Estimation of Early-Age Mortality in Different Districts of Pakistan

By

Shelina Zaman
(00000401543)

Department of Statistics

School of Natural Sciences (SNS)

National University of Sciences and Technology (NUST)

Islamabad, Pakistan

(2024)

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS thesis written by **Shelina Zaman** (Registration No **00000401543)**, 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: 23-08-2024

Signature (HoD): ____________

Date: 26/8/2024

Signature (Dean/Principal): ____________

Date: 26/8/2024

FORM TH-4

National University of Sciences & Technology

MS THESIS WORK

We hereby recommend that the dissertation prepared under our supervision by **Shelina Zaman**, Regn No. **00000401543** Titled **Estimation of Early- age Mortality in Different Districts of Pakistan** 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

Date: 26/8/2024

COUNTERSINGED

Date: 26/8/2024

Dean/Principal

This thesis is dedicated to My Parents,
for their constant support and encouragement.

Acknowledgments

I want to start by expressing my deep gratitude to **ALLAH ALMIGHTY**. His countless blessings and unwavering guidance have been my main sources of strength and have empowered me to pursue my studies at NUST. Without His divine intervention, none of this would have been achievable.
I would also like to express my deepest gratitude to my supervisor, Dr. Shakeel Ahmed, whose extensive knowledge and the invaluable materials he provided were instrumental in guiding me through this research. His unwavering support and insightful advice made the entire process much more manageable, enabling me to successfully complete my work.
I am immensely thankful to my parents and my fiancé for their unshakeable belief in me. Their unconditional love, support, and guidance, along with their patience during my intense focus on studying and thesis work, have been a constant source of strength. Their prayers and encouragement have carried me through the most challenging moments of this journey.
I want to sincerely thank my dear friend Tazeen Zahra for being such a great help as I learned LaTeX. Her guidance and patience made the whole process so much easier and less overwhelming. I'm incredibly grateful for all the time and effort she put into helping me understand it, and I couldn't have done it without her support.
Lastly, I would like to express my deep appreciation to all the individuals whose work I referenced in my thesis. Their invaluable contributions significantly enriched my understanding and were vital to the completion of my research.

Contents

List of Tables

List of Figures

List of Abbreviations

SAE	Small Area Estimation
NMR	Neonatal Mortality Rate
CMR	Child Mortality Rate
IMR	Infant Mortality Rate
SDG's	Sustainable Development Goals
PNMR	Post-Neonatal Mortality Rate
PDHS	Pakistan Demographic and Health Survey
NGOs	Non-Governmental Organizations
CV	Coefficient of Variation
SE	Standard Error
MSE	Mean Square Errors
SF	Survival Function
SRSWOR	Simple Random Sampling Without Replacement
LCI	Lower Confidence Interval
UCI	Upper Confidence Interval

Abstract

Estimating early-age mortality at more localized administrative levels offers public health researchers a deeper insight into infant well-being and aids in developing health policies. To achieve this, design-based strategies are proposed for estimating the survival function, which is then applied to calculate district-level estimates of infant and neonatal mortality rates. The proposed survival function estimate employs an empirical distribution function approach. The study evaluates four different strategies by comparing their relative bias and coefficient of variation. Using data from the Pakistan Demographic and Health Surveys (PDHS) of 2017-18 and a special 2019 survey, the research emphasizes the significance of combining health survey data with administrative records to generate small-area health outcomes. The study introduces four methods for producing small-area estimates by integrating data from consecutive surveys through direct, synthetic, and composite methods. Among these, composite regression under Strategy 2 is identified as the most effective in terms of coefficient of variation and bias. The findings of this study are intended to assist public health policymakers in creating informed policies for areas with limited data and providing a clearer picture of infant mortality rates for governments and NGOs focused on neonatal mortality.

Keywords: *Indicator function, Neonatal mortality, Infant mortality, Direct methods, Indirect methods*

Chapter 1

Introduction and Motivation

1.1 Background of the study

The neonatal mortality rate (NMR) is a crucial indicator under the Sustainable Development Goals, particularly the aim to end preventable deaths of newborns and children under 5 years of age by 2030. The NMR specifically measures the number of infant deaths within the first 28 days of life per 1,000 live births in a given year. This rate is a direct reflection of the standard and availability of maternity and newborn healthcare services, making it an essential measure of the overall health and well-being of a community. High newborn death rates often highlight significant gaps in healthcare systems, such as inadequate facilities, a shortage of skilled medical professionals, limited access to necessary medications and immunizations, and poor maternal nutrition and health. Infant mortality, which is the probability of a child born in a specific year or period dying before reaching the age of 1 year, expressed per 1,000 live births, is another critical indicator. Comprehensive strategies are needed to reduce both neonatal and infant mortality rates. These strategies include enhancing emergency obstetric care, increasing skilled birth attendance, promoting breastfeeding and immunizations, and improving both prenatal and postnatal care. The shared global goal is to reduce neonatal mortality to 12 per 1,000 live births or less. Achieving this target is not only vital for improving child survival rates but also for fostering long-term social and economic development. After all, healthy, happy children are the foundation of prosperous, productive societies[1]. The United Nations' Sustainable Development Goal (SDG) 3 focuses on ensuring healthy lives and promoting well-being for people of all ages. To achieve this, it is essential to gather detailed information at more granular administrative levels using Small Area Estimation (SAE) techniques. Many of the indicators relevant to SDG 3 are derived from surveys or administrative data sources. In the United States, a study applied the SAE method to the Behavioral Risk Factor Surveillance System (BRFSS) data from 1999 to 2005, estimating the prevalence of obesity across 398 communities within the Commonwealth of Massachusetts[2]. Rao and Molina (2015) stress the importance of small area estimation (SAE) techniques in enhancing the accuracy of estimates for subpopulations with limited sample sizes. These methods involve combining survey data with additional information from sources such as administrative records or censuses to generate more precise estimates. SAE techniques, including empirical best linear unbiased predictors (EBLUP) and hierarchical Bayes methods, are especially helpful in dealing with the challenges posed by small sample sizes in specific areas. By utilizing these advanced statistical methods, policymakers can obtain reliable estimates at a more detailed geographic level, which is crucial for effective resource allocation and policy interventions. For instance, SAE can help identify areas with high disease prevalence that were previously undetected due to inadequate data, enabling targeted health interventions. Additionally, Rao and Molina emphasize the potential of model-based approaches in enhancing the efficiency and accuracy of health surveys, thereby supporting more informed decision-making processes in public health. This comprehensive approach ensures that health interventions are not only timely and effective but also distributed equitably, ultimately contributing to improved health outcomes across all regions [3]. Rao and Molina discuss different models and techniques for improving the accuracy of small area estimates. These methods include using mixed-effects models, which consider both fixed and random effects, providing a strong foundation for estimating small area parameters. The authors also investigate the use of benchmarking procedures to ensure that small area estimates align with known large area totals, thus enhancing the credibility of the estimates. They stress the importance of validating small area models using real-world data and simulations to evaluate their performance and reliability.Rao and Molina discuss different models and techniques for improving the accuracy of small area estimates. These methods include using mixed-effects models, which consider both fixed and random effects, providing a strong foundation for estimating small area parameters. The authors also investigate the use of benchmarking procedures to ensure that small area estimates align with known large area totals, thus enhancing the credibility of the estimates. They stress the importance of validating small area models using real-world data and simulations to evaluate their performance and reliability.[3]. Rao and Molina highlighted an important development in SAE, by emphasizing the use of spatial and spatio-temporal models. These models incorporate geographic and temporal correlations into the estimation process. They are particularly useful in public health for tracking the spread of diseases and changes in health indicators over time. By incorporating spatial and temporal data, these models can generate more accurate and timely estimates, which are essential for planning effective interventions and allocating resources.. [3]. Furthermore, integrating SAE techniques with R programming enables the visualization of health data at detailed levels, which helps in better understanding and communication of health disparities. Policymakers and public health officials can utilize these visual tools to pinpoint areas with health issues, strategize targeted interventions, and track the effects of health policies over time. [3]. Rao and Molina discuss not only the technical advancements but also the practical implementation of SAE methods in various national health surveys and censuses. They provide case studies from different countries to demonstrate the successful application of SAE in improving the accuracy and utility of health data. These case studies highlight the versatility of SAE methods in addressing diverse health data challenges, from estimating vaccination coverage in remote areas to monitoring chronic disease prevalence in urban settings. [3].

1.2 Definition of terminologies

1.2.1 Small Area Estimation

Small Area Estimation (SAE) is a statistical method that uses survey data along with additional information to produce more accurate and precise estimates for specific sub-populations or small geographic areas with limited sample sizes. These sub-populations can include geographic regions like districts or socioeconomic subgroups such as the age of a child at death. [4].

1.2.2 Direct method

When an estimate for a particular variable of interest in a population is acquired by directly sampling from that population, it is referred to as a direct estimator. This method does not rely on any information from other areas or external sources.[4].

1.2.3 Indirect method

The indirect method in Small Area Estimation (SAE) combines data from the small area with information from other areas to improve estimates. This method uses statistical models to make the estimates more accurate and reliable[4].

1.3 Neonatal Mortality

The probability of a newborn dying within the first month of life.

$$\text{Neonatal Mortality Rate (NMR)} = \frac{\text{Number of deaths of infants aged 0-28 days}}{\text{Total number of live births}} \times 1000$$

1.3.1 Infant Mortality

The chance of an infant dying between birth and their first birthday.

$$\text{Infant Mortality Rate (IMR)} = \frac{\text{Number of deaths of infants aged 0-1 year}}{\text{Total number of live births}} \times 1000$$

