

Estimation of Health Indicator at Dis-aggregated Geographic Levels



By

Iqra Soomro

Reg. # 00000402927

Department of Mathematics and Statistics

School of Natural Sciences

National University of Sciences and Technology

H-12, Islamabad, Pakistan

August 2024

Estimation of Health Indicator at Dis-aggregated Geographic Levels



By

Iqra Soomro
Reg.# 00000402927

A thesis submitted in partial fulfilment of the requirements
for the degree of **Master of Science**
in
Statistics

Supervised by: Dr. Shakeel Ahmed

Department of Mathematics and Statistics

School of Natural Sciences
National University of Sciences and Technology
H-12, Islamabad, Pakistan

August 2024

Student Sign
Missing
Joan TH-B

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS thesis written by **Iqra Soomro** (Registration No **00000402927**), of **School of Natural Sciences** has been vetted by undersigned, found complete in all respects as per NUST statutes/regulations, is free of plagiarism, errors, and mistakes and is accepted as partial fulfillment for award of MS/M.Phil degree. It is further certified that necessary amendments as pointed out by GEC members and external examiner of the scholar have also been incorporated in the said thesis.

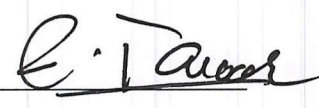
Signature: 

Name of Supervisor: Dr. Shakeel Ahmed

Date: 09-08-2024

Signature (HoD): 

Date: 12-8-2024

Signature (Dean/Principal): 

Date: 12.08.2024

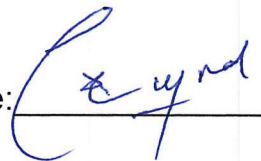
National University of Sciences & Technology

MS THESIS WORK

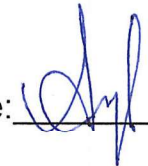
We hereby recommend that the dissertation prepared under our supervision by Iqra Soomro, Regn No. 00000402927 Titled Estimation of Health Indicator at Dis-Aggregated Geographic Levels be Accepted in partial fulfillment of the requirements for the award of MS degree.

Examination Committee Members


1. Name: PROF. TAHIR MEHMOOD

Signature: 

2. Name: DR. AYESHA NAZUK

Signature: 


Supervisor's Name: DR. SHAKEEL AHMED

Signature: 


Head of Department

12-8-2024
Date

COUNTERSIGNED

Date:  12-08-2024


Dean/Principal

*"This Thesis is dedicated to my
beloved Parents, Brothers, and my
respected supervisor Dr. Shakeel
Ahmed"*

Acknowledgements

Utmost Sense of obligation to Allah Almighty for giving me an insight and strength to undertake this research objectively and efficaciously.

"He [Allah] grants wisdom to whom He pleases; and He to whom wisdom is granted indeed receive a benefit overflowing. But none will grasp the message except men of understanding"

I consider it an honour to express my heartfelt appreciation to my respected supervisor **Dr. Shakeel Ahmed** for his continued support, guidance and encouragement during this strenuous work has been a constant source of learning during my stay at School of Natural Sciences (SNS), NUST. One could not have asked for a more amicable and supportive mentor and supervisor. I would like to express my sincere thanks to the members of my Guidance and Examination Committee (GEC), **Dr. Tahir Mehmood** and **Dr. Ayesha** who were extremely reliable source of practical knowledge.

I take pride in acknowledging the insightful guidance and support of **My Parents** for their selfless love and showing faith in me. Heartfelt gratitude to my brothers **Shahzad Soomro** and **Sheezan Soomro** for their cherished prayers and their backing which made me confident throughout my journey towards this degree. I would never be able to pay back the love and affection showered upon by my Parents and Brothers.

Abstract

This study leverages data from the Demographic and Health Surveys (DHS) conducted in 2017-18 and 2019 to examine the utilization of antenatal care (ANC) services among women across various districts in Pakistan. The analysis employs a mixed-methods approach, integrating both area-level and unit-level auxiliary information to enhance the robustness of the findings. Specifically, the Fay-Herriot (FH) model is utilized for area-level data, capturing district-level healthcare infrastructure and socio-economic indicators, while the Battese-Harter-Fuller (BHF) model is applied to unit-level data, encompassing individual socio-demographic factors such as age, education, and wealth index. By merging these models, the study aims to provide a comprehensive understanding of ANC utilization patterns and their determinants at multiple levels.

The combined approach not only enables the identification of nuanced relationships between various factors influencing ANC utilization but also facilitates the generation of more reliable and precise estimates. The results indicate significant variations in ANC coverage, highlighting critical areas where interventions are needed to improve maternal health outcomes. The integration of both FH and BHF models allows for a detailed examination of both district-level and individual-level factors, offering valuable insights for policymakers and healthcare providers aiming to enhance maternal healthcare services in Pakistan. This study underscores the importance of utilizing advanced statistical models to address data limitations and improve the accuracy of small area estimates. **Keywords:** Small area estimation, Antenatal care, Area level models, Unit level models.

Contents

List of Figures	iv
List of Tables	v
List of Abbreviations	vi
1 Introduction of the Study	1
1.1 Background	1
1.2 Definition of Terminologies	3
1.2.1 Small Area Estimation	3
1.2.2 Model-based Estimation	3
1.2.3 Area Level Model	4
1.2.4 Unit Level Model	4
1.2.5 Antenatal Care	4
1.3 Review of Literature	5
1.4 Research Gap	9
1.5 Objective of the Study	9
2 Methodology	11
2.0.1 Area Level Model	11
2.0.2 Unit Level Model	16
2.1 Area-cum-Unit level model for SAE	21

3	Results and Discussions	28
3.1	Description of study	28
3.2	Results	29
4		54
4.1	Conclusion	54
	Bibliography	55

List of Figures

3.1	Flowchart for Obtaining Direct Estimate	31
3.2	Flowchart for Area Level Model	32
3.3	QQ Plots for Realized Residuals and Random Effects in Fay-Herriot Model	33
3.4	Density Plot of Standardized Realized Residuals Divided by Square Root of Direct Variance	34
3.5	Comparison of Direct and Fay-Herriot Model-Based Estimates	34
3.6	Flowchart for Unit Level Model	36
3.7	Flowchart illustrating the integration of surveys and application of the ACU method.	37
3.8	Flowchart of SAE under ACU model	38
3.9	Comparison of Proportion Estimates Using Different Methods	44
3.10	MSE of Different Methods	45
3.11	CV of Different Methods	46
3.12	ANC distribution by Districts in Pakistan. Source: PDHS 2017-18 and PMMS 2019.The districts are represented by their codes and details are given in Appendix	52

List of Tables

3.1	Variables Discription	30
3.2	Summary Proportion of estimates under different methods	40
3.3	Summary MSE of estimates under different methods	42
3.4	Summary CV of estimates under different methods	43
3.5	District wise Estimates	47

List of Abbreviations

SDG	Sustainable Development Goals
SAE	Small Area Estimation
EBLUP	Empirical Best Linear Unbiased Predictor
EB	Empirical Bayes
FH	Fay-Herriot
BHF	Battese, Harter and Fuller
DHS	Demographic and Health Survey
REG1 SYN	Regression Synthetic
MSE	Mean Square Error
BP	Best Predictor
BLUP	Best Linear Unbiased Predictor
REML	Restricted Maximum Likelihood
ANOVA	Analysis of Variance
ML	Maximum Likelihood
SRS	Simple Random Sampling
ACU	Area-Cum-Unit
ANC	Antenatal Care
CV	Coefficient of Variation
SD	Standard Deviation
LNOB	Leave No One Behind
LWB	Low Birth Weight
GIS	Geographic Information System
ME	Measurement Error

NR

Nonresponse Error

HIV

Human Immunodeficiency Virus

Chapter 1

Introduction of the Study

1.1 Background

With the adoption of the 2030 Agenda for Sustainable Development, the importance of generating high-quality, disaggregated estimates for Sustainable Development Goal (SDG) indicators has significantly increased. Often, sample surveys face constraints such as insufficient sample size to provide reliable direct estimates for all sub-populations or the inability to cover all potential disaggregation domains. To overcome these constraints, indirect estimation methods like small area estimation (SAE) techniques are utilized[1]. The SDGs are crafted to be ambitious enough to drive transformative change while allowing each country to implement them according to their unique domestic contexts. In this context, 'leave no one behind', a central commitment to the SDGs [2]. The SDG outcome document articulates this commitment as follows: "As we embark on this great collective journey, we pledge that no one will be left behind. Recognizing that the dignity of the human person is fundamental, we wish to see the Goals and targets met for all nations and peoples and for all segments of society. And we will endeavor to reach the furthest behind first" (UNGA Resolution, 2015). The 'leave no one behind' pledge addresses two interconnected issues: ending absolute poverty in all its forms and ensuring that those who have been left behind (in relative or absolute terms) can catch up with those who have experienced more progress[3]. In adherence to the Sustainable Development Goals (SDGs) principle of data disaggregation, it is emphasized that indicators should be broken down by various demographic factors such

as income, gender, age, race, ethnicity, migration status, disability, and geographic location. This disaggregation is crucial for obtaining a comprehensive understanding of development progress and ensuring inclusivity. Indicator 17.18.1 specifically measures the proportion of sustainable development indicators produced at national level with full disaggregation, aligning with the Fundamental Principles of Official Statistics, examples of SDG indicators with specific disaggregations illustrate the importance of detailed data analysis for effective policymaking and program implementation. For instance, Indicator 1.1.1 assesses the proportion of the population below the international poverty line, disaggregated by sex, age group, employment status, and urban/rural location. Similarly, Indicator 4.1.1 evaluates the proficiency levels of children and young people in the reading and mathematics, with disaggregations including sex, location, and wealth. Indicator 8.5.1 examines average hourly earnings by occupation, age group, and disability status, while Indicator 10.2.1 analyzes the proportion of people living below 50 per cent of median income, disaggregated by age group, sex, and disability status. These examples underscore the necessity of detailed data disaggregation in monitoring and addressing development challenges effectively. Household surveys serve as a primary source of data regarding living conditions and are invaluable for producing disaggregated information across various population groups. These surveys are meticulously planned to generate data for specific study domains, such as the unemployment rate in urban and rural areas. However, there are instances where the predefined study domains may not cover all population subgroups of interest. For example we may want to examine school attendance among children aged 6-12 years, categorized by income quintile, a subgroup not initially addressed in the survey design. In such cases, disaggregation may face challenges. These challenges include a lack of information due to no observed cases and low precision resulting from small sample sizes or high coefficients of variation. These limitations highlight the importance of carefully designing surveys to capture a comprehensive representation of the population and ensuring the reliability and usefulness of data [4].

1.2 Definition of Terminologies

1.2.1 Small Area Estimation

Small area estimation refers to the statistical methods employed to derive reliable estimates for specific geographic or socio-demographic domains when the available sample size within those domains is insufficient for direct estimation. These domains, often termed "small areas," encompass regions like counties, municipalities, or subpopulations defined by age, sex, or race. In situations where direct estimation isn't feasible due to limited sample sizes, small area estimation employs indirect methods, leveraging information from related areas or time periods to enhance the precision of estimates. These indirect estimators, including domain, time, and domain-time combinations, rely on models that incorporate supplementary data such as census counts or administrative records. By borrowing strength from similar areas or historical data, small area estimation enables the generation of precise estimates even when direct sampling yields inadequate representation within specific domains.[\[4\]](#).

1.2.2 Model-based Estimation

In small area estimation, it's become clear that using specific models for each area is the way to go when we're estimating indirectly. These models decide how we use related data to make our estimates. This method has a few advantages: it helps us find the best estimates under the model we've chosen, lets us look at the variability of each area separately, and we can check if our models work using real data. We can use different models depending on the data and how complex it is. One popular method we talk about is called EBLUP, which helps estimate things like averages in small areas using special math. Another method is EB, which can be used more widely. With EBLUP, we first find the best guess for our numbers, then adjust it using some math to make it even better. But sometimes, even if our estimates are the best, they might not match exactly what's really going on. This is where things like rankings or finding the most extreme areas come in. So, we need to find a good balance between getting the numbers right and looking at other important factors. It's like trying to

make everyone happy, which isn't always easy. But we have ways to do it, like making sure our estimates are close to what's really happening while still considering other important stuff.[4].

1.2.3 Area Level Model

The regression–synthetic estimators rely on a linear regression model incorporating auxiliary information. Specifically, the REG1-SYN estimator is applied when only aggregate, or area-level, auxiliary data is accessible. In this context, we designate x_d as the vector encompassing p available auxiliary variables. It is assumed that the target indicator, δ_d (for instance, the mean of the area), consistently changes concerning these aggregated data x_d across all areas. The actual values of the indicator for the areas, which represent the target parameters, are not accessible. Instead of these true values, direct estimators denoted as $\hat{\delta}_d$, where d ranges from 1 to D , are utilized. Therefore, the area–level model operates under the assumption of employing these estimated values rather than the unobserved true parameters.[4]

1.2.4 Unit Level Model

Model-based estimators are categorized as indirect methods since they rely on support from different regions. These estimators factor in variations between areas that cannot be entirely clarified by the existing auxiliary variables. The area random effects within the model account for this unexplained variability, providing a more comprehensive understanding of the data.[4]

1.2.5 Antenatal Care

Antenatal care (ANC) refers to the medical attention given by skilled healthcare professionals to pregnant women and adolescent girls to maintain optimal health for both the mother and the baby throughout pregnancy. ANC involves several key components: identifying potential risks, preventing and managing pregnancy-related or concurrent illnesses, and providing health education and promotion. Antenatal care (ANC) helps

lower maternal and perinatal morbidity and mortality by directly identifying and treating pregnancy-related complications. Additionally, it indirectly aids by spotting women and girls who are at higher risk of complications during labor and delivery, ensuring they are referred to the appropriate level of care[5].

1.3 Review of Literature

Small area estimation (SAE) has become increasingly significant in survey sampling due to the rising need for accurate small area statistics from both the public and private sectors. Direct survey estimates for small areas often have large standard errors because of limited sample sizes. To enhance accuracy, methods that "borrow strength" from related areas are employed. These methods include synthetic estimation, sample size-dependent methods, empirical best linear unbiased prediction (EBLUP), empirical Bayes, and hierarchical Bayes estimation. In 1994, Ghosh and Rao emphasized the need for reliable small area statistics and the limitations of direct survey estimates. Their evaluation of various SAE methods using synthetic data indicated that EBLUP and empirical and hierarchical Bayes methods generally offer significant advantages over other methods for most purposes[6]. Rao and Yu (1994) further investigated SAE methods, highlighting the necessity of borrowing strength from related areas. They evaluated various estimation methods using synthetic data, concluding that EBLUP and empirical and hierarchical Bayes methods generally provide more accurate estimates compared to other methods[7]. In 1983, Rachel Margaret Harter proposed a prediction approach for small area estimation using survey and auxiliary data based on a nested-error model assumption. This model, which considers both the mean squared error and mean squared conditional bias of the predictor, forms the basis for the best linear unbiased predictor of the small area mean, taking the form of a James-Stein estimator. This work underscored the importance of integrating different data sources to improve the accuracy and reliability of small area estimates[8]. In 1987, MacGibbon focused on small area estimates of proportions using empirical Bayes techniques. Through a simulation involving a two-stage sample, MacGibbon demonstrated that

empirical Bayes estimators perform better overall compared to classical and synthetic estimators. While the classical estimator is design-unbiased, it suffers from large root mean squared errors (RMSEs) due to limited data. The synthetic estimator, although more stable, fails to adjust for local differences and provides misleading measures of uncertainty. The empirical Bayes estimator balances these approaches by reducing biases, pooling information, and offering useful measures of uncertainty, making it a robust solution in two-stage sampling simulations[9]. In 1988, Battese, Harter, and Fuller developed an error-components model for predicting county crop areas using survey and satellite data. Their study focused on 12 Iowa counties, using data from the 1978 June Enumerative Survey of the U.S. Department of Agriculture and satellite data from LANDSAT. They emphasized predicting the area under corn and soybeans, specifying a linear regression model for the relationship between reported hectares in the survey and satellite-determined areas. A nested-error model was used to define the correlation structure among reported crop hectares. The proposed model showed that the mean hectares per segment, estimated using satellite data and a random county effect, had a considerably lower standard error compared to traditional survey regression predictors[10]. In 2020, Jonathan Wakefield explored SAE for disease prevalence mapping. He investigated both design-based and model-based approaches, particularly Bayesian spatial models, with an emphasis on health applications. Wakefield illustrated these techniques with a case study on HIV prevalence in Malawi and considered their potential application for COVID-19 outcomes, showcasing the adaptability of SAE methods in public health[11]. Lin and Yue (2024) advanced SAE by developing a synthetic population dataset for the United States to address confidentiality concerns. This dataset includes detailed socio-demographic characteristics aggregated at the census block group level, based on the 2010 Census. Their results validated the synthetic data and demonstrated its practical applications in SAE, highlighting its potential for accurate and confidential small area analyses[12]. In 2022, Clara Aida Khalil integrated surveys with geospatial data through SAE to disaggregate SDG indicators, focusing on SDG Indicator 2.3.1. Using a Fay-Herriot area-level SAE model, she demonstrated that modern geospatial information systems facilitate the produc-

