University of Southampton

Faculty of Social Sciences

Social Statistics and Demography

Statistical Estimation and Inference with Aggregated and Displaced Georeferenced Data

by

Md Jamal Hossain

ORCID ID 0000-0002-2728-1055

Thesis for the degree of Doctor of Philosophy

November 2023

Abstract

University of Southampton

Faculty of Social Sciences

Social Statistics and Demography

Thesis for the degree of Doctor of Philosophy

Statistical Estimation and Inference with Aggregated and Displaced Georeferenced Data

by Md Jamal Hossain

The thesis addresses the problem of statistical inference for data affected by geocoordinate random displacement or a combination of aggregation and random displacement, which is often used to preserve respondents' confidentiality. However, the distortion induced in the location of the observations may compromise the validity of location-dependent estimates. This thesis explores various situations where such trade-offs may arise, including: 1) population density estimation, 2) estimation of health and demographic indicators for lower geographical domains, 3) regression analyses involving lower geographical domains (e.g., within the context of multilevel models), and 4) regression analyses incorporating spatial covariates calculated or linked from external data sources based on geocoordinates.

A measurement error model (MEM) is developed for the case of random displacement or aggregation. It is demonstrated under the Demographic and Health Survey (DHS) random displacement process by devising a new probability distribution for the displaced coordinates. Two methods, Kernel Density Estimation-based (KDE) and External Data-based Classification (EDC), are proposed within the MEM framework to approximate the conditional distribution of the true coordinates given the ones subject to displacement. Additionally, a novel method, KDE-ED, that combines KDE and EDC is proposed to address both aggregation and random displacement issues in approximating the conditional distribution of true coordinates. The KDE method uses kernel density estimates for approximating the unknown marginal distribution of true location coordinates and is implemented using the Stochastic Expectation-Maximization (SEM) algorithm. The EDC approximates the marginal distribution of true coordinates using external data sources and implements estimation through numerical integration.

The MEM and the two proposed KDE and EDC methods are used to address all four location-dependent statistical estimation issues mentioned. The KDE and EDC can be directly used to estimate population densities or domain parameter estimates accounting for random displacement or both aggregation and displacement errors. Apart from the EDC (or KDE)-based algorithm, a new method incorporating a parametric Bootstrap Bias correction (BC) is proposed to obtain improved estimates of the parameters in the linear mixed model, correcting misplacement error due to random displacement. Furthermore, the EDC (or KDE) can be used under regression calibration (EDC-RC) to improve the estimation of spatial covariate effects in a linear regression model under random displacement. An alternative estimator using only a non-parametric Bootstrap Bias correction over the usual OLS estimators is also proposed for the latter situation. The performance of all estimators developed, as well as the variance estimators proposed for them, is assessed via simulation exercises and illustrated using real data from the 2011 Bangladesh DHS.

Contents

List of Tables

List of Figures

Declaration of Authorship

Name: Md Jamal Hossain

Title of thesis: Statistical Estimation and Inference with Aggregated and Displaced Georeferenced Data

I declare that this thesis and the work presented in it is my own and has been generated by me as the result of my own original research. I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. None of this work has been published before submission

Signature: Date: November 2023

Acknowledgements

I would like to take this opportunity to express my deepest gratitude to all those who have contributed to the successful completion of my PhD studies.

First and foremost, I am grateful to my supervisors, Professor Dr. Nikos Tzavidis and Dr. Angela Luna Hernandez, for their unwavering support, guidance, and encouragement throughout my research journey. Their expertise, feedback, and constructive criticism have shaped my research work and helped me grow as a researcher.

I would also like to extend my heartfelt thanks to my family, friends, and loved ones, who have supported and motivated me during this time. Special thanks go to my beloved mother, father, grandmother and grandfather for their unconditional love, understanding, and encouragement.

I am also grateful to my wife, Farhana, for her unwavering support, patience, and understanding throughout this journey. Her love and belief in me have been my driving force in achieving my academic goals. I want to express my sincere appreciation to my son, Ayan, who has been a constant source of joy and inspiration throughout my studies, and to my daughter, Arisha, who was born during my studies and has brought immeasurable joy to my life.

I would like to thank the examiners and reviewers who evaluated my thesis for their valuable feedback and constructive criticism. Their insights and comments have helped me improve the quality of my work and prepare me for future research endeavours. I would also like to express my gratitude to my colleagues and the staff in the Department of Agricultural and Applied Statistics, Bangladesh Agricultural University, for their support and suggestions throughout the journey.

Finally, I am grateful for the generous support of the Commonwealth Scholarship Commission (CSC), which made this research possible. I am honoured to have been a recipient of this scholarship and thankful for the opportunity to contribute to advancing knowledge in my field. It has been a privilege to undertake this research, and I look forward to contributing to the academic community in the future.

Abbreviations

AB	Absolute Bias
ARB	Absolute Relative Bias
BBS	Bangladesh Bureau of Statistics
BDHS	Bangladesh Demographic and Health Survey
BC	Bootstrap Bias Correction
DHS	Demographic and Health Survey
EA	Enumeration Area
EDC	External Data-based Classification
EDC-RC	External Data-based on the regression calibration
ELL	Elbers,Lanjouw, and Lanjouw
GIS	Geographic Information Systems
GPS	Global Positioning System
KDE	Kernel Density Estimation
KDE-ED	External Data and Kernel Density-based
LMM	Linear Mixed Models
LRM	Linear Regression Model
MPC	Maximum Probability Covariate
MPD	Most Probable Domain
PSU	Primary Sampling Units
RC	Regression Calibration
RB	Relative Bias
RMISE	Root-Mean-Integrated-Squared-Error
RMSE	Root-Mean-Squared-Error
SDG	Sustainable Development Goals
WI	Wealth Index

Chapter 1

Introduction

1.1 Background of the research and problem statement

The thesis addresses the problem of obtaining accurate statistical estimates and inferences with aggregated and randomly displaced georeferenced data. Georeferenced data collected through various register systems or surveys are extensively used in spatial data analysis and statistical modelling. The Demographic and Health Survey (DHS) is an example of a widely-used survey that collects georeferenced data. Researchers use the nationally representative georeferenced population-based DHS dataset to examine geographic variations and inequalities in wealth (poverty), health condition, and access to resources within nations to achieve Sustainable Development Goals (SDGs) (Burgert et al., 2018). However, it is essential to ensure the confidentiality of respondents while making georeferenced data accessible to analysts, researchers, and policymakers.

One way of preserving register data or survey respondents' confidentiality involves introducing measurement errors to location coordinates through aggregation and random displacement processes, such as the DHS random Enumeration Area (EA) displacement process, described in the literature review Chapter 2 of this thesis. An EA is a small geographical area of the country with 100-300 households. Nevertheless, the aggregation and random displacement process can introduce measurement errors in the resulting data or estimates that depend on location. Therefore, it is crucial to investigate the impact of the aggregation and displacement error on the estimates and then use appropriate statistical methods accounting for the measurement error to derive better estimates. In particular, two questions need to be addressed: a) what is the impact of aggregation and displacement on statistical estimation and inference? and b) can we derive precise and consistent

estimates accounting for the measurement error caused by the aggregation and displacement?

As aforementioned, various statistical analyses of aggregated and randomly displaced location data can be affected, for example, density estimates of the data. The precision and smoothness of the density estimates of observations depend on the availability of the exact coordinates of observations. The aggregation concentrates data at the centroid of the EA if the centroid is used as the reference point, while the random displacement moves the original location from one place to another. Therefore, the density estimates of observations under aggregation and displacement may be biased and differ from the true estimates. To the best of our knowledge, past studies (for example, Groß et al., 2017, 2020) attempted to obtain corrected density estimates of data affected by only aggregation error under a measurement error model and did not consider the random displacement or the more complex aggregation and random displacement problem.

Furthermore, estimates of health and demographic indicators at the sub-national level based on survey data are crucial for targeting interventions and making policies. In particular, most low-income countries have decentralised provision and funding of public services, including health and nutrition. However, routine (and survey) data representative of administrative units relevant to service provision/delivery are unavailable. Survey estimates derived, for example, using geospatial modelling used by the United Nations and national governments to monitor subnational development indicators tend to use approaches not accounting for displaced EAs, potentially resulting in a misdirection of resources and funding at subnational levels. Specifically, any estimates at an administrative level lower than the random displacement designated administrative boundary (referred to as a domain) may be biased due to the random displacement process. This is because an EA can be placed outside of its original domain area due to the random displacement process, referred to as an EA misplacement. Previous studies in the literature have paid less attention to the domain estimates under a misplacement error caused by the random displacement process. Thus, in this thesis, we aim to explore the impact of the random point displacement on estimates at lower admin levels (domain) and develop methods under a measurement error model for displaced coordinates, considering any additional information available to improve the estimates.

In the case of analytical inference, if a random EA displaced coordinates are used to link variables from different data sources, for instance, one of which is a response variable available at DHS EA points, and another is a covariate available at the disaggregated level of the geography from a different survey, then the regression coefficient estimates may be biased (Warren et al., 2016b; Wilson and Wakefield, 2021). A covariate derived from the true location coordinate is hereafter called a spatial covariate. An example is the distance from the EA centroid to the country's nearest

health facility. Measurement errors can arise when calculating such spatial covariates (e.g., distance or environmental condition), using displaced location coordinate. A regression model with measurement error in variables (covariates) can result in biased estimates of its regression coefficients (Aigner, 1973; Carroll et al., 2006, section 3.2; Grace, 2017, section 2.2). Past studies, for example, Warren et al. (2016b) worked on correctly linking variables from different data sources under the DHS random displacement, and Warren et al. (2016a); Perez-Heydrich et al. (2016); Grace et al. (2019) attempted to obtain estimates of spatial covariates by correcting the displacement error. The methodological details of these studies are discussed and reviewed in Chapter 2. Most of the studies used the likely true locations, which were obtained by displacing the observed displaced coordinates using the random displacement process, in their estimation methods in different ways to correct the displacement error. Enhanced location-dependent estimates can be achieved by generating more likely true locations using all available information. Therefore, this thesis aims to develop a measurement error model to generate more likely true locations using additional information from external data sources, which are likely to be associated with the distribution of the unknown true locations.

Multilevel models are used to analyze grouped observations or when a survey shows a hierarchical data structure, such as the case with the Demographic and Health Surveys (DHS). Random effects in multilevel models are specified at the group level to capture the grouping structure of observations. For instance, the groups can be defined as the EAs or administrative units of the country in the DHS data. When the random effects in mixed models are specified at below the random displacement restricted administrative units, such as the sub-district level or admin 3 of the country, where EAs can be misplaced or mixed among sub-districts due to a random displacement process, the estimates of multilevel model parameters, especially the random variance components, and associated model-based estimates of finite population parameters (e.g., group/sub-district means) may be biased. When random effects are specified at progressively lower administrative levels, the potential for misplaced EAs increases, leading to greater bias in estimates. Thus, in situations with high misplacement errors, failing to account for this error can result in misleading inferences. To the best of our knowledge, previous studies did not consider this problem. This thesis aims to contribute by addressing these gaps in the literature by developing bias-corrected methods of the linear mixed model parameters.

1.1.1 Research objectives

The specific research objectives of this thesis are:

i. To develop measurement error models and estimation methods for the aggregated and randomly displaced georeferenced data.

ii. To develop methods for accurate estimation of population density of interest and health and demographic indicators at lower geographical levels accounting for aggregation and displacement errors.

iii. To develop methods for estimating fixed effects and variance components in mixed models and estimating variances of the model parameters estimates accounting for the random displacement processes.

iv. To propose methods for correcting bias in the estimates of spatial covariate effects within a linear regression model under random displacement, and to estimate the variances of the model parameter estimates under the proposed methods.

1.2 Data sources

The thesis evaluates the proposed methods and demonstrates their practical applications using data from various sources. To illustrate the methodology, simulation data and real survey data are used. The simulation data is generated to reflect different real survey and census data characteristics. To provide a complete understanding of the data and characteristics used in the thesis, they are discussed in detail below:

1.2.1 Bangladesh demographic and health survey (BDHS) data

The 2011 BDHS (NIPORT, 2013) publicly available data is used in this thesis to show the application of the proposed methods under the random displacement process. The survey includes Enumeration Area (EA) centroids that have been displaced to protect respondents' confidentiality and enable researchers to explore spatial variations at disaggregated geographical levels. The nationally representative survey data was collected using a two stage stratified sample that covered all 7 divisions (admin-1), 64 districts (admin-2), and 396 (out of 544) Upazilas (admin-3) of the country.

Following the BDHS 2011 documentation (NIPORT, 2013), the survey utilised the list of enumeration areas (EAs) prepared for the Bangladesh Population and Housing Census 2011 as its sampling frame. The Bangladesh Bureau of Statistics (BBS) created EAs that, on average, comprised approximately 120 households each. Details about the creation of EAs are described in section 1.2.3. However, during the first stage of sampling, 600 EAs, also known as clusters, were chosen, with

the probability proportional to the number of households within each EA. These EAs were stratified by division (admin 1) and rural-urban classification, with 393 EAs in rural regions and 207 EAs in urban regions. In the next stage, a sample of 30 households was selected systematically from each EA, based on the order in the initial household listing. The chosen households were likely to be dispersed throughout the entire EA. Then, the entire sample was georeferenced at the centre of the corresponding EA.

While EAs are generally designed to focus on population count rather than geographic size (measured in square meters), their specific sizes can vary based on population density and whether the area is urban or rural. For example, an EA in an urban area such as Dhaka, the capital of Bangladesh, might be considerably smaller in size than an EA located in a more rural region of the country. However, to preserve the confidentiality of respondents, all EA coordinates in the 2011 survey are displaced using the DHS displacement algorithm, which is described in section 3.2.1, before being provided to data analysts. The DHS displacement process involves a random direction and a random distance. Displacement distances, determined by population density, vary between urban and rural areas. Specifically, within a given circle with a radius equal to the maximum displacement distance, urban locations are displaced by 0-2 kilometres, while rural locations are displaced by 0-5 kilometres. Furthermore, 1% of rural locations are randomly displaced by 0-10 kilometres. Notably, these urban-rural distance parameters remain consistent across all DHS surveys (Burgert et al., 2013). Nevertheless, data analysts are not informed about which rural locations undergo this 0-10 kilometre displacement. The EA displacement is constrained within the district (admin 2) boundaries. Consequently, in the case of the BDHS 2011, the displaced EA coordinates may be wrongly positioned at lower administrative levels in Bangladesh, for example, the upazila level (admin 3). It is worth noting that only the displaced EA centroid coordinates are available in the survey GPS data.

By mapping the displaced EA coordinates to their respective geographical regions, we can determine the corresponding upazila identifiers. However, due to the use of displaced coordinates, these identifiers may not be accurate if the EA coordinates are misplaced. These are hereafter referred to as “displaced upazila IDs”. Although the true EA coordinates are not provided in the GPS data, the correct upazila identifiers (known as upazila IDs, distinct from coordinates) for each EA and household are available in the BDHS 2011 survey data. This inclusion was not continued in surveys after 2011. These identifiers are hereafter referred to as “true upazila IDs”. These true upazila IDs can be used to validate any estimates at the upazila level under the DHS random displacement. To illustrate the performance of the proposed methods using real survey data, the thesis aims to

estimate poverty, i.e., the proportion of poor households at the upazila level in Bangladesh, using the 2011 BDHS Wealth Index (WI) data. This is discussed in Chapter 3. The WI is an asset-based indicator of people's living standards and is an alternative measure of poverty (Falkingham and Namazie, 2002, section 4.1). In this thesis, a household is considered relatively poor if its wealth index falls below the 40th percentile of the wealth indices for all surveyed households.

1.2.2 WorldPop gridded population count data

The population count at a finer spatial resolution may be more closely associated with the unknown distribution of the true coordinates of units, e.g., households or Enumeration Areas (EA). Additionally, following the 2011 BDHS documentation, sampled EAs were selected with a probability proportional to the 2011 census EA size (i.e. the number of households) (NIPORT, 2013). This suggests that the gridded population count may be a good predictor for selecting the likely true EA points from the correct administrative units. Although the population count for the true potential location data is not always available, estimates can be sourced from the WorldPop database that provides annual gridded 100m x 100m population counts for each country (Bondarenko et al., 2020). In this thesis, to apply the proposed method, we consider the WorldPop gridded population counts as fixed, assuming they are measured without error.

The WorldPop top down disaggregation method uses a two-step dasymetric mapping technique to estimate population counts at a higher spatial resolution than is available from admin units (Sorichetta et al., 2015; Stevens et al., 2015). First, a weighting layer is computed using the random forest machine learning algorithm based on the relationship between population count and high-resolution geospatial covariates, such as lights at night, building locations and road networks at the administrative level. In the second step, the weighting layer is used to dis-aggregate projected census totals and estimate population counts for smaller areas, such as 100m grid cells. The population density for each grid is calculated by dividing the gridded population count by the corresponding gridded area (Leasure et al., 2021; Lazar et al., 2021).

Two different types of estimation are adopted by the top-down disaggregation method: unconstrained estimation across all land grid squares worldwide and constrained estimation within areas identified as containing built settlements using satellite-derived building footprint data. In addition to building footprint data, a water mask is also used to identify areas that are unlikely to contain residential populations. The constrained approach achieves a more precise population distribution as it avoids predicting small population numbers in areas that are likely uninhabited but accurately mapped with settlements or buildings (Stevens et al., 2015).