1.4 Literature Review

Small area estimation (SAE) has gained significant importance because there is a growing need for reliable small area statistics, even when only very small samples are available. Traditional survey methods often struggle to provide precise estimates for small areas because of limited sample sizes, leading to high variability and potential bias. To overcome these challenges, model-based methods in SAE have been developed. The methods mentioned above help improve the precision and reliability of estimates for small areas by using additional information from sources such as census and administrative records. This is crucial for ensuring the quality of statistical data, which is necessary for making effective policies and allocating resources. One of the significant advancements in Small Area Estimation (SAE), highlighted by Pfeffermann in 2007, is the use of hierarchical and empirical Bayes methods. These methods allow statisticians to create complex models to predict quantities in target areas and evaluate their mean square errors (MSE). For instance, the hierarchical Bayes approach involves setting prior distributions for the parameters and then updating them with observed data, resulting in more accurate estimates. This method is particularly useful when traditional survey techniques are not sufficient, such as when dealing with very small sample sizes or zero sample scenarios. Pfeffermann also emphasizes the importance of accounting for correlations among small area random effects, which represent unexplained variations in the target quantities. Incorporating these correlations into the models can significantly improve the precision of the estimates. Time series models and discrete measurement models are particularly useful in this regard, where time series models use data from previous occasions to strengthen current estimates, and discrete measurement models handle categorical or binary data commonly found in many surveys. Another important aspect of SAE is the use of synthetic and composite estimators. Synthetic estimators utilize information from larger, assumed-to-be-homogeneous areas to provide estimates for smaller areas. While this approach can reduce variance, it may introduce bias if the assumption of homogeneity is incorrect. Composite estimators address this issue by combining direct and synthetic estimates to minimize mean square error, striking a balance between reducing variance and controlling bias. These advanced SAE methodologies have extensive practical applications and are increasingly used by national statistical agencies to meet the increasing demand for detailed local-level data. For instance, estimates of drug use, employment rates, and other socio-economic indicators at the state or sub-state level heavily rely on these methods. By integrating various data sources and sophisticated modeling techniques, SAE provides high-quality and actionable statistical insights even with limited or complex data. This capability is crucial for supporting effective policy-making, resource distribution, and regional planning, ensuring that decisions are informed by the most accurate and reliable data available. [5]. Pfeffermann (2013) made significant strides in the field by highlighting hierarchical and empirical Bayes methods. These methods improve the accuracy of estimates for small geographic areas by using complex models to predict area-specific quantities and assess their mean square errors (MSE). They are especially effective in considering correlations among small area random effects, which enhances the reliability of estimates. Furthermore, the practical applications of small area estimation (SAE) have expanded to include both design-based and model-dependent approaches, which are essential for generating dependable small area statistics. These statistics are crucial for policymaking, resource allocation, and regional planning as they offer high-quality, actionable data even with small or complex sample sizes. [6]. In the field of health decision-making, small area estimation (SAE) plays a significant role. Research conducted by various authors illustrates how SAE methodologies can produce reliable health statistics for smaller geographic areas, enabling targeted health interventions and well-informed policy-making. However, when dealing with small or zero population sample sizes, traditional direct estimators are ineffective in providing reliable estimates (Cochran, 1977). This issue is addressed by synthetic estimators (Ghosh and Rao, 1994; Gonzalez et al., 1996; Purcell and Kish, 1973), which use larger area estimates to represent similar characteristics of relevant small areas. In 1968, the National Center for Health Statistics in the United States pioneered the use of synthetic estimation through the National Health Interview Survey to overcome the challenge of small sample sizes in accurately estimating state statistics (Gonzalez et al., 1973). Synthetic estimation involves post-stratification, where a non-homogeneous population is classified into homogeneous sub-populations, thereby improving the accuracy of estimates by using information from larger populations. While synthetic methods effectively reduce variance, they may introduce bias in small area estimates. To balance this, composite estimators, which blend direct and indirect estimation techniques, are employed to trade-off between variance and bias, thus enhancing the reliability of small area statistics (Holt and Smith, 1979) [7]. Inference regarding survey sample has frequently concentrated on the design-based (or randomisation) method. This attention to detail has been extended to small area estimation (SAE). Compared to model-based approaches, which constitute the mainstay of mainstream spatial statistics, this approach is very different. Skinner and Wakefield (2017) analyse both inferential approaches. Design-based approaches average over all potential samples that could have been drawn using the given sampling design in order to evaluate the frequentist qualities of estimators. According to this paradigm, the population's response values are fixed rather than random. Bayesian or frequentist methods can be used in model-based strategies. A probabilistic model is given for the replies, which are now considered to be random variables, if the hypothetical infinite population model-based method is chosen. Within the framework of the design-based paradigm, modelling can be accomplished through model-assisted ways (Särndal et al., 1992). In this approach, desirable design-based qualities are preserved even in cases where the model is misspecificated. Lehtonen and Veijanen (2009) present a cautious viewpoint that suggests that while a model-based method might be required in cases of sparse data, design-based (including model-assisted) inference might be dependable in circumstances involving large or medium samples. Datta (2009) reviews and expresses greater enthusiasm about model-based techniques in a companion piece. In this sense, one can ignore a simple random sample (SRS). Nonetheless, non-ignorable designs are present in the majority of real-world household surveys, and the majority of SAE models rely on design-specific assumptions. Design needs to be incorporated into the model when it is not ignorable. The appropriate design elements, such as design weighting, clustering, non-response corrections, and weight modifications, should ideally be incorporated. Many characteristics of the sample frame, for example, the locations of every cluster in a design using cluster sampling, are usually not available, at least not in a way that makes them usable. However, this information may be available. It is usually possible to obtain stratification, clustering, and estimation weights for surveys like the Demographic and Health Surveys (DHS), which are conducted widely in low- and middle-income countries (LMIC). Limited data may be accessible for surveys conducted in developed nations (Wakefield, Okonek, and Pedersen, 2020) [8]. The concept of cumulating survey data over time, initially proposed by Leslie Kish in the mid-20th century, has evolved significantly. Kish advocated for the "rolling sample" design, which uses non-overlapping monthly panels aggregated over varying time periods based on the size of the analytical domain. This approach addresses the need for detailed spatial and temporal data, contrasting with traditional methods that often overlook temporal variations. The use of rolling samples, such as the American Community Survey (ACS) and the Health Care Survey of Department of Defense Beneficiaries (HCSDB), demonstrates the effectiveness of this method in obtaining current and accurate data without sacrificing the necessary sample sizes for estimating data in small areas. The shift towards rolling samples is motivated by the demand for frequent and precise data, especially in small geographical areas. While traditional large-scale surveys like the decennial census provide detailed and unbiased estimates, they lack timeliness. On the other hand, rolling samples, conducted on a monthly or quarterly basis, offer more frequent estimates, allowing researchers to identify trends and changes more easily. This approach is particularly advantageous for capturing seasonal trends or sudden shifts, providing more reliable average estimates over time. For example, the Adult HCSDB switched to quarterly surveys in 2001 to ensure current information on military health system beneficiaries, combining quarterly data into annual datasets to maintain the necessary sample sizes for analysis of small domains. Careful consideration of weighting techniques is required when combining data from multiple small surveys into comprehensive datasets. Kish suggested various methods, including giving full weight to the most recent year, equally weighting each year, or applying monotonically non-decreasing weights based on recency or other criteria. The kind of data and the estimation objectives play a major role in the technique selection. In the case of the Adult HCSDB, equal weighting of quarterly surveys was implemented, assuming that variations between quarters were due to sampling rather than actual differences. This method was found to be effective in providing precise combined estimates by averaging quarterly data, thereby enhancing the reliability of small area estimates. [9]. The combination of data from various surveys can greatly improve the accuracy of prevalence estimates for specific regions, especially in health-related studies. Traditional methods, such as multiple-frame and statistical matching, require individual-level data, which may not always be available. In these cases, aggregate estimates from different sources become crucial, despite potential biases and inconsistencies in methodology. Bayesian hierarchical models provide a strong solution for bias adjustment as they can integrate information from all available sources to give more precise estimates. This method is particularly useful for estimating smoking prevalence across different local authorities, where data from multiple surveys can differ in time, sample size, and transparency of methodology (Manzi et al., 2011). Classical small area estimation methods often have limitations when individual-level data is not accessible. In practical research scenarios, aggregate data from commercial surveys may be more readily available and frequently updated compared to official surveys. However, these commercial surveys often lack transparency in their methodology, which can lead to biases. Bayesian models address these issues by allowing for additional biases within and between data sources. This modeling approach was employed for smoking prevalence data from seven different surveys, adjusting for biases and integrating information from all sources to produce more reliable estimates. These estimates are crucial for public health officials and policymakers to develop effective health promotion strategies tailored to specific regions (Manzi et al., 2011). Moreover, the Bayesian framework enables the assessment of the correlation between different data sources, which is important when sources share similar methodologies or underlying data. For example, estimates derived from the Health Survey for England were found to be more reliable than those from commercial sources due to their known sampling plans and methodologies. However, commercial surveys, despite their biases, provided valuable trend information and finer temporal resolution. By integrating these diverse data sources, Bayesian models can offer comprehensive and nuanced prevalence estimates, essential for addressing area-specific health concerns and improving public health interventions (Manzi et al., 2011). [10]. Small area estimation (SAE) is an important method for obtaining accurate socioeconomic and health statistics for small geographical areas when survey data alone is not sufficient. By combining auxiliary information, primarily from administrative records, SAE enhances the precision of estimates by using related data. Administrative records, derived from government programs, offer valuable data that can improve inferences from survey data. However, there are practical considerations for identifying and preparing administrative records for use in small area estimation models. One major challenge is ensuring the quality and relevance of the covariates from these records. While administrative records cover large populations and are cost-effective, they may not accurately represent the population of interest or measure the desired quantities directly. For example, data from the IRS can provide covariates for estimating poverty rates, but they may exclude low-income households that do not file tax returns, leading to measurement errors. It is now feasible to link administrative records with sample survey and census data thanks to recent developments in computing. This makes it possible to develop sophisticated model-based methods for small area estimation, which by combining data from many sources, can improve estimate accuracy. These methods can also consider spatial and temporal variations, leading to more detailed and reliable estimates. The evolution of SAE methodologies highlights the importance of using administrative records to meet the increasing demand for detailed and accurate statistics at the small area level. [11]. The basic SAE methods, such as synthetic estimators, utilize broader area-level estimates to represent small areas, thus improving the reliability of estimates. Composite estimators, which combine direct and synthetic methods, offer a balanced trade-off between bias and variance, making them particularly useful in contexts with limited data. These methods have been applied effectively in various fields, including health decision-making, where they support policy implementation by providing detailed insights into population characteristics at smaller geographical scales. The study by Ahmed (2024) introduces innovative strategies to enhance the performance of SAE, particularly through the integration of auxiliary information and successive surveys, which significantly improve the accuracy and reliability of estimates in different sub-populations [12]. In Section 2, we employed a two-occasion Small Area Estimation (SAE) approach for analyzing the survival function. Section 3 delves into the proposed SAE strategies, which include direct, synthetic, and composite methods. By applying these strategies, we compared the efficiency of various estimators and conducted parameter estimations to obtain reliable estimates of child health indicators at the district level. The study concludes with a discussion in Section 5, offering recommendations for future research and practice.

1.5 Problem Statement

Surveys are essential for gathering health data but they can be costly and are primarily useful for large populations. They do not provide detailed information for smaller groups or sub-populations, making it difficult to create targeted policies. While survey data can help guide policies for an entire population, it is insufficient for addressing the specific needs of smaller sub-populations.

Advancements in statistical methods have improved our ability to study the effects of the built environment on health outcomes. However, geographic health researchers still struggle to obtain reliable estimates for smaller areas such as districts or regions due to the absence of detailed data. When survey data is limited by small sample sizes or insufficient detail, researchers turn to small area estimation (SAE) techniques to produce more accurate estimates.
As the use of SAE in health research grows, it is crucial for researchers to understand the methods used to obtain these estimates as well as their strengths and limitations. Therefore, there is a need for a more robust, data-driven approach to generate reliable health estimates at smaller area levels, ensuring effective policy implementation for all population segments.

1.6 Objectives of the Study

- The study aims to create accurate estimates of infant and neo-natal deaths at more smaller administrative levels by combining data from Survey-1 and Survey-2.
- The main goal is to improve methods for calculating Early-age mortality rates in small regions and among various demographic groups.
- The Small Area Estimation (SAE) method will be used for this purpose.
- The results will help in developing better public health strategies and assist governments and NGOs in tackling neonatal mortality.

Chapter 2

Material and Methods

2.1 Coefficient of Variation of Estimators

Another useful indicator of the precision of an estimation is the coefficient of variation. The coefficient of variation is a measure of error relative to an estimator, defined as:

$$\mathrm{CV}(\hat{\theta}) = \frac{\mathrm{SE}(\hat{\theta})}{\hat{\theta}} \tag{2.1}$$

2.2 Small Area Direct Estimator for Survival Function

Consider a sample s of size n from a population U of size N using some sampling design P. Let the variable T represent the time to event variable with value t_j for the jth ($j = 1, 2, \ldots, N$) population unit. An empirical Cumulative Distribution Function (CDF) based measure of the population survival function at time t is defined as:

$$F(t) = \sum_{j \in U} \frac{I(t_j < t)}{N} \tag{2.2}$$

A sample version, known as the empirical CDF estimator, is then obtained as:

$$\hat{F}(t) = \sum_{j \in s} \frac{I(t_j < t)}{N} \tag{2.3}$$

Let $U_1, U_2, \ldots, U_m$ be m domains contained in U such that $\bigcup_{i=1}^{m} U_i = U$ with sizes N_i for the ith domain ($i = 1, 2, \ldots, m$). The parameter of interest is the survival function at time t in the ith domain U_i, i.e.,

$$F_i(t) = \sum_{j \in U_i} \frac{I(t_j < t)}{N_i} \quad \text{for } i = 1, 2, \ldots, m \tag{2.4}$$

Here, N and n can be regarded as the population and sample at risk at time point t. Let s_i be the set of n_i units in the sample belonging to U_i such that $\bigcup_{i=1}^{m} s_i = s$. A sample version of the equation 2 is then obtained as:

$$\hat{F}_i(t) = \sum_{j \in s_i} \frac{I(t_j < t)}{n_i} \quad \text{for } i = 1, 2, \ldots, m \tag{2.5}$$

The estimator in this equation 2.5 assumes that the sample is taken using simple random sampling without replacement (SRSWOR) and that equal weights are assigned to each unit. Assuming π_{ij} is the inclusion probability of the jth unit in the ith domain in the sample, the sample weight can be expressed as $w_{ij} = 1/\pi_{ij}$. A weighted version of the estimator given in the equation is obtained as:

$$\hat{F}_{w_i}(t) = \sum_{j \in s_i} \frac{w_{ij} I(t_j < t)}{\sum_{j \in s_i} w_{ij}} \tag{2.6}$$

A reliable weight w_{ij} can be obtained by using an adjustment factor g_{ij} and the final weight is updated to $w^*_{ij} = w_{ij} g_{ij}$ for $j \in s_i$ and $i = 1, 2, \ldots, m$. A weight-adjusted version of $\hat{F}_i(t)$ can be obtained after replacing w^*_{ij} with w_{ij} in the equation 2.6. One feasible way to adjust the weights is post-stratification. Let U_h (for $h = 1, 2, \ldots, H$) with size N_{+h} be another partitioning of U independent of domain membership. Further, s_{+h} be the set of units in the sample belonging to stratum h. A basic direct estimate of N_{+j} is $\hat{N}_{+h} = \sum_{i=1}^{m} \sum_{j \in s_{+h}} w_{ij}$, leading to an adjustment factor $g_{ij} = N_{+j}/N_{+j}$.

When SRSWOR is performed, the weight w_{ij} simplifies to N_i/n_i for $i = 1, 2, \ldots, m$, and equation 2.6 simplifies to equation 2.5. The survival function estimator $\hat{F}_{w_i}(t)$ is unbiased, with variance given by:

$$V[\hat{F}_i(t)] = \sum_{j \in s_i} \frac{w_{ij}(w_{ij} - 1)}{w^2_{ij}} I^2(t_j < t) \tag{2.7}$$

Under SRSWOR, the variance in equation 2.7 simplifies to:

$$V[\hat{F}_i(t)] = \frac{N_i - n_i}{N_i - 1} F_i(t) \left[1 - F_i(t)\right] / n_i \tag{2.8}$$

A sample version of this variance, given in equation 2.8, is:

$$v[\hat{F}_i(t)] = \frac{N_i - n_i}{N_i - 1} \hat{F}_i(t) \left[1 - \hat{F}_i(t)\right] / n_i \tag{2.9}$$

To derive the variance estimator in equation 2.9, the number of people at risk in the population N_i must be known, which is often challenging in practical situations. Nevertheless, an estimate $\hat{N}_i$ can be obtained using the relation:

$$\hat{N}_i = N \times \frac{n}{n_i} \tag{2.10}$$

where N is assumed to be known beforehand. In unplanned domains, obtaining a reliable estimate of the survival function is difficult using equation 2.5 when the domain-specific sample size is very low or zero in extreme cases. These unplanned domains are referred to as small areas. The synthetic method, an indirect approach, leverages known auxiliary data from related areas to enhance the efficiency of small area estimators. Despite its sophistication in improving efficiency, the synthetic method can introduce bias due to the incorporation of information from related areas. Therefore, a composite method, combining direct and synthetic methods in a weighted manner, is a superior approach for estimating parameters in small areas. Additionally, increasing the sample size is another strategy to produce reliable estimates in small areas.