tion of high-quality disaggregated estimates of SDG indicators. This research highlights the potential of SAE methodologies in advancing the measurement and monitoring of SDGs[1]. Green et al (2020) explored SAE in the inventory of loblolly pine, aiming to improve precision in volume estimates. They developed area-level SAE models incorporating lidar height percentiles and stand thinning status as auxiliary information. Their findings demonstrated greater precision improvements compared to models using only lidar data, underscoring the potential of SAE methods to reduce uncertainty in forest assessments[13]. Antenatal care and its impact on the occurrence of low birth weight (LBW) delivery among women in the remote mountainous region of Chitral, Pakistan, was studied by Ahmed, Khoja, and Tirmizi. Their mixed methodology study involved structured data collection from medical records of 1316 mothers, followed by interviews and focus group discussions. The study found significant associations between LBW and maternal and paternal education and occupation. It highlighted that mothers receiving antenatal care were more likely to deliver normal weight babies, with those having more than four visits being six times more likely to have normal weight deliveries. Key facilitators for using antenatal services included information from health center staff, advice from family members, and media programs. Barriers included high costs, lack of transport, and low awareness of antenatal care benefits. The study recommended strategies to increase antenatal care awareness among women in remote areas and emphasized the need to address these limitations to reduce LBW deliveries in Pakistan[14]. Fatmi and Avan (2002) conducted a study to determine the factors affecting the utilization of antenatal care by women in a rural area of Sindh, Pakistan. Through a cross-sectional study involving 222 women, they found that 29.3 percent of the women utilized antenatal care during their most recent pregnancy, with 72.3 percent receiving it from government health care providers. Key factors associated with higher utilization of antenatal care included the presence of electricity in the household and the husband's occupation. The study concluded that social status and economic conditions are significant determinants of antenatal care utilization, highlighting the need for improvements in these areas to enhance antenatal and perinatal care utilization[15]. In 2017, a technical consultation hosted by USAID's global Mater-

nal and Child Survival Program generated recommendations for applying Geographic Information System (GIS) technology to improve maternal and newborn health. The meeting, attended by 72 participants from over 25 global health organizations, emphasized how mapping could contribute to the post-2015 UN's Sustainable Development Goals (SDGs) and improve maternal and neonatal health outcomes. The recommendations were categorized into ancillary geospatial and MNH data sources, technical and human resource needs, and community participation, addressing the need for improved spatial investigation at subnational levels[16]. Building on these advancements, Wulandari et al. (2023) applied hierarchical Bayesian analysis to small area models with overdispersed count data. Using Markov Chain Monte Carlo methods, they developed an area-level model to predict the under-five mortality rate at the district level in Java Island, Indonesia. Their study found that the zero-inflated negative binomial model yielded reduced relative standard error and relative mean squared error compared to district estimates and other models such as the zero-inflated generalized Poisson and Poisson models[17]. Another recent study by Neri (2023) focused on the total bias in income surveys caused by correlated nonresponse and measurement errors. This research utilized a standard sample selection model within a total survey error framework to address the correlation between nonresponse error (NR) and measurement error (ME) in estimating average household income. The study, which analyzed data from the Italian Survey on Income and Wealth linked with administrative income data from tax returns, found a positive correlation between NR and ME. Households at the extremes of the income distribution primarily drove this association. The results indicated that ME contributed more to the total error than NR and that efforts to reduce nonresponse rates would be effective mainly for nonrespondents in the lowest estimated response propensity group[18].

1.4 Research Gap

Despite significant advancements in small area estimation (SAE) techniques, including the use of both area-level models like the Fay-Herriot model and unit-level models such as the Battese model, there remain notable gaps in the application and integration of these approaches. Previous studies have demonstrated the individual strengths of these models: area-level models efficiently incorporate auxiliary data to enhance precision, while unit-level models provide detailed estimates at a finer granularity. However, a comprehensive approach that leverages the strengths of both models simultaneously has not been thoroughly explored. This gap is particularly evident in contexts requiring both high precision and detailed granularity, such as in the inventory of forest characteristics or disease prevalence mapping. My research addresses this gap by developing a hybrid model that combines the Fay-Herriot and Battese models, aiming to improve estimation accuracy and reliability by utilizing the complementary strengths of both approaches. This integration is expected to offer enhanced precision and granularity, providing a more robust tool for small area estimation in various applications.

1.5 Objective of the Study

The objective of this study is to analyze antenatal care utilization among women in Pakistan, utilizing district-wise data from the Demographic and Health Surveys (DHS) conducted in 2017-18 and 2019. This analysis aims to provide a comprehensive overview of antenatal care coverage across various districts in Pakistan, considering both unit-level and area-level information. By incorporating unit-level data, such as individual women's age, sex, education, and wealth index, alongside area-level data, such as district-level healthcare infrastructure and socio-economic indicators, the study aims to identify factors influencing antenatal care utilization at different levels. Furthermore, this study integrates both Battese-Fuller and Fay-Herriot models. The Battese-Fuller model allows for the incorporation of unit-level covariates, enabling the examination of individual-level factors influencing antenatal care utilization. On the other hand, the Fay-Herriot model facilitates the utilization of area-level information, capturing the

overall district-level characteristics that may affect antenatal care utilization patterns. By combining these models, the study aims to provide a more comprehensive understanding of the determinants of antenatal care utilization, considering both individual-level and district-level factors simultaneously. This approach enables the identification of nuanced relationships between socio-demographic characteristics, healthcare infrastructure, and antenatal care utilization, thereby informing targeted interventions and resource allocation strategies to improve maternal health outcomes in Pakistan.

Chapter 2

Methodology

2.0.1 Area Level Model

The area-level model assumes.

$$\hat{\delta}_d = x_d\alpha + \epsilon_d, \quad d = 1, \dots, D. \quad (2.1)$$

In our model, we estimate the indicator values, represented by $\hat{\delta}_d = x_d\alpha + \epsilon_d$. Here, x_d is the population value, and since it's fixed for each area, it has no variability or zero variance. The ϵ_d represents the errors or discrepancies in our estimates. We assume these errors are independent, meaning the inaccuracy in one area's estimate doesn't affect another's. They also have an average error of zero and a known variability, $var(\hat{\delta}_d)$, for each area from 1 to D . In practical terms, we determine these variances by analyzing the detailed survey data, allowing us to understand how much our estimates might deviate from the true values for each area. This information helps us gauge the reliability of our estimates in real-world situations.

We can write the regression-synthetic estimator at area level as,

$$\hat{\delta}_d^{\text{REG1-SYN}} = x'_d\hat{\alpha}, \quad (2.2)$$

$$\hat{\alpha} = \left(\sum_{d=1}^D \psi^{-1} x'_d x_d \right)^{-1} \left(\sum_{d=1}^D \psi^{-1} x_d \hat{\delta}_d \right) \quad (2.3)$$

For the parameter α in the context of the REG1-SYN estimator, the bias is expressed

as $x'_d \alpha - \delta_d$, where x_d represents the auxiliary variables for the area, and δ_d is the true value of the indicator for that area.

What's notable is that this bias remains constant and doesn't rely on the sample size, denoted as n_d , for the area. This means that increasing the sample size for a particular area doesn't reduce or eliminate this bias. The bias is inherent to the method and doesn't diminish with more extensive data collection in that specific area[19].

The Fay-Herriot Model: A Bayesian Model

Sampling Distribution: $\hat{\delta}_d^{DIR} | \delta_d \sim^{ind} N(\delta_d, \psi_d)$;

prior distribution: $\delta_d \sim^{ind} N(x'_d \beta, \sigma_\mu^2)$

where

D : number of small areas;

$\hat{\delta}_d$: direct survey estimate of δ_d ;

δ_d : true mean for area d ;

x_d : $p \times 1$ vector of known auxiliary variables;

ψ_d : known sampling variance of the direct estimate;

β and model σ_μ^2 are unknown.

A Linear Mixed Model

The Fay–Herriot model can be viewed as the following two–level model

Level 1 : $\hat{\delta}_d^{DIR} = \delta_d + e_d$

Level 2 : $\delta_d = x'_d \beta + \mu_d$

where $\{\mu_d, d = 1, \dots, D\}$ and $\{e_d, d = 1, \dots, D\}$ are independent with $\mu_d \sim N(0, \psi_d)$ and $e_d \sim N(0, \sigma_\mu^2)$

Thus, the Fay-Herriot model is essentially the following simple linear mixed model:

$$\hat{\delta}_d^{DIR} = x'_d \beta + \mu_d + e_d, \quad d = 1, \dots, D. \quad (2.4)$$

The Bayesian Estimator of δ_d

Inferences based on the posterior distribution of δ_d can be described in the following manner:

Given the model:

$$\delta_d \mid \hat{\delta}_d^{DIR}\beta, \sigma_\mu^2 \sim N(\delta_d^\beta, g_{d1}(\sigma_\mu^2));$$

where,

$$\delta_d^\beta = \gamma_d \hat{\delta}_d^{DIR} + (1 - \gamma_d)x'_d\beta$$

$$\gamma_d = \frac{\sigma_\mu^2}{\sigma_\mu^2 + \psi_d}$$

$$g_{d1}(\sigma_\mu^2) = \gamma_d\psi_d$$

The Best Predictor (BP) of δ_d

The best predictor of δ_d in this context is the posterior mean $\hat{\delta}_d$. This is because the posterior mean minimizes the mean squared error (MSE) among all possible predictors, making it the optimal choice for predicting δ_d . Thus, $\hat{\delta}_d$ serves as the best predictor, effectively balancing the direct estimate and the additional information provided by the model. Define the mean Squared Error (MSE) of a predictor $\hat{\delta}_d$ of δ_d as:

$$MSE(\hat{\delta}_d) = E(\hat{\delta}_d - \delta_d)^2, \quad (2.5)$$

where the expectation is taken over the linear mixed model. Minimizing the MSE among all predictors of δ_d , we obtain the best predictor of δ_d . as, $MSE(\hat{\delta}_d) = g_{d1}(\sigma_\mu^2)$ where, $g_{d1}(\sigma_\mu^2) = \gamma_d\psi_d$.

Best Linear Unbiased Predictor (BLUP)

When β is unknown but σ_u^2 is known, β can be estimated using the weighted least squares method. This estimator, denoted as $\hat{\beta}(\sigma_u^2)$, incorporates σ_u^2 in its calculation. By substituting $\hat{\beta}(\sigma_u^2)$ for β , we derive an empirical Bayes estimator or empirical best predictor (referred to as EB) of δ_d . This EB estimator is notable for its property of achieving the minimum mean squared error (MSE) among all linear unbiased predictors

of δ_d . Consequently, it is termed the Best Linear Unbiased Predictor (BLUP) of δ . The BLUP of δ_d is therefore defined as follows. The BLUP estimator of δ_d is given by:

$$\hat{\delta}_{\text{BLUP}}(\sigma_u^2) = \gamma_d \hat{\delta}_{\text{DIR}d} + (1 - \gamma_d) x_{0d} \hat{\beta}(\sigma_u^2), \quad (2.6)$$

where $\gamma_d = \frac{\sigma_\mu^2}{\sigma_\mu^2 + \psi_d}$. The mean squared error (MSE) of the BLUP is given by:

$$\text{MSE}(\hat{\delta}_{\text{BLUP}}) = g_{d1}(\sigma_u^2) + g_{d2}(\sigma_u^2), \quad (2.7)$$

where

$$g_{d1}(\sigma_u^2) = \gamma_d \psi_d, \quad (2.8)$$

$$g_{d2}(\sigma_u^2) = (1 - \gamma_d)^2 \mathbf{x}'_d \left(\sum_{d=1}^D (\sigma_u^2 + \psi_d) \mathbf{x}_d \mathbf{x}'_d \right)^{-1} \mathbf{x}_d. \quad (2.9)$$

It's important to note that $g_{d2}(\sigma_u^2)$ typically becomes negligible as D grows large, while $g_{d1}(\sigma_u^2)$ remains dominant even in such cases. An interesting observation emerges when considering a flat prior distribution for β . In this scenario, both the posterior mean and posterior variance of δ_d in a Bayesian framework, serving as the point estimate and measure of uncertainty respectively, coincide precisely with $\hat{\delta}_{\text{BLUP}}$ and $\text{MSE}(\hat{\delta}_{\text{BLUP}})$.

Empirical Bayes or Empirical Best Linear Unbiased Predictor

The Best Linear Unbiased Predictor (BLUP) of δ_d relies on the actual variance σ_u^2 of the random effects u_d , which is often unknown in practical scenarios. Let $\hat{\sigma}_u^2$ represent a reliable estimator for σ_u^2 . With this, we derive the Empirical BLUP (EBLUP) of δ_d as:

$$\hat{\delta}_d^{\text{EB}} = \hat{\gamma}_d \hat{\delta}_d^{\text{DIR}} + (1 - \hat{\gamma}_d) x'_{d} \hat{\beta}, \quad (2.10)$$

The estimator $\hat{\gamma}_d$ is computed as $\hat{\gamma}_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \psi_d}$, where $\hat{\gamma}_d$ is calculated as the ratio of $\hat{\sigma}_u^2$ to the sum of $\hat{\sigma}_u^2$ and a known constant ψ_d [20]. The estimator $\hat{\beta}$, denoted as $\hat{\beta}(\hat{\sigma}_u^2)$, is defined as:

$$\hat{\beta} \equiv \hat{\beta}(\hat{\sigma}_u^2) = \left(\sum_{d=1}^D \frac{1}{\hat{\sigma}_u^2 + \psi_d} \mathbf{x}_d \mathbf{x}'_d \right)^{-1} \left(\sum_{d=1}^D \frac{1}{\hat{\sigma}_u^2 + \psi_d} \mathbf{x}_d \hat{\delta}_d^{\text{DIR}} \right) \quad (2.11)$$

MSE of EBLUP

A second-order unbiased estimator of the Mean Squared Error (MSE) of the Empirical Best Linear Unbiased Predictor (EBLUP) is then expressed as:

$$\text{MSE}(\hat{\sigma}_d^{\text{FH}}) = g_{d1}(\hat{\sigma}_u^2) + g_{d2}(\hat{\sigma}_u^2) + 2g_{d3}(\hat{\sigma}_u^2), \quad (2.12)$$

where $g_{d1}(\hat{\sigma}_u^2)$, $g_{d2}(\hat{\sigma}_u^2)$, and $g_{d3}(\hat{\sigma}_u^2)$ represent different terms depending on the estimated variance $\hat{\sigma}_u^2$.

$$g_{d1}(\sigma_u^2) = \gamma_d \psi_d \quad (2.13)$$

$$g_{d2}(\sigma_u^2) = (1 - \gamma_d)^2 \mathbf{x}'_d \left(\sum_{d=1}^D (\sigma_u^2 + \psi_d) \mathbf{x}_d \mathbf{x}'_d \right)^{-1} \mathbf{x}_d. \quad (2.14)$$

$$g_{d3}(\sigma_u^2) = (1 - \gamma_d)^2 (\sigma_u^2 + \psi_d)^{-1} \text{var}(\hat{\sigma}_u^2), \quad (2.15)$$

$$\text{var}(\hat{\sigma}_u^2) = I^{-1}(\sigma_u^2) = 2 \left(\sum_{d=1}^D (\sigma_u^2 + \psi_d)^{-2} \right)^{-1}, \quad (2.16)$$

where I denotes the Fisher information matrix for a Restricted Maximum Likelihood (REML) estimator[4]. Smoothing techniques can be applied to obtain stable sampling variances ψ_d .

Certain extensions of the Fay-Herriot model can partially address the underreporting of MSE estimates.

Bayesian methods can be developed to approximate good classical solutions when σ_u^2 is unknown.

The adjusted likelihood method provides a classical approach that has been successfully applied in real-world applications.

For large D , the terms $g_{2d}(\sigma_u^2)$ and $g_{3d}(\sigma_u^2)$ are negligible, simplifying the estimation process.

MSE estimates in the Fay-Herriot model are often underreported due to the variability of sampling variance estimates.

There is no guarantee that EBLUP and associated MSE estimates will match the corresponding posterior mean and variance in Bayesian settings when σ_u^2 is unknown.

Zero ML, REML, and ANOVA estimates can lead to substantial overshrinking problems in estimating small area means.

