigma}^2_{\eta}$.

(b) The variance component $\sigma^2_{e,di}$ is estimated using equation (3.6), yielding an estimate of $\hat{\sigma}^2_{e,di}$.

(c) Assuming that η_d and e_{di} are independent and normally distributed with a mean of 0, new values for u_{di} are generated using equation (3.2).

(d) A new estimate of σ_η^2 is computed using equation (3.3).

(e) Steps 1-4 are repeated, and all simulated values of σ_η^2 are retained.

From the simulated values of σ_η^2, the sampling variance of $\hat{\sigma}_\eta^2$, denoted as $Var(\hat{\sigma}_\eta^2)$, can be obtained directly.

Under the second method, the sampling variance is calculated using the approximation technique provided by ELL (2002), as appended below:

$$Var(\hat{\sigma}_\eta^2) = \sum_d 2\left\{a_d^2\left[(\hat{\sigma}_\eta^2)^2 + (\hat{\tau}_d^2)^2 + 2\hat{\sigma}_\eta^2\hat{\tau}_d^2\right] + b_d^2\frac{(\hat{\tau}_d^2)^2}{n_d-1}\right\} \qquad (3.7)$$

Where $a_d = \frac{\omega_d}{\sum_d \omega_d(1-\omega_d)}$ and $b_d = \frac{\omega_d(1-\omega_d)}{\sum_d \omega_d(1-\omega_d)}$. This formula in (3.7) accounts for both the variability in estimating the between-cluster variance ($\hat{\sigma}_\eta^2$) and the within-cluster variance ($\hat{\tau}_d^2$), as well as the uncertainty arising from the finite sample size within each cluster. The resulting sampling variance allows for the construction of confidence intervals for σ_η^2, enabling assessment of the statistical significance of the observed between-cluster heterogeneity in the variable of interest.

Having estimated the variances associated with both the location and household effects, the subsequent step entails the estimation of the GLS estimator. The ELL methodology originally proposed a GLS estimator that leveraged the estimated variance components to construct a variance-covariance matrix (Ω) accounting for both heteroskedasticity and cluster-level correlation. In the variance-covariance matrix,

the off-diagonal elements represent the shared variance component due to the location effect ($\widehat{\sigma}_{\eta}^{2}$), while the diagonal elements capture the total variance for each household, comprising both the location effect and the household-specific error variance ($\widehat{\sigma}_{\eta}^{2} + \hat{\sigma}_{e_{di}}^{2}$).

The variance-covariance matrix of the random effects for the entire dataset, denoted as $\hat{\Omega}$, where along the diagonal represents the variance-covariance matrix specific to a particular cluster/small area, accommodating potential heterogeneity in correlation structures across areas. The off-diagonal blocks, filled with zeros, underscore the assumption of independence between different clusters. However, a major demerit of this approach is the asymmetry arising in the variance-covariance matrix when weights differ across observations. This traditional GLS estimator utilized by (Elbers et al., 2003) is also used in our analysis to compare the results.

We employ GLS estimation to address the potential heteroscedasticity and correlation among the error terms in our regression model. The GLS approach leverages the variance-covariance matrix of the errors (Ω) to weight the observations appropriately, leading to more efficient and unbiased estimates of the model parameters.

Following the ELL (2003) framework, the GLS estimates, and their variances are obtained using the following equations:

$$\hat{\beta}_{GLS} = (X'W\Omega^{-1}X)^{-1}X'W\Omega^{-1}Y \tag{3.8}$$

$$Var(\hat{\beta}_{GLS}) = (X'W\Omega^{-1}X)^{-1}(X'W\Omega^{-1}WX)(X'W\Omega^{-1}X)^{-1} \tag{3.9}$$

where W represents the diagonal matrix of sampling weights. It is crucial to recognize that the product $W\Omega^{-1}$ may not always yield an asymmetric matrix. To address the

potential asymmetry in the variance-covariance matrix arising from unequal weights within clusters, the approach recommended by (SJ Haslett & Jones, 2010) can be adopted, which involves averaging the variance-covariance matrix and its transpose. Furthermore, the incorporation of survey weights into the variance-covariance matrix (Ω) is refined, drawing upon the advancements proposed by (R. Huang & Hidiroglou, 2003) and detailed in (Van der Weide, 2014)[9].

These methodological enhancements may ensure a more robust and accurate estimation procedure, effectively addressing the complexities introduced by unequal weighting within a cluster[10] and correlation structures within the data. To facilitate the estimation of the variance components associated with the GLS estimator, it is imperative to first modify the model by centering the dependent variable within each cluster.

The estimation of variance components, crucial for obtaining reliable GLS estimates, involves a series of computations incorporating the transformed dependent and independent variables, along with the sampling weights and cluster-specific weights. This process effectively accounts for the complex survey design and the inherent clustered structure of the data.

The Sum of Squared Errors (SSE), and trace terms (t_2, t_3, and t_4) are utilized to calculate the residual variance ($\hat{\sigma}_e^2$), and the area-level random effects ($\widehat{\sigma}_\eta^2$).

[9] We compare both approaches: GLS parameter estimates using ELL (2003) and Henderson method as elaborated by (Van der Weide, 2014), in our analysis.
[10] In the case of HIES 2018-19, the weights assigned within a cluster are the same.

The estimation of area-level random effects ($\hat{\eta}_d$) is performed according to the methodology outlined by (Van der Weide, 2014). For each cluster 'd', a shrinkage factor $(\gamma_{d,\omega})$ is calculated using the following equation:

$$\gamma_{d,\omega} = \frac{\hat{\sigma}_{\eta}^2}{\hat{\sigma}_{\eta}^2 + \hat{\sigma}_e^2 \left(\frac{\sum_i \omega_{di}}{\sum_i \omega_{di}^2} \right)} \qquad (3.10)$$

This shrinkage factor balances the cluster-specific information with the overall average, taking into account the estimated variances of the area-level random effects $(\hat{\sigma}_{\eta}^2)$ and the residual errors ($\hat{\sigma}_e^2$), as well as the sampling weights (ω_{di}).

Subsequently, the area-level random effect is estimated by combining the cluster-specific weighted average of predicted random effects with the overall average, weighted by the shrinkage factors, as shown below:

$$\hat{\eta}_d = \gamma_{d,\omega} \sum_i \omega_{di} \hat{u}_{di} - \frac{1}{D} \sum_d \gamma_{d,\omega} \left(\sum_i \omega_{di} \hat{u}_{di} \right) \qquad (3.11)$$

Following the estimation of $\hat{\eta}_d$, the individual-level residual errors $\hat{e}_{di}$ are updated using the following equation:

$$\hat{e}_{di} = \hat{u}_{di} - \hat{\eta}_d - \sum_{di} (\hat{u}_{di} - \hat{\eta}_d) \qquad (3.12)$$

This equation subtracts both the estimated area-level random effect and the average deviation of predicted random effects from the original predicted random effect ($\hat{u}_{di}$). Subsequently, these residuals ($\hat{e}_{di}$) are scaled to ensure their variance aligns with the estimated residual variance ($\hat{\sigma}_e^2$), maintaining consistency between the model assumptions and the estimated parameters.

To address the potential heteroscedasticity (unequal variances) in the individual-level residual errors (e_{di}), once again a modeling approach akin to Equation (3.5) is employed. This equation transforms the variance model into a linear form, facilitating the estimation of the parameters α that govern the relationship between the residual variance and a set of explanatory variables (Z_{di}). The additional error term r_{di} accounts for the inherent variability in this transformation. Subsequently, observation-specific variance estimates ($\hat{\sigma}^2_{e_{di}}$) are obtained using the approximation provided by Equation.

A modified GLS estimator for β, as detailed in (Van der Weide, 2014), is utilized, incorporating survey weights into the variance-covariance matrix. This adjustment enhances the robustness of the estimation procedure in the presence of complex survey designs. The variance-covariance matrices $\hat{\Omega}$ and $\hat{V}$ are constructed as block-diagonal matrices. Each block along the diagonal corresponds to the variance-covariance matrix for a specific cluster, allowing for heterogeneity in correlation structures across different areas. The off-diagonal blocks, filled with zeros, maintain the assumption of independence between areas.

The diagonal elements of $\hat{\Omega}_d$ represent the total variance for each observation within the cluster, incorporating both the estimated area-level variance ($\hat{\sigma}^2_{\eta}$), scaled by a factor considering the survey weights, and the observation-specific residual variance ($\hat{\sigma}^2_{e_{di}}$), adjusted for the sampling weight. The off-diagonal elements capture the covariance between observations within the same cluster, stemming from the shared area-level random effect.

The overall variance-covariance matrix $\hat{\Omega}$ for the entire dataset is constructed as a block diagonal matrix, where each block along the diagonal corresponds to the $\hat{\Omega}_d$ matrix for a specific cluster d. This structure reflects the assumption of independence between different areas while allowing for heterogeneity in the correlation structures within each area.

Similarly, the variance-covariance matrix $\hat{V}_d$ for each cluster d, akin to the original ELL methodology is constructed.

Also, the overall $\hat{V}$ matrix for the entire dataset is constructed as a block diagonal matrix with each block corresponding to the $\hat{V}_d$ matrix for a specific cluster d.

Finally, the GLS estimator for β and its variance-covariance matrix is obtained using the following equations:

$$\hat{\beta}_{GLS} = \left(X'\hat{\Omega}^{-1}X\right)^{-1}X'\hat{\Omega}^{-1}Y \tag{3.13}$$

$$Var\left(\hat{\beta}_{GLS}\right) = \left(X'\hat{\Omega}^{-1}X\right)^{-1}\left(X'\hat{\Omega}^{-1}\hat{V}\hat{\Omega}^{-1}X\right)\left(X'\hat{\Omega}^{-1}X\right)^{-1} \tag{3.14}$$

These methodological enhancements, encompassing the modeling of heteroscedasticity, incorporation of survey weights, and utilization of a modified GLS estimator, collectively contribute to a more robust and accurate estimation procedure, effectively addressing the complexities inherent in the data[11].

3.1.2 Second Stage – Montecarlo Simulations

Following the Elbers et al. (2003) framework, Monte Carlo simulations are employed, to derive the expected welfare measures conditional on the first-stage model. This

[11] Again, it depends upon the sampling design and how weights are assigned.

process involves applying the parameter and error estimates obtained from the survey data (HIES) to the census data (PSLM). The objective is to generate a sufficiently large number of simulations to achieve robust and dependable levels of welfare estimation. Empirical Bayes prediction, as elucidated in (Van der Weide, 2014), is also introduced as a precursor to the Monte-Carlo- simulations. This approach leverages both, the prior information from the model, and the observed data to derive posterior estimates of the area-level parameters, assuming normality of the random effects. This technique leverages survey data to refine the predictions of location-specific effects, thereby enhancing the accuracy of estimates for areas included in the survey. Under the assumption of normality for both area level random effects ($\hat{\eta}_d$) and the individual level residuals (e_{di}), the conditional distribution of η_d given e_d is also normally distributed, as indicated by (Van der Weide, 2014). This normality assumption is pivotal in deriving the distribution of the random location effects. To facilitate the EB prediction process, (Van der Weide, 2014) defines the following shrinkage factor:

$$\gamma_{d,\omega} = \frac{\sigma_\eta^2}{\sigma_\eta^2 + \sum_i \omega_{di}^2 \left(\sum_i \omega_{di} \sum_i \frac{\omega_{di}}{\sigma_{e,di}^2} \right)^{-1}} \qquad (3.15)$$

This shrinkage factor balances the information from the cluster-level predicted random effects with the overall average, incorporating the estimated variances and survey weights. When the sample size within an area is large and the individual-level residuals are reliable (i.e., have small variances), the shrinkage factor will be closer to 1, giving more weight to the cluster-specific information. Conversely, when the sample size is small or the residuals are less reliable (i.e., have large variances), the shrinkage factor will be closer to 0, giving more weight to the overall average across all areas.

Subsequently, the expected value (posterior mean) of the location effect, conditional on the residuals of the households within the location, is obtained as:

$$E[(\eta_d|e_d)] = \hat{\eta}_d = \gamma_{d,\omega}\left(\frac{\sum_i\left(\frac{\omega_{d,i}}{\sigma^2_{e_{di}}}\right)e_{di}}{\sum_i\frac{\omega_{d,i}}{\sigma^2_{e_{di}}}}\right) \qquad (3.16)$$

The equation essentially states that the estimated area-level random effect $(\hat{\eta}_d)$ is a weighted combination of two pieces of information, the prior mean of the random effects, and the weighted average of the individual-level residuals within the area. The shrinkage factor $(\gamma_{d,\omega})$ controls the balance between these two sources of information. Additionally, the variance of this estimate is given by:

$$Var[\hat{\eta}_d] = \sigma^2_\eta - \gamma^2_{d,\omega}\left(\sigma^2_\eta + \sum_i\left(\frac{\frac{\omega_{d,i}}{\sigma^2_{e_{di}}}}{\sum_i\frac{\omega_{d,i}}{\sigma^2_{e_{di}}}}\right)^2\sigma^2_{e_{di}}\right) \qquad (3.17)$$

The equation essentially states that the variance of the estimated area-level random effect $(Var[\hat{\eta}_d])$ is equal to the prior variance (σ^2_η) minus a reduction in variance achieved through the EB prediction process. This reduction is influenced by the shrinkage factor $(\gamma^2_{d,\omega})$, the higher the shrinkage factor implies more borrowing of information from other areas leading to a greater reduction in variance. The individual-level residuals and their variances within the cluster contribute to the estimation of $\hat{\eta}_d$ and thus affect the reduction in variance. The sampling weights influence the contribution of each observation to the estimation process, and consequently, the variance reduction.

This simulation process involves applying the parameter and error estimates obtained from the survey data to the census data. The objective is to generate a sufficiently large number of simulations to achieve robust and dependable levels of welfare estimation. The simulation process encompasses the following steps:

Step-1. This step involves drawing simulated values for the regression coefficients (β) from a multivariate normal distribution.

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

The mean of this distribution is the estimated GLS coefficients $\left(\hat{\beta}_{GLS}\right)$ obtained in the first stage, and the variance-covariance matrix is the estimated variance-covariance matrix of the GLS estimator $\left(Var\left(\hat{\beta}_{GLS}\right)\right)$. The mean and variance-covariance matrix used in this step are derived from the survey data analysis in the first stage.

Step-2. Under the ELL variance estimation approach, the cluster-level random effect $(\tilde{\eta}_d)$ is drawn from a normal distribution with mean, 0, and variance $\hat{\sigma}_\eta^2$.

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

The variance of $\hat{\sigma}_\eta^2$ is drawn from a Gamma distribution to account for its uncertainty. $Var\left(\hat{\sigma}_\eta^2\right)$ is estimated by using the equation 3.7[12].

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

[12] It can also be estimated by using simulations.

The location effect $\tilde{\eta}_d$ is drawn from a normal distribution with mean 0 and the overall estimated variance $\hat{\sigma}_\eta^2$. But it can also be drawn directly from the first-stage estimates (semi-parametric approach) $\tilde{\eta}_d \sim N(\eta_d, \hat{\sigma}_\eta^2)$.

Under Henderson's Method III (H3), the distribution for the parameter $\hat{\sigma}_\eta^2$ is not explicitly defined. As a result, bootstrapping the survey data is employed to estimate $\hat{\sigma}_\eta^2$ along with all other relevant parameters for each simulation iteration. This bootstrapping approach is essential, particularly when Empirical Bayes (EB) is utilized. When EB is selected, the bootstrap generates a vector of η_s values and estimates of $\hat{\sigma}_{\eta d}^2$ for each cluster represented in the survey data. Consequently, for each simulation, we sample $\tilde{\eta}_d \sim N(\eta_d, \hat{\sigma}_{\eta d}^2)$ for clusters included in the survey. For those clusters not included, the samples are drawn from $\tilde{\eta}_d \sim N(0, \hat{\sigma}_\eta^2)$.

The individual-level residual errors ($\tilde{e}_{di}$) are drawn from a normal distribution with mean 0 and variance $\tilde{\sigma}_{e,di}^2$. The variance is estimated using Equation 3.6, which involves drawing α from a normal distribution and using it with the census data to obtain a new D. This D, along with the first-stage estimates of A and $\widehat{Var}(r)$, is used to calculate $\tilde{\sigma}_{e,di}^2$.

$$\tilde{\sigma}_{e,di}^2 \approx \left[\frac{A\tilde{D}}{1+\tilde{D}}\right] + \frac{1}{2}\widehat{Var}(r)\left[\frac{A\tilde{D}(1-\tilde{D})}{(1+\tilde{D})^3}\right] \qquad (3.18)$$

This step simulates the individual-level variations that are not explained by the model's fixed effects or the area-level random effects.

Finally, by imputing all the parameters, we can generate a welfare vector as follows:

$$\tilde{Y} = X\tilde{\beta}_{GLS} + \tilde{\eta}_d + \tilde{e}_{di} \qquad (3.19)$$

This equation generates the simulated values of the dependent variable $(\tilde{Y})$ for each observation in the census data. It combines the fixed effects $(X\tilde{\beta}_{GLS})$, the simulated area-level random effect $(\tilde{\eta}_d)$, and the simulated individual-level residual error $\tilde{e}_{di}$. The simulated welfare $(\tilde{Y})$ are used to obtain M simulated estimates of the variable of interest. The mean indicator and its standard error are then calculated for each domain of interest.

In model 3.1, the general parameter $\delta_d = \delta_d(y_d)$ represents as the ELL estimator, and it provides the numerical approximation of the theoretical ELL estimator which is given as the marginal expectation of ${\delta_d}^{ELL} = E(\delta_d)$. Similarly, the same procedure is used to get an estimate of the MSE of the ELL estimator.

Steps involved in replication procedure:

a. By using the residuals of the fitted model in (3.1), random effects η_d^* (for each area $d = 1, \ldots, D$) and errors e_{di}^* (for each unit $i = 1, \ldots, N_d$) are generated.

b. By using the values of auxiliary variables along with the estimator $\hat{\beta}$ of the regression parameter β, bootstrap values of the response variable are generated for all population units, as shown below through the generation process.

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

c. Through the above steps we get the census of the response variable that can be used to estimate the indicator of any kind based on the welfare of individuals. The process of generation is repeated for $m = 1, \ldots, M$ times to get the M full censuses. After that for each single census m, we calculate the indicator of

interest[13] $\delta_d^{*(m)} = \delta_{\mathrm{d}} Y_d^{*(m)}$, where $Y_d^{*(m)} = Y_{d1}^{*(m)}, \ldots., Y_{dN_d}^{*(m)}$ are the values of the response variable in the area d that obtained through bootstrap census m.

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

$$\hat{\delta}_d^{ELL} = \frac{1}{M}\sum_{m=1}^{M} \delta_d^{*(m)} \qquad (3.20)$$

e. To estimate MSE under ELL

$$mse_{ELL}\left(\hat{\delta}_d^{ELL}\right) = \frac{1}{M}\sum_{m=1}^{M}\left(\delta_d^{*(m)} - \hat{\delta}_d^{ELL}\right)^2 \quad (3.21)$$

We may observe from the whole procedure that if areas have large population N_d and we are interested in calculating the ELL estimator of the area mean $\bar{Y}_d$ through averaging $\bar{Y}_d^{*(m)} \approx \bar{\mathrm{x}}_d'\hat{\beta} + u_d^{*(m)}$ across the M censuses, then the average of the random effects of the bootstrap effects $u_d^{*(m)}$ within the bootstrap replicates is $\frac{1}{M}\sum_{a=1}^{M} u_d^{*(m)} \approx E(u_d) = 0$. As a result, the ELL estimator of the area mean $\hat{\bar{Y}}_d^{ELL} = E(\bar{Y}_d)$, turn into the synthetic regression estimator.

$$\hat{\bar{Y}}_d^{ELL} = \bar{\mathrm{x}}_d'\hat{\beta}$$

The main reason behind this is that the marginal expectation $E(\delta_d)$ does not use the sample observations (unlike the conditional expectation) and, therefore, sticks to the prediction obtained through linear regression without considering the area effects. So, the ELL estimator has the same problem, as the synthetic estimators, of not considering the area effects, particularly, the ELL estimator can be biased, if the considered

[13] For the variable of interest measurement, either poverty or food insecurity headcount, we used Foster, Green, and Thorbecke's (1984) FGT poverty measure in its simplest form with $\alpha = 0$.

auxiliary variables cannot fully explain the between-area heterogeneity of the response variable.

3.2 Empirical Best (EB) Method

Molina and Rao (2010) suggested estimating general nonlinear indicators by using the Bayes Predictor (Best Predictor) which is based on the nested error model, as elaborated in the ELL procedure (see model 3.1). However, the location effects are originally defined for the areas of interest. The main difference with the ELL approach is that the EB method of MR (2010) conditions the survey sample data and thus makes more efficient use of the survey data, which contains only available information on actual welfare. Nevertheless, conditioning on the survey data requires areas and households across the survey and census to be matched. In contrast with this procedure, the traditional ELL approach does not require linking the census and survey observations.

3.2.1 The Nested Error Model

EB method assumes that the variable $Y_{di} = E_{di} + c$ adhere to the model (3.1) with normally distributed random effects u_{di}. Random effects are further divided into area-specific random effects η_d, and household-specific random effects e_{di}[14]. These η_d and e_{di} are independent and follow a normal distribution:

$$\eta_d \; ind\sim N\left(0, \sigma_\eta^2\right), \quad \mathrm{e}_{di} \; ind\sim N(0, \sigma_e^2 k_{di}^2)$$

[14] η_d and e_{di} are also, known as location and household-specific idiosyncratic errors.

For household-specific idiosyncratic errors, k_{di} represents possible heteroskedasticity[15], Finally, we have a nested error model:

$$l_n Y_{di} = X'_{di}\beta + \eta_d + e_{di} \qquad (3.22)$$

Under this model, the variable vector $y_d = \left(Y_{d1}, \ldots, Y_{dN_d}\right)'$ for each area $d = 1, \ldots., D$,[16] are independent and verify $y_d \sim N(\mu_d, V_d)$. The mean vector $\mu_d = X_d\beta$, where $X_d = \left(X_{d1}, \ldots., X_{dN_d}\right)'$ and covariance matrix:

$$V_d = \sigma_\eta^2 1_{N_d} 1'_{N_d} + \sigma_e^2 A_d \qquad (3.23)$$

Here $\sigma_\eta^2 1_{N_d} 1'_{N_d}$ is the constant covariances for all the pairs of units belonging to the same area and represents that any unit within the same area has the same area effect. Where $A_d = diag(k_{di}^2)$ with $i = 1, \ldots., N_d$[17] represents the diagonal matrix with heteroskedasticity weights, that are equal to one, in the case of homoscedasticity. It means variances are $\sigma_\eta^2 + \sigma_e^2$, and covariances are σ_η^2, in case of homoscedasticity.

3.2.2 Empirical Best / Bayes Predictor

The target indicator is defined as a generic function of y_d, that is $\delta_d(y_d)$. By decomposing the area vector as $y_d = (y'_{ds}, y'_{dr})'$, where y_{ds} contains the sample observations and y_{dr} the out-of-sample observations. Typically, the non-sampled households are much larger than the sampled households. It may also happen that $n_d = 0$ for some areas, it means the area is not sampled in the survey. The empirical best (EB) predictor for an area d that is sampled $(n_d > 0)$ is defined as the conditional

[15] Molina and Rao (2010) assumed a homoscedastic model that is $k_{di} = 1$ for all i and d.
[16] D represents the total number of locations in which the population is divided.
[17] N_d denotes the total number of households in a location d.

expectation $E[\delta_d(y_d)|y_{ds};\theta]$. For an out-of-sample area ($n_d = 0$), the empirical best (EB) predictor is $E[\delta_d(y_d);\theta]$ which is similar to the ELL estimator. The best predictor is the one that minimizes MSE and is given by:

$$\tilde{\delta}_d^{EB} = E_{y_{dr}}[\delta_d(y_d)|y_{ds};\theta] \qquad (3.24)$$

Here the expectation is taken with respect to y_{dr} (distribution of out-of-sample values) from area d, given y_{ds} (the sample values). This conditional distribution depends on the true value of the model parameters $\theta = \left(\beta', \sigma_\eta^2, \sigma_e^2\right)'$ for y_{ds}.

In the second step, these parameters are estimated by substituting a consistent estimator $\hat{\theta}$ (using REML) in place of θ in the best predictor that is Empirical Bayes/Empirical Best predictor.

$$\hat{\delta}_d^{EB} = \tilde{\delta}_d^{EB}(\hat{\theta})$$

Usually, the REML fitting model built on the normal likelihood function gives consistent estimators even if the normality assumption does not hold under some particular regulatory conditions.

3.2.3 Required Conditional Distribution

Under the nested error model (3.27), the distribution of $y_{dr}|\mathrm{y_{ds}}$ is obtained as follows:

First, we decompose the X_d and V_d matrices within the sample and the out-of-sample parts in the same way as we decomposed y_d, as shown below:

$$y_d = \begin{pmatrix} y_{ds} \\ y_{dr} \end{pmatrix},\ \mathrm{X_d} = \begin{pmatrix} X_{ds} \\ X_{dr} \end{pmatrix},\ V_d = \begin{pmatrix} V_{ds} & V_{drs} \\ V_{drs} & V_{dr} \end{pmatrix}$$

Where subscript s refers to the sample unit and r to the out-of-sample units. The matrix X_d represents auxiliary variables for all the units in the population. Similarly, the variance-covariance matrix V_d incorporates variances within the diagonal, and covariances for the pair of sampled and out of sampled units. As y_d follows normal distribution, then its conditional also follows a normal distribution, particularly,

$$y_{dr}|y_{ds}\ ind{\sim}N(\mu_{dr|s}, V_{dr|s}) \qquad (3.25)$$

Where the conditional mean vector and corresponding covariance matrix get the form as shown in (3.26) and (3.27).

$$\mu_{dr|s} = X_{dr}\beta + \gamma_d(\bar{y}_{da} - \bar{x}_{da}^T\beta)1_{N_d - n_d} \qquad (3.26)$$

The conditional mean $(\mu_{dr|s})$ for non-sampled units is the combination of the regression part $(X_{dr}\beta)$ and the predicted area effect $\left(\gamma_d(\bar{y}_{da} - \bar{x}_{da}^T\beta)\right)$ which is the same for all non-sampled units in the area d. In case of heteroskedasticity, the weighted average of the transformed variable $\bar{y}_{da}$ is given as under:

$$\bar{y}_{da} = \frac{1}{a_{d\cdot}}\sum_{i\epsilon s_d} a_{di}y_{di}$$

Where $a_{d\cdot} = \sum_{i\epsilon s_d} a_{di}$ and $a_{di} = k_{di}^{-2}$. The vector of the weighted average of the auxiliary variables for the units in the sample is given as:

$$\bar{x}_{da} = \frac{1}{a_{d\cdot}}\sum_{i\epsilon s_d} a_{di}x_{di}$$

In homoscedastic case, $\bar{y}_{da}$is the sample mean of the response variable in the area d.

$$\bar{y}_{da} = \frac{1}{n_d}\sum_{i\epsilon s_d} y_{di}$$

While $\bar{x}_{da}$ is a vector of sample means of the auxiliary variables in the area d.

$$\bar{x}_{da} = \frac{1}{n_d}\sum_{i\epsilon s_d} x_{di}$$

As area effect is same for all units within the same area, then the predicted area effect is also the same for all units, that is why the same is multiplied by the vector of ones of the size equal to non-sampled units. Where $\gamma_d = \frac{\sigma_\eta^2}{\sigma_\eta^2 + \frac{\sigma_e^2}{n_d}}$ (in homoscedastic case) is a correction that ranges from 0 to 1. This γ_d is very large if areas are very heterogeneous. This means that we do not have all the covariates that can explain the heterogeneity of the welfare variable across areas. If the sample size is small, more weight is given to the regression part $(X_{dr}\beta)$ because it borrows strength from all the areas due to common β for those areas. The larger the sample size of the area, the more weight is given to the correction expression $\left(\gamma_d(\bar{y}_{da} - \bar{X}_{da}^T\beta)\right)$.

$$V_{dr|s} = \sigma_\eta^2(1-\gamma_d)1_{N_d-n_d}1_{N_d-n_d}^T + \sigma_e^2 diag_{i\in r_d}(k_{di}^2) \qquad (3.27)$$

The conditional Covariance matrix has the same shape as that of V_{ds}. The only difference is σ_η^2 is multiplied by a factor $(1-\gamma_d)$ which is also an interval of either 1 or 0. This means conditioning $y_{dr}|y_{ds}$ reduces the uncertainty with the factor $\sigma_\eta^2(1-\gamma_d)$.

For an individual household where $i \in r_d$, the cumulative distribution is the normal:

$$Y_{di}|y_{ds} \sim N(\mu_{di|s}, \sigma_{di|s}^2)$$

Where the mean and conditional variance are given by:

$$\mu_{di|s} = x'_{di}\beta + \gamma_d(\bar{y}_{da} - \bar{x}_{da}^T\beta)$$

$$\sigma^2_{di|s} = \sigma^2_{\eta}(1-\gamma_d) + \sigma^2_e k^2_{di}$$

3.2.4 Monte-Carlo Approximation of EB Predictor

For the general indicator $\delta_d = \delta_d(y_d)$ with complex shape, it is not possible to calculate the expectation that defines the best predictor analytically, so to resolve this problem, the best predictor can be approximated empirically using Monte Carlo simulation. To approximate the conditional expectation of the EB predictor, $\hat{\delta}_d^{EB} = E y_{di}\left[\delta_d(y_d)|y_{ds};\hat{\theta}\right]$, with respect to the non-sample distribution, given the sample data[18]. The following steps are involved in the process.

a. By using survey data fit the nested error model (3.22) via REML,[19] and obtain the estimator $\hat{\theta} = \left(\hat{\beta}', \sigma^2_{\eta}, \sigma^2_e\right)'$ of the parameter vector $\theta = \left(\beta', \sigma^2_{\eta}, \sigma^2_e\right)'$.

b. By using the parameter estimates obtained in step a as true values, generate $m = 1, \ldots, M$ vectors of the values of response variables for out-of-sample units in area d, $y_{dr}^{(m)}$, from the distribution of $y_{dr}|y_{ds}$. In this step, it is very expensive and even impossible to generate out-of-sample vectors $y_{dr}^{(m)}$ of size $N_d - n_d$ for area d because the distribution of non-sample vectors is very large $(N_d - n_d)$, and this might be very large for real areas. So instead of generating

[18] If this conditional distribution is normal, it only requires generating vectors many times from it.
[19] Molina & Rao (2010) used REML by considering the homoscedastic model, and Corral et al. (2021) used the Henderson III (H3) method to account for heteroskedasticity in the model based on (Van der Weide, 2014) elaboration. The H3 method does not require to specify a distribution (see section 3.1.1 for Henderson III method).

it through multivariate normal, it is better to generate the random vector $y_{dr}^{(m)}$ through modeling. Because the covariance matrix of this vector is the covariance matrix of $y_{dr}^{(m)}$. The welfare from each out-of-sample household is generated as follows:

$$l_n\left(y_{di,r}^*\right) = x_{di,r}\beta + \eta_d^* + e_{di,r}^*$$

The area effect η_d^* for an area d that is included in the sample, is generated as:

$$\eta_d^* \sim N\left(\hat{\eta}_d, \hat{\sigma}_\eta^2(1-\hat{\gamma}_d)\right)$$

Where $\hat{\eta}_d = \hat{\gamma}_d\left(\bar{y}_{d,s} - \bar{x}_{d,s}\hat{\beta}\right)$, and $\hat{\gamma}_d = \frac{\hat{\sigma}_\eta^2}{\hat{\sigma}_\eta^2 + \frac{\hat{\sigma}_e^2}{n_d}}$.

In case if area d is not sampled then the area effect is generated as $\eta_d^* \sim N\left(0, \hat{\sigma}_\eta^2\right)$. The household error $e_{di,r}^*$ is generated as $e_{di,r}^* \sim N(0, \hat{\sigma}_e^2)$.[20]

c. Extend the generated vector $y_{dr}^m$with the sample data y_{ds}, to form a census vector $y_d^{(m)} = \left(y_{ds}', \left(y_{dr}^{(m)}\right)'\right)'$ for the area d. Now by using $y_d^{(m)}$, the indicator of interest[21] is computed $\delta_d^{(m)} = \delta_d\left(y_d^{(m)}\right)$ and repeat the process for $m = 1, \ldots., M$.

d. Finally, the Monte Carlo approximation of the EB predictor of the indicator δ_d is obtained by averaging the indicators for the M simulated censuses.

$$\hat{\delta}_d^{(EB)} = \frac{1}{M}\sum_{m=1}^{M}\delta_d^{(m)} \qquad (3.28)$$

[20] In case of heteroskedasticity, $e_{di,r}^* \sim N(0, \hat{\sigma}_e^2)diag_{i\in r_d}(k_{di}^2)$.

[21] For the variable of interest, either poverty or food insecurity headcount, we used Foster, Green, and (Foster, Greer, & Thorbecke, 2010) FGT poverty measure in its simplest form with $\alpha = 0$.

3.2.5 Bootstrap Mean Squared Error Estimator

For the MSE estimation, it is impossible to calculate the closed-form expression, so typically, the bootstrap procedure is applied. MR (2010) propose the parametric bootstrap method of (González-Manteiga, Lombardía, Molina, Morales, & Santamaría, 2008) which involves the following steps:

Step 1: Using the survey data $y_s = (y'_{1s}, \ldots, y'_{Ds})$, fit the nested error model (3.22) for all areas by using the REML method[22]. This yields the estimates of the model parameters, $\hat{\beta}, \hat{\sigma}^2_{\eta}$ and $\hat{\sigma}^2_e$.

Step 2: Utilizing the model parameter estimates from step 1, simulate the census welfare from the nested error fitted model

$$l_n(y^*_{di}) = x_{di}\hat{\beta} + \eta^*_d + e^*_{di}$$

Where bootstrap area effects are generated as $\eta^{*(b)}_d \; iid\sim N\left(0, \hat{\sigma}^2_{\eta}\right)$, $d = 1, \ldots, D$, and bootstrap model errors are generated $e^{*(b)}_{di} \; iid\sim N(0, \hat{\sigma}^2_e)$, $\; i = 1, \ldots, N_d, \; d = 1, \ldots, D$.

Step 3: Once area effects and random errors are generated, replace them in the model including the value of $\beta's$, the values of the response variable y^*_d are obtained for all the units in the census (A bootstrap census).

Step 4: Once we have the simulated census vector y^*_d, the next step is to calculate the true value of the indicator of interest for each area d.

[22] Molina and Rao (2010) implemented the REML.

$$\delta_d^{*(b)} = \delta_d\left(y_d^{*(b)}\right), \text{ where } y_d^{*(b)} = \left(Y_{d1}^{*(b)}, \ldots, Y_{dN_d}^{*(b)}\right)'$$

Step 5: From the vector $y_s^{*(b)} = \left(\left(Y_{1s}^{*(b)}\right)', \ldots, \left(Y_{Ds}^{*(b)}\right)'\right)'$ with bootstrap observations for the sample units, fit again the nested error model to the bootstrap data $y_s^{*(b)}$ and obtain bootstrap EB predictor of the target indicator, $\hat{\delta}_d^{EB*(b)}$.

Step 6: Repeat steps 2 to 5 for $b = 1, \ldots, B$ for B large times to obtain bootstrap target indicators $\delta_d^{*(b)}$, and corresponding EB predictors $\hat{\delta}_d^{EB*(b)}$.

Step 7: The naive bootstrap MSE estimator of the EB predictor $\hat{\delta}_d^{EB}$ is given by:

$$mse_B\left(\hat{\delta}_d^{EB}\right) = B^{-1}\sum_{b=1}^{B}\left(\hat{\delta}_d^{EB*(b)} - \delta_d^{*(b)}\right)^2$$

3.2.6 Census EB Predictor

The EB predictor assumes that the survey is part of the census and that survey units can be identified in the census file. However, linking the survey and census data is not always possible[23]. A variation of the Empirical Best Predictor is called the census EB predictor which assumes that survey data is not a part of the census but the augmented vector of Census and survey data $(y'_{ds}, y'_d)'$ follows the nested error model. Correa, Molina and Rao (2012) proposed the "census best" predictor which is calculated using the conditional expectation

$$\tilde{\delta}_d^B(\theta) = Ey_{di}[\delta_d(y_d)|y_{ds}; \theta]$$

[23] In this study, HIES represents survey data and PSLM represents administrative records. HIES is not a subset of PSLM.

for all the units in the area d. So, the conditional expectation is predicted without separating the sample units from non-sampled units.

$$\tilde{\delta}_d^{CB}(\theta) = \frac{1}{N_d}\sum_{i=1}^{N_d}\tilde{\delta}_d^{B}(\theta)$$

The steps are the same as the welfare for each household i in the census is generated from the model:

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

Where η_d^* and e_{di}^* are obtained in the same way as in step b. The resulting vector is a full census vector for the area d which is used to compute the variable of interest or target indicator.

While the conventional Census EB method assumes homoscedasticity, (Van der Weide, 2014) provides a robust framework to accommodate heteroscedasticity by employing a GLS model. This approach estimates variance parameters using an extension of Henderson's method III (1953), adapted to incorporate both heteroscedasticity and survey weights (R. Huang & Hidiroglou, 2003). This GLS-based Census EB method yields estimates for areas represented in both the survey and the census[24], if weights and heteroskedasticity are not specified, then the results are very close to the CEB estimates (Corral et al., 2021).

It is important to note that for non-linear area indicators δ_d, both ELL and EB estimators require a census with the unit-level value of the auxiliary variables $(\bar{x}_{d,s})$.

[24] This study employs the Census EB method, deemed suitable in the context of the data set, and HIES and PSLM structure.

But additionally, the EB method requires identifying the survey units in the census of auxiliary variables, in order to identify sample and out-of-sample units. However, Census EB (CEB) is similar to the EB method but it relaxes the assumption of like units in survey and census. EB and CEB methods are considered superior to the ELL method because they address the issues of auxiliary variables related to the variable of interest. It integrates a fitted area-level model for a variable of interest and estimates the area location effect. It also uses different approaches (as discussed above) to calculate uncertainty with more precision as compared to ELL. This study also utilizes the Census EB method for the generation poverty and food-insecurity data. The comprehensive simulation process is discussed as under:

To get the poverty and food insecurity estimates, the survey data is used to fit the nested error model via REML and obtain the estimator $\hat{\theta} = \left(\hat{\beta}', \sigma_{\eta}^2, \sigma_e^2\right)'$. By utilizing the parameter estimates, we generated $m = 1, \ldots, M$ vectors[25] of the values of response variables for PSLM units in the area d, from the distribution of $y_{di}|y_{ds}$. Since we are dealing with PSLM data, there is no distinction between sampled and non-sampled units within an area. Thus, the generation of the response variable vectors for each household in area d is simplified[26]:

$$l_n(y_{di}^*) = x_{di}\hat{\beta} + \eta_d^* + e_{di}^* \quad (3.29)$$

The area effect η_d^* for all areas is generated as $\eta_d^* \sim_{iid} N\left(0, \hat{\sigma}_{\eta}^2\right)$, and household errors are generated as $e_{di,r}^* \sim_{iid} N(0, \hat{\sigma}_e^2)$.[27]

[25] Where M=200, it means we have generated 200 values of response variable for each household in area d.

[26] There are 160,654 household units in PSLM with 5,673 areas/PSUs. Each household is simulated 200 times.

[27] In case of heteroskedasticity, $e_{di,r}^* \sim N(0, \hat{\sigma}_e^2) diag_{i \in r_d}(k_{di}^2)$.

We use the complete data from area d to form the vector $y_d^{(m)}$ based on PSLM. Since we have full data, this vector includes all units in the area without needing to combine sampled and non-sampled data. Then we have calculated the indicator of interest[28] (poverty or food insecurity) directly from the full census data: $\delta_d^{(m)} = \delta_d\left(y_d^{(m)}\right)$ by replicating the results 200 times based on previous steps. Finally, the Monte Carlo approximation of the census EB predictor of the indicator δ_d is obtained by averaging the indicators for the M simulated censuses $\hat{\delta}_d^{(EB)} = \frac{1}{M}\sum_{m=1}^{M}\delta_d^{(m)}$.

The mean squared error is given as under:

$$mse_B\left(\hat{\delta}_d^{EB}\right) = B^{-1}\sum_{b=1}^{B}\left(\hat{\delta}_d^{EB*(b)} - \delta_d^{*(b)}\right)^2 \qquad (3.30)$$

Where B represents the number of bootstrap simulations performed. $\hat{\delta}_d^{EB*(b)}$ is the Monte Carlo estimate for each domain in every stage of the process, obtained from the extracted sample. $\delta_d^{*(b)}$ is the true value of the indicator obtained from the b^{th} simulated census.

For the variable of interest, which is poverty (and food insecurity) in this case, we used Foster, Greer, and Thorbecke's (1984) FGT poverty measure. As we are interested in headcount poverty, so we used the simplest form of FGT poverty indicator with $\alpha = 0$.

$$F_0(y_d^*) = \frac{1}{N_d}\sum_{i=1}^{N_d} I(Y_{di} < z) \qquad (3.31)$$

Here Y_{di} is the per equivalent adult monthly expenditure and z represents per equivalent adult monthly expenditure. If per equivalent adult monthly expenditure is less than the targeted poverty line, then the household is said to be poor and I takes the

[28] FGT poverty indicator formula is used with $\alpha = 0$, represents poverty or food insecurity headcount.

value as 1, otherwise I takes the value as 0. So, the poverty headcount for the area d is the sum of poor households divided by the total number of households in that particular area. Equation (3.31) can be represented as $F_0(y_d^*) = \delta_d(y_d)$ gives us the general indicator of interest. This general indicator may be poverty or any other variable of interest like headcount food insecurity in our study. Along with poverty, we utilized the equation (3.31) for estimating headcount food insecurity wherein Y_{di} in equation (3.31) this time represents "Kilocalories per capita per day" and z represents Minimum Dietary Energy Requirements (MDER) at the household level. If Kilocalories per capita per day are less than the MDER, the household is said to be food insecure and vice versa.

3.3 Research Strategy

The data on poverty and food insecurity will be generated at the district level (by using HIES and PSLM), based on small area unit level estimation technique under the ELL and extended ELL method with two model settings criteria,[29] and the EB method[30] to identify the state of poverty at the administrative level. Also, by using the data generated on poverty, explore the relationship between rural poverty and rural transformation, (2) rural household income, and agriculture credit at the district level.

[29] The ELL method with error decomposition with ELL (2002), and the Extended ELL method with error decomposition using Henderson-III (H3) method along with Empirical Bayes Prediction by (Van der Weide, 2014).

[30] EB method here refers to the Census EB method which is a variation of the EB method (Molina & Rao, 2010).

CHAPTER 4

THE ELL METHOD: BASICS, DIMENSION, AND EXTENSION

The ELL method offered by Elbers et al. (2003) has been widely used for estimating monetary poverty indicators (as discussed in the Chapter 3). For the application of the traditional ELL method, we required at least two databases typically, a household survey (HIES) and a population census (PSLM) with common variables. We consider only those questions that are defined in a similar way in both data sources with like distributions. In addition, we use PSU as a location variable to link the PSLM and HIES at the cluster level to explain the location residual, as well as between area variability. After the identification of comparable variables in two datasets (e.g., HIES and PSLM) representative at different administrative levels, we estimate an imputation model (predicting household welfare) from HIES data using regression methods and then impute household consumption expenditure/income into each PSLM household to estimate poverty and food insecurity headcount ratio along with standard errors for small areas using imputed household welfare of each household.

While generating the poverty and food insecurity data, we divide the whole process into 03 major steps:

Step-1. Processing HIES 2018-19 and PSLM 2019-20.

Step-2. Modeling Consumption & Simulation.

Step-3. Diagnostics, Validation, and Visualization of Results.

4.1 Processing HIES and PSLM

The data preparation stage is concerned with identifying the common variables that exist between the HIES and the PSLM. These variables form the "bridge" that allows us to predict the consumption levels in the PSLM. For the generation of poverty and food insecurity data, these linking variables are identically defined in the two datasets.

Welfare Variables Creation: The first step towards the application of the ELL method is to create a variable of interest (poverty or food insecurity) by using HIES/survey data.

Definition Matching: We create common variables in PSLM and HIES with nearly the same questions, definitions of variables and categories, etc.

Statistical Matching: Then we ensure that statistical properties (usually means and variances) of the common variables are the same between PSLM and HIES data.

Location Matching: We ensure the matching of PSU variable at the cluster level within PSLM and HIES (consistency of location codes).

4.1.1 Estimation of Poverty

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 by using HIES 2018-19 as proposed by the Planning Commission of Pakistan. The Pakistan Household Integrated Economic Survey (HIES) 2018-19 serves as a crucial instrument for measuring poverty at both national and provincial levels in Pakistan. This comprehensive survey provides a wealth of data on household income and consumption patterns, enabling a nuanced understanding of poverty dynamics across the country. To measure poverty using the HIES 2018-19, a multi-faceted

approach is employed. Firstly, the survey data is used to construct an aggregate consumption measure that encompasses all household expenditures on food and non-food items, harmonized into a consistent time frame, typically monthly. Initially, the monetary value of all expenditure categories within the HIES 2018-19 dataset is meticulously aggregated for each item, consolidating diverse expenses into a cohesive framework. Subsequently, to ensure temporal consistency and facilitate meaningful comparisons, these item-level expenditures are standardized to a uniform time unit. In the context of Pakistan, the aggregate consumption measure encompasses not only food expenditures but also a wide array of non-food essentials. These key non-food components include expenditures on clothing and footwear, housing (including rental value), education, healthcare, fuel and utilities, transportation, recreational activities, and communication services. In the estimation of the consumption aggregate, it is important to note that certain expenditures are excluded. Specifically, expenses related to house and property taxes, fees, infrequent repairs, and maintenance are not factored into the calculation. Additionally, expenditures on durable goods, which are typically characterized by their long lifespan and infrequent purchase, are also omitted from the consumption aggregate. This deliberate exclusion ensures that the aggregate accurately reflects recurring consumption patterns rather than infrequent or one-time expenses, thereby providing a more reliable basis for poverty analysis and comparison. The HIES 2018-19 survey collects expenditure information from households across Pakistan at various frequencies: fortnightly, monthly, and yearly. To establish a standardized basis for comparison and analysis, these expenditures are converted into a common unit, aligning with Pakistan's utilization of a monthly consumption aggregate. The

conversion of expenditures to a monthly basis requires a specific treatment for fortnightly data. In the Pakistani context, fortnightly expenditures are assumed to represent consumption over 14 days. Consequently, these figures are multiplied by a factor of 2.17, an approximation of 30.5 days (average days in a month) divided by 14 days, to derive the equivalent monthly expenditure.

This comprehensive approach ensures that the consumption aggregate reflects the diverse spending patterns of Pakistani households and provides a more holistic view of their overall well-being. This crucial step culminates in the derivation of the aggregate nominal consumption expenditure, a pivotal metric representing the total monetary value of household consumption within the specified timeframe. This aggregate consumption serves as a fundamental indicator of household well-being and is a critical input for subsequent poverty calculations.