2.3 Two-Occasion SAE of Survival Function

In small area estimation (SAE), the objective is to improve the accuracy of survival function estimates by utilizing data collected over two distinct occasions. This approach is particularly beneficial in situations where data from a single occasion may be inadequate or unreliable due to small sample sizes. By combining data from two different time periods, we can leverage the additional information to enhance the precision and reliability of the estimates.

Let us denote the sample selected from the population U for the ith area on the kth occasion as $s_i^{(k)}$ for $k = 1, 2$. Furthermore, let $s_{im}^{(k)}$ represent the set of matched samples in $s_i^{(k)}$ such that $s_{im}^{(1)} = s_{im}^{(2)}$. Additionally, let $s_{iu}^{(k)}$ denote the unmatched part of the sample selected on the kth occasion, such that $s_{iu}^{(k)} = s_i^{(k)} - s_{im}^{(k)}$. These notations help in clearly distinguishing between the matched and unmatched portions of the samples from both occasions.

2.3.1 Strategy 1 (S1):

The first strategy, referred to as $S1$, involves pooling data from both occasions to create a unified sample. This method incorporates the unmatched part of the first occasion's survey along with the complete sample from the second occasion, ensuring that there is no overlap of data points. The combined sample for the ith area, denoted as $s_i^{(c)}$, is formed as follows:

$$s_i^{(c)} = s_i^{(2)} \cup s_{iu}^{(1)} \tag{2.11}$$

where $s_i^{(2)} \cap s_{iu}^{(1)} = \emptyset$. This combination ensures that the pooled sample $s_i^{(c)}$ does not include any repeated observations from both occasions, thus maintaining the integrity of the data.

The direct estimator for the survival function using this pooled method is given by:

$$\hat{F}_{S1,i}(t) = \frac{\sum_{j \in s_i^{(c)}} w_{ij}^{(c)} I(t_j < t)}{\sum_{j \in s_i^{(c)}} w_{ij}^{(c)}} \tag{2.12}$$

In this equation, $\hat{F}_{S1,i}(t)$ represents the estimated survival function for the ith area at time t. The combined sample $s_i^{(c)}$ is utilized, and $w_{ij}^{(c)}$ are the expansion weights for the combined sample units. The indicator function $I(t_j < t)$ is used to indicate whether the event time t_j exceeds t.

For simple random sampling, the expansion weights $w_{ij}^{(c)}$ simplify to:

$$w_{ij}^{(c)} = \frac{N_i}{n_i^{(c)}} \tag{2.13}$$

where N_i is the total population size of the ith area, and $n_i^{(c)}$ is the size of the combined sample for the ith area. By pooling data from both occasions, the $S1$ strategy increases the effective sample size, thereby enhancing the precision and reliability of the survival function estimates. This approach provides a more robust estimation for small areas, making it a valuable technique in small area estimation.

where $s_i^{(2)} \cap s_{iu}^{(1)} = \emptyset$. This combination ensures that the pooled sample $s_i^{(c)}$ does not include any repeated observations from both occasions, thus maintaining the integrity of the data.

Assuming that the population mean of the study character is stable over time, the survival function estimator $\hat{F}_{S1,i}(t)$ is unbiased. The variance of this estimator can be expressed as follows:

$$V[\hat{F}_{S1,i}(t)] = \sum_{j \in s_i^{(c)}} \frac{w_{ij}(w_{ij} - 1)}{w_{ij}^2} I^2(t_j < t) \tag{2.14}$$

In the context of simple random sampling without replacement (SRSWOR), the variance of $\hat{F}_i(t)$ simplifies to:

$$V\left[\hat{F}_i^{S1}(t)\right] = \frac{N_i - n_i^{(c)}}{N_i - 1} \frac{F_i(t)\left[1 - F_i(t)\right]}{n_i} \tag{2.15}$$

Here, the finite population correction (fpc) factor uses $n_i^{(c)}$ instead of n_i, leading to a reduced variance since $n_i^{(c)} \geq n_i$. The equality holds when there are no observations for the ith area on the first occasion.

A sample estimate of the variance given in equation 2.15 is obtained using the following expression:

$$v\left[\hat{F}_i^{S1}(t)\right] = \frac{N_i - n_i^{(c)}}{N_i - 1} \frac{\hat{F}_i^{S1}(t)\left[1 - \hat{F}_i^{S1}(t)\right]}{n_i^{(c)}} \tag{2.16}$$

When the number of people at risk in the population N_i is unknown, an estimate $\hat{N}_i$ can be obtained using the relation:

$$\hat{N}_i = N \times \frac{n}{n_i^{(c)}} \tag{2.17}$$

where $n = n_1 + n_2$ and N are fixed in advance. This approach allows for the estimation of the population size based on the combined sample size from both occasions, thus improving the reliability of the survival function estimates for small areas.

2.3.1.1 Synthetic Method under S2

The weighted combination of direct estimates increases efficiency by incorporating information related to the ith area from two surveys. However, when the sample sizes in these surveys, especially the current one, are insufficient, producing reliable estimates for areas with low sample sizes becomes challenging. To address this issue, strength can be borrowed from related areas to generate estimates for those with smaller sample sizes. One approach is to obtain an unbiased estimator for a relatively broader area (B) and use it to derive estimates for the smaller areas of interest. This approach assumes that the small areas share the same characteristics as the larger area, and these estimates are classified as synthetic estimates.

Assuming an implicit model that the survival function of the ith area is equal to the overall survival function, we have:

$$F_i(t) = F_B(t) \tag{2.18}$$

where $F_i(t)$ is the true survival function for the ith area, and $F_B(t)$ is the survival function for the broader area B.

Under this assumption, the survival function estimator for the ith area on the second occasion is given by:

$$\hat{F}_i^{(2)\text{syn}}(t) = \frac{\sum_{j \in s_B} w_{Bj} I(t_j < t)}{\sum_{j \in s_B} w_{Bj}} \tag{2.19}$$

Here, s_B is the set of units selected in the sample from the broader area B, such that $s_B \subseteq s$ and $s_i \subseteq s_B$. Further, $w_{Bj} = \pi_{Bj}^{-1}$ is the inclusion probability of the jth unit in the broader area B, and $I(t_j < t)$ is an indicator function that equals 1 if the event time t_j exceeds t, and 0 otherwise.

Under simple random sampling, the synthetic estimator for the survival function of the population is:

$$\hat{F}_i^{(2)\text{syn}}(t) = \hat{F}_B(t) = \frac{1}{n_B^{(2)}} \sum_{j \in s_B} I(t_j < t) \tag{2.20}$$

where $n_B^{(2)}$ is the sample size of the broader area B, such that $n_i^{(2)} \leq n_B^{(2)} \leq n^{(2)}$.

A synthetic estimator using the implicit model above under $S2$ is obtained by replacing $\hat{F}_i^{(2)}(t)$ with $\hat{F}_i^{(2)\text{syn}}(t)$ in the combined estimator:

$$\hat{F}_i^{S2\text{-syn}}(t) = \alpha_i \hat{F}_i^{(1)}(t) + (1 - \alpha_i)\hat{F}_i^{(2)\text{syn}}(t) \tag{2.21}$$

where $\hat{F}_i^{S2\text{-syn}}(t)$ is the synthetic estimator for the survival function using the weighted combination under $S2$, $\hat{F}_i^{(1)}(t)$ is the direct estimator for the survival function on the first occasion, and α_i is the weight assigned to the estimate from the first occasion.

The bias and mean squared error (MSE) of $\hat{F}_i^{S2\text{-syn}}(t)$ are given by:

$$\text{Bias}\left[\hat{F}_i^{S2\text{-syn}}(t)\right] = \alpha_i E\left[\hat{F}_i^{(1)}(t) - F_i(t)\right] + (1 - \alpha_i) E\left[\hat{F}_i^{(2)\text{syn}}(t) - F_i(t)\right] \tag{2.22}$$

where $E[\cdot]$ denotes the expected value operator.

$$\text{MSE}\left[\hat{F}_i^{S2\text{-syn}}(t)\right] = \alpha_i^2 E\left[\hat{F}_i^{(1)}(t) - F_i(t)\right]^2 + (1-\alpha_i)^2 E\left[\hat{F}_i^{(2)\text{syn}}(t) - F_i(t)\right]^2 \quad (2.23)$$

The first term in the bias reduces to zero when the survival function is stable over k. The second term goes to zero when the survival function of the ith area coincides with that of the broader area B. However, in practice, it is difficult to maintain the relationship given by the model. Assuming the survival function is stable over k, the MSE can be expressed as:

$$\text{MSE}\left[\hat{F}_i^{S2\text{-syn}}(t)\right] = \alpha_i^2 V\left[\hat{F}_i^{(1)}(t)\right] + (1-\alpha_i)^2 \text{MSE}\left[\hat{F}_i^{(2)\text{syn}}(t)\right] \quad (2.24)$$

where

$$V\left[\hat{F}_i^{(1)}(t)\right] = \frac{N-n_1}{N-1}\frac{\hat{F}_i^{(1)}(t)\left[1-\hat{F}_i^{(1)}(t)\right]}{n_1} \quad (2.25)$$

and

$$\text{MSE}\left[\hat{F}_i^{(2)\text{syn}}(t)\right] = E\left[\hat{F}_i^{(2)\text{syn}}(t) - \hat{F}_i^{(2)}(t)\right]^2 - V\left[\hat{F}_i^{(2)}(t)\right] \quad (2.26)$$

with sample estimates

$$v\left[\hat{F}_i^{(1)}(t)\right] = \frac{N-n_1}{N-1}\frac{F_i(t)\left[1-F_i(t)\right]}{n_1} \quad (2.27)$$

and

$$\text{mse}\left[\hat{F}_i^{(2)\text{syn}}(t)\right] = \left(\hat{F}_i^{(2)\text{syn}}(t) - \hat{F}_i^{(2)}(t)\right)^2 - v\left[\hat{F}_i^{(2)}(t)\right] \quad (2.28)$$

2.3.1.2 Composite Method under S2

The synthetic method addresses the problem of small sample sizes at the area level on the current occasion, enhancing efficiency while making some compromise on bias. However, the direct method provides unbiased estimates in small areas, but with an inflated coefficient of variation (CV). To balance efficiency and bias, Royall (1973) and

Schaible (1977) suggested a weighted combination of the direct and synthetic estimators to obtain a composite estimator from the second occasion sample. We propose using a composite estimator for the sample obtained at the current occasion:

$$\hat{F}_i^{(2)\text{com}}(t) = \lambda_i \hat{F}_i^{(2)}(t) + (1 - \lambda_i)\hat{F}_i^{(2)\text{syn}}(t) \tag{2.29}$$

where $\hat{F}_i^{(2)\text{com}}(t)$ is the composite estimator for the survival function, $\hat{F}_i^{(2)}(t)$ is the direct estimator for the survival function on the second occasion, $\hat{F}_i^{(2)\text{syn}}(t)$ is the synthetic estimator, and λ_i is the weight assigned to the direct estimator.

The resulting estimator can be written as:

$$\hat{F}_i^{S2\text{-com}}(t) = \alpha_i \hat{F}_i^{(1)}(t) + (1 - \alpha_i)\hat{F}_i^{(2)\text{com}}(t) \tag{2.30}$$

where α_i is the weight assigned to the estimate from the first occasion.

The bias of the composite estimator is expressed as:

$$\text{Bias}[\hat{F}_i^{S2\text{-com}}(t)] = \alpha_i E[\hat{F}_i^{(1)}(t) - F_i(t)] + (1 - \alpha_i)(1 - \lambda_i)E[\hat{F}_i^{(2)\text{syn}}(t) - F_i(t)] \tag{2.31}$$

Assuming that the survival function is stable over k, the bias of the composite estimator can be simplified as:

$$\text{Bias}[\hat{F}_i^{S2\text{-com}}(t)] = (1 - \alpha_i)(1 - \lambda_i)E[\hat{F}_i^{(2)\text{syn}}(t) - F_i(t)] = (1 - \lambda_i)\text{Bias}[\hat{F}_i^{(2)\text{syn}}(t)] \tag{2.32}$$

It is evident from this expression that the bias of the composite estimator is always smaller than that of the synthetic estimator since $\lambda_i < 1$.