Zero estimates can result in very small MSE estimates due to the g_{1d} term being zero[21].

2.0.2 Unit Level Model

BLUP based on the BHF model

The nested-error linear regression model, introduced by Battese, Harter, and Fuller (1977), aims to estimate corn and soy production in U.S. counties. In this model, the values of a target variable Y_{di} for individual i within area d are related to p auxiliary variables. Unlike the Fay–Herriot model, which relates direct estimators to auxiliary variables, the nested-error model focuses on the relationship between individual-level observations and auxiliary variables to predict the target variable. The model is defined by the following equation:

$$Y_{di} = \mathbf{x}'_{di} + u_d + e_{di}, \quad i = 1, \dots, N_d, \quad d = 1, \dots, D. \quad (2.17)$$

The random effects are regarded as being independent of the errors, meaning that they are assumed to vary independently from the individual-level errors in the model.

$$u_d \sim \mathcal{N}(0, \sigma_u^2)$$

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

The $k_{di} \geq 0$ are predetermined constants that denote the potential presence of heteroscedasticity in the model. The average value of an area can be broken down into the sum of observations from within the sample and observations from outside the sample.

$$\bar{Y} = N_d^{-1} \left(\sum_{i \in s_d} Y_{di} + \sum_{i \in r_d} Y_{di} \right) \quad (2.18)$$

The process of obtaining the BLUP estimator within the nested-error linear regression model involves fitting the model to the available sample data and then using this

model to predict the values for observations that were not part of the original sample (out-of-sample values).

$$\tilde{Y}_d^{\text{BLUP}} = N_d^{-1} \left(\sum_{i \in s_d} Y_{di} + \sum_{i \in r_d} \tilde{Y}_{di}^{\text{BLUP}} \right) \quad (2.19)$$

The BLUP of Y_{di} under this model is given by,

$$\tilde{Y}_{di}^{\text{BLUP}} = \mathbf{x}'_{di} \tilde{\beta} + \tilde{u}_d \quad (2.20)$$

where,

$$\tilde{u}_d = \gamma_d (\bar{y}_{da} - \bar{\mathbf{x}}'_{da} \tilde{\beta})$$

Further

$$\begin{aligned} \gamma_d &= \frac{\sigma_u^2}{\sigma_u^2 + \frac{\sigma_{\varepsilon}^2}{a_d}} \\ \bar{y}_{da} &= a_d^{-1} \sum_{i \in s_d} a_{di} Y_{di} \\ \bar{\mathbf{x}}_{da} &= a_d^{-1} \sum_{i \in s_d} a_{di} \mathbf{x}_{di} \end{aligned}$$

are the weighted sampling means of the target variable and auxiliary variables, respectively with weights $a_{di} = k_{di}^{-2}$, where $a_d = \sum_{i \in s_d} a_{di}$. For areas with $\frac{n_d}{N_d} \approx 0$, the Best Linear Unbiased Predictor (BLUP) for \tilde{Y}_d is approximately:

$$\tilde{Y}_d^{\text{BLUP}} \approx N_d^{-1} \sum_{U_d} \tilde{Y}_{di}^{\text{BLUP}}. \quad (2.21)$$

For areas with $\frac{n_d}{N_d} \approx 0$, the Best Linear Unbiased Predictor (BLUP) for \tilde{Y}_d is approximately:

$$\tilde{Y}_d^{\text{BLUP}} \approx N_d^{-1} \sum_{U_d} \left(\mathbf{x}'_{di} \hat{\beta} + \gamma_d (\bar{y}_{da} - \bar{\mathbf{x}}'_{da} \hat{\beta}) \right) \quad (2.22)$$

This can be further simplified as:

$$\tilde{Y}_d^{\text{BLUP}} \approx \gamma_d \left(\bar{y}_{da} + (\bar{\mathbf{X}}_d - \bar{\mathbf{x}}_{da})' \tilde{\beta} \right) + (1 - \gamma_d) \bar{\mathbf{X}}'_d \tilde{\beta} \quad (2.23)$$

The expression represents a weighted average between the "survey regression" estimator on the left and the regression-synthetic estimator on the right. The "survey regression"

estimator is derived by fitting a similar model where the area effects u_d are treated as fixed (using dummy variables) rather than random. Consider a homoscedastic model where $k_{di} = 1$. In this scenario, the weight factor γ_d is given by:

$$\gamma_d = \frac{\sigma_u^2}{\sigma_u^2 + \frac{\sigma_e^2}{n_d}}. \quad (2.24)$$

If the sample size n_d is small, γ_d is close to zero, and the BLUP for that area approaches the regression-synthetic estimator. Conversely, if σ_u^2 is large compared to $\frac{\sigma_e^2}{n_d}$, the BLUP approximates the "survey regression" estimator. Using matrix notation, let $\mathbf{y}_d = (Y_{d1}, \dots, Y_{dN_d})'$ represent the population vector containing the target variable in area d , and let $\mathbf{X}_d = (x_{d1}, \dots, x_{dN_d})'$ represent the corresponding matrix of auxiliary variables. Under the nested-error linear regression model, we have \mathbf{y}_d independently distributed as $\mathbf{y}_d \stackrel{\text{ind}}{\sim} (\mathbf{X}_d\beta, \mathbf{V}_d)$ for $d = 1, \dots, D$, where

$$\mathbf{V}_d = \sigma_u^2 \mathbf{1}_{N_d} \mathbf{1}'_{N_d} + \sigma_e^2 \mathbf{A}_d, \quad (2.25)$$

with $\mathbf{A}_d = \text{diag}(k_{di}^2; i = 1, \dots, N_d)$. We have

$$\mathbf{V}_d = \sigma_u^2 \mathbf{1}_{N_d} \mathbf{1}'_{N_d} + \sigma_e^2 \mathbf{A}_d = \begin{pmatrix} \mathbf{V}_{ds} & \mathbf{V}_{dsr} \\ \mathbf{V}_{drs} & \mathbf{V}_{dr} \end{pmatrix}, \quad (2.26)$$

where s represents the in-sample observations and r represents the out-of-sample observations. Then, the weighted least squares estimator of β is given by

$$\tilde{\beta} = \left(\sum_{d=1}^D \mathbf{X}_{ds} \mathbf{V}_{ds}^{-1} \mathbf{X}'_{ds} \right)^{-1} \left(\sum_{d=1}^D \mathbf{X}_{ds} \mathbf{V}_{ds}^{-1} \mathbf{y}_{ds} \right), \quad (2.27)$$

Both $\tilde{\beta}$ and \tilde{u}_d depend on the true variance components $\theta = (\sigma_u^2, \sigma_e^2)'$. Replacing those with their estimates $\hat{\theta} = (\hat{\sigma}_u^2, \hat{\sigma}_e^2)'$, we obtain:

$$\hat{\beta} = \left(\sum_{d=1}^D \mathbf{X}_{ds} \hat{\mathbf{V}}_{ds}^{-1} \mathbf{X}'_{ds} \right)^{-1} \left(\sum_{d=1}^D \mathbf{X}_{ds} \hat{\mathbf{V}}_{ds}^{-1} \mathbf{y}_{ds} \right), \quad (2.28)$$

where $\hat{\mathbf{V}}_{ds} = \hat{\mathbf{V}}_{ds}(\hat{\sigma}_u^2, \hat{\sigma}_e^2)$. The estimator for u_d is given by:

$$\hat{u}_d = \hat{\gamma}_d (\bar{y}_{da} - \bar{\mathbf{x}}'_{da} \hat{\beta}), \quad (2.29)$$

with

$$\hat{\gamma}_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2 / a_d}. \quad (2.30)$$

EBLUP based on the BHF model

With these elements, an Empirical Best Linear Unbiased Predictor (EBLUP) of Y_d is given by:

$$\hat{Y}_d^{\text{EBLUP}} = N_d^{-1} \left(\sum_{i \in s_d} Y_{di} + \sum_{i \in r_d} \hat{Y}_{di}^{\text{EBLUP}} \right), \quad (2.31)$$

where

$$\hat{Y}_{di}^{\text{EBLUP}} = \mathbf{x}'_{di} \hat{\beta} + \hat{u}_d. \quad (2.32)$$

The EBLUP, similar to the BLUP, remains unbiased within the model's framework. However, neither BLUP nor EBLUP are unbiased with respect to the design. Despite this, both BLUP and EBLUP estimators generally offer greater efficiency compared to direct estimators because they leverage information from multiple areas. Additionally, they often outperform Fay-Herriot (FH) estimators as they can utilize more detailed data. For a non-sampled area, it is not possible to calculate \hat{u}_d , so we set $\gamma_d = 0$ to obtain the regression-synthetic estimator $\bar{X}'_d \hat{\beta}$. Under simple random sampling (SRS), where $k_{di} = 1$ for all i, d and $n_d/N_d \approx 0$, the absolute relative bias (ARB) of the BLUP under the design is smaller than that of the regression-synthetic estimator for the same vector of regression coefficients.

$$(1 - \gamma_d) \frac{\bar{Y}_d - \bar{\mathbf{X}}'_d \beta}{\bar{Y}_d} \leq \frac{\bar{Y}_d - \bar{\mathbf{X}}'_d \beta}{\bar{Y}_d},$$

provided $\gamma_d > 0$. To estimate the MSE of \hat{Y}_d^{EBLUP} , we can use a parametric bootstrap procedure: 1. Fit a suitable nested-error model $Y_{di} = \mathbf{x}'_{di} \beta + u_d + e_{di}$ to the sample data and obtain the estimates $\hat{\beta}$, $\hat{\sigma}_u^2$, and $\hat{\sigma}_e^2$. 2. Generate the area effects in the form $u_d^{*(b)} \stackrel{\text{iid}}{\sim} N(0, \hat{\sigma}_u^2)$ for $d = 1, \dots, D$.

3. Generate, independent of the area's effects for the sample area units, $e_{di}^{*(b)} \stackrel{\text{iid}}{\sim} N(0, \hat{\sigma}_e^2)$ for $i \in s_d$. Also generate the population means of the errors in the areas, $\bar{E}_d^{*(b)} \stackrel{\text{iid}}{\sim} N(0, \hat{\sigma}_e^2/N_d)$ for $d = 1, \dots, D$.

4. Calculate the true bootstrap means of the areas,

$$\bar{Y}_d^{*(b)} = \bar{\mathbf{X}}'_d \hat{\beta} + u_d^{*(b)} + \bar{E}_d^{*(b)}, \quad d = 1, \dots, D. \quad (2.33)$$

Note that the computation of the mean does not require the unit values x_{di} for out-of-sample observations.

5. Using the vectors with the auxiliary variables for the sample units, generate the target variables

$$Y_{di}^{*(b)} = x'_{di}\hat{\beta} + u_d^{*(b)} + e_{di}^{*(b)}, \quad i \in s_d, \quad d = 1, \dots, D. \quad (2.34)$$

6. For the original sample $s = s_1 \cup \dots \cup s_D$, let

$$y_s^{*(b)} = \left((y_{1s}^{*(b)})', \dots, (y_{Ds}^{*(b)})' \right)' \quad (2.35)$$

be the bootstrap vector of sample values. Fit the nested-error model to the bootstrap data $y_s^{*(b)}$ and obtain the parametric bootstrap $\hat{Y}_d^{\text{EBLUP}*(b)}$, for $d = 1, \dots, D$. 7. Repeat steps 2-6 for $b = 1, \dots, B$. The MSE “naive bootstrap” estimators of the EBLUPs \hat{Y}_d^{EBLUP} is given by:

$$\text{MSE}_B(\hat{Y}_d^{\text{EBLUP}}) = \frac{1}{B} \sum_{b=1}^B \left(\hat{Y}_d^{\text{EBLUP}*(b)} - \bar{Y}_d^{*(b)} \right)^2, \quad d = 1, \dots, D. \quad (2.36)$$

This bootstrap estimator is not second-order but first-order unbiased, i.e., its bias does not decrease faster than D^{-1} when the number of areas D increases[22].

The total sample size corresponds to the number of observations used to fit the model. Therefore, utilizing unit-level data allows for more efficient estimation of the model parameters compared to area-level models. Additionally, significant efficiency gains over direct estimators are expected.

The BHF model accounts for between-area heterogeneity beyond what can be explained by fixed effects, potentially resulting in a smaller bias.

The resulting EBLUP is a composite estimator that can draw strength from multiple areas as needed. It converges to the “survey regression” estimator as the area size increases.

Unlike the FH model, this method does not require knowledge of sampling variances. The proposed model-based MSE estimator is a stable estimator of the design MSE and is design-unbiased when averaged over a large number of areas.

Estimates can be disaggregated for any desired subdomain or subarea within the larger areas, including at the unit level. It is possible to generate estimates for out-of-sample areas.

Estimates are based on a model, necessitating model checking through residual analysis.

The BHF model does not incorporate the sampling design, resulting in estimates that are not design-unbiased and are more suitable for simple random sampling (SRS).

Estimates may be sensitive to outliers or violations of normality assumptions. Additionally, the requirement for microdata to obtain estimates may lead to data access issues due to confidentiality concerns from data owners.

The widely used analytic MSE estimator by Prasad and Rao is second-order unbiased under the assumption of normality in the model, but it is not design-unbiased for the design MSE of any given area.

[23].

2.1 Area-cum-Unit level model for SAE

Let U_d be a population where d represents area of population where, $d = 1, \dots, D$. we examine two distinct samples: Sample 1 $S^{(1)}$ and Sample 2 $S^{(2)}$. In Sample 1, we gather auxiliary information at the area level, denoted as X_d for $d = 1, \dots, D$. Conversely, Sample 2 includes auxiliary information at the unit level, represented as $X_{d,i}$ where i range from 1 to n_d . To enhance our analysis, we merge these samples into a combined sample, S_d^+ , by uniting $S_d^{(1)}$ and $S_d^{(2)}$. The out-of-sample data is then defined as $S_d = U_d - S_d^+$, where U_d encompasses the entire dataset, and $Y_{d,i}$, serves as the target variable for unit level information from area d of each individuals i by using the corresponding information. Our methodology employs two statistical approaches: Fay-Herriot model for area-level data and the Battese-Fuller model for unit-level data. The objective is to integrate these models, crafting a unified method capable of analyzing both sample types, $S_d^{(1)}$ and $S_d^{(2)}$, effectively. This integrated approach aims to leverage the strengths of both models improving efficiency of area-level and unit-level information.

The objective is to proposed the Fay-Herriot (FH) and Battese-Harter-Fuller (BHF) models to obtain more robust small area estimates, where δ_d^{FH} is obtain from the survey conducted on 1st occasion while δ_d^{BHF} is obtain from the survey conducted on 2nd occasions. The combined estimate δ_d^{ACU} is calculated as a weighted average of the FH and BHF estimates.

The proposed estimate for a small area d is given by:

$$\delta_d^{\text{ACU}} = \lambda_d \delta_d^{\text{FH}} + (1 - \lambda_d) \delta_d^{\text{BHF}} \quad (2.37)$$

where λ_d is a weighting parameter that balances the contribution of the FH and BHF estimates. The detailed formulation for the combined estimate is:

$$\delta_d^{\text{ACU}} = \lambda_d (\gamma_d \hat{\delta}_d^{\text{DIR}} + (1 - \gamma_d) x_d' \beta) + (1 - \lambda_d) \left(\gamma_d (\bar{y}_{da} + (\bar{X}_d - \bar{x}_{da}) \tilde{B}) + (1 - \gamma_d) \bar{x}' \tilde{B} \right) \quad (2.38)$$

where,

- D : Number of small areas.
- $\hat{\delta}_d$: Direct survey estimate of δ_d .
- δ_d : True mean for area d .
- x_d : $p \times 1$ vector of known auxiliary variables.
- ψ_d : Known sampling variance of the direct estimate.
- β and σ_μ^2 : Unknown parameters to be estimated.
- γ_d : Shrinkage factor.