Secondly, recognizing that the cost of living can vary significantly across different regions within Pakistan, the HIES 2018-19 data is utilized to compute a spatial price index. This index accounts for regional disparities in prices, ensuring that consumption comparisons are made on a level playing field, regardless of geographic location. In the HIES 2018-19 report, a Paasche formula is utilized to construct this index, estimated at the level of each primary sampling unit (PSU). The PSU-level Paasche index is meticulously calculated using a subset of items for which both quantity and expenditure data are readily available within the HIES dataset. The spatial price index is constructed using three key components:

PSU-level Budget Shares: The budget share of each consumption item at the PSU level is calculated as a weighted average of the budget shares within that PSU, with the

weights determined by the PSU-level population estimates. This weighting ensures that the budget shares accurately reflect the consumption patterns of the population residing in each PSU.

PSU-level Median Unit Values**:** Unit values, which serve as proxies for prices, are computed for each item at the PSU level. These unit values are obtained by dividing the total expenditure on an item by the quantity consumed of that item. It is important to note that this calculation is only feasible for items where both expenditure and quantity information are available in the HIES dataset. In the context of the HIES 2018-19, this applies to most food items, tobacco and chewing products, and certain fuel and lighting items. Once these item-level unit values are estimated, the median unit value for each item at the PSU level is incorporated into the spatial price index.

National-level Median Unit Values: Like the PSU-level unit values, median unit values are also calculated for each item at the national level. These national-level median unit values act as reference prices in the estimation of the PSU-level Paasche index, providing a benchmark against which price variations across PSUs can be assessed.

The integration of these three components, namely PSU-level budget shares, PSU-level median unit values, and national-level median unit values, results in a robust and comprehensive spatial price index. This index effectively captures the relative cost of living in different PSUs, enabling for spatial price variations in their analysis of household consumption and poverty, enhancing the accuracy of poverty assessments. During the calculation of the spatial price index, all item budget shares, and median prices are weighted based on the PSU's population, ensuring accurate representation of consumption patterns. Items lacking quantity data are systematically excluded to

maintain data integrity. Furthermore, grouped item categories are also omitted due to the inherent variability in unit values within these groups, which often stems from the inclusion of diverse items rather than genuine spatial price differences. To ensure robustness, the final set of items utilized in estimating the spatial price index exclusively comprises those with complete quantity and expenditure information. Additionally, a household's contribution to the index is contingent upon their consumption of at least five of the included items. This criterion enhances the reliability of the index by filtering out households with limited consumption data. Ultimately, each PSU is assigned a unique price index, reflecting its relative cost of living. In the final stage, the index is normalized by its mean to facilitate meaningful comparisons across different PSUs.

Thirdly, the HIES 2018-19 data is leveraged to calculate a spatially adjusted monthly per-adult equivalent consumption expenditure. This metric considers not only the total consumption of a household but also adjusts for differences in household size and composition. To ensure equitable comparisons and account for these variations, adjustments are made to the welfare aggregate. These adjustments consider age and, in some cases, the gender distribution within the household. This approach recognizes that larger households may benefit from economies of scale, potentially accessing bulk-purchasing discounts or sharing certain expenses. By employing equivalent scales, these complexities are addressed, ensuring that the welfare aggregate accurately reflects the diverse needs and consumption patterns of different household types. In Pakistan, a standardized adult equivalence scale is employed to account for the varying consumption needs of individuals within a household. This scale assigns a weight of

0.8 to each member under the age of 18, acknowledging their lower consumption requirements compared to adults. Conversely, individuals aged 18 and above are assigned a weight of 1, reflecting their full adult consumption needs. By applying this equivalence scale, the welfare aggregate is adjusted to a per-adult equivalent basis, ensuring equitable comparisons across households with diverse age compositions. This spatially adjusted monthly per-adult equivalent consumption expenditure serves as the definitive measure of welfare for poverty assessment in Pakistan. By standardizing consumption on a per-adult basis and incorporating spatial price adjustments, this measure provides a more equitable and precise basis for poverty comparisons across diverse households.

Finally, a poverty line is established, typically using methodologies such as the CBN approach, which is employed by the Planning Commission of Pakistan for the HIES 2018-19 survey. This poverty line, expressed in monetary terms per adult equivalent per month, serves as a critical threshold. Households whose spatially adjusted monthly per-adult equivalent consumption expenditure falls below this line are classified as poor. By comparing the distribution of household consumption to the poverty line, the HIES 2018-19 enables the estimation of poverty rates at both national and provincial levels, providing invaluable insights into the prevalence and depth of poverty across Pakistan.

4.1.2 Estimation of Food Insecurity

HIES 2018-19 collects information on the consumption of various food items[31] on a fortnightly and monthly basis listed in the female questionnaire of the survey. In this

[31] More than 120 food items are reported in HIES 2018-19.

part of the questionnaire, respondents were asked to recall all the food items, drinks, etc., their family had consumed during the last 14 or 30 days. The survey questionnaire includes detailed food consumption data to assess the situation of food security at the household level. The questionnaire includes food items that households consumed in the form of "paid and consumed", "unpaid and consumed"[32]. These food items are then converted into kilocalories (kcal) according to the Pakistan Dietary Guidelines for Better Nutrition (PDGN) developed by the Planning Commission of Pakistan-2018[33]. The food insecurity status of each household is calculated through a daily calorie intake approach.

To ascertain the food insecurity status of each household, the first step is to estimate the Dietary Energy Consumption (DEC). DEC is measured as the average number of kcal) consumed daily by an individual. Initially, quantities of food items consumed by each household on a monthly or fortnightly basis are converted into daily food consumption expenditure. For this we divide all fortnightly food items by 14, and all monthly food items by 30.4167. Then these daily food consumption expenditures are converted into per capita by dividing household size. Then each food item is scrutinized for outliers by establishing the interquartile range. If quantities in each food item are less than then the minimum quartile range minus 2 times of interquartile range, or maximum quartile range plus 2 times of interquartile range, treated as outliers, and the same values are replaced with the median value of that food item at four different

[32] Unpaid and consumed includes wages & salaries in kind consumed, own produced and consumed, and receipt from assistance, gift, dowry, inheritance, or other sources.
[33] See Annexure-B for the calories table.

levels provided that threshold of at least 30 values[34]. If there are less than 30 values at level 1, then the median quantity is considered at level 2, and so on. After calculating per capita daily food quantities from the HIES 2018-19 data, the next step is to standardize all quantities into a consistent unit of measurement. This is essential because the data is reported in various units, including kilograms (kg) for weight-based items like pulses or meat, units for individual items like eggs, and liters (L) for liquid-based items like milk or oil. We convert all these diverse quantities into grams. PDGN tables provide the calorie amount of each food item in 100 grams of edible portion. Therefore, for each food item, Kcal/capita/day is computed:

$$tkcal = (food\ quantity * edible\ portion * density) * \left(\frac{kcal100}{100}\right)$$

Whereas $tkcal$ is kilo calories per capita per day, food quantity represents food quantity consumed in grams, the edible portion represents an item's comestible portion, and density differs across items. For instance, if an item is in liquid form like oil, then the density is 1.03, and if it is in the form of wheat then the density is 1. Kcal100 represents kilo calories per 100 grams of an item. Finally, we consider those food items where monetary value is reported but quantity is missing. To find the missing quantities, we first see the price per calorie (ppcal) which is the sum of the monetary value of daily per capita food expenditure at the household level divided by the sum of calories at the household level. Then we find the median ppcal by province and region. Finally, we get the $tkcal$ for missing items by dividing their monetary value of food by the median

[34] Level-1 is deciles, level-2 is within provinces, level-3 is within regions, and level-4 is at national level.

ppcal. Finally, to get the tkcal for a household we sum the tkcal for all household items. This $tkcal$ is DEC at the household level.

Following the estimation of DEC, the subsequent step involves establishing a threshold for Dietary Energy Requirement (DER) to assess nutritional adequacy. In this domain, researchers have utilized various threshold levels over time to evaluate nutritional intake. For this study, the benchmark established by the Food and Agriculture Organization (FAO, 2017) for 2018 specifically for Pakistan, defined as 1732 kcal/capita/day, is the normative reference for sufficient nutrition. This benchmark is called the MDER. The household is said to be food insecure if DEC is less than the MDER. As the focus of this study is to find the proportion of food insecure households (headcount) at the district level in Pakistan by using the small area modeling technique (ELL/Census EBP methods), so information on DEC and MDER is required.

4.1.3 Definition Matching HIES 2018 and PSLM 2019

Consumption in the PSLM is predicted using variables available in both the HIES and PSLM survey data. These common variables are known as auxiliary variables. These auxiliary variables are divided into four major classes:

- Household Head Characteristics (see Table 1 in Appendix A)
- Household Characteristics (see Table 2 in Appendix A)
- Dwelling Characteristics (see Table 3 in Appendix A)
- Household Asset Possession (see Table 4 in Appendix A)

These common variables between HIES and PSLM, essential for subsequent analyses and comparisons, are defined and coded systematically based on shared characteristics

and a uniform coding scheme. This standardized approach ensures consistency across the dataset, enhancing the reliability and validity of the findings while facilitating meaningful interpretation and comparison of results.

In alignment with the methodological frameworks proposed by (Elbers et al. 2003) and (Molina & Rao, 2010) for small area estimation, the selection of variables in this analysis is guided by their potential relevance as covariates of consumption. Both ELL and EBP emphasize the importance of choosing auxiliary variables that are likely to be correlated with the variable of interest. The ELL methodology focuses on selecting variables that are theoretically and empirically linked to consumption patterns, aiming to capture the underlying heterogeneity in consumption across different areas or groups. Similarly, the EBP approach emphasizes using auxiliary variables that are informative about the distribution of the outcome variable at the small area level. Therefore, common variables found in HIES and PSLM data are technically included. The focus is made on identifying those variables that are most likely to be informative about consumption patterns based on the principles of both ELL and EBP. This approach aims to enhance the predictive power and accuracy of the SAE model by utilizing variables that effectively capture the spatial heterogeneity in consumption across different areas. By adhering to this principle, the analysis deliberately includes the variables that are expected to influence consumption significantly, ensuring the robustness and reliability of the estimated small-area consumption measures. This rigorous variable selection process, informed by both ELL and EBP, contributes to the overall validity and precision of the findings of this study regarding consumption in small areas.

4.1.4 Statistical Matching

After establishing common variables across the HIES and PSLM surveys, the subsequent phase involves statistical matching of these variables. A comprehensive list of candidate variables is compiled, encompassing all common variables present in both datasets (Appendix A). Binary variables with limited applicability[35] are removed to ensure relevance and avoid undue influence in the matching process. Any variable having missing values for more than 1% of the observations is discarded to maintain data quality and prevent potential biases in subsequent analyses. To ensure comparability, the statistical properties of the remaining candidate variables are assessed for similarity between the two datasets. For both continuous and categorical variables, a meticulous comparison of the ratios of means and standard deviations is conducted. Weighted means and standard deviations are calculated for each variable once in the PSLM and once in the HIES, considering survey weights. If the ratios of the weighted means and the standard deviations between the two surveys fall within the range of 0.95 to 1.05, the variable is deemed suitable for inclusion in the regression model. This rigorous selection process, incorporating both distribution-based and summary-statistic-based comparisons, is designed to maximize the validity and reliability of the subsequent analysis by ensuring that the included variables consistently reflect the intended constructs across both the HIES and PSLM surveys. Table 4.1 shows the final list of variables that are considered for modeling after dropping the statistically mismatched variables.

[35] Those that apply to less than 5% or more than 95% of households

Table 4.1 List of variables after dropping statistically mismatched variables

	Variables	Description
1	sexratio	No. of men (15-65 yrs old)/HH size
2	workratio_adult	No. of adults employed (15-65 yrs old)/HH size
3	employ_1	Employment status of the household head
4	washingmachine	1 = household owns Washing Machine; 0 otherwise
5	dryer	1 = household owns Dryer; 0 otherwise
6	fan	1 = household owns Fan; 0 otherwise
7	microwave	1 = household owns Microwave; 0 otherwise
8	iron	1 = household owns Iron; 0 otherwise
9	table	1 = household owns Table; 0 otherwise
10	ups	1 = household owns UPS; 0 otherwise
11	geaser	1 = household owns Geaser; 0 otherwise
12	resibuilding	1 = household owns Residential Building; 0 otherwise
13	edu	Educational attainment
14	highestedu	Highest level of educational attainment in the HH
15	room	Number of Rooms in a House
16	floor	1=pacca floor; 0=otherwise
17	roof	1=pacca roof; 0=otherwise
18	wall	1=pacca wall; 0=otherwise
19	cooking	1=improved cooking source; 0=otherwise
20	water	Main source of drinking water
21	cleanwater	1=improved drinking water source; 0=otherwise
22	toilet_detail	Kind of toilet facility household use
23	toilet_type1	1=HH has flush toilet connected to source; 0=otherwise
24	internet	1=HH has an internet connection; 0=otherwise
25	lnage	Age in natural logarithm
26	urban	1=urban; 0=rural
27	edu_1	1=No Schooling; 0=otherwise
28	edu_4	1=Lower secondary (grade 9-10); 0=otherwise
29	highestedu_2	1=High Education Primary (grade 1-5); 0=otherwise
30	highestedu_4	1= High Education Lower secondary (grade 9-10); 0=otherwise
31	highestedu_6	1= High Education Undergraduate; 0=otherwise
32	toilet_detail_2	1=flush connected to public sewerage, 0=otherwise
33	toilet_detail_4	1=flush connected to pit, 0=otherwise
34	toilet_detail_7	1=dry pit latrine, 0=otherwise
35	ownership_1	1=owned house; 0=otherwise
36	ownership_2	1=rented house; 0=otherwise
37	water_2	1=hand-pump; 0=otherwise
38	water_5	1=river/spring/tanker/others; 0=otherwise
39	toilet_1	1=flush toilet; 0=otherwise
40	toilet_3	1= no toilet; 0=otherwise
41	language_new_4	1=pashto; 0=otherwise (Language of Interview)
42	marital_2	1=married; 0=otherwise

Note: See Appendix A, Tables 1 to 4 for a detailed description of variables.

Initially, a pool of 130 common variables is identified across the HIES 2018-19 and PSLM 2019-20 surveys. After a rigorous statistical matching process, which includes filtering out variables with limited applicability, excessive missing values, and significant mismatching in means and standard deviations between the two datasets, a refined set of 42 variables is deemed suitable for further analysis.

4.1.5 Location Code Matching

In socio-economic surveys like the HIES 2018 and PSLM 2019-20, precise geographical identification is paramount for granular analysis and comparison. This necessitates the utilization of hierarchical location codes, which provide a systematic framework for categorizing households based on their geographical location. These codes typically incorporate various levels of administrative division, including province, region (urban/rural), district, and primary sampling unit (PSU). To achieve this, hierarchical location codes are employed, systematically categorizing households based on their precise geographical location. These codes encompass multiple tiers of administrative division, including province, region (urban/rural), district, and PSU.

A meticulous harmonization of location codes is imperative to foster seamless integration and comparability between the HIES 2018-19 and PSLM 2019-20 datasets. While the HIES 2018-19 employs an 8-digit PSU code, the PSLM 2019-20 utilizes a similar scheme but omits the fifth digit, representing the quarter of the year. To circumvent this inconsistency and ensure a standardized level of geographical granularity, a strategic choice is to adopt a 7-digit location matching code, effectively discarding the non-essential quarter-of-the-year numeral from the HIES 2018-19 code. The resulting 7-digit code, encompassing the province, strata (division/district), region

(rural or urban), and PSU, establishes a robust and unified framework for geographical identification. By precisely aligning and standardizing location codes, this methodological approach enhances the accuracy and reliability of subsequent analyses, enabling meaningful comparisons and insights to be drawn from the integrated HIES 2018-19 and PSLM 2019-20 datasets.

4.2 Modeling Consumption & Simulation

After the meticulous data processing phase, the analysis proceeded to the modeling of consumption patterns utilizing the survey data [36]. The estimated parameters are then integrated into the PSLM framework. By employing Monte Carlo simulations within the PSLM, robust estimates of poverty and food insecurity are generated

4.2.1 Estimating Beta Model:

The process starts with estimating the Beta Model using OLS (see equation 3.1) as elaborated in Chapter 3. Before the estimation of the Beta model, there are some perquisites/steps:

Step 1: Adjust the Beta model for multicollinearity by using the Variance Inflation Factor (VIF).

While multicollinearity may not hinder the predictive capability of a model, it can lead to instability in the estimated coefficients. This instability can translate into less precise predictions, particularly when generalizing to new, unseen data, especially if the collinearity patterns differ between the training and target datasets. The presence of

[36] This stage, referred to as stage 1, leverages the HIES (Household Integrated and Economic Survey) data to derive crucial parameters of interest.

strong correlations among predictors can complicate model interpretation and lead to substantial shifts in coefficient estimates and inflated standard errors when variables are added or removed. In practice, a degree of collinearity often exists among predictors. To mitigate the challenges posed by multicollinearity, we adopt the VIF as proposed by (James, Witten, Hastie, & Tibshirani, 2013). Conventionally, VIF values surpassing 5 or 10 signal problematic multicollinearity. In our analysis, we employ a more conservative threshold of 7. Predictors exceeding this VIF value are excluded from the final Beta model.

Step 2: Stepwise Backward and Forward induction using p-values.

Stepwise regression employs a pre-defined significance threshold to guide the parsimonious selection of predictors for inclusion in the final model. In the present study, we adopt a stringent alpha level of 0.05. The backward elimination procedure is initialized by fitting a comprehensive regression model encompassing all pre-specified candidate predictors. Subsequently, the predictor exhibiting the highest p-value (i.e., the least statistically significant) is identified. If this p-value exceeds the 0.05 threshold, the predictor is systematically excluded from the model. The procedure then iterates, re-estimating the model parameters with the reduced set of predictors and re-evaluating their statistical significance. This iterative elimination process continues until all remaining predictors in the model exhibit p-values below the 0.05 threshold, ensuring that only statistically significant predictors are retained. In contrast to the backward elimination approach, the forward stepwise procedure adopts an additive strategy, commencing with an empty model devoid of predictors. The process is initiated by running k individual regressions, each focusing on one of the k independent variables

under consideration. Subsequently, the variable exhibiting the lowest absolute t-statistic is identified. The p-value associated with x_i is then compared to the pre-defined significance threshold. If the p-value is below 0.05, x_i is provisionally included in the model. Next, $k-1$ new regression fitted, each incorporating x_i along with one of the remaining $k-1$ predictors. At each step, the predictor demonstrating the most substantial improvement in model fit is identified and provisionally included. The statistical significance of all included predictors is then re-evaluated, and any predictor whose p-value surpasses the pre-defined 0.05 threshold is subsequently removed. This iterative process of inclusion and potential exclusion persists until no further predictors meet the criteria for inclusion, culminating in a final model, comprising only statistically significant predictors.

Step-3: Least Absolute Shrinkage and Selection Operator (Lasso).

In stepwise regression, when two or more predictors exhibit substantial collinearity, the model may retain only the predictor entered first into the analysis. This occurs because the first predictor captures a significant portion of the shared variance, rendering the subsequent collinear predictors statistically redundant. Consequently, the model may exhibit suboptimal predictive performance due to the exclusion of potentially relevant information. While this challenge pertains to various analytical approaches, it is particularly pronounced in certain contexts. Notably, when there is a temporal mismatch between the survey (HIES) and census (PSLM) in the context of the reference year, the issue of collinearity and its impact on model selection can be exacerbated. To address such challenges, modern estimation techniques like "post-lasso" have emerged as part of the growing field of data science. Lasso, is a shrinkage

estimator (James et al., 2013) that effectively handles collinearity by shrinking the coefficients of less important predictors towards zero, thereby performing variable selection and enhancing model interpretability. The fundamental principle underpinning lasso regression is the imposition of penalty on the absolute magnitude of coefficient estimates. This penalty effectively discourages the proliferation of superfluous predictors in the model, thereby fostering sparsity. The degree of shrinkage, and consequently the number of variables retained in the final model, is governed by the value of the tuning parameter, λ (lambda). In the context of lasso regression, the tuning parameter lambda (λ) governs the degree of regularization imposed on the model's coefficients. For a fixed number of predictors, there exists a specific λ value that corresponds to the inclusion of all variables in the model. This scenario, where λ equals zero, effectively reduces lasso regression to ordinary least squares regression, as no penalty is applied to the coefficients. Conversely, as λ increases, the penalty on the coefficients intensifies, leading to their shrinkage towards zero. At sufficiently high λ values, all predictors are effectively eliminated from the model, leaving only the intercept term. The selection of an optimal λ value is paramount for effective lasso implementation. K-fold cross-validation offers a robust approach to address this challenge, mitigating the risk of overfitting. This procedure involves partitioning the dataset into k mutually exclusive and collectively exhaustive subsets. In each iteration, one set is designated as the holdout or validation set, while the remaining $k - 1$ folds constitute the training set. A lasso model is then trained across a predefined grid of λ values using the training data. Subsequently, the trained model's predictive performance is evaluated on the held-out validation set, typically using MSE as the

performance metric. This process is repeated k times ensuring that each fold serves as the validation set once. Finally, the value that yields the lowest average cross-validation error across all folds is selected as the optimal tuning parameter for the final lasso model.

Step-4. Stepwise model adjustment by using adjusted R-squared, and BIC to find the best regression model.

Having established a preliminary set of predictors through stepwise and lasso regression, we proceed to refine the model further. This refinement entails an additional stepwise selection procedure, employing both the adjusted R-squared and the Bayesian Information Criterion (BIC) as a model selection criterion. The BIC, a widely recognized measure of model fit, balances goodness-of-fit against model complexity. It is defined as:

$$BIC = \mathrm{l}_n(n)\, k - 2\ln(\hat{L})$$

Where $\hat{L}$ represents the maximized likelihood of the model, n denotes the number of observations, and k signifies the number of estimated parameters. The objective is to identify the model that minimizes the BIC, thereby achieving a parsimonious data representation. Reducing the number of predictors (k) generally contributes to a lower BIC.

In conjunction with BIC, we also utilize the adjusted R-squared as a selection criterion. This metric, which accounts for the number of predictors in the model, quantifies the proportion of variance in the dependent variable explained by the independent variables. It is expressed as:

$$\bar{R}^2 = 1 - \frac{n-1}{n-k}(1 - R^2)$$

where n represents the number of observations and k denotes the number of independent variables. This phase aims to maximize the adjusted R-squared, thereby selecting the model that explains the most variance in the dependent variable while penalizing excessive model complexity. In the backward stepwise model refinement, the process initiates by fitting a full model encompassing all k predictors and recording its goodness-of-fit metric (either adjusted R-squared or BIC). Subsequently, each of the k predictors are individually removed from the model, and the goodness-of-fit is reassessed for the resulting $k-1$ models. If removing a predictor does not substantially diminish the model's fit, that predictor is eliminated, starting with the one exhibiting the least impact. This iterative elimination process is repeated, progressively reducing the set of predictors until all remaining variables significantly contribute to the model's overall fit.

Conversely, the forward stepwise procedure begins with an empty model. Individual regressions are performed for each k candidate predictor, and their respective goodness-of-fit values are recorded. The predictor demonstrating the most substantial improvement in model fit (x_i) is then selected and added to the model. Subsequently, each of the remaining $k-1$ predictors are individually added to this nascent model, and the resulting goodness-of-fit is evaluated. The predictor yielding the greatest improvement is incorporated into the model. This iterative inclusion process continues until no further predictors significantly enhance the model's overall fit. Finally, we compare the number of predictors under each method (lasso, stepwise backward, and

forward) based on Adjusted R-squared and BIC. The predictors that are robust and statistically significant under the specific method and threshold are included in the final model[37].

4.2.2 Estimating Alpha Model

Upon finalizing the predictor set for the Beta model, we estimate an OLS regression and extract the residuals $\hat{e}_{di}$. These residuals are then subjected to further modeling, utilizing a set of auxiliary variables not included in the Beta model specification. We construct interaction terms between these auxiliary variables and both the estimated residuals and their squared counterparts, derived from the stage-1 OLS. These interaction terms serve as regressors in the subsequent modeling phase. The resulting model is then scrutinized for multicollinearity and undergoes a refined variable selection process. Multicollinearity is assessed using the VIF criterion, as elaborated upon in the Beta model section. To optimize the Alpha model's parsimony and predictive accuracy, we employed a backward stepwise selection procedure with the adjusted r-squared criteria as the guiding metric. The detailed estimation procedure for the final model, including variance estimation, is comprehensively discussed in Chapter 3[38].

4.2.3 GLS Estimation:

While Ordinary Least Squares regression operates under the assumption of homoscedasticity, where the error variance is constant across all observations, there

[37] When the variable of interest is poverty or food insecurity, the results indicate that the step-wise model under the stepwise-backward method with adjusted r-squared criteria is the more appropriate choice.
[38] Specifically, within equations 3.4 to 3.6

may be the possibility of heteroscedasticity arising from the HIES sampling design. Specifically, the distribution of errors may vary across both households and geographical areas. To address this, we employ GLS estimation, which explicitly incorporates the varying error structures into the model. This approach yields more efficient estimates compared to OLS, as it not only provides point estimates for the regression coefficients but also accounts for the underlying distribution of both coefficients and errors. Moreover, the literature frequently highlights the phenomenon of heteroscedasticity in household expenditure data, with affluent households often exhibiting greater variance in expenditures compared to their less affluent counterparts. The ELL method further enhances our modeling framework by enabling the explicit incorporation of household-specific variances. By accommodating these varying error structures, we strive to explain the maximum possible variation in the dependent variable, thereby enhancing the precision and reliability of our estimates. However, GLS estimates based on the ELL (2003) method may potentially lead to an asymmetric variance-covariance matrix due to unequal weights within clusters[39]. To mitigate this issue, the approach recommended by (Haslett et al. 2010) is adopted, which involves averaging the matrix with its transpose to ensure symmetry. Furthermore, the incorporation of survey weights into the variance-covariance matrix is refined, drawing upon the advancements proposed by (Huang and Hidiroglou, 2003) and elaborated upon by (Van der Weide, 2014). Consequently, the estimation of GLS parameters necessitates a recalculation of the variance components. For this purpose, a modified version of Henderson's Method III (Henderson, 1953) is also employed, specifically

[39] This is not the case with HIES 2018-19 because weights remain the same for a specific cluster.

adapted to accommodate survey weights, as presented by (Huang & Hidiroglou, 2003) and further detailed by (Van der Weide, 2014). In light of the HIES sampling design, error decomposition is conducted using both the original ELL (2003) method and the modified Henderson III (H3) method. The results are described in section 4.2.4.

4.2.4 Montecarlo Simulations

In the present analysis, both the parametric and bootstrap approaches to Monte Carlo simulation are employed for estimating welfare measures in small areas[40]. The simulation results reported in this study are derived from the crucial Second Stage of the SAE framework, employing Monte Carlo simulations repeated M = 200 times. This iterative process is designed to derive the expected welfare measures conditional on the Stage 1 model and propagate the uncertainty of the model parameters into the final predictions for the target population (PSLM data). The simulation is based on the unit-level mixed model defined by the Stage 1 estimation (Eq. 3.1).

The core of the MC simulation involves generating the simulated log-welfare $(\tilde{Y})$ for all PSLM households by drawing new values for the fixed effects $\tilde{\beta}_{GLS}$, the area effect $\tilde{\eta}_d$, and the household residual $\tilde{e}_{di}$ in each replication. The specific method for drawing or re-estimating the parameters depends on the methodological scheme used. The Model setting 1 uses the parametric draws to account for parameter uncertainty, following the original Elbers et al. (2003) framework. The simulated values are drawn from a multivariate normal distribution. The variance component $\hat{\sigma}_{\eta}^2$ is drawn from a Gamma distribution to account for its uncertainty. The area effect $\tilde{\eta}_d$ is drawn from

[40] For details, see Chapter 3, section 3.1.2

normal distribution. The household residuals $\tilde{e}_{di}$ are also drawn from a normal distribution with its variance $\tilde{\sigma}^2_{e,di}$ is modeled heteroskedastically (Eq. 3.18). In the case of the extended ELL method, fixed ($\tilde{\beta}_{GLS}$) and variance parameters ($\hat{\sigma}^2_{\eta}$) are re-estimated via bootstrapping the HIES survey data in each replication, as H3 does not explicitly define the distribution for $\hat{\sigma}^2_{\eta}$. The area effects are derived as the expected value (posterior mean) conditional on the residuals using the Empirical Bayes prediction. This prediction uses the shrinkage factor $\gamma_{d,\omega}$ (Eq. 3.15) which balances the cluster-specific information with the overall average. For clusters included in the survey, $\hat{\eta}_d$ is sampled $N(\eta_d, \hat{\sigma}^2_{\eta d})$ and for those not included, samples are drawn from $N(0, \hat{\sigma}^2_{\eta})$.

The PSLM covariates, the chosen model specification, and the welfare thresholds are held fixed across all M replications. The estimation sample size for HIES is 22,677 households, and the prediction (target) sample size for PSLM is 160,654 households. The simulation's primary purpose is to obtain the final ELL estimator, $\hat{\delta}_d^{ELL}$, by averaging the calculated indicators $\delta_d^{*(m)}$ across the M simulated censuses (Eq. 3.20), and to estimate its MSE (Eq. 3.21).

Finally, the MC simulations are used to test the comparative performance of the ELL method and the Extended ELL method. This assessment is conducted by evaluating their associated uncertainty measures, specifically the Mean Squared Error (MSE) or the Standard Error (SE).

Regardless of the specific simulation technique employed, the Monte Carlo framework serves as a rigorous tool for quantifying uncertainty in small-area estimates of poverty

and food insecurity. By generating a distribution of plausible outcomes, these simulations enable the derivation of robust indicators that account for both observed and unobserved sources of variability. This comprehensive approach enhances the precision and reliability of the small area estimates, ultimately contributing to a more nuanced understanding of the underlying socioeconomic dynamics.

4.3 Diagnostics, Results and Validation

Post-model fitting diagnostics are essential for ensuring the robustness and reliability of regression analysis. Beyond addressing issues like multicollinearity during data preprocessing, the focus shifts to identifying outliers and influential observations that can unduly skew model parameters and predictions. Influence analysis, employing metrics such as Cook's distance, leverage, and studentized residuals systematically quantifies the impact of individual data points. Visualization of diagnostics like leverage-residual plots further aids in pinpointing potential outliers. However, any decision to remove observations must be prudent, weighing the benefits of improved model -fit against the potential loss of valuable information. Once the model-fitting is done and outliers are addressed, the next step is to interpret the final results based on stage 1 and stage 2 of the method. The empirical analysis encompasses a comparative assessment of model performance under two distinct methodological frameworks: the ELL method along with the Empirical Bayes (EB) prediction approach for location coupled with the Henderson III (H3) variance estimation technique, and the alternative Elbers, Lanjouv, and Lanjouv (Elbers et al., 2003) methodology for error decomposition. Furthermore, the investigation extends to two distinct outcome variables of critical socio-economic significance: poverty and food insecurity. This

multifaceted approach facilitates a comprehensive evaluation of the robustness and sensitivity of the model's predictive capabilities across varying methodological and substantive contexts. Analogously, the validation of results is an important phase that entails a rigorous assessment of the model's outputs through comparative benchmarking against established ground truth. Upon refining the model and mitigating the impact of unduly influential data points, the subsequent stage involves evaluating the concordance between model-derived estimates (aggregated at the national level from census data, such as PSLM) and corresponding survey-based estimates (at the national level from sources like HIES). This empirical validation serves as a critical appraisal of the model's efficacy in accurately capturing real-world phenomena.

4.3.1 Diagnostics Check[41]

While proceeding toward diagnostics, the initial step is to transform the dependent variable to some suitable scale for analysis. The analysis in this study used the natural log transformation based on the positively skewed nature of the dependent variable. The figure 4.1 below demonstrates the shape of the original and log-transformed dependent variable against the normal transformation.

[41] Figures 4.2 to 4.9 represent diagnostics when the dependent variable is poverty based on per-adult equivalent consumption expenditure, and Figures 4.11 to 4.18 represent diagnostics when the dependent variable is food insecurity based on Kilo Calories per capita per day.

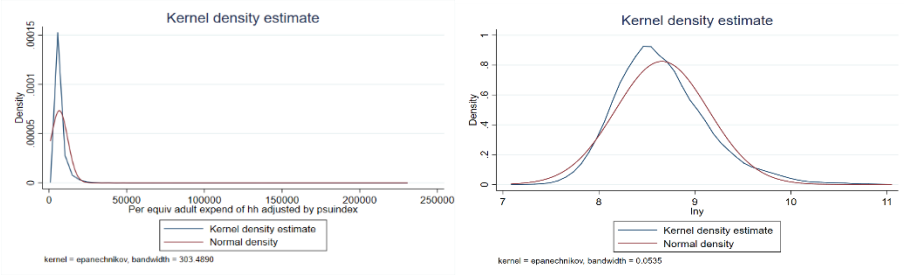


Figure 4.1 **Kernel Density Plot based on actual data of per equivalent adult expenditure**

Source: Author's construct

The kernel density plot, illustrated in blue, indicates a distribution that closely resembles a normal curve after natural logarithm transformation (on the right side). While before transformation, the dependent variable exhibits positively skewed shape (on the left side). Transformations of the dependent variable are frequently employed to enhance the alignment of empirical quantiles with their theoretical counterparts. As highlighted by Marhuenda et al. (2017), achieving a precise match to a distribution in real-world scenarios is rarely accomplished. Figure 4.1 demonstrates that data exhibits greater adaptability compared with a normal distribution with similar parameters after log transformation.

A preliminary step in diagnostic checking involves the identification of influential observations and outliers that may unduly impact the model's fit and stability. Cook's Distance, a metric that integrates both leverage[42] and residual information for each observation, provides a valuable tool for this purpose. As articulated by Cook (1977) and further elaborated by Rao and Molina (2015), Cook's Distance plot quantifies the extent to which the estimated coefficients are perturbed upon the removal of a specific observation. A conventional heuristic for flagging potentially influential observations

[42] An observation exhibiting an extreme value on a predictor variable is characterized as a leverage point.

is an absolute Cook's D value exceeding $\frac{4}{N}$, where N denotes the total number of observations in the dataset. The appended Cook's distance plot facilitates the visual identification of such observations:

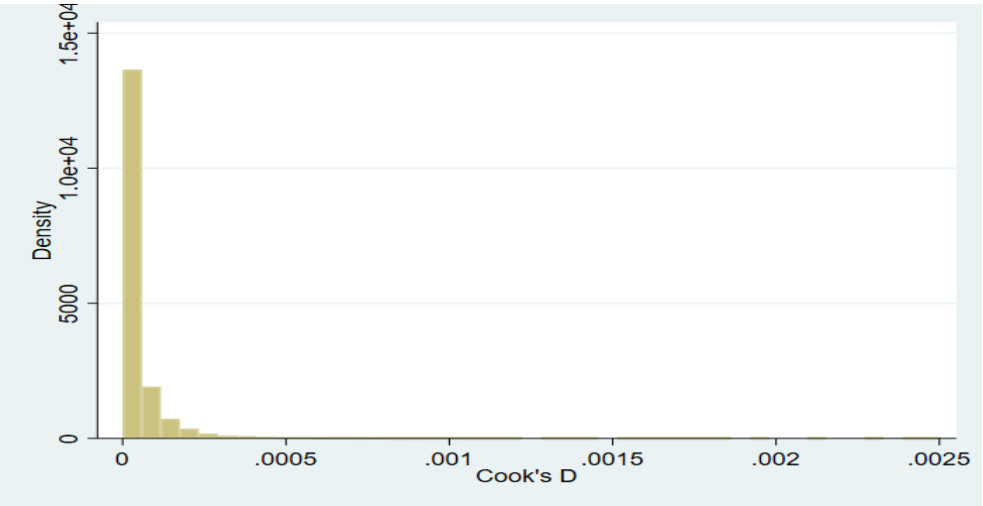


Figure 4.2 **Cook's Distance Plot**

Source: Author's construct

Cook's distance plot reveals that the vast majority of observations exhibit Cook's D values well below the threshold, suggesting a limited influence on the model's overall fit. However, a few observations with slightly elevated Cook's D values warrant further scrutiny to ascertain their potential impact and determine appropriate remedial actions. In general, having most Cook's D values near zero is desirable. This suggests that the model is not overly reliant on any single observation, and its results are more likely to be stable and generalizable.

Studentized residuals aid in pinpointing observations that deviate significantly from the model's expectations. Large absolute values of studentized residuals signal potential outliers or influential points that might disproportionately sway the model's estimates, especially in small areas where individual observations carry greater weight.

Observations with studentized residuals exceeding a predetermined threshold of ± 2 warrant further scrutiny as potential outliers or influential points.

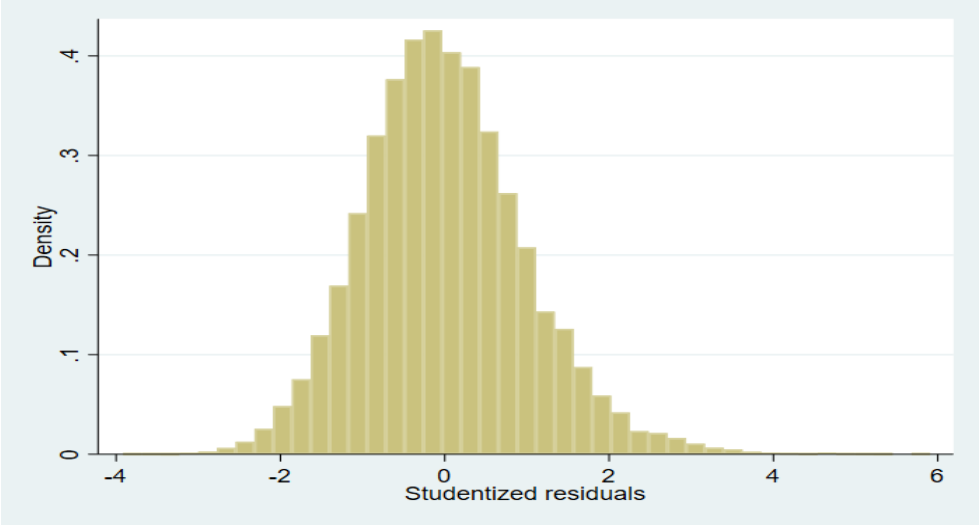


Figure 4.3 **Studentized Residuals**

Source: Author's construct

The distribution of studentized residuals closely approximated a normal distribution centered at zero, providing evidence for the validity of the normality and homoscedasticity assumptions underlying the regression model. While the values less than -2 and greater than +2 of pronounced tails in the residual histogram suggest the presence of influential observations that required further investigation.

Leverage quantifies the extent to which an independent variable's value diverges from its mean. Consequently, high leverage points possess the potential to exert a substantial influence on the estimation of regression coefficients, thereby impacting the overall model fit.

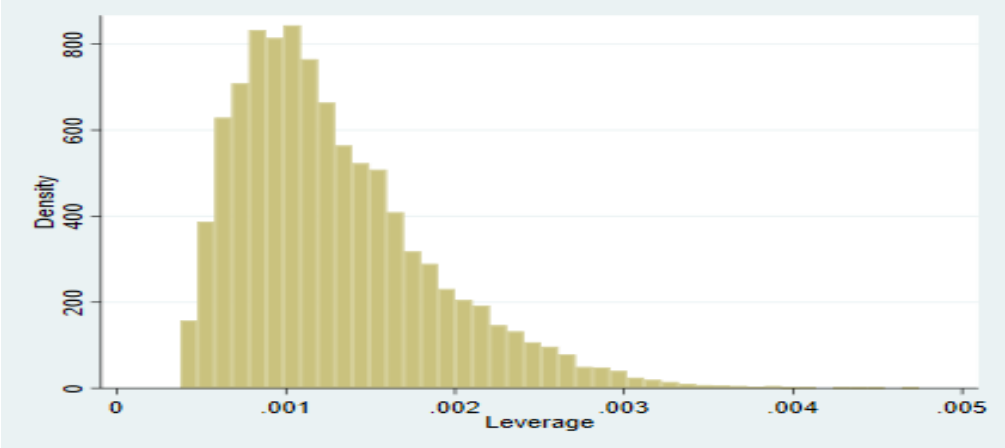


Figure 4.4 **Leverage**

Source: Author's construct

The plot reveals a right-skewed distribution, indicating that most observations have low leverage, which is generally desirable. However, the extended tail suggests the presence of some high-leverage points that warrant further investigation.

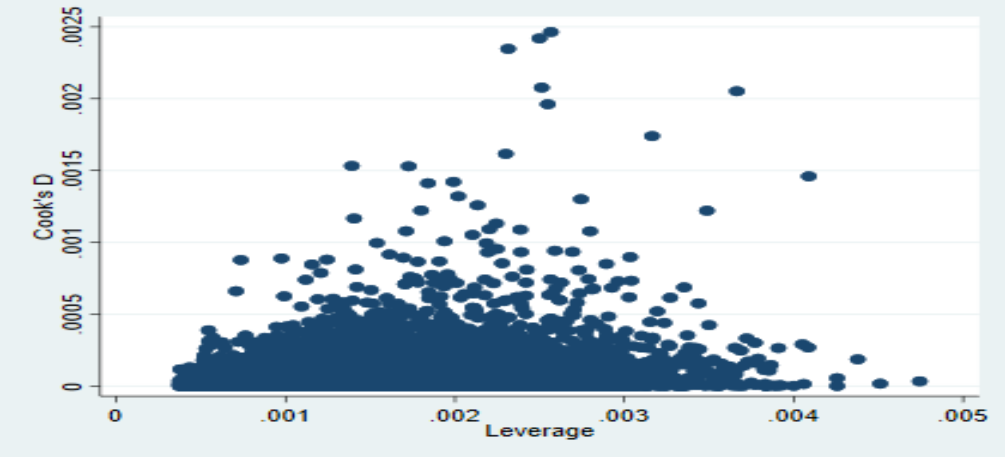


Figure 4.5 **Scatter Plot Leverage and Cook's Distance**

Source: Author's calculations

The leverage plot indicated a few data points with high leverage. In conjunction, the plot examining the relationship between Cook's distance and leverage revealed some observations with notably high Cook's distances, even at moderate leverage levels.

These findings underscore the presence of influential observations that could impact the stability and reliability of the regression model.

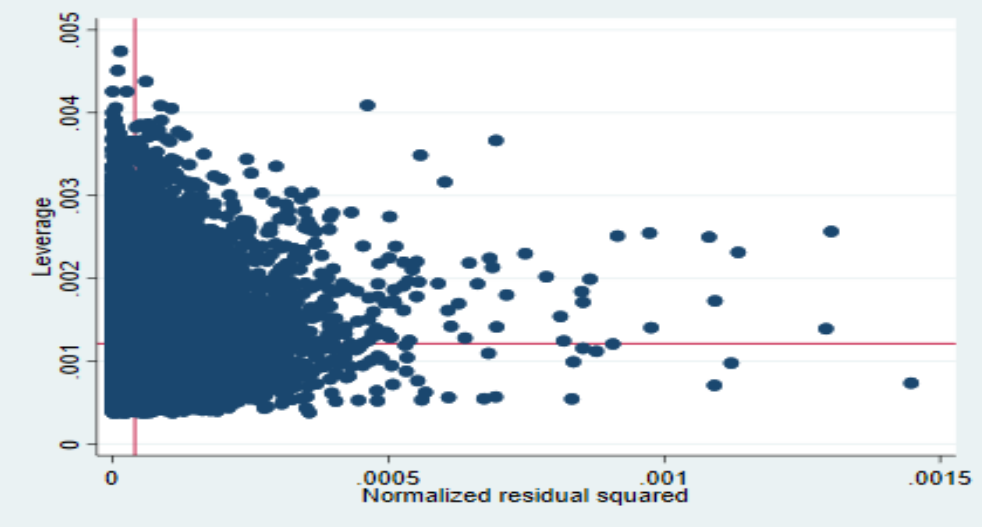


Figure 4.6 **Scatter Plot Linear Prediction and Residuals**

Source: Author's calculations

The diagnostic plot of normalized residual squared values against leverage reveals a predominantly well-behaved dataset, with most observations exhibiting low leverage and small residuals. However, a few high-leverage points warrant attention as they could potentially exert undue influence on the regression estimates. Additionally, some observations, despite having low leverage, display relatively large residuals, suggesting further investigation for potential outliers.

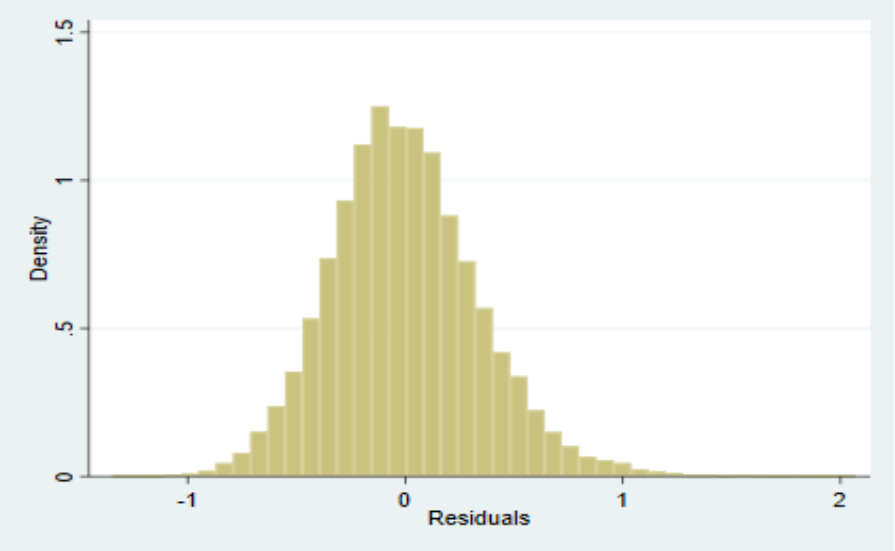


Figure 4.7 **Residuals Based on Poverty**

Source: Author's calculations

The residual analysis (Figure 4.7) reveals that the model's residuals exhibit an approximately normal distribution[43], with a mean of zero and a relatively narrow dispersion. This favorable outcome indicates that the regression model is statistically valid and reliable, thereby enhancing its predictive accuracy and credibility. However, somehow heavy tails indicate the presence of extreme values in the data which requires further investigation.

To make a more definitive conclusion about normality, we are also examining the QQ plots of the residuals. The quantile-quantile plot juxtaposes the sample quantiles of residuals with theoretical quantiles of the normal distribution, revealing deviations from the expected linear relationship on the top extreme.

[43] As Marhuenda et al. (2017) aptly highlighted real-world data often exhibit only approximate adherence to theoretical assumptions.

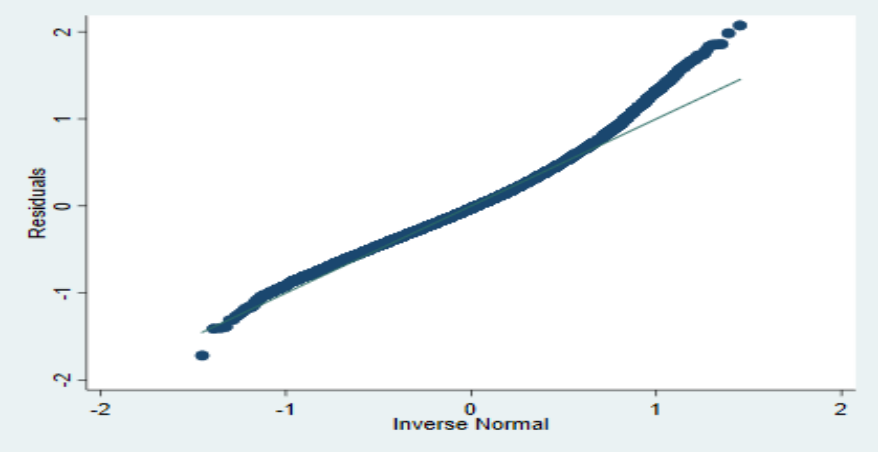


Figure 4.8 **Sample Quantiles of residuals with the theoretical normal distribution**

Source: Author's Calculations

The result suggests that the residuals do not conform to a perfectly normal distribution. The upward curvature in the tails of the plot indicates that the residual distribution exhibits heavier tails than a normal distribution, implying the presence of more extreme values on the top than anticipated under a normality assumption.