The Mean Squared Error (MSE) of the composite estimator is given by:

$$\text{MSE}[\hat{F}_i^{S2\text{-com}}(t)] = \alpha_i^2 E[\hat{F}_i^{(1)}(t) - F_i(t)]^2 + (1 - \alpha_i)^2 \text{MSE}[\hat{F}_i^{(2)\text{com}}(t)] \tag{2.33}$$

Here, $\hat{F}_i^{(1)}(t)$ is the direct estimator on the first occasion, $\hat{F}_i^{(2)}(t)$ is the direct estimator on the second occasion, and $\hat{F}_i^{(2)\text{syn}}(t)$ is the synthetic estimator on the second occasion. The weights λ_i and α_i are assigned to balance the efficiency and bias of the estimators.

2.3.2 Strategy 3 (S3)

The synthetic estimators of survival function utilize the sample information from the current survey to estimate the parameters related to the auxiliary variable. However, estimates for the parameters of the auxiliary variable can also be obtained using the S2 strategy. A regression-type estimator for the survival function in the ith area can be formulated as follows:

$$\hat{F}_i^{S3\text{-reg}}(t) = \hat{F}_i^{S2}(t) + \beta_i \left[F_i(x) - \hat{F}_i^{S2}(x) \right] \tag{2.34}$$

In this equation, $\hat{F}_i^{S3\text{-reg}}(t)$ is the regression-type estimator for the survival function under S3. $\hat{F}_i^{S2}(t)$ is the survival function estimator under S2. The coefficient β_i can be derived from the sample on the current occasion. $F_i(x)$ and $\hat{F}_i^{S2}(x)$ represent the true survival function and the S2 estimator at a different time point x, respectively.

Assuming a stable model, the variance of the regression-type estimator under S3, $\hat{F}_i^{S3\text{-reg}}(t)$, is given by:

$$V\left[\hat{F}_i^{S3\text{-reg}}(t)\right] = \left[1 - \rho_{txi}^{S2}\right] V\left[\hat{F}_i^{S2}(t)\right] \tag{2.35}$$

Here, $V\left[\hat{F}_i^{S3\text{-reg}}(t)\right]$ denotes the variance of the regression-type estimator under S3. ρ_{txi}^{S2} is the correlation coefficient between the event time variable and the auxiliary variable under S2. $V\left[\hat{F}_i^{S2}(t)\right]$ is the variance of the survival function estimator under S2. It is evident from this variance expression that the estimator $\hat{F}_i^{S3\text{-reg}}(t)$ is at least as efficient as $\hat{F}_i^{S2}(t)$.

Another approach within S3 is to use a ratio-type estimator. The ratio-type estimator for the survival function in the ith area can be formulated as:

$$\hat{F}_i^{S3\text{-r}}(t) = \hat{F}_i^{S2}(t) \frac{F_i(x)}{\hat{F}_i^{S2}(x)} \tag{2.36}$$

In this formula, $\hat{F}_i^{S3\text{-r}}(t)$ is the ratio-type estimator for the survival function under S3. $\hat{F}_i^{S2}(t)$ and $\hat{F}_i^{S2}(x)$ are the S2 estimators at times t and x, respectively. $F_i(x)$ is the true survival function for the ith area at time x.

The bias and (MSE) of the ratio-type estimator under S3 are given by:

$$\text{Bias}\left[\hat{F}_i^{S3\text{-r}}(t)\right] = \left[\frac{V\left[\hat{F}_i^{S2}(x)\right]}{F_i(x)^2} - \frac{\text{Cov}\left[\hat{F}_i^{S2}(t), \hat{F}_i^{S2}(x)\right]}{F_i(t)F_i(x)} \right] \tag{2.37}$$

Here, Bias $\left[\hat{F}_i^{S3\text{-r}}(t)\right]$ denotes the bias of the ratio-type estimator under S3. $V\left[\hat{F}_i^{S2}(x)\right]$ represents the variance of the survival function estimator under S2 at time x. Cov $\left[\hat{F}_i^{S2}(t), \hat{F}_i^{S2}(x)\right]$ is the covariance between the survival function estimators under S2 at times t and x.

$$\text{MSE}\left[\hat{F}_i^{S3\text{-r}}(t)\right] = \frac{V\left[\hat{F}_i^{S2}(t)\right]}{F_i(t)^2} + \frac{V\left[\hat{F}_i^{S2}(x)\right]}{F_i(x)^2} - 2\frac{\text{Cov}\left[\hat{F}_i^{S2}(t), \hat{F}_i^{S2}(x)\right]}{F_i(t)F_i(x)} \tag{2.38}$$

In this formula, MSE $\left[\hat{F}_i^{S3\text{-r}}(t)\right]$ represents the Mean Squared Error of the ratio-type estimator under S3. $V\left[\hat{F}_i^{S2}(t)\right]$ and $V\left[\hat{F}_i^{S2}(x)\right]$ are the variances of the survival function estimators under S2 at times t and x, respectively. The term Cov $\left[\hat{F}_i^{S2}(t), \hat{F}_i^{S2}(x)\right]$ denotes the covariance between the survival function estimators under S2 at times t and x.

This strategy illustrates the effectiveness of combining information from auxiliary variables to improve the accuracy and efficiency of survival function estimators.

2.3.3 Strategy 4 (S4)

As discussed in the previous subsection, estimates on the parameters of the auxiliary variable can be obtained using S1 instead of S2. These can be used to construct the ratio and regression estimators with direct estimators under S1. A regression-type estimator for the survival function in the ith area under S4 can be defined as:

$$\hat{F}_i^{S4\text{-reg}}(t) = \hat{F}_i^{S1}(t) + \beta_i\left[F_i(x) - \hat{F}_i^{S1}(x)\right] \tag{2.39}$$

In this formula, $\hat{F}_i^{S4\text{-reg}}(t)$ is the regression-type estimator for the survival function under S4. $\hat{F}_i^{S1}(t)$ and $\hat{F}_i^{S1}(x)$ are obtained by replacing t by x in the S1 estimator. The coefficient β_i can be obtained from the combined sample as $\hat{\beta}_i$, but in the synthetic method, a known relationship between the two variables for the whole population β is used instead of β_i.

Here, the two estimators $\hat{F}_i^{S1}(t)$ and $\hat{F}_i^{S1}(x)$ can be obtained through the Horvitz-Thompson method or post-stratified method to gain more stability.

Assuming a stable model, the variance of the regression-type estimator under S4, $\hat{F}_i^{S4\text{-reg}}(t)$, is given by:

$$V\left[\hat{F}_i^{S4\text{-reg}}(t)\right] = \left[1 - \rho_{txi}^{S1}\right] V\left[\hat{F}_i^{S1}(t)\right] \tag{2.40}$$

This variance derivation is similar to that of $\hat{F}_i^{S3\text{-reg}}(t)$. From this, it is evident that the estimator $\hat{F}_i^{S4\text{-reg}}(t)$ is at least as efficient as $\hat{F}_i^{S1}(t)$.

Another approach within S4 is to suggest a ratio-type estimator. The ratio-type estimator for the survival function in the ith area can be obtained as:

$$\hat{F}_i^{S4\text{-r}}(t) = \hat{F}_i^{S1}(t)\frac{F_i(x)}{\hat{F}_i^{S1}(x)} \tag{2.41}$$

The bias and Mean Squared Error (MSE) of the ratio-type estimator under S4 are given by:

$$\text{Bias}\left[\hat{F}_i^{S4\text{-r}}(t)\right] = \left[\frac{V\left[\hat{F}_i^{S1}(x)\right]}{F_i(x)^2} - \frac{\text{Cov}\left[\hat{F}_i^{S1}(t), \hat{F}_i^{S1}(x)\right]}{F_i(t)F_i(x)}\right] \tag{2.42}$$

In this formula, $\text{Bias}\left[\hat{F}_i^{S4\text{-r}}(t)\right]$ represents the bias of the ratio-type estimator under S4. $V\left[\hat{F}_i^{S1}(x)\right]$ is the variance of the survival function estimator under S1 at time x. $\text{Cov}\left[\hat{F}_i^{S1}(t), \hat{F}_i^{S1}(x)\right]$ is the covariance between the survival function estimators under S1 at times t and x.

$$\text{MSE}\left[\hat{F}_i^{S4\text{-r}}(t)\right] = \frac{V\left[\hat{F}_i^{S1}(t)\right]}{F_i(t)^2} + \frac{V\left[\hat{F}_i^{S1}(x)\right]}{F_i(x)^2} - 2\frac{\text{Cov}\left[\hat{F}_i^{S1}(t), \hat{F}_i^{S1}(x)\right]}{F_i(t)F_i(x)} \tag{2.43}$$

In this formula, $\text{MSE}\left[\hat{F}_i^{S4\text{-r}}(t)\right]$ represents the Mean Squared Error of the ratio-type estimator under S4. $V\left[\hat{F}_i^{S1}(t)\right]$ and $V\left[\hat{F}_i^{S1}(x)\right]$ are the variances of the survival function estimators under S1 at times t and x, respectively. The term $\text{Cov}\left[\hat{F}_i^{S1}(t), \hat{F}_i^{S1}(x)\right]$ denotes the covariance between the survival function estimators under S1 at times t and x.

Comparing the variance of the direct estimator under S1 with the MSE of the ratio estimator under S4, we have:

$$V\left[\hat{F}_i^{S1}(t)\right] - \text{MSE}\left[\hat{F}_i^{S4\text{-r}}(t)\right] = R_i^2 V\left[\hat{F}_i^{S1}(x)\right] - 2R_i\text{Cov}\left[\hat{F}_i^{S1}(t), \hat{F}_i^{S1}(x)\right] \tag{2.44}$$

Here, $R_i = \frac{F_i(t)}{F_i(x)}$. The right side of this equation is positive when $\rho_{txi}^{S1} > \frac{1}{2}R_i\frac{V\left[\hat{F}_i^{S1}(x)\right]}{\text{Cov}\left[\hat{F}_i^{S1}(t), \hat{F}_i^{S1}(x)\right]}$, indicating the conditional superiority of the synthetic ratio estimator under S4 over the direct estimator under S1.

Chapter 3

Results and Discussions

3.1 Efficiency Comparison of Estimators

The child birth data from the PDHS 2017–18 and PDHS 2019 Special surveys is being used for evaluating the effectiveness of the suggested small area estimate (SAE) techniques. The study populations include 20,227 children from the PDHS 2017−18 survey and 20,895 children from the PDHS Special 2019 survey. Table 3.1 provides a detailed explanation of the variables that will be examined.

Table 3.1. Some important variables used in the study

DHS code	Variable name	Description	Usage
B7	Age at death	Age of the child at the time of death from Survey-1	Response
B8	Current age	Child's age at the time of data collection from Survey-1	Response
Q220C	Age at death	Age of child at the time of death from Survey-2	Response
Q217	Current age	Age of Child at the time of data collection from Survey-2	Response
V214	Pregnancy Duration	Duration of pregnancy for Mother from Survey-1	Auxiliary variable
Q220AC	Pregnancy Duration	Duration of mother's pregnancy from Survey-2	Auxiliary variable
B4	Sex of a Child	Gender of a Child from Survey-1	Stratification
Q213	Sex of a Child	Gender of a Child from Survey-2	Stratification

In this study, an indicator function has also been employed

$$I = \begin{cases} 1 & , \text{Child is died before 12 months} \\ 0 & , \text{Child is not died before 12 months} \end{cases}$$

Age at Death(Months)	< 12	> 12	NA	NA
Current Age (Months)	NA	NA	< 12	> 12
Indicator function	1	0	censord	0

Table 3.2. Survival Indicator for Infant Mortality

The indicator function is derived from the variables "Age at death" and "Current age," as shown in Table 3.2. Children are considered to have died before 12 months if their "age at death" is less than 12 months and their "current age" is not provided. Those with a "current age" not given and an "age at death" over 12 months are considered to have survived. Individuals with an unknown "age at death" and a "current age" under 12 months are excluded from the study. Those with a "current age" over 12 months and an unknown "age at death" are considered to have survived.

3.1.1 Bootstrap Comparison of Strategies

A bootstrapped study was carried out by treating the two surveys as the study population across two consecutive occasions, involving 20,227 children in the PDHS 2017-18 survey and 20,895 children in the PDHS Special 2019 survey, respectively. The bootstrapping process involved the following steps:

1. **Step 1:** A random sample of size n_1 was selected from the first survey (PDHS 2017–18) and n_2 from the second survey (PDHS Special 2019), without replacement. Matching cases from the first sample were excluded.

2. **Step 2a:** Using the separate samples selected in Step 1, estimates were obtained under Strategies 2 and 3, with appropriate choices of the weighting parameter $\lambda_{li}^{(t)}$ (where $t = 1, 2$ and $l = 1, 2, 3$).

3. **Step 2b:** The two samples from Step 1 were pooled to calculate the area-level mean estimators under Strategies 1 and 4.

4. **Step 3:** Steps 1-2 were repeated Q times to determine the Mean Squared Error (MSE) and the Relative Efficiency (RE) of the mean estimators. A larger choice of Q generally resulted in more stable outcomes. Q refers to the number of times the bootstrapping process was repeated. More iterations of Q typically lead to more reliable estimates because they reduce variability in the results.