Bias

It is assumed that the auxiliary information on unit level is available from current sample only, while the auxiliary information at population level is available on first occasion where the first survey was conducted. To analyze the bias of the proposed

estimator, we start with the squared expected value of the difference between the cumulative estimator δ_d^{ACU} and the true parameter δ_d . This helps us understand how close our estimator is to the actual value, the bias is defined as:

$$\text{Bia}(\hat{\delta}_d^{\text{ACU}}) = E[\delta_d^{\text{ACU}} - \delta_d] \quad (2.39)$$

Using the definition of δ_d^{cum} from Equation 2.37, we can express this as:

$$\text{Bia}(\hat{\delta}_d^{\text{ACU}}) = E[\lambda_d \delta_d^{\text{FH}} + (1 - \lambda_d) \delta_d^{\text{BHF}} - \delta_d] \quad (2.40)$$

This expression shows the combined effect of the FH and BHF estimators on the bias. Expanding and simplifying by taking common terms, we get:

$$\text{Bia}(\hat{\delta}_d^{\text{ACU}}) = E[\lambda_d \delta_d^{\text{FH}} + (1 - \lambda_d) \delta_d^{\text{BHF}} - \lambda_d \delta_d - (1 - \lambda_d) \delta_d] \quad (2.41)$$

This reduces to a form where we can analyze the contributions of each component separately:

$$\text{Bia}(\hat{\delta}_d^{\text{ACU}}) = E[\lambda_d (\delta_d^{\text{FH}} - \delta_d) + (1 - \lambda_d) (\delta_d^{\text{BHF}} - \delta_d)] \quad (2.42)$$

We have,

$$\text{Bia}(\hat{\delta}_d^{\text{ACU}}) = \lambda_d E(\delta_d^{\text{FH}} - \delta_d) + (1 - \lambda_d) E(\delta_d^{\text{BHF}} - \delta_d) \quad (2.43)$$

Given that the correlation term $\text{Cov}(\delta_d^{\text{FH}} - \delta_d, \delta_d^{\text{BHF}} - \delta_d)$ reduces to zero due to independent samples, as we are selecting only un-matched part from the sample obtained on occasion 1.

$$\lambda_d E(\delta_d^{\text{FH}} - \delta_d) + (1 - \lambda_d) E(\delta_d^{\text{BHF}} - \delta_d) \quad (2.44)$$

After some simplifications we get:

$$\text{Bia}(\hat{\delta}_d^{\text{ACU}}) = \lambda B(\delta_d^{\text{FH}}) + (1 - \lambda) B(\delta_d^{\text{BHF}}), \quad (2.45)$$

the bias expression in eq reduced to $\lambda B(\delta_d^{\text{FH}})$ when $B(\delta_d^{\text{BHF}})$ reduces to 0 as the unit level model provides unbiased estimates. while it reduce to $(1 - \lambda) B(\delta_d^{\text{BHF}})$ when the area level model gives unbiased result.

MSE

The Mean Squared Error (MSE) of the proposed estimates is a crucial metric to evaluate their accuracy. It represents the average of the squares of the errors, providing a measure of how much the estimated values deviate from the actual values. The MSE can be expressed as:

$$MSE(\hat{\delta}_d^{\text{ACU}}) = E[\hat{\delta}_d^{\text{ACU}} - \delta_d]^2 \quad (2.46)$$

Using Equation (2.37), this becomes:

$$MSE(\hat{\delta}_d^{\text{ACU}}) = E[\lambda_d \hat{\delta}_d^{\text{FH}} + (1 - \lambda) \hat{\delta}_d^{\text{BHF}} - \delta_d]^2 \quad (2.47)$$

By taking some terms common and rearranging, we have:

$$MSE(\delta_d^{\text{ACU}}) = E[\lambda_d \hat{\delta}_d^{\text{FH}} + (1 - \lambda_d) \hat{\delta}_d^{\text{BHF}} - \lambda_d \delta_d - (1 - \lambda_d) \delta_d]^2 \quad (2.48)$$

This further simplifies to:

$$MSE(\delta_d^{\text{ACU}}) = E[\lambda_d (\hat{\delta}_d^{\text{FH}} - \delta_d) + (1 - \lambda_d) (\hat{\delta}_d^{\text{BHF}} - \delta_d)]^2 \quad (2.49)$$

The constant term of expected value is the square of that term:

$$MSE(\delta_d^{\text{ACU}}) = \lambda_d^2 E(\delta_d^{\text{FH}} - \delta_d) + (1 - \lambda_d)^2 E(\delta_d^{\text{BHF}} - \delta_d) + E(\delta_d^{\text{FH}} - \delta_d) E(\delta_d^{\text{BHF}} - \delta_d) \quad (2.50)$$

As the correlation term $Cov(\delta_d^{\text{FH}} - \delta_d, \delta_d^{\text{BHF}} - \delta_d)$ reduces to zero due to independent samples as this is possible only we take the unmatched part of the sample taken from first occasions.

$$MSE(\delta_d^{\text{ACU}}) = \lambda_d^2 E(\delta_d^{\text{FH}} - \delta_d) + (1 - \lambda_d)^2 E(\delta_d^{\text{BHF}} - \delta_d) \quad (2.51)$$

$$(2.52)$$

Expanding the MSE terms, we get:

$$\lambda_d^2 MSE(\hat{\delta}_d^{\text{FH}}) + (1 - \lambda_d)^2 MSE(\hat{\delta}_d^{\text{BHF}}) \quad (2.53)$$

Using Equations (2.12) and (2.36), we get:

$$MSE(\delta_d^{\text{ACU}}) = \lambda_d^2 (g_{d1}(\hat{\sigma}_u^2) + g_{d2}(\hat{\sigma}_u^2) + 2g_{d3}(\hat{\sigma}_u^2)) + (1 - \lambda_d)^2 \left(\frac{1}{B} \sum_{b=1}^B \left(\hat{Y}_d^{\text{EBLUP}*(b)} - \bar{Y}_d^{*(b)} \right)^2 \right) \quad (2.54)$$

A high value of λ gives higher weight to the estimator obtain through FH and lower weight to the estimator obtain through BHF. Hence an optimum value of λ which minimize the MSE expression is required. The mean square error of δ_d^{ACU} can be expressed as a function of g_{d1} , g_{d2} , g_{d3} , g_{d4} as follows:

$$\begin{aligned} g_{d1}(\sigma_u^2) &= \gamma_d \psi_d \\ g_{d2}(\sigma_u^2) &= (1 - \gamma_d)^2 \mathbf{x}'_d \left(\sum_{d=1}^D (\sigma_u^2 + \psi_d) \mathbf{x}_d \mathbf{x}'_d \right)^{-1} \mathbf{x}_d \\ g_{d3}(\sigma_u^2) &= (1 - \gamma_d)^2 (\sigma_u^2 + \psi_d)^{-1} \text{var}(\hat{\sigma}_u^2) \\ g_{d4}(B) &= \frac{1}{B} \sum_{b=1}^B \left(\hat{Y}_d^{\text{EBLUP}*(b)} - \bar{Y}_d^{*(b)} \right)^2 \end{aligned}$$

Thus, the MSE can be written as:

$$MSE(\delta_d^{\text{ACU}}) = (\lambda_d)^2 (g_{d1}(\hat{\sigma}_u^2) + g_{d2}(\hat{\sigma}_u^2) + 2g_{d3}(\hat{\sigma}_u^2)) + (1 - \lambda_d)^2 g_{d4}(B) \quad (2.55)$$

Optimum Value of λ_d

The MSE given in Equation 2.53 relies on the value of λ . To find the optimal value of λ_d , we differentiate the Mean Squared Error (MSE) with respect to λ_d and set it to zero. This helps us in minimizing the MSE for our proposed method, ensuring that the estimator is as accurate as possible. First, we take the derivative of the MSE:

$$\frac{d}{d\lambda_d} MSE(\delta_d^{\text{ACU}}) = 2\lambda_d MSE(\delta_d^{\text{FH}}) - 2(1 - \lambda_d) MSE(\delta_d^{\text{BHF}}) \quad (2.56)$$

Setting this derivative to zero and solving for λ_d , we get:

$$\lambda_d^{\text{opt}} = \frac{MSE(\delta_d^{\text{BHF}})}{MSE(\delta_d^{\text{FH}}) + MSE(\delta_d^{\text{BHF}})} \quad (2.57)$$

This λ_d represents the optimal weight that minimizes the MSE when combining the Fay-Herriot (FH) and Battese-Harter-Fuller (BHF) estimators. Using this optimal λ_d in the MSE expression, we obtain:

$$\begin{aligned} MSE(\delta_d^{\text{ACU}})_{\text{opt}} &= \left\{ \frac{MSE(\delta_d^{\text{BHF}})}{MSE(\delta_d^{\text{FH}}) + MSE(\delta_d^{\text{BHF}})} \right\}^2 (g_{d1}(\hat{\sigma}_u^2) + g_{d2}(\hat{\sigma}_u^2) + 2g_{d3}(\hat{\sigma}_u^2)) \\ &+ \left\{ 1 - \frac{MSE(\delta_d^{\text{BHF}})}{MSE(\delta_d^{\text{FH}}) + MSE(\delta_d^{\text{BHF}})} \right\}^2 g_{d4}(B) \end{aligned} \quad (2.58)$$

To simplify, we recognize that:

$$MSE(\delta_d^{\text{FH}}) = (g_{d1}(\hat{\sigma}_u^2) + g_{d2}(\hat{\sigma}_u^2) + 2g_{d3}(\hat{\sigma}_u^2))$$

and

$$MSE(\delta_d^{\text{BHF}}) = g_{d4}(B)$$

Rewriting the equation using these definitions:

$$\begin{aligned} MSE(\delta_d^{\text{ACU}})_{\text{opt}} &= \left\{ \frac{MSE(\delta_d^{\text{BHF}})}{MSE(\delta_d^{\text{FH}}) + MSE(\delta_d^{\text{BHF}})} \right\}^2 MSE(\delta_d^{\text{FH}}) \\ &\quad + \left\{ 1 - \frac{MSE(\delta_d^{\text{BHF}})}{MSE(\delta_d^{\text{FH}}) + MSE(\delta_d^{\text{BHF}})} \right\}^2 MSE(\delta_d^{\text{BHF}}) \end{aligned} \quad (2.59)$$

Further simplification gives us:

$$\begin{aligned} MSE(\delta_d^{\text{cum}})_{\text{opt}} &= \left\{ \frac{(MSE(\delta_d^{\text{BHF}}))^2}{(MSE(\delta_d^{\text{FH}}) + MSE(\delta_d^{\text{BHF}}))^2} \right\} MSE(\delta_d^{\text{FH}}) \\ &\quad + \left\{ \frac{(MSE(\delta_d^{\text{FH}}))^2}{(MSE(\delta_d^{\text{FH}}) + MSE(\delta_d^{\text{BHF}}))^2} \right\} MSE(\delta_d^{\text{BHF}}) \end{aligned} \quad (2.60)$$

Taking the least common multiple (LCM) of the terms on the right-hand side, we get:

$$MSE(\delta_d^{\text{cum}})_{\text{opt}} = \left\{ \frac{(MSE(\delta_d^{\text{BHF}}))^2 MSE(\delta_d^{\text{FH}}) + (MSE(\delta_d^{\text{FH}}))^2 MSE(\delta_d^{\text{BHF}})}{(MSE(\delta_d^{\text{FH}}) + MSE(\delta_d^{\text{BHF}}))^2} \right\} \quad (2.61)$$

By factoring out the common terms, we obtain:

$$MSE(\delta_d^{\text{cum}})_{\text{opt}} = \frac{MSE(\delta_d^{\text{BHF}})MSE(\delta_d^{\text{FH}}) \times [(MSE(\delta_d^{\text{BHF}}) + MSE(\delta_d^{\text{FH}})]}{(MSE(\delta_d^{\text{FH}}) + MSE(\delta_d^{\text{BHF}}))^2} \quad (2.62)$$

Finally, simplifying further, we get:

$$MSE(\delta_d^{\text{cum}})_{\text{opt}} = \frac{(MSE(\delta_d^{\text{BHF}}))(MSE(\delta_d^{\text{FH}}))}{(MSE(\delta_d^{\text{FH}}) + MSE(\delta_d^{\text{BHF}}))} \quad (2.63)$$

This equation shows the minimized mean squared error for the combined estimator.

Comparison of MSE

To demonstrate the efficiency of the combined method, we compare its MSE to the MSE's obtained the individual FH and BHF methods. The goal is to show that the combined method has a lower MSE. First, we express this as:

$$MSE(\delta_d^{\text{ACU}})_{\text{opt}} \leq MSE(\hat{\delta}_d^{\text{FH}}) \quad (2.64)$$

Using the previous equations, we get:

$$\frac{MSE(\delta_d^{\text{BHF}})MSE(\delta_d^{\text{FH}})}{MSE(\delta_d^{\text{FH}}) + MSE(\delta_d^{\text{BHF}})} \leq MSE(\delta_d^{\text{FH}}) \quad (2.65)$$

Simplifying, we have:

$$\frac{MSE(\delta_d^{\text{BHF}})}{MSE(\delta_d^{\text{FH}}) + MSE(\delta_d^{\text{BHF}})} \leq 1 \quad (2.66)$$

Similarly, we compare with the BHF method:

$$MSE(\delta_d^{\text{ACU}})_{\text{opt}} \leq MSE(\delta_d^{\text{BHF}}) \quad (2.67)$$

Using the previous equations, we get:

$$\frac{MSE(\delta_d^{\text{BHF}})MSE(\delta_d^{\text{FH}})}{MSE(\delta_d^{\text{FH}}) + MSE(\delta_d^{\text{BHF}})} \leq MSE(\delta_d^{\text{BHF}}) \quad (2.68)$$

Simplifying, we have:

$$\frac{MSE\delta_d^{\text{FH}}}{MSE(\delta_d^{\text{FH}}) + MSE\delta_d^{\text{BHF}}} \leq 1 \quad (2.69)$$

This conditions given in (2.66) and (2.67) are always true which demonstrates that the combined estimator provides a more efficient and robust solution by minimizing the overall MSE, leveraging the strengths of both FH and BHF methods.

Chapter 3

Results and Discussions

3.1 Description of study

Pakistan is a developing country situated in South Asia, covering a total area of approximately 881,913 square kilometers. It shares borders with India, Afghanistan, Iran, and China. Despite its strategic location and abundant natural resources, Pakistan faces significant challenges in economic growth and development, particularly in the healthcare sector. According to data from the Demographic and Health Surveys (DHS), Pakistan has made strides in improving maternal healthcare over the years, but there are still significant gaps, especially in antenatal care (ANC) for women. In 2017, only about 80percentage of pregnant women in Pakistan received the recommended four or more ANC visits, which is crucial for monitoring the health of both the mother and the unborn child (DHS, 2017). The low uptake of ANC services can be attributed to various factors, including geographical disparities, socio-economic status, cultural beliefs, and limited access to healthcare facilities, particularly in rural areas. Additionally, gender disparities and patriarchal norms often hinder women's access to healthcare services, including ANC. Pakistan faces challenges such as political instability, terrorism, and natural disasters, which have hindered efforts to improve maternal healthcare. Moreover, inadequate investment in the healthcare system, limited resources, and inefficient healthcare delivery mechanisms contribute to the challenges in providing quality ANC services to pregnant women across the country. Efforts to improve maternal healthcare in Pakistan are underway, with initiatives aimed at increasing awareness, improving ac-

cess to healthcare facilities, training healthcare providers, and addressing socio-cultural barriers. However, there is a need for targeted interventions and investment in healthcare infrastructure to ensure that all women, regardless of their socio-economic status or geographical location, have access to quality ANC services. Utilizing data from the DHS and other relevant sources, researchers and policymakers can formulate models to understand the determinants of ANC utilization among women in Pakistan. By identifying the factors influencing ANC uptake and targeting interventions accordingly, Pakistan can make significant progress in improving maternal and child health outcomes and ultimately reducing maternal and infant mortality rates.

3.2 Results

This study utilized data from the 2017-18 and 2020 Demographic and Health Surveys (DHS). Sample 1 (S1) included auxiliary information at the area level and consisted of 15,068 observations. Sample 2 (S2) included auxiliary information at the unit level and consisted of 15,143 observations. After merging both surveys and excluding overlapping parts, we were left with 6,979 unique observations across 9 variables.

Table 3.1 provides the variable information, the primary response variable, antenatal care (M57AS1) from survey 1 and (Q406) from survey 2, captures information about the antenatal care services received by individuals. The district (SDIST) variable identifies the specific district of the respondent's residence and is used to define the geographical areas for analysis. Age indicates the current age of the respondent and serves as an auxiliary variable to control for age-related differences from both surveys. The wealth index represents the economic status of the respondent's household and is used as an auxiliary variable to account for economic variations from both surveys. The region (V0124) from survey 1 and (QREGION) from survey 2 variable specifies the province or region where the respondent lives, providing a basis for regional analysis. Education denotes the highest level of education attained by the respondent, helping to understand the impact of educational attainment. The residence variable indicates whether the respondent lives in an urban or rural area, distinguishing between different

living environments. Total children shows the number of children ever born to the respondent, controlling for family size and reproductive history. Finally, the weight variable provides the sample weight for each respondent, ensuring that the analysis accurately represents the population by adjusting for different probabilities of selection.