1.2.3 Bangladesh population and housing census data

Some characteristics of the 2011 Bangladesh Population and Housing Census (hereafter, it is referred to as the 2011 census) and the 2011 BDHS EAs are used to generate the model-based simulated data, which is described in Chapter 4 of this thesis and used to evaluate the performance of the proposed methods. In particular, as discussed in Section 1.2.1, the true coordinates of the 2011 BDHS EAs are not available. Also, as a sampling frame, the survey used the list of EAs prepared for the 2011 census. The list of the 2011 census EAs with their number of households is also not available. Thus, we generate EA population data using the WorldPop 100m x 100m gridded population counts following the same principles used to create the 2011 census EAs, which are discussed below:

According to the documentation of the 2011 Bangladesh census (BBS, 2014), Bangladesh was divided into 296,718 EAs, the lowest unit in the census. On average, each EA consisting of approximately 120 households. Intermediate administrative levels, such as Mauza and Mahalla (admin-5) maps, were used to delineate EAs and ensure proper identification, with Mauza being the lowest rural geographic unit and Mahalla being the lowest urban geographic unit. Population counts for Mauza and Mahalla were published in the National census 2011 report, and the maps can be compiled based on publicly available information from the 2011 census report (BBS, 2014) or the Bangladesh Bureau of Statistics website (http://www.bbs.gov.bd/), the download link is **here**. Mauzas with more than 120 households were divided into two or more EAs, while EAs with fewer than 80 households were merged with other adjacent smaller EAs. One or more EAs were also created in urban areas by dividing Mahallas.

1.3 Contributions of the thesis

The thesis addresses the problem of obtaining accurate statistical estimates and inferences using georeferenced data that has either been randomly displaced or aggregated and randomly displaced to preserve respondents' confidentiality, resulting in measurement errors in location-dependent estimates. Therefore, in the case of only random displacement, a key contribution is that we develop a measurement error model (MEM) under the DHS random displacement process, followed by the development of the new probability distribution for the random displacement, which can be potentially extended to other forms of random displacement processes. This is described in Chapter 3. We propose two estimation methods for the MEM to obtain density estimates and domain parameter estimates by drawing likely true location coordinates. The first method is based on Kernel

Density Estimates (KDE) as the marginal distribution of the true location data. The estimation is implemented using the Stochastic Expectation-Maximization (SEM) algorithm. Additionally, we derive the bandwidth of the KDE as part of the estimation process, which simplifies implementation and works with any bandwidth selection method, making it highly flexible. As an alternative to the KDE-based method, a second method called External Data Classification (EDC)-based is proposed, which approximates the unknown underlying distribution of true location data using additional information from external data sources. The estimation of the EDC-based method is implemented through numerical integration, and both methods are described in Chapter 3.

Also, under both aggregation and random displacement, another key contribution of this thesis is that a method combining KDE and EDC, called KDE-ED, is proposed to obtain density estimates and estimates of the domain parameters by developing a MEM. This is also discussed in Chapter 3. The proposed methods' performance is compared to competing approaches in controlled settings through simulation studies and a real survey data application.

The next contribution is described in Chapter 4, where we develop the framework of the linear mixed model under the random displacement process. The proposed model is fitted by an algorithm under the EDC method, which is developed under the measurement error model by bringing additional information from different external sources. We also propose a methodology for estimating the linear mixed model's parameters, fixed effects and variance components, using a parametric bootstrap bias correction (BC) under random displacement. In addition to the point estimates, we develop variance estimation of the fixed and variance component parameters estimates under the proposed EDC and BC methods. Finally, a model-based simulation is used to assess the performance of the proposed methods for fitting the linear mixed model and obtaining estimates of the model parameters, and the model-based group means under random displacement.

The final contribution of this thesis is that we develop the theory and methods for estimating the effects of spatial covariates in a linear regression model under the random displacement process. The model considers two types of spatial covariates: distance-based and linking area-level covariates. This is described in Chapter 5. We propose two methods to fit the model. The first method is the proposed EDC under regression calibration (EDC-RC) method that uses additional information from external data sources and estimates the expected values of the spatial covariates. We also develop a bootstrap bias correction method for correcting the bias of the estimates of the spatial covariate parameters of the linear regression model under the random displacement process. Apart from the point estimates, we develop variance estimation for the model parameter estimates under the proposed EDC-RC and BC methods. This is implemented using model-based bootstrapping

under repeated random displacement. A model-based simulation study is conducted to assess the performance of the proposed methods.

1.4 Ethics approval

The Ethics Committee has approved the use of the dataset (Submission ID: 61348).

1.5 Thesis outline

This thesis is divided into six chapters, starting with a literature review in **Chapter 2**. It examines the primary areas of research of this thesis, including the protection of respondents' confidentiality through aggregation and displacement of geo-referenced data and measurement error models. The literature review also focuses on density estimation with aggregated and random displaced data, spatial statistics, mixed models, regression with spatial covariates and bootstrap bias correction under random displacement. Recent developments and challenges in statistical estimation and inference with aggregated and displaced location data are discussed, along with possible methodological extensions.

Chapter 3 is devoted to developing the measurement error model for randomly displaced georeferenced data. It also describes the proposed estimation method for obtaining density and domain parameter estimates correcting for displacement error based on Multivariate Kernel Density Estimates (KDE). Also, an alternative method based on external data sources is presented to approximate the distribution of the unknown true locations. The chapter also presents a proposed method for obtaining density and domain parameter estimates accounting for aggregation and displacement errors. Simulation designs for analysing population (also, known as register) data under only displacement and both aggregation and displacement are outlined, followed by a discussion of the findings. The chapter concludes with a demonstration of the proposed method using real data and a discussion of the results.

The proposed linear mixed model framework under the random displacement process is described in **Chapter 4**. This chapter also contains methods for model fitting and variance estimation of the model parameter estimates using the EDC method, accounting for uncertainty due to displacement error. The chapter also includes the proposed bootstrap bias correction (BC) method to obtain bias-corrected estimates of the linear mixed model parameters under displacement. The variance estimation of the estimates of the linear mixed model parameters under the BC method is developed, accounting for random displacement errors. The chapter includes model-based simulation studies

to evaluate the impact of the DHS EA displacement on the linear mixed model estimation and the performance of the proposed methods. Finally, the chapter concludes with a discussion of the simulation study results.

Chapter 5 describes the theory and methods for estimating the effects of spatial covariates in the linear regression model under random displacement. The proposed methodology based on EDC under regression calibration for estimating the expected spatial covariate values is discussed. The chapter also includes a bootstrap bias correction method for correcting the bias of the estimates of the spatial covariate parameters of the linear regression model under the random displacement process. In addition to the point estimates, variance estimation for the estimates of the model parameters is developed under the proposed EDC-RC and BC methods. The proposed methods are evaluated through a model-based simulation study, and the results are presented and discussed.

Finally, **Chapter 6** briefly summarises the thesis and outlines potential directions for future research.

Chapter 2

Literature review

2.1 Introduction

This chapter provides an overview of the preservation of respondents' confidentiality through aggregation and displacement of geo-referenced data and measurement error models. We first discuss recent developments and problems in statistical estimation and inference with aggregated and displaced location data to motivate the methods, and then, their possible methodological extensions are reviewed.

2.2 Confidentiality preservation via aggregation and displacement

Collecting census or survey respondents' household location data is essential for various research and policy purposes, such as understanding mobility patterns, access to services, and health outcomes (Wasfi et al., 2013; Galea and Vaughan, 2019; Bruzelius and Shutes, 2022). However, releasing this exact georeferenced data presents a risk to privacy and confidentiality because identities associated with these locations can be identified through reverse geocoding. It may uncover sensitive information about individuals and households, including their whereabouts, vulnerabilities and activities (Hundepool et al., 2010; Vayena et al., 2015). To address the disclosure risks of georeferenced data, it is crucial to implement measures that preserve confidentiality and protect privacy, such as obtaining informed consent, anonymising the data, and storing it securely (Hundepool et al., 2010; Vayena et al., 2015). Under ethical guidelines and relevant data protection legislation, the geoprivacy of respondents must be protected to ensure that individuals cannot be identified from the released locational information (Kwan et al., 2004; Boulos et al., 2009; Dupriez and Boyko, 2010).

Respondents' confidentiality can be preserved by geo-masking the location coordinates before releasing the data to the researchers. Geo-masking is the process of altering the georeferenced data to reduce the likelihood of identifying individuals when the data is released, while still maintaining the relationship between the location and the occurrence of the phenomena (Allshouse et al., 2010; Hundepool et al., 2012, section 1.2). Aggregation and random displacement of location coordinates of individuals or households are standard and widely used geo-masking techniques by different surveys and censuses of a country to preserve the confidentiality of respondents.

2.2.1 Aggregation

Aggregation is a geo-masking technique that involves grouping respondents' household coordinates together and placing them in the centre of an Enumeration Area (EA) or administrative area, such as postal code, neighbourhood, or district. By aggregating the data in this way, the precise location of each respondent's household is obscured, which helps protect the respondents' privacy. Rounding is a kind of aggregation method, also known as grid masking, that involves rounding or truncating the coordinates of data points to a specified number of digits or decimal places (Burgert et al., 2013). This results in the data being concentrated on a grid of points (Groß et al., 2017). The range of likely geographical or aggregation error relies on the number of digits; for example, if the coordinates of data points are rounded to three decimal places, each point will be located at the intersection of a grid with intervals of 0.001 units (Burgert et al., 2013). Rounding is commonly used in geographic information systems (GIS) and other spatial analysis applications to protect the privacy of individuals (Ajayakumar et al., 2019). Many surveys or censuses conducted by different countries collect households' location coordinates and apply aggregation techniques to protect privacy before releasing the data. For instance, according to the documentation (Statistics Canada, 2011), the Canadian Community Health Survey (CCHS) executed by Statistics Canada collects data on healthcare utilisation, health condition and health determinants of the Canadian population. It also collects georeferenced data, i.e., geographic coordinates for each respondent's household. In order to protect respondent privacy, geographic coordinates are aggregated into larger geographic areas, such as health regions, to ensure that no individual or household can be identified from the survey data (Wang et al., 2020). Some other examples of the surveys that endorse the aggregation approach are the International Household Survey Network (IHSN) conducted in low-and middle-income countries, UNICEF's Multiple Indicator Cluster Surveys (MICS) and Demographic Surveillance System (DSS) conducted by the United States Centers for Disease Control and Prevention (CDC) (for details, see Burgert et al., 2013). The geographic aggregation method is also used by the Demographic Health Surveys (DHS) in their first step of disclosure control,

where all household locations within an EA are aggregated and placed at the centroid of the EA (Burgert et al., 2013). Although this reduces the disclosure risk, aggregation induces a loss in data resolution and decreases the performance of spatial pattern detection (Cassa et al., 2006; Hampton et al., 2010). This is a trade-off when using geographic aggregation to preserve respondent confidentiality and protect privacy. Therefore, appropriate statistical methods or modelling approaches need to be considered to account for the loss of data resolution caused by aggregation. We discuss the existing approaches that account for aggregation errors in estimation and any potential further developments in sections 2.3- 2.4.

2.2.2 Random displacement

Random displacement of the respondent or household location coordinates is a geo-masking technique to preserve respondents' confidentiality and protect privacy. The random displacement process displaces the geographic location of an individual or household, typically involving two components: a random distance and a random direction (Allshouse et al., 2010; Seidl et al., 2016).

The random distance is drawn based on a probability distribution, such as a uniform distribution or a normal distribution (in which case, it is referred to as a Gaussian random displacement), and represents the distance that the household will be displaced from its original location. The random direction (or angle) is also generated from a probability distribution, e.g., a uniform distribution. The combination of these two components results in a randomly displaced location for the household (Allshouse et al., 2010; Burgert et al., 2013). Typically, the maximum distance for displacement is determined by considering the density of the local population. This is because when the population density is higher, a smaller distance is needed to preserve the confidentiality of respondents (Zandbergen, 2014). Consequently, confidentiality preservation for respondents in urban EAs requires a smaller displacement distance than that needed for rural EAs.

The displaced location is selected randomly from all potential locations that share similar geographic features, such as residential areas within a pre-specified boundary circle. This process, known as location swapping (Zhang et al., 2017), ensures that the original location is not relocated to uninhabitable areas, for example, large bodies of water. Also, when a displaced location is randomly chosen within a range of minimum and maximum displacement distances, the approach is known as Donut displacement (Hampton et al., 2010). The term originates from the 'donut-shaped' region that is formed between a smaller circle (with a minimum radius) and a larger circle (with a maximum radius) around the original location. This ensures that the displaced location is not too close or too far from the original location.

The random displacement method (the random direction and a random distance) is commonly used by many surveys to protect the confidentiality of survey respondents while preserving the spatial distribution of the data. For example, the World Bank's Living Standard Indicator Survey (LSIS) and the Demographic and Health Surveys (DHS) program conducted by the ICF international use the random displacement process. Within the DHS program, displacement is conducted by randomly selecting a direction uniformly distributed between 0 and 2π, along with a random distance that follows a uniform distribution between 0 and the maximum distance (Karra et al., 2020). The maximum distance is chosen based on the population density of the EA's location (Chao and Colwell, 2017). In particular, urban locations are displaced 0-2 kilometres while rural locations are displaced 0-5 kilometres with 1% of rural ones randomly displaced 0-10 kilometres. All DHS surveys consistently use the same distance parameters for urban and rural areas (Perez-Heydrich et al., 2013). Hereafter, this is referred to as the DHS random displacement process. We have discussed in detail and developed the probability distribution of the DHS random displaced coordinates in Sections 3.2.1 - 3.2.3. To protect respondent confidentiality and mitigate disclosure risk, DHS randomly displaces EA centroids before releasing them to data analysts. This reflects a deliberate trade-off with the precision of location-dependent estimates. The challenges introduced by this lack of precise data aim to enhance the accuracy of estimates without increasing disclosure risk.

The location coordinates of households are aggregated and randomly displaced in many surveys, such as the DHS, to preserve respondent confidentiality. However, the aggregation and displacement process can induce measurement error in the resulting data or estimates that depend on location. Therefore, it is crucial to investigate the impact of the aggregation and displacement error on the estimates and then use appropriate statistical methods accounting for the measurement error to derive better estimates. In particular, two questions need to be addressed: a) what is the impact of aggregation and displacement on statistical estimation and inference? and b) can we derive precise and consistent estimates accounting for the measurement error caused by the aggregation and displacement? We discuss this, review relevant literature, and explore further methodological development in the following sections.

2.3 Measurement error models

Ignoring the error in the georeferenced data due to displacement or aggregation may lead to flawed inference and a misleading conclusion. As the true locations are aggregated or displaced, any location-dependent measures or estimates will contain in measurement error. In such a situation, measurement error models are used to account for the measurement error in the geographic co-

ordinates (Gleser, 1991; Carroll et al., 2006). Measurement error models are used to model the relationship between the true location coordinates and the displaced (or aggregated) location coordinates. Two types of measurement error models are widely used, the classical measurement error model and the non-classical or Berkson measurement error model. A measurement error is considered classical if it is independent of the unobserved true location data; otherwise, it is non-classical (Buonaccorsi, 2010, section 1.4.1; Carroll et al., 2006, section 2.2.2; Deffner et al., 2018). In particular, let $\{Y_1, Y_2, \cdots, Y_N\}$ denote the set of observed outcomes measured at the true household locations $\{T_1, T_2, \cdots, T_N\}$. To preserve the respondents' confidentiality, only the displaced set of household locations $\{W_1, W_2, \cdots, W_N\}$ associated with the outcomes are available. Now, the non-classical or Berkson measurement error model can be expressed as $T_k = W_k + e_k^\star, k = 1, 2, \cdots, N$ where $e_k^\star$ is a random noise. This noise is independent of the observed location W_k, but it depends on the true location T_k. This implies that the error may vary based on the actual location's position, potentially due to issues, such as imprecise measurement equipment (Wilson and Wakefield, 2021). For instance, when the original location is in a region with complex geography, e.g., mountainous areas, obtaining an accurate record of the location becomes challenging. In such scenarios, the observed or measured location might experience errors stemming from the terrain or signal interference. This results in discrepancies between the observed and the true location coordinates. Many past studies have addressed these positional error concerns within the context of a Berkson measurement error model (Fanshawe and Diggle, 2011; Gabrosek and Cressie, 2002). Considering that the processes of displacement and aggregation are independent of the true locations within urban or rural areas, we do not classify this as a Berkson type of error, as explained below.

In contrast, the classical measurement error model can be expressed as $W_k = T_k + e_k, k = 1, 2, \cdots, N$, where e_k is a random noise that is independent of the unobserved true location T_k. Classical measurement errors may arise due to a random displacement process, for instance, the DHS random displacement process, where the random noise e_k is generated through a displacement mechanism independent of the true location coordinates within rural or urban areas. This noise is then added to the true location coordinates to generate the displaced locations. More specifically, if we assume that: 1) the distribution of the random error e_k is known, and 2) within each location type (urban or rural), the error term e_k is independent of the true location T_k, then we can consider the DHS random displacement process as a classical type of error. We have discussed it in detail in Section 3.2 of Chapter 3 of this thesis. Previous studies that have obtained statistical estimates accounting for displacement error under the classical measurement error model are outlined in Sections 2.4 - 2.7.