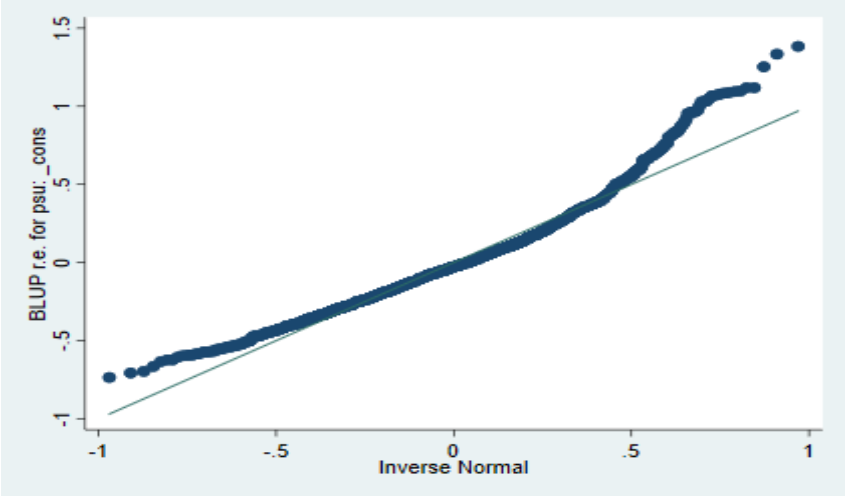


Figure 4.9 **Theoretical Quantiles VS Sample Quantiles of Random Effects**

Source: Author's Calculations

The plot exhibits that the distribution of the random effects does not conform to a perfect normality. The upward curvature in the tails of the plot suggests that the

distribution of the random effects exhibits heavier tails than a normal distribution, implying the presence of more extreme values (outliers) than would be anticipated under a normality assumption.

Figure 4.10 below represents a Kernel Density plot (of food insecurity which is the dependent variable) which is a non-parametric technique used to estimate the probability density function of a random variable of interest. It helps in identifying patterns, trends, and differences in the data. It also helps in identifying the data points that deviate significantly from the main distribution.

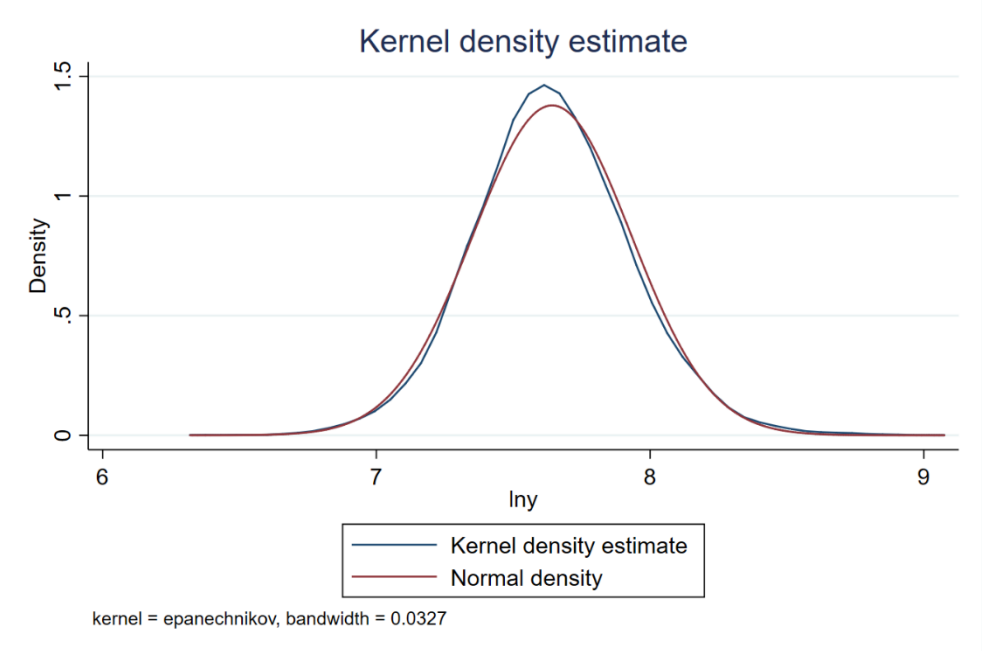


Figure 4.10 **Kernal Density plot based on data of KCAL per day per capita**

Source: Author's calculations

The kernel density estimate (KDE) plot indicates a single-peaked distribution that closely resembles a normal distribution. The application of a logarithmic transformation to the variable has successfully normalized the data, yielding a distribution that aligns well with Gaussian characteristics. Although the KDE offers a

non-parametric approach to estimating the probability density function, its visual resemblance to a normal distribution implies that parametric statistical methods, which assume normality, could be appropriate for analyzing this transformed dataset.

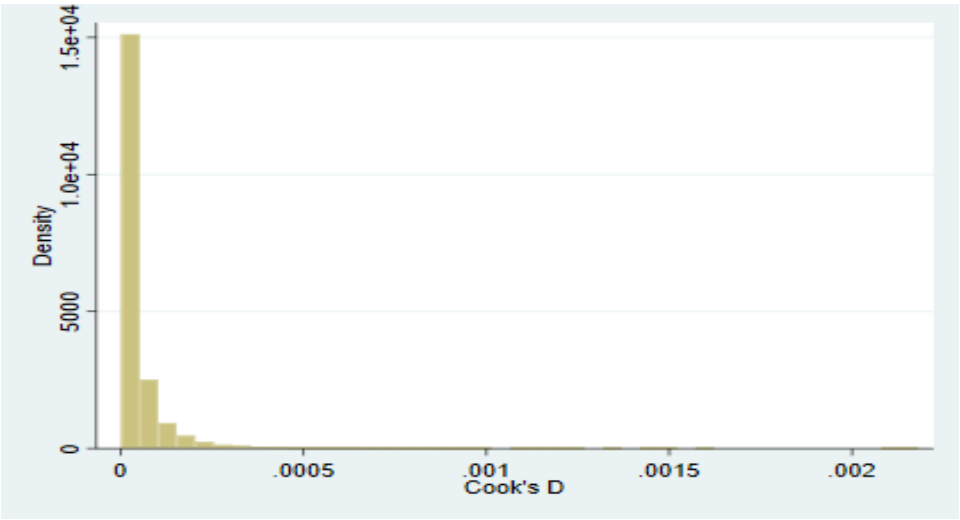


Figure 4.11 **Cook's Distance Plot**

Source: Author's calculations

The Cook's distance plot demonstrates a distribution predominantly skewed towards lower values, highlighting that the majority of observations exert minimal influence on the model's parameter estimates. Conversely, a few observations exhibit higher Cook's Distance values, indicating the presence of influential observations that may disproportionately affect the model. In the realm of small-area estimation, pinpointing and potentially mitigating the impact of these influential observations is essential to maintain the robustness and precision of the model's predictions, particularly in areas with constrained or no sample sizes.

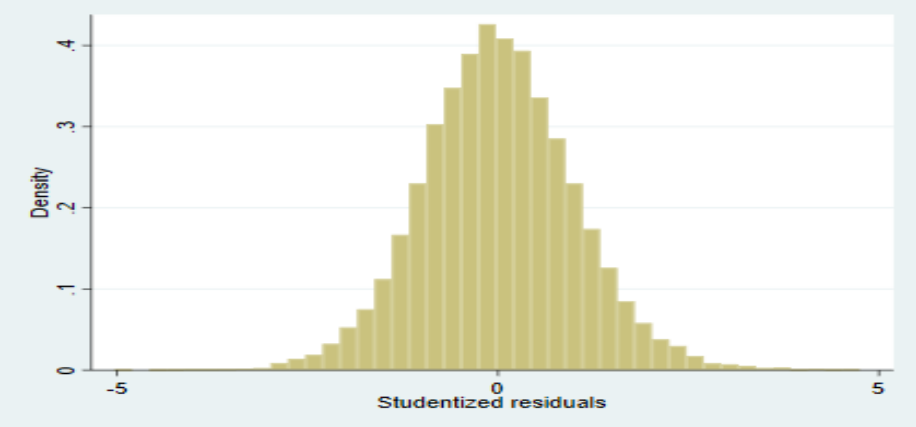


Figure 4.12 **Studentized Residuals**

Source: Author's calculations

The plot presents a histogram of studentized residuals that account for the influence of each observation on the model's fit. The histogram of studentized residuals shows a distribution that is roughly symmetric and centered around zero, suggesting that the model's assumptions of homoscedasticity (constant variance) and normality of residuals are reasonably met. However, there are some outliers visible in the tails of the distribution, indicating that a few observations may have a disproportionate impact on the model's fit.

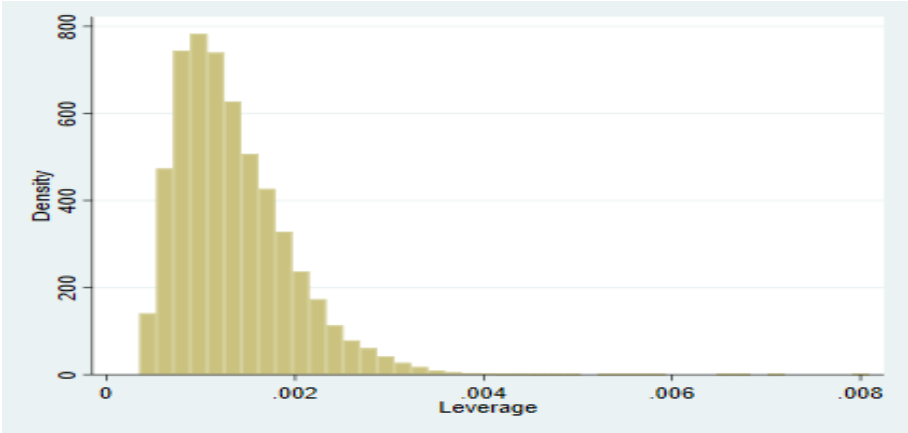


Figure 4.13 **Leverage**

Source: Author's construct

The histogram reveals a right-skewed distribution, with most observations exhibiting low leverage values. Nonetheless, a few observations with notably higher leverage values appear in the right tail of the distribution. These high-leverage data may disproportionately affect the model's fit, potentially introducing bias into the parameter estimates.

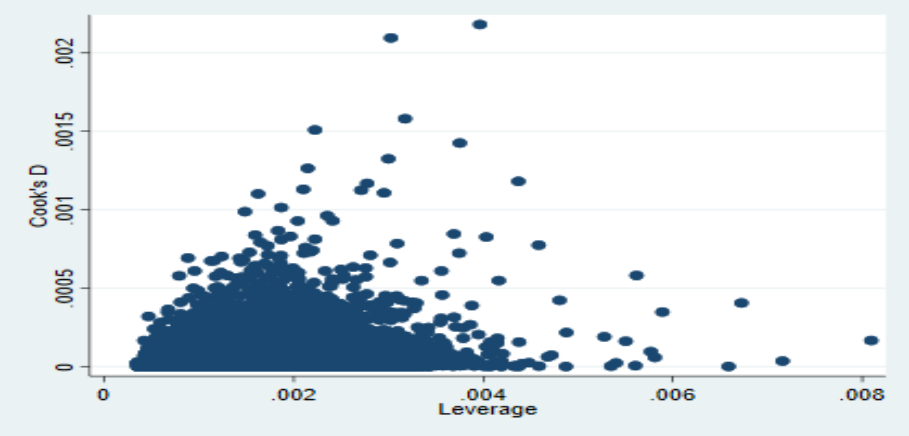


Figure 4.14 **Scatter Plot Leverage and Cook's Distance**

Source: Author's calculations

The scatter plot of Cook's distance versus leverage offers valuable insights into the impact of individual observations on the model's parameter estimates. Observations with high leverage and Cook's distance are more crucial, indicating a significant influence on the model. Additionally, some observations show high Cook's distance even at moderate leverage levels, emphasizing their potential to affect the model's stability and reliability. These points may represent outliers or influential data that can distort the model's fit and bias parameter estimates.

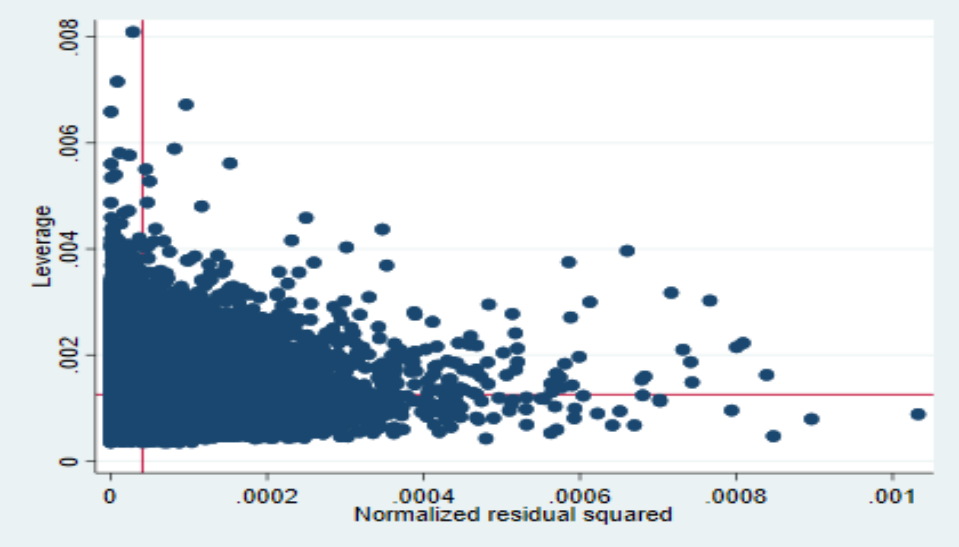


Figure 4.15 **Scatter Plot Linear Prediction and Residuals**
Source: Author's calculations

The vertical axis of the plot represents the squared residuals, which are normalized. Observations with larger values on this axis are poorly fitted by the model. The horizontal axis indicates the leverage of each observation, reflecting its potential impact on the model's fitted values. Observations with higher leverage have a greater capacity to influence the model. Points above the horizontal red line have substantial residuals, signaling a poor fit by the model. The leverage-residual plot identifies several observations with both high leverage and large residuals. These points may be influential and could introduce bias into the model's parameter estimates, necessitating careful examination and potential adjustment to maintain model accuracy and reliability.

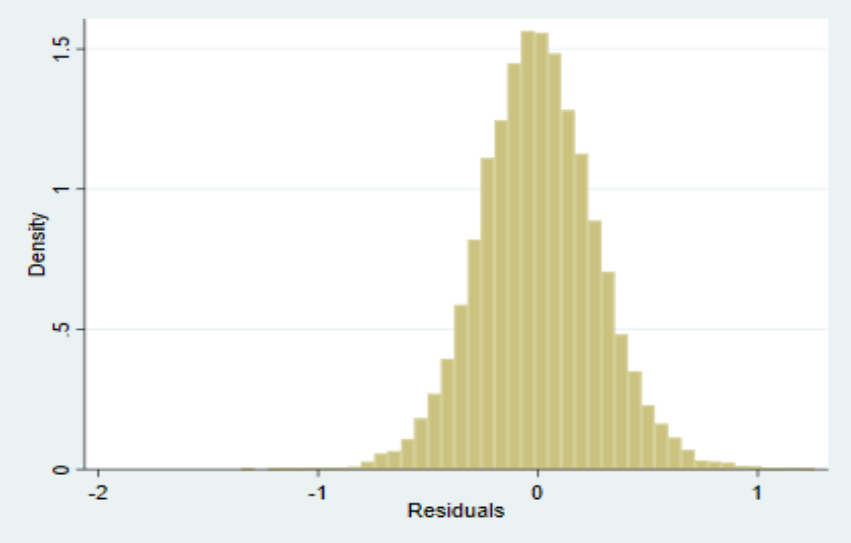


Figure 4.16 **Residuals Based on Food Insecurity**
Source: Author's Calculations

The residual analysis, as depicted in the histogram above, provides crucial insights into the performance and validity of our model. Upon examination, the histogram reveals an approximately Gaussian distribution of residuals, a key indicator of the model's adherence to fundamental assumptions. This near-normal distribution is a positive sign, suggesting robust model performance and reliability of subsequent inferences. The distribution's symmetry around the zero point is particularly noteworthy. This characteristic implies that the model demonstrates minimal systematic bias, avoiding consistent over- or underestimation of the dependent variable. Such balanced residual distribution reinforces the model's predictive accuracy across the spectrum of observations. The overall shape of the distribution closely approximates the classic bell curve, a hallmark of normality in statistical theory. However, it's important to note the presence of some outlying data points, particularly evident in the left tail of the distribution. These extreme values, while not necessarily invalidating the model, warrant further investigation. They may represent anomalies in the data, unique cases

that the model struggles to predict accurately, or potential areas for model refinement. Therefore, the Q-Q Plot is utilized to further test the normality of the data.

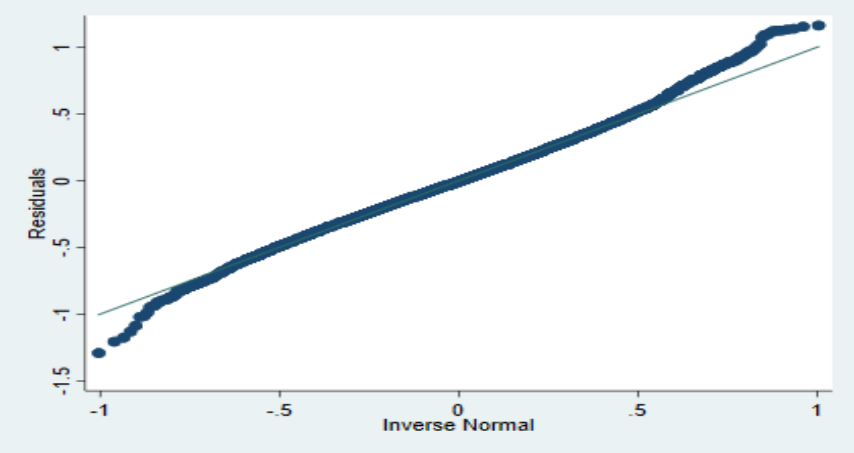


Figure 4.17 **Sample Quantiles of residuals with the theoretical normal distribution**
Source: Author's Calculations

Q-Q plot of quantiles of residuals based on HIES 2018-19 against theoretical quantiles reveals that the residuals conform to an approximately normal distribution, as supported by the empirical evidence. However, a closer examination of the residual plot reveals deviations from normality in the tails, suggesting that the residual distribution exhibits heavier tails than expected under a normality assumption. This indicates the presence of some extreme values that need to be addressed.

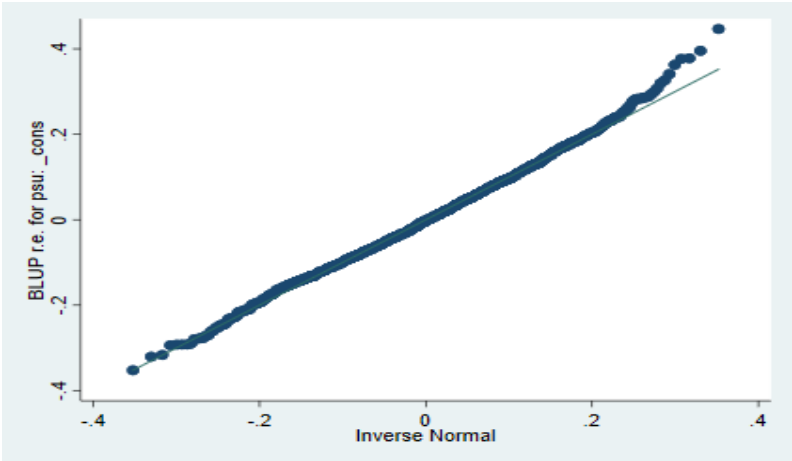


Figure 4.18 **Theoretical Quantiles VS Sample Quantiles of Random Effects**
Source: Author's Calculations

The Q-Q plot comparing the quantiles of predicted random effects from the HIES 2018-19 data with theoretical normal distribution quantiles indicates that the predicted random effects generally follow an approximately normal distribution, as corroborated by empirical data. However, a detailed analysis of the residual plot shows deviations from normality in the tails, particularly with heavier tails in the upper right. This suggests the presence of extreme values that need to be addressed. Diagnostic assessments of the regression model for poverty and food insecurity indicated the presence of influential observations that could potentially impact the model's predictive accuracy and robustness. To mitigate this risk, we adopted a conservative approach, excluding observations that exceeded the following thresholds[44]: Cook's Distance greater than $\frac{4}{N}$, absolute Studentized Residuals surpass 2 in absolute terms, and Leverage exceeds $\frac{2k+2}{N}$. The observations are then analyzed and considered as outliers accordingly. This exclusion strategy aims to enhance the model's reliability and ensure the validity of subsequent inferences.

4.3.2 Stage-1 Results[45]

Following a rigorous diagnostic assessment and the subsequent mitigation of influential observations and outliers, the regression models are re-estimated. The results of the Beta, Alpha, and GLS coefficient estimates, along with the simulation outcomes for the variables of interest (poverty and food insecurity), are presented accordingly[46]. A comparative analysis is provided, contrasting the findings obtained when error

[44] Please see Corral et al. (2021) on guidelines regarding thresholds.
[45] Table 4.3 to 4.7 relates to poverty estimation, and Table 4.8 to 4.12 narrates to food insecurity estimation.
[46] For details of Variables mentioned in Tables, please see Appendix-A, Tables 1 to 4.

decomposition is conducted via the Henderson III (H3) method in conjunction with EB, and when the ELL (2003) method is employed. The detailed results for the Beta model are comprehensively outlined in Table 4.1 for both cases[47].

Table 4.2 *Model Selection – Beta Model Estimates (Poverty)*

Dependent Variable: Per-adult equivalent Consumption Expenditure in logarithm		
Variable	**Description**	**Coefficient**
cleanwater	1=improved drinking water source; 0=otherwise	0.0986***
cooking	1=improved cooking source; 0=otherwise	0.0439***
dryer	1 = household owns Dryer ; 0 otherwise	0.0720***
edu	Educational attainment	0.0654***
edu_1	1=No Schooling; 0=otherwise	0.0571***
edu_4	1=Lower secondary (grade 9-10); 0=otherwise	-0.0163**
fan	1 = household owns Fan; 0 otherwise	0.0337***
floor	1=pacca floor; 0=otherwise	0.0617***
geaser	1 = household owns Geaser ; 0 otherwise	0.1980***
highestedu_4	1= High Education Lower secondary (grade 9-10); 0=otherwise	-0.0167***
highestedu_6	1= High Education Undergraduate; 0=otherwise	0.0447***
internet	1=HH has an internet connection; 0=otherwise	0.0909***
iron	1 = household owns Iron; 0 otherwise	0.0366***
language_new_4	1=Pashto; 0=otherwise (Language of Interview)	-0.0982***
lnage	Age in natural logarithm	-0.0231***
marital_2	1=married; 0=otherwise	-0.0713***
microwave	1 = household owns Microwave; 0 otherwise	0.2388***
ownership_2	1=rented house; 0=otherwise	-0.0605***
prov_2	Punjab	-0.0172
prov_3	Sindh	-0.0445***
prov_4	Balochistan	-0.0710***
roof	1=pacca roof; 0=otherwise	0.0453***
sexratio	No. of men (15-65 yrs old)/HH size	-0.0453***
table	1 = household owns Table; 0 otherwise	0.0897***
toilet_1	1=flush toilet; 0=otherwise	0.0457***
ups	1 = household owns UPS; 0 otherwise	0.1846***
urban	1=urban; 0=rural	0.0225***
wall	1=pacca wall; 0=otherwise	0.0612***
washingmachine	1 = household owns Washing Machine; 0 otherwise	0.0344***
water	Main source of drinking water	-0.0182***
water_2	1=hand-pump; 0=otherwise	-0.0285***
water_5	1=river/spring/tanker/others; 0=otherwise	0.1606***
workratio_adult	No. of adults employed (15-65 yrs old)/HH size	0.3557***
cons	Constant	8.0644***

Note: If p < 0.01 (***), p < 0.05 (**), p < 0.10 (*).

[47] The results of the Beta model are the same for both cases (ELL and Extended ELL Method).

The regression analysis reveals that most of the included predictors exhibit a statistically significant association with the outcome variable at the 5% significance level[48]. Notably, the prov_2 variable did not reach statistical significance. Nevertheless, recognizing the potential influence of provincial origin on the phenomenon under investigation, we retained this variable in the model as a control factor. The province variable has four categories. The prov_1 represents KPK (reference category), prov_2 as Punjab, prov_3 as Sindh, and prov_4 as Balochistan. This strategic inclusion allows us to account for any potential confounding effects associated with provincial variations. This stage of the analysis requires statistically significant coefficients as well as residuals from this model serve as a dependent variable in the subsequent Alpha Model (Please see Chapter 3 for details).

Table 4.3 *Model Selection Alpha Model Estimates*

Dependent Variable: Residual			
Variable	**Description**	**ELL**	**Henderson III (H3)**
		Coefficient	**Coefficient**
employ_1	Employment status of the household head 1=employed; 0=otherwise	-0.1801***	-0.1574***
room	Number of Rooms in a House	0.0477***	0.0397***
toilet_detail_2	1=flush connected to public sewerage, 0=otherwise	-0.0763*	-0.0562
toilet_detail_7	1=dry pit latrine, 0=otherwise	-0.2027***	-0.2728***
_cons	Constant	-3.8204***	-3.5892***

Note: If p < 0.01 (***), p < 0.05 (**), p < 0.10 (*).

As presented in the accompanying Table 4.4, the regression results demonstrate that most of the included predictors exhibit a statistically significant association with the outcome variable at the conventional 5% significance level. An exception is noted for

[48] The statistical significance of the variables under the Beta model is achieved after the rigorous model selection process. Please see section 4.2.1 for details.

the toilet_detail_2 variable which achieves statistical significance at a slightly relaxed threshold of 10%[49] under the ELL model settings. However, the same is insignificant under the H3 model setting. The important thing is that the Adjusted R-squared value should not be negative, otherwise, we need to revise our model with more predictable regressors.

Table 4.4 *Model Selection - GLS Model estimates*

Dependent Variable: Per-adult equivalent Consumption Expenditure in logarithm			
Variable	**Description**	**ELL Coefficient**	**Henderson Coefficient**
cleanwater	1=improved drinking water source; 0=otherwise	0.0788***	0.0781***
cooking	1=improved cooking source; 0=otherwise	0.0545***	0.0552***
dryer	1 = household owns Dryer; 0 otherwise	0.0735***	0.0736***
edu	Educational attainment	0.0623***	0.0620***
edu_1	1=No Schooling; 0=otherwise	0.0526***	0.0524***
edu_4	1=Lower secondary (grade 9-10); 0=otherwise	-0.0170**	-0.0169**
fan	1 = household owns Fan; 0 otherwise	0.0291***	0.0291***
floor	1=pacca floor; 0=otherwise	0.0695***	0.0699***
geaser	1 = household owns Geaser ; 0 otherwise	0.1866***	0.1861***
highestedu_4	1= High Education Lower secondary (grade 9-10); 0=otherwise	-0.0128***	-0.0128**
highestedu_6	1= High Education Undergraduate; 0=otherwise	0.0403***	0.0399***
internet	1=HH has an internet connection; 0=otherwise	0.0949***	0.0950***
iron	1 = household owns Iron; 0 otherwise	0.0353***	0.0351***
language_new_4	1=pashto; 0=otherwise (Language of Interview)	-0.0713***	-0.0686***
lnage	Age in natural logarithm	-0.0258***	-0.0263***
marital_2	1=married; 0=otherwise	-0.0640***	-0.0640***
microwave	1 = household owns Microwave; 0 otherwise	0.2238***	0.2229***
ownership_2	1=rented house; 0=otherwise	-0.0705***	-0.0710***
prov_2	Punjab	0.0049	0.0079
prov_3	Sindh	-0.023	-0.0199
prov_4	Balochistan	-0.0587***	-0.0566***
roof	1=pacca roof; 0=otherwise	0.0486***	0.0486***
sexratio	No. of men (15-65 yrs old)/HH size	-0.0586***	-0.0586***
table	1 = household owns Table; 0 otherwise	0.0800***	0.0798***
toilet_1	1=flush toilet; 0=otherwise	0.0481***	0.0485***
ups	1 = household owns UPS; 0 otherwise	0.1737***	0.1732***
urban	1=urban; 0=rural	0.0189**	0.0185**
wall	1=pacca wall; 0=otherwise	0.0479***	0.0475***
washingmachine	1 = household owns Washing Machine; 0=otherwise	0.0462***	0.0465***
water	Main source of drinking water	-0.0159***	-0.0156***
water_2	1=hand-pump; 0=otherwise	-0.0367***	-0.0369***
water_5	1=river/spring/tanker/others; 0=otherwise	0.1465***	0.1450***
workratio_adult	No. of adults employed (15-65 yrs old)/HH size	0.3619***	0.3590***
_cons	Constant	8.0786***	8.0788***

[49] The statistical significance of the variables under the Alpha model is achieved after the rigorous model selection process. Please see section 4.2.2 for details.

Table 4.5 reveals that all estimated coefficients under both model settings achieve statistical significance at the conventional 5% level except for a province variable, which serve as a control, are insignificant. This step is crucial because the basic assumption under the ELL (2003) is that GLS parameters imputed into the census should be statistically significant.

Table 4.5 *Comparison of OLS vs GLS Estimates*

Dependent Variable: Per-adult equivalent Consumption Expenditure in logarithm				
Variable	**Coefficients**		**Coefficients**	
	OLS	**ELL GLS**	**OLS**	**H3 GLS**
cleanwater	0.0986	0.0788	0.0986	0.0781
cooking	0.0439	0.0545	0.0439	0.0552
dryer	0.0720	0.0735	0.0720	0.0736
edu	0.0654	0.0623	0.0654	0.0620
edu_1	0.0571	0.0526	0.0571	0.0524
edu_4	-0.0163	-0.0170	-0.0163	-0.0169
fan	0.0337	0.0291	0.0337	0.0291
floor	0.0617	0.0695	0.0617	0.0699
geaser	0.1980	0.1866	0.1980	0.1861
highestedu_4	-0.0167	-0.0128	-0.0167	-0.0128
highestedu_6	0.0447	0.0403	0.0447	0.0399
internet	0.0909	0.0949	0.0909	0.0950
iron	0.0366	0.0353	0.0366	0.0351
language_new_4	-0.0982	-0.0713	-0.0982	-0.0686
lnage	-0.0231	-0.0258	-0.0231	-0.0263
marital_2	-0.0713	-0.0640	-0.0713	-0.0640
microwave	0.2388	0.2238	0.2388	0.2229
ownership_2	-0.0605	-0.0705	-0.0605	-0.0710
prov_2	-0.0172	0.0049	-0.0172	0.0079
prov_3	-0.0445	-0.0230	-0.0445	-0.0199
prov_4	-0.0710	-0.0587	-0.0710	-0.0566
roof	0.0453	0.0486	0.0453	0.0486
sexratio	-0.0453	-0.0586	-0.0453	-0.0586
table	0.0897	0.0800	0.0897	0.0798
toilet_1	0.0457	0.0481	0.0457	0.0485
ups	0.1846	0.1737	0.1846	0.1732
urban	0.0225	0.0189	0.0225	0.0185
wall	0.0612	0.0479	0.0612	0.0475
washingmachine	0.0344	0.0462	0.0344	0.0465
water	-0.0182	-0.0159	-0.0182	-0.0156
water_2	-0.0285	-0.0367	-0.0285	-0.0369
water_5	0.1606	0.1465	0.1606	0.1450
workratio_adult	0.3557	0.3619	0.3557	0.3590
_cons	8.0644	8.0786	8.0644	8.0788

A meticulous examination of the coefficients generated through Ordinary Least Squares and GLS regressions, detailed in Table 4.4 reveals a negligible disparity across all variables except the province variable. This analysis utilizes GLS estimates derived from two distinct error decomposition methodologies: the (Elbers et al. 2003) approach and the Henderson Method III (H3). Notably, employing the ELL method for error decomposition yields GLS coefficients that exhibit a greater affinity to the OLS estimates than those obtained using the Henderson Method III (H3).

Table 4.6 *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
Variance of epsilon	0.0687	0.0685
Sampling variance of Sigma eta sq.	0.0000003	N/A

Two distinct methodological approaches for small area estimation are employed and contrasted in this study. As detailed in Table 4.7, the first approach utilizes the original (Elbers et al. 2003) framework for error decomposition, employing a parametric

approach for estimating model parameters. Both area-specific and household-specific errors are assumed to follow a normal distribution. 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 maintains the assumption of normally distributed errors, but it diverges from the original ELL by employing bootstrapping for parameter estimation and utilizing the EB predictor to enhance the accuracy of area-level estimates.

Both approaches share an identical beta model, as this initial stage involves employing Ordinary Least Squares regression on the set of auxiliary variables to select statistically significant predictors for the welfare model. The adjusted R-squared value under the Beta Model is 0.57 which is quite satisfactory and allows us to move forward to the next analysis step. The divergence between the two methods emerges in the subsequent steps, concerning estimation of the alpha model, and the derivation of GLS based parameter estimates. An analysis of the alpha model, typically characterized by low adjusted R-squared values[50], reveals that the adjusted R-squared value for the ELL-based model setting is higher than that of the H3 model setting. Similarly, in terms of the model parameters, the rule of thumb is that the ratio of sigma eta squared (variance of location effect) over MSE should not be very high. This ratio explains the portion of the residual's variance attributable to location. The variation that auxiliary variables cannot explain, is explained by the location effect. It will be artificially low if R-squared is low. In model setting 1, the ratio of sigma eta squared over MSE is lower,

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

as compared to model setting 2 which suggests that the former performs better. However, the difference is minor.

Similarly, to estimate food insecurity headcount, the results of (stage-1) Beta, Alpha, and GLS models along with descriptions are detailed below:

Table 4.7 *Model Selection - Beta Model Estimates (Food Insecurity)*

Dependent Variable: Kilo Calories per Capita per Day		
Variable	**Description**	**Coefficient**
dryer	1 = household owns Dryer ; 0 otherwise	-0.0575***
edu	Educational attainment	0.0185***
edu_1	1=No Schooling; 0=otherwise	0.0115***
geaser	1 = household owns Geaser ; 0 otherwise	0.0855***
highestedu_2	1=High Education Primary (grade 1-5); 0=otherwise	-0.0167***
highestedu_4	1= High Education Lower secondary (grade 9-10); 0=otherwise	-0.0095**
internet	1=HH has an internet connection; 0=otherwise	0.0251***
iron	1 = household owns Iron ; 0 otherwise	0.0239***
language_new_4	1=pashto; 0=otherwise (Language of Interview)	-0.0231***
marital_2	1=married; 0=otherwise	-0.0567***
microwave	1 = household owns Microwave ; 0 otherwise	0.0672***
ownership_2	1=rented house; 0=otherwise	-0.0250***
prov_2	Punjab	-0.0758***
prov_3	Sindh	-0.0214***
prov_4	Balochistan	0.0036***
resibuilding	1 = household owns Residential Building ; 0 otherwise	0.0176***
roof	1=pacca roof; 0=otherwise	0.0187***
room	Number of Rooms in a House	-0.0060***
table	1 = household owns Table ; 0 otherwise	0.0378
toilet_1	1=flush toilet; 0=otherwise	-0.0458***
toilet_detail	Kind of toilet facility household use	0.0099***
toilet_detail_4	1=flush connected to pit, 0=otherwise	0.0470***
toilet_detail_7	1=dry pit latrine, 0=otherwise	-0.0996***
ups	1 = household owns UPS ; 0 otherwise	0.0549***
urban	1=urban; 0=rural	-0.0430***
wall	1=pacca wall; 0=otherwise	0.0265***
washingmachine	1 = household owns Washing Machine ; 0 otherwise	0.0021***
water_5	1=river/spring/tanker/others; 0=otherwise	0.0124***
workratio_adult	No. of adults employed (15-65 yrs old)/HH size	0.3909***
_cons	Constant	7.2935***

Note: If p < 0.01 (***), p < 0.05 (**), p < 0.10 (*)

As depicted in Table 4.8, the regression analysis indicates that most predictors have a statistically significant relationship with the outcome variable at the 5% significance level. This basic step is characterized by a rigorous model selection process as detailed

in section 4.2.1 of this chapter. The main purpose at this point is to get a set of auxiliary variables that are statistically significant. The only insignificant variable is table that requires significance in the subsequent GLS estimation.

Table 4.8 *Model Selection - Alpha Model Estimates*

Dependent Variable: Residual			
Variable	**Description**	**ELL**	**Henderson III (H3)**
		Coefficient	**Coefficient**
cooking	1=improved cooking source; 0=otherwise	0.0331	0.0724*
highestedu	Highest level of educational attainment in the HH	-0.0192*	-0.0298**
highestedu_6	1= High Education Undergraduate; 0=otherwise	0.1037*	0.1001*
toilet_type1	1=HH has flush toilet connected to source; 0=otherwise	-0.0824*	-0.0388**
_cons	Constant	-3.6245***	-3.4373***

Note: If p < 0.01 (***), p < 0.05 (**), p < 0.10 (*)

As depicted in Table 4.9, most predictor variables demonstrate a statistically significant relationship with the dependent variable. However, two notable exceptions emerge from this pattern. Firstly, the cooking-related predictor exhibits varying levels of significance across different model specifications. Specifically, it reaches statistical significance at a more lenient 10% threshold within the H3 model framework yet fails to achieve significance under the ELL model parametersThese findings are based on a rigorous model selection procedure as detailed in section 4.2.2 of this chapter.

Table 4.9 *Model Selection - GLS Model estimates*

Dependent Variable: Kilo Calories per Capita per Day			
Variable	**Description**	**ELL**	**Henderson**
		Coef.	**Coef.**
dryer	1 = household owns Dryer; 0=otherwise	-0.0325***	-0.032***
edu	Educational attainment	0.0169***	0.0168***
edu_1	1=No Schooling; 0=otherwise	0.0101*	0.0103**
geaser	1 = household owns Geaser; 0 otherwise	0.0751***	0.0747***
highestedu_2	1=High Education Primary (grade 1-5); 0=otherwise	-0.0169***	-0.0169***
highestedu_4	1= High Education Lower secondary (grade 9-10); 0=otherwise	-0.0067*	-0.0065*
internet	1=HH has an internet connection;	0.0279***	0.028***
iron	1 = household owns Iron; 0 otherwise	0.0194***	0.0195***
language_new_4	1=pashto; 0=otherwise (Lang. Interview)	-0.0398***	-0.0402***
marital_2	1=married; 0=otherwise	-0.0545***	-0.0545***
microwave	1 = household owns Microwave; 0	0.061***	0.0607***
ownership_2	1=rented house; 0=otherwise	-0.0213***	-0.0213***
prov_2	Punjab	-0.0958***	-0.0958***
prov_3	Sindh	-0.0408***	-0.0407***
prov_4	Balochistan	-0.0131	-0.0133
resibuilding	1 = household owns Residential Building ; 0=otherwise	0.0212***	0.0212***
roof	1=pacca roof; 0=otherwise	0.0193***	0.0192***
room	Number of Rooms in a House	-0.0071***	-0.0071***
table	1 = household owns Table ; 0=otherwise	0.0338***	0.0337***
toilet_1	1=flush toilet; 0=otherwise	-0.0289***	-0.0281***
toilet_detail	Kind of toilet facility household use	0.008***	0.0079***
toilet_detail_4	1=flush connected to pit, 0=otherwise	0.0373***	0.037***
toilet_detail_7	1=dry pit latrine, 0=otherwise	-0.0659***	-0.0645***
ups	1 = household owns UPS ; 0=otherwise	0.0505***	0.0503***
urban	1=urban; 0=rural	-0.0493***	-0.0493***
wall	1=pacca wall; 0=otherwise	0.0238***	0.0236***
washingmachine	1 = household owns Washing Machine ; 0=otherwise	0.0086*	0.0088***
water_5	1=river/spring/tanker/others; 0=otherwise	0.0184***	0.0185***
workratio_adult	No. of adults employed (15-65 yrs	0.3899***	0.389***
_cons	Constant	7.3207***	7.3222***

Note: If p < 0.01 (***), p < 0.05 (**), p < 0.10 (*)

Analysis of the regression outputs, as presented in Table 4.10, indicates robust statistical significance across the majority of estimated coefficients under both modeling frameworks. The sole exception to this trend is the prov_4 variable, which functions as a control measure in the analysis. This high degree of statistical significance among the predictors is crucial for the subsequent phase of the research. It provides a solid foundation for extrapolating food insecurity predictions to small-area levels by applying these GLS parameter estimates derived from the survey (HIES) to the more comprehensive census (PSLM) dataset. This approach leverages the strengths of both survey-based modeling and large-scale census data, potentially enhancing the granularity and reliability of food insecurity projections across diverse geographical units. The consistency of significant results across different model specifications further bolsters confidence in the robustness of these findings for small-area estimation purposes.

Table 4.10 *Comparison of OLS vs GLS Estimates*

Dependent Variable: Kilo Calories per Capita per Day					
Variable	**Coefficients**			**Coefficients**	
	OLS	**ELL**	**GLS**	**OLS**	**H3**	**GLS**
dryer	-0.0575	-0.0325		-0.0575	-0.0320	
edu	0.0185	0.0169		0.0185	0.0168	
edu_1	0.0115	0.0101		0.0115	0.0103	
geaser	0.0855	0.0751		0.0855	0.0747	
highestedu_2	-0.0167	-0.0169		-0.0167	-0.0169	
highestedu_4	-0.0095	-0.0067		-0.0095	-0.0065	
internet	0.0251	0.0279		0.0251	0.0280	
iron	0.0239	0.0194		0.0239	0.0195	
language_new_4	-0.0231	-0.0398		-0.0231	-0.0402	
marital_2	-0.0567	-0.0545		-0.0567	-0.0545	
microwave	0.0672	0.0610		0.0672	0.0607	
ownership_2	-0.0250	-0.0213		-0.0250	-0.0213	
prov_2	-0.0758	-0.0958		-0.0758	-0.0958	
prov_3	-0.0214	-0.0408		-0.0214	-0.0407	
prov_4	0.0036	-0.0131		0.0036	-0.0133	
resibuilding	0.0176	0.0212		0.0176	0.0212	
roof	0.0187	0.0193		0.0187	0.0192	
room	-0.0060	-0.0071		-0.0060	-0.0071	
table	0.0378	0.0338		0.0378	0.0337	
toilet_1	-0.0458	-0.0289		-0.0458	-0.0281	
toilet_detail	0.0099	0.0080		0.0099	0.0079	
toilet_detail_4	0.0470	0.0373		0.0470	0.0370	
toilet_detail_7	-0.0996	-0.0659		-0.0996	-0.0645	
ups	0.0549	0.0505		0.0549	0.0503	
urban	-0.0430	-0.0493		-0.0430	-0.0493	
wall	0.0265	0.0238		0.0265	0.0236	
washingmachine	0.0021	0.0086		0.0021	0.0088	
water_5	0.0124	0.0184		0.0124	0.0185	
workratio_adult	0.3909	0.3899		0.3909	0.3890	
cons	7.2935	7.3207		7.2935	7.3222	

A critical aspect of model validation in predictive analytics is the comparative assessment of Ordinary Least Squares Elbers et al. (2002) and GLS estimates. Optimal predictive power is often associated with the high degree of congruence between these two estimation methods. In this context, Table 4.9 offers valuable insights into the relative performance of different error decomposition techniques within the GLS framework.

The data reveal that GLS coefficients derived using the Elbers, Lanjouw, and Lanjouw (ELL) approach exhibit a stronger concordance with their OLS counterparts as contrasted with those generated via the Henderson Method III (H3). This alignment

suggests that the ELL method may introduce more subtle adjustments to the initial OLS estimates, potentially preserving more of the underlying data structure. However, it is important to note that the discrepancies between the ELL-based and H3-based GLS estimates are relatively modest. This observation implies that both error decomposition techniques yield reasonably consistent results, with only nuanced differences in their adjustments to the OLS baseline. The close alignment between OLS and ELL-based GLS estimates combined with the minor variations observed across different GLS methodologies, reinforces confidence in the model's stability and predictive capabilities. These findings suggest that the chosen modeling approach is robust across various estimation techniques, providing a solid foundation for subsequent predictive analyses and small-area estimations in the field of socioeconomic research.

Table 4.11 *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	23108	23108
Adjusted R-squared	0.1778	0.1778
R-squared	0.1789	0.1789
Root MSE	0.2148	0.2148
F-stat	167.6091	167.6091
Alpha-model diagnostics		
Number of observations	23108	23108
Adjusted R-squared	0.0002	0.0003
R-squared	0.0005	0.0005
Root MSE	2.2656	2.2748
F-stat	5.0937	5.2410
Model Parameters		
Sigma ETA sq.	0.0064	0.0065
Ratio of sigma eta sq over MSE	0.1395	0.1416
Variance of epsilon	0.0397	0.0397
Sampling variance of Sigma eta sq.	0.000000125	N/A

This study employs and compares two sophisticated methodological frameworks for small-area estimation, each offering unique advantages in addressing spatial heterogeneity and enhancing predictive accuracy as detailed in Table 4.12 that adheres to the seminal (Elbers et al. 2003) model for error decomposition. This method adopts a parametric stance in estimating model parameters, operating under the assumption that both area-specific and household-specific error terms conform to a normal distribution. This classic approach provides a robust foundation for small-area estimation, particularly in contexts where normality assumptions hold. In contrast, the second methodology represents an evolution of the ELL framework, as implemented by Nguyen et al. (2018). This advanced approach integrates the Henderson III (H3) model for error decomposition and incorporates Empirical Bayes (EB) prediction techniques. While maintaining the assumption of normally distributed errors, this extended model diverges from its predecessor in two key aspects:

a. It employs bootstrapping methods for parameter estimation, potentially offering greater resilience to departures from distributional assumptions.

b. It utilizes the EB predictor to refine area-level estimates, potentially yielding more precise localized predictions.

Both methods commence with an identical beta model, utilizing Ordinary Least Squares regression to identify statistically significant predictors for the welfare model from a set of auxiliary variables. The differentiation between these approaches becomes apparent in the subsequent stages, particularly in the estimation of the alpha model and the derivation of GLS parameter estimates. A critical examination of the alpha model, typically characterized by modest adjusted R-squared values, reveals a lower adjusted

R-squared for the Elbers, Lanjouw, and Lanjouw (ELL)-based model compared to the Henderson III (H3) model configuration. In assessing model parameters, a key heuristic involves examining the ratio of sigma eta squared (variance of location effect) to MSE. This ratio, which quantifies the proportion of residual variance attributable to location effects, should ideally not be excessively high. Notably, this metric can be artificially suppressed in scenarios with low R-squared values. Comparative analysis shows that Model Setting 1 exhibits a lower ratio of sigma eta squared to MSE relative to Model Setting 2. A higher ratio in this context suggests that the auxiliary variables incorporated in the model may not adequately capture between-area heterogeneity, leaving a substantial portion of spatial variation to be explained by the location effect variable.