3.2 Parameter Estimations in Districts

The DHS reports provide detailed estimates of neonatal and infant mortality rates across various regions in the country, including the four major provinces: Punjab, Sindh, Khyber Pakhtunkhwa (KPK), and Balochistan, as well as important areas like the Federally Administered Tribal Areas (FATA), Islamabad Capital Territory (ICT), Gilgit Baltistan, and Azad Jammu and Kashmir (AJK). To give a clearer picture, the reports also break down these mortality rates at the district level within these provinces. This district-wise analysis helps to highlight specific areas that might need more attention and targeted interventions. By comparing mortality rates across different regions and districts, these reports offer valuable insights that can help shape public health strategies, inform policy decisions, and ultimately improve health outcomes for children across the country. This more localized approach allows for a better understanding of the unique challenges faced by different districts, helping to identify areas where resources and efforts should be concentrated. The reports provide a roadmap for health officials and policymakers to design targeted interventions that address the specific needs of each district.

In this article, we suggested ways to create district-level estimates of these rates by using PDHS 17-18 and PDHS 19 and also combining samples from these surveys. To test these methods, a bootstrap study was conducted with repeated samples of $n(1) = 5000$ from the first survey and $n(2) = 5000$ from the second survey to estimate health indicators at the district level. An R package called "tosae" has been developed to produce these expected sample sizes and is available on GitHub. Table 3.3 shows expected sample sizes in different districts for the current survey ESS2 and the combined survey ESSC, along with CVs of the proposed estimators. Due to unavoidable circumstances, there are no observations available for the districts of Sujawal, Kohistan, Barkhan, Bolan/Kachhi, Gawadar, Nasirabad/Tamboo, Panjgur, Pishin, and Sohbat Pur in both the PDHS 2017-18 and PDHS 2019 (Special) surveys. In this study, we proposed four distinct strategies for Two-Occasion Small Area Estimation (SAE) of the Survival Function. The first strategy, (S1), was based exclusively on the Direct method. For the second strategy, (S2), we employed a more diverse set of five methods: Direct method, Regression, Ratio, Composite Regression, and Composite Ratio. The third strategy, (S3), involved the use of four methods: Regression, Ratio, Composite Regression, and Composite Ratio. Similarly, (S4) also incorporated these four methods: Regression, Ratio, Composite Regression, and Composite Ratio. Upon comparing the Coefficients of Variation (CVs) across the different strategies illustrated in Figures 3.1 and 3.2, it becomes evident that the Composite Regression method under Strategy S2 consistently provides more stable and reliable estimates at the area level, particularly in terms of CV and bias, when compared to the outcomes from Strategies S1, S3, and S4.

Table 3.3. District-wise expected Sample Sizes

Districts	**ESS2**	**ESSC**	**Districts**	**ESS2**	**ESSC**
Attock	17.6	46.9	Kohat	21.6	50.3
Bahawalnagar	37.5	68.7	Kohistan	0.0	0.0
Bahawalpur	56.6	104.2	Lakki Marwat	14.2	25.3
Bhakkar	26.3	56.3	Lower Dir	39.6	57.0
Chakwal	16.6	43.9	Malakand	13.7	34.7
Chiniot	10.4	41.7	Mansehra	18.0	40.7
Dera Ghazi Khan	24.3	44.7	Mardan	32.0	91.7
Faisalabad	70.1	138.9	Nowshera	37.3	86.2
Gujranwala	68.8	111.2	Peshawar	105.2	299.5
Gujrat	21.1	38.0	Shangla	6.4	12.6
Hafizabad	12.7	46.2	Swabi	38.2	83.8
Jhang	29.3	54.3	Swat	37.1	104.4
Jhelum	8.9	25.7	Tank	21.4	38.9
Kasur	35.3	74.5	Tor Ghar	5.7	14.5
Khanewal	28.2	63.0	Upper Dir	25.3	47.0
Khushab	14.6	33.1	Awaran	10.5	27.6
Lahore	143.0	233.8	Barkhan	25.3	25.3
Layyah	17.2	43.7	Bolan/Kachhi	14.7	14.7
Lodhran	17.5	43.7	Chagai	20.4	25.9
Mandi Bahauddin	17.4	26.9	Dera Bugti	2.4	33.7
Mianwali	18.6	23.3	Gawadar	5.9	5.9
Multan	51.3	115.3	Harnai	14.9	37.7
Muzaffargarh	30.5	56.7	Jaffarabad	18.8	26.0
Nankana Sahib	12.3	16.6	Jhal Magsi	10.7	19.1
Narowal	21.9	41.9	Kalat	24.5	56.5
Okara	31.5	65.2	Kech/Turbat	13.4	61.0
Pakpattan	20.0	39.1	Kharan	12.5	28.0
Rahim Yar Khan	62.0	122.7	Khuzdar	40.1	119.9
Rajanpur	20.0	31.5	Killa Abdullah	18.4	39.9
Rawalpindi	51.7	95.9	Killa Saifullah	19.9	28.6
Sahiwal	18.5	49.9	Kohlu	18.7	31.1
Sargodha	35.6	73.8	Lasbela	19.3	69.0
Sheikhupura	36.0	67.6	Loralai	18.9	29.2
Sialkot	41.2	70.4	Mastung	8.5	28.7
Toba Tek Singh	20.7	47.1	Musakhel	12.2	34.6
Vehari	23.8	50.2	Nasirabad/Tamboo	27.9	27.9
Badin	34.2	66.6	Nushki	15.0	28.5

Districts	ESS2	ESSC	Districts	ESS2	ESSC
Dadu	27.3	57.6	Panjgur	8.8	8.8
Ghotki	45.3	81.2	Pishin	41.1	41.1
Hyderabad	42.3	71.6	Quetta	104.0	127.9
Jacobabad	25.2	58.5	Sherani	8.5	105.0
Jamshoro	21.0	39.1	Sibi	25.7	29.7
Karachi Central	44.9	68.7	Washuk	2.2	9.5
Karachi East	42.9	117.4	Zhob	13.0	21.0
Karachi South	25.2	51.3	Ziarat	7.5	24.4
Karachi West	57.7	85.0	Sohbat Pur	23.9	23.9
Kashmore	33.4	106.9	Astore	33.3	49.2
Khairpur	59.1	74.6	Baltistan	78.1	96.8
Larkana	41.4	92.3	Diamir	60.7	112.0
Matiari	28.7	64.5	Ghanche	34.3	79.3
Mirpur Khas	45.6	97.8	Ghizer	47.7	148.2
Naushahro Feroze	38.0	47.8	Gilgit	94.2	120.1
Sanghar	43.9	71.4	Nagar	16.9	31.5
Shahdad Kot	36.3	56.8	Kharmang	22.8	43.7
Nawabshah	32.1	55.6	Hunza	21.6	45.9
Shikarpur	19.6	56.3	Shigar	33.3	87.2
Sukkur	30.3	41.4	Islamabad	249.1	567.1
Tando Alla Yar	30.1	73.4	Bajour	23.8	54.3
Tando Muhammad Khan	24.7	49.9	Khyber	110.0	140.1
Tharparkar	42.9	62.8	Kurram	66.7	171.4
Thatta	34.3	40.5	Mohmand	15.7	68.8
Umer Kot	21.1	42.6	North Waziristan	43.6	59.4
Korangi	26.2	56.3	Orakzai	14.0	61.8
Malair	35.4	48.5	South Waziristan	28.3	38.4
Sujawal	38.1	38.1	Bagh	52.1	105.4
Abbottabad	35.3	59.1	Bhimber	40.1	76.8
Bannu	35.5	54.5	Hattian Bala	39.2	77.1
Batagram	12.4	24.4	Haveli	19.7	43.2
Buner	11.4	26.7	Kotli	83.3	148.6
Charsadda	30.8	72.0	Mirpur	130.8	225.4
Chitral	24.4	42.1	Muzaffarabad	95.7	197.1
D. I. Khan	30.7	81.3	Neelum	8.2	38.5
Hangu	19.9	30.3	Poonch	51.1	132.2
Haripur	23.3	49.8	Sudhonti	29.1	57.9
Karak	14.2	21.8			

For numerical comparison, the "District" variable is used as the domain variable, and "Age of a child at death" is used as the response variable from the PDHS Children's re-code file. The results, including the coefficient of variation (CV), are presented in Table 3.4 for the districts of the four provinces (Punjab, Sindh, KPK, and Balochistan) of Pakistan, as well as FATA (a federally administered region). The weight $\lambda_{li}^{(1)} = 1 - \lambda_{li}^{(2)}$ is set at 0.3, and the weighting parameter Ψ is set at 0.4. After carefully evaluating each approach, we identified the most effective method from each strategy by analyzing the Coefficient of Variation (CV) and Bias. Upon further comparison, we determined that the Composite Regression method under Strategy 2 consistently provided the most reliable estimates in terms of both CV and Bias. Overall, it was observed that across all the strategies, the Composite Regression estimator under Strategy 2 consistently delivered the most precise estimates, particularly in terms of the Coefficient of Variation (CV) and Bias. Tables 3.5 and 3.6 provide detailed information on the estimated proportions, standard errors, and 95 percent confidence intervals for early-age mortality rates across different districts. By using the "sf" package in R, we created geographical maps to visually represent the estimated proportion of infant deaths across various districts in Pakistan. These maps illustrate the impact of different strategies, as shown in the Figure 3.3. The results highlight the most stable and reliable estimates, focusing on the best-performing strategies. Specifically, the maps showcase the Composite Regression Estimators under Strategy S2, providing a clear comparison of infant mortality rates across the regions.

Table 3.4. Summary of CV for different infant mortality estimators

Strategy	Estimate	Minimum	Q1	Mean	Q3	Maximum
S1	Direct	10.8661	25.6954	**40.5404**	51.1239	135.7267
S2	Direct	11.4220	26.2722	38.1577	43.4877	114.4000
	Regression	11.4054	26.3748	38.1558	43.4861	114.6846
	Ratio	11.4946	26.5564	38.3834	43.1031	114.3864
	Composite Regression	11.5324	26.3841	**38.1345**	43.4851	114.4218
	Composite Ratio	11.4414	26.2245	38.2091	43.2450	114.4154
S3	Regression	10.9818	27.5963	43.0604	54.4788	114.7299
	Ratio	15.3614	28.7849	42.0481	50.6925	119.4373
	Composite Regression	10.9035	27.5134	42.9655	54.5723	114.6975
	Composite Ratio	11.7060	28.0135	**41.6670**	51.9620	116.3753
S4	Regression	11.1969	25.6645	40.7095	51.1178	138.9332
	Ratio	15.9942	27.2018	39.9846	47.6601	115.0726
	Composite Regression	11.0823	25.8780	40.4024	50.9082	136.0466
	Composite Ratio	12.8193	25.9346	**39.2993**	47.3781	115.1354

Table 3.5. District-wise estimated proportions, standard errors and confidence intervals for infant mortality

Districts	**Estimates**	**STD**	**LCI**	**UCI**
Attock	0.13108	0.05956	0.01409	0.24807
Bahawalnagar	0.25232	0.05265	0.14890	0.35574
Bahawalpur	0.25888	0.04036	0.17962	0.33815
Bhakkar	0.11992	0.04548	0.03058	0.20925
Chakwal	0.11060	0.05328	0.00595	0.21524
Chiniot	0.17650	0.08723	0.00517	0.34783
Dera Ghazi Khan	0.29069	0.06930	0.15456	0.42681
Faisalabad	0.19965	0.03458	0.13174	0.26756
Gujranwala	0.13199	0.03029	0.07249	0.19149
Gujrat	0.24130	0.06469	0.11423	0.36837
Hafizabad	0.21321	0.07843	0.05916	0.36726
Jhang	0.21083	0.05362	0.10550	0.31615
Jhelum	0.14145	0.09984	0.00000	0.33755
Kasur	0.21847	0.04948	0.12128	0.31565
Khanewal	0.23160	0.05334	0.12683	0.33637
Khushab	0.25520	0.08921	0.07998	0.43043
Lahore	0.14106	0.02035	0.10109	0.18102
Layyah	0.25767	0.07199	0.11628	0.39907
Lodhran	0.23037	0.06759	0.09762	0.36313
Mandi Bahauddin	0.21304	0.08593	0.04427	0.38181
Mianwali	0.13498	0.07940	0.00000	0.29092
Multan	0.24883	0.05339	0.14397	0.35370
Muzaffargarh	0.22156	0.05728	0.10905	0.33407
Nankana Sahib	0.12470	0.10238	0.00000	0.32579
Narowal	0.16998	0.05736	0.05731	0.28265
Okara	0.23273	0.06552	0.10404	0.36143
Pakpattan	0.27414	0.07280	0.13115	0.41713
Rahim Yar Khan	0.20510	0.04417	0.11834	0.29186
Rajanpur	0.23948	0.08284	0.07677	0.40218
Rawalpindi	0.11712	0.03190	0.05447	0.17977
Sahiwal	0.18980	0.06369	0.06471	0.31490
Sargodha	0.20640	0.04673	0.11462	0.29817
Sheikhupura	0.10791	0.04331	0.02286	0.19297
Sialkot	0.14249	0.04417	0.05574	0.22924
Toba Tek Singh	0.14601	0.05340	0.04113	0.25089
Vehari	0.24225	0.06332	0.11787	0.36663