When using survey data along with location coordinates of households, researchers target to obtain different spatial statistics depending on the policy needs. For example, they may need to obtain area-specific density estimates of sub-populations, sub-national level estimates of means of a variable of interest, e.g., income, or the estimate of the impact of the distance between a survey EA and the nearest health facility to study the health outcomes. However, to preserve respondent confidentiality, the true location coordinates of individuals or households are aggregated or displaced, and only the aggregated or displaced data is provided to the data analysts. This introduces measurement error, and suitable statistical methods are required to obtain estimates accounting for the aggregation and displacement error, depending on the nature of the statistical estimation problem.

In the following sections of this chapter, we review the relevant literature and discuss different statistical estimation problems and their potential methodological extensions to address the measurement error induced by aggregation and displacement.

2.4 Density estimation with both aggregated and randomly displaced data

Precise estimates of area-specific densities of target sub-populations of interest, for instance, poor households, obtained through a survey or register data, can help identify poor household hotspots and develop policies and interventions to combat poverty in these areas. Density estimation through kernel methods has received much practical attention in the literature. Kernel density estimation (KDE) is a popular non-parametric method used to estimate the probability density function of a random variable based on observed data. One of its significant features is its ability to produce smooth density estimates, effectively visualizing data distributions (Duong, 2007). While KDE has applications across fields, such as statistics, data science, and machine learning, it is also prominently used in georeference-based point pattern analyses. For instance, in health geography and disease mapping, KDE plays a role in the identification of Enumeration Areas (EAs) (Fotheringham and Zhan, 1996; Rushton and Lolonis, 1996). For an overview, for example, see Simonoff and Simonoff (1996).

The univariate KDE is used when we observe a single random variable based on a set of observations. On the other hand, the bivariate KDE method is used to estimate the joint probability density function of two random variables. For instance, when we observe spatial (continuous) coordinates (latitude and longitude) for each observation, such as poor household, the bivariate KDE can be used to estimate the density of the joint distribution of the latitude and longitude coordinates. This

can provide a better insight into the spatial distribution of the data, such as poor households, by creating a density map showing the areas where poverty is most prevalent. We further discuss this and define mathematically the bivariate KDE of the joint distribution of spatial coordinates of observations in Section 3.3.

The selection of bandwidth, also known as a smoothing parameter, is essential for the performance of a bivariate kernel density estimator. In particular, the bandwidth parameter is a tuning parameter which controls the degree of smoothing in the density estimate. Thus, optimal values of bandwidth parameters need to be chosen. Larger values of the bandwidth parameters may lead to over smoothing the data, while small values lead to under smoothing. In bivariate kernel density estimation for estimating density of the distribution of spatial coordinates of observations, which is one of our interests in this thesis, there are different methods available to choose the values of the bandwidth parameters. Some commonly used methods are rule-of-thumb, and plug-in or cross-validation methods (Wand and Jones, 1994). One simple and easy to use rule-of-thumb method is Silverman's rule of thumb method (Dehnad, 1987). In Silverman's rule, the bandwidth parameter values are determined by using the sample size and the number of dimensions, which is two in case of bivariate density estimation (for details, see Dehnad, 1987). Another rule-of-thumb method is Scott's rule (Scott, 2015, section 6.2). The rule-of-thumb methods are simple and computationally efficient, but not appropriate to provide the best estimate in all cases. Cross-validation or plug-in methods are more sophisticated and provide more accurate estimates than the rule-of-thumb methods, but are computationally intensive. Wand and Jones (1994) described the selection of the bandwidth in the multivariate situation by using a plug-in estimator. In the plug-in or cross-validation method, the data are divided into training and validation sets. The strategy is to choose the bandwidth parameter that minimises a cross-validation criterion, for instance, the asymptotic root-mean integrated squared error (RMISE) or the leave-one-out cross-validation (LOOCV) criterion (Silverman, 2018, section 2.4).

Apart from selecting the optimal values of the bandwidth parameters, the precision and smoothness of the density estimates of observations depend on the availability of the exact coordinates of observations. When data is contaminated with aggregation or rounding error, it can result in a spiky density that deviates from the density of the actual location data (Groß et al., 2017, 2020). Detailed spatial information about the distribution of the true data is lost when the true coordinates are aggregated to a coarser resolution level, leading to bias in density estimates. A higher level of aggregation error leads to more biased density estimates (Groß et al., 2017, 2020).

Scott and Sheather (1985) used naive methods that ignore the rounding error to estimate the den-

sity. Several studies (Härdle and Scott, 1990; Scott, 1985; Minnotte, 1998; Silverman, 1982; Wand, 1994) have proposed using the weighted averages rounded points (WARPing) to estimate kernel density accounting for rounding error. If the data are aggregated into a coarser level, such a method becomes less efficient, as the approximation becomes gradually worse as the level of aggregation increases. In the univariate case, Wang and Wertelecki (2013) developed a parametric and a non-parametric KDE for data with rounding errors. A recent work by Groß et al. (2017), developed a methodology for deriving multivariate non-parametric kernel density estimates of sub-populations of interest in the presence of rounding error under a measurement error model. In particular, Groß et al. (2017) formulated a measurement error model for rounded coordinates that we describe below:

We first define the notations. Let $T_{jk} = (T_{jk1}, T_{jk2})$ represent the true coordinates of the observation, such as the longitude and latitude of the kth $(k = 1, \cdots, N_j)$ unit, which can be a household or individual e.g., poor household/people, in the jth $(j = 1, \cdots, R)$ EA. In this context, we are dealing with a bivariate variable. Let $f(T^\star)$ denote the unknown density of the true (continuous) coordinates of the observations at a specific point $T^\star = (T_1^\star, T_2^\star)$, where $T_1^\star$ and $T_2^\star$ indicate the longitude and latitude, respectively. However, the true coordinates T_{jk} are not available to the data analysts. In cases of only aggregation (or rounding), it is assumed that all households within the jth EA are located at its centroid, represented by $T_j = (T_{j1}, T_{j2})$ for $j = 1, \cdots, R$. Throughout the thesis, when considering both aggregation and random displacement, the aggregated and displaced coordinate of the jth EA is denoted by $W_j = (W_{j1}, W_{j2})$.

Under only aggregation, instead of the true coordinates T_{jk}, data analysts have access only to the rounded (or aggregated) coordinates $T_j = (T_{j1}, T_{j2})$ for $j = 1, \cdots, R$. As a result, density estimation using T_j may lead to spiky estimates depending on the degree of the rounding error. To draw pseudo-samples of the unknown T_{jk} and assuming that the rounding process is known, Groß et al. (2017) formulated the conditional distribution of T_{jk} given T_j using Bayes rule, expressed as:

$$f(T_{jk} \mid T_j) \propto f(T_j \mid T_{jk}) f(T_{jk}), \tag{2.1}$$

where $f(T_j \mid T_{jk})$ is the conditional distribution of the rounded coordinates given the true coordinates of the observations. This distribution can be defined based on the type of EA geography used, such as a square grid. Considering a square grid with side length denoted by γ, the distribution

$f(T_j \mid T_k)$ is given by:

$$f(T_j \mid T_{jk}) = \begin{cases} 1 & \text{for } T_{jk} \epsilon [T_{j1} - \frac{\gamma}{2}, T_{j1} + \frac{\gamma}{2}] \times [T_{j2} - \frac{\gamma}{2}, T_{j2} + \frac{\gamma}{2}], \\ 0 & \text{otherwise.} \end{cases} \tag{2.2}$$

Since both the true data T_{jk} and $f(T_{jk})$ are initially unknown, Groß et al. (2017) proposed a Stochastic Expectation-Maximization (SEM) algorithm for estimation. This algorithm uses an initial estimate of $f(T^{\star})$ based on T_j, followed by alternating simulations of T_{jk} from $f(T_{jk} \mid T_j)$ and re-estimation of $f(T_{jk})$ until convergence is reached. This is considered as a modified version of the expectation-maximization (EM) algorithm (Dempster et al., 1977). The reason for using the SEM algorithm over the classical EM algorithm is that in the classical EM algorithm, the conditional-expectation of T_{jk} given T_j is computed analytically in the E-step, which leads to spiky density estimates as the data are concentrated around the grid centroid due to rounding. The estimator proposed by Groß et al. (2017) allowed the estimation of the bandwidth matrix within the density estimation framework. We refer to the publication by Groß et al. (2017) for the computational details of the proposed method and further details. In the case of aggregated coordinates of observations at the centroid of the higher-level administrative area, Groß et al. (2020) extended the methodology of Groß et al. (2017) by replacing the distribution of T_j given T_{jk} under aggregation. In particular, Groß et al. (2017) generated the simulated coordinates of T_{jk} from the conditional distribution of T_{jk} given T_j using kernel density estimates under aggregation and dis-aggregated the cases from the higher to lower (or non-hierarchical) administrative geography. Using simulation studies, Groß et al. (2020) showed that the proposed SEM algorithm-based method outperforms the naive method that assumes a uniform distribution of the units in the grids instead of re-estimating the densities.

Under random displacement of spatial coordinates, the density estimates that ignore the displacement error may be inaccurate and biased. The random displacement process moves the original location from one place to another, which may alter the original spatial pattern. Shi et al. (2009) investigated the spatial pattern by creating kernel density surfaces with random displaced location data under a series of bandwidths and found that the density surfaces of the true points and the displaced points differ significantly in the case of lower bandwidth. Moreover, the spatial pattern changes more when a greater maximum displacement distance is applied to the true coordinates. However, it is not possible to use directly the method proposed by either Groß et al. (2017) under rounding or Groß et al. (2020) under geographical aggregation to estimate the kernel density ac-

counting for displacement error- it requires methodological extensions. The main reasons for this are a) the aggregation or rounding process is deterministic, while the displacement process with varying maximum displacement distances for geographical regions is stochastic, and b) the conditional distribution of aggregation or rounding error is uniform, which is not the case regarding the random displacement algorithm- involves a random distance and a random angle. Moreover, we can draw better inferences for disclosure preservation under the random displacement process than the deterministic rounding or aggregation process. This is because the random displacement process incorporates the statistical property of randomness. Therefore, the possible methodological extensions for multivariate KDE under the random displacement process are described in Section 3.3, which is one of the novel contributions to the literature of this thesis.

Both aggregation and random displacement are sometimes applied to the true coordinates to protect respondents' privacy in the survey data. For instance, in the DHS, the household locations data within the same EA are first aggregated to a single coordinate, such as the EA centroid. Next, the EA centroid coordinates are displaced using the DHS displacement algorithm. So, only the aggregated and then displaced EA centroid coordinates are provided to researchers. The aggregation and displacement processes induce measurement error in spatial data and that may alter the original spatial pattern (Hampton et al., 2010). Therefore, the kernel density estimates under aggregation and displacement may be biased and differ from the true estimates. To the best of our knowledge, previous studies did not attempt to obtain kernel density estimates by correcting the aggregation and displacement errors under a measurement error model. Therefore, we aim to propose a method to obtain kernel density estimates of the joint distribution of the spatial coordinates of observations accounting for aggregation and displacement error in Section 3.5. This makes a significant and original contribution to the existing literature.

2.5 Spatial statistics under random displacement

Apart from the density estimates of the sub-populations of interest, the random displacement process may affect any other location-dependent or spatial measures of the variables of interest, which may be used in further statistical analyses. For example, distance calculations to the nearest points of interest, e.g., in health facility, or the evaluation of environmental conditions by using displaced coordinates may differ from the results based on the true coordinates. Previous studies (for instance, Feldacker et al., 2010; Lohela et al., 2012) ignored this displacement error and calculated spatial variables, e.g., the distance from a DHS EA location to the nearest resource location. These types of statistical measures of variables can potentially be inaccurate since calculating distances

using displaced locations results in measurement errors (Warren et al., 2016a).

The effect of the EA point random displacement on spatial analyses was studied by (Gething et al., 2015). Gething et al. (2015) used 100 randomly displaced sets of EA location data obtained by applying the DHS random displacement process to the real EA location data to assess the displacement effect on analyses, including several socioeconomic indicators of interest. The impact of displacement on the spatial correlation among data points, environmental covariate relationships and model-derived interpolated surfaces was explored. In the case of the spatial correlation, for each displaced dataset, the empirical variogram was calculated, and then they were combined and compared to the non-displaced variogram. A little effect on the spatial correlation was found. Nevertheless, some differences were found in the relationship between the outcomes and spatial covariates (e.g., population density, human settlement density, distance to the road), which were extracted at the EA centroids. In particular, models with spatial covariates values extracted using displaced EA centroids have a lower predictive power (R^2) than those that used true EA centroids, especially for the case with covariates, which have a strong relationship with the outcomes based on true EA centroids. Moreover, differences to some extent in the model-derived predicted surfaces were observed when using displaced EA coordinates compared to the true EA coordinates.

According to work by Perez-Heydrich et al. (2016) and Gething et al. (2015), the DHS has suggested guidelines to address potential bias in extracting spatial covariate values at the EA centroid resulting from the random displacement process. In particular, the guideline involves using the average value of the covariate in grid cells within a fixed buffer (maximum displacement distance) of the displaced EA centroid in analyses. However, a simple average of the grid cell values indicates that all points within a maximum displacement buffer were considered equally likely to be the true EA point. This is not the case in the DHS random displacement process, which we discuss in section 3.2 by deriving the conditional distribution of the displaced coordinates given the true coordinates.

A recent study by Grace et al. (2019) proposed a method to address potential errors in generating environmental and contextual variables at the DHS EA centroids due to the random displacement process using remotely sensed imagery. In particular, the study suggests selecting a settlement (such as housing structures) near the DHS displaced EA point and measuring the environmental conditions around it, which is the likely true point, using a buffer smaller than the DHS maximum displacement distance. It was found that using a smaller buffer around the settlement produces less error in estimates compared to the buffer with the DHS maximum displacement distance around the displaced point. The main challenge with this method is finding suitable settlements (for instance,

road networks, housing structures, and path networks) that are dependent on the availability of high-quality images, which can be time-consuming. Nevertheless, the authors hypothesised that a short, precise buffer around a wrong settlement provides a more accurate measure of the truth than a big buffer around the displaced point. This is because even though the settlement utilised to provide contextual information may not be the original DHS EA, it is a neighbour of the EA, and neighbouring settlements are usually more similar.

Estimates of demographic and health indicators at the sub-national level based on survey data are crucial for targeting interventions and making policies. However, any estimates at an administrative level lower than the displacement designated administrative boundary (referred to as a domain) may be biased due to the random displacement process. This is because an EA can be placed outside of its original domain area due to the random displacement process, referred to as an EA misplacement. Previous studies in the literature have paid less attention to the domain estimates under a misplacement error caused by the random displacement process. The study by Wilson et al. (2020) aimed to obtain estimates of some commonly used health indicators at the district level (admin 2) in Malawi accounting for the DHS EA displacement error. To address the displacement error, the method by Wilson et al. (2020) calculated the proportion of points generated by displacing the displaced point, which falls within each district boundary intersected with the maximum distance displacement buffer around the displaced point. District estimates of the health indicators were obtained by assigning a) a single district with the greatest likelihood to an EA, b) multiple districts in proportion to their likelihoods, and c) using the naive method that ignores the displacement error. It was found that the district estimates by each method approximate the true estimates (not affected by the displacement process) by comparing the results using confidence intervals and the concordance correlation coefficient. Thus, to obtain district estimates, Wilson et al. (2020) recommended using the simplest naive method that ignores the misplacement error due to the random displacement process. Nevertheless, the inference may differ for the estimates at a level lower than the district level, as more misplacement errors are expected due to the random displacement process. In chapter 3, this thesis aims to address this literature gap by obtaining estimates of interest at a lower administrative level under the different intensities of displacement error.

Researchers are often interested in linking variables (such as poverty levels, population density, and socioeconomic indicators) available at the domain level from an external data source to the EA by using the displaced EA centroids. However, when the EA is misplaced due to the random displacement process, there is a chance for inaccurate variable assignment to the EA, which could

affect the subsequent statistical analysis. Furthermore, the likelihood of EA misplacement errors and variable mis-assignment errors increases when linked with smaller administrative boundaries or geographies. Several studies (for example, those conducted by Balk et al., 2004, Pande et al. (2008), and Feldacker et al. (2010)) have assigned domain variables to the DHS EA centroid while ignoring the misplacement error due to the random displacement process. In order to reduce errors associated with assigning domain variables to the DHS EA, Warren et al. (2016b) suggested a method called the maximum probability covariate (MPC). This method involves calculating the probabilities for each unique domain value within the displacement buffer around the displaced EA centroid, and then assigning the most probable domain to the DHS EA. In particular, Warren et al. (2016b) estimated the domain probabilities by calculating the proportion of likely true points within each domain boundary intersecting with the displacement buffer. The likely true points were generated by displacing the displaced coordinates using the DHS displacement mechanism, i.e., through the conditional distribution of displaced coordinates given the true coordinates expressed by $f(W_k \mid T_k)$ in (3.4). However, Warren et al. (2016b) did not present the mathematical form of $f(W_k \mid T_k)$, which is one of the contributions of this thesis. In addition, Warren et al. (2016b) did not consider $f(T_k)$ in their process and allowed all points within a buffer to be possible true points, but the points in large water bodies e.g., sea, lake and uninhabited areas cannot be possible true locations and have a selection probability equal to 0 (Figure 3.5). Also, the MPC selection method is not a complete solution and can result in biased estimates when the correct domain value is selected with a probability of less than 1. Another drawback of this method is that the domain value assigned to an EA point is deterministic, meaning that only the most probable domain is assigned to the EA location as the correct domain. Therefore, it is possible that the true EA point may belong to another domain when the probability of the selected domain value is less than 1. Moreover, the underlying distribution of true locations, denoted as $f(T_k)$, may play an essential role in improving the accuracy of domain selection probabilities, especially if it is related to the potential true locations. Therefore, this thesis aims to explore possible extensions of the method proposed by Warren et al. (2016b) by incorporating additional information about $f(T_k)$ from external data sources. We provide a detailed description of our approach in Section 3.4 of Chapter 3.