4.3.3 Simulation Results

The second stage of both the ELL and Extended ELL methods involves simulation-based approaches to generate welfare estimates at the district level. Using the parameter estimates (regression coefficients, variance components, and location effects) obtained during the first stage, this step integrates survey data (HIES 2018-19) with auxiliary data from PSLM 2019-20. Monte Carlo simulations are employed to iteratively simulate welfare outcomes for each household in the auxiliary data. The simulation results reported in this study are derived from the crucial Second Stage of the SAE framework, employing Monte Carlo simulations repeated M = 200 times. This iterative process is designed to derive the expected welfare measures conditional on the Stage 1 model and propagate the uncertainty of the model parameters into the final predictions

for the target population (PSLM data). The simulation is based on the unit-level mixed model defined by the Stage 1 estimation (Eq. 3.1):

These simulations allow for calculating welfare indicators and their associated uncertainties, ensuring precise, granular predictions essential for policy formulation and resource allocation[51]. This study segment presents a comprehensive analysis of simulation outcomes focusing on two key socioeconomic indicators: poverty and food insecurity. These constructs are operationalized using distinct metrics to ensure robust measurement:

a. Poverty: Quantified through monthly expenditures per equivalent adult, providing a standardized measure of household economic status.
b. Food Insecurity: Assessed via kilocalories per capita per day, offering a direct measure of nutritional adequacy at the individual level.

The simulation results are systematically reported under dual model configurations[52], allowing for a nuanced comparison of methodological approaches. This comparative framework critically evaluates the model's predictive capabilities and sensitivity to different parameterizations. The study compares estimates derived from two distinct methodological frameworks: the traditional ELL approach and an enhanced version incorporating Henderson III (H3) error decomposition with Empirical Bayes (EB) prediction, referred to as ELL_H3EB.

[51] Please see section 3.1.2 for more details.
[52] Model Settings 1 refers to the ELL framework and Model Settings 2 refers to the extended ELL framework (H3 method with EB prediction), as detailed in tables 4.6 and 4.11.

Figure 4.19 below depicts the poverty headcount for both ELL and extended ELL methods by utilizing the HIES 2018-19 and PSLM (2019-20).

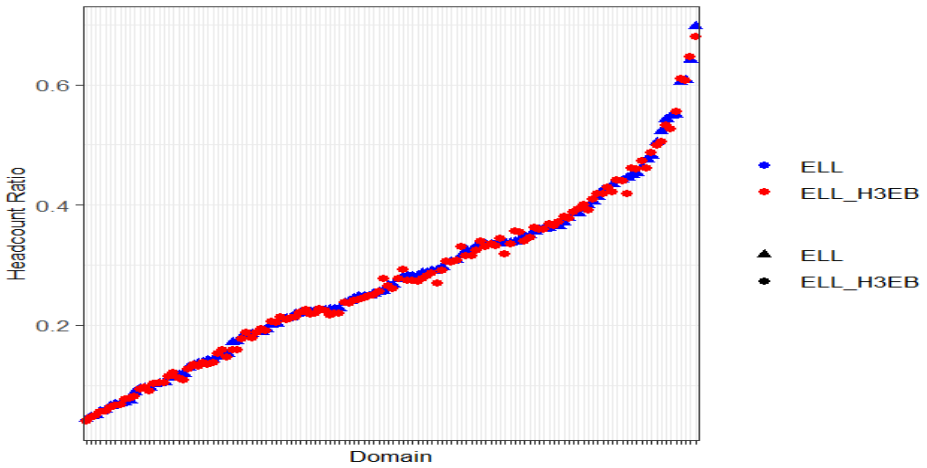


Figure 4.19 **Poverty Headcount at the District level**

Source: Author's calculations

The study employs a visual analytical approach to elucidate the relationship between 126 geographical domains (districts) and their corresponding poverty headcount ratios. Both methodologies capture a consistent overarching trend of escalating poverty levels across domains. However, a more granular examination of the visual data reveals subtle variations among the poverty headcounts under the two different analytical model settings. While these differences are not substantial enough (0.206 vs 0.204) to alter the broader narrative of poverty distribution, they nonetheless warrant further investigation.

To rigorously validate and differentiate between the poverty headcount estimates produced by each model setting, an in-depth examination of the MSE is imperative, as depicted in Figure 4.20:

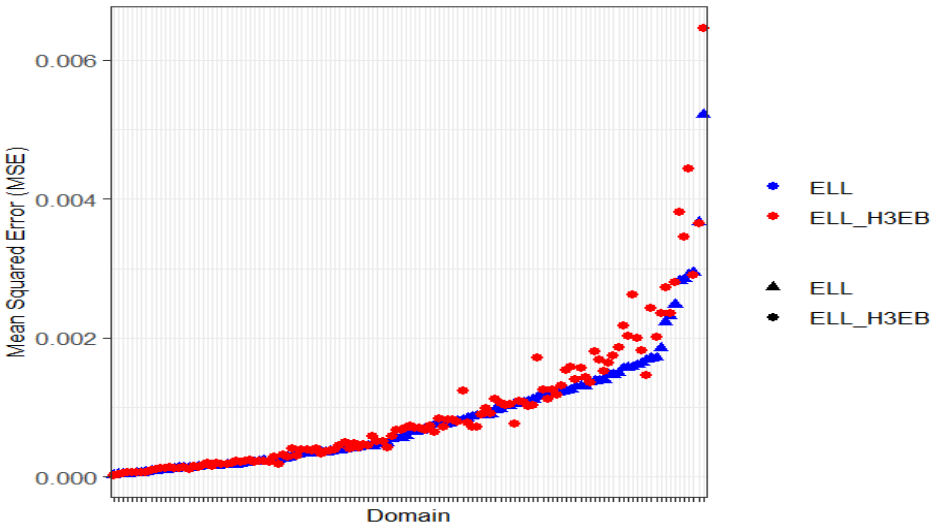


Figure 4.20 **MSE of Poverty Estimates at District level**

Source: Author's calculations

The scatterplot representation orders domains along an ascending MSE scale based on the ELL model configuration, facilitating a systematic comparison of estimation accuracy across methods and spatial units. Both methodologies demonstrate comparable performance in domains characterized by low MSE values, suggesting robust estimation capabilities in areas with more predictable socioeconomic patterns. As MSE values increase, a more pronounced divergence emerges between the two estimation approaches. This pattern indicates heightened prediction uncertainty in more complex socioeconomic environments. Notably, the traditional ELL method consistently exhibits lower MSE values (0.0009 vs 0.0010) across the spectrum of domains, particularly in areas of higher uncertainty. This systematic outperformance suggests that the ELL approach may offer superior accuracy and reliability in estimating poverty levels at the district scale within the context of Pakistan. To further validate the results, a Box-plot presentation is offered in the below appended figure 4.21:

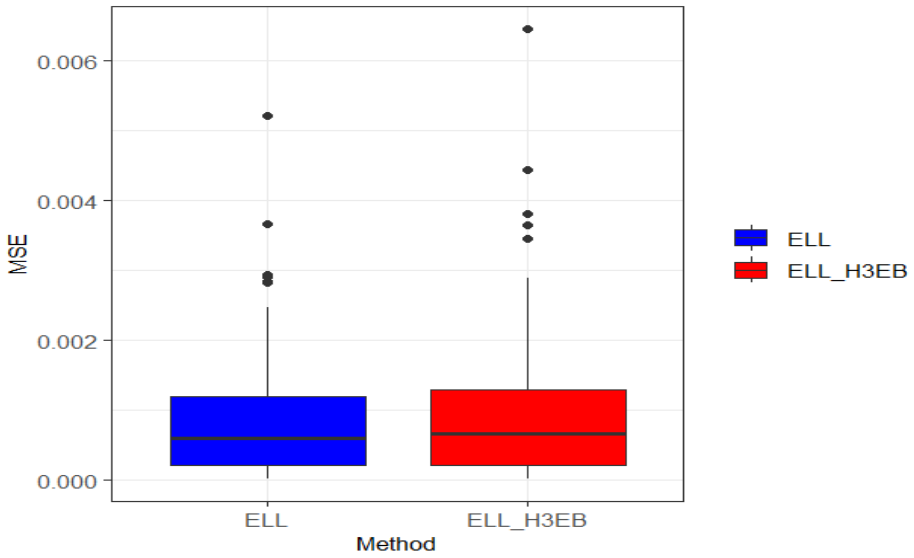


Figure 4.21 **Box Plot of MSE of Poverty Estimates at District level**

Source: Author's calculations

A box plot representation is utilized to elucidate the distribution of MSE across these simulations, offering a comprehensive view of estimation accuracy and consistency. The ELL method consistently demonstrates a lower median MSE (0.00072 vs 0.00077) compared to the ELL_H3EB approach. This indicates a higher overall predictive accuracy, suggesting that the ELL method is more adept at capturing the underlying poverty headcounts at the district level in Pakistan. The ELL method exhibits marginally reduced variability in MSE across simulations. This lower dispersion implies greater consistency in estimation performance, a crucial factor for reliable small-area poverty mapping. Both methodologies display a limited number of high-MSE outliers. These anomalies may be attributed to specific domain characteristics or sampling variability, or due to very high poverty in those areas.

Furthermore, to validate the district level results, the distribution of MSE at PSU level is also observed via Boxplot as shown in Figure 4.22:

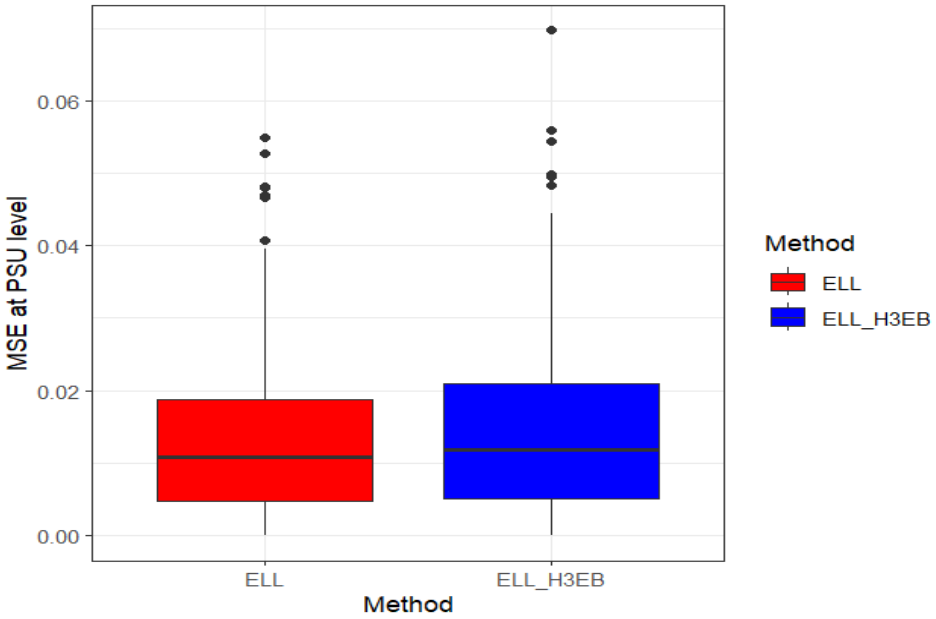


Figure 4.22 **Box Plot of MSE of Poverty Estimates at PSU level**

Source: Author's calculations

The box plot created at the PSU level provides an insightful analysis of the performance of the ELL and ELL_H3EB methodologies. Consistent with the district-level findings, the ELL method shows a lower median MSE (0.011 vs 0.012) and a more concentrated range of MSE values than the ELL_H3EB approach. This evidence supports the assertion that the traditional ELL method outperforms the ELL_H3EB method, making it a more reliable choice for poverty estimation in this context. However, it is noteworthy that the poverty estimates generated by both methods are quite similar, suggesting that either methodology could be effectively implemented.

Figure 4.23 below serves as the focal point for this analysis, offering a comprehensive visual representation of food insecurity prevalence across geographical domains under both ELL and extended ELL methods.

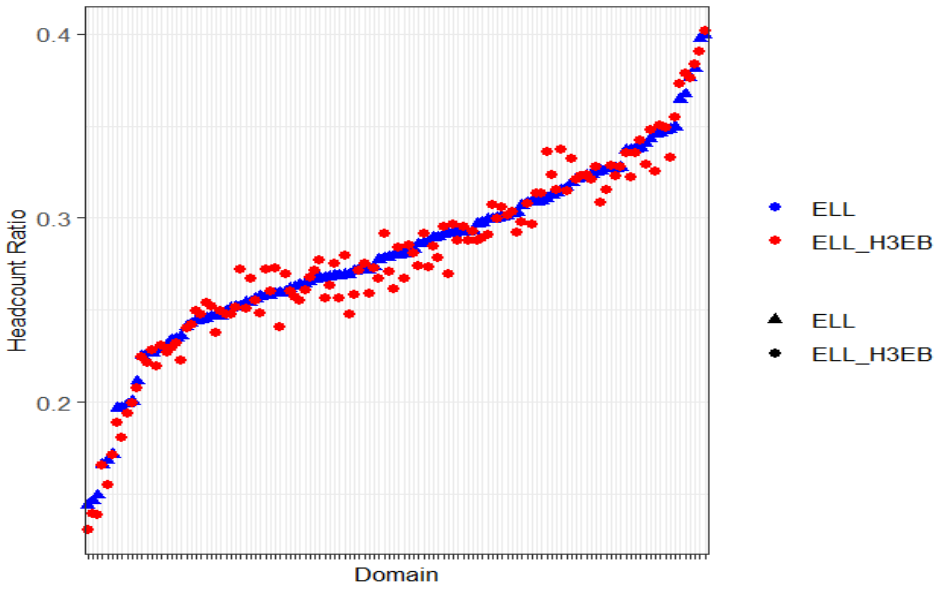


Figure 4.23 **Food Insecurity Headcount at the District level**

Source: Author's calculations

The analysis leverages a robust Monte Carlo simulation framework, encompassing 200 iterations for each household in the PSLM 2019-20 dataset under both methods. The analysis reveals a nuanced relationship between geographical domains and food insecurity prevalence, as quantified by headcount ratios derived from two distinct methodologies: the standard Elbers, Lanjouw, and Lanjouw (ELL) approach and the enhanced ELL_H3EB technique incorporating Henderson III error decomposition and empirical Bayes predictions. When domains are arranged in ascending order based on ELL-estimated headcount ratios, both methods demonstrate a consistent upward trajectory. However, the ELL_H3EB methodology generally produces marginally lower estimates, particularly in regions exhibiting higher levels of food insecurity. This discrepancy becomes more pronounced as the severity of food insecurity increases across domains. Interestingly, the comparative analysis highlights domain-specific variations between the two methodologies. In certain instances, the ELL_H3EB

approach yields substantially lower estimates compared to its ELL counterpart, while the reverse holds true for other domains. To determine the superior methodology, it is crucial to evaluate the MSE associated with each approach. The method yielding lower MSE estimates for food insecurity indicators would be considered more reliable and robust.

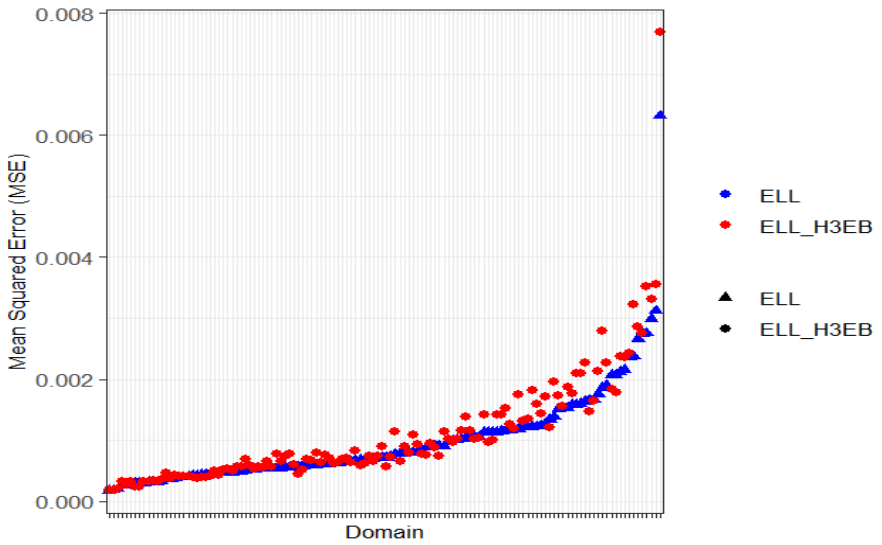


Figure 4.24 **MSE of Food Insecurity Estimates at District level**

Source: Author's calculations

Considering Figure 4.24 above, when domains are arranged in ascending order based on ELL-derived MSE values, both methodologies exhibit a consistent upward trajectory. This pattern indicates a systematic variation in estimation accuracy across different geographical units. Notably, the ELL_H3EB method demonstrates significantly more inflated MSE values than ELL method, particularly in domains where MSE estimates are very high. Interestingly, the analysis reveals that the standard ELL method demonstrates superior performance in terms of MSE reduction across a majority of domains compared to the ELL_H3EB approach. The results suggest that

ELL based food insecurity estimates are more reliable. To enhance the robustness of our findings we have employed Box Plot visualization of MSE. Figure 4.19, appended below, presents a box plot representation of the data distribution. This graphical method offers a nuanced perspective on the central tendencies, dispersion, and potential outliers within our dataset, thereby complementing and corroborating the insights gleaned from previous analyses.

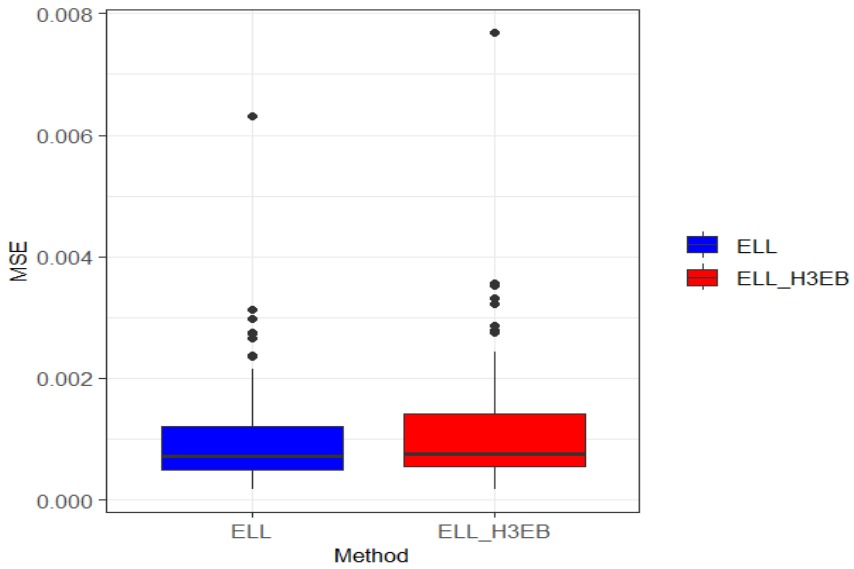


Figure 4.25 **Box Plot of MSE for Food Insecurity Estimates at District level**

Source: Author's calculations

In evaluating predictive accuracy for district-level food insecurity headcounts in Pakistan, the ELL method demonstrates superior performance over the ELL_H3EB method, as evidenced by a lower median MSE. Furthermore, the ELL method shows reduced variability in MSE across simulations, indicating enhanced consistency and reliability in small-area food insecurity estimates. Although both methods encounter some outliers with elevated MSE values, the prevailing trend highlights the ELL method's capacity to deliver more robust and precise estimates. To validate the results

further, a Box Plot presentation of MSE at the lowest aggregation level (PSU) is also observed:

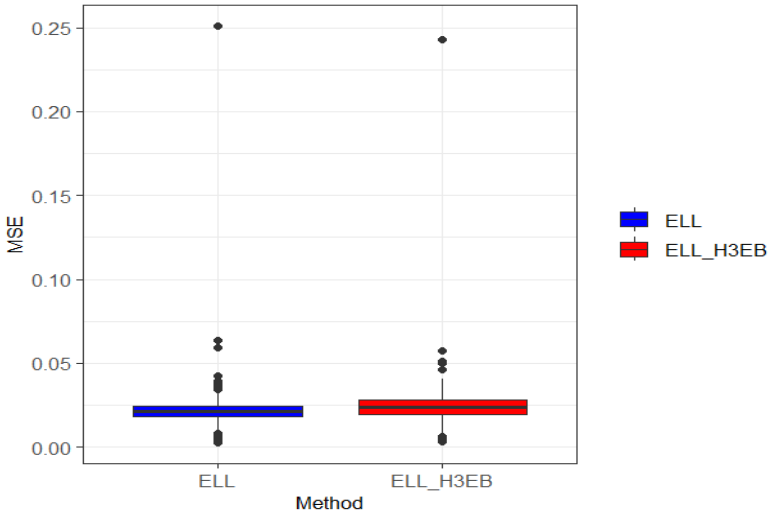


Figure 4.26 **Box Plot of MSE for Food Insecurity Estimates at PSU level**

Source: Author's calculations

The box plot analysis at the PSU level offers valuable insights into the performance of the ELL and ELL_H3EB methodologies. The ELL method demonstrates a lower median MSE and a narrower range of MSE values compared to the ELL_H3EB approach. This evidence supports the conclusion that the traditional ELL method outperforms the ELL_H3EB method, making it a more reliable choice for MSE estimation in this context. However, it is important to note that the MSE estimates produced by both methods are quite similar, indicating that either methodology could be effectively utilized.

4.3.4 Validation

Validation refers to comparing the results of Direct estimates with the Model-based estimates. Results from the previous section suggest that the ELL approach

outperforms the ELL_H3EB approach. We used two different cases. In case 1, Poverty headcount is the dependent variable. In case 2, Food Insecurity is the dependent variable. In both cases, the results suggest that the ELL model is performing better than the ELL_H3EB approach. Now the next step is to compare the direct estimates based on the survey (HIES 2018-19) with the model-based estimates.

In our analysis, we employed a sophisticated approach to calculate direct estimates. This method utilized the mean of expenditure data, which was derived through logarithmic transformation for each specified domain. To ensure accuracy and representativeness, we incorporated household weighting factors and accounted for the complex survey design in our calculations. Additionally, we applied a predetermined threshold to refine our estimates. In simple words, Direct estimates are calculated as the weighted averages of the variable of interest (e.g., income or consumption) within a specific area, using survey weights to ensure representativeness. These estimates are then compared against a predetermined threshold, such as the poverty line or MDER, to classify individuals or households as poor or food insecure. Subsequently, poverty or food insecurity is quantified using the Foster-Greer-Thorbecke (FGT) indicator, such as the headcount ratio to provide a comprehensive measure of the welfare status in the area. For variance/MSE estimation, we implemented a standardized measure that accounted for the intricate structure of the survey data. This approach considered both the weighting scheme and the inherent clustering within the dataset. By doing so, we can compute domain-specific variance estimates that reflect the true variability in our sample while adhering to the principles of robust statistical analysis. More precisely, the MSE of direct estimates is calculated by taking the difference between each

observation in the area of interest and the weighted average (direct estimate) of that area. This difference is then squared and averaged across all observations within the area, incorporating the sampling weights to ensure representativeness. The resulting value quantifies the variability or uncertainty in the direct estimate for the area. Notably direct estimates are calculated for 93 rural districts from HIES 2018-19 due to the structure of HIES[53] 2018-19. Then these direct estimates are compared with model-based estimates under ELL model settings. Figure 4.27 displays the results of MSE based on poverty.

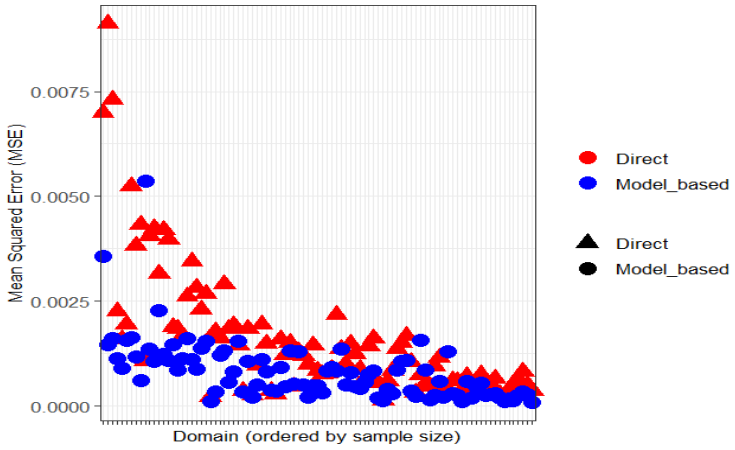


Figure 4.27 **MSE of Poverty Estimates at District level**
Source: Author's calculations

Examination of Figure 4.27 reveals a notable pattern in estimation precision across different sample sizes. In scenarios with limited sample sizes, the MSE associated with direct estimation methods exhibits substantially inflated values (0.00083) compared to the MSE of model-based (0.00088) approaches. This discrepancy in performance is particularly pronounced in the small-sample regime. The observed trend in MSE values

[53] 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.

provides compelling evidence for the superior reliability of model-based estimation techniques, especially when dealing with constrained data availability. The robust performance of model-based estimates in low-sample conditions underscores their efficacy in mitigating the limitations often encountered with direct estimation methods.

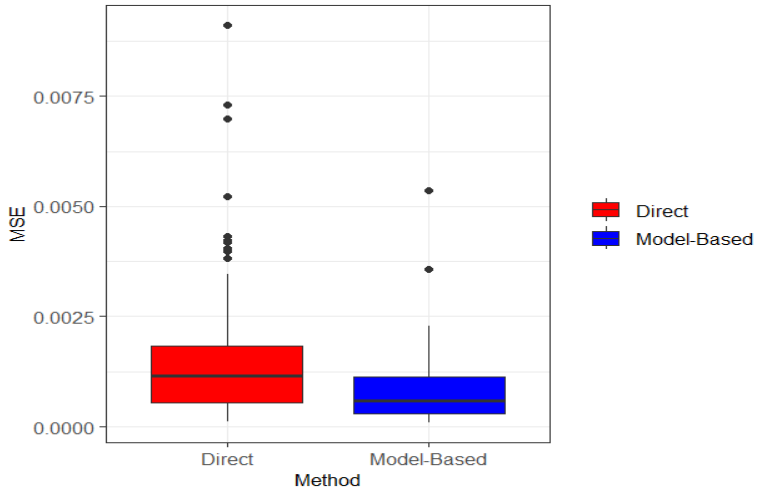


Figure 4.28 **Box Plot of MSE of Poverty Estimates at District level**

Source: Author's calculations

The box plot visualization presented in Figure 4.28 provides further corroboration of the comparative performance between direct and model-based estimation techniques. This graphical representation elucidates the distribution of MSE values across both methodologies. Analysis of the plot reveals that the MSE of the ELL model-based approach exhibits superior performance characteristics. Specifically, the model-based estimates demonstrate a markedly lower median MSE (0.00061), indicating enhanced overall accuracy. Furthermore, the interquartile range of MSE values for the model-based method is notably compressed, suggesting a higher degree of precision and consistency in its estimates. In contrast, the direct estimation method displays a wider dispersion of MSE values, with a higher median value (0.00147) and extended whiskers

in the box plot. This broader distribution implies greater variability and potentially less reliable estimates.

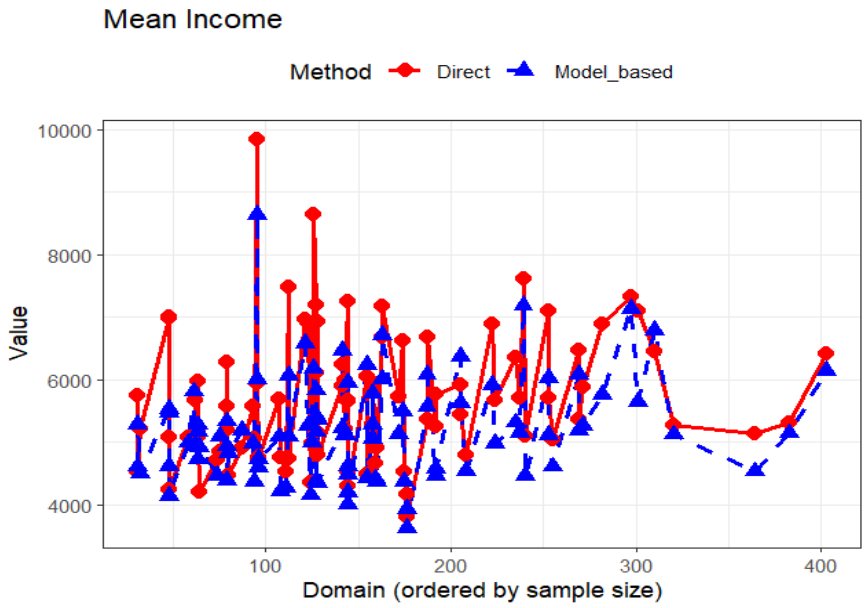


Figure 4.29 **Per equivalent Adult Expenditure at District level**

Source: Author's calculations

In our operationalization of poverty, we utilize the metric of per equivalent adult expenditure as a key indicator. Upon examination of the mean income distribution across districts, a notable pattern emerges in the comparative performance of direct and model-based estimation methods. The analysis reveals that for domains characterized by limited sample sizes, the direct estimation approach yields a substantially wider dispersion of mean income values. This increased spread in the direct estimates suggests a higher degree of volatility and potential unreliability in areas where data availability is constrained. In contrast, the model-based estimation technique demonstrates a more condensed distribution of mean income estimates across these same small-sample domains. This compression in the range of estimates indicates a

higher level of stability and consistency in the model-based approach, even when faced with data scarcity. The comprehensive analysis of our findings provides compelling evidence in favour of the model-based approach to poverty estimation. The results demonstrate a marked superiority in the robustness and reliability of model-based poverty estimates compared to alternative methods. This validation is underpinned by several key observations show that the model-based poverty estimates consistently exhibit lower variability across different domains, indicating a higher degree of precision. In scenarios with limited data availability, the model-based approach maintains its integrity, showcasing resilience to sample size constraints. The methodology appears to mitigate potential biases that often plague traditional estimation techniques, particularly in heterogeneous populations. Across various metrics and visualizations, the model-based poverty estimates display a coherent and stable pattern, reinforcing their dependability.

Finally, another imperative step is to identify the poverty status of households via poverty headcount at district level in Pakistan. The culmination of our analysis involved the application of the highly robust ELL) method to determine household poverty status through district-level poverty headcount estimation in Pakistan. This final step is crucial for generating a comprehensive poverty landscape across the country.

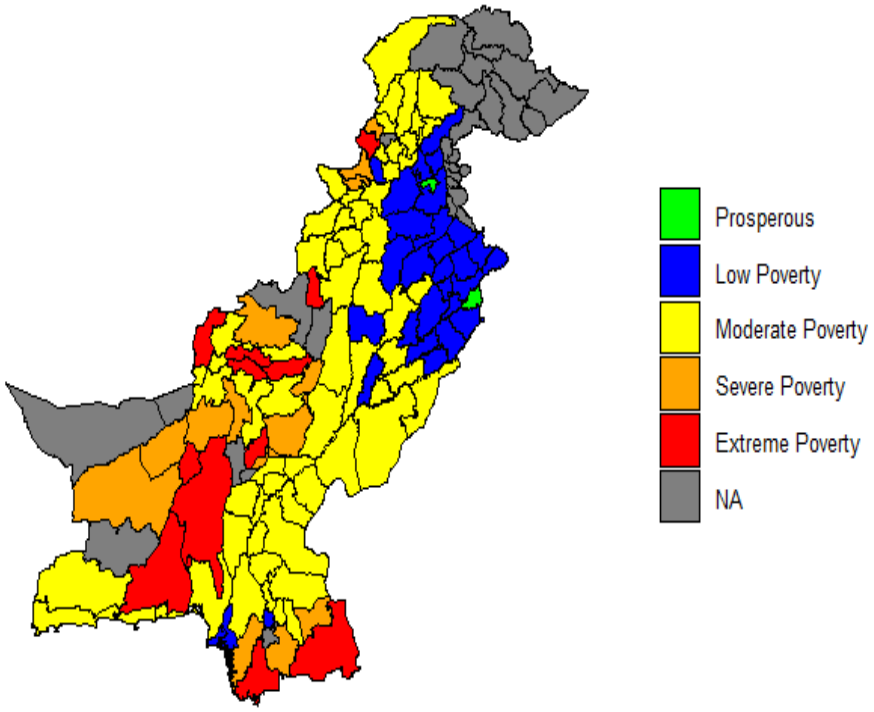


Figure 4.30 **Poverty Mapping at the District level in Pakistan – ELL Approach**

Source: Author's calculations

Figure 4.30 presents a detailed poverty mapping at the district level, offering a nuanced visualization of poverty distribution throughout Pakistan [54] . The graphical representation reveals distinct patterns of poverty intensity across various regions. Our findings indicate the presence of extreme poverty in 13 districts: Sheerani, Awaran, Tharparkar, Khuzdar, Shaheed Sikandarabad, Ziarat, Killa Abdullah, Harnai,

[54] 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.

Mohmand, Kalat, Sujawal, Duki and Nasirabad. These areas represent the most economically challenged regions within the country, necessitating targeted interventions.

Furthermore, the analysis identifies 19 districts experiencing severe poverty: Bajur, Tando Muhammad Khan, Barkhan, Killa Saifullah, Kachhi, Kharan, Dera Bugti, Nuski, Washuk, Thatta, Jaffarabad, Umerkot, Khyber, Orakzai, South Waziristan, Badin, Mirpur Khas, Shaheed Benazirabad, and Sohbatpur. While not at the extreme level, these districts still face significant economic hardships and require focused attention in poverty alleviation strategies. Interestingly, our analysis also reveals a subset of districts that, while categorized under Moderate Poverty, are precariously close to the Severe Poverty threshold. These districts include Rajanpur, Sanghar, Khairpur and Shikarpur. Their proximity to the severe poverty category underscores the need for preventive measures to forestall potential deterioration in their economic conditions.

This granular mapping of poverty intensity across districts provides insights for policymakers and development practitioners. It enables the identification of priority areas for intervention and allows for the tailoring of poverty reduction strategies to the specific needs and circumstances of each district.

To further corroborate the efficacy of our analytical approach, we conducted a comparative validation of food insecurity estimates derived from both direct and model-based methodologies. This validation process is crucial for verifying the robustness of our findings and ensuring the reliability of our conclusions.

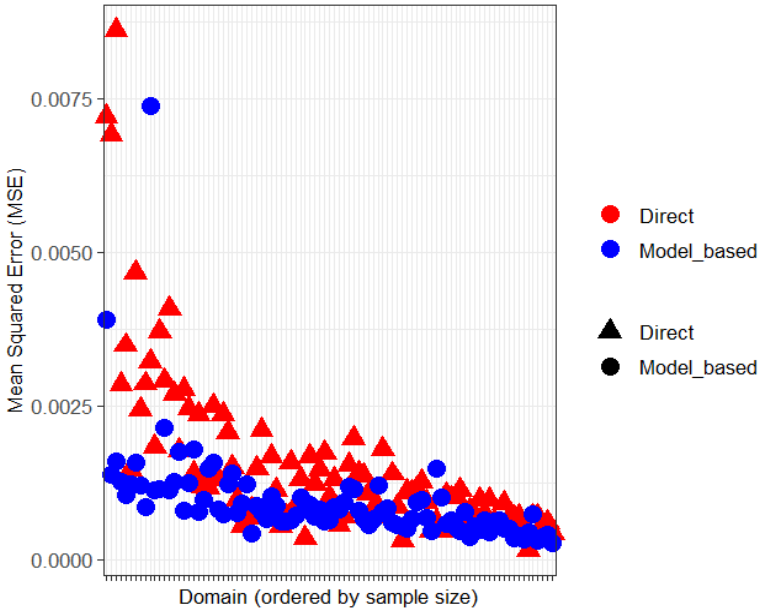


Figure 4.31 **MSE of Food Insecurity Estimates at District level**

Source: Author's calculations

Figure 4.31 presents a visual juxtaposition of the direct and model-based estimates, offering insights into their relative performance. The graphical representation unequivocally demonstrates that the direct estimation method yields substantially inflated MSE values across the spectrum of districts analysed. A particularly noteworthy observation is the heightened volatility of direct estimates in districts characterized by smaller sample sizes. This increased variability in low-sample domains suggests a potential vulnerability of the direct method to data scarcity, potentially compromising the reliability of estimates in these areas. In stark contrast, the model-based food insecurity estimates exhibit markedly superior characteristics. These estimates consistently demonstrate lower MSE values, indicating a higher degree of precision and reliability. The stability of the model-based estimates is particularly evident across varying sample size conditions, suggesting a robust performance even in data-constrained scenarios. The consistency of low MSE values in the model-based

approach underscores its resilience to sample size fluctuations and its capacity to produce dependable estimates across diverse district contexts. This uniformity in performance is indicative of the method's ability to leverage underlying data structures effectively, mitigating the impact of sampling variability on estimate quality.

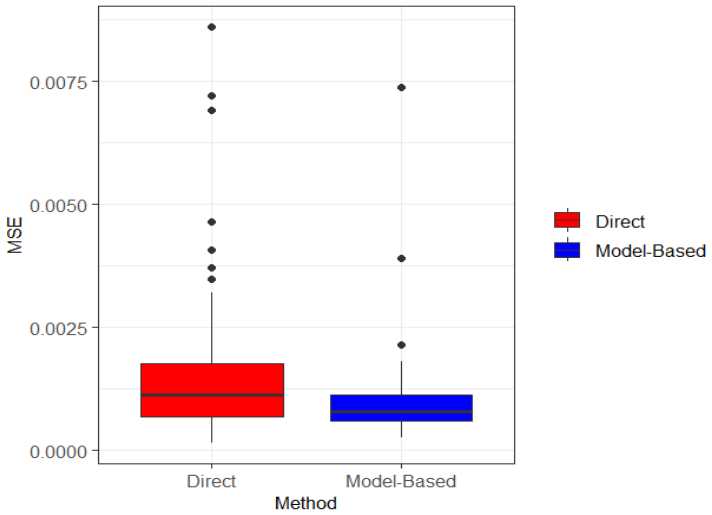


Figure 4.32 **MSE of Food Insecurity Estimates at PSU level**

Source: Author's calculations

The box plot depicted in Figure 4.32 offers substantial evidence regarding the comparative efficacy of direct versus model-based estimation techniques. This visualization highlights the distribution of MSE values for both methods. Examination of the plot indicates that the model-based approach outperforms its counterpart. Notably, the model-based method shows a significantly lower median MSE for food insecurity estimates, reflecting improved accuracy. Additionally, the interquartile range for the model-based MSE is considerably narrower, pointing to increased precision and consistency. Conversely, the direct estimation method reveals a broader

MSE distribution, characterized by a higher median and extended whiskers, indicating greater variability and less reliability in its estimates.

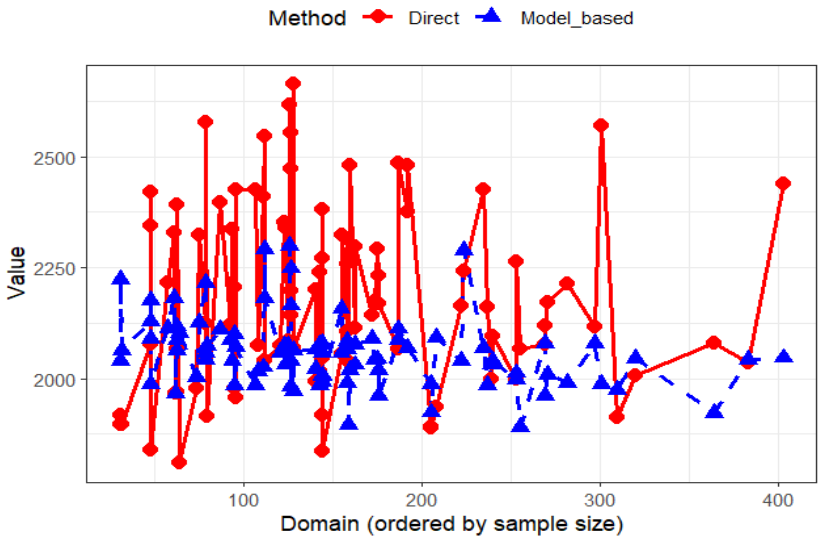


Figure 4.33 **Kilo Calories per Capita per Day at District level**

Source: Author's calculations

In assessing food insecurity status, we employ kilo calories per capita per day as a primary indicator. Analyzing the mean calorie consumption distribution across districts reveals distinct patterns in the effectiveness of direct versus model-based estimation methods. The findings indicate that in domains with smaller sample sizes, the direct estimation method results in a significantly wider spread of mean calorie consumption values. This increased variability suggests greater volatility and potential unreliability in data-constrained areas. Conversely, the model-based estimation technique displays a more compact distribution of mean calorie consumption estimates in these small-sample domains, signifying greater stability and consistency. This comprehensive analysis supports the model-based approach, highlighting its superior robustness and reliability in estimating food insecurity compared to the direct method.

This study provides a granular assessment of food insecurity in Pakistan by estimating the headcount at the district level. Utilizing the Elbers, Lanjouw, and Lanjouw (ELL) method, a robust approach widely employed for district-level food insecurity estimation, we generated a comprehensive picture of food insecurity across the nation. This methodology allowed for the identification of spatial patterns and disparities in food insecurity, offering valuable insights into the challenges faced by different districts. The resulting data contributes to a deeper understanding of the food insecurity landscape in Pakistan, informing targeted interventions and policy decisions aimed at alleviating hunger and promoting food security at the local level.

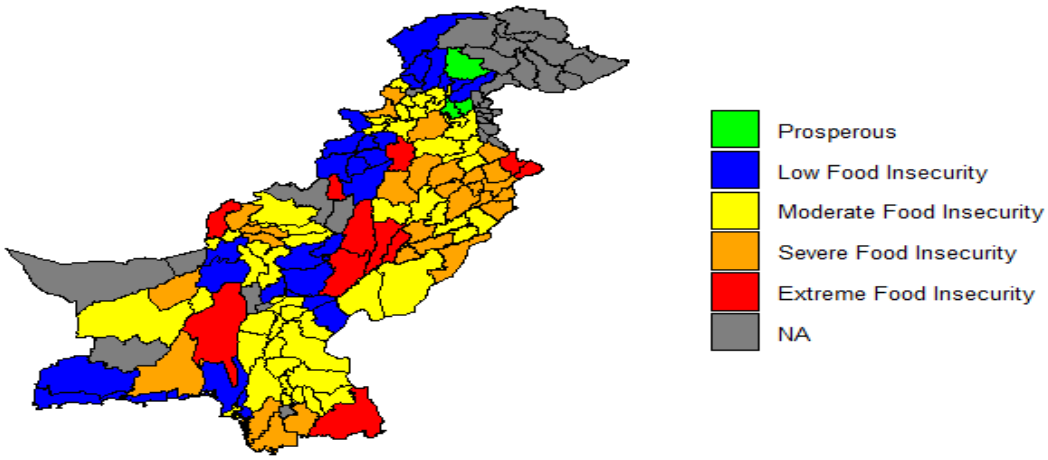


Figure 4.34 **Food Insecure Mapping at the District level in Pakistan – ELL Approach**

Source: Author's calculations

Figure 4.28 presents a detailed cartographic representation of food insecurity across Pakistan's districts. This visualization offers a nuanced perspective on the spatial

distribution of food insecurity, revealing stark disparities among different regions. Our analysis uncovered alarming concentrations of extreme food insecurity in 11 districts: Killa Abdullah, Sheerani, Khuzdar, Tharparkar, Rajanpur, Dera Ghazi Khan, Muzaffargarh, Narowal, Multan, Sialkot, Mianwali and Hafizabad. These areas represent the epicentres of nutritional vulnerability, demanding immediate and targeted interventions to address this critical issue. Furthermore, the study identified a broader swath of severe food insecurity encompassing 30 districts: Awaran, Kharan, Pishin, Harnai, Ziarat, Badin, Tando Muhammad Khan, Sujawal, Thatta, Mohmand, Khyber, Chiniot, Gujranwala, Lodhran, Khushab, Khanewal, Bhakkar, Sargodha, Sheikhupura, Kasur, Bahawalnagar, Nankana Sahib, Vehari, Gujrat, Faisalabad, Attock, Rahim Yar Khan, Mandi-Bahauddin and Bahawalpur. While not at the extreme level, these regions face significant challenges in ensuring food security for their populations, necessitating comprehensive strategies for improvement.

A critical finding of this research is the identification of districts experiencing concurrent challenges of poverty and food insecurity. This intersectionality of deprivation demands heightened attention and targeted interventions. Our analysis, leveraging the synthesized data on poverty and food insecurity, reveals a subset of districts that exhibit particularly acute vulnerability. These areas are characterized by either extreme or severe manifestations of both poverty and food insecurity, creating a compounded socioeconomic challenge. Specifically, our findings highlight fourteen districts as epicentres of this dual vulnerability: Sheerani, Tharparkar, Khuzdar, Killa Abdullah, Awaran, Ziarat, Harnai, Mohmand, Sujawal, Badin, Tando Allah Yar, Kharan, Thatta and Khyber. These regions emerge as priority areas where the

convergence of economic deprivation and nutritional insecurity creates a particularly pressing humanitarian concern. The co-occurrence of extreme or severe poverty and food insecurity in these districts suggests a complex, mutually reinforcing cycle of deprivation. This interplay between economic hardship and nutritional vulnerability underscores the need for integrated, multidimensional approaches to address these interconnected challenges.

These findings have significant implications for policy formulation and resource allocation. They emphasize the necessity for holistic interventions that simultaneously address both the economic and nutritional dimensions of deprivation. Such an approach could potentially yield synergistic benefits, creating a more substantial impact than addressing each issue in isolation. Moreover, identifying these high-priority districts provides a crucial focal point for both governmental and non-governmental organizations engaged in poverty alleviation and food security initiatives. It offers a data-driven basis for strategically deploying resources and designing tailored programs that address the specific needs of these highly vulnerable populations. In conclusion, our research underscores the importance of recognizing and addressing the overlap between poverty and food insecurity at the district level. These 14 districts emerge as key areas requiring immediate and comprehensive interventions to break the cycle of deprivation and improve the overall well-being of their populations.

CHAPTER 5

EMPIRICAL BAYES (EB) PREDICTION AND COMPARATIVE ANALYSIS

After producing district-level poverty estimates in Chapter 4, this chapter delves into a comparative analysis of the EB prediction method and the ELL method, evaluating their performance in estimating poverty and food insecurity, along with their associated standard errors, within an empirical context. Specifically, we juxtapose the EB prediction approach, grounded in a nested-error regression model, with the ELL approach (as elaborated in Chapter 4).