Table 3.1: Variables Discription

DHS code	Variable name	Discription	Usage
M57AS1 ¹	Antenatal Care	Antenatal Care	Response Variable
SDIST ¹	District	District	Domain
V012 ¹	Age	Individual's Current Age	Auxiliary Variable
AWFACTW ¹	Wealth Index	Wealth Index	Auxiliary Variable
V0124 ¹	Region	Provinces	Auxiliary Variable
V106 ¹	Education	Highest Education	Auxiliary Variable
V025 ¹	Residence	Urban/Rural	Auxiliary Variable
V201 ¹	Total Children	Total Children Ever Born	Auxiliary Variable
V005 ¹	Weight	Sample Weight	Sample Weight
Q103 ²	Age	Women's Age	Auxiliary Variable
Q406 ²	ANC	Antenatal Care	Response Variable
QDIST ²	District	District	Domain
QWEIGHT ²	Weight	Sample Weight	Sample Weight
QWFWLTH ²	Wealth Index	Wealth Index	Auxiliary Variable
QREGION ²	Region	Provinces	Auxiliary Variable
Q105 ²	Education	Highest Education	Auxiliary Variable
QTYPE ²	Residence	Urban/Rural	Auxiliary Variable
Q208A ²	Total Children	Total Children Ever Born	Auxiliary Variable

In this study, we employ four methods to analyze the data. The direct method uses information directly obtained from the sample and three are model based methods. The Fay-Herriot method leverages area-level information for its analysis. The Battese-Harter-Fuller method utilizes unit-level information. Our proposed method, Area cum Unit, integrates both area-level and unit-level information from the surveys, combining the strengths of both models for a comprehensive analysis.

In 3.1 the process begins with the survey data. Following data, weights are assigned to the survey data based on the sampling design to ensure representatives. After the weights are assigned, the weighted average of the responses is calculated. This calculated weighted average represents the direct estimate. The process concludes with the acquisition of this direct estimate.

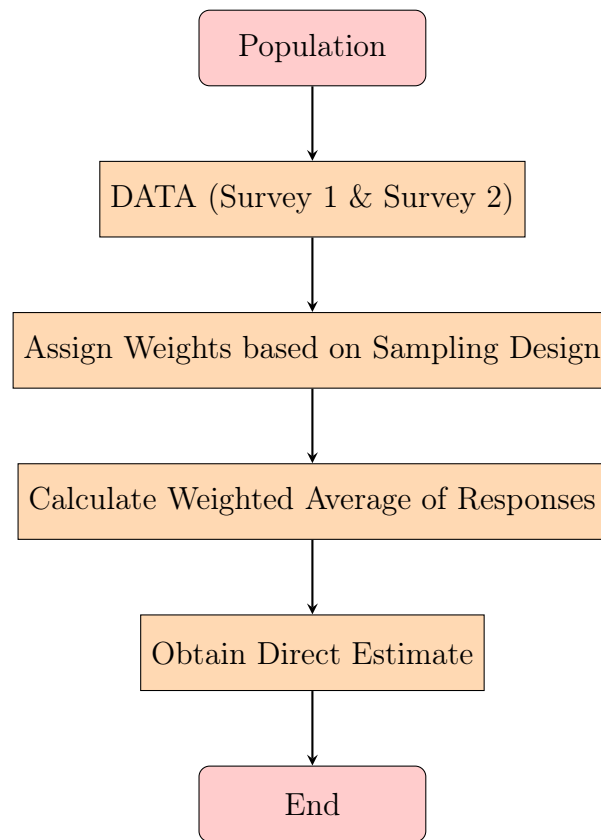


Figure 3.1: Flowchart for Obtaining Direct Estimate

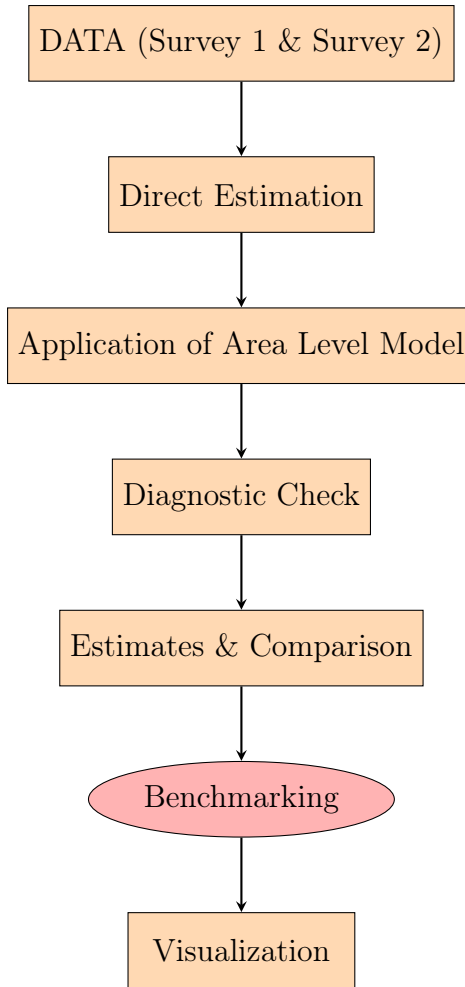


Figure 3.2: Flowchart for Area Level Model

In 3.2, the process begins by loading the necessary libraries: ‘survey’, ‘reshape2’, and ‘emdi’, ensuring that the ‘fh’ function is available. Data from Survey 1 and Survey 2 is read into R. The datasets are combined. A sampling design is defined using the ‘svydesign’ function with weights assigned based on the sampling design. Direct estimates for various variables from table 3.1 are calculated using ‘svyby’. The sample size and effective sample size are calculated. Finally, the Fay-Herriot model is applied using the ‘fh’ function, with the model’s class checked and a summary of the model generated. The next step involves conducting a diagnostic check to ensure the model’s accuracy. After the diagnostic check, estimates are generated and compared. The

benchmarking step follows to verify the accuracy of the estimates.

In 3.3 QQ plots obtained from applying the Fay-Herriot model illustrate the distri-

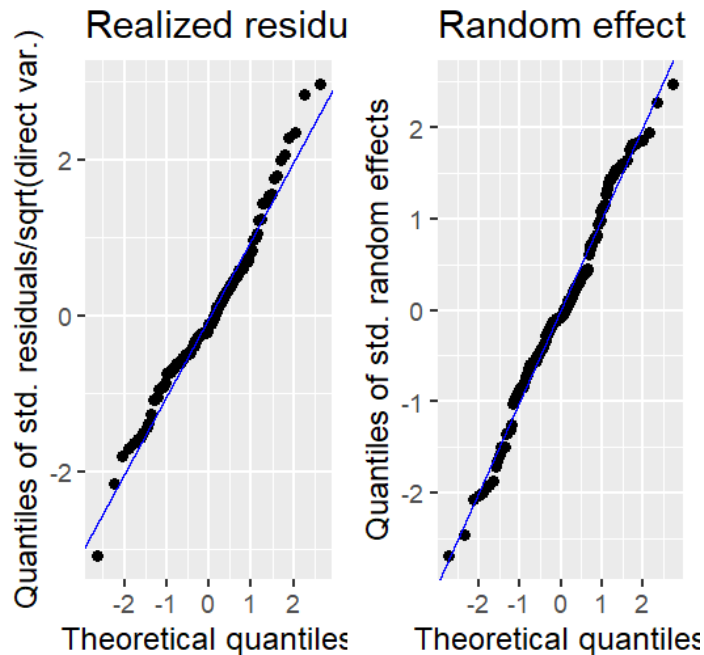


Figure 3.3: QQ Plots for Realized Residuals and Random Effects in Fay-Herriot Model

bution of the realized residuals and random effects against the theoretical quantiles of a normal distribution. The left panel shows the realized residuals, and the right panel displays the random effects. Both panels aim to assess the normality of these components. The alignment of points along the diagonal blue line indicates adherence to a normal distribution. The observed points in both plots mostly follow the diagonal line, suggesting that the residuals and random effects conform reasonably well to the normality assumption, although some deviations at the tails indicate slight departures from perfect normality.

In 3.4, the density plot shows the distribution of the standardized realized residuals divided by the square root of the direct variance from the Fay-Herriot model. The black line represents the theoretical normal density curve, while the shaded area depicts the actual density of the standardized residuals. The plot aims to assess the normality of the residuals. The observed residuals generally follow the shape of the theoretical

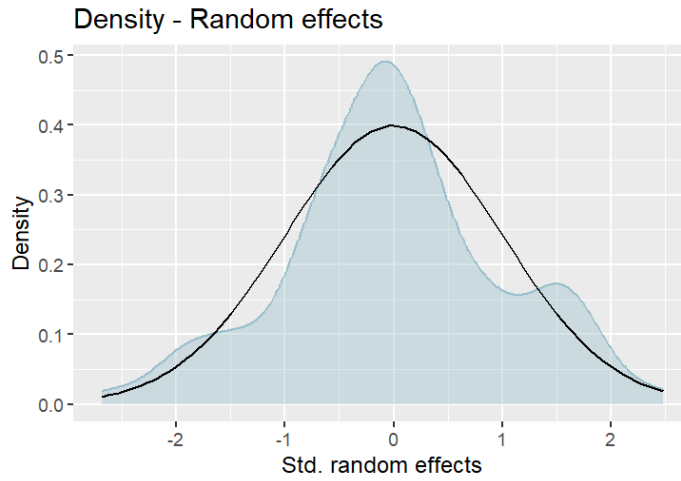


Figure 3.4: Density Plot of Standardized Realized Residuals Divided by Square Root of Direct Variance

normal distribution, indicating that the residuals approximate a normal distribution. However, there are slight deviations, especially at the tails, suggesting minor departures from normality.

In 3.5, the scatter plot compares the direct estimates to the model-based Fay-Herriot

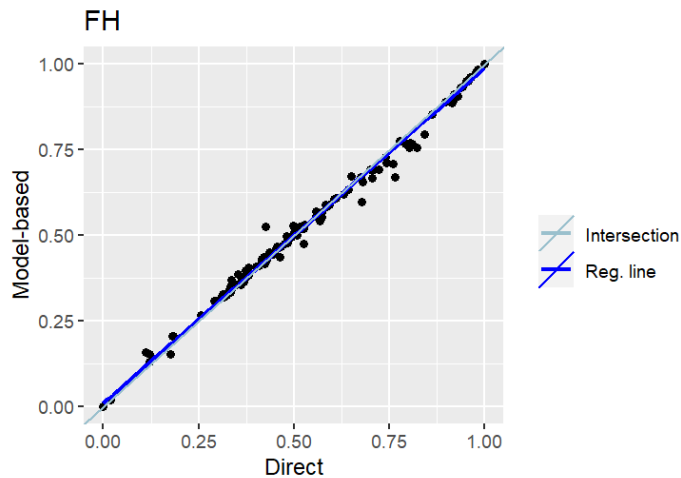


Figure 3.5: Comparison of Direct and Fay-Herriot Model-Based Estimates

(FH) estimates. Each point represents a comparison between the direct estimate (x-axis) and the FH model-based estimate (y-axis) for a given domain. The 45-degree

line (gray) indicates perfect agreement between the two estimates. The blue regression line provides a visual indication of the overall relationship between the direct and model-based estimates. The plot shows a strong linear relationship, suggesting that the FH model provides estimates that are generally consistent with the direct estimates, but with some deviations indicating areas where the model might be improving the precision of the estimates.

In 3.6, In this analysis, several packages are utilized to perform various steps of the unit-level model. The ‘survey’ package is used to define the sampling design and calculate direct estimates based on the weighted data. The ‘emdi’ package is essential for applying small area estimation models, specifically the Empirical Best Prediction (EBP) models. The ‘dplyr’ package aids in data manipulation and merging datasets. For visualization, the ‘ggplot2’ package is employed to create comprehensive plots of the results. The ‘summarytools’ package helps in generating summaries and checking data quality. Spatial data processing and mapping are accomplished using the ‘maptools’ and ‘sf’ packages, which facilitate the reading and plotting of shapefiles to visualize the geographical distribution of the estimates. These packages together enable a systematic approach to analyze and visualize the unit-level data, from data preparation to model application and result visualization.

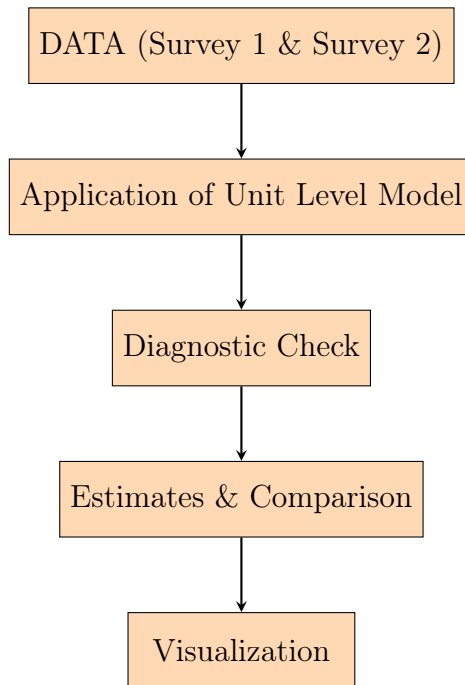


Figure 3.6: Flowchart for Unit Level Model

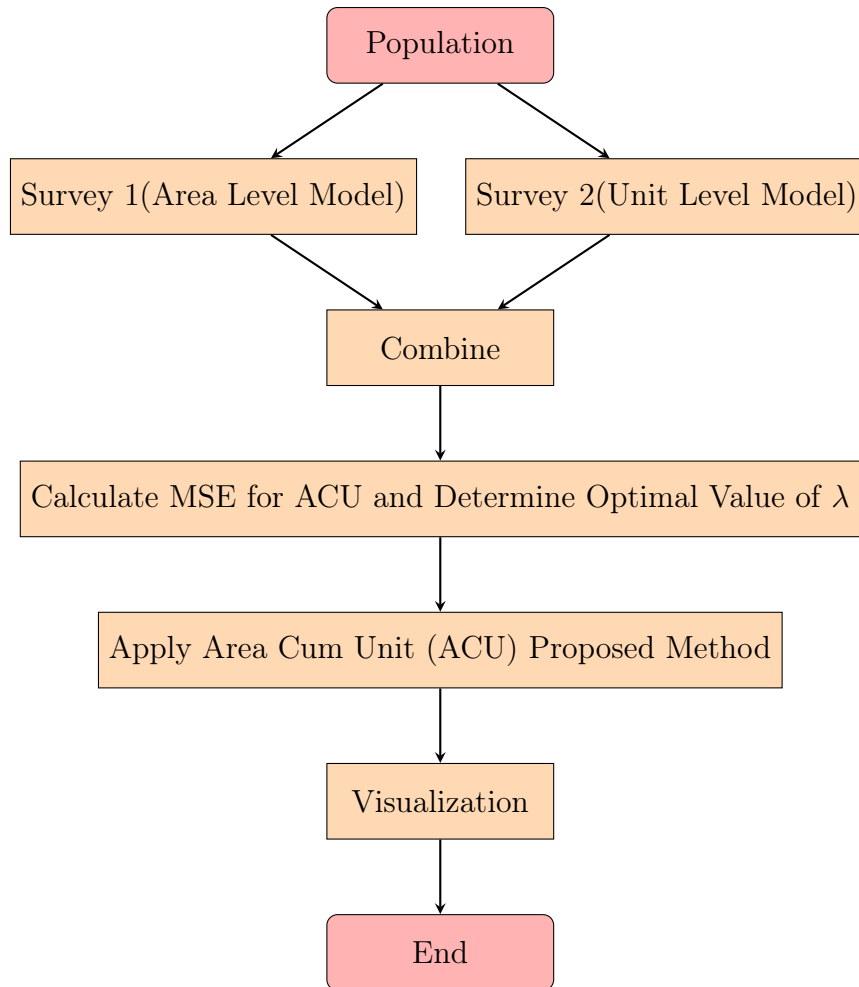


Figure 3.7: Flowchart illustrating the integration of surveys and application of the ACU method.

As illustrated in Figure 3.7, the process begins with conducting two separate surveys. Survey 1 collects auxiliary information at the area level, while Survey 2 gathers auxiliary information at the unit level. The data from these surveys are then merged, ensuring that the response variable remains consistent and that the auxiliary variables are utilized appropriately for their respective levels. Subsequently, the Mean Squared Error (MSE) for the Area Cum Unit (ACU) method is calculated, and the optimal value of the parameter λ is determined. The proposed ACU method, which combines both area-level and unit-level information, is then applied. Finally, benchmarking is performed to ensure the accuracy and reliability of the ACU method.