2.6 Mixed models under random displacement

Multilevel models, also known as mixed effects models or hierarchical linear models, are a type of statistical model used to analyze data with a hierarchical or clustered structure (Melamed and Vuolo,

2019). In a multilevel model, the observations are grouped into different levels, each representing a separate analysis unit. For example, in a study of individuals in different Enumeration Areas (EAs) based on the Demographic and Health Surveys (DHS), where individuals are nested within the EAs, and EAs are nested within the administrative boundaries of the country. This hierarchical structure highlights the necessity of employing multilevel regression modelling to account for the nested nature of the data and obtain accurate results (Islam, 2005; Alom et al., 2012; Imam et al., 2018; Bhuiyan et al., 2020). Moreover, when individuals' responses over time are correlated, as in longitudinal studies, multilevel modelling is used to analyze the data. Multilevel models allow for the estimation of both within-group and between-group effects. Specifically, the error variance is partitioned into a between-group component (representing the variance of the EA-level errors) and a within-group component (representing the variance of the individual-level errors). These models can account for the similarity of observations within each group, which may differ from observations in other groups.

In multilevel models, there are two types of effects: random and fixed. The fixed effects help estimate the overall relationship between the independent and outcome variables using data from all groups in the analysis. On the other hand, the random effects allow for modelling the variability between groups and accounting for differences in the data due to group-level factors. In particular, random effects vary across groups and are specific to each group in the analysis. They represent a particular group's deviation from the independent variable's overall average effect on the outcome variable. Also, random effects are estimated using mainly data from the relevant groups in the analysis (see, for an overview of multilevel modelling Leyland and Goldstein, 2001; Leyland and Groenewegen, 2020, chapter 3; Fielding et al., 2003; Gelman and Hill, 2006, part 2A).

The random effects in multilevel models are specified at the group level to capture the grouping structure of observations. For example, in the DHS data, these groups can be defined by EAs or the country's administrative units. Without *exact* or accurate information on observation groups, observations can be mixed or misclassified among groups, compromising the precision of group-level variation estimates. The term *exact* refers to information obtained based on true location coordinates of observations, which may be unknown to the data analysts and displaced to preserve the respondents' confidentiality. In the case of DHS data, the true DHS EA centroids are displaced within the designated higher administrative boundary, e.g., admin 2 the district of the country, using the random displacement mechanism and only displaced EA centroid coordinates along with the survey EA identifiers are given in the survey data (Burgert et al., 2013). Therefore, displacement is not an issue for the use of multilevel models for the DHS data with the random effect specified

at the EA level (for example, see Kasaye et al., 2019; Ogbo et al., 2018; Gidado, 2012) or at the displacement-restricted higher administrative level (for instance, see Smith and Shively, 2019; Das et al., 2020a).

However, if the random effect in the mixed models is specified at below the displacement restricted administrative level, such as the sub-district level of the country, where EAs can be misplaced due to a random displacement process, the estimates of multilevel model parameters, especially the random variance components, and associated model-based estimates of finite population parameters (e.g., group/sub-district means) may be biased. A few studies have used DHS data to fit multilevel models with random effects specified at the sub-district level (for instance, Das et al., 2019a,b, 2020b), ignoring the DHS EA misplacement error uncertainty in the estimation methods. These naive estimates ignoring the misplacement error may mislead the inferences. This emphasises the need to investigate the effect of misplacement due to random displacement on the estimation of multilevel models, e.g., the linear mixed model (LMM) and hence, to develop methods for model fitting and uncertainty estimation of the estimates accounting for misplacement error. This thesis aims to contribute by addressing these gaps in the literature. We have discussed this and proposed methods in Chapter 4.

2.7 Regression with spatial covariates under random displacement

Researchers are often interested in examining how spatial covariates, e.g., the distance from the Enumeration Area (EA) centroid to the nearest health facility of the country, affect the outcome of interest in regression models. For example, Feldacker et al. (2010) studied whether HIV infection in rural Malawi is related to the distance from the Demographic and Health Survey (DHS) EA centroid to major roads, health clinics and major cities. The DHS true EA coordinates are not available to the data analysts and are displaced using the displacement algorithm to protect the respondents' confidentiality. Therefore, the statistical analyses with a spatial covariate that is constructed using the randomly displaced location can potentially be inaccurate, and the random displacement process induces a measurement error in covariate problems in regression models, e.g., a linear regression. In previous studies, for example, those by Feldacker et al. (2010) and Lohela et al. (2012), randomly displaced EA locations were used to calculate spatial covariates (e.g., distance) at the EA points, ignoring the displacement error and investigating how these spatial covariates influenced the outcome of interest at the EA level.

A measurement error in a covariate in a regression model leads to a biased estimate of its coefficient

(Aigner, 1973; Carroll et al., 2006). In order to form the problem mathematically, let T_k denote the true location coordinates of the kth $(k = 1, \cdots, N)$ unit. The unit of the analysis can be the individual, household or Enumeration Area (EA). The true value T_k that can be represented as $T_k = (T_{k1}, T_{k2})$, where T_{k1} and T_{k2} denote the exact geographical longitude and latitude coordinates of the kth unit. Let $X_{1k}(T_k)$ be the value of a spatial covariate which is constructed or measured based on the true location coordinates T_k. We aim to estimate a relationship described by a linear regression model formulated as follows:

$$y_k = \beta_0 + \beta_1 X_{1k}(T_k) + \varepsilon_k, \tag{2.3}$$

where y_k is the outcome, β_0 is the unknown intercept, β_1 is the unknown coefficient of the spatial covariate $X_{1k}(T_k)$, and ε_k is a random-error term which is independent of $X_{1k}(T_k)$. In what follows, we suppose, for simplicity, that the random errors ε_k in (2.3) are independent and identically distributed with mean 0 and constant variance σ_ε^2. The unknown parameters of the model (2.3) are denoted by $\theta = (\beta_0, \beta_1, \sigma_\varepsilon^2)$. The true spatial covariate values $X_{1k}(T_k)$ in (2.3) are unknown to the data analysts, as the true coordinates T_k are unknown. Only the displaced (contaminated by measurement error) coordinates denoted by $W_k = (W_{k1}, W_{k2})$ are available to data analysts. Let $X_{1k}(W_k)$ be the value of a spatial covariate which is constructed based on W_k. Therefore, the use of $X_{1k}(W_k)$ instead of $X_{1k}(T_k)$ in the model (2.3) can be considered a measurement error in the covariate problem, and the regression model can be expressed as follows:

$$y_k = \beta_0 + \beta_1 X_{1k}(W_k) + \varepsilon_k. \tag{2.4}$$

When fitting the model (2.4) using ordinary least squares (OLS) and obtaining the estimated model parameters denoted by $\hat{\theta} = \left(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}_\varepsilon^2\right)$, these naive estimates of the model parameters, particularly the estimates of the regression coefficient β_1, will be biased due to random displacement (Carroll et al., 2006; Warren et al., 2016a; Karra et al., 2020).

Various methods have been proposed in the literature to deal with the general issue of measurement errors in covariates, including maximum likelihood methods (Rabe-Hesketh et al., 2003), Bayesian Markov chain Monte Carlo algorithms (Goldstein and Shlomo, 2020) and regression calibration (Hardin et al., 2003; Spiegelman et al., 1997). In particular, the regression calibration (RC) approach replaces the spatial covariate values measured with errors by their expected values given the displaced locations to obtain unbiased and consistent estimates (Carroll et al., 2006; Karra et al., 2020).

Studies by Warren et al. (2016a) and Karra et al. (2020) focused on the subsequent statistical analysis of interest that includes the distance covariates and the potential impact of the EA location displacement on the bias of the resulting distance covariate parameter estimates. Warren et al. (2016a) claimed that the given (observed) distance covariate $X_{1k}(W_k)$ follows the classical measurement error model (MEM) form as proposed by Carroll et al. (2006) and is represented as follows:

$$X_{1k}(W_k) = X_{1k}(T_k) + u_k, \tag{2.5}$$

where u_k is the random measurement error. It was assumed that u_k is independent and identically distributed for each EA location scenario (e.g., urban or rural) and follows a normal distribution with $E(u_k) = 0$ and $\text{Var}(u_k) = \sigma_u^2$. Then, Warren et al. (2016a) applied the standard RC method (Carroll and Stefanski, 1990) to estimate the expected value of $X_{1k}(T_k)$ given $X_{1k}(W_k)$ expressed as

$$E\left\{X_{1k}(T_k) \mid X_{1k}(W_k)\right\} = \mu_t + \frac{\sigma_t^2}{\sigma_t^2 + \sigma_u^2}[X_{1k}(W_k) - \mu_t], \tag{2.6}$$

where $\mu_t = E\left\{X_{1k}(T_k)\right\}$ and $\sigma_t^2 = \text{Var}\left\{X_{1k}(T_k)\right\}$ and $\sigma_u^2 = \text{Var}(u_k)$ in (2.5). As the values of $X_{1k}(T_k)$ are unknown, μ_t, σ_t^2 and σ_u^2 are unknown and need to be estimated. The estimates of μ_t and σ_t^2 are obtained by calculating $\hat{\mu_t} = \frac{1}{n}\sum_{k=1}^{n} X_{1k}(W_k)$ and $\hat{\sigma}_t^2 = \frac{1}{n-1}\sum_{k=1}^{n}\left\{X_{1k}(W_k) - \hat{\mu_t}\right\}^2 - \hat{\sigma}_u^2$. The $\hat{\sigma}_u^2$ is computed based on a set of validation data that are obtained by displacing the displaced coordinates W_k under the DHS displacement process (for details, see Warren et al., 2016a). Finally, the unknown $X_{1k}(T_k)$ in the model (2.3) were replaced by the estimates of their expected values from (2.6), and the model was fitted using the standard least square method to obtain the estimates of the model parameters. The results were compared to the naive method that ignores the displacement error. However, the assumptions of independent and normal distribution for u_k in (2.5) were criticised in the work of Karra et al. (2020), Arbia et al. (2015) and Elkies et al. (2015), questioning their validity under the random displacement process.

As an alternative to the Warren et al. (2016a) method, Karra et al. (2020) eliminated the normality assumption and estimated the conditionally expected distance $E\left\{X_{1k}(T_k) \mid X_{1k}(W_k)\right\}$ in (2.6) based on a numerical integration by drawing sample locations over all possible true locations within the maximum distance displacement buffer around the displacement point. In particular, the likely true locations were drawn from the conditional distribution of true coordinates given the displaced coordinates, which can be expressed as $f(T_k \mid W_k) \propto f(W_k \mid T_k)f(T_k)$. Karra et al. (2020) compared the performance of their method for correcting the bias of the distance covariate esti-

mates over the naive method that ignores the displacement error through a simulation study. It can be noted that the Karra et al. (2020) method's performance was not assessed by applying the exact DHS random displacement process, which allows varying displacement parameters for rural and urban locations, with the displacement being restricted within a designated administrative boundary. However, in the simulation study by Karra et al. (2020), the form of the distribution of $f(T_k)$ in the model was not essential as the coordinates of population observations were generated based a uniform distribution. While Karra et al. (2020) applied their proposed method to real data from a project survey with known true household coordinates and sourced $f(T_k)$ from the WorldPop unconstrained population density map, the methodology for obtaining $f(T_k)$ was not explicitly detailed. This lack of detail could pose challenges for external researchers aiming to replicate the method. In addition, Karra et al. (2020) method did not consider any boundary corrections, for example, the points in large water bodies e.g., sea, lake and uninhabited areas cannot be possible true locations and have a selection probability equal to 0 (Figure 3.5). Therefore, to improve the selection probabilities for likely true locations, the underlying distribution of true locations can be approximated by using available information (e.g., population density, designated administrative boundary restriction, rural urban boundaries, boundary corrections for non-settlement places) from different external data sources. We have discussed this and developed a flexible framework in Section 3.4 for approximating the distribution using information from multiple external sources that improve the selection probabilities for likely true locations. We have used this framework to propose a method in Section 5.2.1 for estimating the effect of spatial covariates under the random displacement process.

2.8 Bootstrap bias correction under random displacement

As discussed in Sections 2.6 and 2.7, the estimates of location-dependent model parameters, specifically a) random variance components in multilevel models when the random effect is specified below the random displacement restricted administrative level and b) regression coefficients for spatial covariates in regression models, are biased under the random displacement process. Therefore, an attempt can also be made to obtain the estimates of the bias, followed by developing a bias-corrected estimator of the model parameters. In particular, one could potentially think about bootstrap bias correction methods under the random displacement process. To the best of our knowledge, none of the previous studies attempted to use a bootstrap bias correction method for correcting the bias of the model parameters under the random displacement process.

The bootstrap technique is a resampling method that samples with replacement from the original

sample and is used to estimate the bias of a (biased) estimator. Modern computers have allowed statisticians to use bootstrapping as a powerful tool for making statistical inferences and estimating the accuracy of sample statistics concerning the population parameter (Higgins, 2004, section 8.1; Chen, 2018, p. 10). Bootstrapping is an application of the plug-in principle, which involves using the sample to estimate parameters for the population (Efron and Tibshirani, 1992, p.35). In the case of bootstrapping, the existing sample data is treated as if it were the population, and samples are taken out of the existing dataset as if to sample from the population (Efron and Tibshirani, 1992, p.139). This approach enables researchers to estimate the distribution of the statistic of interest and to obtain valuable information about the sampling variability of the statistic (Higgins, 2004).

In the context of the linear regression model parameters under random displacement, the bootstrap technique involves repeatedly resampling the original dataset with replacement according to the model (2.4). The observed (displaced) points are then displaced using the random displacement algorithm to generate bootstrap samples. The model parameters are estimated using the same method as in the original sample dataset for each model-based bootstrap sample dataset. The difference between the average estimates over bootstrap samples and the estimates from the (given) original sample provides an estimate of the bias in the original estimates (for an overview, see section 10.2 of Efron and Tibshirani (1992)].

Regarding the bias of the model parameters under displacement, the bootstrap bias estimation method should work to satisfy the two important assumptions. The first assumption is that the functional form of the underlying model, e.g., the linear regression model (2.3) is appropriate to the data. The second but most crucial assumption is that the magnitude of the bias of the naive estimates that ignore the displacement error of the model parameters under displacement is the same as the magnitude of the bias of the estimates under bootstrap samples and repeated displacement process to the displaced locations.

The bootstrap samples are generated under the linear regression model given in Equation (2.4) using the naive estimates $\hat{\theta}$ that ignore the displacement error. Regarding the generation of model-based bootstrap samples, either a parametric or non-parametric bootstrap approach can be used (Efron and Tibshirani, 1992, p. 297). Depending on the model assumptions, we have described the process of generating a parametric bootstrap sample under the linear mixed model in Section 4.3 and a non-parametric bootstrap sample under the linear regression model in Section 5.2.2. When a parametric distribution of the random error, such as the one assumed in the linear mixed model, is used, the bootstrap random errors can be generated under the assumed distribution. We have discussed this process in Section 4.3. In the case of the non-parametric bootstrap, the bootstrapped

random errors (residuals) of the model are obtained by resampling with replacement from the estimated random errors of the model in Equation (2.4). Although this approach does not assume a parametric distribution of the model errors, it implicitly assumes that the functional form of the underlying model is appropriate for the data. In particular, the model errors are assumed to be independent and have constant variance. However, under the linear regression model (2.4), for each bootstrap sample $b; b = 1, 2, \cdots, B$, we estimate the model parameters using the same method as in the original sample and denote the resulting estimate as $\hat{\theta}^{*b}$. Next, we calculate the bootstrap bias estimate by taking the difference between the average of the bootstrap estimates and the naive estimate: $\widehat{\text{Bias}}\left(\hat{\theta}, \theta\right) = \sum_{b=1}^{B} \hat{\theta}^{*b}/B - \hat{\theta}$. Finally, we obtain the bootstrap bias-corrected estimator by subtracting the estimated bias from the naive estimate: $\hat{\theta}_{\text{cor}} = \hat{\theta} - \widehat{\text{Bias}}\left(\hat{\theta}, \theta\right)$. To implement the bootstrap bias correction method, a large number of bootstrap samples B are typically needed to obtain a stable estimate of the bias. Specifically, the number of bootstrap samples required for accurate estimation depends on the size and complexity of the dataset, as well as the available computing and processing power. Efron (1979) suggests that a value between 200 and 500 bootstrap samples is generally sufficient for most applications. Nevertheless, with modern computing capabilities, many researchers use 1000 or more bootstrap samples to improve the accuracy of the estimates.