The chapter unfolds as follows: Section 5.1 provides a concise overview of the EB prediction approach for poverty and food insecurity estimation. Section 5.2 presents a comparative assessment of the SAE estimates derived from both the (census) EB predictor and ELL methods. Section 5.2 further synthesizes the study's findings, highlighting the most vulnerable districts in terms of poverty and food insecurity within the country by utilizing the superior modeling approach.

5.1 EB Prediction

To get the welfare (poverty and food insecurity) estimates under EB prediction at the district level in Pakistan, HIES 2018-19 and PSLM 2019-20 are utilized. As briefly discussed in Chapter 4 about the structure of HIES and PSLM, section 4.1 and sub-section 4.2.1 remain the same for the estimation of the (Census) EB predictor. In section 4.1, the following sub-sections 4.1.1, 4.1.2, 4.1.3, 4.1.4, and 4.1.5 remain the same.

- Estimation of poverty by utilizing (HIES 2018-19) survey (sub-section 4.1.1).

- Estimation of food insecurity by utilizing (HIES 2018-19) survey (sub-section 4.1.2).
- Definition Matching HIES 2018-19 and PSLM 2019-20 (sub-section 4.1.3).
- Statistical Matching (sub-section 4.1.4).
- Geographical Codification and Location Code Matching (sub-section 4.1.5).

Similarly, in section 4.2 (modeling consumption and simulation), the initial steps are the same which include the estimation of linear model 3.1 for the selection of covariates (or auxiliary variables).

Once the final set of covariates is selected, the next step is to estimate the model by using the appropriate method.

5.1.1 Leveraging Restricted Maximum Likelihood

This study employs Restricted Maximum Likelihood as a pivotal step in enhancing the accuracy of small area poverty estimates, specifically utilizing a Census-based Empirical Bayes (Census EB) approach inspired by Molina & Rao (2010) and informed by the observations of (Molina, 2019). As Molina (2019) highlights, REML provides a crucial refinement over standard Maximum Likelihood (ML) estimation by correcting for the degrees of freedom lost when estimating regression coefficients. This correction yields a less biased estimator of the variance components, particularly beneficial in finite sample settings commonly encountered in small area estimation. REML addresses a critical limitation of standard ML estimation in the context of mixed-effects models, which are essential for capturing the inherent hierarchical structure prevalent in small-area data. While ML estimation yields unbiased fixed effects estimates, it tends to underestimate the variance components associated with random effects. This

underestimation arises from ML's inability to account for the degrees of freedom lost during the estimation of fixed effects, leading to potentially biased variance components and subsequently, underestimated standard errors. REML circumvents this issue by maximizing the likelihood of a transformed dataset that is invariant to the fixed effects, effectively decoupling the estimation of fixed effects and variance components. This leads to more accurate variance component estimates. Under REML, the log transformation is chosen to ensure the residual likelihood remains independent of the fixed effects. Then an appropriate interval for the variance component is determined with an initial value within the predefined interval. Subsequently, the residual log-likelihood function is maximized to estimate model parameters, holding the variance component constant. This process is repeated, iteratively updating until the residual log-likelihood function attains its maximum value, yielding the optimal variance component estimate. By integrating REML within a Census EB framework, this study ensures more precise variance component estimation, resulting in more accurate standard errors and ultimately, more robust and defensible poverty and food insecurity estimates at the small area level.

5.1.2 Model-based Simulations

To get the poverty and food insecurity estimates, the survey data is used to fit the nested error model via REML and obtain the parameter estimates. By utilizing the parameter estimates, we generated M vectors of the values of response variables for PSLM. .

The area effect and household errors are generated using normal distribution.

Then we have calculated the indicator of interest (poverty or food insecurity) by replicating the results 200 times by averaging the same. The mean squared error is calculated by utilizing the bootstrap simulations B.

For the variable of interest, which is poverty (and food insecurity) in this case, we used Foster, Greer, and Thorbecke's (1984) FGT poverty measure. As we are interested in headcount poverty, so we used the simplest form of FGT poverty indicator with $\alpha = 0$. For Poverty headcount, if per equivalent adult monthly expenditure is less than the targeted poverty line, then the household is said to be poor. For the Food Insecurity headcount, if Kilocalories per capita per day are less than the MDER, the household is said to be food insecure and vice versa.

5.1.3 Results, Validation, and Benchmarking

Our analytical framework encompasses a multifaceted approach to evaluate and validate the Census Empirical Best (EB) prediction method within the context of Small Area Estimation. Before parameter estimation, we implemented a rigorous variable selection protocol, as detailed in Chapter 4. This process involved advanced statistical techniques to identify the most relevant and predictive auxiliary variables. Utilizing the REML method, we conducted a comprehensive estimation of model parameters. To thoroughly assess the efficacy of the Census EB estimates we applied simulation and bootstrap procedures explained in section 5.1.2 of Chapter 5. Finally, a critical component of our study involved a comparative analysis between direct estimation methods and our model-based approach. This comparison was further enriched by a detailed examination of the respective MSE associated with each method. This benchmarking process quantifies the potential gains in estimation precision offered by

the SAE approach and helps in identifying the systematic differences or biases that arise in direct estimates. By integrating these diverse analytical components, from initial model specification through to comparative benchmarking, this study offers a comprehensive evaluation of the Census EB prediction method.

Table 5.1 below depicts the parameter estimates by utilizing the REML method when poverty is a dependent variable.

Table 5.1 *Restricted Maximum Likelihood Model estimates*

Dependent Variable: Per-adult equivalent Consumption Expenditure in logarithm		
Variable	**Description**	**Coefficient**
cleanwater	1=improved drinking water source; 0=otherwise	0.0744***
cooking	1=improved cooking source; 0=otherwise	0.0666***
dryer	1 = household owns Dryer; 0 otherwise	0.0833***
edu	Educational attainment	0.0629***
edu_1	1=No Schooling; 0=otherwise	0.0476***
edu_4	1=Lower secondary (grade 9-10); 0=otherwise	-0.016**
fan	1 = household owns Fan; 0 otherwise	0.0352***
floor	1=pacca floor; 0=otherwise	0.0709***
geaser	1 = household owns Geaser ; 0 otherwise	0.1818***
highestedu_4	1= High Education Lower secondary (grade 9-10);	-0.0199***
highestedu_6	1= High Education Undergraduate; 0=otherwise	0.0483***
internet	1=HH has an internet connection; 0=otherwise	0.0933***
iron	1 = household owns Iron; 0 otherwise	0.0307***
language_new_4	1=pashto; 0=otherwise (Language of Interview)	-0.0663***
lnage	Age in natural logarithm	-0.0226***
marital_2	1=married; 0=otherwise	-0.0753***
microwave	1 = household owns Microwave; 0 otherwise	0.2225***
ownership_2	1=rented house; 0=otherwise	-0.0757***
prov_2	Punjab	-0.0017
prov_3	Sindh	-0.0307**
prov_4	Balochistan	-0.0776***
roof	1=pacca roof; 0=otherwise	0.0492***
sexratio	No. of men (15-65 yrs old)/HH size	-0.0375***
table	1 = household owns Table; 0 otherwise	0.077***
toilet_1	1=flush toilet; 0=otherwise	0.0533***
ups	1 = household owns UPS; 0 otherwise	0.1617***
urban	1=urban; 0=rural	0.0231***
wall	1=pacca wall; 0=otherwise	0.0421***
washingmachine	1 = household owns Washing Machine; 0 otherwise	0.0445***
water	Main source of drinking water	-0.0139***
water_2	1=hand-pump; 0=otherwise	-0.0392***
water_5	1=river/spring/tanker/others; 0=otherwise	0.1326***
workratio_adult	No. of adults employed (15-65 yrs old)/HH size	0.4167***
_cons	Constant	8.1329***

In our analysis, we implemented a hierarchical linear model, specifically a mixed-effects framework, to account for the complex nested structure of our data. This approach, which is particularly well-suited for small area estimation scenarios, was fitted using the REML method. REML is preferred in this context for its ability to produce unbiased estimates of variance components in unbalanced designs. The model's computational efficiency was notable, with the Expectation-Maximization (Beck, Demirgüç-Kunt, & Levine, 2007) algorithm converging rapidly after a single iteration. This swift convergence suggests a well-specified model and a stable solution, indicating that the algorithm quickly identified the optimal parameter estimates. The log-restricted likelihood value of -3376.5356 serves as a metric for assessing model fit. While the negative value is typical in statistical modeling, it provides a baseline for potential model comparisons. More importantly, the Wald chi-squared test yielded a p-value of 0.0000, providing strong evidence for the overall statistical significance of the model. This result indicates that the fixed effects included in the model collectively contribute meaningful explanatory power. Our dataset comprised 22,677 observations[55] nested within 1,802 primary sampling units (PSUs), reflecting the hierarchical nature of the sampling design. The average cluster size is 12.4 observations per PSU, with a range from 1 to 16. This variation in cluster size underscores the importance of using a mixed-effects approach, which can effectively handle such unbalanced designs. The model's structure accounts for both fixed effects, which capture the overall relationships in the population, and random effects, which allow for PSU-specific deviations from these population-level trends. This dual-level modeling

[55] A total of 2,132 observations are excluded in the process from the survey data (HIES 2018-19) for poverty.

strategy enables us to partition variance between and within PSUs, providing a more nuanced understanding of the data's hierarchical structure. Examination of Table 5.1 reveals crucial insights into the statistical significance of our model parameters. The coefficient estimates derived through the REML method demonstrate robust statistical significance at the 5% level of significance across most variables. This high level of significance underscores the relevance and predictive power of the selected covariates in our model. It is noteworthy that the province variable, while not meeting the conventional threshold for statistical significance, has been retained in the model as a control variable. This inclusion aligns with contemporary statistical practices that recognize the importance of accounting for potential spatial heterogeneity, even when the direct effect may not reach traditional significance levels.

The incorporation of a PSU Identity-based random intercept in our hierarchical model allows for the accommodation of unobserved heterogeneity across Primary Sampling Units (PSUs). This methodological approach enables the model to capture variations in baseline effects among different PSUs, thereby accounting for cluster-level differences that may not be explicitly measured. In our multilevel analysis, the inter-cluster heterogeneity is quantified by the standard deviation of random intercepts (sd(_cons)) at the PSU level, estimated at 0.0988 with a 95% confidence interval of (0.0936, 0.1043). This parameter elucidates the extent of variability between PSUs. The narrow confidence bounds indicate high estimation precision, while the non-zero value substantiates significant inter-PSU differences, thus validating our hierarchical modeling approach. Similarly, the intra-cluster variability is represented by the residual standard deviation (sd(Residual)), calculated at 0.2685 with a 95% confidence interval

of (0.2660, 0.2711). This metric captures the within-PSU variation. The minimal standard error (0.0013) associated with this estimate underscores its high precision. The Likelihood Ratio (LR) test, yielding a highly significant p-value (0.0000), provides robust evidence supporting the superiority of the mixed-effects model over a simpler linear alternative. This statistical confirmation underscores the necessity of incorporating random effects to adequately address the hierarchical data structure. By simultaneously accounting for individual-level and cluster-level variations, our mixed-effects model enhances the accuracy and reliability of parameter estimates. The significant random effects underscore the substantial variability existing between PSUs, further validating the importance of this multilevel approach in improving estimation precision.

To further validate our methodological approach, we conducted a comprehensive simulation study comparing Direct estimates with Census Empirical Bayes (EB) Model-based estimates. Figure 5.1 illustrates the MSE of poverty estimates providing a visual representation of their relative performance across varying sample sizes within domains.

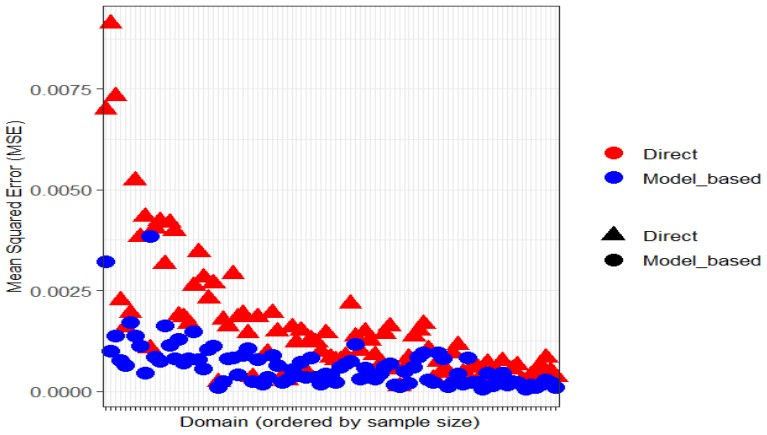


Figure 5.1 **MSE of Poverty Estimates at District level**

The domains in our analysis are arranged in ascending order of sample size, revealing a striking pattern in estimation precision. In domains characterized by smaller sample sizes, the MSE associated with Direct estimates exhibits inflated values (0.00159 vs 0.00063) compared to their Census EB model-based counterparts. This disparity displays the limitations of traditional Direct estimation techniques when applied to areas with limited data size. As we progress towards domains with larger sample sizes, an interesting convergence emerges. The MSE values of Direct and model-based estimates begin to align more closely. However, MSE based on Census EB estimates are still at lower side. This observed trend has profound implications for small area estimation in the context of poverty assessment. The results unequivocally demonstrate the superior reliability of Census EB model-based estimates, particularly in scenarios where direct sampling is constrained. This finding is especially pertinent for policy-makers and researchers focused on estimating poverty rates at granular administrative levels, such as districts, where sample sizes are often limited. The robustness of the Census EB methodology in domains with smaller sample sizes can be attributed to its ability to borrow strength across areas and integrate auxiliary information effectively. This capability enables the production of more stable and precise estimates in data-scarce environments, a critical advantage when dealing with the inherent challenges of small area estimation.

To further corroborate the results, a box plot presentation of MSE for direct and census EB model-based estimates is depicted in below mentioned figure 5.2

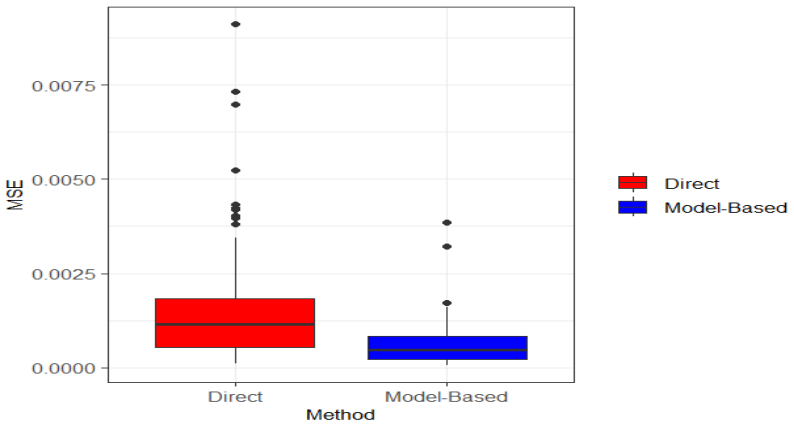


Figure 5.2 **MSE of Poverty Estimates at the District level**

Source: Author's calculations

Analysis from the plot indicates that the census EB model-based estimates demonstrate higher accuracy overall, as evidenced by a substantially lower median MSE (0.00046 vs 0.00147). Additionally, the model-based method exhibits greater precision and consistency in its estimates, reflected in a noticeably smaller interquartile range of MSE values. Conversely, the direct estimation method yielded more varying and potentially less reliable estimates. This conclusion is supported by the wider dispersion of MSE values observed for the direct estimation method, characterized by a higher median MSE and a broader range of values overall.

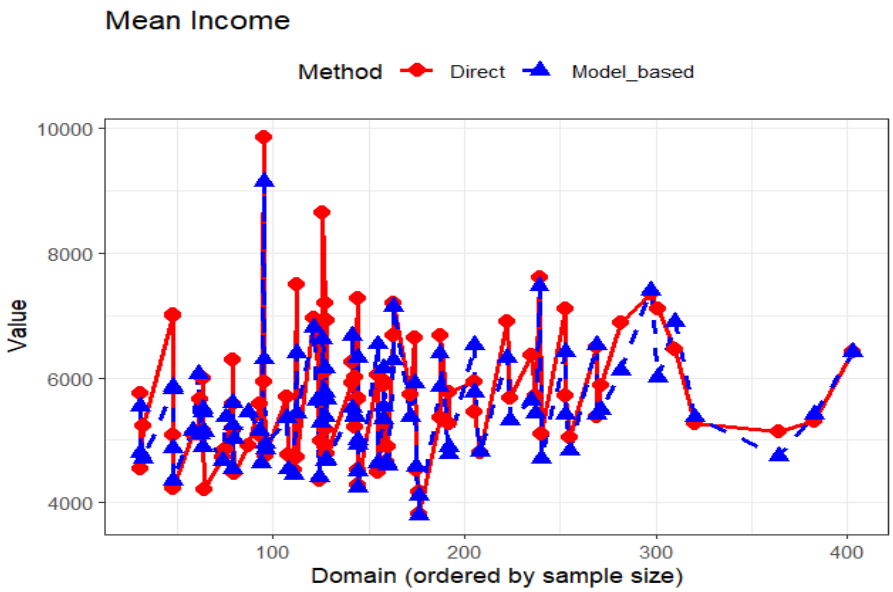


Figure 5.3 **Per equivalent Adult Expenditure at District level**

Source: Author's calculations

This visualization underscores the differential performance of direct estimates and Census EB model-based estimates when applied to per-equivalent adult expenditure data. The domains characterized by smaller sample sizes, we observe a pronounced difference between the direct estimates and their Census EB model-based counterparts. This heightened variation in smaller domains highlights the challenges associated with traditional direct estimation techniques when applied to areas with limited sample size. Conversely, as we examine domains with larger sample sizes, a notable convergence emerges between the two estimation methods. The reduced variation in these data-rich domains suggests that both approaches tend to yield more consistent results when supplied with ample observational data. This observed trend has significant implications for small-area estimation, particularly in contexts where sample sizes vary considerably across domains. The enhanced stability of Census EB estimates in domains with limited samples shows the method's capacity to mitigate the effects of

data scarcity through the incorporation of auxiliary information and borrowing strength across areas. Our findings strongly support the adoption of Census EB methodology as a superior alternative to direct estimation, especially in scenarios where sample sizes are constrained. This approach offers a more reliable framework for estimating mean income based on per equivalent adult expenditure, particularly in domains where direct sampling may be insufficient or impractical.

Table 5.2 below displays the parameter estimates based on the REML method When food insecurity, is a dependent variable.

Table 5.2 *Restricted Maximum Likelihood Model estimates*

Dependent Variable: Kilo Calories per Capita per Day		
Variable	**Description**	**Coefficient**
dryer	1 = household owns Dryer; 0=otherwise	-0.0279***
edu	Educational attainment	0.0159***
edu_1	1=No Schooling; 0=otherwise	0.0070**
geaser	1 = household owns Geaser; 0 otherwise	0.0671***
highestedu_2	1=High Education Primary (grade 1-5); 0=otherwise	-0.0205***
highestedu_4	1= High Education Lower secondary (grade 9-10);	-0.0126***
internet	1=HH has an internet connection; 0=otherwise	0.0277***
iron	1 = household owns Iron; 0 otherwise	0.0157***
language_new_4	1=pashto; 0=otherwise (Lang. Interview)	-0.0256**
marital_2	1=married; 0=otherwise	-0.0647***
microwave	1 = household owns Microwave; 0 otherwise	0.0661***
ownership_2	1=rented house; 0=otherwise	-0.0204***
prov_2	Punjab	-0.1028***
prov_3	Sindh	-0.0371***
prov_4	Balochistan	-0.0334**
resibuilding	1 = household owns Residential Building; 0=otherwise	0.022***
roof	1=pacca roof; 0=otherwise	0.0227***
room	Number of Rooms in a House	-0.008***
table	1 = household owns Table; 0=otherwise	0.0314***
toilet_1	1=flush toilet; 0=otherwise	-0.0261***
toilet_detail	Kind of toilet facility household use	0.0091***
toilet_detail_4	1=flush connected to pit, 0=otherwise	0.034***
toilet_detail_7	1=dry pit latrine, 0=otherwise	-0.0665***
ups	1 = household owns UPS; 0=otherwise	0.0454***
urban	1=urban; 0=rural	-0.0415***
wall	1=pacca wall; 0=otherwise	0.021***
washingmachine	1 = household owns Washing Machine; 0=otherwise	0.0089**
water_5	1=river/spring/tanker/others; 0=otherwise	0.0182***
workratio_adult	No. of adults employed (15-65 yrs old)/HH size	0.4187***
_cons	Constant	7.329***

Our analysis based on food insecurity as a dependent variable, employed the Expectation-Maximization (Beck et al., 2007) algorithm, a numerical optimization technique, which rapidly converged after a single iteration, indicating that the model

has reached a stable solution. The model considered 23,108 observations[56] after removing the outliers. The log-restricted likelihood value of 2627.5765 serves as a metric for assessing model fit, with higher values indicating a better alignment with the data. Furthermore, the Wald chi-squared test yielded a p-value of 0.0000, signifying that the overall model is statistically significant. This suggests that the fixed effects in the model differ significantly from zero, implying that the model is capable of capturing a substantial proportion of the variability in the data. The results of the REML analysis reveal that most of the variables exhibit statistically significant coefficients at the 5% level, indicating a strong relationship between the variables and the outcome including the province variable, included as a control factor, demonstrates a statistically significant effect.

The standard deviation of the random intercepts sd(_cons) at the PSU level is estimated at 0.0848, with a 95% confidence interval of (0.0806, 0.0891). This parameter quantifies the between-PSU variability. The relatively narrow confidence interval suggests a good level of precision in this estimate. The non-zero value indicates significant variation across PSUs, justifying the use of a multilevel model. The residual standard deviation (sd Residual) is estimated at 0.2043, with a 95% confidence interval of (0.2024, 0.2063). This represents the within-PSU variability. The small standard error of (0.001) indicates high precision in this estimate. The magnitude of this value relative to the PSU-level variance suggests that while there is substantial within-PSU variation, the between-PSU differences are also meaningful. The LR test statistic (chibar2(01) = 1356.3) with a p-value of 0.0000 strongly supports the use of the mixed-

[56] A total of 1701 observations are excluded in the process from the survey data (HIES 2018-19) in case of food insecurity.

effects model over a simpler linear model. This confirms that accounting for the hierarchical structure in the data significantly improves the model fit. The use of Restricted Maximum Likelihood for parameter estimation is appropriate here, as it provides unbiased estimates of variance components, which is crucial in small-area estimation.

The analysis of food insecurity estimates, quantified through daily per capita kilocalorie consumption, reveals a notable pattern when comparing direct estimates and census-based empirical Bayes (EB) model-based estimates (Figure 5.4).

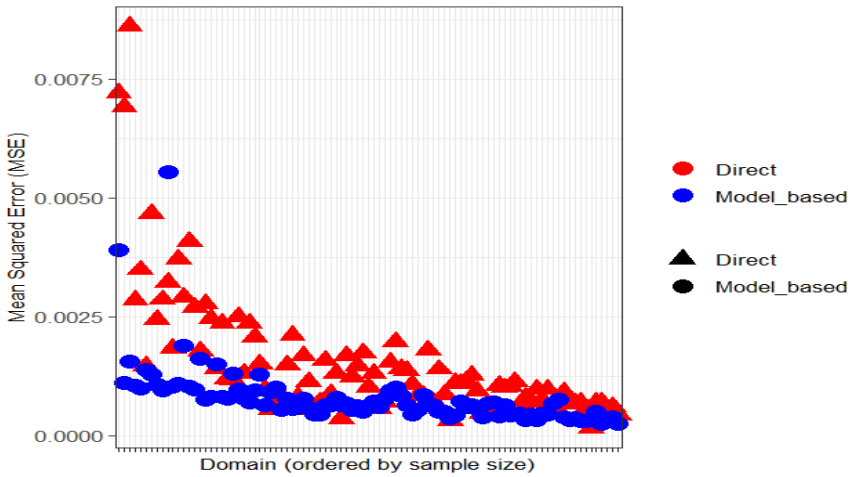


Figure 5.4 **MSE of Food Insecurity Estimates at District level**

Source: Author's calculations

Consistent with previous findings on poverty estimates, the results demonstrate that MSE of direct estimates tend to be higher for districts with smaller sample sizes, while the census EB model-based estimates exhibit lower MSE values. Interestingly, as the sample size increases for certain districts, there is a convergence between the direct and census EB model-based estimates. This convergence suggests that both methods tend

to align more closely when working with larger datasets. Overall, the census EB model-based estimates demonstrate superior reliability, as evidenced by their consistently lower MSE values. This finding underscores the potential advantages of employing advanced statistical techniques, such as empirical Bayes modeling, in conjunction with census data to generate more accurate and robust estimates of food insecurity across diverse geographical regions.

The analysis of the box plot (figure 5.5) indicates that the model-based estimates using the census EB model demonstrate superior performance in terms of MSE.

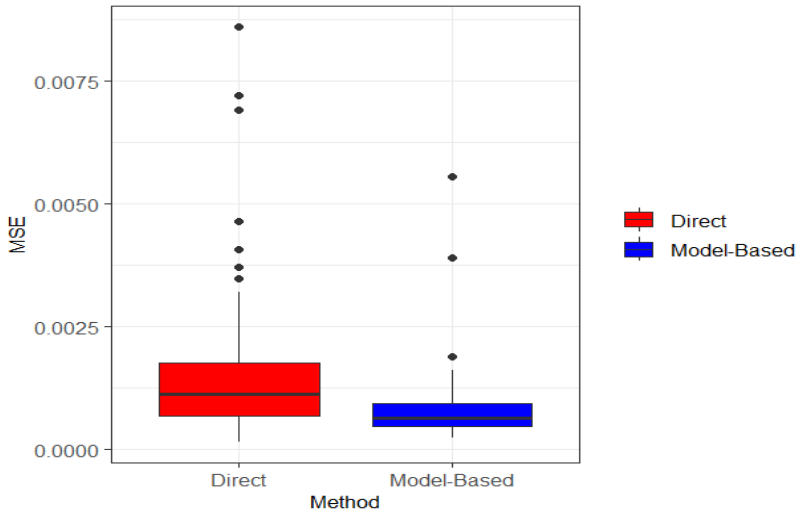


Figure 5.5 **MSE of Food Insecurity Estimates at District level**

Source: Author's calculations

Notably, these estimates show a significantly lower median MSE, reflecting improved overall accuracy. Additionally, the interquartile range of MSE values for the model-based approach is significantly narrower, highlighting greater precision and consistency. In comparison, the direct estimation method has a wider spread of MSE

values, with a higher median and longer whiskers, indicating increased variability and potentially less reliable estimates.

Figure 5.6 below clearly demonstrates that the spread of direct estimates is very high compared to census EB (model-based) estimates, particularly within the domains where the sample size is small.

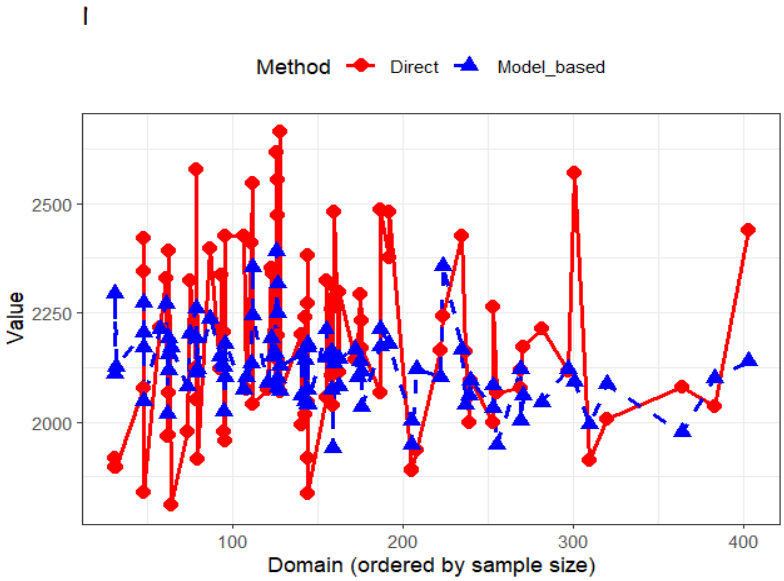


Figure 5.6 **Kilo Calories per Capita per Day at District level**

Source: Author's calculations

The findings precisely demonstrate the superiority of the census-based estimation approach (model-based) over the direct method, particularly in domains with limited sample sizes. This superiority is consistently observed across both scenarios, where poverty and food insecurity serve as the variables of interest.

5.2 Comparative Assessment of the Census EB and ELL Approach

Our comprehensive analysis has progressed through several methodological comparisons to identify the most reliable approach for poverty and food insecurity

mapping at the district level in Pakistan. This section aims to synthesize our findings and conduct a final comparative assessment between the Census Empirical Bayes (EB) approach and the Elbers, Lanjouw, and Lanjouw (ELL) methodology.

In Chapter 4, we conducted an in-depth examination of the ELL methodology, benchmarking it against an extended ELL approach as proposed by (Nguyen et al., 2018). Our analysis revealed the superiority of the standard ELL method over its extended counterpart. Subsequently, we validated the ELL model-based estimates through a comparative analysis with direct estimates. This validation process substantiated the enhanced reliability of ELL model-based estimates, particularly in domains characterized by limited sample sizes.

Chapter 5 (Section 5.1) introduced and elaborated on the Census EB method. Our rigorous evaluation demonstrated that model-based estimates derived from the Census EB approach exhibited greater reliability for poverty and food insecurity than direct method estimates. This advantage was especially pronounced in domains with constrained sample sizes, underscoring the method's robustness in data-scarce environments.

Building upon these findings, the current section aims to conduct a critical comparison between the Census EB approach and the ELL methodology. This final assessment is crucial for determining the optimal method for constructing high-resolution poverty and food insecurity maps at the district level in Pakistan. By juxtaposing these two advanced small-area estimation techniques, we seek to identify which approach offers the most precise, stable, and reliable estimates across various domain sizes and data availability scenarios.

5.2.1 Comparison of Poverty Estimates Via CEBP and ELL Approach

In our study, we conducted a comparative analysis of two Small Area Estimation techniques, the CEBP method, which is an extension of the unit-level model proposed by (Molina & Rao, 2010), and the ELL method (2003), widely used for poverty mapping. The reliability of these methodologies is assessed using MSE as the primary evaluation metric. To account for potential variability between different aggregation levels, we present our findings at both the district and PSU scales. This multi-level approach provides a more comprehensive understanding of the methods' performance under various spatial resolutions. Based on the comparative MSE analysis, we identified the most statistically robust method for poverty mapping in our study area. Using this optimal technique, we proceeded further, and generated poverty maps, to identify the most vulnerable districts within our research framework.

The comparative analysis of MSE for poverty estimates at the district level is visualized in Figure 5.7. The districts are arranged in ascending order based on the MSE values derived from the ELL method.

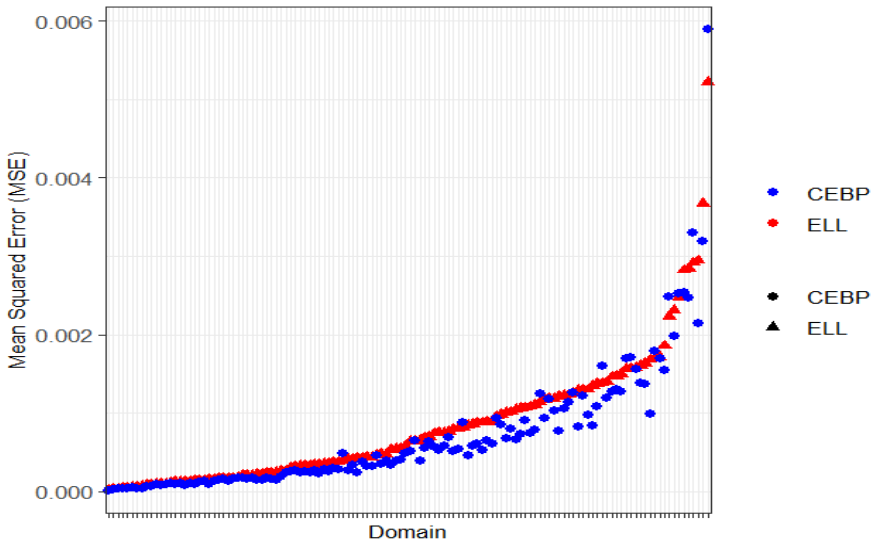


Figure 5.7 **MSE of Poverty Estimates at District level**

Source: Author's calculations

This graphical representation reveals a compelling pattern: in districts where the MSE of poverty estimates is relatively low, there is a close alignment between the MSE values of both methodologies under examination. However, as the MSE increases, a divergence becomes apparent between the two approaches. Notably, the CEBP method consistently demonstrates lower MSE values (0.00072 vs 0.00090) compared to the traditional ELL approach across the spectrum of districts. This trend becomes more pronounced in districts with higher overall MSE values, where the gap between the two methodologies widens significantly. These findings provide strong evidence for the superior performance of the CEBP method in estimating poverty at the district level. The consistently lower MSE values associated with CEBP suggest that this approach yields more precise and reliable poverty estimates compared to the ELL method.

Figure 5.8 presents a comparative box plot analysis of the MSE associated with the CEBP and ELL methods, further corroborating our earlier findings.

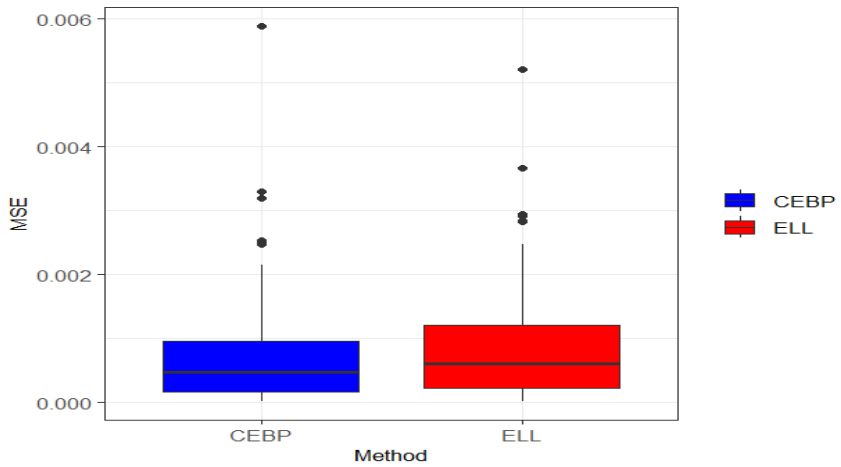


Figure 5.8 **Box Plot of MSE for Poverty Estimates at District level**

Source: Author's calculations

The visualization demonstrates a clear distinction between the two approaches in terms of their estimation precision. The CEBP method exhibits notably lower MSE values (0.00047 vs 0.00072) across the distribution, as evidenced by the position of its box plot relative to that of the ELL method. Moreover, the interquartile range of the CEBP's MSE estimates is substantially more compact, indicating a higher degree of consistency in its predictions across different districts. A key indicator of the CEBP's superior performance is the position of its median MSE, which is significantly lower than that of the ELL method. This lower median suggests that, on average, the CEBP approach yields more accurate poverty estimates across the majority of districts in our study area. The reduced spread of the CEBP's MSE distribution, as shown by the shorter whiskers and more condensed box, further underscores its reliability. This tighter distribution implies that the CEBP method produces more stable and dependable estimates, with fewer extreme outliers compared to the ELL approach. In summary, the box plot analysis provides compelling visual evidence of the CEBP method's enhanced

performance in poverty estimation. The combination of lower overall MSE values, a more compressed distribution, and a lower median MSE collectively demonstrate the CEBP's superior precision and consistency in poverty mapping at the district level.

To address potential variations arising from different aggregation levels, we extended our analysis to compare the MSE of poverty estimates at the PSU level, the smallest area considered in our study. This additional examination serves to validate and reinforce our district-level findings.

Figure 5.9 illustrates a comparative box plot analysis of MSE values for both the CEBP and Elbers, Lanjouw, and Lanjouw (ELL) methods at the PSU level.

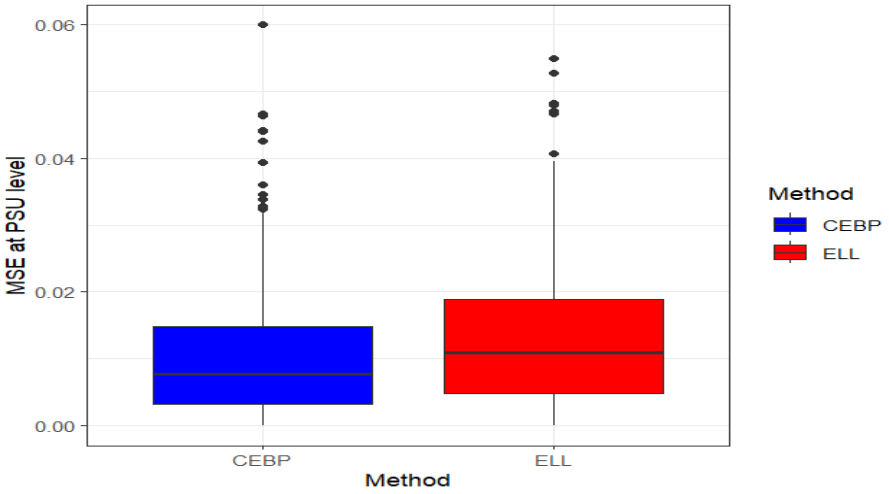


Figure 5.9 **Box Plot of MSE for Poverty Estimates at PSU level**

Source: Author's calculations

The visualization reveals a consistent pattern with our district-level results, further substantiating the superiority of the CEBP approach. At the PSU level, the CEBP method continues to demonstrate lower MSE values across the distribution. The box plot for CEBP exhibits a more compact interquartile range and shorter whiskers, indicating a higher degree of precision and consistency in its estimates compared to the

ELL method. A critical observation is the notably lower median MSE (0.0077 vs 0.011) for the CEBP method at the PSU level. This lower median underscores the CEBP's ability to produce more accurate poverty estimates across a majority of PSUs, mirroring its performance at the district level. The compressed distribution of MSE values for the CEBP method, as evidenced by the tighter box plot, suggests a reduction in estimation variability across different PSUs. This consistency is particularly valuable when working with smaller geographical units, where local variations can significantly impact poverty estimates. In conclusion, the PSU-level analysis corroborates our district-level findings, providing robust evidence for the CEBP method's superior performance in poverty estimation. The consistently lower MSE values, more compact distribution, and lower median MSE at both district and PSU levels collectively affirm the CEBP approach as the more reliable and robust methodology for poverty mapping in our study area. These results strongly support the adoption of CEBP for precise and dependable small-area poverty estimation.

Figure 5.10 presents a comparative visualization of the poverty headcount ratios derived from the CEBP and Elbers, Lanjouw, and Lanjouw (ELL) modeling approaches. This analysis serves as the culmination of our methodological comparison, leveraging the insights gained from our comprehensive MSE assessments.

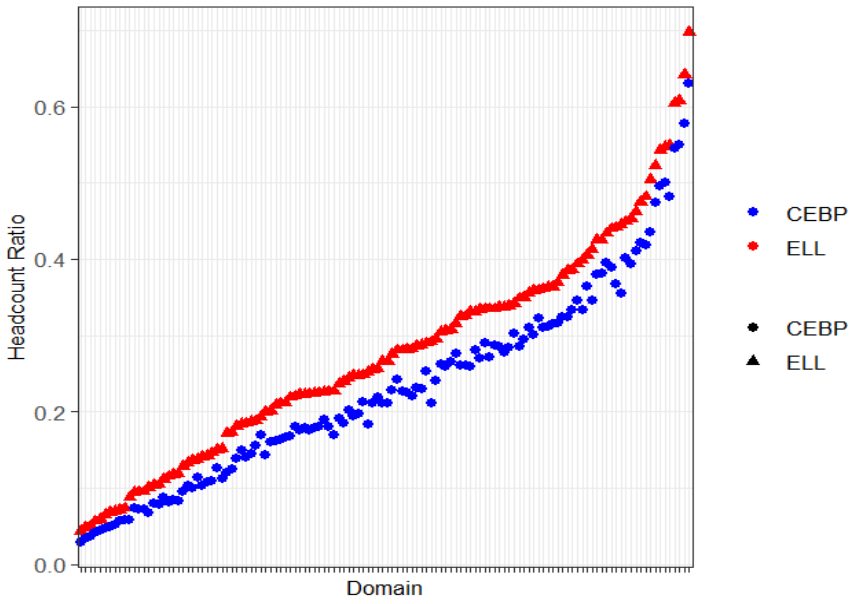


Figure 5.10 **Poverty Headcount Estimates at the District level**

Source: Author's calculations

The poverty estimates generated by the CEBP method, which our previous analyses identified as the superior approach, demonstrate a high degree of reliability and precision. These estimates provide a nuanced and accurate representation of poverty distribution across the study area. A notable finding from this comparison is the tendency of the ELL method to overestimate poverty levels relative to the CEBP approach. This overestimation is not uniform across all areas but exhibits a pattern correlated with the magnitude of MSE. Specifically, the discrepancy between the two methods' estimates becomes more pronounced in areas characterized by higher MSE values. This divergence in estimates, particularly in high-MSE regions, underscores the importance of methodological choice in poverty mapping. The CEBP method's more conservative estimates, coupled with its consistently lower MSE values, suggest that it provides a more accurate reflection of the true poverty landscape. The observed pattern of increasing divergence between CEBP and ELL estimates in areas with higher MSE

values further validates our earlier conclusions about the CEBP method's superior performance. It indicates that the CEBP approach is particularly adept at maintaining estimation accuracy even in challenging areas where traditional methods like ELL tend to lose precision.

Table 5.3 displays the national poverty estimates based on CEBP and ELL modeling approaches and compares them with the direct estimates.

Table 5.3 *Comparison of Poverty Model-based Estimates with Direct Estimates*

Provincial/ National	Poverty CEBP Model Estimates	Poverty ELL Model Estimates	Poverty Direct Estimates
KPK	0.20	0.25	0.28
Punjab	0.12	0.15	0.16
Sindh	0.20	0.24	0.24
Balochistan	0.33	0.37	0.42
Pakistan	0.16	0.20	0.21

Source: Author's Calculation

As presented in Table 5.3, the poverty estimates based on the CEBP approach are the most reliable. The results indicate that actual poverty based on ELL[57] and direct poverty estimates are somehow over-estimated, which is aligned with the literature (Corral et al., 2022). However, if the user is interested in poverty estimates that are more closely aligned with national estimates then ELL poverty estimates are desirable.

In conclusion, this analysis of poverty headcount ratios not only reaffirms the CEBP method's superiority but also highlights the practical implications of methodological choices in poverty estimation. The more reliable estimates provided by the CEBP

[57] The user may use the ELL method for Poverty Mapping in Pakistan if the purpose is to get Poverty estimates aligned more closely to HIES direct estimates.

approach offer a solid foundation for targeted policy interventions and resource allocation, potentially leading to more effective poverty alleviation strategies.

Our comprehensive analysis has conclusively demonstrated the superior performance of the CEBP method over the traditional Elbers, Lanjouw, and Lanjouw (ELL) approach. The CEBP method consistently yielded more reliable and precise poverty estimates across various spatial scales and evaluation metrics. Building upon these findings, we proceeded to leverage the CEBP method's robustness to conduct a detailed poverty mapping exercise. This crucial step aims to provide a high-resolution assessment of poverty distribution across Pakistan at the district level, as well as to identify the most economically vulnerable districts within the region.

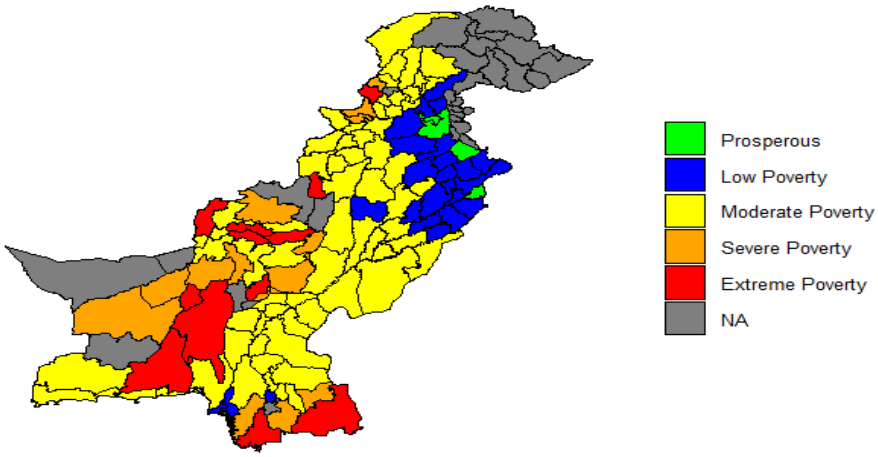


Figure 5.11 **Poverty Mapping at the District level in Pakistan**
Source: Author's calculations

Figure 5.11 presents a visual representation of the poverty landscape in Pakistan, derived from our application of the CEBP methodology. This poverty map offers a

nuanced and spatially explicit depiction of poverty incidence across the districts, capitalizing on the enhanced accuracy and reliability of the CEBP estimates. The poverty map visualization highlights a concentration of extreme poverty primarily in Balochistan, followed by Sindh, while Punjab and KPK emerge as the region with comparatively lower poverty rates. The analysis further identifies 16 districts experiencing extreme poverty conditions including Sheerani, Awaran, Barkhan, Khuzdar, Killa Abdullah, Killa Saifullah, Shaheed Sikandarabad, Ziarat, Harnai, Duki, Mohmand, Kalat, Tharparkar, Sujawal, Nasirabad and Badin. These districts represent areas of critical concern, requiring immediate and targeted interventions to address severe economic deprivation. Furthermore, there are 16 districts categorized under severe poverty: Kharan, Tando Muhammad Khan, Washuk, Dera Bugti, Nuski, Kachhi, Sohbatpur, Jaffarabad, Umerkot, Thatta, Shaheed Benazirabad, Mirpurkhas, Bajur, Khyber, Orakzai and South Waziristan. While not at the extreme level, these districts face significant economic challenges and warrant focused attention in poverty alleviation strategies. Additionally, our analysis reveals four districts, Sanghar, Kech, Rajanpur, and Khairpur, which, although classified under moderate poverty, are perilously close to the severe poverty threshold. These districts require careful monitoring and preemptive interventions to prevent further economic deterioration. A comparative analysis of poverty classifications derived from the CEBP and ELL methods reveals notable consistencies and some key divergences in district-level poverty assessments. There is a high degree of concordance between the CEBP and ELL methods in identifying districts experiencing extreme poverty. Some exceptions are Badin, Barkhan, and Killa Saifullah, which the CEBP method classifies under

extreme poverty, while the ELL approach categorizes it under severe poverty. This consistency in extreme poverty identification across methodologies underscores the robustness of our findings for the most economically vulnerable areas. Both methodologies demonstrate substantial agreement in the classification of severely impoverished districts. Specifically, 16 districts are consistently categorized as experiencing severe poverty under both the CEBP and ELL approaches[58]. This alignment further validates the reliability of our poverty assessments for a significant portion of the study area. The analysis reveals notable discrepancies in the classification of four specific districts: Tando Muhammad Khan, Nuski, Shaheed Benazirabad and Mirpurkhas. For these districts, the CEBP method yields a classification of severe poverty, whereas the ELL approach categorizes them under moderate poverty. Similarly, provincial analysis reveals that the most deprived districts under both ELL and CEBP are:

Punjab: Rajanpur, Muzaffargarh, Dera Ghazi Khan, Bhakkar, Bahawalnagar, Mianwali, Rahim Yar Khan, Khanewal, Lodhran and Vehari.