Districts	Estimates	STD	LCI	UCI
Badin	0.19423	0.05137	0.09333	0.29512
Dadu	0.14717	0.06669	0.01618	0.27815
Ghotki	0.27251	0.04715	0.17991	0.36511
Hyderabad	0.15037	0.04058	0.07066	0.23008
Jacobabad	0.20459	0.06106	0.08467	0.32451
Jamshoro	0.18533	0.05970	0.06806	0.30260
Karachi Central	0.08042	0.04307	0.00000	0.16501
Karachi East	0.11792	0.04018	0.03900	0.19684
Karachi South	0.11905	0.05294	0.01507	0.22302
Karachi West	0.09741	0.02863	0.04119	0.15364
Kashmore	0.23895	0.06353	0.11417	0.36372
Khairpur	0.16051	0.04292	0.07620	0.24482
Larkana	0.11305	0.04036	0.03377	0.19232
Matiari	0.17394	0.06521	0.04585	0.30202
Mirpur Khas	0.18649	0.04064	0.10668	0.26631
Naushahro Feroze	0.22573	0.05912	0.10961	0.34186
Sanghar	0.14986	0.04051	0.07029	0.22943
Shahdad Kot	0.20450	0.04791	0.11038	0.29861
Nawabshah/Shaheed Benazir Abad	0.22530	0.07292	0.08207	0.36852
Shikarpur	0.21610	0.06163	0.09505	0.33716
Sukkur	0.18471	0.05521	0.07627	0.29316
Tando Alla Yar	0.18065	0.05866	0.06543	0.29586
Tando Muhammad Khan	0.22511	0.05779	0.11160	0.33862
Tharparkar	0.18914	0.04759	0.09567	0.28261
Thatta	0.31734	0.11452	0.09241	0.54228
Umer Kot	0.25350	0.07630	0.10363	0.40337
Korangi	0.15785	0.05272	0.05430	0.26140
Malair	0.10350	0.04066	0.02363	0.18336
Sujawal	-	-	-	-
Abbottabad	0.15211	0.04297	0.06772	0.23650
Bannu	0.13690	0.04163	0.05513	0.21868
Batagram	0.11963	0.07230	0.00000	0.26164
Buner	0.10241	0.06699	0.00000	0.23399
Charsadda	0.16587	0.04673	0.07408	0.25766
Chitral	0.07711	0.04436	0.00000	0.16424
D. I. Khan	0.27531	0.05339	0.17045	0.38017
Hangu	0.12923	0.07793	0.00000	0.28230
Haripur	0.12462	0.04682	0.03267	0.21658
Karak	0.12659	0.07717	0.00000	0.27816

Districts	Estimates	STD	LCI	UCI
Kohat	0.16999	0.05855	0.05499	0.28499
Kohistan	-	-	-	-
Lakki Marwat	0.17902	0.07267	0.03629	0.32175
Lower Dir	0.12349	0.03958	0.04575	0.20123
Malakand Protected Area	0.07611	0.04935	0.00000	0.17304
Mansehra	0.12255	0.05934	0.00600	0.23910
Mardan	0.12543	0.04263	0.04169	0.20918
Nowshera	0.15208	0.04139	0.07079	0.23337
Peshawar	0.13867	0.02178	0.09590	0.18145
Shangla	0.10217	0.09215	0.00000	0.28317
Swabi	0.15610	0.03933	0.07884	0.23335
Swat	0.08531	0.03459	0.01738	0.15325
Tank	0.23083	0.06310	0.10689	0.35478
Tor Ghar	0.12825	0.10173	0.00000	0.32806
Upper Dir	0.10630	0.04211	0.02358	0.18902
Awaran	0.09253	0.06820	0.00000	0.22648
Barkhan	-	-	-	-
Bolan/Kachhi	-	-	-	-
Chagai	0.20664	0.08633	0.03707	0.37622
Dera Bugti	0.02111	0.02415	0.00000	0.06854
Gawadar	-	-	-	-
Harnai	0.18808	0.09760	0.00000	0.37977
Jaffarabad	0.33648	0.10024	0.13959	0.53337
Jhal Magsi	0.33588	0.10676	0.12619	0.54558
Kalat	0.14536	0.04765	0.05176	0.23895
Kech/Turbat	0.05588	0.04001	0.00000	0.13447
Kharan	0.29779	0.12019	0.06172	0.53386
Khuzdar	0.14176	0.05612	0.03152	0.25199
Killa Abdullah	0.22277	0.10220	0.02203	0.42351
Killa Saifullah	0.22522	0.09186	0.04478	0.40565
Kohlu	0.23590	0.10968	0.02046	0.45134
Lasbela	0.14117	0.05413	0.03485	0.24750
Loralai	0.20748	0.08955	0.03160	0.38336
Mastung	0.17062	0.11166	0.00000	0.38994
Musakhel	0.06034	0.04497	0.00000	0.14866
Nasirabad/Tamboo	-	-	-	-
Nushki	0.22428	0.07855	0.06999	0.37857
Panjgur	-	-	-	-
Pishin	-	-	-	-

Districts	**Estimates**	**STD**	**LCI**	**UCI**
Quetta	0.16201	0.03799	0.08739	0.23662
Sherani	0.10121	0.06580	0.00000	0.23046
Sibi	0.16982	0.09054	0.00000	0.34765
Washuk	0.14123	0.15812	0.00000	0.45180
Zhob	0.06902	0.06847	0.00000	0.20350
Ziarat	0.06750	0.06361	0.00000	0.19244
Sohbat Pur	-	-	-	-
Astore	0.16141	0.05591	0.05159	0.27123
Baltistan	0.20255	0.05755	0.08951	0.31560
Diamir	0.22176	0.03926	0.14464	0.29888
Ghanche	0.18134	0.05768	0.06805	0.29463
Ghizer	0.10827	0.03055	0.04826	0.16828
Gilgit	0.11907	0.03952	0.04144	0.19670
Nagar	0.13927	0.06816	0.00539	0.27316
Kharmang	0.26132	0.06655	0.13062	0.39203
Hunza	0.14417	0.05804	0.03018	0.25816
Shigar	0.18931	0.04455	0.10180	0.27682
Islamabad	0.12176	0.01404	0.09418	0.14933
Bajour	0.16153	0.05553	0.05247	0.27059
Khyber	0.16392	0.03358	0.09797	0.22987
Kurram	0.17180	0.03622	0.10066	0.24293
Mohmand	0.13543	0.05623	0.02498	0.24588
North Waziristan	0.16409	0.04821	0.06939	0.25878
Orakzai	0.18191	0.07533	0.03394	0.32987
South Waziristan	0.14925	0.07318	0.00551	0.29300
Bagh	0.14856	0.03771	0.07449	0.22263
Bhimber	0.06790	0.03574	0.00000	0.13809
Hattian Bala	0.11594	0.03598	0.04526	0.18661
Haveli	0.23261	0.06422	0.10648	0.35874
Kotli	0.10027	0.02287	0.05536	0.14519
Mirpur	0.10908	0.01983	0.07013	0.14804
Muzaffarabad	0.15449	0.02528	0.10484	0.20413
Neelum	0.15633	0.08395	0.00000	0.32123
Poonch	0.17589	0.04386	0.08974	0.26204
Sudhonti	0.14406	0.04519	0.05530	0.23282