Chapter 3

Density and domain parameter estimation under displaced and aggregated geo-referenced data

3.1 Introduction

The aggregation and displacement of location coordinates induce measurement errors in data, which may alter the original spatial pattern. Moreover, a location coordinate can be misplaced when it is displaced outside of its original domain boundary. A domain is an administrative unit whose geographical boundary is nested within the displacement-restricted administrative boundary. Ignoring aggregation and displacement errors may lead to inaccurate density and domain (parameter) estimates. Therefore, in this chapter, we aim to develop methods to obtain 1) density estimates corresponding to the distribution of the location data and 2) estimates of the domain parameters under a measurement error model. We illustrate the theory using the DHS aggregation and random displacement process; however, other aggregation and random displacement processes can be used, such as the ones described in Section 2.2 of the literature review. In this chapter, first, we focus on the issue of the random displacement of location coordinates. Next, we focus on the combined aggregation and displacement issue.

The organisation of this chapter is as follows. In Section 3.2, we formally state the random displacement problem and propose a measurement error model, followed by developing the target probability distribution of the unknown true coordinates given the displaced coordinates. Section 3.3 describes the proposed estimation method based on Multivariate Kernel Density Estimates

(KDE) as the marginal distribution of the true location data, accounting for displacement error to obtain density estimates and domain parameters estimates. An alternative to the KDE-based method, which is based on external data sources to approximate the underlying distribution of the true location data, is presented in Section 3.4. Section 3.5 presents the proposed method to obtain density and domain parameter estimates accounting for both aggregation and random displacement errors. Section 3.6 outlines simulation designs for analysing register data under a) only displacement and b) both aggregation and displacement and then discusses the findings of the simulation study. Section 3.7 shows the application of the proposed method under displacement using real data, and finally, we conclude with a discussion in Section 3.8.

3.2 A measurement error model for randomly displaced data

Let T_k denote the true location coordinates of the kth $(k = 1, \cdots, N)$ unit. The unit of analysis can be the individual, household or Enumeration Area (EA). The true value $T_k = (T_{k1}, T_{k2})$, where T_{k1} and T_{k2} denote the exact geographical longitude and latitude coordinates of the kth unit, which is lost due to the displacement. Therefore, only the displaced (contaminated by measurement error) coordinates $W_k = (W_{k1}, W_{k2})$ are observed. This can be expressed as

$$W_k = T_k + e_k, \tag{3.1}$$

where $e_k = (e_{k1}, e_{k2})$ is a random noise or error term added to the true location by the displacement process. As a result, the points move to other locations from their original locations, and may even cross the administrative boundaries of a country e.g., upazilas in Bangladesh which are lower than the displacement restricted higher administrative boundaries.

The true location T_k is unknown. Under the assumptions that 1) the distribution of the random error e_k is known, and 2) within a location type (urban or rural), the error term e_k is independent of the true location T_k, we can consider the model (3.1) to be a classical measurement error model. The T_k's in the model (3.1) might be either random variables or fixed quantities. The measurement error is called structural in the first case, whereas this is defined as functional in the second case (Arima and Polettini, 2019; Ybarra and Lohr, 2008). In our case, for a given unit k, the true location T_k will be considered a fixed quantity but unknown. Therefore, model (3.1) is considered to be a functional measurement error model.

For location point displacement, the measurement error means displacement error. Also, the

discretised nature of disaggregated domain assignment to the unit using displaced coordinates may lead to domain misclassification error. For example, the displaced unit may be misplaced at the lower administrative level. Therefore, in this case, the measurement error model (3.1) can be defined in terms of the misclassification probabilities for each domain e.g., upazila. The misclassification probabilities are unknown. However, these probabilities can be estimated through simulated location coordinates from the location displacement probability model defined as

$$f(W_k \mid T_k), k = 1, \cdots, N, \tag{3.2}$$

which is the conditional distribution of the displaced coordinates given the true coordinates. T_k is fixed but unknown, and the distribution of the displacement random component e_k in (3.1) is used to develop the distribution $f(W_k \mid T_k)$ in (3.2). This distribution will be derived analytically later in this section and is a key novel contribution in this chapter.

To derive $f(W_k \mid T_k)$, in this thesis, we consider the most popular and frequently used DHS displacement process involving a random distance and a random angle. However, this can be extended to other displacement error processes, for example, the Gaussian Displacement process. In sub-section 3.2.1, we discuss the DHS random displacement process. Sub-section 3.2.2 explores the distribution of the displaced coordinates given the true coordinates using simulation. The probability distribution of the displaced coordinates given the true coordinates is derived in sub-section 3.2.3. Finally, we present the target probability distribution of the unknown true coordinates given the displaced coordinates in sub-section 3.2.4.

3.2.1 The DHS displacement process

The displaced coordinates are generated by applying the DHS displacement algorithm (Burgert et al., 2013) to the original coordinates. Each true location is displaced independently. The steps in the DHS location data displacement are as follows:

1) Generate a random direction (angle) which is uniformly distributed between 0 and 2π that is $2\pi\upsilon_1$, where $\upsilon_1 \sim \text{Uniform}(0, 1)$.

2) Generate a random distance which is uniformly selected between 0 and δkm that is $\delta\upsilon_2$, where $\upsilon_2 \sim \text{Uniform}(0, 1)$ and the maximum displaced distance, $\delta = 2$km for urban locations and $\delta = 5$km for rural locations with 1% of rural locations randomly being given $\delta = 10$km distance.

3) Generate the xOffset and yOffset for T_k by combining steps 1 and 2 as follows: xOffset: $e_{k1} = sin(2\pi\upsilon_1) \times \delta\upsilon_2$ and yOffset: $e_{k2} = cos(2\pi\upsilon_1) \times \delta\upsilon_2$, where the error term $e_k = (e_{k1}, e_{k2})$ is the Offset or Circular 'Error'.

4) The new location is generated by adding the offsets to the actual coordinates, resulting in a new latitude and longitude for the location, as expressed below: $W_{k1} = T_{k1} + e_{k1}$ and $W_{k2} = T_{k2} + e_{k2}$.

5) The newly generated location is examined to guarantee its placement within the designated restricted higher administrative boundaries. In the case of Bangladesh, displaced locations must not cross district boundaries (Admin 2), but they can cross lower administrative boundaries, e.g., upazila. If it does not fall within the designated higher boundaries, a new displaced location is generated by repeating Steps 1-4. However, the displaced location is not limited to areas outside of large water bodies such as lakes, rivers, and oceans.

We observe that the generation of offsets is a crucial component of the DHS displacement process, which is later added to the true coordinates to obtain the displaced coordinates. We want to highlight the similarities between the way used to generate these offsets and the well-known Box-Muller algorithm (Muller and Muller, 1958) for generating observations from a bivariate normal distribution, using two independent uniformly distributed random variables υ_1 and υ_2. Indeed, if the DHS used $e_{k1} = sin(2\pi\upsilon_1) \times \delta\sqrt{-2\ln\upsilon_2}$ and $e_{k2} = cos(2\pi\upsilon_1) \times \delta\sqrt{-2\ln\upsilon_2}$ instead of $e_{k1} = sin(2\pi\upsilon_1) \times \delta\upsilon_2$ and $e_{k2} = cos(2\pi\upsilon_1) \times \delta\upsilon_2$, the resulting offsets would have been normally distributed, centered around the original point.

3.2.2 Exploring the distribution of the displaced coordinates by simulation

Before working on theoretical developments, we use simulation to explore the distribution of the displaced coordinates. Suppose we have a true point at latitude 0 and longitude 0, denoted as $T = (0, 0)$. The coordinates (latitude and longitude) in the simulated data set are then displaced using the DHS random displacement algorithm without any restrictions for administrative units. In urban areas, the coordinates are displaced with a maximum distance of 2km, while in rural areas, the coordinates are displaced by 5km, with an additional random displacement of 10km for 1% of the rural locations.

Figure 3.1 shows histograms representing the distribution of displacement distances and the displaced coordinates $x$offset and $y$offset for urban locations. The figure also demonstrates the location of the 5,000 displaced coordinates along with the original coordinate. The distribution of

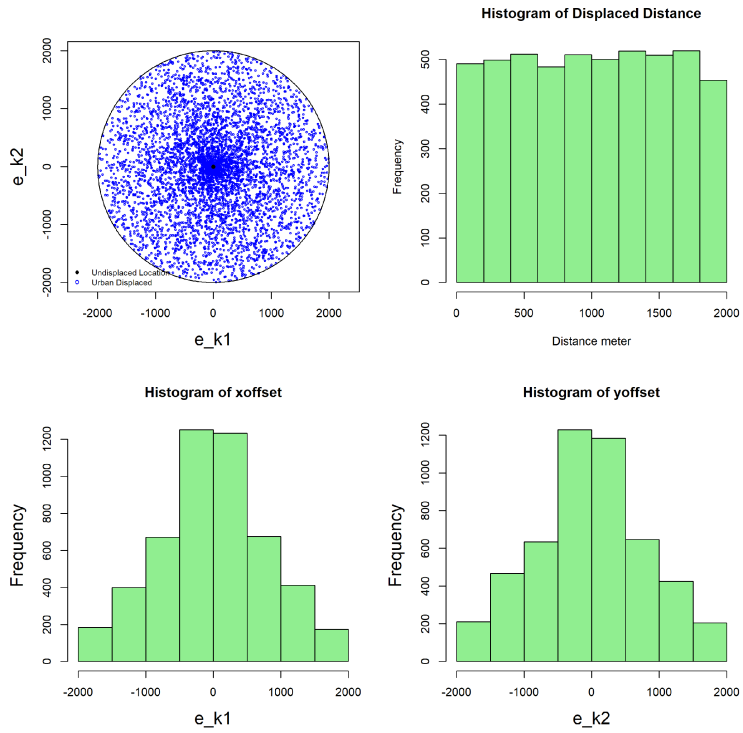

Figure 3.1: Simulated distributions of urban displaced points and distances from circle centre (0,0); 5000 displaced points, and 2000m buffer (average displaced distance = 998m, range = (1.3m-1999.2m)).

distances for urban locations appears largely uniform, while the distribution of offsets (representing random noise or errors) is approximately symmetric. The distribution of displacement random errors is non-uniform over the circle.

Figure 3.2 presents histograms indicating the distribution of displacement distances and the displaced coordinates $x$offset and $y$offset for rural locations, using a simulated dataset of 5,000 coordinates. Also, the figure shows the spatial structure of these displaced coordinates around the true rural location coordinate. The distribution of distances for rural locations demonstrates a relatively uniform pattern within the 5-kilometre buffer. Only a few coordinates (50 points) are scattered within the 10-kilometre buffer. Among these 50 points, only 22 fall between the 5-10km boundaries. Furthermore, the distribution of offsets (representing errors) exhibits an approximate symmetry.

3.2.3 Probability distribution of displaced coordinates given the true coordinates

The DHS displacement process is independent from each true location to location. Therefore, we develop the distribution of the displaced coordinates for a particular location, namely unit k and its two-dimensional offset/error, and displaced coordinates can be written from (3.1) as follows:

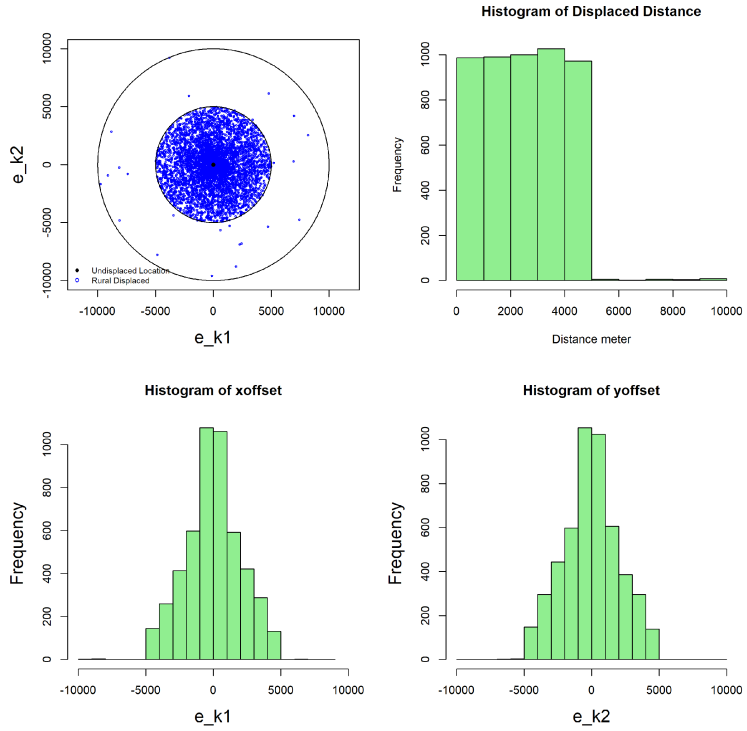

Figure 3.2: Simulated distributions of rural displaced points and distances from circle centre (0,0); 5000 displaced points, 5000m and 10000m buffers (average displaced distance = 2524m, Range = (3m, 9965m)).

$W_{k1} = T_{k1} + e_{k1}$ and $W_{k2} = T_{k2} + e_{k2}$, where, W_{k1} and W_{k2} are displaced x and y coordinates, T_{k1} and T_{k2} are true x and y coordinates. Also, e_{k1} and e_{k2} are xOffset and yOffset or circular errors expressed as follows: xOffset: $e_{k1} = sin(2\pi\upsilon_1) \times \delta\upsilon_2$ and yOffset: $e_{k2} = cos(2\pi\upsilon_1) \times \delta\upsilon_2$, where $\upsilon_1, \upsilon_2 \sim$ Uniform$(0, 1)$ and δ is the maximum displaced distance. Therefore, $2\pi\upsilon_1$ and $\delta\upsilon_2$ are the random angle and random distance over the interval $(0, 2\pi)$ and $(0, \delta)$ respectively.

At first, we develop the distribution of the random error term or offset (e_{k1}, e_{k2}). Next, we consider that the true coordinates (T_{k1}, T_{k2}) are given, and develop the distribution of the displaced coordinates (W_{k1}, W_{k2}).

Let $f_{\upsilon_1,\upsilon_2}(\upsilon_1, \upsilon_2)$ be the joint distribution of the random variables υ_1 and υ_2, which are uniformly distributed over (0,1), and it can be expressed as $f_{\upsilon_1,\upsilon_2}(\upsilon_1, \upsilon_2) = 1$ when $0 < \upsilon_1 < 1$ and $0 < \upsilon_2 < 1$, and 0 otherwise. The joint distribution of e_{k1} and e_{k2} is denoted as $f_{e_{k1},e_{k2}}(e_{k1}, e_{k2})$, and it is derived through the transformation of variables from $\upsilon = (\upsilon_1, \upsilon_2)$ to $e_k = (e_{k1}, e_{k2})$.

Let g_1 and g_2 represent transformation functions, expressed as $g_1(\upsilon_1, \upsilon_2) = e_{k1} = sin(2\pi\upsilon_1) \times \delta\upsilon_2$ and $g_2(\upsilon_1, \upsilon_2) = e_{k2} = cos(2\pi\upsilon_1) \times \delta\upsilon_2$ respectively. Also, g_1^{-1} and g_2^{-1} represent the respective inverse transformations for g_1 and g_2, mapping e_{k1} and e_{k2} back to υ_1 and υ_2.

The joint distribution of e_{k1} and e_{k2} is given by

$$\begin{aligned} f_{e_{k1},e_{k2}}(e_{k1},e_{k2}) &= f_{\upsilon_1,\upsilon_2}(g_1^{-1}(e_{k1},e_{k2}),g_2^{-1}(e_{k1},e_{k2})) \times \left|J(g_1^{-1},g_2^{-1})\right| \\ &= 1 \times \left|J(g_1^{-1},g_2^{-1})\right| \\ &= \left|J(g_1^{-1},g_2^{-1})\right| \end{aligned}$$

where $\left|J(g_1^{-1},g_2^{-1})\right|$ is the determinant of the Jacobian matrix of the inverse transformation, which is calculated using the partial derivatives of the inverse functions g_1^{-1} and g_2^{-1} with respect to e_{k1} and e_{k2}, and expressed as follows:

$$\left|J(g_1^{-1},g_2^{-1})\right| = \left|\frac{\partial(g_1^{-1},g_2^{-1})}{\partial(e_{k1},e_{k2})}\right| = \begin{vmatrix} \frac{\partial g_1^{-1}}{\partial e_{k1}} & \frac{\partial g_1^{-1}}{\partial e_{k2}} \\ \frac{\partial g_2^{-1}}{\partial e_{k1}} & \frac{\partial g_2^{-1}}{\partial e_{k2}} \end{vmatrix}$$

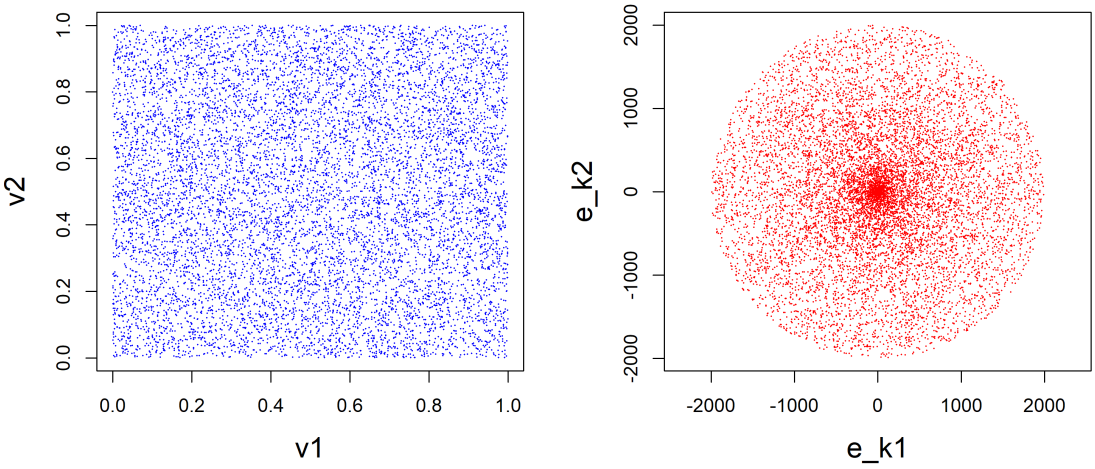

Figure 3.3: Sketch of the transformation of υ onto e_k for $T_k = (T_{k1}, T_{k2}) = (0,0)m$ and distance = 2000m.