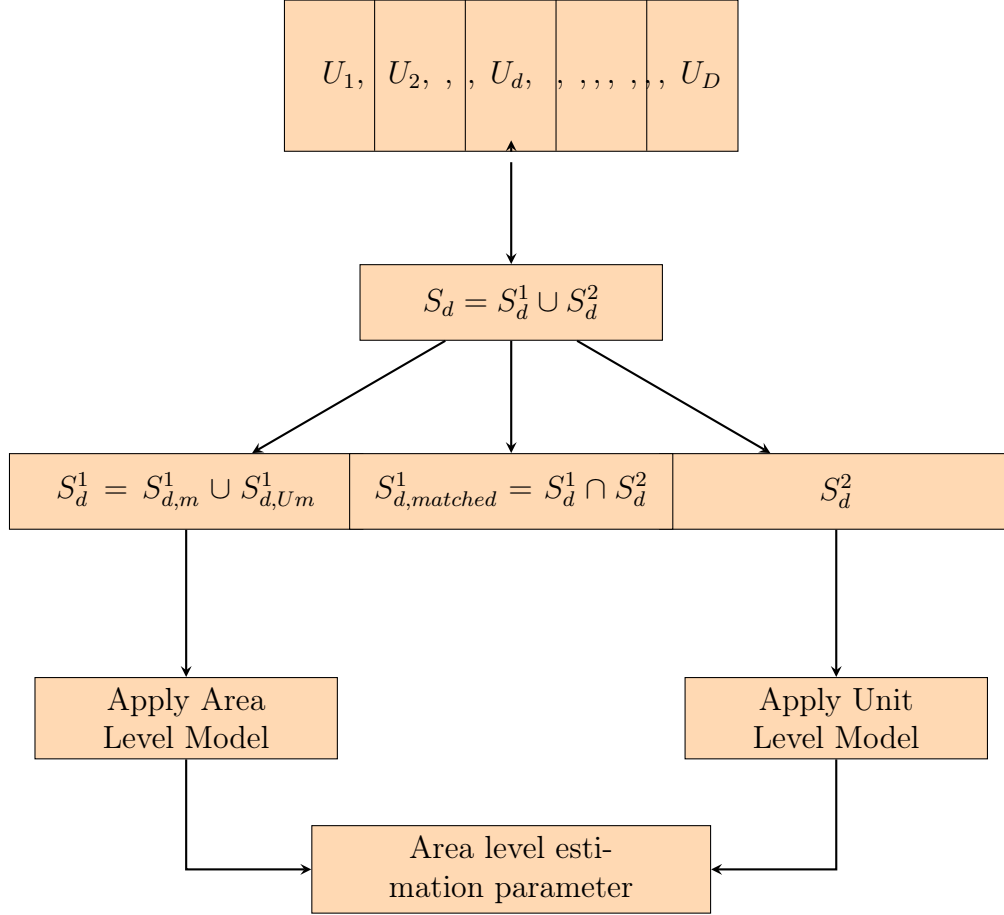


Figure 3.8: Flowchart of SAE under ACU model

The flow chart (3.8) illustrates the structure and relationships between the population and the samples used in our study. The entire square represents the population U_d , which is divided into random partitions for $d = 1, \dots, D$. Within this population, two distinct samples are drawn: Sample 1 ($S_d^{(1)}$), represented by the union of sample 1 and matched part of both samples ($S_d = S_d^1 \cup S_d^2$), which includes auxiliary information at the area level at occasion 1. Sample 2 ($S_d^{(2)}$), represented by the $S_d^2 = S_d^2 - S_{d,matched}^1$, which includes auxiliary information at the unit level at occasion 2. The intersection of these two samples is indicated as the combined sample ($S_{d,matched}^1$), which is derived by merging $S_d^{(1)}$ and $S_d^{(2)}$. This combined sample ensures that the response variable remains consistent while appropriately utilizing the auxiliary variables from each occasion. It is essential to acknowledge the independence of the two surveys. In the

case of Sample 2, the intersection is excluded, ensuring that $S_d^{(2)}$ contains only unique observations without the overlapping section. Conversely, Sample 1 incorporates the entire survey, including both its unique observations and the intersection with Sample 2. This structure highlights the comprehensive approach of integrating both area-level and unit-level information to enhance the accuracy and reliability of our analysis.

Table 3.2: Summary Proportion of estimates under different methods

	MIN	Q1	MEAN	Q3	MAX	SD
Direct	0.0181	0.4004	0.5174	0.7927	1.0000	0.2522
FH	0.0190	0.4069	0.5240	0.7555	1.0000	0.2437
BHF	0.0931	0.3845	0.4709	0.5803	1.0037	0.1908
ACU	0.1217	0.7096	0.8273	0.9214	1.0000	0.1538

The table 3.2 provides a comparative summary of the proportions for different small area estimation methods, including Direct, Fay-Herriot (FH), Battese-Harter-Fuller (BHF), and Area Cum Unit (ACU). Key statistical measures such as minimum (MIN), first quartile (Q1), mean, third quartile (Q3), maximum (MAX), and standard deviation (SD) are reported for each method. The Direct method shows a mean proportion of 0.5174 with a standard deviation of 0.2522. The range of values spans from a minimum of 0.0181 to a maximum of 1.0000. The first quartile and third quartile are 0.4004 and 0.7927, respectively, indicating a fairly wide distribution of the proportions. The FH method demonstrates a mean proportion of 0.5240 and a standard deviation of 0.2437. The proportions range from 0.0190 to 1.0000. The first and third quartiles are 0.4069 and 0.7555, respectively. This method has a slightly lower standard deviation compared to the Direct method, indicating a modest reduction in variability. The BHF method has a mean proportion of 0.4709 and a standard deviation of 0.1908. The values range from a minimum of 0.0931 to a maximum of 1.0037. The first quartile is 0.3845, and the third quartile is 0.5803. This method shows the least variability among the three traditional methods, with a tighter range of proportions. The ACU method exhibits a mean proportion of 0.8273 with a standard deviation of 0.1538. The range is from a minimum of 0.1217 to a maximum of 1.0000. The first quartile is 0.7096 and the third quartile is 0.9214. This method not only provides the highest mean proportion but also shows the lowest variability among all methods, indicating more consistent and higher estimates of the proportions. In summary, the ACU method stands out for its superior performance, providing higher mean proportions and demonstrating the least variability compared to the Direct, FH, and BHF methods. This suggests

that the ACU method may offer more reliable and consistent estimates for small area proportions.

Table 3.3: Summary MSE of estimates under different methods

	MIN	Q1	MEAN	Q3	MAX	SD
Direct	0.0002	0.0037	0.0074	0.0152	0.0799	0.0128
FH	0.0002	0.0036	0.0069	0.0131	0.0443	0.0082
BHF	0.0009	0.0090	0.0201	0.0407	0.3113	0.0463
ACU	0.0000	0.0008	0.0030	0.0059	0.0222	0.0042

The table 3.3 presents a comparative summary of the mean squared error (MSE) for different small area estimation methods, including Direct, Fay-Herriot (FH), Battese-Harter-Fuller (BHF), and Area Cum Unit (ACU). The key statistical measures reported are the minimum (MIN), first quartile (Q1), mean, third quartile (Q3), maximum (MAX), and standard deviation (SD). The Direct method shows a mean MSE of 0.0074 with a standard deviation of 0.0128. The range extends from a minimum of 0.0002 to a maximum of 0.0799, with the first and third quartiles at 0.0037 and 0.0152, respectively, indicating a wide distribution of errors. The FH method has a mean MSE of 0.0069 and a standard deviation of 0.0082, with values ranging from 0.0002 to 0.0443. The first and third quartiles are 0.0036 and 0.0131, respectively, showing a slight reduction in error variability compared to the Direct method. The BHF method, however, exhibits a higher mean MSE of 0.0201 and a significantly larger standard deviation of 0.0463. The MSE values for this method range from 0.0009 to 0.3113, with the first quartile at 0.0090 and the third quartile at 0.0407, indicating higher variability and larger errors. In contrast, the ACU method demonstrates superior performance with the lowest mean MSE of 0.0030 and a standard deviation of 0.0042. The MSE values for ACU range from a minimum of 0.0000 to a maximum of 0.0222. The first and third quartiles are 0.0008 and 0.0059, respectively, reflecting a narrower error distribution and greater consistency. Overall, the ACU method stands out as the most reliable, showing the lowest mean MSE and variability among all methods, which suggests it provides more accurate and consistent estimates in small area estimation.

Table 3.4: Summary CV of estimates under different methods

	MIN	Q1	MEAN	Q3	MAX	SD
Direct	0.0139	0.1238	0.1824	0.2576	1.0873	0.1893
FH	0.0139	0.1232	0.1737	0.2397	1.0190	0.1590
BHF	0.0702	0.1972	0.2751	0.4348	1.5944	0.3488
ACU	0.0037	0.0573	0.0782	0.1058	0.2474	0.0411

The table 3.4 presents a comparative summary of the coefficient of variation (CV) for different small area estimation methods, including Direct, Fay-Herriot (FH), Battese-Harter-Fuller (BHF), and Area Cum Unit (ACU). The key statistical measures reported are the minimum (MIN), first quartile (Q1), mean, third quartile (Q3), maximum (MAX), and standard deviation (SD). The Direct method shows a mean CV of 0.1824 with a standard deviation of 0.1893. The range extends from a minimum of 0.0139 to a maximum of 1.0873, with the first and third quartiles at 0.1238 and 0.2576, respectively, indicating a considerable spread in the CV values. The FH method has a mean CV of 0.1737 and a standard deviation of 0.1590, with values ranging from 0.0139 to 1.0190. The first and third quartiles are 0.1232 and 0.2397, respectively, showing a slight reduction in variability compared to the Direct method. The BHF method exhibits a higher mean CV of 0.2751 and a significantly larger standard deviation of 0.3488. The CV values for this method range from 0.0702 to 1.5944, with the first quartile at 0.1972 and the third quartile at 0.4348, indicating greater variability and higher coefficients of variation. In contrast, the ACU method demonstrates superior performance with the lowest mean CV of 0.0782 and a standard deviation of 0.0411. The CV values for ACU range from a minimum of 0.0037 to a maximum of 0.2474. The first and third quartiles are 0.0573 and 0.1058, respectively, reflecting a narrower distribution and greater consistency. Overall, the ACU method stands out as the most reliable, showing the lowest mean CV and variability among all methods, suggesting it provides more accurate and consistent estimates in small area estimation.

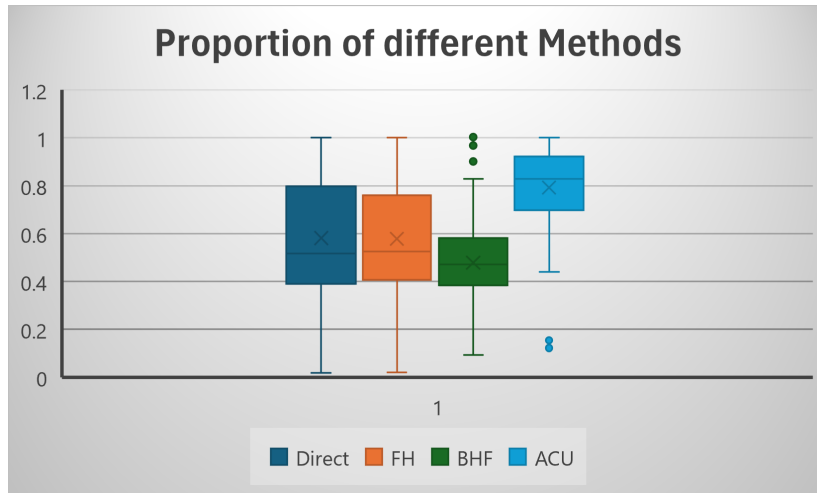


Figure 3.9: Comparison of Proportion Estimates Using Different Methods

In 3.9, the box plot compares the proportions estimated by four different methods: Direct, Fay-Herriot (FH), Battese-Harter-Fuller (BHF), and Area Cum Unit (ACU). The Direct method shows a wide interquartile range and several outliers, indicating higher variability and less stable estimates. The FH method exhibits slightly narrower variability than the Direct method but still maintains some level of variability. The BHF method demonstrates the least variability among the traditional methods, suggesting more stable and consistent estimates. However, the ACU method, which integrates both area-level and unit-level information, shows the narrowest interquartile range and fewer outliers. This indicates that the ACU method provides the most precise and reliable estimates among the four methods, with less variability and more consistent results..

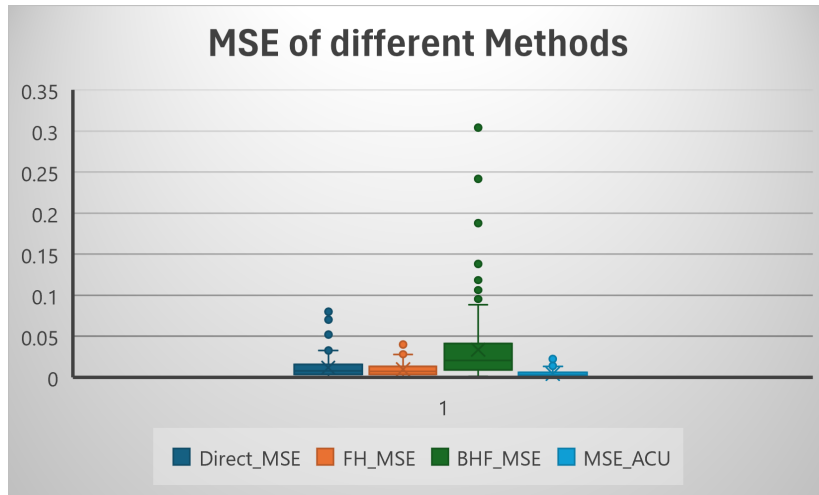


Figure 3.10: MSE of Different Methods

The box plot in Figure 3.10 illustrates the Mean Squared Error (MSE) for four different estimation methods: Direct, Fay–Herriot (FH), Battese-Harter-Fuller (BHF), and Area Cum Unit (ACU). The Direct method shows a relatively low MSE with a few outliers, indicating reasonable accuracy but with some variability. The FH method also displays low MSE values, similar to the Direct method but with slightly higher consistency. The BHF method, however, exhibits a higher MSE with significant variability, suggesting less accurate estimates and more inconsistency. In contrast, the ACU method has the lowest MSE among the methods, with minimal variability, indicating it produces the most accurate and reliable estimates. Overall, the ACU method outperforms the other methods in terms of minimizing estimation error.

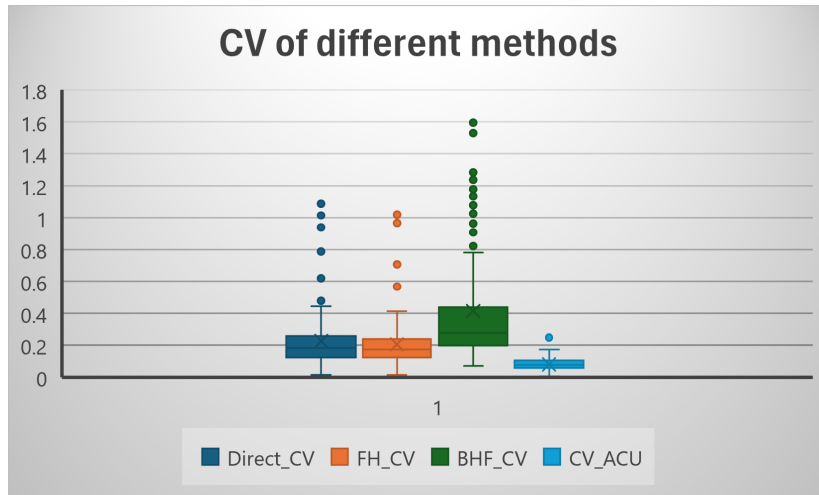


Figure 3.11: CV of Different Methods

The box plot in Figure 3.11 presents the Coefficient of Variation (CV) for four different estimation methods: Direct, Fay-Herriot (FH), Battese-Harter-Fuller (BHF), and Area Cum Unit (ACU). The Direct method shows a moderate CV with several outliers, indicating variability in the estimates. The FH method exhibits slightly lower CV values with fewer outliers, suggesting more consistent estimates compared to the Direct method. The BHF method, however, displays higher CV values and significant variability, indicating less reliable estimates. The ACU method stands out with the lowest CV values and minimal variability, reflecting highly consistent and reliable estimates. Overall, the ACU method demonstrates superior performance in terms of producing stable and precise estimates, outperforming the other methods in this comparison.