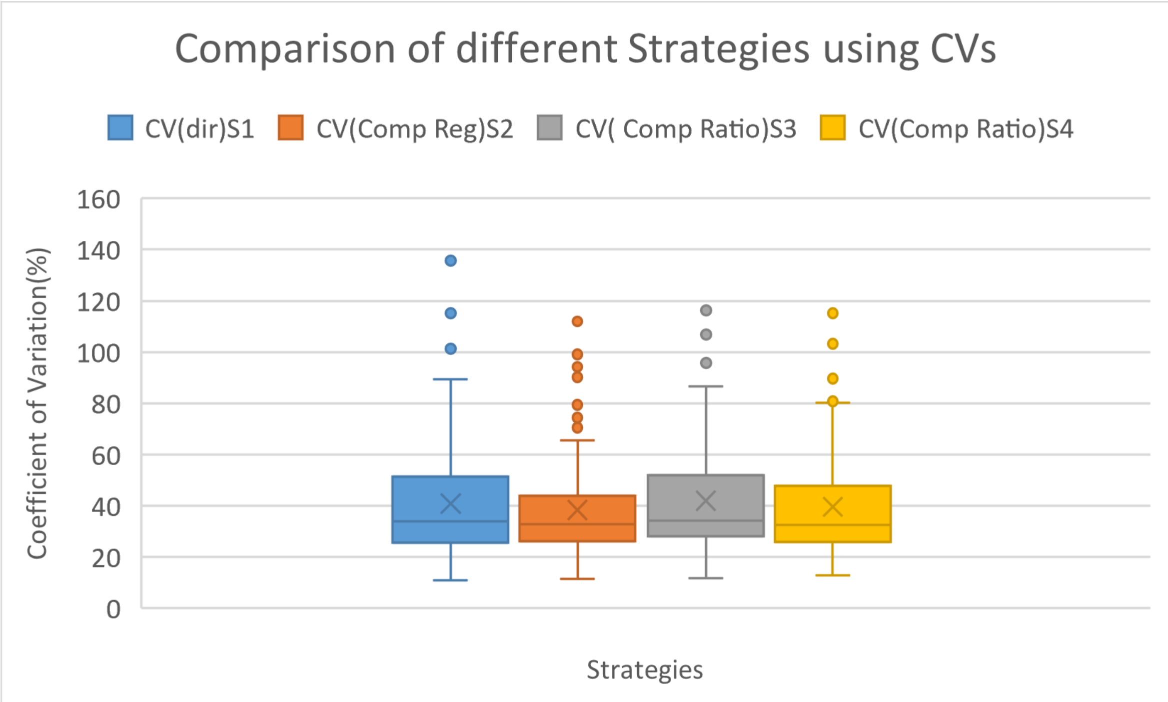

Figure 3.1: Comparison of different Strategies using CVs

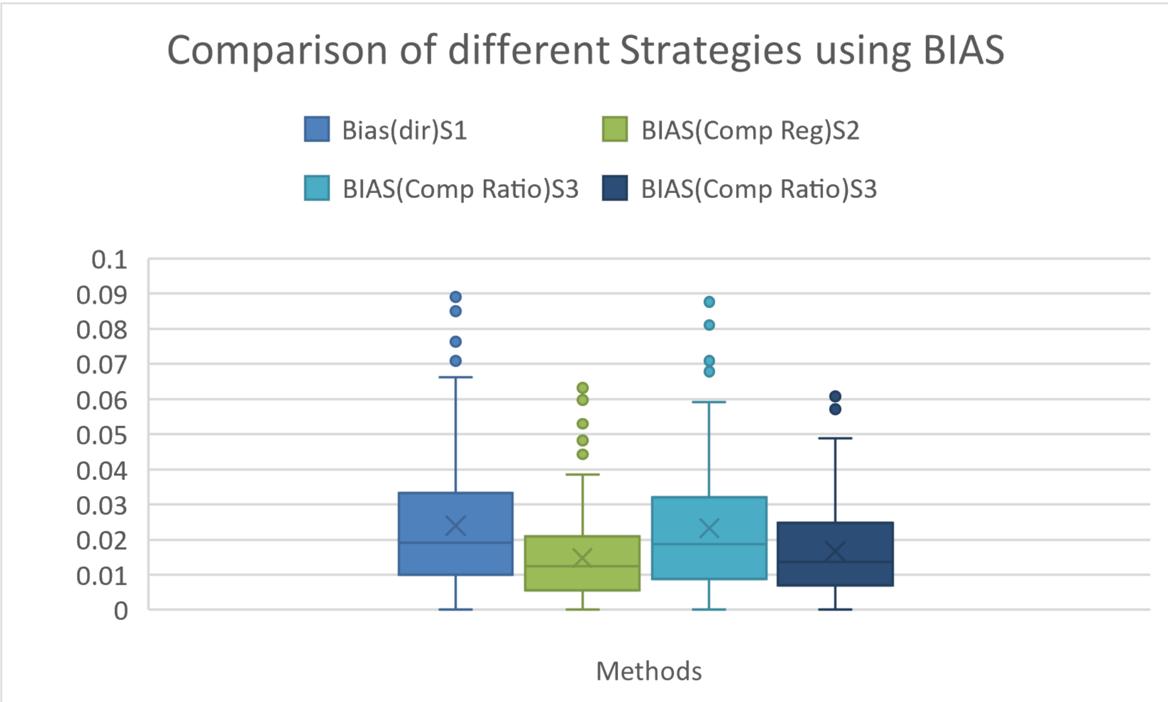

Figure 3.2: Comparison of different Strategies using BIAS

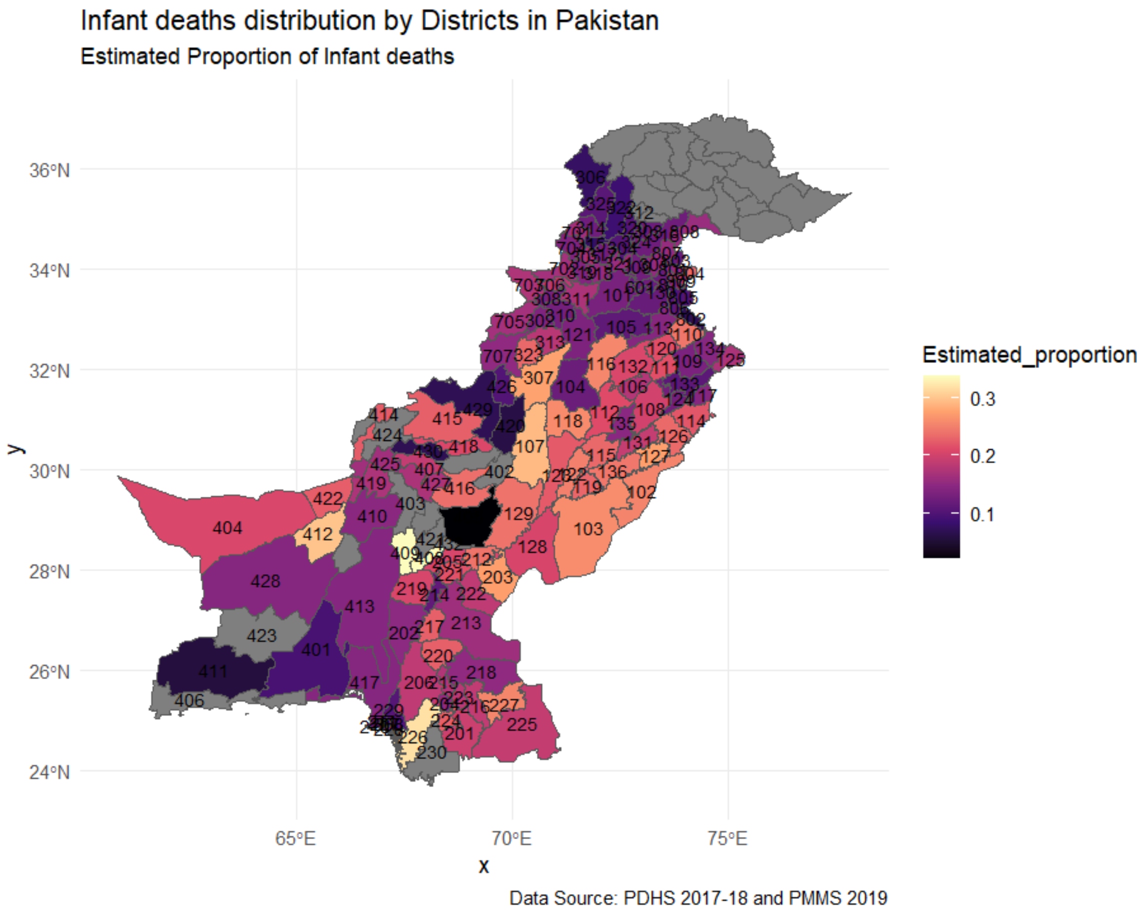

Figure 3.3: Infant Deaths Distribution by Districts in Pakistan

Table 3.6. District-wise estimated proportions, standard errors and confidence intervals for neo natal mortality

Districts	Estimates	STD	LCI	UCI
Attock	0.04304	0.03563	0.02695	0.11302
Bahawalnagar	0.10859	0.04418	0.02181	0.19537
Bahawalpur	0.10197	0.03233	0.03846	0.16548
Bhakkar	0.06515	0.04284	0.01899	0.14929
Chakwal	0.06097	0.04382	0.02509	0.14704
Chiniot	0.04538	0.03702	0.02733	0.11809
Dera Ghazi Khan	0.08129	0.03894	0.00481	0.15778
Faisalabad	0.06891	0.02176	0.02617	0.11165
Gujranwala	0.05149	0.02209	0.00810	0.09489
Gujrat	0.06662	0.03816	0.00833	0.14156
Hafizabad	0.07854	0.04654	0.00000	0.16995
Jhang	0.10771	0.04867	0.01213	0.20330
Jhelum	0.02976	0.04409	0.00000	0.11637
Kasur	0.06848	0.02897	0.01158	0.12538
Khanewal	0.08497	0.03797	0.01040	0.15954
Khushab	0.10125	0.05517	0.00000	0.20961
Lahore	0.05353	0.01618	0.02175	0.08530
Layyah	0.11405	0.05484	0.00633	0.22177
Lodhran	0.09305	0.05821	0.00000	0.20738
Mandi Bahauddin	0.08197	0.06066	0.00000	0.20112
Mianwali	0.03906	0.04193	0.00000	0.12142
Multan	0.09209	0.03175	0.02974	0.15445
Muzaffargarh	0.05605	0.02971	0.00000	0.11441
Nankana Sahib	0.02045	0.03369	0.00000	0.08661
Narowal	0.05576	0.03853	0.00000	0.13143
Okara	0.09697	0.04354	0.01145	0.18249
Pakpattan	0.10243	0.05141	0.00145	0.20341
Rahim Yar Khan	0.09558	0.04377	0.00961	0.18155
Rajanpur	0.09571	0.05477	0.00000	0.20329
Rawalpindi	0.04723	0.02441	0.00000	0.09517
Sahiwal	0.06223	0.03707	0.01059	0.13505
Sargodha	0.08186	0.03286	0.01732	0.14640
Sheikhupura	0.04783	0.03345	0.01787	0.11354
Sialkot	0.05439	0.03165	0.00778	0.11655
Toba Tek Singh	0.05405	0.04030	0.02510	0.13321
Vehari	0.10096	0.04332	0.01588	0.18603

Districts	**Estimates**	**STD**	**LCI**	**UCI**
Badin	0.06868	0.03500	0.00006	0.13741
Dadu	0.04465	0.03875	0.03146	0.12077
Ghotki	0.11074	0.03489	0.04222	0.17926
Hyderabad	0.04703	0.02940	0.00000	0.10478
Jacobabad	0.09224	0.04844	0.00000	0.18737
Jamshoro	0.05971	0.03858	0.00000	0.13548
Karachi Central	0.03130	0.02310	0.00000	0.07667
Karachi East	0.03882	0.01975	0.00003	0.07762
Karachi South	0.04658	0.03532	0.00000	0.11595
Karachi West	0.02815	0.01937	0.00000	0.06619
Kashmore	0.06801	0.04078	0.00000	0.14810
Khairpur	0.07571	0.03146	0.01391	0.13751
Larkana	0.04004	0.02087	0.00000	0.08103
Matiari	0.05951	0.04881	0.00000	0.15538
Mirpur Khas	0.06811	0.02839	0.01235	0.12387
Naushahro Feroze	0.07628	0.03974	0.00000	0.15433
Sanghar	0.05158	0.02763	0.00000	0.10585
Shahdad Kot	0.06390	0.03027	0.00445	0.12334
Nawabshah/Shaheed Benazir Abad	0.06164	0.04040	0.00000	0.14100
Shikarpur	0.07518	0.04102	0.00000	0.15574
Sukkur	0.08006	0.04116	0.00000	0.16090
Tando Alla Yar	0.06830	0.03801	0.00635	0.14295
Tando Muhammad Khan	0.07174	0.03707	0.00106	0.14455
Tharparkar	0.09090	0.03940	0.01350	0.16829
Thatta	0.06501	0.03614	0.00598	0.13600
Umer Kot	0.08740	0.05004	0.01089	0.18569
Korangi	0.03932	0.02558	0.01092	0.08955
Malair	0.03390	0.02634	0.00000	0.08564
Sujawal	0.10132	0.05077	0.00160	0.20103
Abbottabad	0.04458	0.02687	0.00819	0.09735
Bannu	0.04490	0.03031	0.01464	0.10444
Batagram	0.06466	0.06451	0.06206	0.19137
Buner	0.03720	0.04169	0.04469	0.11909
Charsadda	0.05830	0.02984	0.00031	0.11691
Chitral	0.02381	0.02574	0.02674	0.07435
D. I. Khan	0.10519	0.03449	0.03746	0.17293
Hangu	0.04039	0.04197	0.00000	0.12283
Haripur	0.04954	0.03460	0.01841	0.11749
Karak	0.04523	0.04515	0.04346	0.13391

Districts	**Estimates**	**STD**	**LCI**	**UCI**
Kohat	0.06892	0.03728	0.00431	0.14215
Kohistan	-	-	-	-
Lakki Marwat	0.05715	0.04311	0.02753	0.14183
Lower Dir	0.04235	0.02542	0.00758	0.09228
Malakand Protected Area	0.03252	0.04080	0.04761	0.11265
Mansehra	0.03796	0.03325	0.02735	0.10326
Mardan	0.04029	0.02314	0.00517	0.08574
Nowshera	0.05304	0.02845	0.00283	0.10892
Peshawar	0.04400	0.01722	0.01018	0.07783
Shangla	0.03381	0.04439	0.05338	0.12099
Swabi	0.07078	0.03089	0.01011	0.13145
Swat	0.04657	0.03637	0.02487	0.11802
Tank	0.07092	0.03742	0.00000	0.14441
Tor Ghar	0.04026	0.04821	0.05443	0.13495
Upper Dir	0.02691	0.02275	0.01777	0.07159
Awaran	0.00953	0.02378	0.03718	0.05624
Barkhan	0.08096	0.05020	0.01765	0.17957
Bolan/Kachhi	0.02765	0.03467	0.00000	0.09576
Chagai	0.06584	0.05127	0.03485	0.16653
Dera Bugti	0.03057	0.04429	0.05642	0.11756
Gawadar	0.00000	0.00000	0.00000	0.00000
Harnai	0.04568	0.04977	0.05208	0.14344
Jaffarabad	0.12233	0.06304	0.00148	0.24615
Jhal Magsi	0.10832	0.07020	0.02955	0.24620
Kalat	0.04206	0.02467	0.00640	0.09052
Kech/Turbat	0.02110	0.02070	0.01956	0.06176
Kharan	0.07944	0.05364	0.02591	0.18479
Khuzdar	0.06320	0.03800	0.01143	0.13784
Killa Abdullah	0.02917	0.03460	0.03879	0.09714
Killa Saifullah	0.06243	0.04179	0.01965	0.14450
Kohlu	0.04649	0.03643	0.02507	0.11805
Lasbela	0.04675	0.02588	0.00409	0.09759
Loralai	0.03901	0.03152	0.00000	0.10093
Mastung	0.04564	0.05110	0.05473	0.14600
Musakhel	0.01446	0.02330	0.03131	0.06023
Nasirabad/Tamboo	0.04435	0.03572	0.00000	0.11450
Nushki	0.07043	0.05122	0.03016	0.17103
Panjgur	0.02096	0.04448	0.06641	0.10832
Pishin	0.04656	0.03110	0.00000	0.10764

Districts	**Estimates**	**STD**	**LCI**	**UCI**
Quetta	0.04113	0.01619	0.00933	0.07293
Sherani	0.04486	0.04966	0.05268	0.14240
Sibi	0.06684	0.04741	0.02629	0.15996
Washuk	0.03287	0.05633	0.00000	0.14350
Zhob	0.02901	0.04115	0.05181	0.10983
Ziarat	0.01619	0.02369	0.03034	0.06273
Sohbat Pur	0.06253	0.04974	0.03516	0.16023
Astore	0.05497	0.03462	0.01304	0.12297
Baltistan	0.06561	0.02456	0.01737	0.11384
Diamir	0.05953	0.02760	0.00533	0.11374
Ghanche	0.04633	0.02290	0.00134	0.09131
Ghizer	0.05068	0.02179	0.00788	0.09348
Gilgit	0.03491	0.01724	0.00104	0.06878
Nagar	0.05611	0.04669	0.03560	0.14782
Kharmang	0.10310	0.04956	0.00576	0.20044
Hunza	0.05638	0.03728	0.01684	0.12960
Shigar	0.07211	0.02886	0.01544	0.12879
Islamabad	0.04014	0.00873	0.02299	0.05729
Bajour	0.03854	0.02539	0.01132	0.08840
Khyber	0.04630	0.01748	0.01198	0.08062
Kurram	0.05053	0.02069	0.00988	0.09117
Mohmand	0.02781	0.01896	0.00943	0.06506
North Waziristan	0.04786	0.02390	0.00092	0.09480
Orakzai	0.04247	0.02810	0.01272	0.09766
South Waziristan	0.05406	0.04034	0.02519	0.13330
Bagh	0.06174	0.03266	0.00241	0.12589
Bhimber	0.02623	0.02564	0.02412	0.07658
Hattian Bala	0.04356	0.02722	0.00990	0.09702
Haveli	0.07975	0.03803	0.00506	0.15443
Kotli	0.04358	0.01845	0.00734	0.07982
Mirpur	0.05248	0.01610	0.02086	0.08411
Muzaffarabad	0.05239	0.01721	0.01859	0.08618
Neelum	0.07321	0.05008	0.02516	0.17158
Poonch	0.05317	0.02043	0.01304	0.09330
Sudhonti	0.05880	0.03633	0.01256	0.13015