From Figure 3.3, we note that $\upsilon = (\upsilon_1, \upsilon_2)$ are all pairs of points over (0,1) and transformed pairs of points $e_k = (e_{k1}, e_{k2})$ are over $(-\delta, \delta)$ bounded within the circle. The transformation can generate the same pair of values (e_{k1}, e_{k2}) with different signs for (υ_1, υ_2). As an illustration, let $\delta = 1$ then $(\upsilon_1, \upsilon_2) = (\frac{1}{8}, 1) \rightarrow (e_{k1}, e_{k2}) = (0.7, 0.7)$, and $(\upsilon_1, \upsilon_2) = (\frac{3}{8}, 1) \rightarrow (e_{k1}, e_{k2}) = (0.7, -0.7)$. For the transformation $\upsilon \rightarrow e_k$, each point in υ will correspond to just one point in e_k and the reverse is true. Therefore, the transformation is one-to-one.

The domain of the function $f_{e_{k1},e_{k2}}(e_{k1},e_{k2})$ is a bounded set of all points within a circle of the radius δ with a centre at the origin in the (e_{k1}, e_{k2}) plane in Figure 3.3 and can be expressed as

dom$(f_{e_{k1},e_{k2}}(e_{k1},e_{k2})) = \{(e_{k1},e_{k2})|0 < (e_{k1}^2 + e_{k2}^2) < \delta^2\}$.

Now, we are going to calculate $\left|J(g_1^{-1},g_2^{-1})\right|$ as follows:

$\frac{e_{k1}}{e_{k2}} = tan(2\pi\upsilon_1) \Rightarrow \upsilon_1 = g_1^{-1}(e_{k1},e_{k2}) = \frac{1}{2\pi}arctan(\frac{e_{k1}}{e_{k2}})$ and $e_{k1}^2 + e_{k2}^2 = \delta^2\upsilon_2^2 \Rightarrow \upsilon_2 = g_2^{-1}(e_{k1},e_{k2}) = \frac{(e_{k1}^2+e_{k2}^2)^{\frac{1}{2}}}{\delta}$

We obtain the partial derivatives as follows: $\frac{\partial g_1^{-1}}{e_{k1}} = \frac{1}{2\pi} \times \frac{1}{1+\left(\frac{e_{k1}}{e_{k2}}\right)^2} \times \frac{1}{e_{k2}} = \frac{1}{2\pi}\frac{e_{k2}}{e_{k1}^2+e_{k2}^2}$ and $\frac{\partial g_1^{-1}}{e_{k2}} = \frac{1}{2\pi} \times \frac{1}{1+\left(\frac{e_{k1}}{e_{k2}}\right)^2} \times \frac{-e_{k1}}{e_{k2}^2} = \frac{1}{2\pi}\frac{-e_{k1}}{e_{k1}^2+e_{k2}^2}$.

$\frac{\partial g_2^{-1}}{\partial e_{k1}} = \frac{1}{\delta} \times \frac{1}{2(e_{k1}^2+e_{k2}^2)^{\frac{1}{2}}} \times 2e_{k1} = \frac{e_{k1}}{\delta(e_{k1}^2+e_{k2}^2)^{\frac{1}{2}}}$ and $\frac{\partial g_2^{-1}}{\partial e_{k2}} = \frac{1}{\delta} \times \frac{1}{2(e_{k1}^2+e_{k2}^2)^{\frac{1}{2}}} \times 2e_{k2} = \frac{e_{k2}}{\delta(e_{k1}^2+e_{k2}^2)^{\frac{1}{2}}}$.

Now,

$$\left|J(g_1^{-1},g_2^{-1})\right| = \begin{vmatrix} \frac{\partial g_1^{-1}}{\partial e_{k1}} & \frac{\partial g_1^{-1}}{\partial e_{k2}} \\ \frac{\partial g_2^{-1}}{\partial e_{k1}} & \frac{\partial g_2^{-1}}{\partial e_{k2}} \end{vmatrix} = \begin{vmatrix} \frac{1}{2\pi}\frac{e_{k2}}{e_{k1}^2+e_{k2}^2} & \frac{1}{2\pi}\frac{-e_{k1}}{e_{k1}^2+e_{k2}^2} \\ \frac{e_{k1}}{\delta(e_{k1}^2+e_{k2}^2)^{\frac{1}{2}}} & \frac{e_{k2}}{\delta(e_{k1}^2+e_{k2}^2)^{\frac{1}{2}}} \end{vmatrix} = \frac{1}{2\pi\delta(e_{k1}^2+e_{k2}^2)^{\frac{1}{2}}}$$

Using the determinant of the Jacobian transformation, the joint distribution of e_{k1} and e_{k2} is expressed as follows:

$$f_{e_{k1},e_{k2}}(e_{k1},e_{k2}) = \begin{cases} \frac{1}{2\pi\delta(e_{k1}^2+e_{k2}^2)^{\frac{1}{2}}} & ,\left\{(e_{k1},e_{k2}) : 0 < (e_{k1}^2+e_{k2}^2)^{\frac{1}{2}} < \delta\right\} \\ 0 & \text{otherwise} \end{cases} \tag{3.3}$$

Our goal is to obtain the joint distribution of displaced coordinates (W_{k1}, W_{k2}) given the true coordinates (T_{k1}, T_{k2}). Therefore, using the distribution of the random error or offset in (3.3), we eventually develop the distribution for the displaced coordinates. To recall, we have $W_{k1} = T_{k1} + e_{k1}$ and $W_{k2} = T_{k2} + e_{k2}$. This implies that the generated random error is added to the true coordinates to obtain the displaced coordinates. Also, from these equations, we can write $e_{k1} = W_{k1} - T_{k1}$ and $e_{k2} = W_{k2} - T_{k2}$. This transformation is also one-to-one and is presented in Figure 3.4.

Therefore, conditional on $T_k = (T_{k1}, T_{k2})$, developing the distribution of displaced coordinates $W_k = (W_{k1}, W_{k2})$ using the distribution of $e_k = (e_{k1}, e_{k2})$ from (3.3) is straightforward. The joint distribution of (W_{k1}, W_{k2}) given (T_{k1}, T_{k2}) is given by

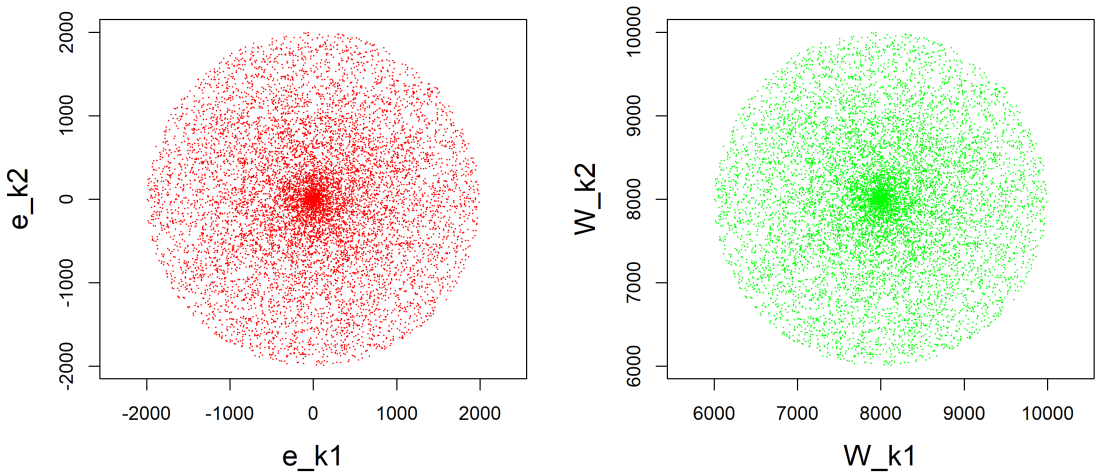

Figure 3.4: Sketch of the transformation of e_k onto W_k for $T_k = (T_{k1}, T_{k2}) = (8000, 8000)m$ and distance = 2000m.

$$f\left\{(W_{k1}, W_{k2}) \mid (T_{k1}, T_{k2})\right\}$$

$$= \begin{cases} \frac{1}{2\pi\delta\{(W_{k1}-T_{k1})^2+(W_{k2}-T_{k2})^2\}^{\frac{1}{2}}} & , \left\{(W_{k1}, W_{k2}) : 0 < \left\{(W_{k1}-T_{k1})^2 + (W_{k2}-T_{k2})^2\right\}^{\frac{1}{2}} < \delta\right\} \\ 0 & \text{otherwise} \end{cases}$$

Hence, the above probability distribution of the displaced coordinates given the true coordinates can be expressed as follows:

$$f(W_k|T_k) = \{2\pi\delta \text{dist}(W_k, T_k)\}^{-1} I(0 < \text{dist}(W_k, T_k) \leq \delta), \tag{3.4}$$

where $\text{dist}(W_k, T_k) = \sqrt{(W_{k1} - T_{k1})^2 + (W_{k2} - T_{k2})^2}$ is the Euclidean distance between the displaced location $W_k = (W_{k1}, W_{k2})$ and true location $T_k = (T_{k1}, T_{k2})$ and $I(.)$ is an indicator function.

3.2.4 Development of the distribution of the true coordinates given the displaced coordinates

Our target is to reconstruct the unknown distribution of $f(T_k \mid W_k)$ and using Bayes rule this can be expressed as follows:

$$f(T_k \mid W_k) \propto f(W_k \mid T_k) f(T_k). \tag{3.5}$$

Using (3.4) we can express the above unknown distribution $f(T_k|W_k)$ as

$$f(T_k|W_k) \propto \{2\pi\delta \text{dist}(W_k, T_k)\}^{-1} I(0 < \text{dist}(W_k, T_k) \leq \delta) f(T_k). \tag{3.6}$$

Now, regarding the second term on the right hand side of (3.6), since T_k is also unobserved here, the distribution of the true location data, $f(T_k)$ is unknown. If we knew it, we could generate pseudo samples of the unobserved true data given the observed data using the conditional distribution in (3.6).

In this situation, a non-parametric Multivariate Kernel Density Estimation (KDE) can be used to estimate the density corresponding to the spatial distribution of the unobserved true coordinates $f(T_k)$. Also, as true coordinates are unobserved, the target conditional distribution $f(T_k|W_k)$ in (3.6) can be used to obtain kernel density estimates. This is implemented using the Stochastic Expectation-Maximization (SEM) algorithm. The details are described in Section 3.3 below. However, if the underlying probability distribution of the true location data is known or can be approximated through external data e.g., WorldPop gridded population density, this can be incorporated into the estimation. In that case, an alternative to the KDE method, external data based estimation method will be described in Section 3.4.

3.3 Multivariate KDE in the presence of DHS displacement error

In this section, we propose a method of multivariate density estimation for estimating $f(T_k)$ accounting for the DHS random displacement error under the measurement error model in (3.6), followed by the estimation of domain parameters. This method is an extension of the proposed method by Groß et al. (2017), which is used for estimating a multivariate kernel density in the presence of a deterministic rounding (or aggregation) error. Notice that the method by Groß et al. (2017) is not directly applicable in the case of the DHS random displacement process that we extensively discussed in Chapter 2. To remind the reader, there are two primary reasons for this. First, the aggregation or rounding process is deterministic, whereas the DHS displacement process is stochastic. The later process has varying maximum displacement distances for urban and rural areas. Second, the conditional distribution of the aggregation or rounding error is uniform. This is not the case for the DHS random displacement algorithm. We have developed its probability distribution in section 3.2.3.

Multivariate KDE, which is a non-parametric approach, plays a significant role in estimating the

joint probability distribution of more than one continuous variable. Let $T = \{T_1 \dots, T_N\}$ denote the true (continuous) coordinates of the observations, such as longitude and latitude, with $T_k = (T_{k1}, T_{k2})$ where $k = 1, \dots, N$. In that case, we would have a multivariate (bi-variate) variable. Let $f(T^\star_{qs})$ be the unknown density corresponding to the distribution of the true (continuous) coordinates of the observations $T = \{T_1 \dots, T_N\}$. The continuous surface of all possible true coordinates are discretized into a finite set of points indexed by q and s, each ranging from 1 to L. Specifically, the grid point $T^\star_{qs} = (T^\star_{q1}, T^\star_{s2})$ represents one such discretized coordinate. In order to estimate the density $f(T^\star_{qs})$, a multivariate kernel density estimator is used, which is given as follows:

$$\hat{f}_H(T^\star_{qs}) = \frac{1}{N|H|^{1/2}} \sum_{k=1}^{N} K\left\{H^{-1/2}(T^\star_{qs} - T_k)\right\}, \tag{3.7}$$

where, $K\{.\}$ represents a multivariate kernel function, H indicates a symmetric positive definite bandwidth matrix, and $|.|$ represents the determinant. The selection of the bandwidth, known as a smoothing parameter, plays a crucial role in determining the performance of a kernel density estimator. Choosing an appropriate value for the bandwidth is essential, as excessively small or large values are not desirable. When the bandwidth, H, is too small, the resulting estimates exhibit spikiness due to under-smoothing. On the other hand, larger values of H lead to over-smoothing of the estimates. We have mentioned different approaches for bandwidth selection in Chapter 2. In this work, we adopt an approach described by Wand and Jones (1994), where the bandwidth selection in the multivariate situation is achieved using a plug-in estimator. The typical strategy involves selecting H by minimising the asymptotic root mean integrated squared error (RMISE) through the use of plug-in or cross-validation techniques (Silverman, 2018, section 2.4).

Instead of T, only the randomly displaced (contaminated by measurement error) coordinates of the observations $W = \{W_1 \dots, W_N\}$ are available, where $W_k = (W_{k1}, W_{k2})$. However, the displacement of true points may alter the original spatial pattern. Hence, the density estimates of the observations using a naive density estimator that discards the displacement error by substituting the true values T_k with the displaced values W_k in equation (3.7) may not be close to the density of the true data.

However, we still target to estimate the density $f_H(T^\star_{qs})$ by using only the observed displaced values W under the measurement error model $f(T|W) \propto f(W|T)f(T)$, where $f(W|T) = \prod_{k=1}^{N} f(W_k|T_k)$ and $f(T) = \prod_{k=1}^{N} f(T_k)$. The true coordinates are displaced independently for each location. Therefore, we estimate $f(T|W)$ under the assumption of factorising the density for

each independent location. Following the equation (3.6), we can write

$$f(T_k = T^{\star}_{qs}|W_k) \propto \left\{2\pi\delta \text{dist}(W_k, T^{\star}_{qs})\right\}^{-1} I(0 < \text{dist}(W_k, T^{\star}_{qs}) \leq \delta) f(T_k = T^{\star}_{qs}), \quad (3.8)$$

where $f(T_k = T^{\star}_{qs})$ is evaluated at grid point $T^{\star}_{qs}$ defined earlier.

Since the true data T and consequently $f(T) = \prod_{k=1}^{N} f(T_k)$ are initially unknown, Groß et al. (2017) proposed a Stochastic Expectation-Maximization (SEM) algorithm for estimation. This algorithm utilizes an initial estimate of $f_H(T^{\star}_{qs})$ based on W, followed by alternating simulations of T from $f(T|W)$ and re-estimation of $f(T)$ until convergence is achieved. This can be considered as a modified version of the expectation-maximization (EM) algorithm (Dempster et al., 1977; Groß et al., 2017). In this thesis, we extend the Groß et al. (2017) procedure for the case of the DHS random displacement. We present the exact implementation of the algorithm in the following subsection.

3.3.1 Estimation using the SEM algorithm

T_k are repeatedly drawn from $f(T_k = T^{\star}_{qs}|W_k)$ followed by estimation of $f(T_k = T^{\star}_{qs})$, by employing a multivariate kernel density estimator on the T_k. The computational steps of the algorithm are described below:

- Step 1: Generate an evenly spaced fine grid $G = T^{\star}_{q1} \times T^{\star}_{s2}$ for $q = 1, \dots, L$ and $s = 1, \dots, L$, with grid size L, grid width $L_{g1} = \frac{\max_k(W_{k1}) - \min_k(W_{k1}) + 2\delta}{L-1}$, $L_{g2} = \frac{\max_k(W_{k2}) - \min_k(W_{k2}) + 2\delta}{L-1}$; $T^{\star}_{q1} = \left\{\min_k(W_{k1}) - \delta, \min_k(W_{k1}) - \delta + L_{g1}, \cdots, \max_k(W_{k1}) + \delta\right\}$, also $T^{\star}_{s2} = \left\{\min_k(W_{k2}) - \delta, \min_k(W_{k2}) - \delta + L_{g2}, \cdots, \max_k(W_{k2}) + \delta\right\}$
  where $k = 1, \dots, N$ and δ denotes the displacement parameter (distance) that is introduced in the DHS displacement process in subsection 3.2.1.