Table 3.5: District wise Estimates

Domain	Direct		FH		BHF		ACU	
	Estimates	SE	Estimates	SE	Estimates	SE	Estimates	SE
N0rthWazirstn	1.0000	0.0000	1.0000	0.0000	1.0000	0.7953	1	0.0168
Owshera	0.9568	0.0423	0.9484	0.0419	0.9484	0.9686	0.949716908	0.0320
Abbottabad	0.5607	0.0754	0.5565	0.0732	0.5565	0.4526	0.849250712	0.0530
Attock	0.6100	0.0860	0.6074	0.0828	0.6074	0.3853	0.759444345	0.0649
Awaran	0.4810	0.1975	0.4961	0.1656	0.4961	0.4110	0.764807009	0.1178
Badin	0.3349	0.1052	0.3673	0.0994	0.3673	0.4084	0.715523015	0.0739
Bagh	1.0000	0.0000	1.0000	0.0000	1.0000	0.4592	1	0.0037
Bahawalnagar	0.5840	0.0743	0.5820	0.0722	0.5820	0.5000	0.877998161	0.0517
Bahawalpur	0.3292	0.0585	0.3358	0.0575	0.3358	0.4233	0.624584846	0.0444
Bajour	0.5831	0.1809	0.5869	0.1555	0.5869	0.6654	0.830798512	0.1102
Bannu	0.5719	0.1472	0.5528	0.1325	0.5528	0.4456	0.807615871	0.0953
Barkhan	0.9147	0.0751	0.8857	0.0729	0.8857	0.6654	0.915077582	0.0555
Batagram	0.5719	0.1407	0.5651	0.1278	0.5651	0.5902	0.900089208	0.0904
Bhakkar	0.4004	0.0839	0.4079	0.0809	0.4079	0.4347	0.820671585	0.0583
Bhimber	0.9633	0.0290	0.9594	0.0289	0.9594	0.5462	0.934801512	0.0265
Bolan/Kachhi	0.7220	0.1486	0.6901	0.1334	0.6901	0.5805	0.837331882	0.0944
Buner	0.4979	0.1301	0.5280	0.1197	0.5280	0.5373	0.788499118	0.0865
Chagai	0.7927	0.1710	0.7641	0.1493	0.7641	0.6398	0.921385817	0.1050
Chakwal	0.4216	0.0816	0.4339	0.0788	0.4339	0.4108	0.693455397	0.0598
Charsadda	0.6289	0.1092	0.6179	0.1027	0.6179	0.4773	0.768505682	0.0769
Chiniot	0.4302	0.1051	0.4391	0.0993	0.4391	0.4225	0.876144759	0.0705
Chitral	0.4069	0.1036	0.4087	0.0981	0.4087	0.5339	0.64555725	0.0747
D. I. Khan	0.4982	0.0971	0.5010	0.0924	0.5010	0.3359	0.874240158	0.0663
Dadu	0.4544	0.1464	0.4598	0.1320	0.4598	0.5525	0.881912891	0.0937
Dera Bugti		0.0000		0.0000		0.3948		0.0073
Dera Ghazi Khan	0.6053	0.1000	0.6043	0.0951	0.6043	0.5414	0.753020051	0.0704

Faisalabad	0.4272	0.0444	0.4279	0.0439	0.4279	0.3971	0.481526523	0.0480
Gawadar	0.5246	0.1979	0.4752	0.1670	0.4752	0.5370	0.955466038	0.1171
Ghotki	0.4998	0.1241	0.5209	0.1150	0.5209	0.6344	0.955679578	0.0808
Gujranwala	0.3805	0.0429	0.3813	0.0424	0.3813	0.4815	0.55445296	0.0350
Gujrat	0.3146	0.0491	0.3180	0.0485	0.3180	0.4067	0.579758631	0.0381
Hafizabad	0.3127	0.0778	0.3263	0.0753	0.3263	0.2796	0.578982656	0.0591
Hangu	0.9166	0.0859	0.8902	0.0829	0.8902	0.6846	0.903091484	0.0596
Haripur	0.4392	0.0718	0.4473	0.0699	0.4473	0.4891	0.864916908	0.0500
Harnai	1.0000	0.0000	1.0000	0.0000	1.0000	0.6276	1	0.0070
Hattian Bala	0.8630	0.0716	0.8507	0.0697	0.8507	0.4812	0.943127952	0.0491
Haveli	1.0000	0.0000	1.0000	0.0000	1.0000	0.6419	1	0.0156
Hyderabad	0.7054	0.0790	0.6891	0.0765	0.6891	0.7712	0.719495073	0.0842
Islamabad	0.4655	0.0218	0.4654	0.0218	0.4654	0.4166	0.580197121	0.0193
Jacobabad	0.7059	0.1455	0.6646	0.1312	0.6646	0.6464	0.802516108	0.0964
Jaffarabad	1.0000	0.0000	1.0000	0.0000	1.0000	0.6862	1	0.0130
Jamshoro	0.3538	0.1087	0.3851	0.1023	0.3851	0.5431	0.671507645	0.0759
Jhal Magsi	0.1754	0.1701	0.1516	0.1545	0.1516	0.4285	0.859392288	0.1097
Jhang	0.4016	0.0781	0.4069	0.0756	0.4069	0.3820	0.831113055	0.0543
Jhelum	0.4339	0.0921	0.4397	0.0881	0.4397	0.5573	0.831519498	0.0635
Kalat	0.7585	0.1275	0.7075	0.1177	0.7075	0.5622	0.915875809	0.0825
Karachi Central	0.3324	0.0419	0.3328	0.0415	0.3328	0.3246	0.565377202	0.0331
Karachi East	0.3740	0.0733	0.3764	0.0714	0.3764	0.3167	0.769253244	0.0510
Karachi South	0.3235	0.0609	0.3250	0.0597	0.3250	0.3493	0.765864878	0.0434
Karachi West	0.4444	0.0465	0.4435	0.0460	0.4435	0.4012	0.610762447	0.0377
Karak	0.5066	0.0993	0.4993	0.0945	0.4993	0.4969	0.80061585	0.0694
Kashmore	0.1811	0.0883	0.2054	0.0848	0.2054	0.3221	0.559607176	0.0639
Kasur	0.5136	0.0595	0.5147	0.0583	0.5147	0.5168	0.820564172	0.0424
Kech/Turbat	0.8093	0.0990	0.7646	0.0942	0.7646	0.7240	0.871980786	0.0681
Khairpur	0.3635	0.0695	0.3697	0.0677	0.3697	0.3947	0.644536446	0.0525
Khanewal	0.5151	0.0853	0.5244	0.0821	0.5244	0.4880	0.679706893	0.0623

Kharan	0.9287	0.0557	0.9043	0.0548	0.9043	0.6082	0.91321562	0.0420
Khushab	0.2549	0.0735	0.2656	0.0715	0.2656	0.3930	0.919313246	0.0504
Khuzdar	0.1826	0.0876	0.2035	0.0842	0.2035	0.2700	0.568806401	0.0644
Khyber	0.9819	0.0150	0.9807	0.0150	0.9807	1.0037	0.973404257	0.0140
Killa Abdullah	0.1123	0.1221	0.1577	0.1140	0.1577	0.3979	0.769077487	0.0812
Killa Saifullah	0.3654	0.1629	0.3736	0.1436	0.3736	0.3574	0.502942634	0.1244
Kohat	0.4231	0.0820	0.4168	0.0792	0.4168	0.4233	0.884235097	0.0555
Kohlu	0.5679	0.1832	0.5408	0.1574	0.5408	0.6263	0.827352582	0.1129
Korangi	0.3193	0.0566	0.3202	0.0556	0.3202	0.3017	0.554863463	0.0429
Kotli	0.9827	0.0171	0.9812	0.0171	0.9812	0.5238	0.976726033	0.0186
Kurram	1.0000	0.0000	1.0000	0.0000	1.0000	0.8286	1	0.0185
Lahore	0.4756	0.0378	0.4739	0.0375	0.4739	0.4908	0.588203231	0.0333
Lakki Marwat	0.4635	0.1937	0.4346	0.1634	0.4346	0.4850	0.889558617	0.1157
Larkana	0.4826	0.0947	0.4768	0.0904	0.4768	0.4097	0.729926751	0.0674
Lasbela	0.0182	0.0184	0.0191	0.0184	0.0191	0.1707	0.121730268	0.0200
Layyah	0.3634	0.0787	0.3747	0.0761	0.3747	0.3979	0.604777804	0.0595
Lehri		0.0000		0.0000		0.4093		0.0199
Lodhran	0.3329	0.0863	0.3480	0.0830	0.3480	0.4420	0.929646212	0.0584
Loralai	0.3382	0.1271	0.3470	0.1172	0.3470	0.3265	0.799290246	0.0848
Lower Dir	0.6418	0.0795	0.6329	0.0769	0.6329	0.6066	0.734831229	0.0590
Malair	1.0000	0.0000	1.0000	0.0000	1.0000	0.9682	1	0.0084
Malakand Protected	0.5237	0.1234	0.5258	0.1144	0.5258	0.3105	0.726876217	0.0849
Mandi Bahauddin	0.4344	0.1252	0.4481	0.1158	0.4481	0.6145	0.940539459	0.0813
Mansehra	0.2909	0.0841	0.3085	0.0811	0.3085	0.3471	0.767459166	0.0591
Mardan	0.5174	0.0714	0.5172	0.0696	0.5172	0.4347	0.868925033	0.0498
Mastung	0.3691	0.1309	0.3636	0.1203	0.3636	0.4106	0.767067003	0.0881
Matiali	0.4267	0.1808	0.5241	0.1565	0.5241	0.5478	0.77537051	0.1129
Mianwali	0.4144	0.1019	0.4288	0.0966	0.4288	0.5948	0.799575849	0.0700
Mirpur	0.9608	0.0198	0.9589	0.0198	0.9589	0.4769	0.954450539	0.0278
Mirpur Khas	1.0000	0.0000	1.0000	0.0000	1.0000	0.9180	1	0.0257

Mohmand	0.6505	0.2274	0.6703	0.1833	0.6703	0.5869	0.949102639	0.1292
Multan	0.4924	0.0557	0.4921	0.0547	0.4921	0.5097	0.60331388	0.0486
Musakhel	0.7646	0.1797	0.6677	0.1549	0.6677	0.5268	0.795020924	0.1181
Muzaffarabad	0.9506	0.0234	0.9482	0.0234	0.9482	0.5009	0.938884927	0.0178
Muzaffargarh	0.5219	0.0763	0.5271	0.0740	0.5271	0.5525	0.829046074	0.0523
Nankana Sahib	0.7396	0.0943	0.7275	0.0901	0.7275	0.7057	0.948782471	0.0634
Narowal	0.4612	0.0863	0.4628	0.0831	0.4628	0.4650	0.873595972	0.0589
Nasirabad/Tambo	0.8012	0.1363	0.7535	0.1253	0.7535	0.6555	0.881140165	0.0904
Naushahro Feroze	1.0000	0.0000	1.0000	0.0000	1.0000	0.9007	1	0.0111
Nawabshah/SBA ¹	0.8217	0.1169	0.7555	0.1091	0.7555	0.7484	0.882559457	0.0770
Neelum	0.9826	0.0136	0.9818	0.0136	0.9818	0.4407	0.975471002	0.0147
Nushki	0.8980	0.0881	0.8881	0.0847	0.8881	0.6857	0.930669317	0.0599
Okara	0.4351	0.0793	0.4425	0.0768	0.4425	0.5254	0.602145391	0.0637
Pakpattan	0.7419	0.1146	0.7093	0.1072	0.7093	0.7315	0.977159741	0.0754
Panjgur	1.0000	0.0000	1.0000	0.0000	1.0000	0.6729	1	0.0165
Peshawar	0.5584	0.0536	0.5548	0.0528	0.5548	0.3732	0.66881177	0.0433
Pishin	0.3723	0.0919	0.3953	0.0880	0.3953	0.4576	0.486201034	0.0850
Poonch	0.9754	0.0220	0.9734	0.0220	0.9734	0.4137	0.957688917	0.0176
Quetta	0.6740	0.0537	0.6669	0.0529	0.6669	0.5217	0.824440829	0.0398
Rahim Yar Khan	0.4185	0.0646	0.4249	0.0632	0.4249	0.5454	0.781635835	0.0464
Rajanpur	0.8041	0.1269	0.7677	0.1171	0.7677	0.7015	0.845204749	0.0831
Rawalpindi	0.4133	0.0431	0.4146	0.0426	0.4146	0.4473	0.698387695	0.0321
Sahiwal	0.5058	0.0810	0.5057	0.0782	0.5057	0.5181	0.763115875	0.0581
Sanghar	0.4502	0.0901	0.4552	0.0864	0.4552	0.5578	0.697157703	0.0664
Sargodha	0.3483	0.0555	0.3533	0.0546	0.3533	0.4097	0.439052146	0.0548
Shahdad Kot	1.0000	0.0000	1.0000	0.0000	1.0000	0.9037	1	0.0183
Shangla		0.0000		0.0000		0.3091		0.0224
Sheikhupura	0.5932	0.0589	0.5883	0.0579	0.5883	0.5437	0.656290737	0.0513
Shikarpur	0.3204	0.1223	0.3233	0.1136	0.3233	0.3941	0.849927967	0.0800

¹Shaheed Benazir Abad

Sialkot	0.3311	0.0474	0.3344	0.0469	0.3344	0.4783	0.559724987	0.0378
Sibi	0.6810	0.1081	0.6545	0.1019	0.6545	0.7384	0.709611436	0.0883
Sohbat Pur	0.8426	0.1430	0.7934	0.1302	0.7934	0.7434	0.795050403	0.0953
South Waziristan	1.0000	0.0000	1.0000	0.0000	1.0000	0.5803	1	0.0168
Sudhonti	0.9360	0.0428	0.9298	0.0424	0.9298	0.4100	0.943470962	0.0326
Sujawal	0.7765	0.0752	0.7732	0.0731	0.7732	0.5767	0.844387061	0.0590
Sukkur	0.5652	0.0902	0.5547	0.0865	0.5547	0.5292	0.710371896	0.0641
Swabi	0.2963	0.0763	0.3067	0.0740	0.3067	0.3475	0.595795638	0.0575
Swat	0.5263	0.0813	0.5189	0.0785	0.5189	0.3872	0.97247532	0.0551
Tando Alla Yar	0.9411	0.0574	0.9310	0.0564	0.9310	0.9173	0.947948942	0.0410
Tando M.Khan ²	0.3799	0.0972	0.4057	0.0927	0.4057	0.5119	0.691424416	0.0681
Tank	0.1206	0.0748	0.1282	0.0727	0.1282	0.2561	0.587039592	0.0536
Tharparkar	0.6986	0.1263	0.6899	0.1167	0.6899	0.7542	0.82999085	0.0845
Thatta	0.4552	0.1236	0.4645	0.1145	0.4645	0.3571	0.806684155	0.0831
Toba Tek Singh	0.3080	0.0602	0.3143	0.0591	0.3143	0.3079	0.723846078	0.0433
Tor Ghar		0.0000		0.0000		0.4462		0.0139
Umer Kot	0.9184	0.0618	0.9103	0.0606	0.9103	0.7469	0.872409502	0.0457
Upper Dir	0.5571	0.1629	0.5677	0.1437	0.5677	0.4504	0.819784604	0.1029
Vehari	0.5272	0.0868	0.5307	0.0834	0.5307	0.5159	0.873099371	0.0593
Washuk	0.3594	0.2827	0.3553	0.2104	0.3553	0.4399	0.871316135	0.1491
Zhob	0.1220	0.1145	0.1523	0.1075	0.1523	0.2635	0.580756961	0.0800
Ziarat	0.6763	0.2645	0.5962	0.1996	0.5962	0.5176	0.931723161	0.1402