Sindh: Tharparkar, Sujawal, Badin, Tando Muhammad Khan, Thatta, Umerkot, Shaheed Benazirabad, Mirpur Khas and Sanghar.

KPK: Mohmand, Bajur, Khyber, Orakzai, South Wazirastan, Tor Ghar, Buner and Hangu.

Balochistan: Sheerani, Awaran, Khuzdar, Shaheed Sikandarabad, Ziarat, Killa Abdullah, Harnai, Duki and Nasirabad.

[58] Please see Appendix C for detail of Poverty Status.

In conclusion, while both methods largely agree on the identification of the most impoverished areas, the CEBP approach appears to offer a more conservative and potentially more comprehensive assessment of poverty, particularly for districts at the margins of poverty categories. This nuanced understanding is crucial for developing targeted and effective poverty alleviation strategies.

5.2.2 Food Insecurity Estimates Via CEBP and ELL Approach

A comparative evaluation of two distinct Small Area Estimation methodologies is made: the CEBP approach, which builds upon the unit-level model initially proposed by (Molina & Rao, 2010), and the Elbers, Lanjouw, and Lanjouw (ELL) method, recognized for its application in poverty mapping since 2003. To evaluate the efficacy of these techniques, we employed MSE as our primary metric for comparison. To address potential variations that may arise across different levels of data aggregation, we presented our results at both the district level and the PSU scale. This multi-tiered analysis allows for a more refined understanding of each method's performance across various spatial resolutions. Our comparative analysis of the MSE reveals the statistically most robust technique for poverty mapping within the context of our study area. Leveraging this optimal method, we subsequently developed food insecure mapping that further highlights the most vulnerable districts, thereby enhancing the overall framework of our research.

The comparative assessment of MSE for food insecurity estimates at the district level is illustrated in Figure 5.12.

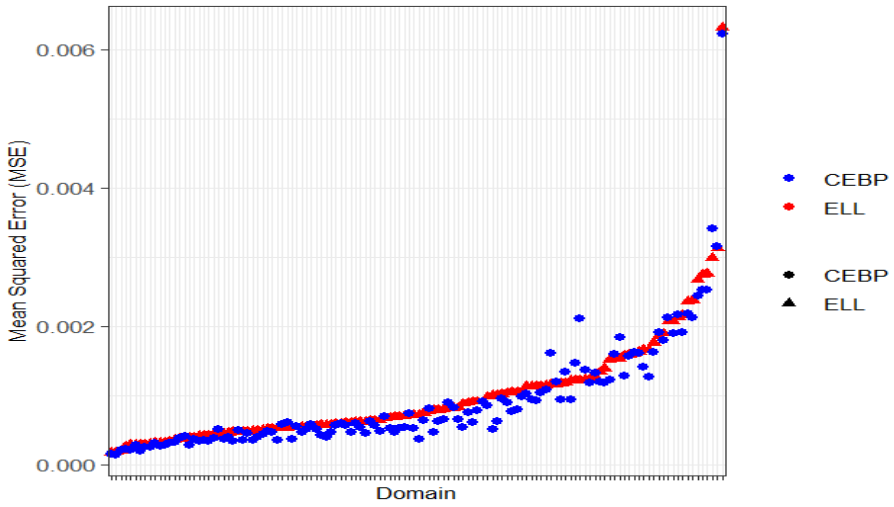


Figure 5.12 **MSE of Food Insecurity Estimates at District level**

Source: Author's calculations

The districts are organized in ascending order according to the MSE values obtained from the ELLmethod. The visual representation reveals a notable pattern in the estimation of food insecurity. In districts with a low MSE, both the CEBP and ELL methodologies demonstrate closely aligned performance. However, as the MSE values escalate, a distinct divergence emerges, with the CEBP method consistently yielding lower MSE values across the majority of districts. This consistent outperformance provides compelling evidence for the enhanced accuracy and reliability of the CEBP method in district-level food insecurity assessments, reinforcing its superior utility over the ELL approach. Figure 5.13 showcases a comparative box plot analysis of the MSE for both the CEBP and the Elbers, Lanjouw, and Lanjouw (ELL) methods, further validating our previous findings.

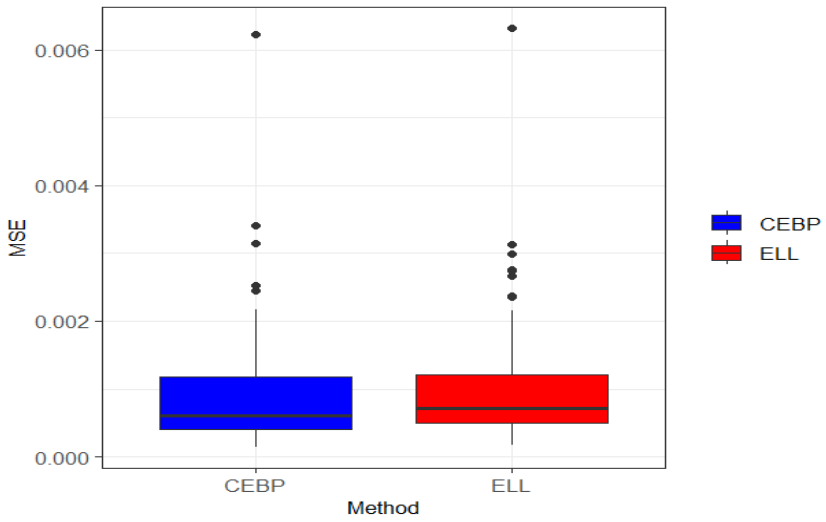


Figure 5.13 **Box Plot of MSE of Food Insecurity Estimates at District level**

Source: Author's calculations

This visualization clearly illustrates a significant difference in estimation precision between the two approaches. The CEBP method consistently demonstrates lower MSE values throughout its distribution, as indicated by the positioning of its box plot concerning that of the ELL method. A crucial indicator of the CEBP method's superior performance is the median MSE, which is markedly lower than that of the ELL method. This reduced median indicates that, on average, the CEBP approach provides more accurate estimates of food insecurity across the majority of districts analyzed. Overall, the box plot analysis offers compelling visual evidence that underscores the enhanced effectiveness of the CEBP method in estimating food insecurity.

Figure 5.14 presents a comparative box plot analysis of MSE values for both the CEBP method and the Elbers, Lanjouw, and Lanjouw (ELL) method at the PSU level.

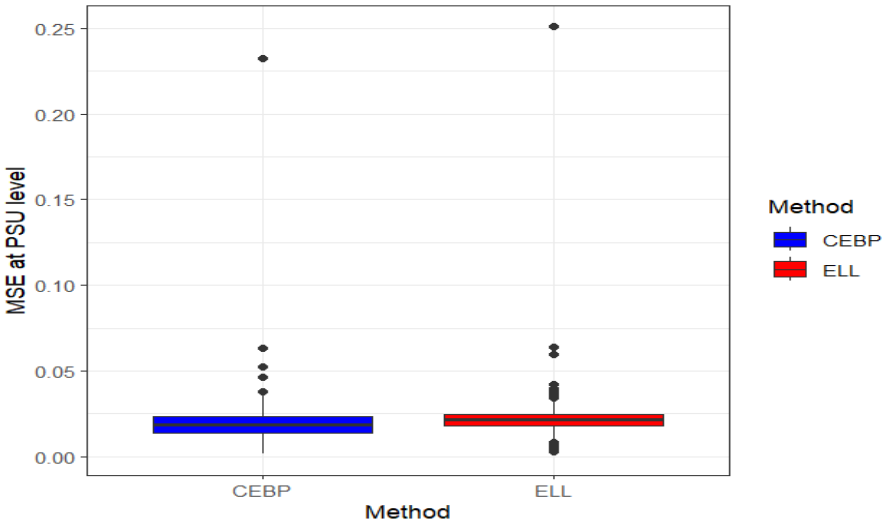


Figure 5.14 **Box Plot of MSE for Food Insecurity Estimates at PSU level**

Source: Author's calculations

This visualization reveals a consistent pattern that aligns with our findings at the district level, further reinforcing the superiority of the CEBP approach. At the PSU level, the CEBP method consistently exhibits lower MSE values throughout its distribution. A notable aspect of this analysis is the significantly lower median MSE associated with the CEBP method at the PSU level. This lower median highlights the capability of CEBP to deliver more accurate estimates of food insecurity across the majority of PSUs, reflecting its strong performance at the district level. In summary, the results at the PSU level corroborate our earlier district-level findings, providing robust evidence for the CEBP method's enhanced effectiveness of CEBP method in estimating food insecurity. These outcomes strongly advocate for the adoption of the CEBP approach as a reliable and precise method for small-area food insecurity estimation.

Figure 5.15 presents a comparative visualization of food insecurity headcount ratios derived from the CEBP method and the Elbers, Lanjouw, and Lanjouw (ELL) modeling approach.

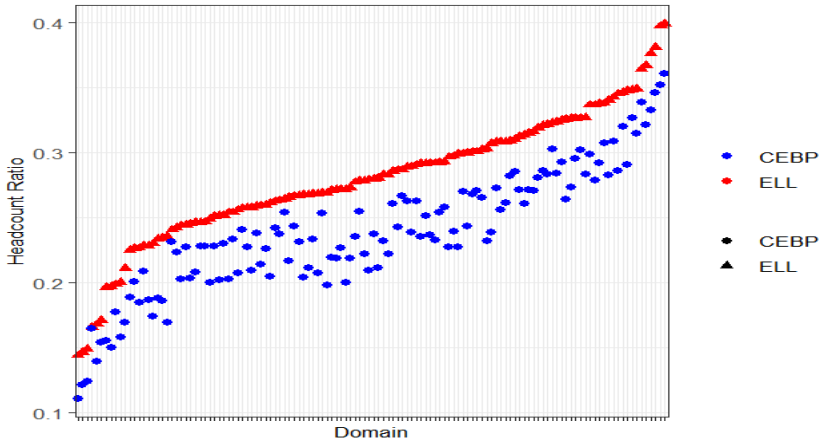


Figure 5.15 **Food Insecurity Estimates at District level**
Source: Author's calculations

This analysis serves as the culmination of our methodological comparison, utilizing insights gained from our comprehensive evaluations of MSE. The food insecurity estimates generated by the CEBP approach provide a nuanced and accurate representation of food insecurity distribution across the study area. A key finding from this analysis is the tendency of the ELL method to overestimate food insecurity levels compared to the CEBP approach. This overestimation is not consistent across all regions; instead, it reveals a pattern. The estimates are initially aligned but later they increase as the Food insecurity headcount across the domain increases. Instead, it reveals a pattern that correlates with the magnitude of MSE. Specifically, the differences in estimates between the two methods become more pronounced in areas with higher MSE values. The more conservative estimates under CEBP method, along with its consistently lower MSE values, indicate that it provides a more accurate portrayal of the true food insecurity status.

Our extensive study has shown that the CEBP method significantly surpasses the traditional Elbers, Lanjouw, and Lanjouw (ELL) approach in performance. The CEBP

method consistently generates more accurate and dependable estimates of food insecurity across various spatial dimensions and assessment metrics. Building on these insights, we employed the strengths of CEBP method to perform a comprehensive food insecurity mapping analysis. This vital step seeks to provide a detailed, high-resolution overview of food insecurity distribution at the district level within Pakistan, while also identifying the districts that are more vulnerable in terms of food insecurity in the region.

Figure 5.16 provides a visual depiction of the food insecurity landscape in Pakistan, based on our implementation of the CEBP methodology.

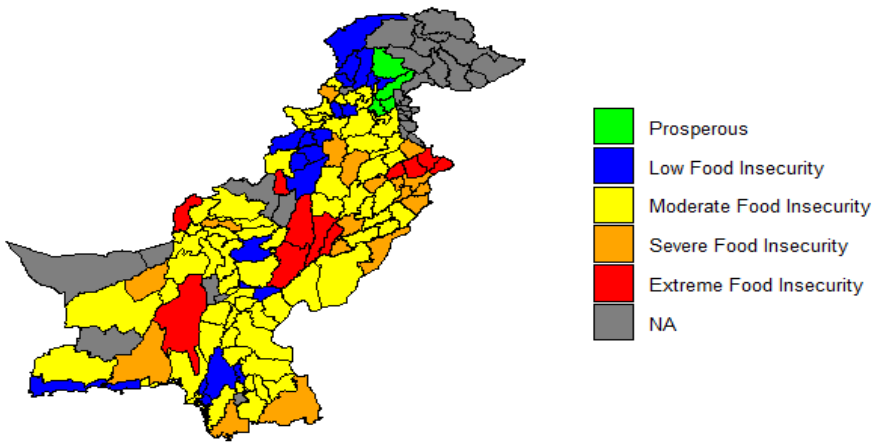


Figure 5.16 **Food Insecurity Mapping at the District level**
Source: Author's calculations

This food insecurity map offers a detailed and spatially explicit representation of food insecurity levels across various districts, leveraging the enhanced accuracy and

reliability of the CEBP estimates. The visualization reveals a concentration of severe food insecurity predominantly in, Punjab followed by Balochistan and Sindh, while KPK stands out as the region with relatively lower rates of food insecurity. Our analysis identifies 13 districts facing conditions of extreme food insecurity: Awaran, Kharan, Killa Abdullah, Khuzdar, Sheerani, Rajanpur, Dera Ghazi Khan, Narowal, Muzaffargarh, Sialkot, Multan, Hafizabad and Gujranwala. These districts are of critical concern, necessitating immediate and targeted interventions to combat food insecurity. Additionally, 21 districts are classified under severe food insecurity: Pishin, Harnai, Ziarat, Duki, Mohmand, Khanewal, Sheikhupura, Lodhran, Mianwali, Nankana Sahib, Kasur, Bahawalnagar, Chiniot, Gujrat, Khushab, Bhakkar, Attock, Vehari, Mandi Bahauddin, Tharparkar and Sujawal. Although these districts do not experience the most extreme levels, they still confront significant economic challenges that require focused attention in food insecurity alleviation efforts. Out of the 13 districts identified as experiencing extreme food insecurity, 10 are consistently recognized by both the CEBP and the Elbers, Lanjouw, and Lanjouw (ELL) methodologies: Killa Abdullah, Khuzdar, Sheerani, Rajanpur, Dera Ghazi Khan, Narowal, Muzaffargarh, Sialkot, Multan and Hafizabad. Furthermore, there are three districts classified under extreme food insecurity solely by the CEBP method that do not appear in the ELL analysis: Awaran, Kharan and Gujranwala. This difference underscores the enhanced reliability of the CEBP approach in generating more accurate estimates.

5.2.3 Summarizing Results

In this study, we employed the EBP approach to estimate poverty and food insecurity at the district level in Pakistan. This method, an extension of (Molina & Rao, 2010) work, is designed for small-area estimation. Our analysis involved 200 Monte Carlo simulations to generate final estimates of the variables of interest, coupled with 200 bootstrap simulations for each household to calculate these estimates' Mean Square Error. We then conducted a comparative analysis between the Census EBP approach, and the traditional ELL method. Our findings demonstrate that the Census EBP approach outperforms the ELL method in terms of reliability, and precision, as evidenced by lower MSE values for both poverty and food insecurity estimates. Our research identified several districts facing critical levels of poverty and food insecurity. Notably, Awaran, Khuzdar, Killa Abdullah, and Sheerani exhibited extreme levels in both poverty and food insecurity categories. Additionally, Duki, Harnai, Kharan, Ziarat, Mohmand, Sujawal, and Tharparkar are found to be either extremely or severely affected by poverty or food insecurity. These districts require immediate and focused attention from policymakers and stakeholders. The study underscores the crucial role of government intervention in addressing these pressing issues. Targeted efforts to alleviate poverty and food insecurity in these high-risk areas are essential for Pakistan to make progress towards achieving the SDGs.

Our findings provide valuable insights for policymakers to prioritize resources and implement effective strategies in the most vulnerable districts based on deprivation status, potentially leading to more equitable development across the country. Funds for social protection programs (e.g., cash transfers, subsidies, health coverage), major

infrastructure investments (e.g., schools, hospitals, roads, electrification), and economic opportunity (micro-credit access, market linkages, and agricultural extension services, etc) can now be allocated directly according to the specific needs category of each district (Extremely, Severely, and Moderately deprived, etc.). The granular data allows the government to use the severely and moderately deprived districts as an early warning system. By monitoring the trends in districts bordering the Extremely Deprived category, the government can do proactive interventions by implementing preventative measures to prevent them from sliding into the worst poverty tier, saving high costs and suffering. Finally, this work empowers the government to move from generalized policymaking to strategic, locale-specific, and evidence-based planning.

CHAPTER 6

APPLICATION OF POVERTY/INCOME DATA FOR SECTORAL ANALYSIS

The empirical evidence presented in Chapter 5 provides the foundational analytical framework for the subsequent investigation in this chapter. Specifically, these findings motivate and substantiate the quantitative assessment of the impact of rural transformation and agricultural credit instruments on the amelioration of rural poverty and augmentation of household income across Pakistan. Furthermore, this chapter scrutinizes the differential effects and spatial heterogeneity of the district-level transformation pace on these critical socioeconomic indicators.

In the preceding chapters, we elucidated the methodologies and analytical frameworks employed to generate high-resolution data on poverty and food insecurity at the granular level in Pakistan. Through rigorous evaluation, we identified the most robust and reliable method for data generation. This chapter pivots to the practical application of the newly generated poverty data[59], based on the findings of Chapter 5, with a particular emphasis on exploring the intricate interplay between poverty, income dynamics (quantified through monthly per equivalent adult expenditure), and the multifaceted process of rural transformation. Furthermore, we delve into the potential of formal agricultural credit as a catalyst for enhancing rural household incomes. This chapter is structured into two primary sections, each addressing a critical aspect of our research. Section 6.1 examines the nexus between rural transformation and poverty reduction at the district level in Pakistan. This analysis leverages spatial econometric

[59] The data used for analysis is generated from the most reliable Census EBP method of small area estimation.

techniques and district-level visualization to uncover patterns and trends in the evolving rural landscape. We explore how various dimensions of rural transformation, including agricultural modernization, and non-farm diversification correlate with changes in poverty indices across different geographic contexts. Section 6.2 investigates the role of formal agricultural credit in elevating the socio-economic status of rural households at the district level. Utilizing multivariate econometric modeling, we assess the impact of credit access on key welfare of households. Then through visualization, we check the districts with high agriculture credit and their role in enhancing the household income.

6.1 Regional Rural Poverty through the Lens of Rural Transformation

This comprehensive analysis explores regional rural poverty through the lens of rural transformation, structured into four interconnected parts that progress logically from conceptual framing to empirical analysis and synthesis of findings. We begin by establishing the theoretical underpinnings and contextual background of our investigation, elucidating the multifaceted nature of rural poverty and the potential of rural transformation as a catalyst for socioeconomic change. Following this, we provide a detailed exposition of our data acquisition strategy and analytical methodology. This encompasses a comprehensive overview of the diverse data repositories, including government publications, household surveys, and specialized reports. We explain how key variables, particularly those related to rural transformation, are operationalized and measured. The summary statistics provide an overview of the data distribution, and characteristics related to key variables are explained. This section also expounds on the econometric models employed, including panel data analysis, fixed or random effects

estimation, or the application of specialized statistical tests or procedures. The core of our analysis presents the results of our econometric investigation and offers a nuanced interpretation of these findings. The results of the econometric output, including coefficient estimates, standard errors, and measures of model fit are presented. A critical discussion of the results follows, contextualized within the broader literature on rural transformation and poverty alleviation. In the final segment, we synthesize the key findings from our analysis and extrapolate their implications for policy and practice. This section further provides a concise recapitulation of the main empirical results and their significance, followed by a discussion of how these findings can update rural development strategies and poverty alleviation programs. We identify knowledge gaps and potential avenues for further investigation, concluding with a reflection on the broader significance of our findings within the context of sustainable rural development and poverty reduction initiatives.

6.1.1 Introduction

The concept of RT has gained significant attention in recent years as a crucial process for poverty alleviation and economic development, particularly in agrarian economies. As defined by (J. Huang & Shi, 2021), RT encompasses agricultural transformation and the expansion of non-farm employment opportunities for rural labor. This multifaceted process is intricately linked to technological advancements and industrialization (Johnston, 1970; O'Brien, 1977), encompassing not only agricultural evolution but also the emergence of diverse livelihood and income-generating prospects in the rural non-farm sector (FAO, 2017; IFAD, 2016). In the context of global challenges such as rapid population growth, food insecurity, urbanization, and

extreme poverty, the focus of development strategies has increasingly shifted towards rural areas (Chigbu, 2015). This shift is particularly pertinent given the United Nations' projection that approximately 66% of the world's population will reside in urban areas by 2050 (K. Malik, 2014). The success of China's rural transformation initiatives over the past four decades, which have lifted millions out of poverty, serves as a compelling example of the potential impact of well-implemented RT strategies (FAO, 2017; IFAD, 2016; Huang and Shi, 2021). The significance of rural transformation in enhancing the collective GDP of agriculture-based economies cannot be overstated. The rate of poverty reduction is directly linked to the rate of agricultural commercialization and the speed of rural transformation. Public investment plays a crucial role in this process. Studies on successful rural transformation in China have revealed that the central focus is on farmers, with flexible land handling, market liberalization, and non-interference of the state supporting the transformation process (J. Huang & Shi, 2021). Technological innovation and education have also contributed to better and constant growth rates.

Farooq and Farah (2023) present a comprehensive framework of rural transformation concerning poverty reduction at the district level in Pakistan. The intricate relationship between rural poverty and rural transformation has been extensively studied by international organizations, with a consensus that institutionalizing rural transformation is a practical pathway for alleviating rural poverty (FAO, 2017; IFAD, 2016). J. Huang (2018) asserts that successful rural transformation is not accidental but rather a result of deliberate efforts in infrastructural development, off-farm employment creation, and technological and scientific advancements. Key determinants of rural

transformation include the diversification of off-farm employment sources and the shift from traditional grain-based crop production to commercial and high-value crops competitive in both national and international markets (Fan, 2019; Timmer, 2017). This transition has the potential to enhance farmers' incomes, but also to optimize agricultural input costs (Timmer, 2017). The development of the rural non-farm economy has been recognized as a crucial strategy for creating better export markets and stimulating the services sector. This functional aspect of the rural economy has accelerated structural transformation, reducing the total agricultural share in comparison to off-farm manufacturing and services (Jayne, Mather, & Mghenyi, 2010; Malik, 2008). The shift of resources from agriculture to secondary and tertiary sectors contributes to overall economic stabilization through poverty reduction.

The process of rural transformation faces numerous challenges, including land fragmentation, climate change impacts, lack of employment opportunities, extreme poverty, inadequate infrastructure, and increased urban migration, particularly in developing countries (Chigbu, 2015; Krause, Lotze-Campen, Popp, Dietrich, & Bonsch, 2013). In Pakistan, the process of business development in rural areas is relatively slow due to various socio-economic and structural constraints. Rapid rural transformation cannot be possible without business development in such areas, highlighting the crucial role of rural entrepreneurship in alleviating rural poverty and speeding up the whole process of rural transformation (Coad & Tamvada, 2012). Keeping in view these problems, this study has an important role in exploring how rural transformation is linked with rural poverty. What are the most vulnerable districts where the transformation speed is very slow concerning rural poverty alleviation? The objective

of this study is to explore the relationship between rural poverty and rural transformation. The analysis further leads to checking the speed of poverty reduction in relation to the RT through a typology analysis to identify the most vulnerable districts where the speed of rural poverty reduction and rural transformation is very slow. In the case of Pakistan, an agrarian economy since its inception, the link between rural poverty and rural transformation trends is particularly relevant. The gradual shift from a predominantly agricultural economy to a more diversified one is evident in the decrease in agriculture's contribution to the total GDP from 40% in 1960 to 19.2% in 2021 (PES, 2021)[60]. This transformation, though gradual, reflects a continuous process of structural change at the rural grassroots level. Therefore, the main focus of this section of the study is to examine the role of rural transformation in rural poverty reduction at the district level in Pakistan, considering the changing dynamics of rural non-farm employment and high-valued agriculture crops.

This comprehensive approach to rural transformation in Pakistan, encompassing shifts in economic focus, technological adoption, entrepreneurship, and women's participation, presents a complex but promising pathway for poverty reduction and economic development. The multifaceted nature of these changes underscores the need for targeted policies and investments that can effectively harness the potential of rural areas and their populations. As Pakistan continues to navigate these transformations, the interplay between rural poverty reduction and rural transformation will remain a critical area for research and policy intervention.

[60] Pakistan Economic Survey (PES) 2021-2022: https://www.finance.gov.pk/survey_2022.html

6.1.2 Data and Methodology

Drawing upon the extant literature, we conceptualize RT through two distinct metrics, RT1 and RT2, in alignment with the frameworks proposed by (J. Huang & Shi, 2021) and (Farooq & Farah, 2023). RT1 is operationalized as the proportion of High-Value Agriculture products (HVAP) to the total value of Agricultural products. HVAP encompasses a diverse array of agricultural products, including dairy, meat, aquaculture, horticulture, livestock, cotton, and sugar crops[61]. Conversely, RT2 is defined as the ratio of non-farm employment to total employment, where non-farm employment is calculated by subtracting agricultural employment from the aggregate employment figures. Similarly, we have a set of control variables (CV) that include rural household size, labor force participation which is defined as the proportion of individuals aged 10 years and older who are either currently employed or actively looking for employment, relative to the total population within that age demographic. This metric provides insight into the engagement levels of the working-age population in the labor market. Similarly, the Gross Enrollment Rate is an indicator that measures the overall participation in a given level of education. It is calculated by taking the total number of students enrolled in primary and middle school, regardless of their age, and dividing that number by the total population of the official age group for those levels (typically ages 5–9 for primary and 10–12 for middle). The Net Enrollment Rate for the primary education sector is calculated by dividing the number of enrolled primary children aged 5 to 9 years by the total number of children in that same age group. Female non-vulnerable job is the percentage of the employed female population (10

[61] For a detailed definition of RT1, see (Farooq & Farah, 2023).

years and above) that are either paid workers, employers, or self-employed in the non-agriculture sector, divided by the total employed female population. Table 6.1 provides the summary statistics of the variables considered for the analysis:

Table 6.1 *Summary Statistics*

Variables	Mean	Standard Deviation	Minimum	Maximum
Poverty Headcount	32.7651	14.2977	5.2689	64.5137
RT1	80.5364	12.5469	38.5342	99.4660
RT2	50.6657	18.8522	10.1465	91.2567
Household Size	7.2899	1.3530	4.3871	11.1907
Labor force Participation	35.1442	10.0723	14.0636	78.2947
Gross Enrolment Rate - Primary	81.5395	20.5202	31.0687	129.3978
Gross Enrolment Rate - Middle	47.8745	19.7173	7.2336	100.0000
Net Enrolment Rate - Primary	50.9633	12.8133	17.6064	82.2060
Non-Vulnerable Jobs - Female	37.0233	26.8598	0.3610	100.0000

The primary objective of this study is to utilize econometrically modeled poverty data to elucidate the intricate relationship between Rural Poverty and RT, with a particular focus on how the RT process catalyzes poverty alleviation. Our analytical framework extends beyond this initial investigation to examine the velocity of Rural Transformation vis-à-vis poverty reduction dynamics. Furthermore, we employ a rigorous Typology Analysis to identify and categorize districts based on their growth trajectories and poverty levels. This multifaceted approach allows for the identification of high-growth, low-poverty districts as well as areas characterized by persistent poverty, thereby facilitating targeted interventions and enhanced monitoring of poverty reduction initiatives through the lens of rural transformation.

This research employs a longitudinal district-level analysis, focusing on two temporal points: 2008 and 2019. The study leverages data primarily from the PSLM, a comprehensive national survey that provides a rich source of socio-economic indicators. However, it is noteworthy that the RT1 data is sourced from alternative repositories[62]. Our sampling frame encompasses 87 districts, which effectively represent 95 districts out of the total 126 districts as delineated in the PSLM 2019-20. The exclusion of 31 districts from the analysis is predicated on several factors like the predominantly urban character of certain districts, the absence of agriculture-centric regions, and instances of data insufficiency or inconsistency. To rigorously analyze the relationship between Rural Poverty and Rural Transformation, we employ a sophisticated multivariate econometric model. This model is designed to capture the complex interplay between our variables of interest while controlling for potential confounding factors. The general form of our econometric specification can be expressed as follows:

$$Rural\ Poverty_{it} = \lambda_1 RT1_{it} + \lambda_2 RT2_{it} + \beta_n CV_{it} + v_i + \eta_{it} \qquad (6.1)$$

Where i indicates the district i, t indicates the year t. Also, λ_1 and λ_2 are coefficients to the respective variables in above model, and βn represents coefficients of control variables in the model. Our methodological approach incorporates a series of rigorous diagnostic procedures to ensure the robustness and reliability of our econometric model. The analytical process is structured so that we initially employ the VIF to

[62] The data sources for RT1 are: the Directorate of Agriculture Punjab, the Ministry of Food Security and Research, the Agricultural Census Organization of the Government of Pakistan, the Provincial Crop Reporting Services, Provincial Agricultural Marketing Departments, Agriculture Marketing Information Services, the Pakistan Census of Livestock and Special Reports. These resources collectively contribute to the generation of data for RT1.

scrutinize the potential presence of multicollinearity among our predictor variables. This crucial step helps identify and mitigate any excessive correlation between independent variables, which could otherwise lead to inflated standard errors and unreliable coefficient estimates. Variables exhibiting VIF values exceeding a predetermined threshold of 7 are carefully evaluated for their theoretical importance and potential redundancy in the model. Following the multicollinearity assessment, we implement a sophisticated variable selection strategy utilizing stepwise forward and backward procedures. This dual-directional approach allows for a comprehensive exploration of the variable space, ensuring the identification of the most economical yet explanatory powerful set of predictors.

6.1.3 Results and Discussions

This section aims to apply poverty data generated via an indirect modeling technique to reveal the relationship with rural transformation. In this respect, a panel multivariate econometric modeling based on panel data is utilized to explore the relationship between rural poverty and rural transformation in Pakistan. Table 6.1 displays the results:

Table 6.2 *Impact of Rural Transformation on Rural Poverty – A Panel Fixed Effect Model*

Dependent Variable: Poverty Headcount (Rural)				
Variables	**Coefficient**	**Standard Error**	**T-Stats**	**P-Value**
RT1	-0.5247	0.1674	-3.13	0.0020
RT2	-0.3058	0.1229	-2.49	0.0150
Household Size	3.0273	0.9196	3.29	0.0010
Laborforce Participation	-0.4858	0.1220	-3.98	0.0000
Gross Enrolment Rate - Primary	-0.5110	0.1598	-3.2	0.0020
Gross Enrolment Rate - Middle	-0.2880	0.1053	-2.74	0.0080
Net Enrolment Rate - Primary	-0.6415	0.2158	-3.97	0.0000
Non-Vulnerable Jobs - Female	0.1322	0.0445	2.97	0.0040
Constant	103.3864	17.9086	5.77	0.0000
R - Squared	**0.80**	**F - Stat**	**40.44**	**0.0000**

This panel fixed effects model[63] explores the impact of rural transformation on rural poverty at the district level in Pakistan. Given the two-period panel, this approach is methodologically robust as it effectively removes time-invariant, district-specific unobserved heterogeneity, which is a primary source of omitted variable bias. By focusing on within-district changes over the decade, the model provides a reliable estimate of the relationship between changes in rural transformation and changes in poverty. In contrast, the pooled OLS specification is deliberately excluded, due to a high risk of omitted variable bias, which would compromise the reliability of the estimates. RT1 and RT2 are the main explanatory variables while all other explanatory variables serve to control various socioeconomic factors. The analysis reveals that both measures of rural transformation (RT1 and RT2) have statistically significant negative associations with rural poverty headcount, highlighting the role of rural transformation in poverty reduction. RT1 exhibits a negative relationship with poverty (β = -0.5247 and p-value < 0.01). This indicates that with an increase of 1% in RT1, the rural poverty

[63] The significant p-value (at 1% level of significance) of Hausman test suggest that fixed effect model is an appropriate choice for this panel analysis.

headcount is expected to decrease by approximately 0.52 percentage points, holding other factors constant. The substantial magnitude and statistical significance of this coefficient emphasize the importance of the aspects of rural transformation captured by RT1 in alleviating poverty. RT2 also demonstrates a significant negative association with poverty, albeit with a magnitude (β = -0.3058 and p-value < 0.05). This suggests that a 1% increase in RT2 is associated with a reduction of about 0.31 percentage points in the rural poverty headcount, ceteris paribus. The differential impacts of RT1 and RT2 suggest that various dimensions of rural transformation may have varying degrees of influence on poverty reduction. This finding highlights the complexity of rural transformation processes and their effects on socioeconomic outcomes. The model incorporates several control variables to isolate the effects of rural transformation indices. These controls, including household size, labor force participation, educational enrollment rates, and female employment in non-vulnerable jobs, all show statistically significant relationships with poverty. Their inclusion enhances the model's robustness and helps to mitigate potential omitted variable bias, thereby increasing confidence in the estimated effects of RT1 and RT2. The high R-squared value of 0.80 indicates that the model explains a substantial portion (80%) of the variation in rural poverty headcount across districts and over time. This strong explanatory power, combined with the highly significant F-statistic (40.44, p < 0.0001), suggests that the model is well-specified and provides the best fit for data. These findings provide support for the poverty-reducing potential of rural transformation in Pakistan[64]. The significant

[64] This analysis is based on limited two-period data and lacks a comprehensive exploratory analysis within specific districts to establish any causal link. Additionally, as the purpose of Chapter 6 is the application of data, a claim of causality cannot be made at this stage.

negative coefficients of both RT1 and RT2 suggest that policies and interventions targeted for promoting various aspects of rural transformation could be effective strategies for combating rural poverty. It is important to note that a proper stepwise model selection procedure is applied, which includes testing for non-linear relationships. The square terms of both RT1 and RT2 are statistically insignificant, indicating the absence of non-linear relationships between rural transformation (RT1 and RT2) and rural poverty. This finding strengthens the validity of the linear model specification, and it suggests that the poverty-reducing effects of rural transformation occur consistently and linearly across the observed range of RT1 and RT2 values.

Following the validation of results through econometric modeling techniques, our subsequent objective is to visualize the findings within the framework of RT and its impact on rural poverty alleviation. We aimed to assess the velocity of RT concerning rural poverty reduction at the district level in Pakistan. To quantify the pace of RT1, RT2, and the composite RT Index vis-à-vis rural poverty, we analyzed data from two temporal points: 2008 and 2019. The annual rate of change is computed based on these temporal benchmarks. We then employed scatter plot visualizations to juxtapose each RT metric (RT1, RT2, and the RT Index) against rural poverty indicators. In these visualizations, median values served as thresholds for the mean annual fluctuations in the variables under scrutiny. This approach allowed for a subtle interpretation of the dynamic interaction between rural transformation processes and poverty reduction trajectories across Pakistan's diverse districts. The scatter plot visualizations are structured to delineate four distinct quadrants, each representing a unique interplay between rural transformation metrics and poverty reduction dynamics. These quadrants

offer nuanced insights into district-level development trajectories. The High-Performance Quadrant encompasses districts exhibiting above-median poverty reduction coupled with above-median rural transformation indicators (RT1, RT2, or RT Index), signifying areas of comprehensive progress. Contrastingly, the Poverty-Focused Quadrant represents districts achieving above-median poverty reduction despite below-median rural transformation metrics, suggesting targeted poverty alleviation efforts. The Transformation-Centric Quadrant includes districts with below-median poverty reduction but above-median rural transformation indicators, indicating structural changes that have yet to translate into significant poverty reduction. Lastly, the Lagging Quadrant comprises districts demonstrating below-median performance in poverty reduction and rural transformation metrics, highlighting areas requiring intensified development interventions. This visualization is a powerful tool for spatially mapping the heterogeneous development trajectories across districts, offering insights into the complex interchange between rural transformation processes and poverty alleviation outcomes. Figure 6.1 presents a scatter plot illustrating the relationship between RT1 and rural poverty reduction. The figure's X-axis represents the mean annualized growth in RT1, while the Y-axis plots the corresponding average annual change in poverty, serving as a proxy for poverty reduction.

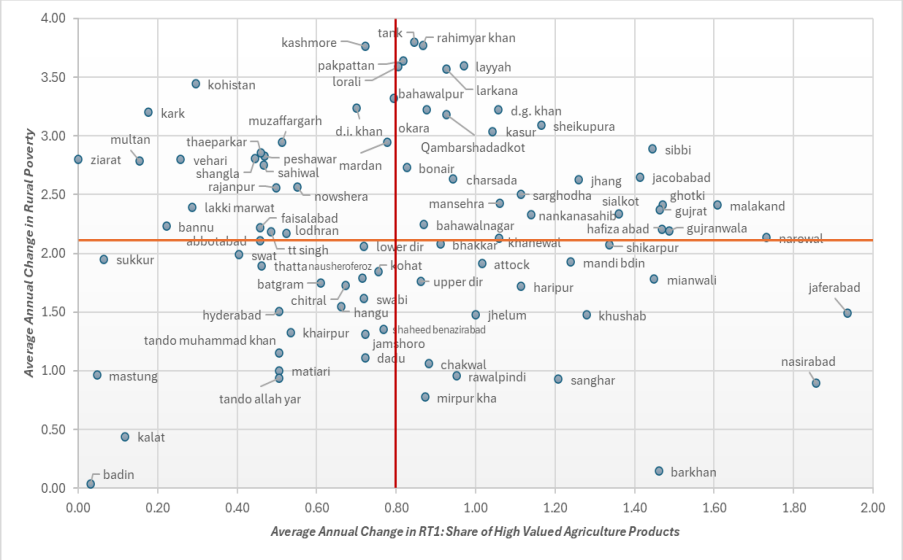


Figure 6.1 **Speed of RT1 and Rural Poverty for the Last Decade**

High-Performance Quadrant: The districts include Bahawalnagar, Bonair, Charsada, D.G. Khan, Ghotki, Gujrat, Jacobabad, Jhang, Kasur, Larkana, Layyah, Malakand, Mansehra, Nankana Sahib, Okara, Pakpattan, Qambar Shahdadkot, Rahimyar Khan, Sargodha, Sheikhupura, Sialkot, Sibbi and Tank. Over the past decade, these districts have witnessed significant poverty alleviation and RT1 is a crucial driver in reducing poverty levels across these regions.

Poverty-Focused Quadrant: The districts include Bahawalpur, Bannu, D.I. Khan, Faisalabad, Kark, Kashmore, Kohistan, Lakki Marwat, Lorali, Mardan, Multan, Muzaffargarh, Nowshera, Peshawar, Rajanpur, Sahiwal, Shangla, Tharparkar, Vehari and Ziarat. Over the past decade, these districts have witnessed significant poverty alleviation and RT1 is not a main driver in reducing poverty levels across these regions.

Transformation-Centric Quadrant: The districts include Attock, Barkhan, Bhakkar, Chakwal, Gujranwala, Hafizabad, Haripur, Jaferabad, Jhelum, Khanewal, Khushab, Mandi Bahauddin, Mianwali, Mirpur Khas, Narowal, Nasirabad, Rawalpindi, Sanghar,

Shikarpur and Upper Dir. Over the past decade, these districts have witnessed significantly low poverty reduction although RT1 is high across these regions.

Lagging Quadrant: The districts include Badin, Batgram, Chitral, Dadu, Hangu, Hyderabad, Jamshoro, Kalat, Khairpur, Kohat, Lodhran, Lower Dir, Mastung, Matiari, Naushero Feroz, Shaheed Benazirabad, Sukkur, Swabi, Swat, Tando Allah Yar, Tando Muhammad Khan, Thatta, Toba Tek Singh. Over the past decade, these districts have witnessed significantly low poverty reduction followed by low RT1. Due to low RT1, poverty reduction is also low, showing that RT1 can be a main driver of poverty reduction. The district Abbotabad is exceptional in the sense that it is one of the top-ranked districts in 2008-09 and 2019-20. The district has already achieved maturity, so it falls in the lagging quadrant in terms of poverty and RT1 during the last decade. Figure 6.2 displays a scatter plot illuminating the association between RT2 and rural poverty reduction.

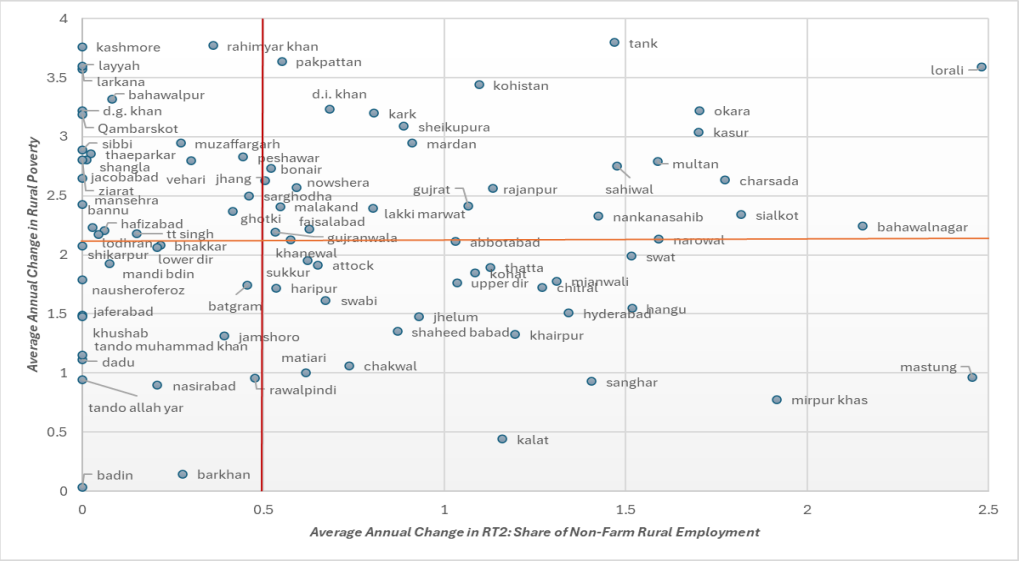


Figure 6.2 **Speed of RT2 and Rural Poverty for the Last Decade**
Source: Author's Calculations

High-Performance Quadrant: The districts include Lorali, Bahawalnagar, Sialkot, Charsada, Okara, Kasur, Multan, Sahiwal, Tank, Nankana Sahib, Rajanpur, Kohistan, Gujrat, Mardan, Sheikhupura, Kark, Lakki Marwat, D.I. Khan, Faisalabad and Nowshera. Over the last decade, these districts have viewed substantial poverty alleviation and RT2 is a fundamental driver in reducing poverty levels across these domains.

Poverty-Focused Quadrant: The districts include Pakpattan, Malakand, Bonair, Jhang, Sargodha, Peshawar, Ghotki, Rahimyar Khan, Vehari, Muzaffargarh, Bahawalpur, Bannu, Tharparkar, Shangla, D.G. Khan, Jacobabad, Kashmore, Larkana, Layyah, Mansehra, Qambar Shahdadkot, Sibbi and Ziarat. Over the past decade, these districts have witnessed significant poverty alleviation but RT2 is not a leading handler in reducing poverty levels across these territories.

Transformation-Centric Quadrant: The districts include Mastung, Mirpur Khas, Narowal, Hangu, Swat, Sanghar, Hyderabad, Mianwali, Chitral, Khairpur, Kalat, Thatta, Kohat, Upper Dir, Abbotabad, Jhelum, Shaheed Benazirabad, Chakwal, Swabi, Attock, Sukkur, Matiari and Khanewal. Over the last decade, these districts have seen notably low poverty reduction although RT2 is high across these regions implying that RT2 is not a leading factor in reducing rural poverty.

Lagging Quadrant: The districts include Batgram, Jamshoro, Barkhan, Bhakkar, Nasirabad, Lower Dir, Toba Tek Singh, Mandi Bahauddin, Hafizabad, Lodhran, Badin, Dadu, Jaferabad, Khushab, Nausheroferoz, Shikarpur, Tando Allah Yar and Tando Muhammad Khan. Over the last decade, these districts have observed considerably low poverty reduction followed by low RT2. Due to low RT2, poverty reduction is also

low, showing that RT2 can be a main driver of poverty reduction. Following districts, Haripur, Gujranwala, Rawalpindi, are exceptional in the sense that they are the top-ranked districts in 2008-09 and 2019-20. These districts have already achieved maturity, so they fall in the lagging quadrant in terms of poverty and RT2 during the last decade.

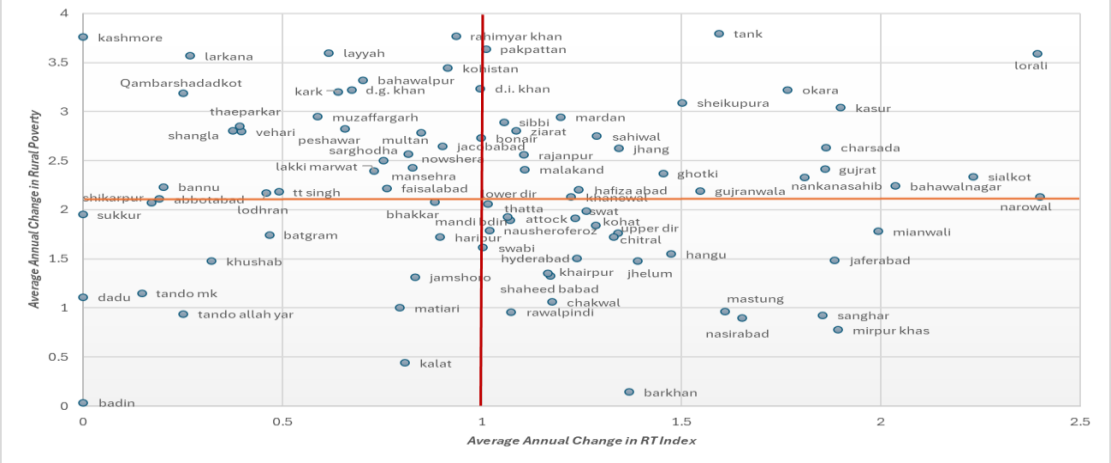


Figure 6.3 **Speed of RT (Index)[65] and Rural Poverty for the Last Decade**
Source: Author's Calculations

High-Performance Quadrant: The districts include Lorali, Sialkot, Bahawalnagar, Kasur, Charsada, Gujrat, Nankana Sahib, Okara, Tank, Sheikhupura, Ghotki, Jhang, Sahiwal, Mardan, Malakand, Rajanpur, Ziarat and Sibbi. Over the preceding decade, these districts have perceived substantial poverty alleviation and Rural Transformation is a fundamental driver in reducing poverty levels across these spheres.

Poverty-Focused Quadrant: The districts include Pakpattan, Bonair, D.I. Khan, Rahimyar Khan, Kohistan, Jacobabad, Multan, Mansehra, Nowshera, Faisalabad, Sargodha, Lakki Marwat, Bahawalpur, D.G. Khan, Peshawar, Kark, Layyah,

[65] The RT Index is constructed by computing a weighted average of RT1 and RT2, with each component assigned equal weighting to reflect its equivalent influence.