- Step 2: Obtain a pilot estimate $\hat{f}_H(T^{\star}_{qs})$ of $f_H(T^{\star}_{qs})$ using the displaced data W with an initial bandwidth matrix

$$H^{(0)} = \begin{pmatrix} l & 0 \\ 0 & l \end{pmatrix},$$

  where l takes a value that is sufficiently large, such as twice the range of the observation coordinates.

- Step 3 (S-step): For $k = 1, 2, \dots, N$, a pseudo-sample $T^{\star}_k = (T^{\star}_{k1}, T^{\star}_{k2})$ of the unknown true points $T_k = (T_{k1}, T_{k2})$ is selected by sampling from the conditional distribution $f(T_k = T^{\star}_{qs}|W_k)$ in (3.8). In this step, the estimation of the conditional distribution of displaced

coordinates $f(W_k|T_k)$ in (3.8) is restricted to the circular region centered around $W_k = (W_{k1}, W_{k2})$, i.e., $0 < \sqrt{(W_{k1} - T^{\star}_{q1})^2 + (W_{k2} - T^{\star}_{s2})^2} \leq \delta$, and is also confined within the designated higher administrative boundaries (e.g., a district or admin 2 of the country). This implies that any coordinates falling outside of both the circular buffer and the designated administrative boundary will have a 0 probability of being selected as the true coordinates.

- Step 4 (M-step): Obtain the estimated bandwidth matrix H by using the multivariate plug-in estimator of Wand and Jones (1994) and re-compute the bivariate kernel density $\hat{f}_H(T^{\star}_{qs})$ using the pseudo-sample points from Step 3. Alternative bandwidth selection techniques can also be considered. When the goal is to estimate the density and the estimation includes regions with large water bodies, such as seas, applying boundary corrections is crucial: eliminate points in water and rescale land points to preserve accurate density representation.

- Step 5: Repeat Steps 3-4 for a total of A_1 (burn-in iterations) + A_2 (sample iterations).

- Step 6: Exclude the A_1 burn-in density estimates and calculate the final density estimate of $f_H(T^{\star}_{qs})$ by taking the average of the remaining A_2 density estimates $\hat{f}_H(T^{\star}_{qs})$. Larger values for A_1 and A_2, such as $A_1 = 20$, and $A_2 = 200$, enhance the accuracy in realizing the unknown true coordinates of observations and subsequently improve the density estimates by accounting for random displacement errors. The sensitivity of the results to different combinations of A_1 and A_2 can be assessed by evaluating the quality of the density estimates using measures, such as the RIMSE, as defined in (3.27).

3.3.2 Domain estimation using randomly displaced data

As we stated, our next aim is to develop a method to obtain estimates of the domain (e.g., upazila) parameters of interest accounting for misplacement uncertainty due to random displacement. Although the random displacement process is restricted within a higher administrative boundary ,e.g., district (admin 2) in Bangladesh, the displaced points can cross any administrative boundary, for instance, upazila (admin 3) in Bangladesh, which is lower than the restricted boundary. As we mentioned earlier, in this thesis, we consider upazila as the target domain. Therefore, when obtaining any (direct) estimates at the upazila level using georeferenced household location data and spatial overlaid polygon regions, the random displacement of household locations can result in misclassification or misassignment of households to the correct upazila boundaries, especially when the true household location is displaced outside of its associated polygon region. The probability of misclassifying a point to an administrative boundary (polygon) feature depends on the size and shape of the overlaid feature. Larger polygon regions generally have a lower chance of incorrect

assignment. Also, the true point certainly lies within a buffer surrounding the displaced point, with a radius equal to the maximum displacement distance (e.g., 5km). This buffer is referred to as the displacement buffer. Moreover, if the displacement buffer falls completely within an administrative (upazila) boundary, there is no chance of misclassification i.e., the point is always classified to the correct upazila with a probability equal to 1. On the other hand, if the displacement buffer intersects with an administrative boundary, the point is known as a potentially misplaced point. In that case, the true point lies within one of these intersecting administrative areas. The chance of originating from one admin boundary is not proportional to the intersecting boundary area with the displacement buffer because it depends on the random displacement process, and the distribution of random displaced points is not uniform (Figure 3.1). For instance, if most of the likely true points from an upazila within the displacement buffer have a short distance from the displaced point, then the upazila will be correctly assigned to the household location with a high probability. Therefore, we propose a domain estimator taking into account the misplacement uncertainty under the measurement error model (3.8), which is described as follows:

Let U be the population of units e.g., households in the country and the units denoted by $k = 1 \dots, N$. These units are grouped in domains $i = 1, \dots, M$, such that each unit belongs to one and only one domain and they add up to the population i.e., $\cup_{i=1}^{M} U_i = U$, where U_i denotes the ith domain population. Associated to unit k, we have the data vector $D_k = (Y_k, T_k)$, where Y_k denotes a variable measured at the true location T_k, for instance, household income (or whether the household is poor), and $T_k = (T_{k1}, T_{k2})$ are the true coordinates of the kth unit.

We are interested in estimating a parameter defined for domain U_i e.g., the proportion of poor households expressed as

$$\bar{Y_i} = \frac{\sum_U Y_k I_i(T_k)}{\sum_U I_i(T_k)}, i = 1, \dots, M; k = 1, \dots, N, \tag{3.9}$$

where $I_i(T_k) = 1$ if $T_k \in U_i$, otherwise it takes value 0.

This task would be straightforward if we had knowledge of the coordinates T_k for all units. Unfortunately, as a way to reduce the disclosure risk, the coordinates T_k have been displaced to location W_k using the random mechanism explained in subsection 3.2.1, and only the coordinates W_k are available to the data analysts. Even in the case of having data for the whole population of units (no sampling), using a similar expansion to (3.9) with W_k in place of T_k may lead to bias in estimates whenever the displaced W_k belongs to a different domain than T_k (misplacement).

Although the displacement mechanism is random and independent of Y_k, it is possible to use the

displaced coordinates W_k to inform about the true location of the household via the distribution $f(T_k|W_k)$. However, we still aim to estimate the parameter $\bar{Y}_i$ using the values W_k accounting for misplacement uncertainty via $f(T_k \mid W_k)$ in (3.6) that corrects for misplacement error. In particular, we can generate pseudo samples of the unobserved true data given the observed data from $f(T_k \mid W_k)$ but we need to estimate $f(T_k)$. In that case, we propose an estimator of $\bar{Y}_i$ using pseudo samples $T_k^\star$ of the true unknown points T_k drawn by using the proposed kernel density based SEM algorithm that we presented in subsection 3.3.1. Notice that one obtains simulated coordinates for each unit as a by-product of the routine. The proposed domain estimator is expressed as

$$\hat{\bar{Y}}_{i(\text{KDE})} = \frac{\sum_U Y_k I_i(T^\star_{k(\text{KDE})})}{\sum_U I_i(T^\star_{k(\text{KDE})})}, i = 1, \dots, M; k = 1, \dots, N, \tag{3.10}$$

where $I_i(T^\star_{k(\text{KDE})}) = 1$ if $T^\star_{k(\text{KDE})}$ belongs to the domain i, otherwise it takes value 0.

Iterative samples of true location data out of all possible true locations are needed for a better understanding of the true location. Therefore, for each complete set of simulated coordinates $T_k^\star = (T_{k1}^\star, T_{k2}^\star)$, $k = 1, 2, \dots, N$ from Step 3 in the proposed SEM algorithm in 3.3.1, obtain the domain parameter estimates $\hat{\bar{Y}}_{i(\text{KDE})}$ in (3.10) and obtain the final estimate by taking an average of estimates over A_2 samples.

Now, in the case of sample data, let $s \subset U$ be a sample of units $(k = 1 \dots, n)$ selected according to a given sampling design. Let $s_i = s \cap U_i$ denote the sampled units in (small) domain $i (i = 1, \dots, m)$ and let $\pi_k = P(k \in s)$ denote the inclusion probabilities, and b_k denote the sampling weights, which are calculated as the inverse of the inclusion probabilities. By using sampling weights and pseudo-samples of true coordinates, and following the process described by Walter et al. (2022), we can express the estimators corresponding to $\bar{Y}_i$ in (3.9) and $\hat{\bar{Y}}_{i(\text{KDE})}$ in (3.10) using sample data s as follows:

$$\hat{\bar{Y}}_i^{\,s} = \frac{\sum_s Y_k b_k I_i(T_k)}{\sum_s b_k I_i(T_k)}, i = 1, \dots, m; k = 1, \dots, n, \tag{3.11}$$

$$\hat{\bar{Y}}^{\,s}_{i(\text{KDE})} = \frac{\sum_s Y_k b_k I_i(T^\star_{k(\text{KDE})})}{\sum_s b_k I_i(T^\star_{k(\text{KDE})})}, i = 1, \dots, m; k = 1, \dots, n, \tag{3.12}$$

where b_k are the sampling weights.

The KDE estimator, used to estimate density, may not work well when applied to a sample. Specif-

ically, using a sample to estimate $f(T_k)$ may produce a spike density when a sample has a sort of aggregation. For example, when a household represents, e.g., 20 households, according to sampling weights, they may be expected to be distributed throughout the EA rather than geographically centred in it. Furthermore, the method cannot differentiate between the absence of sample data and the absence of population. Also, under the random displacement process, false in-sample and out-of-sample domains can arise. A false in-sample domain refers to a domain that gains EAs due to the displacement process, even though originally no EAs were sampled from that domain. Conversely, a false out-of-sample domain means that all of the sampled EAs from that domain are lost due to the random displacement process. However, the method may work under certain sampling designs that cover the entire population area, such as a simple random sample over the whole area or a geographically stratified sample. On the other hand, it may work badly in the presence of multistage sampling, as entire areas would not be selected.

3.4 External data based estimation method in the presence of random displacement error

In Section 3.3, we proposed an estimation method for drawing likely pseudo-samples of true coordinates under the measurement error model in (3.6), where we used kernel density estimates as the marginal distribution of the true location data. The target of the estimation remains the same as before. However, the underlying distribution of true location data can be approximated using available information from external data sources. The information may be useful for drawing the best possible true locations. Therefore, in this section, we aim to propose an estimation method for drawing likely pseudo-samples of true coordinates from the model by using external data for approximating the distribution of the true location data. As our primary interest, in this section, we focus only on domain parameter estimates accounting for misplacement error due to the displacement process. The method we are proposing here has similarities with the method proposed by Warren et al. (2016b) which we described in detail in Chapter 2. As a reminder to the reader, Warren et al. (2016b) generated the likely true coordinates by displacing the displaced coordinates using the DHS displacement mechanism i.e., through the distribution $f(W_k \mid T_k)$ in (3.4). However, Warren et al. (2016b) did not present the mathematical form of $f(W_k \mid T_k)$ which is one of the contributions of this thesis. In addition, Warren et al. (2016b) did not consider $f(T_k)$ in their model and allowed all points within a buffer to be possible true points, but the points in large water bodies e.g., sea, lake and uninhabited areas, such as large green or forest lands, cannot be possible true locations and have a selection probability equal to 0 (Figure 3.5). Therefore, we extend

the method of Warren et al. (2016b) by bringing additional information for $f(T_k)$ from external data sources. Hereafter, we refer to the proposed method as the external data based classification (EDC) method. Specifically, in the proposed EDC method, we use our developed target distribution $f(T_k \mid W_k)$ in (3.6). Another distinct feature of the proposed method is that it accounts for the uncertainty of all probable domains where the true location point may belong. In contrast, in the method by Warren et al. (2016b), only the most probable domain was picked and assigned to the EA location as the correct domain. We describe the proposed method using appropriate notations as follows:

Following equation (3.6), we can express the unknown conditional distribution $f(T_k \mid W_k)$ as

$$f(T_k = T_{kq}^{\star} \mid W_k) \propto [2\pi\delta \text{dist}(W_k, T_{kq}^{\star})]^{-1} I(0 < \text{dist}(W_k, T_{kq}^{\star}) \leq \delta) f(T_k = T_{kq}^{\star}), \quad (3.13)$$

where $T_{kq}^{\star}$ denotes the qth $(q = 1, 2, \cdots, Q_k)$ potential true location for the kth household within the displacement buffer around the W_k, and $f(T_k = T_{kq}^{\star})$ is the distribution of the potential true locations $T_{kq}^{\star}$. Although the T_k is continuous, we estimate $f(T_k = T_{kq}^{\star})$ based on a discrete approximation using external grid centroids described below.

The distribution of the potential true locations, $f(T_k = T_{kq}^{\star})$ is unknown. This underlying distribution can be approximated by using available information from different external data sources. The information may be useful for classifying the households to correct domains with higher classification probabilities. Therefore, we choose the information carefully and their possible relations with the correct domains/locations are presented in Figure 3.5 and justified below:

1) Population density: The household population density may be a good predictor for selecting the likely true points from the correct domain within the displacement buffer. Moreover, in some cases, there may be a direct relationship between the population density and true locations. For example, as we have discussed in detail in Chapter 1, in the 2011 BDHS, 600 true Enumeration Area (EA) centroid locations were selected with probability proportional to the 2011 census EA size (number of households) (NIPORT, 2013). While this thesis does not directly address the probability proportional to size (PPS) sampling approach, the effect of unequal probability selection is inherently reflected when using household population density to estimate $f(T_k)$. Thus, locations with higher population densities are more likely to be potential true locations. The population density of the true potential location data is not known. However, these population densities can be approximated by using alternative sources of population data. For example, the WorldPop database which produces the gridded

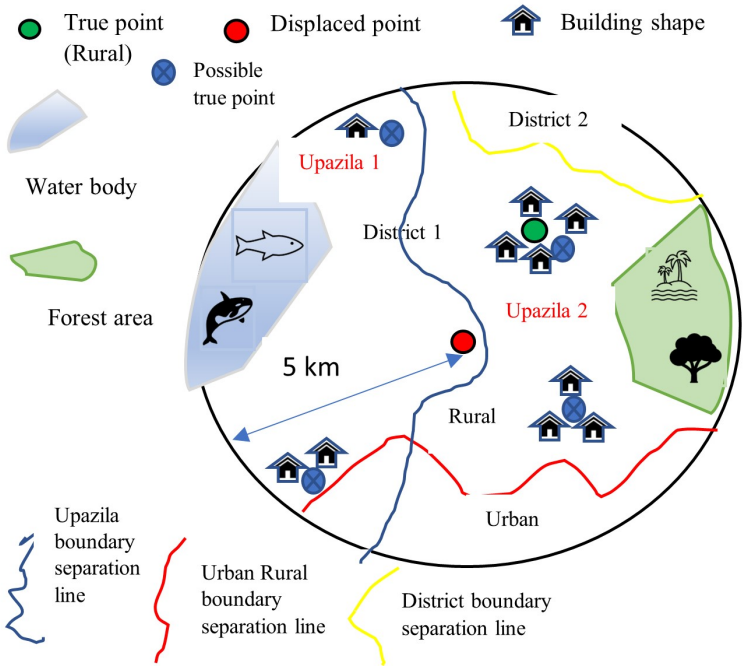

Figure 3.5: Schematic of a hypothetical scenario where a rural household location coordinate is misclassified to an administrative unit (upazila) due to the random displacement (where building shape refers to building footprints obtained through satellite imagery).

population density estimates for each country (Bondarenko et al., 2020). Therefore, in this work, we use the 100m $\times$ 100m boundary corrected gridded population density 2020 data of Bangladesh from WorldPop. We have discussed the WorldPop gridded population density estimation method in Chapter 1. To remind the reader, the gridded population densities for a given country are estimated using Random Forests machine learning methods, which incorporate survey/census and remote sensing data, such as the Lights at Night (LAN) data. The method employs a top-down approach to disaggregate administrative unit-based census and projection counts to grid cell-based counts, using geospatial datasets such as building footprint data and high-resolution road networks.

2) Designated admin boundary restriction: As the true points are displaced within a higher administrative boundary e.g., admin 2 (district) in Bangladesh, the displaced points can not cross their original district boundaries. Therefore, the points outside the original district boundary of the true household within the displacement buffer are given 0 probability of being selected as the true likely locations.

3) Rural Urban boundaries: The DHS random displacement process is not restricted within the rural urban locations. Therefore, it is possible that a household that originated from a rural location can be displaced to an urban location and vice versa. However, if we know a) whether the household (or EA) locations are originally from urban or rural areas, as is

the case with the DHS e.g., the 2011 BDHS and b) the rural/urban boundaries of a country, this additional information can be used while selecting the likely true household location within the displacement buffer around the displaced point. In the case of Bangladesh, the rural/urban boundaries information can be obtained from the population census data that we discussed in detail in Chapter 1.