²Tando Muhammad khan

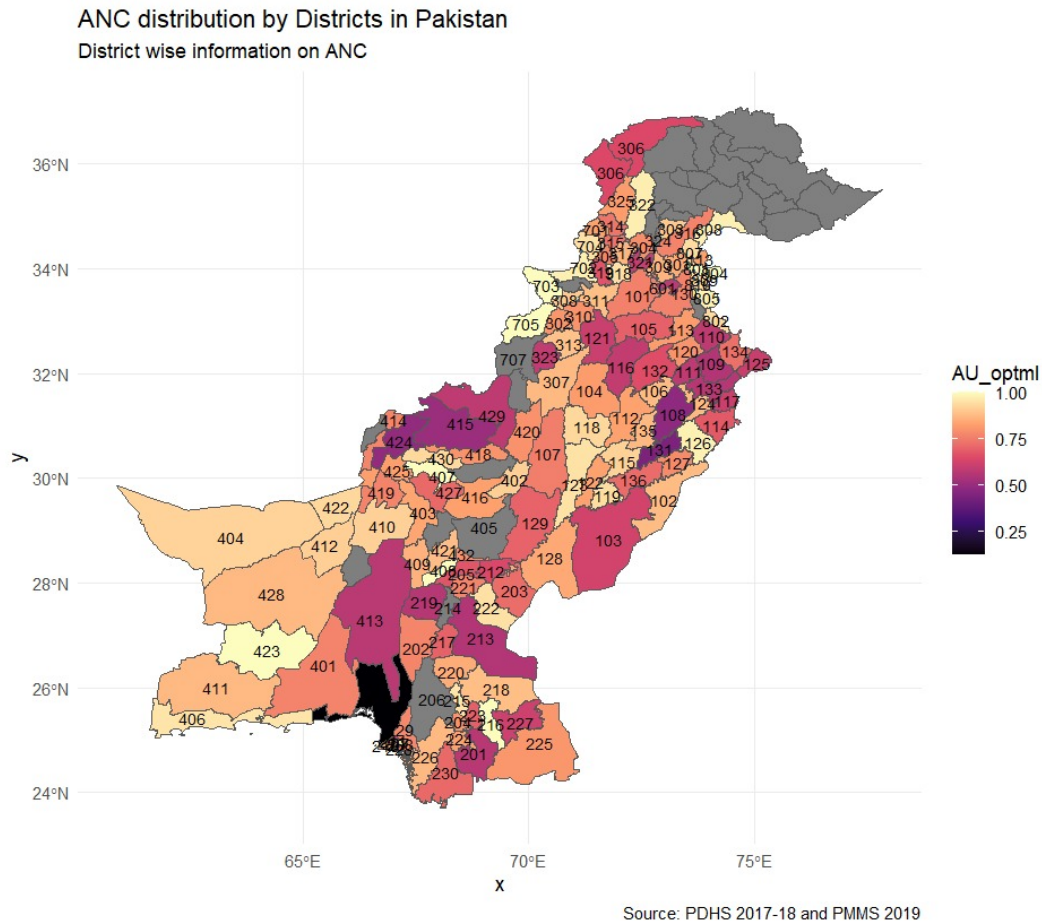


Figure 3.12: ANC distribution by Districts in Pakistan. Source: PDHS 2017-18 and PMMS 2019. The districts are represented by their codes and details are given in Appendix

The map 3.12 "ANC Distribution by Districts in Pakistan" provides a detailed visual representation of the distribution of Antenatal Care (ANC) services across various districts in Pakistan. This visual tool is essential for identifying regional disparities in ANC coverage and understanding the varying levels of healthcare services available to pregnant women in different parts of the country. Covering the entire nation, the map is divided into administrative districts. It includes latitude and longitude coordinates ranging from approximately 24°N to 37°N and 60°E to 75°E, respectively, providing a precise geographical context. Districts are color-coded based on the AUC value, which likely reflects an optimal ANC coverage index. The legend on the right side of the map

uses a gradient scale from 0.25 to 1.00. Dark purple indicates the lowest ANC coverage, while yellow signifies the highest coverage. Intermediate colors such as orange, pink, and various shades of purple represent varying levels of ANC coverage. Each district is labeled with a numerical code, which helps in identifying and cross-referencing the districts. These codes can be matched with corresponding district names in an accompanying dataset or list. The data sources, cited as PDHS (Pakistan Demographic and Health Survey) 2017-18 and PMMS (Pakistan Maternal Mortality Survey) 2019, add credibility and reliability to the information presented. Districts with high ANC coverage, represented in yellow, are dispersed across different regions, indicating some level of equitable distribution while highlighting areas with particularly strong ANC services. Notably, some northern and central districts exhibit high ANC coverage. Conversely, districts with low ANC coverage, shown in dark purple, are mainly concentrated in specific regions such as parts of Balochistan and Khyber Pakhtunkhwa. This suggests significant regional disparities and points to areas where healthcare interventions and ANC programs are most urgently needed. The majority of districts fall within the intermediate range, indicated by shades of orange and pink. These areas have moderate ANC coverage, underscoring the need for sustained or improved healthcare services to reach optimal levels. This map is a critical tool for policymakers and healthcare planners, enabling them to identify regions with insufficient ANC services. It can guide the allocation of resources, the establishment of new healthcare facilities, and targeted healthcare programs to improve ANC coverage in underserved districts. The map also highlights the need for further research to understand the factors contributing to regional disparities in ANC coverage. Exploring socioeconomic, cultural, and infrastructural variables can help develop comprehensive strategies for enhancing maternal healthcare services. In summary, the ANC Distribution Map by Districts in Pakistan offers an in-depth visual analysis of the state of antenatal care across the country. By showcasing regions with varying levels of ANC coverage, it serves as a vital resource for healthcare professionals, policymakers, and researchers working to close gaps in maternal healthcare services and achieve equitable healthcare for all pregnant women in Pakistan.

Chapter 4

4.1 Conclusion

This study has made significant contributions to the field of small area estimation (SAE) by developing a hybrid model that integrates both the Fay-Herriot and Battese models. The innovative Area Cum Unit (ACU) method presented here combines the strengths of area-level and unit-level information to produce more precise and reliable estimates. This approach addresses the inherent limitations of using either model independently and demonstrates improved estimation accuracy and reduced variability. The comprehensive analysis of antenatal care utilization among women in Pakistan, using district-wise data from the Demographic and Health Surveys (DHS) of 2017-18 and 2019, underscores the importance of integrating multiple data sources to enhance the quality of small area estimates. By incorporating both individual-level factors such as age, sex, education, and wealth index, and district-level characteristics such as healthcare infrastructure and socio-economic indicators, this study has provided a nuanced understanding of the determinants influencing antenatal care utilization. The results indicate that the ACU method significantly outperforms traditional methods like the direct estimation, Fay-Herriot, and Battese-Harter-Fuller models in terms of minimizing mean squared error (MSE) and coefficient of variation (CV). This improved performance highlights the potential of the ACU method to provide more accurate and reliable estimates, which is crucial for effective policy-making and resource allocation, particularly in the context of Sustainable Development Goals (SDGs) and public health interventions. Furthermore, the application of the ACU method in this study

has demonstrated its adaptability and effectiveness in various contexts, from predicting disease prevalence to assessing forest characteristics. The methodological advancements presented here pave the way for future research to explore and refine hybrid models further, enhancing their applicability and robustness across different domains. In conclusion, the development and application of the ACU method represent a significant step forward in small area estimation techniques. This study has not only addressed existing gaps in the literature but also provided practical solutions for improving the accuracy and reliability of small area estimates. The findings underscore the importance of integrating diverse data sources and employing advanced statistical methods to meet the growing demand for high-quality, disaggregated data in both public and private sectors.

Bibliography

- [1] Clara Aida Khalil, Stefano Di Candia, Piero Demetrio Falorsi, and Pietro Gennari. Integrating surveys with geospatial data through small area estimation to disaggregate sdg indicators: A practical application on sdg indicator 2.3. 1. *Statistical Journal of the IAOS*, 38(3):879–891, 2022.
- [2] Elizabeth Stuart and Emma Samman. Defining ‘leave no one behind’. *ODI Briefing Note*. London: Overseas Development Institute, 2017.
- [3] NAUSHEEN SARWAR. Sustainable development goal 6: Clean water and sanitation in pakistan.
- [4] John NK Rao and Isabel Molina. *Small area estimation*. John Wiley & Sons, 2015.
- [5] World Health Organization et al. *WHO recommendations on antenatal care for a positive pregnancy experience*. World Health Organization, 2016.
- [6] Malay Ghosh and John NK Rao. Small area estimation: an appraisal. *Statistical science*, 9(1):55–76, 1994.
- [7] Jon NK Rao and Mingyu Yu. Small-area estimation by combining time-series and cross-sectional data. *Canadian Journal of Statistics*, 22(4):511–528, 1994.
- [8] Rachel Margaret Harter. *SMALL AREA ESTIMATION USING NESTED-ERROR MODELS AND AUXILIARY DATA (LANDSAT)*. Iowa State University, 1983.

- [9] Brenda MacGibbon and Thomas J Tomberlin. *Small area estimates of proportions via empirical Bayes techniques*. Faculty of Commerce and Administration, Concordia University, 1987.
- [10] George E Battese, Rachel M Harter, and Wayne A Fuller. 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, 1988.
- [11] Jonathan Wakefield, Taylor Okonek, and Jon Pedersen. Small area estimation for disease prevalence mapping. *International Statistical Review*, 88(2):398–418, 2020.
- [12] Yue Lin. Synthetic population data for small area estimation in the united states. *Environment and Planning B: Urban Analytics and City Science*, 51(2):553–562, 2024.
- [13] P Corey Green, Harold E Burkhart, John W Coulston, and Philip J Radtke. A novel application of small area estimation in loblolly pine forest inventory. *Forestry: An International Journal of Forest Research*, 93(3):444–457, 2020.
- [14] Zafar Ahmed, Shariq Khoja, and S Suha Tirmizi. Antenatal care and the occurrence of low birth weight delivery among women in remote mountainous region of chitral, pakistan. 2012.
- [15] Z Fatmi and Bilal Iqbal Avan. Demographic, socio-economic and environmental determinants of utilisation of antenatal care in a rural setting of sindh, pakistan. *Journal of Pakistan Medical Association*, 52(4):138, 2002.
- [16] Yordanos B Molla, Barbara Rawlins, Prestige Tatenda Makanga, Marc Cunningham, Juan Eugenio Hernández Ávila, Corrine Warren Ruktanonchai, Kavita Singh, Sylvia Alford, Mira Thompson, Vikas Dwivedi, et al. Geographic information system for improving maternal and newborn health: recommendations for policy and programs. *BMC pregnancy and childbirth*, 17:1–7, 2017.

- [17] Ita Wulandari, Khairil Anwar Notodiputro, Anwar Fitrianto, and Anang Kurnia. Hierarchical bayesian models for small area estimation under overdispersed count data. *Engineering Letters*, 31(4), 2023.
- [18] Andrea Neri and Eleonora Porreca. Total bias in income surveys when nonresponse and measurement errors are correlated. *Journal of Survey Statistics and Methodology*, page smad027, 2023.
- [19] Paul Corral, Isabel Molina, Alexandru Cojocaru, and Sandra Segovia. *Guidelines to small area estimation for poverty mapping*. World Bank Washington, 2022.
- [20] Wenceslao González-Manteiga, Maria J Lombardía, Isabel Molina, Domingo Morales, and Laureano Santamaría. Bootstrap mean squared error of a small-area eblup. *Journal of Statistical Computation and Simulation*, 78(5):443–462, 2008.
- [21] Isabel Molina, Paul Corral, and Minh Nguyen. Estimation of poverty and inequality in small areas: review and discussion. *Test*, 31(4):1143–1166, 2022.
- [22] María Guadarrama, Isabel Molina, and JNK Rao. Small area estimation of general parameters under complex sampling designs. *Computational Statistics & Data Analysis*, 121:20–40, 2018.
- [23] Isabel Molina and Jon NK Rao. Small area estimation of poverty indicators. *Canadian Journal of statistics*, 38(3):369–385, 2010.
- [24] Hyeong Choi and Richard Schoen. The Space of Minimal Embeddings of a Surface into a 3–dimensional Manifold of Positive Ricci Curvature. *Inventiones Mathematicae*, **81**:387–394, 1985.
- [25] Robert Osserman. *A Survey of Minimal Surfaces*. Van Nostrand Reinhold Company, 1969.
- [26] Thomas H. Colding and William P. Minicozzi. *Minimal Surfaces*. Courant Lecture Notes in Math, 1999.

- [27] Robert Gulliver. Removability of Singular Points on Surfaces of Bounded Mean Curvature. *The Journal of Differential Geometry*, **11**:345–350, 1976.
- [28] Yong You and Beatrice Chapman. Small area estimation using area level models and estimated sampling variances. *Survey Methodology*, 32(1):97, 2006.
- [29] Geografía e Informática (México) Instituto Nacional de Estadística. *ENIGH-92: encuesta nacional de ingresos y gastos de los hogares*. INEGI, 1993.
- [30] Matthew J Gurka, Lloyd J Edwards, Keith E Muller, and Lawrence L Kupper. Extending the box–cox transformation to the linear mixed model. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 169(2):273–288, 2006.
- [31] JNK Rao and I Molina 2nd. Small area estimation, a john wiley & sons. *Inc, New Jersey*, 2003.
- [32] Ann-Kristin Kreutzmann, Sören Pannier, Natalia Rojas-Perilla, Timo Schmid, Matthias Templ, and Nikos Tzavidis. The r package emdi for estimating and mapping regionally disaggregated indicators. *Journal of Statistical Software*, 91, 2019.
- [33] Sylvia Harmening, Ann-Kristin Kreutzmann, Sören Schmidt, Nicola Salvati, and Timo Schmid. A framework for producing small area estimates based on area-level models in r. *R Journal*, 15(1), 2023.
- [34] Gary Brown, Ray Chambers, Patrick Heady, and Dick Heasman. Evaluation of small area estimation methods—an application to unemployment estimates from the uk lfs. In *Proceedings of statistics Canada symposium*, volume 2001, pages 1–10. Statistics Canada, 2001.
- [35] Carolina Casas-Cordero Valencia, Jenny Encina, and Partha Lahiri. Poverty mapping for the chilean comunas. *Analysis of poverty data by small area estimation*, pages 379–404, 2016.

- [36] Sandra Hadam, Nora Würz, Ann-Kristin Kreutzmann, and Timo Schmid. Estimating regional unemployment with mobile network data for functional urban areas in germany. *Statistical Methods & Applications*, pages 1–29, 2023.
- [37] Juan Galeano, Albert Esteve, Anna Turu, Joan García-Roman, Federica Becca, Huifen Fang, Maria Pohl, and Rita Trias-Prats. Coresidence: National and sub-national data on household size and composition around the world, 1964–2021. *Scientific data*, 11(1):145, 2024.
- [38] Huilin Li and Partha Lahiri. An adjusted maximum likelihood method for solving small area estimation problems. *Journal of multivariate analysis*, 101(4):882–892, 2010.
- [39] Thomas Lumley. Analysis of complex survey samples. *Journal of statistical software*, 9:1–19, 2004.
- [40] Alessandra Petrucci and Nicola Salvati. Small area estimation for spatial correlation in watershed erosion assessment. *Journal of agricultural, biological, and environmental statistics*, 11:169–182, 2006.
- [41] Timo Schmid, Fabian Bruckschen, Nicola Salvati, and Till Zbiranski. Constructing sociodemographic indicators for national statistical institutes by using mobile phone data: estimating literacy rates in senegal. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 180(4):1163–1190, 2017.
- [42] Eric V Slud and Tapabrata Maiti. Mean-squared error estimation in transformed fay–herriot models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(2):239–257, 2006.
- [43] Sebastian Warnholz. *Small Area Estimation Using Robust Extensions to Area Level Models: Theory, Implementation and Simulation Studies*. PhD thesis, 2016.
- [44] Lynn MR Ybarra and Sharon L Lohr. Small area estimation when auxiliary information is measured with error. *Biometrika*, 95(4):919–931, 2008.

- [45] Masayo Yoshimori and Partha Lahiri. A new adjusted maximum likelihood method for the fay–herriot small area model. *Journal of Multivariate Analysis*, 124:281–294, 2014.
- [46] María Guadarrama, Isabel Molina, and JNK Rao. A comparison of small area estimation methods for poverty mapping. *Statistics in Transition new series*, 17(1):41–66, 2016.
- [47] Natalia Rojas-Perilla, Sören Pannier, Timo Schmid, and Nikos Tzavidis. Data-driven transformations in small area estimation. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 183(1):121–148, 2020.
- [48] Yong You and JNK1944372 Rao. A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. *Canadian Journal of Statistics*, 30(3):431–439, 2002.
- [49] Paul Corral Rodas, Isabel Molina, and Minh Nguyen. Pull your small area estimates up by the bootstraps. *Journal of Statistical Computation and Simulation*, 91(16):3304–3357, 2021.
- [50] Takaaki Masaki, David Newhouse, Ani Rudra Silwal, Adane Bedada, and Ryan Engstrom. Small area estimation of non-monetary poverty with geospatial data. *Statistical Journal of the IAOS*, 38(3):1035–1051, 2022.
- [51] Roy Van der Weide. 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), 2014.
- [52] Nikos Tzavidis, Li-Chun Zhang, Angela Luna, Timo Schmid, and Natalia Rojas-Perilla. 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, 2018.

- [53] Robert E Fay III and Roger A Herriot. 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, 1979.
- [54] Andrés Gutiérrez, Álvaro Fuentes, Xavier Mancero, and Felipe López. Criterios de calidad en la estimación de indicadores a partir de encuestas de hogares.
- [55] Michel A Hidioglou, Jean-François Beaumont, and Wesley Yung. Development of a small area estimation system at statistics canada. *Survey Methodology*, 45(1):101–126, 2019.
- [56] Jiming Jiang, P Lahiri, Shu-Mei Wan, and Chien-Hua Wu. Jackknifing in the fay-herriot model with an example. *Proc. Sem. Funding Opportunity in Survey Research*, pages 75–97, 2001.
- [57] Kirk M Wolter and Kirk M Wolter. The bootstrap method. *Introduction to variance estimation*, pages 194–225, 2007.