Table 3.7. District-wise Comparison of different Strategies using CVs

Districts	$CV\hat{(T_i)}^{S1(dir)}$	$CV\hat{(T_i)}^{S2((comp\_reg)}$	$CV\hat{(T_i)}^{S3((comp\_ratio)}$	$CV\hat{(T_i)}^{S4(comp\_ratio)}$
Attock	52.411	45.439	41.623	41.741
Bahawalnagar	21.474	20.867	27.962	28.610
Bahawalpur	15.225	15.588	15.397	15.636
Bhakkar	39.324	37.927	43.741	45.066
Chakwal	48.870	48.173	51.877	56.059
Chiniot	63.283	49.420	43.339	44.574
Dera Ghazi Khan	25.484	23.841	22.730	21.325
Faisalabad	18.283	17.318	15.258	15.193
Gujranwala	23.355	22.951	30.310	28.300
Gujrat	25.283	26.811	28.849	27.610
Hafizabad	30.445	36.786	29.749	26.708
Jhang	24.914	25.435	30.534	30.585
Jhelum	135.727	70.588	80.943	106.898
Kasur	23.737	22.648	20.470	19.887
Khanewal	20.803	23.032	21.817	21.882
Khushab	44.422	34.957	34.214	34.630
Lahore	14.137	14.426	17.510	16.809
Layyah	25.194	27.937	28.931	31.667
Lodhran	25.240	29.338	28.009	31.269
Mandi Bahauddin	43.622	40.333	46.983	36.999
Mianwali	51.963	58.821	76.832	52.928
Multan	33.335	21.456	21.141	20.279
Muzaffargarh	28.497	25.853	24.094	23.324
Nankana Sahib	78.703	82.103	95.697	80.832
Narowal	34.440	33.746	32.283	31.218
Okara	40.320	28.153	28.569	30.058

Districts	$CV(\hat{T_i})^{S1(dir)}$	$CV(\hat{T_i})^{S2((comp\_reg)}$	$CV(\hat{T_i})^{S3((comp\_ratio)}$	$CV(\hat{T_i})^{S4(comp\_ratio)}$
Pakpattan	27.539	26.555	24.216	24.190
Rahim Yar Khan	25.753	21.537	32.066	33.265
Rajanpur	35.587	34.591	45.377	40.937
Rawalpindi	26.855	27.234	32.804	31.665
Sahiwal	31.938	33.554	28.015	27.142
Sargodha	21.714	22.638	21.063	21.244
Sheikhupura	44.824	40.130	50.734	50.909
Sialkot	34.358	30.997	31.371	28.768
Toba Tek Singh	32.946	36.571	32.587	32.172
Vehari	27.199	26.141	23.518	22.992
Badin	28.433	26.449	34.094	35.176
Dadu	78.385	45.315	57.301	61.818
Ghotki	17.047	17.300	23.308	22.148
Hyderabad	27.413	26.988	34.498	33.079
Jacobabad	33.821	29.843	37.478	42.382
Jamshoro	31.070	32.215	30.779	30.329
Karachi Central	56.107	53.557	67.126	59.726
Karachi East	55.490	34.075	34.544	38.512
Karachi South	50.584	44.466	55.598	56.474
Karachi West	29.196	29.385	32.383	29.901
Kashmore	52.078	26.586	26.096	32.638
Khairpur	24.248	26.743	30.748	22.950
Larkana	42.555	35.703	45.577	49.355
Matiari	61.989	37.492	42.947	47.247
Mirpur Khas	21.796	21.790	19.050	19.065
Naushahro Feroze	23.978	26.190	31.178	23.356
Sanghar	28.213	27.033	26.779	25.702
Shahdad Kot	23.165	23.430	25.886	24.617

Districts	$CV(\hat{T_i})^{S1(dir)}$	$CV(\hat{T_i})^{S2((comp\_reg)}$	$CV(\hat{T_i})^{S3((comp\_ratio)}$	$CV(\hat{T_i})^{S4(comp\_ratio)}$
Nawabshah/Shaheed Benazir Abad	41.842	32.366	40.337	32.524
Shikarpur	22.711	28.519	25.924	27.296
Sukkur	29.340	29.890	33.753	28.779
Tando Alla Yar	51.304	32.471	32.902	36.657
Tando Muhammad Khan	24.276	25.672	24.950	23.836
Tharparkar	25.222	25.160	26.103	23.136
Thatta	27.698	36.087	52.498	29.039
Umer Kot	37.429	30.099	30.866	28.295
Korangi	36.794	33.398	30.362	29.770
Malair	38.542	39.287	44.613	36.884
Sujawal	22.939	**NA**	**NA**	23.272
Abbottabad	28.094	28.247	32.587	31.175
Bannu	30.051	30.410	34.184	31.151
Batagram	64.227	60.440	69.065	69.546
Buner	69.246	65.407	71.536	75.501
Charsadda	27.876	28.175	24.493	24.281
Chitral	65.358	57.526	66.011	59.281
D. I. Khan	16.982	19.391	17.803	19.328
Hangu	62.433	60.307	73.019	65.887
Haripur	35.164	37.566	39.518	40.010
Karak	61.349	60.959	72.698	64.769
Kohat	36.670	34.443	40.443	44.061
Kohistan	**NA**	**NA**	**NA**	**NA**
Lakki Marwat	38.165	40.593	42.349	41.016
Lower Dir	31.810	32.049	39.470	33.096
Malakand Protected Area	65.233	64.837	66.687	71.215
Mansehra	60.385	48.419	49.131	50.382
Mardan	42.807	33.990	30.531	30.463

Districts	$CV(\hat{T_i})^{S1(dir)}$	$CV(\hat{T_i})^{S2((comp\_reg)}$	$CV(\hat{T_i})^{S3((comp\_ratio)}$	$CV(\hat{T_i})^{S4(comp\_ratio)}$
Nowshera	28.886	27.213	33.153	36.634
Peshawar	13.696	15.705	18.652	23.646
Shangla	81.039	90.199	85.489	79.610
Swabi	22.409	25.197	24.198	25.226
Swat	48.425	40.540	48.635	55.679
Tank	25.676	27.337	28.168	27.663
Tor Ghar	60.818	79.321	70.366	64.465
Upper Dir	36.834	39.619	40.488	38.995
Awaran	87.749	73.701	73.086	74.632
Barkhan	29.540	**NA**	**NA**	30.033
Bolan/Kachhi	61.297	**NA**	**NA**	61.843
Chagai	36.275	41.779	54.000	34.945
Dera Bugti	115.135	114.422	116.375	115.135
Gawadar	**NA**	**NA**	**NA**	**NA**
Harnai	117.437	51.893	73.420	89.760
Jaffarabad	28.542	29.792	34.616	25.918
Jhal Magsi	29.732	31.785	35.130	33.501
Kalat	28.847	32.784	30.327	30.814
Kech/Turbat	69.670	71.590	68.317	75.440
Kharan	70.404	40.360	48.248	55.538
Khuzdar	54.074	39.591	52.217	61.080
Killa Abdullah	89.523	45.877	62.400	68.044
Killa Saifullah	40.518	40.789	54.134	35.721
Kohlu	61.744	46.496	68.983	52.001
Lasbela	32.322	38.345	31.419	27.554
Loralai	48.490	43.158	55.295	41.282
Mastung	75.606	65.442	73.514	80.206
Musakhel	73.166	74.518	74.004	77.724

Districts	$CV\hat{(T_i)}^{S1(dir)}$	$CV\hat{(T_i)}^{S2((comp\_reg)}$	$CV\hat{(T_i)}^{S3((comp\_ratio)}$	$CV\hat{(T_i)}^{S4(comp\_ratio)}$
Nasirabad/Tamboo	32.134	**NA**	**NA**	32.122
Nushki	34.863	35.024	33.322	32.266
Panjgur	**NA**	**NA**	**NA**	**NA**
Pishin	25.226	**NA**	**NA**	25.121
Quetta	19.447	23.449	27.670	17.769
Sherani	45.076	65.017	54.169	56.712
Sibi	35.153	53.312	86.674	33.413
Washuk	76.301	111.957	86.489	
Zhob	101.408	99.197	106.961	103.223
Ziarat	65.919	94.245	76.439	69.566
Sohbat Pur	47.851	**NA**	**NA**	47.378
Astore	35.164	34.640	46.981	38.747
Baltistan	23.780	28.414	43.326	25.935
Diamir	18.572	17.705	25.075	24.752
Ghanche	50.436	31.807	33.954	34.800
Ghizer	30.051	28.219	33.199	39.463
Gilgit	27.095	33.194	44.663	24.393
Nagar	52.765	48.943	57.891	57.718
Kharmang	26.515	25.465	23.599	23.109
Hunza	45.115	40.256	49.418	51.524
Shigar	21.126	23.535	25.056	29.038
Islamabad	10.866	11.532	11.706	12.819
Bajour	40.046	34.376	31.098	31.714
Khyber	18.724	20.484	32.148	21.201
Kurram	29.948	21.081	18.085	17.437
Mohmand	27.225	41.520	34.041	30.416
North Waziristan	28.571	29.382	33.766	25.935
Orakzai	53.007	41.412	35.740	35.649

Districts	$CV\hat{(}T_i)^{S1(dir)}$	$CV\hat{(}T_i)^{S2((comp\_reg)}$	$CV\hat{(}T_i)^{S3((comp\_ratio)}$	$CV\hat{(}T_i)^{S4(comp\_ratio)}$
South Waziristan	45.032	49.033	68.788	41.251
Bagh	28.546	25.384	34.550	34.982
Bhimber	59.487	52.636	64.192	63.918
Hattian Bala	31.106	31.035	36.322	36.750
Haveli	25.437	27.607	29.849	31.011
Kotli	22.612	22.806	27.547	26.769
Mirpur	18.807	18.180	17.075	16.978
Muzaffarabad	16.091	16.362	15.681	16.247
Neelum	36.321	53.703	46.252	46.031
Poonch	38.232	24.937	24.110	24.032
Sudhonti	30.271	31.369	35.095	36.623

Chapter 4

CONCLUSIONS AND FUTURE RECOMMENDATION

4.0.1 CONCLUSIONS

The study focuses on improving the accuracy of survival function estimates in small areas, particularly for infant mortality. Traditional direct estimation methods often fall short due to small or zero sample sizes in sub-populations. This research addresses these limitations by proposing enhanced indirect methods, such as synthetic estimation and composite estimation. These new estimators leverage indirect approaches to improve the precision of survival function estimates by integrating various data sources and survey information. The findings highlight significant variability in survival functions across different geographic sub-populations, emphasizing the need for tailored estimation methods that consider the unique demographic characteristics of each area. The improved estimators provide more accurate insights into survival patterns, which are crucial for public health policymakers. These accurate estimates enable better addressing of health disparities and more effective allocation of resources to areas in greatest need. The study contributes to the existing body of knowledge by refining and validating indirect estimation techniques, representing a significant advancement in small area estimation. The findings have practical applications for stakeholders, including government agencies and non-governmental organizations working on infant mortality. Overall, this research advances the field of small area estimation by overcoming the limitations of traditional direct estimation approaches, providing more reliable estimates of survival functions in small areas, and offering valuable insights for public health policy and future studies.

4.0.2 Future Recommendations

- Enhance the robustness of synthetic and composite estimators by exploring advanced statistical techniques for greater precision with smaller sample sizes.
- Integrate a broader range of auxiliary data sources, including real-time data and administrative records, to improve the accuracy and granularity of survival function estimates.
- Extend the application of the proposed estimation methods to other health outcomes, such as maternal mortality and disease prevalence, to gain a wider understanding of health disparities in small areas.
- Collaborate with policymakers to implement data-driven public health interventions based on improved estimates and assess their impact on health outcomes in small areas.
- Establish guidelines for the ethical use and privacy protection of data in small area estimation, ensuring responsible and secure handling of individual data.

Bibliography

[1] Asian Development Bank. *Tajikistan: Country Gender Assessment*. Asian Development Bank, 2020.

[2] Wenjun Li, Jennifer L Kelsey, Zi Zhang, Stephenie C Lemon, Solomon Mezgebu, Cynthia Boddie-Willis, and George W Reed. Small-area estimation and prioritizing communities for obesity control in massachusetts. *American journal of public health*, 99(3):511–519, 2009.

[3] John NK Rao and Isabel Molina. *Small area estimation*. John Wiley & Sons, 2015.

[4] John NK Rao and Isabel Molina. *Small area estimation*. John Wiley and Sons, Inc., 2nd edition, 2015.

[5] Danny Pfeffermann. Small area estimation-new developments and directions. *International Statistical Review*, 70(1):125–143, 2002.

[6] Danny Pfeffermann. New important developments in small area estimation. 2013.

[7] Kingsley Davis and Wilbert E Moore. Some principles of stratification. In *Kingsley Davis*, pages 221–231. Routledge, 2017.

[8] Jonathan Wakefield, Taylor Okonek, and Jon Pedersen. Small area estimation for disease prevalence mapping. *International Statistical Review*, 88(2):398–418, 2020.

[9] Giancarlo Manzi, David J Spiegelhalter, Rebecca M Turner, Julian Flowers, and Simon G Thompson. Modelling bias in combining small area prevalence estimates from multiple surveys. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 174(1):31–50, 2011.

[10] Esther M Friedman, Don Jang, and Thomas V Williams. Combined estimates from four quarterly survey data sets. In *Proceedings of the american statistical*

association joint statistical meetings—section on survey research methods, pages 1064–1069. American Statistical Association. Alexandria, VA, 2002.

[11] Andreea L Erciulescu, Carolina Franco, and Partha Lahiri. Use of administrative records in small area estimation. *Administrative records for survey methodology*, pages 231–267, 2021.

[12] Shakeel Ahmed. Some new strategies for estimating area level parameters using information from successive surveys. *Quality & Quantity*, pages 1–45, 2024.