Muzaffargarh, Vehari, Tharparkar, Shangla, Larkana, Qambar Shahdadkot, Bannu, Kashmore. Over the prior decade, these districts have viewed significant poverty alleviation, but Rural Transformation is not a major driver in reducing poverty levels across these territories.

Transformation-Centric Quadrant: The districts include Narowal, Mianwali, Mirpur Khas, Jaferabad, Sanghar, Nasirabad, Mastung, Gujranwala, Hangu, Jhelum, Barkhan, Upper Dir, Chitral, Kohat, Swat, Hafizabad, Hyderabad, Attock, Khanewal, Chakwal, Khairpur, Shaheed Benazirabad, Rawalpindi, Thatta, Mandi Bahauddin. Over the last decade, these districts have seen conspicuously low poverty reduction. However, Rural Transformation is high across these counties implying that RT is not a leading component in reducing rural poverty.

Lagging Quadrant: The districts include Nausheroferoz, Lower Dir, Swabi, Haripur, Bhakkar, Jamshoro, Kalat, Matiari, Toba Tek Singh, Batgram, Lodhran, Khushab, Tando Allah Yar, Abbotabad, Shikarpur, Tando Muhammad Khan, Badin, Dadu, Sukkur. Over the last decade, these districts have observed considerably low poverty reduction followed by low Rural Transformation. Due to low RT, poverty reduction is also low showing that RT can be a main driver of poverty reduction.

To further elucidate the dynamics of rural poverty alleviation concerning the rural transformation process, we conducted an advanced typological analysis examining the velocities of rural transformation and rural poverty for the last decade. This refined approach employs a six-category classification system, expanding upon previous four-category models. The methodology involves establishing nuanced thresholds for annual changes in rural poverty over the past decade. We defined two critical

thresholds: $T1 = minimum\ value + \Delta x/3$ and $T2 = minimum\ value + 2\Delta x/3$ where Δx represents the range $maximum\ value - minimum\ value$. Subsequently, we categorized the speed of rural poverty reduction using these thresholds. If rural poverty change is below $T1$ is classified as "Slow", those between $T1$ and $T2$ as "Moderate", and those exceeding $T2$ as "Fast". We utilized the median value as the demarcation point for the Rural Transformation indicators (RT1, RT2, and RT Index). Values above the median were classified as "Fast". At the same time, those below are designated as "Slow". This sophisticated typology integrates three categories of rural poverty reduction with two categories of RT1, yielding a comprehensive six-category framework. Table 6.1 presents a typological analysis, illustrating the districts under each of six categories based on rural poverty and RT1 speed.

Table 6.3 *Typology Analysis: Rural Poverty and RT1 for the Last Decade (2008 and 2019)*

		Speed of Rural Poverty Reduction		
		Fast	Moderate	Slow
Speed of RT1	Fast	Bonair, Charsada, Dera Ghazi Khan, Jacobabad, Jhang, Kasur, Larkana, Layyah, Okara, Pakpattan, Qambar Shahdadkot, Rahimyar Khan, Sheikhupura, Sibbi, Tank, Lorali	Ghotki, Gujrat, Malakand, Mansehra, Nankana Sahib, Sargodha, Sialkot, Attock, Bhakkar, Gujranwala, Hafizabad, Haripur, Jaferabad, Jhelum, Khanewal, Khushab, Mandi Bahauddin, Mianwali, Narowal, Shikarpur, Upper Dir and Bahawalnagar.	Barkhan, Chakwal, Mirpurkhas, Nasirabad, and Sanghar.
	Slow	Bahawalpur, D.I. Khan, Kark, Kashmore, Kohistan, Mardan, Multan, Muzaffargarh, Nowshera, Peshawar, Rajanpur, Sahiwal, Shangla, Tharparkar, Vehari, Ziarat	Bannu, Lakki Marwat, Batgram, Chitral, Hangu, Hyderabad, Jamshoro, Khairpur, Kohat, Lodhran, Lower Dir, Nausheroferoz, Shaheed Benazirabad, Sukkur, Swabi, Swat, Thatta, Toba Tek Singh	Badin, Dadu, Kalat, Mastung, Matiari, Tando Allah Yar, Tando Muhammad Khan

In our refined typological analysis, a comprehensive classification system that explains the intricate relationship between rural poverty reduction and the speed of RT1 is developed[66]. Category 1 encompasses districts exhibiting faster poverty alleviation coupled with fast RT1. These areas represent optimal scenarios where socioeconomic progress and structural changes are occurring synergistically. Districts falling under Category 2 demonstrate moderate poverty reduction alongside fast RT1. This pattern suggests a positive trajectory, albeit with the potential for further enhancement in poverty alleviation efforts. Category 3 includes districts characterized by slow poverty reduction despite fast rural transformation. This discrepancy highlights potential inefficiencies or barriers in translating structural changes into tangible improvements in living standards. In Category 4, we observe districts achieving fast poverty reduction despite slower RT1. This intriguing pattern suggests that RT1 is not a major driver in these areas to combat poverty effectively. Category 5 comprises districts with moderate poverty reduction rates and slower RT1. This interesting pattern advocates that RT1 is not a key handler in these regions to reduce poverty effectively. Faisalabad and Abbotabad are mature districts that fall in this category. Finally, Category 6 represents the most challenging scenario, where both poverty reduction and rural transformation progress are at a slow pace. These districts likely require comprehensive, multifaceted approaches to stimulate development and improve socioeconomic conditions. The role of RT1 may be vital in the anticipated scenarios.

Table 6.4 demonstrates the speed of rural poverty concerning the speed of RT2 for the past decade.

[66] The categories explain here are same for RT2 and RT Index relation with rural poverty.

Table 6.4 *Typology Analysis: Rural Poverty and RT2 for the Last Decade (2008 and 2019)*

		Speed of Rural Poverty Reduction		
		Fast	Moderate	Slow
Speed of RT2	Fast	Lorali, Charsada, Okara, Kasur, Multan, Sahiwal, Tank, Rajanpur, Kohistan, Mardan, Sheikupura, Kark, D.I. Khan, Nowshera, Pakpattan	Bahawalnagar, Sialkot, Nankana Sahib, Gujrat, Lakki Marwat, Faisalabad, Narowal, Hangu, Swat, Hyderabad, Mianwali, Chitral, Khairpur, Thatta, Kohat, Upper Dir, Abbotabad, Jhelum, Shaheed Benazirabad, Swabi, Attock, Sukkur	Mastung, Mirpurkhas, Sanghar, Kalat, Chakwal, Matiari
	Slow	Bonair, Jhang, Peshawar, Rahimyar Khan, Vehari, Muzaffargarh, Bahawalpur, Tharparkar, Shangla, D.G. Khan, Jacobabad, Kashmore, Larkana, Layyah, Qambar Shahdadkot, Sibbi, Ziarat	Malakand, Ghotki, Bannu, Mansehra, Batgram, Jamshoro, Bhakkar, Lower Dir, Toba Tek Singh, Mandi Bahauddin, Hafizabad, Lodhran, Jaferabad, Khushab, Nausheroferoz, Shikarpur	Barkhan, Nasirabad, Badin, Dadu, Tando Allah Yar, Tando Muhammad Khan

Category 1 encompasses districts exhibiting faster poverty alleviation coupled with fast RT2. These areas represent optimal scenarios where socioeconomic progress and structural changes in terms of non-farm employment are occurring synergistically. Districts falling under Category 2 demonstrate moderate poverty reduction alongside fast RT2. This pattern suggests a positive route, albeit with the potential for further enhancement in poverty alleviation efforts alongside RT2. Category 3 includes districts characterized by slow poverty reduction despite fast rural transformation highlighting potential inefficiencies or barriers in interpreting structural changes into tangible improvements in living standards. In Category 4, we witnessed districts achieving fast poverty reduction despite slower RT2. This intriguing pattern suggests that RT2 is not a foremost factor in these domains to address poverty effectively. Category 5 encompasses districts with moderate poverty reduction followed by slower RT2.

Haripur, Gujranwala, and Sargodha are the top-ranked districts having achieved maturity results in the fall to this category. This interesting pattern advocates that a slow pace of non-farm employment results in a moderate pace of poverty reduction in these areas. Finally, Category 6 represents the most challenging scenario, where both poverty reduction and RT2 progress are at a slow pace. These districts likely require comprehensive, multifaceted approaches to stimulate development and improve socioeconomic conditions. In this respect, the role of RT2 can be crucial in the anticipated scenarios.

Finally, Table 6.5 combines both RT1 and RT2 to form an RT index and then its relation is explored with rural poverty.

Table 6.5 *Typology Analysis: Rural Poverty and RT Index for the Last Decade (2008 and 2019)*

		Speed of Rural Poverty Reduction		
		Fast	Moderate	Slow
Speed of RT	Fast	Lorali, Kasur, Charsada, Okara, Tank, Sheikhupura, Jhang, Sahiwal, Mardan, Rajanpur, Ziarat, Sibbi	Sialkot, Bahawalnagar, Gujrat, Nankana Sahib, Ghotki, Malakand, Narowal, Mianwali, Jaferabad, Gujranwala, Hangu, Jhelum, Upper Dir, Chitral, Kohat, Swat, Hafizabad, Hyderabad, Attock, Khanewal, Khairpur, Shaheed Benazirabad, Thatta, Mandi Bahauddin, Nausheroferoz	Mirpur Khas, Sanghar, Nasirabad, Mastung, Barkhan, Chakwal, Rawalpindi
	Slow	Pakpattan, Bonair, D.I. Khan, Rahimyar Khan, Kohistan, Jacobabad, Multan, Nowshera, Bahawalpur, D.G. Khan, Peshawar, Kark, Layyah, Muzaffargarh, Vehari, Thaeparkar, Shangla, Larkana, Qambar Shahdadkot, Kashmore	Mansehra, Lakki Marwat, Bannu, Lower Dir, Swabi, Bhakkar, Jamshoro, Tt Singh, Batgram, Lodhran, Khushab, Shikarpur, Sukkur	Kalat, Matiari, Tando Allah Yar, Tando Muhammad Khan, Badin, Dadu

Category 1 incorporates districts revealing faster poverty alleviation linked with fast RT. These domains are characterized by optimum settings where socioeconomic advancement and underlying changes in terms of RT1 and RT2 are happening collectively. Districts falling under Category 2 reveal moderate poverty reduction alongside fast RT. This pattern indicates an optimistic path, granting the potential for further augmentation in poverty alleviation attempts. Category 3 includes districts portrayed by slow poverty reduction despite fast rural transformation emphasizing potential inadequacies or hurdles in interpreting basic transformations into tangible improvements in living standards. In Category 4, we observed districts achieving fast poverty reduction despite slower RT. This stimulating pattern suggests that RT is not a foremost factor in these domains to address poverty effectively. Category 5 involves districts with moderate poverty reduction trailed by slower RT. This interesting pattern advocates that a slow pace of RT results in a moderate pace of poverty reduction in these areas. Abbotabad, Haripur, Faisalabad, and Sargodha are the mature districts[67] that fall in this category. Finally, Category 6 represents the most challenging scenario, where both poverty reduction and RT2 progress are at a slow pace. These districts likely necessitate thorough, manifold techniques to stimulate development and improve socioeconomic conditions. In this regard, the role of Rural Transformation might be crucial.

[67] Mature districts means those districts that are already well established in the previous decade. As a result, the average annual change in these districts is lower as compared to those districts that are not matured a decade ago. Consequently, having high annual change or growth.

6.1.4 Conclusion

This study provides compelling evidence for the significant role of rural transformation in alleviating poverty across Pakistan's diverse districts. Through rigorous econometric modeling and advanced typological analysis, we have uncovered subtle relationships between various dimensions of rural transformation and poverty reduction dynamics.

Our panel fixed effects model revealed statistically significant negative associations between both measures of rural transformation (RT1 and RT2) and rural poverty headcount. RT1 and RT2 demonstrated a particularly strong impact, with a 1% increase corresponding to a 0.52 and 0.31 percentage point decrease in rural poverty respectively. This finding underscores the critical importance of structural changes captured by RT1 and RT2 in poverty alleviation efforts.

The subsequent quadrant analysis provided a spatial mapping of development trajectories across districts. This approach illuminated the heterogeneous nature of rural transformation and poverty reduction processes, identifying high-performing districts where RT and poverty reduction synergize effectively, as well as areas where these processes are decoupled or lagging.

Our sophisticated six-category typology further elucidated the complex interplay between the velocity of rural transformation and poverty reduction. This analysis revealed optimal scenarios where rapid RT coincides with accelerated poverty alleviation, as well as challenging contexts where both processes stagnate, necessitating targeted interventions.

Importantly, the study found that the poverty-reducing effects of rural transformation occur consistently and linearly across the observed range of RT1 and RT2 values,

validating our linear model specification. This finding suggests that policies promoting various aspects of rural transformation could be effective strategies for combating rural poverty across different contexts.

The differential impacts of RT1 and RT2 on poverty reduction highlight the multifaceted nature of rural transformation processes. This delicate understanding calls for tailored policy approaches that consider the specific dimensions of RT most relevant to each district's socioeconomic context.

The government must shift from uniform welfare programs to targeted interventions informed by district-level dynamics. The policy focus for optimal districts is on sustaining momentum and replicating their success factors nationally. Districts with high RT but low poverty reduction require urgent review to address structural barriers and inequality leakage. Similarly, districts with low RT but high poverty reduction require human capital investment (e.g., vocational training). Districts with slow poverty and RT require the most intensive interventions. They should be designated as Special Development Zones receiving comprehensive, multi-sectoral stimulus packages to address fundamental deficits in both economic structure and welfare.

In conclusion, this research provides a comprehensive framework for understanding and addressing rural poverty through the lens of rural transformation. The findings underscore the importance of promoting structural changes in rural economies, enhancing non-farm employment opportunities, and implementing context-specific interventions to accelerate poverty reduction. Future policy initiatives should leverage these insights to design more effective, targeted strategies for sustainable rural development across Pakistan's diverse landscape. This study contributes significantly

to the growing body of literature on rural development and poverty alleviation, offering a methodologically robust approach to analyzing the complex dynamics of rural transformation. Further research could explore the specific mechanisms through which different aspects of rural transformation impact poverty reduction, potentially identifying key leverage points for policy intervention. Building upon our analysis based on nationally representative surveys (PSLM 2008-09 and PSLM 2019-20), a valuable next step would be to conduct in-depth qualitative studies for each district included in this research. While our quantitative approach provides robust, generalizable findings, district-specific qualitative analyses could offer complementary, context-rich information about how rural transformation processes are experienced at the household and community levels. This further uncovers district-specific factors influencing the relationship between rural transformation and poverty reduction, offering subtle perspectives on local dynamics. Identify local best practices and challenges in implementing rural transformation initiatives, potentially explaining variations in poverty reduction outcomes across districts. Finally, captures the lived experiences of community members, offering a human-centered perspective on the impacts of rural transformation.

6.2 Regional Rural Income and Agriculture Credit

This study investigates the impact of agricultural credit on rural household income at the district level in Pakistan. The study is structured into four cohesive sections that progress from theoretical foundations to empirical analysis and synthesis of findings. Section 6.2.1 describes the contextual background and illuminates the complex nature of rural income dynamics and the potential of agricultural credit as a catalyst for economic growth. This section explores the theoretical underpinnings of the relationship between credit accessibility and income generation in rural settings, setting the stage for our empirical investigation. Section 6.2.2 provides a comprehensive overview of the data acquisition strategy and analytical approach. The diverse data sources utilized, including government databases, household surveys, and specialized agricultural reports are detailed in this section. The operationalization and measurement of key variables related to agricultural credit and rural income are explicated along with summary statistics. The econometric model employed, such as advanced panel data analysis techniques, fixed or random effects estimations, and other specialized statistical procedures tailored to our research objectives is elucidated. Section 6.2.3 presents the outcomes of our econometric investigation, offering an interpretation of the findings. The econometric output, including coefficient estimates, standard errors, and model fit indicators, is then presented. Finally, a visual analysis is conducted to explore the mechanisms through which credit acquisition influences household income in rural districts. In this final segment (section 6.2.4), we synthesize the key insights from our analysis and extrapolate their implications for policy and practice. We provide a concise recapitulation of the main empirical results and their

significance, followed by a discussion on how these findings can inform and refine rural credit and income enhancement strategies.

6.2.1 Introduction

Credit is widely recognized as the backbone of economies, particularly in developing countries like Pakistan. The primary purpose of lending credit is to improve the living standards of people by enhancing their income and interrupting the vicious cycle of poverty (Khandker & Faruqee, 2003; Lipton & Sepp, 2009). Numerous studies have demonstrated a positive relationship between credit accessibility and household expenditure, suggesting that credit support for deprived households can contribute to increased income and improved well-being (Armendariz, 2010; Beck et al., 2007; Lipton & Sepp, 2009). In Pakistan, agriculture remains a crucial sector, contributing approximately 19.2% to the total GDP (PES, 2021). With over 117.48 million people living in rural areas, the role of agricultural credit is pivotal in supporting agricultural activities, especially for small and medium-scale farmers. However, the process of seeking agricultural credit is complex, particularly for small-scale farmers who often hesitate to take substantial loans due to their poor socio-economic conditions (S. R. Das & Hanouna, 2009; Pandey, 2016).

This research is grounded in the broader context of rural development and agricultural economics. It examines the intricate relationship between rural household income and agricultural credit, focusing on how agriculture credit disbursements impact the well-being (income) of rural households. Various studies in the past consider diverse factors influencing credit accessibility, including land ownership, past loan repayment performance, farming experience, exposure to agricultural technologies, and the age of

the borrower (Akudugu & Development, 2012; Hadley, 2006; Mbam & Edeh, 2011). The framework also incorporates the concept of rural transformation, recognizing the shift in employment trends from farm to off-farm activities in rural areas over the past four decades. This transformation is essential in understanding the changing dynamics of rural economies and the role of credit in supporting both agricultural and non-agricultural activities[68].

Despite the government's efforts to assist marginalized and deprived farmers in poverty-stricken rural areas, significant challenges persist in the accessibility and effective utilization of agricultural credit. Only 16% of farmers in Pakistan receive institutional credit, often due to political references and strong social backgrounds, while 31% struggle to obtain credit from commercial banks under strict regulations (Chandio, Jiang, Wei, & Guangshun, 2018). This disparity in credit access exacerbates existing inequalities and hinders rural development. Furthermore, the complex lending process, coupled with the hesitation of small-scale farmers to take substantial loans, creates a barrier to financial inclusion. The lack of proper management and utilization of credit facilities often fails to translate into increased rural household income, perpetuating the cycle of poverty (Khandker & Faruqee, 2003; Lipton & Sepp, 2009).

The relationship between rural income and rural credit is evident. Due to rural credit, the overall growth rate increases (Chandio et al., 2018; Iqbal, Ahmad, Abbas, & Mustafa, 2003; Kouamé, 2010; Rehman, Muhammad, Sarwar, & Raz, 2019; Sial, Awan, & Waqas, 2011). Even though there are different campaigns on credit loans by

The categories explain here are same for RT2 and RT Index relation with rural poverty.

[68] Mature districts means those districts that are already well established in the previous decade. As a result, the average annual change in these districts is lower as compared

the government yet there are some illiterate farmers who are unable to understand this lending facility (Khalid Bashir, Gill, & Hassan, 2009). However, for big farmers loan schemes are different and they get more benefits from lending credit. It is important to highlight that sometimes farmers get into big losses and cannot pay back their loans on time, which complicates the situation for farmers, and they get into the vicious cycle of poverty.

The rural income is reflective of the socio-economic conditions of marginalized families who are struggling with food insecurity, malnutrition, mortalities, vulnerabilities, and natural disasters. In such scenarios, the need of the hour is to understand this deeper strategy of empowering farmers and non-farm agriculturists by making them independent in their decision-making regimes via effective credit policies (Ahmed et al., 2016; Asad, Mehdi, Ashfaq, Hassan, & Abid, 2019; Barrett, 2008; Sitko & Jayne, 2014).

The primary objective of this study is to explore the relationship between rural household income[69] and agricultural credit in Pakistan. Specifically, the research aims to examine the impact of agricultural credit on rural household income and overall well-being. Also, identify the most vulnerable districts that are under severe poverty and also have no or less access to credit. Finally, conduct a typology analysis to find credit growth and income growth for 2019 (using data points from 2008 and 2019) to observe the speed of credit concerning rural household income. This analysis will help to identify districts where income and credit growth are low in terms of speed.

to those districts that are not matured a decade ago. Consequently, having high annual change or growth. od at the district level in Pakistan.

The research aims to inform government policymakers about the current landscape of credit accessibility and its impact on rural household incomes across different districts in Pakistan. The information enables the government about high-risk areas characterized by low incomes and inadequate credit availability. By highlighting vulnerable districts through visualization and typology analysis, the study assists the government in targeted planning and implementation of measures to address disparities in credit access and income levels leading to more effective resource allocation and policy formulation for rural development. The study's findings will provide financial institutions with a comprehensive view of agriculture credit distribution. This information guides these institutions in focusing their efforts on more vulnerable areas, potentially leading to the development of specialized loan schemes with potential collaborations between the government and financial institutions in designing and implementing tailored credit programs. These programs would aim to enhance rural household incomes in identified high-risk districts, thereby contributing to broader well-being efforts of rural households. By examining credit and income growth patterns over time (2008-2019), the study offers insights into the long-term effectiveness of existing credit policies and programs. This temporal perspective will aid both government and financial institutions in refining their approaches to rural credit. Furthermore, this study aims to bridge the gap between research, policy, and practice in the realm of agricultural credit and rural development in Pakistan, ultimately contributing to more informed decision-making and targeted interventions for improving rural livelihoods.

6.2.2 Data and Methodology

The purpose of this section of the study is to explore the relationship between the per capita income of households and agriculture credit disbursement, that is to investigate how the process of Credit disbursement enhances household income at the district level in Pakistan. The analysis further leads to checking the speed of credit disbursement concerning the rural income of households and conducting typology analysis to identify the district where household level income is low due to limited/non-availability of credit facilities, so that corrective measures can be taken to enhance the income of households in those districts through learning from fast-growing districts. The agriculture credit data is taken from SBP, and data on control variables is taken from nationally representative surveys HIES and PSLM. Agriculture credit is our main explanatory variable which is a combination of farm and non-farm credit. The farm credit includes loans for inputs like seeds, fertilizers, and pesticides, as well as working capital for farming expenses. Farm credit loans are usually self-liquidating, meaning the loan is repaid from the sale of the crop after harvesting. On the other hand, non-farm credit includes loans for non-agricultural activities like livestock, dairy, poultry, fisheries, apiculture, sericulture, and floriculture. The rural income and agriculture credit data is deflated to account for possible inflation. Other than the main explanatory variables, some control variables are also used to address socioeconomic factors. These control variables are defined as follows:

The gross enrolment rate in the middle is defined as the total number of students enrolled in middle education, divided by the total population of children within the middle, aged between 10 to 12 years. The Net Enrollment Rate for the primary

education sector is calculated by dividing the number of children aged 5 to 9 years enrolled by the total number of children in the same age group. Non-vulnerable jobs for females are the percentage of the employed female population (10 years and above) that are either paid workers, employers, or self-employed in the non-agriculture sector, divided by the total employed female population. Table 6.6 displays the summary statistics of the variables used in the analysis:

Table 6.6 *Summary Statistics*

Variables	Mean	Standard Deviation	Minimum	Maximum
Rural Income (in logarithm)[70]	8.1903	0.3083	7.7003	9.0003
Agriculture Credit (in logarithm)	6.2198	2.7954	-1.6045	11.1612
Household Size	5.9805	0.9527	3.7277	8.3959
Gross Enrolment Rate - Middle	47.5242	20.1993	7.9686	100.0000
Male Net Enrolment Rate - Primary	56.2977	12.1648	21.1561	84.9643
Non-Vulnerable Jobs - Female	39.3837	27.7369	0.0000	100.0000

Source: Author's Calculation

Once the variable of interest is calculated at the district level, then to explore the relationship between the variables, a panel analysis is conducted.

$$Y_{it} = \lambda_1 Credit_{it} + \beta_n CV_{it} + v_i + \eta_{it} \qquad (8)$$

where i indicates the district i, t indicates the year t. Also λ_1 and λ_2 are the coefficients to the respective variables in model, and β_n represents coefficients of control variables in model. The fixed or random effect model is finalized based on the Hausman test.

[70] Rural Income in logarithm is the log of average per adult equivalent consumption expenditure.

6.2.3 Results and Discussions

This section examines the relationship between rural income, operationalized through per equivalent adult expenditure data derived from an advanced indirect modeling approach[71], and formal agricultural credit. To investigate the dynamics between rural income and credit in Pakistan, we employ a sophisticated panel fixed effect, multivariate econometric model[72] leveraging longitudinal data. This methodological approach allows for a delicate analysis of temporal and cross-sectional variations[73]. The findings of this analysis are presented in Table 6.7, which elucidates the complex interaction between these critical socioeconomic variables.

Table 6.7 *Impact of Agriculture Credit on Rural Income – A Panel Fixed Effect Model*

Dependent Variable: Rural Income (in logarithm)				
Variables	**Coef.**	**Std. Err.**	**T-Stats**	**P-Value**
Agriculture Credit (in logarithm)	0.0612	0.0301	2.04	0.0450
Household Size	-0.2285	0.0232	-9.83	0.0000
Gross Enrolment Rate - Middle	0.0115	0.0023	5.05	0.0000
Male Net Enrolment Rate -	0.0155	0.0030	5.1	0.0000
Non-Vulnerable Jobs - Female	-0.0010	0.0119	2.01	0.0409
Constant	9.5421	0.2836	33.65	0.0000
R - Squared	**0.74**	**F - Stat**	**46.66**	**0.0000**

The coefficient for agricultural credit (0.0612) is positive and statistically significant (p-value of 0.0450). In this log-log model, the coefficient can be interpreted as elasticity. Specifically, a 1% increase in agricultural credit is associated with a 0.0612% increase in rural income, holding other factors constant. This elasticity suggests that while agricultural credit is positively associated with rural income, the relationship is

[71] The data on rural income at the district level is generated via the Census EBP method of Small Area Estimation.
[72] The Hausman Test suggests that the Panel Fixed Effect model is appropriate.
[73] This analysis applies the same robustness criteria detailed in Section 6.2, including the stepwise selection of regressors, VIF-based multicollinearity, and endogeneity checks.

inelastic, indicating that large changes in credit availability are needed to produce substantial changes in rural income[74]. Several factors from the literature could explain this inelastic relationship. (Binswanger & Khandker, 1995) suggest that farmers may not always use credit optimally due to a lack of financial literacy or limited access to complementary inputs. This inefficiency could dampen the impact of credit on income. The lack of proper management and utilization of credit facilities often fails to translate into increased rural household income, perpetuating the cycle of poverty (Khandker & Faruqee, 2003; Lipton & Sepp, 2009). (Feder, Lau, Lin, & Luo, 1990) argue even with access to credit, farmers may face constraints in input and output markets, limiting their ability to translate credit into higher incomes. (Rosenzweig & Wolpin, 1993) argue that agricultural investments may exhibit diminishing returns, particularly in areas with limited technological adoption, resulting in smaller income gains as credit increases. (Stiglitz & Weiss, 1981) discuss how credit rationing in agricultural markets can lead to suboptimal credit allocation, potentially reducing its effectiveness in boosting incomes. (Sharma & Zeller, 1997) point out that some farmers may use agricultural credit for smooth consumption rather than productive investments, limiting their impact on long-term income growth. (M. R. Carter & Olinto, 2003)highlight how structural barriers in rural areas, such as poor infrastructure or limited access to markets, can constrain the income-generating potential of agricultural credit.

The gross enrollment rate at the middle school level shows a positive and significant association with rural income (coefficient = 0.0115, p-value < 0.0001). A one percentage point increase in this enrollment rate is associated with approximately a

[74] This analysis is based on limited two-period data and lacks a comprehensive exploratory analysis within specific districts to establish any causal link.

1.15% increase in rural income, highlighting the potential returns to secondary education in rural areas. Similarly, the male net enrollment rate at the primary level demonstrates a positive relationship with rural income (coefficient = 0.0155, p-value < 0.0001). A one percentage point increase in this rate is associated with a 1.55% increase in rural income. This result may reflect complex socioeconomic dynamics or opportunity costs associated with primary education in rural settings. The proportion of females in non-vulnerable jobs shows a slight negative but statistically significant relationship with rural income (coefficient = -0.0010, p-value = 0.0409). A one percentage point increase in this proportion is associated with approximately a 0.10% decrease in rural income. While the effect is small, this counterintuitive finding might suggest complexities in rural labor markets or potential trade-offs between job security and income levels for women in rural areas.

The model demonstrates strong explanatory power with an R-squared value of 0.74, indicating that the included explanatory variables explain approximately 74% of the variation in log rural income. The F-statistic (46.66) with a p-value < 0.0001 confirms the overall significance of the model[75]. These results provide delicate insights into the determinants of rural income, with agricultural credit emerging as a positive but inelastic influencer. The findings also highlight the complex interplay between household characteristics, education, and employment factors in shaping rural economic outcomes. Policymakers and rural development practitioners may find these results valuable in designing targeted interventions to enhance rural incomes through

[75] The model is thoroughly checked for multicollinearity using VIF criteria, and explanatory variables are selected through the robust analysis using stepwise forward and backward model selection criteria based on p-value.

strategic credit allocation and complementary investments in education and labor market development while being aware of the potentially complex relationships between these factors and income levels.

The econometric modeling reveals an interaction between formal agriculture credit and rural income across Pakistan's districts. To better understand this relationship, we employed scatter plot visualizations to juxtapose credit growth against rural income growth. The resulting plots were structured into four quadrants, each representing a distinct dynamic between agriculture credit and rural income. The High-Performance Quadrant comprises districts exhibiting above-median rural income growth and above-median agriculture growth, indicating comprehensive progress. In contrast, the Income-Focused Quadrant represents districts achieving above-median rural income growth despite below-median agriculture credit growth, suggesting targeted rural income elevation efforts. The Credit-Centric Quadrant includes districts with below-median rural income growth but above-median agriculture credit growth, indicating structural changes that have yet to translate into significant rural income elevation. Finally, the Lagging Quadrant comprises districts demonstrating below-median performance in both rural income growth and agriculture credit growth, highlighting areas requiring intensified financial and development interventions.

Figure 6.4 visualizes the correlation between the dynamics of formal agriculture credit disbursements and rural household income trajectories, providing a graphical representation of the relationship between these two variables.

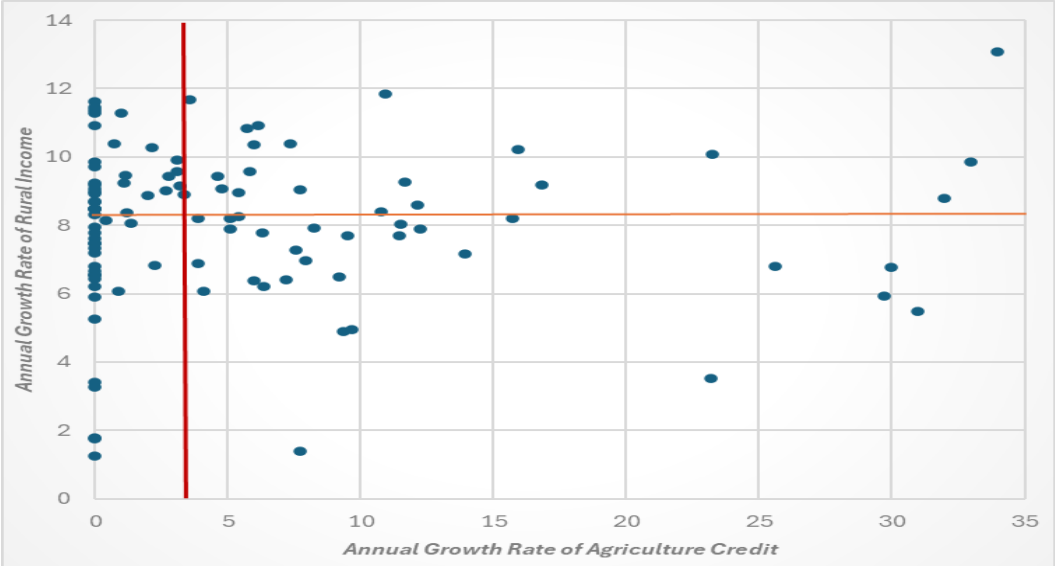


Figure 6.4 **Speed of Credit and Rural Income for the Last Decade**

Source: Author's Calculation

High-Performance Quadrant: The districts of Islamabad, Kech/Turbat, Karachi, Lorali, Qambar Shahdadkot, Sahiwal, Muzaffargarh, Mansehra, Layyah, Rajanpur, Multan, Sialkot, Pakpattan, Rahimyar Khan, Faisalabad, Kasur, Jhang, Hafizabad, Nankana Sahib, Okara, T.T. Singh, Abbotabad and Larkana have experienced significant rural income elevation over the past decade, with agriculture credit playing a crucial role in elevating income levels across these regions.

Income-Focused Quadrant: The districts of Mandi Bahauddin, Sargodha, Gujranwala, Peshawar, Vehari, Jacobabad, Bahawalpur, D.G. Khan, Gujrat, D.I. Khan, Bannu, Bonair, Charsada, Dera Bugti, Jhelum, Kark, Kashmore, Kohistan, Kohlu, Lakki Marwat, Lower Dir, Malakand, Mardan, Nowshera, Nuski, Shangla, Sheikhupura, and Tank have witnessed significant rural income elevation over the past decade, with agriculture credit not being a primary driver of income elevation in these regions.

Credit-Centric Quadrant: The districts of Kalat, Khairpur, Tando Allah Yar, Mastung, Tharparkar, Bahawalnagar, Naushero Feroz, Haripur, Rawalpindi, Sukkur, Killa

Saifullah, Lodhran, Badin, Sanghar, Narowal, Hyderabad, Sibbi, Bhakkar, Tando Muhammad Khan, Mirpurkhas, Khanewal, Matiari, Mianwali, Attock, Shikarpur, Dadu, Ghotki and Khushab have experienced low rural income elevation over the past decade, despite having high levels of agriculture credit.

Lagging Quadrant: The districts of Jaferabad, Lasbela, Kachi/Bolan, Chakwal, Awaran, Batgram, Chitral, Gwadar, Hangu, Jamshoro, Kharan, Khuzdar, Killa Abdullah, Kohat, Nasirabad, Peshin, Shaheed Benzirabad, Swabi, Swat, Thatta, Upper Dir, Washuk and Ziarat have experienced low rural income elevation and low levels of agriculture credit over the past decade, indicating that agriculture credit can be a key driver of rural income elevation in these regions. The districts classified under the Lagging Quadrant are characterized by sluggish or declining agriculture credit growth, concurrent with low rural income growth. These districts are particularly vulnerable, necessitating swift and targeted government interventions to stimulate economic development and elevate the rural incomes of the households within these regions.

Figure 6.4 presents a visual representation of the current dynamics between formal agriculture credit disbursements and rural poverty levels, as depicted through a scatter plot.

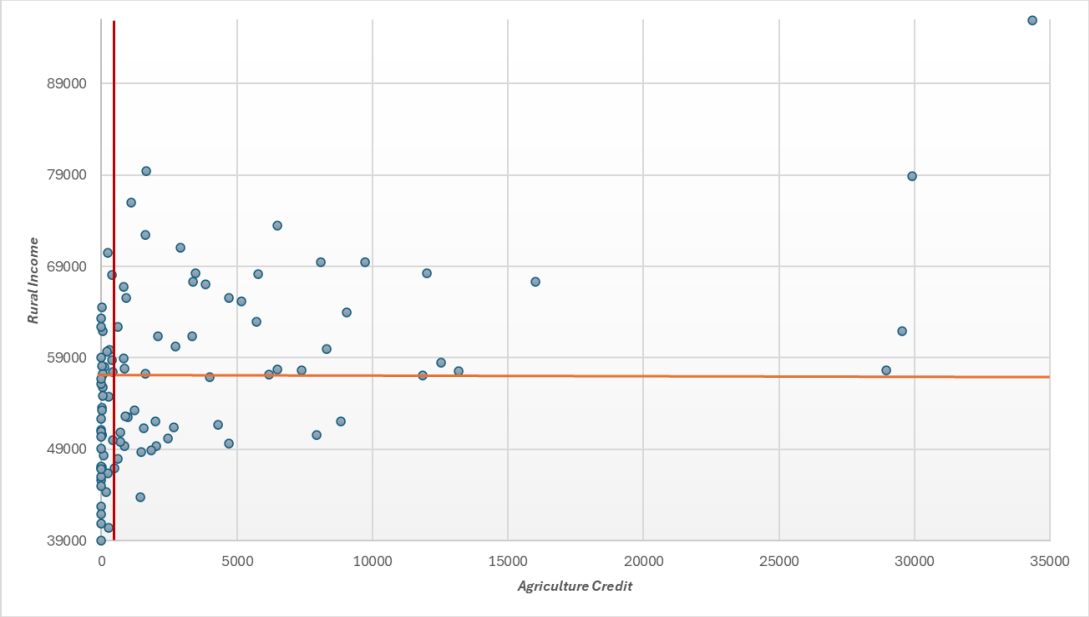


Figure 6.5 **Current Status of Agriculture Credit and Rural Income**

Source: Author's Calculation

High-Performance Quadrant: All districts in this quadrant are characterized by high rural income followed by high credit disbursements. These include Islamabad, Karachi, Multan, Rawalpindi, Sahiwal, Rahimyar Khan, Bahawalnagar, Faisalabad, Okara, Sargodha, Jhang, Gujranwala, Vehari, Khanewal, Sialkot, Lodhran, Toba Tek Singh, Pakpattan, Kasur, Layyah, Hafizabad, Sheikhupura, Nankana Sahib, Narowal, Mandi Bahauddin, Mianwali, Khushab, Gujrat, Dera Ismail Khan, Jhelum, Chakwal, Mansehra and Kashmore. This shows that agriculture credit plays a vital role in elevating rural household income in these districts.

Income-Focused Quadrant: The districts within this quadrant are characterized by elevated rural income levels, accompanied by relatively low levels of formal agriculture credit disbursements. This subset includes Attock, Mardan, Peshawar, Charsada, Swabi, Haripur, Nowshera, Abbotabad, Chitral, Bannu, Malakand, Lower

Dir, Quetta, Shangla, Kark, Batgram, Kohlu, and Tank. Notably, while agriculture credit is limited in these districts, rural household income remains relatively high.

Credit-Centric Quadrant: The districts within this quadrant are characterized by low rural income levels, concomitant with elevated levels of formal agriculture credit disbursements. This subset includes Bahawalpur, Hyderabad, Muzaffargarh, Rajanpur, Sukkur, Bhakkar, Dera Ghazi Khan, Khairpur, Sanghar, Naushero Feroz, Mirpurkhas, Ghotki, Shaheed Benazirabad, Badin, Larkana, Qambar Shahdadkot, Matiari, and Tando Allah Yar. Notably, while formal agriculture credit disbursements are abundant in these districts, rural household income remains relatively stagnant.

Lagging Quadrant: The districts within this quadrant are characterized by low rural income levels, concomitant with limited or no formal agriculture credit disbursements. This subset includes Dadu, Shikarpur, Thatta, Tando Muhammad Khan, Jacobabad, Tharparkar, Swat, Jaferabad, Nasirabad, Jamshoro, Kohat, Lakki Marwat, Bonair, Lorali, Lasbella, Kech/Turbat, Killa Saifullah, Upper Dir, Hangu, Sibbi, Peshin, Kalat, Kachi/Bolan, Khuzdar, Mastung, Kharan, Kohistan, Ziarat, Awaran, Dera Bugti, Gwadar, Killa Abdullah, Nuski, and Washuk. The districts classified under the Lagging Quadrant are characterized by vulnerable regions where not only the current state of income of rural households is extremely low, but also the availability of credit is severely limited, exacerbating the challenges faced by these communities.

To gain a deeper understanding of the dynamics of rural income development concerning agricultural credit expansion, we conducted a typological analysis examining the speed of agricultural credit and rural income over the past decade. This advanced approach builds upon previous frameworks by introducing a six-category

classification system. Our methodology involved establishing precise benchmarks for annual growth in rural income over the past decade. We defined two critical thresholds: $T1 = minimum\ value + \Delta x/3$ and $T2 = minimum\ value + 2\,\Delta x/3$, where Δx represents the range $maximum\ value - minimum\ value$. Subsequently, we categorized the pace of rural income growth using these thresholds. If rural income growth falls below $T1$, it is classified as "Slow", while growth between $T1$ and $T2$ is classified as "Moderate", and growth exceeding $T2$ is classified as "Fast". We employed the median value as the demarcation point for agricultural credit growth, with values above the median classified as "Fast" and those below designated as "Slow". This typology integrates three categories of rural income growth with two categories of agricultural credit growth, resulting in a comprehensive six-category framework. Table 6.7 presents a typological analysis, illustrating the districts under each of six categories based on rural income and agricultural credit growth.

Table 6.8 *Typology Analysis: Rural Income and Credit for the Last Decade (2008 and 2019)*

		Speed of Rural Income		
		Fast	Moderate	Slow
Speed of Agriculture Credit	Fast	Islamabad, Kech/Turbat, Lorali, Qambar Shahdadkot, Sahiwal, Mansehra, Layyah, Sialkot, Pakpattan, Rahimyar Khan, Faisalabad, Kasur, Nankana Sahib, Okara, Abbotabad, Larkana, Mandi Bahauddin	Karachi, Muzaffargarh, Rajanpur, Multan, Jhang, Hafizabad, T.T. Singh, Kalat, Khairpur, Tando Allah Yar, Mastung, Bahawalnagar, Naushero Feroz, Haripur, Rawalpindi, Sukkur, Lodhran, Sanghar, Narowal, Hyderabad, Bhakkar, Tando Muhammad Khan, Mirpurkhas, Khanewal, Matiari, Mianwali, Attock, Shikarpur, Dadu, Ghotki, Khushab	Tharparkar, Killa Saifullah, Badin, Sibbi
	Slow	Sargodha, Peshawar, Bahawalpur, D.G. Khan, Gujrat, D.I. Khan, Charsada, Kark, Kashmore, Kohistan, Kohlu, Mardan, Nowshera, Sheikhupura, Tank	Gujranwala, Vehari, Jacobabad, Bannu, Bonair, Dera Bugti, Jhelum, Lakki Marwat, Lower Dir, Malakand, Nuski, Shangla, Jaferabad, Lasbela, Kachi/Bolan, Chakwal, Batgram, Chitral, Gwadar, Hangu, Jamshoro, Kharan, Kohat, Nasirabad, Quetta, Shaheed Benzirabad, Swabi, Swat, Thatta, Upper Dir, Washuk	Awaran, Khuzdar, Killa Abdullah, Peshin, Ziarat

To assess the contemporary state of rural income and agricultural credit, we applied the thresholds of RT1 and T2 to the current data on rural income, in conjunction with the median threshold for agricultural credit data from 2019. The resulting categorization is presented in Table 6.8, which illustrates the districts falling within each of the six categories.

Table 6.9 *Typology Analysis: Rural Income and Credit for the year 2019*

		Current Status of Rural Income		
		High	Moderate	Low
Current Status of Agriculture Credit	High	Islamabad, Rawalpindi, Gujrat	Karachi, Sahiwal, Bahawalnagar, Faisalabad, Okara, Sargodha, Jhang, Gujranwala, Sialkot, Toba Tek Singh, Pakpattan, Kasur, Layyah, Hafizabad, Sheikhupura, Nankana Sahib, Narowal, Mandi Bahauddin, Mianwali, Khushab, Jhelum, Chakwal, Mansehra, Attock	Multan, Rahimyar Khan, Vehari, Khanewal, Lodhran, D.I. Khan, Kashmore, Bahawalpur, Hyderabad, Muzaffargarh, Rajanpur, Sukkur, Bhakkar, D.G. Khan, Khairpur, Sanghar, Naushero Feroz, Mirpurkhas, Ghotki, Shaheed Benazirabad, Badin, Larkana, Qambar Shahdadkot, Matiari, Tando Allah Yar
	Low		Mardan, Peshawar, Swabi, Haripur, Nowshera, Abbotabad, Chitral, Bannu, Malakand, Shangla, Kark, Batgram, Kohlu, Tank	Charsada, Lower Dir, Quetta, Dadu, Shikarpur, Thatta, Tando Muhammad Khan, Jacobabad, Tharparkar, Swat, Jaferabad, Nasirabad, Jamshoro, Kohat, Lakki Marwat, Bonair, Lorali, Lasbela, Kech/Turbat, Killa Saifullah, Upper Dir, Hangu, Sibbi, Peshin, Kalat, Kachi/Bolan, Khuzdar, Mastung, Kharan, Kohistan, Ziarat, Awaran, Dera Bugti, Gwadar, Killa Abdullah, Nuski, Washuk

In this typological analysis, a notable subset of districts exhibits heightened vulnerability, including Charsada, Lower Dir, Quetta, Dadu, Shikarpur, Thatta, Tando Muhammad Khan, Jacobabad, Tharparkar, Swat, Jaferabad, Nasirabad, Jamshoro, Kohat, Lakki Marwat, Bonair, Lorali, Lasbela, Kech/Turbat, Killa Saifullah, Upper Dir, Hangu, Sibbi, Peshin, Kalat, Kachi/Bolan, Khuzdar, Mastung, Kharan, Kohistan, Ziarat, Awaran, Dera Bugti, Gwadar, Killa Abdullah, Nuski, and Washuk. These districts are characterized by a high-risk profile due to the persistently low levels of rural income and limited access to agriculture credit, which exacerbates their economic vulnerability. A high level of government intervention is required in these districts to

boost the income of rural households, thereby mitigating the effects of poverty and promoting sustainable economic development.

6.2.4 Conclusion

This study's rigorous examination of the relationship between formal agricultural credit and rural income in Pakistan, utilizing advanced econometric modeling and sophisticated typological analysis, has yielded nuanced insights into the complex dynamics of rural economic development. The findings reveal a positive yet inelastic association between agricultural credit and rural income, with a 1% increase in credit corresponding to a 0.0612% rise in income. This inelasticity underscores the multifaceted nature of rural poverty alleviation, suggesting that while credit access is beneficial, it is not a panacea for rural economic challenges.

The study's quadrant analysis and six-category typology have illuminated significant spatial heterogeneity in the credit-income nexus across Pakistan's districts. This granular approach has identified regions of high performance, where credit and income growth are synergistic, as well as areas of disconnect, where high credit disbursement has not translated into commensurate income gains. Particularly concerning are the 'Lagging Quadrant' districts, characterized by both low credit access and stunted income growth, signaling acute vulnerability and the urgent need for targeted interventions.

Educational factors emerged as significant correlates of rural income, with middle school enrolment positively associated with income growth. However, the counterintuitive negative relationship between primary male enrolment and income hints at complex opportunity costs in rural education. These findings underscore the

need for a holistic approach to rural development that considers the interplay between financial services, education, and labor market dynamics.

The study's results have profound implications for policy formulation and rural development strategies. They suggest that while expanding agricultural credit is beneficial, its effectiveness may be constrained by structural barriers, market inefficiencies, and limitations in complementary inputs. Therefore, policymakers should consider a multi-pronged approach that combines credit provision with initiatives to enhance financial literacy, improve market access, and address infrastructural deficits in rural areas.

Future research directions should explore several key areas to build upon the findings of this study. To gain a more delicate understanding of the complex credit dynamics in rural areas, future research should employ qualitative methods at the district level, focusing on the informal credit systems along with formal credit that play a vital role in rural economies of developing countries. This could lead to a more comprehensive understanding of the rural financial ecosystem and potentially uncover novel avenues for income augmentation. By integrating both quantitative and qualitative approaches, future studies can provide a more detailed and accurate portrayal of the rural financial landscape, ultimately informing evidence-based policies and interventions that address the unique challenges faced by rural households in Pakistan.

In conclusion, this study contributes to the growing body of literature on rural finance and development by providing a subtle, data-driven perspective on the role of agricultural credit in rural income generation. The findings emphasize the need for tailored, context-specific interventions that go beyond credit provision to address the

multidimensional challenges of rural poverty. As Pakistan continues to strive for inclusive economic growth, these insights offer a valuable foundation for evidence-based policymaking and targeted rural development initiatives.

CHAPTER 7

CONCLUSION, RECOMMENDATIONS, AND LIMITATIONS

This chapter synthesizes the comprehensive findings of our investigation, offering a critical analysis of policies about the accuracy of welfare estimations, and culminating in forward-looking recommendations. Section 7.1 presents a holistic conclusion, encapsulating the key insights derived from this research endeavor. Section 7.2 delves into a nuanced examination of policy frameworks relevant to the precision of welfare estimates. This policy review scrutinizes the sampling methodology employed in the HIES, facilitating a rigorous comparative analysis between HIES direct estimates and model-based projections. Furthermore, it explores potential enhancements to the food consumption model utilized in HIES, intending to refine food insecurity assessments through optimal utilization of HIES data. Finally, this section delineates evidence-based recommendations emerging from this study, providing a roadmap for future research and policy interventions in this domain.

7.1Conclusion

This comprehensive research study delves into the intricate world of Small Area Estimation techniques, focusing on their application in generating poverty and food insecurity data at the district level in Pakistan. The study's primary objective is to compare and evaluate three prominent SAE methods: the Elbers, Lanjouw, and Lanjouw (ELL) method (2003), the Extended ELL method (Nguyen et al., 2018), and the Census Empirical Best (EB) method, which is an extension of the EBP approach by Molina and Rao (2010). The research addresses a critical gap in the availability of

sub-population level data, which has long impeded researchers and policymakers from making accurate inferences about welfare situations at small area levels. This limitation has consequently hindered the development of targeted and effective policies aimed at reducing poverty and food insecurity, ultimately impacting sustainable economic growth. This work further shows that poverty is not uniform across the regions. A province with less poverty may contain several districts that are Extremely Poor, alongside districts that have minimal poverty. This exposes pockets of acute deprivation that were previously obscured by the larger, better-off populations in the same region.

Furthermore, the study extends its scope beyond mere estimation to explore the relationships between rural poverty, rural transformation, and agricultural credit at the district level. This multi-faceted approach provides a holistic view of the socio-economic dynamics at play in rural Pakistan, offering valuable insights for policymakers and researchers alike. The significance of this research lies in its potential to revolutionize the way welfare data is generated and analyzed in Pakistan. By providing more accurate and localized estimates of poverty and food insecurity, it empowers government policymakers to craft well-designed, targeted interventions that address the specific needs of different districts. This granular approach to policy-making has the potential to significantly enhance the effectiveness of poverty reduction strategies and contribute to more sustainable and inclusive economic growth in Pakistan. The district-level poverty estimates shift the understanding of poverty in Pakistan from broad patterns (national or provincial-specific averages) to actionable, district-specific insights. It reveals not only where deprivation is concentrated (such as

extremely deprived areas) but also why certain districts remain trapped in deprivation (for example, due to a lack of rural transformation). This deeper spatial and structural understanding supports more effective, evidence-based policymaking for balanced regional development.

The research employs a rigorous model selection process, utilizing both Ordinary Least Squares and GLS regressions to estimate per-adult equivalent consumption expenditure and food insecurity headcount (measured by kilocalories per day per capita) at the district level. This dual approach allows for a comprehensive comparison of different estimation techniques and their suitability for small-area estimation in the Pakistani context. For the poverty model, the study finds that most predictors are statistically significant at the 5% level, with the exception of some province-level controls. This high level of significance is crucial for the ELL (2003) method, which relies on the statistical significance of GLS parameters when imputed into census data. The comparison between OLS and GLS estimates reveals minimal disparities across most variables, except for province-level indicators. Notably, the ELL method for error decomposition produces GLS coefficients that are more closely aligned with OLS estimates compared to the Henderson Method III (H3).

In the case of food insecurity estimation, a similar pattern emerges. Most predictors show statistical significance. The strong statistical significance of most coefficients provides a solid foundation for extrapolating food insecurity predictions to small-area levels. The study employs and contrasts two distinct methodological approaches for small area estimation: the original ELL (2003) framework and an extended ELL method used by Nguyen et al. (2018). While both approaches share an identical beta

model, they differ in their treatment of the alpha model and the derivation of GLS-based parameter estimates. The extended ELL method incorporates the Henderson III (H3) model for error decomposition and implements Empirical Bayes (EB) prediction, potentially enhancing the accuracy of area-level estimates. A critical finding is that the ELL-based model setting demonstrates a higher adjusted R-squared value of 0.57 for the Beta model when poverty is the dependent variable, and R -squared value is just 0.18 when dependent variable is food insecurity. The finding suggests that food consumption model in HIES 2018-19 does not comprehensively cover all dimensions necessary to measure food insecurity based on the caloric method. Similarly, R-squared for the ELL method is higher in the alpha model compared to the H3 model setting. Furthermore, the ratio of sigma eta squared over MSE is lower in the ELL-based model, suggesting better performance, although the difference is marginal. The comparative assessment of OLS and GLS estimates reveals that GLS coefficients derived using the ELL approach exhibit stronger concordance with OLS counterparts compared to those generated via the H3 method. This alignment suggests that the ELL method may introduce more subtle adjustments to the initial OLS estimates, potentially preserving more of the underlying data structure. These findings underscore the robustness of the chosen modeling approach across various estimation techniques, providing a solid foundation for subsequent predictive analyses and small-area estimations in socioeconomic research. The consistency of results across different model specifications bolsters confidence in the reliability of these findings for small-area estimation purposes, particularly in the context of poverty and food insecurity measurement in Pakistan. For poverty estimation, the analysis reveals that while both

methods produce similar overall patterns of poverty distribution, the traditional ELL method consistently demonstrates lower MSE values across domains. This suggests the superior accuracy and reliability of the ELL approach in estimating district-level poverty in Pakistan. Similarly, for food insecurity estimation, measured via kilocalories per capita per day, the ELL method outperforms the ELL_H3EB approach. It shows a lower median MSE and reduced variability across simulations, indicating enhanced consistency and reliability in small-area food insecurity estimates.

The study also compares these model-based estimates with direct estimates from the Household Integrated Economic Survey (HIES 2018-19). In scenarios with limited sample sizes, the model-based approaches, particularly the ELL method, demonstrate substantially lower MSE than direct estimation methods for poverty and food insecurity indicators. This finding underscores the value of small-area estimation techniques in producing more reliable estimates, especially in areas with limited survey coverage. The research provides detailed poverty and food insecurity mapping at the district level, offering nuanced visualizations of these socioeconomic challenges across Pakistan. Key findings include that 16 districts identified with extreme poverty, 19 districts experiencing severe poverty and Several districts categorized under moderate poverty but close to the severe poverty threshold. Similarly, in the case of food insecurity, 11 districts are observed with extreme food insecurity, and 30 districts face severe food insecurity. Similarly, 14 districts are identified as epicenters of dual vulnerability, experiencing both extreme or severe poverty and food insecurity These granular insights into the spatial distribution of poverty and food insecurity provide crucial information for policymakers and development practitioners. They highlight areas

requiring urgent attention and targeted interventions, allowing for more efficient allocation of resources and tailored policy responses. The study's methodology and findings represent a significant advancement in the application of small-area estimation techniques in Pakistan. By demonstrating the superior performance of the ELL method in this context, the research provides a robust framework for future welfare studies in the country and potentially in similar developing economies. Moreover, the identification of districts facing compounded challenges of poverty and food insecurity offers a novel perspective on vulnerability. This intersectional approach to analysing deprivation can inform more holistic and effective intervention strategies, addressing multiple dimensions of socioeconomic challenges simultaneously. In conclusion, this research not only contributes to the methodological discourse on small area estimation but also provides actionable insights for poverty alleviation and food security efforts in Pakistan. The study's findings can serve as a valuable tool for evidence-based policymaking, enabling more targeted and effective interventions to address these critical socioeconomic issues at the district level.

The research further delves into the application and comparison of advanced SAE techniques, specifically the Census EB method and the ELL methodology. This comprehensive analysis provides crucial insights into the efficacy of these methods for estimating poverty and food insecurity at the district level in Pakistan. The study's methodological rigor is evident in its multifaceted approach to model development and validation. The implementation of the REML method for parameter estimation, coupled with a thorough variable selection process, underscores the study's commitment to statistical robustness. The swift convergence of the Expectation-

Maximization algorithm, particularly in the poverty model, indicates a well-specified model and stable solution. A key strength of this research lies in its comparative analysis of estimation methods. The study convincingly demonstrates the superiority of model-based estimates, particularly the Census EB approach, over direct estimation methods. This is especially notable in areas with small sample sizes, where the MSE of direct estimates tends to be significantly higher. The visual representation of these findings through box plots and figures provides clear evidence of the Census EB method's enhanced reliability and precision. The research also contributes significantly by addressing the often-overlooked issue of outliers in SAE. By employing a conservative approach based on multiple criteria (Cook's distance, studentized residuals, and leverage values), the study ensures a more accurate identification of true outliers, thereby enhancing the overall model reliability. In the context of food insecurity estimation, the study reveals similar patterns of improved performance for the Census EB model-based estimates. This consistency across different welfare indicators (poverty and food insecurity) reinforces the robustness of the Census EB method in various SAE applications.

The comparative analysis between the Census EB approach and the ELL methodology represents a critical advancement in the field of SAE. By building upon previous findings and conducting a systematic comparison, the study provides valuable insights into the relative strengths of these methods. This comprehensive evaluation offers practitioners and policymakers a clear framework for selecting the most appropriate estimation technique for welfare mapping in Pakistan and potentially in similar contexts. The research presents a comprehensive analysis of poverty and food

insecurity in Pakistan, utilizing advanced small-area estimation techniques. The study's findings demonstrate the superior performance of the CEBP method over the ELL methodology in estimating poverty at the district level. This superiority is evidenced by consistently lower MSE values associated with CEBP, indicating more precise and reliable poverty estimates. The study under CEBP identified 16 districts experiencing extreme poverty conditions and 16 districts categorized under severe poverty. Similarly, 13 districts are found to face extreme food insecurity, with an additional 21 districts classified under severe food insecurity. Notably, several districts, including Awaran, Khuzdar, Killa Abdullah, and Sheerani, exhibit extreme levels in both poverty and food insecurity categories, highlighting areas of critical concern requiring immediate intervention. Finally, 13 districts that both ELL and Census EB methods considered as extremely poor, and 3 districts that are specifically considered as "Extremely Poor" by CEBP require urgent interventions. The research also explores the relationship between RT and poverty alleviation. Through econometric modeling and typological analysis, the study reveals a significant negative association between measures of rural transformation and rural poverty headcount. This finding underscores the critical importance of structural changes in rural areas for poverty reduction efforts. Furthermore, the study examines the impact of formal agricultural credit on rural income. The findings indicate a positive yet inelastic relationship between credit access and income growth, suggesting that while beneficial, credit alone is not sufficient to address rural economic challenges comprehensively. The research highlights significant spatial heterogeneity in the credit-income nexus across Pakistan's districts, identifying both high-performing areas and regions where credit disbursement has not

translated into commensurate income gains. These findings have profound implications for policy formulation and rural development strategies in Pakistan. They underscore the need for targeted interventions in the most vulnerable districts, a holistic approach to rural development that considers multiple factors beyond credit access, and tailored policies that account for the specific socioeconomic contexts of different regions. The study provides valuable insights for policymakers to prioritize resources and implement effective strategies in the most vulnerable districts, potentially leading to more equitable development across the country and progress toward achieving the Sustainable Development Goals.

7.2 Recommendations

This study advocates for the adoption of advanced small-area estimation techniques, particularly the Census Empirical Bayes (Census EB) method, which has demonstrated superior performance in estimating poverty and food insecurity at the district level. However, it is recommended to use the ELL method to align estimates and corroborates the findings of superior method, especially when aiming to address user requirements. To facilitate this, capacity-building initiatives for government statisticians and researchers in these advanced techniques are recommended.

To address the spatial heterogeneity in development across Pakistan's districts, the study suggests developing a spatial framework for policy implementation. This framework should account for the varying impacts of credit and rural transformation across different regions, enabling more targeted and effective interventions. The government must adopt strategic, locale-specific planning by prioritizing resource allocation based on district deprivation. Funds for social protection and infrastructure

should be allocated disproportionately to districts classified as Extremely Poor and Severely Poor. Interventions must be customized. Extremely Deprived districts need Survival/Safety Nets; severely deprived districts need Human Capital investments; and Moderately Deprived districts need a focus on economic opportunity. The granular data serves as an early warning system to implement proactive measures, preventing moderate districts from declining into the worst poverty tier.

Improved small area estimates offer clear policy value by enabling a shift from broad provincial targeting to precise district- and sub-district-level interventions. This enhanced granularity strengthens the effectiveness of social protection programs, resource allocation, and development planning in contexts like Pakistan, where spatial disparities in poverty and food insecurity are pronounced. More reliable SAEs facilitate better geographic targeting of cash transfers, nutrition initiatives, and poverty-alleviation measures by identifying pockets of deprivation that national surveys often overlook. They also support evidence-based budgeting, allowing resources to be allocated according to the actual severity of deprivation across districts. Finally, credible SAEs improve policy accountability by enabling finer-scale monitoring of development outcomes and progress toward SDG indicators.

As highlighted by typology analysis RT and rural poverty, the government must shift from uniform welfare programs to targeted interventions informed by district-level dynamics. For optimal districts where both RT and rural poverty reduction are high, the budgetary allocation should focus on maintaining and upgrading existing infrastructure (e.g., road networks, electricity stability) that facilitates these growth drivers. For challenging districts (where either poverty or RT is Slow but the other is

high), deploy demand-driven vocational training centers directly linked to the sectors driving the RT (e.g., logistics, small-scale manufacturing, or value-added agriculture) to upskill the local poor population. Provide micro-finance access specifically for starting businesses in the new RT sectors (for slow RT). Similarly, for slow rural poverty reduction areas, implement policies to strengthen the social safety net and ensure transparent public expenditure. Focus on empowerment programs rather than just welfare handouts, such as securing land tenure or strengthening local institutions representing the poor. Finally, for deprived regions where both RT and rural poverty reduction are slow, the government must treat these areas as priority zones with specific, ring-fenced budgetary allocations, bypassing typical bureaucratic inertia. Interventions cannot be single-sector. A simultaneous approach is needed, including Basic Physical Infrastructure (roads, water, power) to enable RT, Human Capital Investment (new schools, training centres) to enable poverty reduction, and Financial Inclusion (micro-credit access).

The positive yet inelastic relationship of agricultural credit highlights that institutional credit alone is insufficient (due to informal credit, which is easy to grab, and complications associated with institutional credit in Pakistan). The government should initiate interest-free loans for the deprived districts for low-income and credit growth regions, parallel investments in supporting infrastructure, to overcome the structural barriers that limit credit utilization. Credit-centric districts need policies focused on financial literacy and market access to ensure credit is used productively. The practices of highly performing regions (where both credit and income growth are high) should be adopted in the high-risk districts (where both credit and income growth are slow).

Regular updates of district-level poverty and food insecurity maps using the Census EB and ELL method are recommended to track progress and adjust interventions accordingly. Additionally, the integration of these district-level insights into national strategies for achieving the Sustainable Development Goals is advised to ensure more effective poverty and food security alleviation.

The study strongly recommends a significant overhaul of the current sampling and estimation methodologies employed in Pakistan's national surveys, particularly the HIES. A key proposal is the harmonization of HIES and PSLM surveys by establishing administrative districts as independent strata for both rural and urban domains across all provinces, including Balochistan. This alignment would substantially improve the comparability of direct and model-based estimates at the district level, thereby enhancing the overall robustness of small area estimation techniques.

In conclusion, these recommendations aim to significantly enhance the quality and reliability of small-area estimates for poverty and food insecurity in Pakistan. By addressing methodological limitations in current surveys, adopting advanced estimation techniques, and promoting a more nuanced understanding of regional variations, these proposals offer a pathway to more effective, evidence-based policymaking and targeted interventions in the country's ongoing efforts to combat poverty and food insecurity.

7.3 Limitations of the Study

The limitations of this study primarily stem from data constraints and methodological considerations inherent in small area estimation techniques. A key limitation is the utilization of the PSLM survey data in place of census data for small area estimation,

diverging from the conventional approach of pairing survey data with census information. While this substitution is necessitated by data availability, it introduces potential biases and uncertainties that must be acknowledged.

The study operates under the assumption that the PSLM survey, with its larger sample size of 160,654 households across more than 5,600 Primary Sampling Units (PSUs), offers greater reliability compared to the HIES, which encompasses only 24,809 households across 1,802 PSUs. This assumption, while reasonable given the sample sizes, may not fully account for potential differences in survey design, implementation, and specific variables collected between the two surveys.

A notable limitation arises from the independent and random sample selection processes of HIES and PSLM. Unlike scenarios where one survey is a subset of another, the independence of these surveys may introduce additional variability and potential inconsistencies in the estimation process. This challenge is compounded by the exclusion of five Balochistan districts present in HIES but absent from PSLM 2019-20, potentially impacting the comprehensiveness of the small area estimates for that region.

One of the limitations is that the food consumption module of HIES 2018-19 consider only food items consumed or ready-made food. It does not consider data on out-of-home food consumption by working individuals and food items brought home by household members employed in other households.

From a methodological perspective, while powerful, the unit-level methods are susceptible to violations of normality assumptions and the presence of outliers. As highlighted by Rao and Molina (2015), such violations can significantly affect the

reliability of small-area estimates. Moreover, the potential for design bias under informative sampling designs, as noted by Pfeffermann and Sverchkov (2007), presents another layer of complexity that may influence the accuracy of the results. The SAE model-based estimates are design-biased and can have a considerable bias under informative sampling designs or if the defined auxiliary variables do not adequately capture the required phenomena.

These limitations, while significant, do not diminish the value of the study's contributions. Rather, they highlight areas for future research and methodological refinement in the application of small area estimation techniques to poverty and food insecurity analysis in Pakistan and similar contexts.

REFERENCES

Ahmed, U., Ying, L., Bashir, M., Abid, M., Elahi, E., Iqbal, M. J. J. o. A., & Sciences, P. (2016). Access to output market by small farmers: The case of Punjab, Pakistan. *26*(3), 787-793.

Akudugu, M. A. J. A. J. o. A., & Development, R. (2012). Estimation of the determinants of credit demand by farmers and supply by rural banks in Ghana's Upper East Region. *2*(2), 189-200.

Albacea, Z. (1999). Small Area Estimation of Poverty Incidence in the Philippines. *Unpublished PhD Dissertation, University of the Philippines Los Banos*.

Armendariz, B. (2010). *The economics of microfinance*: MIT press.

Asad, M., Mehdi, M., Ashfaq, M., Hassan, S., & Abid, M. J. P. J. A. S. (2019). Effect of marketing channel choice on the profitability of citrus farmers: evidence form Punjab-Pakistan. *56*, 1003-1011.

Barrett, C. B. (2008). Spatial market integration.

Battese, G. E., Harter, R. M., & Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association, 83*(401), 28-36.

Beck, T., Demirgüç-Kunt, A., & Levine, R. J. J. o. e. g. (2007). Finance, inequality and the poor. *12*, 27-49.

Binswanger, H. P., & Khandker, S. R. J. T. J. o. D. S. (1995). The impact of formal finance on the rural economy of India. *32*(2), 234-262.

Booth, J. G., & Hobert, J. P. (1998). Standard errors of prediction in generalized linear mixed models. *Journal of the American Statistical Association, 93*(441), 262-272.

Brackstone, G. J. (1987). Issues in the use of administrative records for statistical purposes. *Survey Methodology, 13*(1), 29-43.

Bui, T. D., & Nguyen, C. V. (2017). Spatial poverty reduction in Vietnam: An application of small area estimation. *Economics Bulletin, 37*(3), 1785-1796.

Butar, F. B., & Lahiri, P. (2003). On measures of uncertainty of empirical Bayes small-area estimators. *Journal of Statistical Planning and Inference, 112*(1-2), 63-76.

Carter, G. M., & Rolph, J. E. (1974). Empirical Bayes methods applied to estimating fire alarm probabilities. *Journal of the American Statistical Association, 69*(348), 880-885.

Carter, M. R., & Olinto, P. J. a. J. o. a. E. (2003). Getting institutions "right" for whom? Credit constraints and the impact of property rights on the quantity and composition of investment. *85*(1), 173-186.

Casas-Cordero Valencia, C., Encina, J., & Lahiri, P. (2016). Poverty mapping for the Chilean comunas. *Analysis of poverty data by small area estimation*, 379-404.

Chambers, R., & Tzavidis, N. (2006). M-quantile models for small area estimation. *Biometrika, 93*(2), 255-268.

Chandio, A. A., Jiang, Y., Wei, F., & Guangshun, X. J. A. F. R. (2018). Effects of agricultural credit on wheat productivity of small farms in Sindh, Pakistan: are short-term loans better? *, 78*(5), 592-610.

Chigbu, U. E. J. D. i. p. (2015). Ruralisation: a tool for rural transformation. *25*(7), 1067-1073.

Christiaensen, L., Lanjouw, P., Luoto, J., & Stifel, D. (2012). Small area estimation-based prediction methods to track poverty: validation and applications. *The Journal of Economic Inequality, 10*, 267-297.

Coad, A., & Tamvada, J. P. J. S. B. E. (2012). Firm growth and barriers to growth among small firms in India. *39*, 383-400.

Corral, P., Himelein, K., McGee, K., & Molina, I. (2021). A Map of the Poor or a Poor Map? *Mathematics, 9*(21), 2780.

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

Correa, L., Molina, I., & Rao, J. (2012). Comparison of methods for estimation of poverty indicators in small areas. *Unpublished report*.

Cressie, N. (1993). Aggregation in geostatistical problems. In *Geostatistics Tróia'92: Volume 1* (pp. 25-36): Springer.

Das, K., Jiang, J., & Rao, J. (2004). Mean squared error of empirical predictor.

Das, S., & Chambers, R. (2017). Robust mean-squared error estimation for poverty estimates based on the method of Elbers, Lanjouw and Lanjouw. *Journal of the Royal Statistical Society Series A: Statistics in Society, 180*(4), 1137-1161.

Das, S. R., & Hanouna, P. J. J. o. F. I. (2009). Hedging credit: Equity liquidity matters. *18*(1), 112-123.

Efron, B., & Morris, C. (1973). Stein's estimation rule and its competitors—an empirical Bayes approach. *Journal of the American Statistical Association, 68*(341), 117-130.

Efron, B., & Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association, 70*(350), 311-319.
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.
Fan, S. (2019). Some lessons from a life in food policy. *Global Food Security, 22*, 33-36.
FAO, F. (2017). The State of Food and Agriculture. Leveraging Food Systems for Inclusive Rural Transformation. In: FAO Rome.
Farooq, S., & Farah, N. J. J. o. I. A. (2023). Developing strategy for rural transformation to alleviate poverty in Pakistan: Stylized facts from panel analysis. *22*(12), 3610-3623.
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.
Feder, G., Lau, L. J., Lin, J. Y., & Luo, X. J. A. J. o. A. E. (1990). The relationship between credit and productivity in Chinese agriculture: A microeconomic model of disequilibrium. *72*(5), 1151-1157.
Foster, J., Greer, J., & Thorbecke, E. (2010). The Foster–Greer–Thorbecke (FGT) poverty measures: 25 years later. *The Journal of Economic Inequality, 8*, 491-524.
Fujii, T. (2014). Dynamic poverty decomposition analysis: An application to the Philippines.
Ghosh, M. (2020). Small area estimation: Its evolution in five decades. *Statistics in Transition. New Series, 21*(4), 1-22.
Ghosh, M., & Rao, J. N. (1994). Small area estimation: an appraisal. *Statistical science, 9*(1), 55-76.
González-Manteiga, W., Lombarda, M., Molina, I., Morales, D., & Santamaría, L. (2010). Small area estimation under Fay–Herriot models with non-parametric estimation of heteroscedasticity. *Statistical Modelling, 10*(2), 215-239.
González-Manteiga, W., Lombardía, M. J., Molina, I., Morales, D., & Santamaría, L. (2008). Bootstrap mean squared error of a small-area EBLUP. *Journal of Statistical Computation and Simulation, 78*(5), 443-462.
Hadley, D. J. J. (2006). Patterns in technical efficiency and technical change at the farm-level in England and Wales, 1982–2002. *57*(1), 81-100.
Hájek, J. (1971). Comment on "an essay on the logical foundations of survey sampling, part one". *The foundations of survey sampling, 236*.
Haslett, S., & Jones, G. (2010). Small-area estimation of poverty: the aid industry standard and its alternatives. *Australian & New Zealand Journal of Statistics, 52*(4), 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, C. R. (1953). Estimation of variance and covariance components. *Biometrics, 9*(2), 226-252.
Hirose, M. Y., & Lahiri, P. (2018). Estimating variance of random effects to solve multiple problems simultaneously. *The Annals of Statistics, 46*(4), 1721-1741.
Holan, S. H., & Wikle, C. K. (2016). Hierarchical dynamic generalized linear mixed models for discrete-valued spatio-temporal data. *Handbook of Discrete-Valued Time Series*, 327-348.
Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association, 47*(260), 663-685.
Huang, J. (2018). Facilitating inclusive rural transformation in the Asian developing countries. *World Food Policy, 4*(2), 31-55.
Huang, J., & Shi, P. (2021). Regional rural and structural transformations and farmer's income in the past four decades in China. *China agricultural economic review*.
Huang, R., & Hidiroglou, M. (2003). Design consistent estimators for a mixed linear model on survey data. *Proceedings of the Survey Research Methods Section, American Statistical Association (2003)*, 1897-1904.
IFAD. (2016). Rural development report 2016: Fostering inclusive rural transformation. In: International Fund for Agricultural Development Rome.
Iqbal, M., Ahmad, M., Abbas, K., & Mustafa, K. J. T. P. D. R. (2003). The impact of institutional credit on agricultural production in Pakistan [with comments]. 469-485.
Jamal, H. (2007). Income poverty at district level: An application of small area estimation technique.
James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112): Springer.

Jayne, T. S., Mather, D., & Mghenyi, E. J. W. d. (2010). Principal challenges confronting smallholder agriculture in sub-Saharan Africa. *38*(10), 1384-1398.
Jiang, J., & Lahiri, P. (2001). Empirical best prediction for small area inference with binary data. *Annals of the Institute of Statistical Mathematics, 53*, 217-243.
Jiang, J., & Lahiri, P. (2006). Mixed model prediction and small area estimation. *Test, 15*, 1-96.
Jiang, J., Lahiri, P., & Wan, S.-M. (2002). A unified jackknife theory for empirical best prediction with M-estimation. *The Annals of Statistics, 30*(6), 1782-1810.
Johnston, B. F. J. J. o. E. l. (1970). Agriculture and structural transformation in developing countries: A survey of research. *8*(2), 369-404.
Khalid Bashir, M., Gill, Z. A., & Hassan, S. J. C. A. E. R. (2009). Impact of credit disbursed by commercial banks on the productivity of wheat in Faisalabad district. *1*(3), 275-282.
Khandker, S. R., & Faruqee, R. R. J. A. e. (2003). The impact of farm credit in Pakistan. *28*(3), 197-213.
Kouamé, E. B.-H. (2010). *Risk, risk aversion and choice of risk management Strategies by cocoa farmers in western Cote D'ivoire.* Paper presented at the CSAE conference.
Krause, M., Lotze-Campen, H., Popp, A., Dietrich, J. P., & Bonsch, M. J. L. U. P. (2013). Conservation of undisturbed natural forests and economic impacts on agriculture. *30*(1), 344-354.
Lahiri, D., Utsuki, T., Chen, D., Farlow, M., Shoaib, M., Ingram, D., & Greig, N. (2002). Nicotine reduces the secretion of Alzheimer's β-amyloid precursor protein containing β-amyloid peptide in the rat without altering synaptic proteins. *Annals of the New York Academy of Sciences, 965*(1), 364-372.
Lahiri, P., & Rao, J. (1995). Robust estimation of mean squared error of small area estimators. *Journal of the American Statistical Association, 90*(430), 758-766.
Li, H., & Lahiri, P. (2010). An adjusted maximum likelihood method for solving small area estimation problems. *Journal of Multivariate Analysis, 101*(4), 882-892.
Lipton, A., & Sepp, A. J. T. J. o. C. R. (2009). Credit value adjustment for credit default swaps via the structural default model. *5*(2), 127-150.
Malik, K. (2014). *Human development report 2014: Sustaining human progress: Reducing vulnerabilities and building resilience*: United Nations Development Programme, New York.
Malik, S. J. J. L. J. o. E. (2008). Rethinking development strategy–The importance of the rural non farm economy in growth and poverty reduction in Pakistan. *13*(Special Edition), 189-204.
Marhuenda, Y., Molina, I., & Morales, D. (2013). Small area estimation with spatio-temporal Fay–Herriot models. *Computational Statistics & Data Analysis, 58*, 308-325.
Mbam, B., & Edeh, H. (2011). Determinants of farm productivity among smallholder rice farmers in Anambra State, Nigeria.
Molina, I. (2019). Desagregación de datos en encuestas de hogares: metodologías de estimación en áreas pequeñas.
Molina, I., Nandram, B., & Rao, J. (2014). Small area estimation of general parameters with application to poverty indicators: a hierarchical Bayes approach.
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.
Molina, 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).
O'Brien, P. K. J. T. E. H. R. (1977). Agriculture and the industrial revolution. *30*(1), 166-181.
Opsomer, J. D., Claeskens, G., Ranalli, M. G., Kauermann, G., & Breidt, F. J. (2008). Non-parametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 70*(1), 265-286.
Pandey, G. K. J. I. J. o. A. E. (2016). Extent, magnitude and determinants of indebtedness among farmers in eastern India: a survey-based study. *71*(4), 450-462.
Pfeffermann, D., & Ben-Hur, D. (2019). Estimation of randomisation mean square error in small area estimation. *International Statistical Review, 87*, S31-S49.
Pfeffermann, D., & Correa, S. (2012). Empirical bootstrap bias correction and estimation of prediction mean square error in small area estimation. *Biometrika, 99*(2), 457-472.
Pfeffermann, D., & Tiller, R. (2005). Bootstrap approximation to prediction mse for state–space models with estimated parameters. *Journal of Time Series Analysis, 26*(6), 893-916.
Purcell, N. J., & Kish, L. (1980). Postcensal estimates for local areas (or domains). *International Statistical Review/Revue Internationale de Statistique*, 3-18.

Rao, J. N., & Molina, I. (2015). *Small area estimation*: John Wiley & Sons.
Rehman, Z. U., Muhammad, N., Sarwar, B., & Raz, M. A. J. F. I. (2019). Impact of risk management strategies on the credit risk faced by commercial banks of Balochistan. *5*, 1-13.
Rosenzweig, M. R., & Wolpin, K. I. J. J. o. p. e. (1993). Credit market constraints, consumption smoothing, and the accumulation of durable production assets in low-income countries: Investments in bullocks in India. *101*(2), 223-244.
Sharma, M., & Zeller, M. J. W. d. (1997). Repayment performance in group-based credit programs in Bangladesh: An empirical analysis. *25*(10), 1731-1742.
Sial, M. H., Awan, M. S., & Waqas, M. (2011). Institutional credit and agricultural production nexus.
Sitko, N. J., & Jayne, T. S. J. F. P. (2014). Structural transformation or elite land capture? The growth of "emergent" farmers in Zambia. *48*, 194-202.
Souza, C. (2015). Breaking the boundary: pro-poor policies and electoral outcomes in Brazilian sub-national governments. *Regional & Federal Studies, 25*(4), 347-363.
Stiglitz, J. E., & Weiss, A. J. T. A. e. r. (1981). Credit rationing in markets with imperfect information. *71*(3), 393-410.
Tarozzi, A., & Deaton, A. (2009). Using census and survey data to estimate poverty and inequality for small areas. *The review of economics and statistics, 91*(4), 773-792.
Tarozzi, A., Desai, J., & Johnson, K. (2015). The impacts of microcredit: Evidence from Ethiopia. *American Economic Journal: Applied Economics, 7*(1), 54-89.
Timmer, C. P. (2017). Food security, structural transformation, markets and government policy. *Asia & the Pacific Policy Studies, 4*(1), 4-19.
Torabi, M., & Rao, J. (2008). Small area estimation under a two-level model. *Survey Methodology, 34*(1), 11.
Torabi, M., & Rao, J. (2014). On small area estimation under a sub-area level model. *Journal of Multivariate Analysis, 127*, 36-55.
Tzavidis, N., Marchetti, S., & Chambers, R. (2010). Robust estimation of small-area means and quantiles. *Australian & New Zealand Journal of Statistics, 52*(2), 167-186.
Tzavidis, N., Zhang, L.-C., Luna, A., Schmid, T., & Rojas-Perilla, N. (2018). From start to finish: a framework for the production of small area official statistics. *Journal of the Royal Statistical Society Series A: Statistics in Society, 181*(4), 927-979.
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*(7028).
Zhang, L.-C., & Fosen, J. (2012). A modeling approach for uncertainty assessment of register-based small area statistics. *Journal of the Indian Society of Agricultural Statistics, 66*, 91-104.
Zhang, L. C., & Giusti, C. (2016). Small area methods and administrative data integration. *Analysis of poverty data by small area estimation*, 61-82.

Appendix A

Table 1 *Household Head Characteristics*

Variables	**Definition**
marital	Marital status
marital_1	1=never married; 0=otherwise
marital_2	1=married; 0=otherwise
marital_3	1=divorced/widowed; 0=otherwise
employ	Employment status of the household heads
employ_1	1=employed; 0=otherwise
employ_2	1=unemployed; 0=otherwise
employ_3	1=not in labor-force; 0=otherwise
sector	Employment sector of the household head
sector_1	1=not in labor-force; 0=otherwise
sector_2	1=unemployed; 0=otherwise
sector_3	1=agriculture; 0=otherwise
sector_4	1=unpaid family worker; 0=otherwise
sector_5	1=paid employee; 0=otherwise
sector_6	1=employer/self-employed; 0=otherwise
edu	Educational attainment
edu_1	1=No Schooling; 0=otherwise
edu_2	1=Primary (grade 1-5); 0=otherwise
edu_3	1=Middle (grade 6-8); 0=otherwise
edu_4	1=Lower secondary (grade 9-10); 0=otherwise
edu_5	1=Upper secondary (grade 11-12); 0=otherwise
edu_6	1=Undergraduate; 0=otherwise
edu_7	1=Post-graduate; 0=otherwise
age	Age of head of households
lnage	Age in natural logarithm
age2	age^2/100
age3	age^3/1000

Appendix A

Table 2 *Household Characteristics*

Variables	**Definition/Coding**
hhsize	Household Size
hhsize2	hhsize^2/100
hhsize3	hhsize^3/1000
lnsize	Household size in natural logarithm
Sexratio_f	No. of female/HH size
depratio	No. of kids and elderly (<15 or >65 yrs old)/HH size
adult	No. of adults (15-65 yrs old)
men	No. of men (15-65 yrs old)
sexratio	No. of men (15-65 yrs old)/HH size
noadult	1=HH has no adult (15-65 yrs old); 0=otherwise
kid	No. of children (<15 yrs old)
boy	No. of male children (<15 yrs old)
girl	No. of female children (<15 yrs old)
girlratio	No of female children/total no. of children
nokid	1=HH has no kid (<15 yrs old); 0=otherwise
nogirl	1=HH has no girl (<15 yrs old); 0=otherwise
workratio_adult	No. of adults employed (15-65 yrs old)/HH size
highestedu	Highest level of edu attainment in the HH
highestedu_1	1=No schooling, 0=0therwise
highestedu_2	1=Primary (grade 1-5), 0=otherwise
highestedu_3	1=Middle (grade 6-8), 0=otherwise
highestedu_4	1=Lower secondary (grade 9-10), 0=otherwise
highestedu_5	Upper secondary (grade 11-12)
highestedu_6	Undergraduate
highestedu_7	Post-graduate
eduratio_1	Proportion of HH member having no schooling
eduratio_2	Proportion of HH members having primary schooling
eduratio_3	Proportion of HH members having middle schooling
eduratio_4	Proportion of HH members having secondary schooling
kidinschool	No. of children (<15 yrs old) attending school
boyinschool	No. of boys (<15 yrs old) attending school
girlinschool	No. of girls (<15 yrs old) attending school
schoolratio	No. of kids(<15 yrs old) attending school/No. of kids
schoolratio_f	No. of female kids in school/No. of kids in school
language	Language of interview
language_new_1	1=urdu; 0=otherwise
language_new_2	1=punjabi; 0=otherwise
language_new_3	1=sindhi; 0=otherwise
language_new_4	1=pashto; 0=otherwise
language_new_5	1=balochi/balti/hindko/sariki/other; 0=otherwise

Appendix A

Table 3 *Household Dwelling Characteristics*

Variables	**Definition**
room	How many separate rooms are there in your dwelling?
toilet_detail	What kind of toilet facility does your household use?
toilet_detail_1	1=no toilet, 0=otherwise
toilet_detail_2	1=flush connected to public sewerage, 0=otherwise
toilet_detail_3	1=flush connected to septic tank, 0=otherwise
toilet_detail_4	1=flush connected to pit, 0=otherwise
toilet_detail_5	1=flush connected to open drain, 0=otherwise
toilet_detail_6	1=dry raised latrine, 0=otherwise
toilet_detail_7	1=dry pit latrine, 0=otherwise
toilet_detail_8	1=composting toilet, 0=otherwise
toilet_detail_9	1=other specific, 0=otherwise
distancewater	How far is it from here to reach the nearest drinking water?
distancewater_1	1= 01 to 15 minutes, 0=otherwise
distancewater_2	1= 16 to 30 minutes, 0=otherwise
distancewater_3	1= 31 to 45 minutes, 0=otherwise
distancewater_4	1= 46 to 60 minutes, 0=otherwise
distancewater_5	1= 60+ minutes, 0=otherwise
Ownership	Dwelling ownership (owned, rented or rent-free)
ownership_1	1=owned house; 0=otherwise
ownership_2	1=rented house; 0=otherwise
ownership_3	1=rent-free house; 0=otherwise
floor	1=pacca floor; 0=otherwise
roof	1=pacca roof; 0=otherwise
wall	1=pacca wall 0=otherwise
cooking	1=improved cooking source; 0=otherwise
lighting	1=improved lighting_source; 0=otherwise
water	Main source of drinking water
water_1	1=piped/bottled/filtered water; 0=otherwise
water_2	1=hand-pump; 0=otherwise
water_3	1=motorized-pump/tube-well water; 0=otherwise
water_4	1=well water; 0=otherwise
water_5	1=river/spring/tanker/others; 0=otherwise
pipedwater	1= household has piped water; 0 = otherwise
cleanwater	1=improved drinking water source; 0=otherwise
toilet	Type of toilet
tiolet_1	1=flush toilet; 0=otherwise
toilet_2	1= dry/pit latrine; 0=otherwise
tiolet_3	1= no toilet; 0=otherwise
toilet_type1	1=HH has flush toilet connected to sewer; 0=otherwise
phone	1=HH has a phone connection; 0=otherwise
mobile	1=HH has a mobile; 0=otherwise
internet	1=HH has an internet connection; 0=otherwise
urban	1=urban; 0=rural
computer (including	1=HH has a computer; 0=otherwise

Appendix A

Table 4 *Household Asset Possession*

Variables	**Definition**
agriland	1=Own agricultural land; 0=otherwise
nonagriland	1=Own non-agricultural land; 0=otherwise
resibuilding	1=Own residential building; 0=otherwise
combuilding	1=Own commercial building; 0=otherwise
remit_inside	1=received remittance from w/i Pakistan; 0=otherwise
remit_outside	1=received remittance from outside Pakistan;
radio	1 = household owns item ; 0 otherwise
televsion	1 = household owns item ; 0 otherwise
lcd/Led	1 = household owns item ; 0 otherwise
refrigerator	1 = household owns item ; 0 otherwise
freezer	1 = household owns item ; 0 otherwise
washingmachine	1 = household owns item ; 0 otherwise
dryer	1 = household owns item ; 0 otherwise
airconditioner	1 = household owns item ; 0 otherwise
aircooler	1 = household owns item ; 0 otherwise
fan	1 = household owns item ; 0 otherwise
stove	1 = household owns item ; 0 otherwise
cookingrange	1 = household owns item ; 0 otherwise
microwave	1 = household owns item ; 0 otherwise
sewing Machine	1 = household owns item ; 0 otherwise
knitting Machine	1 = household owns item ; 0 otherwise
iron	1 = household owns item ; 0 otherwise
water Filter	1 = household owns item ; 0 otherwise
donkey Pump	1 = household owns item ; 0 otherwise
turbine	1 = household owns item ; 0 otherwise
chair	1 = household owns item ; 0 otherwise
table	1 = household owns item ; 0 otherwise
ups	1 = household owns item ; 0 otherwise
generator	1 = household owns item ; 0 otherwise
solar Panel	1 = household owns item ; 0 otherwise
heater	1 = household owns item ; 0 otherwise
geaser	1 = household owns item ; 0 otherwise
bicycle	1 = household owns item ; 0 otherwise
motorcycle	1 = household owns item ; 0 otherwise
car	1 = household owns item ; 0 otherwise
van_truck_bus	1 = household owns item ; 0 otherwise
motorboat	1 = household owns item ; 0 otherwise
tractor_trolley	1 = household owns item ; 0 otherwise
watch	1 = household owns item ; 0 otherwise
animaldrawncart	1 = household owns item ; 0 otherwise

Appendix B

Table-1 *Calories Table*

Description	Unit	KCAL	KCAL 100	Edible Portion	Density
Almond, Walnut, Pistachio, Kaju,	Gm	5.71	591.94	0.91	1
Alou Bukhara (Plum) , Apricot (Khobani) , Guava	Kg	570.95	62.43	0.91	1
Apple	Kg	513	57	0.9	1
Bananas	No	113.66	96	0.74	1
Barfi, Jaleebi, Halwa & other Sweetmeats	Kg	4152.5	415.25	1	1
Beans (lobia red and white)	Kg	3600	360	1	1
Beef Fresh/Frozen	Kg	1830	244	0.75	1
Biscuits-packed/Bakery(Loose)	Gm	2.63	263	1	1
Bitter Gourd, Lady finger, Brinjal, Cucumber	Kg	220	25	0.88	1
Bread plain	No	263	263	1	1
Butter /Margarine (Loose/ packed)	Gm	7.21	721	1	1
Butter products (Butter oil/Desi Ghee)	Kg	9000	900	1	1
Cake , Bakar khani	Gm	2.63	2.63	1	1
Canned Fruits	Gm	0.49	57.14	1	1
Canned Vegetables /Olives/ Sweet corn, Mushrooms etc	Gm	0.34	34.2	1	1
Cardamom (Loose/Packed)	Gm	3.26	326	1	1
Cauliflower	Kg	220	25	0.88	1
Cheese	Gm	7.21	721	1	1
Chicken Meat (Fresh/ Frozen)	Kg	1346.4	187	0.72	1
Chillies (Loose/Packed)	Gm	0.35	35	1	1
Coffee	Gm	2.92	292	1	1
Cold Drink (Carbonated)	Ltr	602.55	58.5	1	1.03
Cooking Oil (Tin/Loose)	Ltr	9000	900	1	1
Coriander Seeds, Turmeric Powder, Curry powder	Gm	3.46	346	1	1
Cumin Seeds, Pepper Black, Cloves	Gm	3.03	302.67	1	1
Curd / Yogurt (Loose/packed)	Kg	710	71	1	1
Custard Powder /Jelly powder	Gm	4.15	415.25	1	1
Dry milk powder for children	Gm	4.66	466	1	1
Eggs	No	80.91	155	0.87	1
Fish (fresh,frozen, dried)	Kg	787.8	101	0.78	1
Fruit Juice Fresh/ sugarcane juice	Ltr	602.55	58.5	1	1.03
Fruit Juice Packed	Ltr	602.55	58.5	1	1.03
Garlic	Gm	1.03	121	0.85	1
Ginger (Powder /Paste/whole)	Gm	0.45	53	0.85	1
Glucose , Energile , Tang, Lemon Pani, etc.	Gm	0.58	58.5	1	1
Gram whole Black/White	Kg	3600	360	1	1
Grapes	Kg	703	74	0.95	1
Gur, Shaker	Kg	3405	340.5	1	1
Honey	Gm	3.1	310	1	1
Ice Cream (Cup only)	No	830.5	415.25	1	1
Jam /Marmalade / Jellies	Gm	3.9	390	1	1
Lassi made with yogurt ,Olala , umang,milo, etc.	Ltr	319.3	31	1	1.03
Lemon	Gm	247.62	29.83	0.83	1
Maida, Suji, Besan	Kg	3600	360	1	1

Source: Planning Commission of Pakistan

Appendix C

Table 1 Poverty Status of Deprived Districts 2019-20

Status	Common Districts under both ELL and CEBP Methods	Change in Status of Districts under ELL Or CEBP
EXTREME POVERTY	Sheerani, Awaran, Khuzdar, Killa Abdullah, Shaheed Sikandarabad, Ziarat, Harnai, Duki, Mohmand, Kalat, Tharparkar, Sujawal and Nasirabad.	Badin, Barkhan, and Killa Saifullah, reported as Extremely Poor by CEBP.
SEVERE POVERTY	Bajaur, Dera Bugti, Jaffarabad, Kachhi, Kharan, Khyber, Mirpur Khas, Nuski, Orakzai, Shaheed Benazir Abad, Sohbatpur, South Waziristan, Tando Muhammad Khan, Thatta, Umerkot, Washuk	Badin, Barkhan, and Killa Saifullah, reported as Severely Poor by ELL.
MODERATE POVERTY	Bahawalnagar, Bahawalpur, Bannu, Batagram, Bhakkar, Buner, Charsadda, Chiniot, Chitral, Dadu, Dera Ghazi Khan, Dera Ismail Khan, Ghotki, Gwadar, Hangu, Jacobabad, Jamshoro, Jhang, Karak, Kashmore, Kech, Khairpur, Khanewal, Khushab, Kohat, Kohistan, Kohlu, Kurram, Lakki Marwat, Larkana, Lasbela, Layyah, Lodhran, Loralai, Lower Dir, Malakand PA, Mardan, Mastung, Matiari, Mianwali, Multan, Muzaffargarh, Narowal, Naushahro Feroze, North Waziristan, Nowshera, Peshawar, Pishin, Quetta, Rahim Yar Khan, Rajanpur, Sanghar, Shahdadkot, Shangla, Shikarpur, Sibi, Sukkur, Swabi, Swat, Tando Allahyar, Tank, Tor Ghar, Upper Dir, Vehari.	No Changes