4) Boundary corrections for non-settlement places: True household locations can not be in uninhabited places. However, the DHS random displacement process is not restricted to inhabited places only. Therefore, it is possible that the true household points can be displaced to uninhabited places e.g., large green or forest areas or in large water bodies such as the sea, lakes etc. For instance, we see by creating a map of the 2011 BDHS displaced EA coordinates that some EA coordinates were displaced to non-settlement places. Thus, in order to select the likely true household locations through the proposed method, any points in uninhabited places within the displacement buffer are given 0 probability of being selected. We implement these boundary corrections in the proposed method by using the boundary corrected gridded population data from WorldPop which we discussed in Chapter 1. WorldPop produces the boundary corrected gridded population by using remote sensing data based higher resolution satellite images for the settlement places and/or building footprints.

Therefore, using the above information which is available from different external data sources, the distribution of potential true locations $f\left(T_k = T_{kq}^{\star}\right)$ can be approximated as follows:

$$\hat{f}\left(T_k = T_{kq}^{\star}\right) = \frac{N_{kq} \times I_{\text{DRA}}(T_{kq}^{\star}) \times I_{\text{UR}}(T_{kq}^{\star}) \times I_{\text{NS}}(T_{kq}^{\star})}{\sum_q [N_{kq} \times I_{\text{DRA}}(T_{kq}^{\star}) \times I_{\text{UR}}(T_{kq}^{\star}) \times I_{\text{NS}}(T_{kq}^{\star})]}, q = 1, 2, \cdots, Q_k, \quad (3.14)$$

where N_{kq} is the number of people at location $T_{kq}^{\star}$; $I_{\text{DRA}}(T_{kq}^{\star}) = 1$ if $T_{kq}^{\star}$ falls within the displacement restricted administrative boundaries, otherwise it takes value 0; $I_{\text{UR}}(T_{kq}^{\star}) = 1$ if $T_{kq}^{\star}$ falls in the same region (either urban or rural) from which the household (or EA) point actually originated, otherwise it takes value 0, and $I_{\text{NS}}(T_{kq}^{\star}) = 1$ if $T_{kq}^{\star}$ does not fall in the non-settlement places e.g., a large water body, forest area etc, otherwise it takes value 0.

3.4.1 Estimation

It is not possible to obtain the target conditional distribution of the true data $f(T_k|W_k)$ in (3.13) analytically. As the true location T_k is unobserved, we can use numerical integration for estimation. In particular, we draw Monte Carlo samples of potential true locations using (3.13) given the observed/displaced locations. This works by replacing the unobserved true point T_k in the target

distribution by generating pseudo-samples of the unobserved true point given the observed data and then, obtaining estimates of the target parameters e.g., domain means/proportions $\hat{\bar{Y}}_i$ with the updated (new) location samples. The process is repeated a large number of times and we take the average of the estimates. However, the domain mean estimates converge when the proportions of the realizations (pseudo-sample points) within a domain converge. Studies by Warren et al. (2016b) and Wilson et al. (2020) recommended drawing a large number of samples for better realizations of the true location. The computational steps of the algorithm for the external data based proposed method are described below:

- Step 1: For each household $k = 1, 2, \ldots, N$ take the WorldPop grid of points $(T^{\star}_{kq1}, T^{\star}_{kq2}), q = 1, 2, \ldots, Q_k$ around (W_{k1}, W_{k2}) of width 2δ that covers the complete (T_{k1}, T_{k2}) space, where δ is the maximum displacement distance.

- Step 2: For each household $k = 1, 2, \ldots, N$ pseudo-samples $T^{\star}_k = (T^{\star}_{k1}, T^{\star}_{k2})$ of the true unknown points $T_k = (T_{k1}, T_{k2})$ are drawn from $f(T_k = T^{\star}_{kq}|W_k)$ in (3.13) with $\hat{f}\left(T_k = T^{\star}_{kq}\right) = \frac{N_{kq} \times I_{\text{DRA}}(T^{\star}_{kq}) \times I_{\text{UR}}(T^{\star}_{kq}) \times I_{\text{NS}}(T^{\star}_{kq})}{\sum_q [N_{kq} \times I_{\text{DRA}}(T^{\star}_{kq}) \times I_{\text{UR}}(T^{\star}_{kq}) \times I_{\text{NS}}(T^{\star}_{kq})]}, q = 1, 2, \cdots, Q_k$ in (3.14).

- Step 3: For $i = 1, 2, \ldots, M$ and $k = 1, 2, \ldots, N$ assign domain i to the sample household points $T^{\star}_k$ from Step 2 that is, let $\text{Area}_i(.)$ be a domain assignment function such that $\text{Area}_i(T^{\star}_k) = i$ if the point $T^{\star}_k$ belongs to domain (or area) i.

- Step 4: Iterate Steps 2-3 $B_1 + B_2$ times with $B_1 = 200$ (large number of iterations) and $B_2 = 20$ (additional iterations).

- Step 5: In order to set the convergence criteria, let $\hat{P_{ki}} = P(\text{Area}_i(T_k = T^{\star}_k) = i|W_k)$ be the estimated domain membership probabilities for the kth household based on the pseudo-samples of the true location data $T^{\star}_k$, where $\text{Area}_i(.)$ is a domain assignment function defined in Step 3. If the buffer around W_k does not intersect with any domain boundary (i.e., non-misplaced location), there is no need for iteration. However, a large number of iterations should be done over all possible domain boundaries that intersect with the displacement buffer. Thus, compute the domain membership probabilities for the kth household $\hat{P_{ki}}^{(B_1)}$ and $\hat{P_{ki}}^{(B_1+B_2)}$ by using the B_1 and $(B_1 + B_2)$ iterative sample steps respectively, where $\sum_{i=1}^{M} \hat{P_{ki}}^{(B_1)} = 1$ and $\sum_{i=1}^{M} \hat{P_{ki}}^{(B_1+B_2)} = 1$.

- Step 6: If $|\hat{P_{ki}}^{(B_1+B_2)} - \hat{P_{ki}}^{(B_1)}| < \omega$, where ω is a vector of small quantities, the domain membership probabilities $\hat{P_{ki}}^{(B_1)}$ have converged. Alternatively, we can create trace plots for the probabilities of domain membership to see whether the two estimates, $\hat{P_{ki}}^{(B_1+B_2)}$ and $\hat{P_{ki}}^{(B_1)}$, are reasonably close. If they are, then we can consider that the algorithm has

converged. Otherwise, return to Step 2 and repeat the subsequent steps, increasing B_1 by a predefined amount (e.g., $B_1 = B_1 + 20$), while retaining the constant value $B_2 = 20$, until the convergence criteria are met.

- Step 7: For each household $k = 1, 2, \ldots, N$ select domains i by sampling from the domain membership probability distribution $\hat{P}_{ki}^{(B_1)}; i = 1, 2, \ldots, M$ in Step 6.
- Step 8: For each complete set of selected domains from Step 7, obtain the domain parameter estimates $\hat{\bar{Y}}_i, (i = 1, 2, \ldots, M)$ expressed in Subsection 3.3.2.
- Step 9: Repeat Steps 7 and 8 $B_3 = 100$ times and obtain the final estimates by averaging the derived B_3 estimates.

3.5 Estimation under aggregation and displacement errors

Suppose that to preserve the confidentiality of respondents, the population (e.g., register or census) data release policy uses two types of protection. First, the household location data within each Enumeration Area (EA) is aggregated at the centroid of the EA. Next, the EA centroid coordinates are displaced using a displacement algorithm, e.g., the DHS random displacement algorithm discussed in Subsection 3.2.1. So, only displaced EA centroid coordinates are provided to researchers. Notice that, as in the DHS, the aggregation of coordinates is restricted to the boundaries of the target domain. This domain level determines where we aim to obtain estimates for the indicator of interest. However, the aggregation and displacement processes induce measurement errors in data, which may alter the original spatial pattern. In particular, the aggregation concentrates the data into a single location, and the data can be misplaced when the EA is displaced outside of its original domain boundary. Disregarding the measurement error may result in incorrect density and domain parameter estimates. To the best of our knowledge, previous studies did not attempt to obtain estimates by taking into account the uncertainty due to both aggregation and displacement. Therefore, in this section, we aim to propose a method to obtain a) density estimates corresponding to the distribution of the true (continuous) coordinates of the observations (e.g., household or poor household) and b) estimates of the domain parameters under a measurement error model accounting for both aggregation and displacement errors. This is one of the key novel contributions in this chapter.

3.5.1 Multivariate KDE in the presence of aggregation and displacement errors

In this subsection, we propose a method for multivariate density estimation accounting for both aggregation and displacement errors. Let U be a population of observations spread over space. Let, $T_{jk} = (T_{jk1}, T_{jk2})$ be the true coordinates, such as longitude and latitude of the kth ($k = 1, \cdots, N_j$) unit, which can be household or individual e.g., poor household/people, in the jth ($j = 1, \cdots, R$) EA. In that case, we would have a multivariate (bi-variate) variable. Let $f(T^{\star}_{qs})$ be the unknown density corresponding to the distribution of the true (continuous) coordinates of the observations $T = (T_{11}, \ldots, T_{1N_1}, \ldots, T_{R1}, \ldots, T_{RN_R})$, where $T^{\star}_{qs} = (T^{\star}_{q1}, T^{\star}_{s2})$ for $q = 1, \cdots, L$ and $s = 1, \cdots, L$ denote the grid points that cover the complete $T_{jk} = (T_{jk1}, T_{jk2})$ space i.e., all possible true locations and L is the grid size, denoting the equal length of sequences of points on both axes. In order to estimate the density $f(T^{\star}_{qs})$, a multivariate kernel density estimator is utilised. This estimator is represented by the following equation:

$$\hat{f}_H(T^{\star}_{qs}) = \frac{1}{N|H|^{1/2}} \sum_{U} K\left\{H^{-1/2}(T^{\star}_{qs} - T_{jk})\right\}, \tag{3.15}$$

where $K\{.\}$ indicates a multivariate kernel function, H denotes a symmetric positive definite bandwidth matrix and $|.|$ represents the determinant. The selection of the bandwidth, a critical smoothing parameter, significantly impacts the performance of a kernel density estimator. We have discussed the bandwidth selection approaches in Section 3.3.

To preserve the confidentiality of respondents, the true location coordinates T_{jk} are first subjected to an aggregation process and subsequently to a random displacement. As a result, data analysts only have access to these aggregated and displaced coordinates. To describe this process mathematically, we define specific notations. Under only aggregation, it is assumed that all households are aggregated to the center of the jth EA, sharing its centroid coordinate. This is denoted by $T_j = (T_{j1}, T_{j2})$, where T_{j1} and T_{j2} represent the geographical longitude and latitude coordinates of the EA centroid, respectively. We should clearly note that T_{jk} represents the true coordinates, whereas T_j denotes the aggregated centroid coordinates. However, these aggregated EA centroid coordinates are lost due to subsequent displacement. Specifically, the random displacement process is applied to T_j. Consequently, data analysts can only observe the coordinates that are both aggregated and displaced (affected by measurement error) represented by $W_j = (W_{j1}, W_{j2})$. Using a naive kernel density estimator that disregards the measurement error and substitutes the true values T_{jk} with the displaced values W_j in equation (3.15) can result in a density with abrupt peaks (referred to as "spiky") that deviates from the density of the actual data.

We assume that the aggregation and displacement processes are known. Therefore, we can formulate a measurement error model for aggregated and randomly displaced coordinates. In particular, the conditional distribution of true coordinates T_{jk} given the aggregated and displaced coordinates T_j and W_j can be expressed as follows:

$$\begin{aligned} f(T_{jk} \mid T_j, W_j) &\propto f(W_j \mid T_{jk}, T_j) f(T_{jk}|T_j) f(T_j) \\ &\propto f(W_j \mid T_j) f(T_j) f(T_{jk}|T_j) \\ &\propto f(T_{jk}|T_j) f(T_j \mid W_j), \end{aligned} \tag{3.16}$$

where W_j is independent of T_{jk} conditional on T_j and $f(T_j \mid W_j) \propto f(W_j \mid T_j) f(T_j)$, which we can write based on the conditional distribution of true coordinates given the displaced coordinates in (3.5). The first part of the right hand side of equation (3.16) indicates the aggregation issue and the last part indicates the displacement issue.

Under EA displacement, following (3.5) and (3.6) that we developed in Section 3.2, the conditional distribution of the *unknown* EA coordinates T_j given the displaced EA coordinates W_j can be expressed as follows:

$$\begin{aligned} f(T_j \mid W_j) &\propto f(W_j \mid T_j) f(T_j) \\ &\propto [2\pi\delta \text{dist}(W_j, T^\star_{jq})]^{-1} I(0 < \text{dist}(W_j, T^\star_{jq}) \leq \delta) f(T_j = T^\star_{jq}), \end{aligned} \tag{3.17}$$

where δ is the maximum displaced distance, $T^\star_{jq} = (T^\star_{jq1}, T^\star_{jq2})$ is the qth $(q = 1, 2, \cdots, Q_j)$ potential true location for the jth EA within the displacement buffer around the W_j, $\text{dist}(W_j, T^\star_{jq}) = \sqrt{(W_{j1} - T^\star_{jq1})^2 + (W_{j2} - T^\star_{jq2})^2}$ is the Euclidean distance between the displaced location $W_j = (W_{j1}, W_{j2})$ and true possible location $T^\star_{jq} = (T^\star_{jq1}, T^\star_{jq2})$, $I(.)$ is an indicator function, and $f(T_j = T^\star_{jq})$ is the distribution of the potential true locations.

However, the true location T_j in (3.17) is unobserved and also, the distribution of the true location data, $f(T_j = T^\star_{jq})$ is unknown. This underlying distribution can be approximated by using available information from different external data sources. Using the available information (e.g., population density, designated administrative boundary restriction, rural urban boundaries, boundary corrections for non-settlement places) that we discussed in detail in Section 3.4, the distribution of potential true locations $f\left(T_j = T^\star_{jq}\right)$ can be approximated as follows:

$$\hat{f}\left(T_j = T^\star_{jq}\right) = \frac{N_{jq} \times I_{\text{DRA}}(T^\star_{jq}) \times I_{\text{UR}}(T^\star_{jq}) \times I_{\text{NS}}(T^\star_{jq})}{\sum_q [N_{jq} \times I_{\text{DRA}}(T^\star_{jq}) \times I_{\text{UR}}(T^\star_{jq}) \times I_{\text{NS}}(T^\star_{jq})]}, q = 1, 2, \cdots, Q_j, \tag{3.18}$$

where N_{jq} is the number of people at location $T^{\star}_{jq}$; $I_{\text{DRA}}(T^{\star}_{jq}) = 1$ if $T^{\star}_{jq}$ falls within the displacement restricted administrative boundaries, otherwise it takes value 0; $I_{\text{UR}}(T^{\star}_{jq}) = 1$ if $T^{\star}_{jq}$ falls in the same region (either urban or rural) from which the EA centroid actually originated, otherwise it takes value 0, and $I_{\text{NS}}(T^{\star}_{jq}) = 1$ if $T^{\star}_{jq}$ does not fall in the non-settlement places e.g., a large water body, forest area etc, otherwise it takes value 0.

Now, under only aggregation, following Groß et al. (2017), the conditional distribution of the *unknown* true coordinates T_{jk} given the aggregated coordinates T_j can be expressed as follows:

$$f(T_{jk} \mid T_j) \propto f(T_j \mid T_{jk}) f(T_{jk}), \tag{3.19}$$

where $f(T_j \mid T_{jk})$ is the conditional distribution of the aggregated coordinates given the true coordinates of the observations. This distribution can be defined on the basis of the type of geography of EA used such as administrative boundary (e.g., mauza-admin 5 in Bangladesh) or a square grid.

If for simplicity we consider a square grid with a side length denoted by γ as the geography of the EA, the distribution $f(T_j \mid T_{jk})$ can be expressed as follows:

$$f(T_j \mid T_{jk}) = \begin{cases} 1 & \text{for } T_{jk} \,\epsilon [T_{j1} - \frac{\gamma}{2}, T_{j1} + \frac{\gamma}{2}] \times [T_{j2} - \frac{\gamma}{2}, T_{j2} + \frac{\gamma}{2}], \\ 0 & \text{otherwise.} \end{cases} \tag{3.20}$$

In the case of the administrative boundary (e.g., mauza-admin 5 in Bangladesh) as the geography of the EA, the above distribution is expressed as follows:

$$f(T_j \mid T_{jk}) = \begin{cases} 1 & \text{for} \quad T_{jk} \,\epsilon \text{ Boundary}(T_j), \\ 0 & \text{otherwise.} \end{cases} \tag{3.21}$$

As the true data T_{jk} and hence, $f(T_{jk})$ is initially unknown, Groß et al. (2017) proposed a Stochastic Expectation-Maximization (SEM) algorithm for estimation. The algorithm employs an initial estimation of $f_H(T^{\star}_{qs})$ based on the T_j values, followed by iterative simulations of T_{jk} from $f(T_{jk} \mid T_j)$ and subsequent re-estimation of $f(T_{jk})$ until convergence is achieved. This approach can be seen as a variation of the expectation-maximization (EM) algorithm (Dempster et al., 1977).

However, under both aggregation and displacement, the aggregated coordinates T_j are unknown. Therefore, the Groß et al. (2017) method cannot be used. A methodological extension is required to deal with the problems of aggregation and displacement. Thus, under both aggregation and

displacement, for estimating the density $\hat{f}_H(T^{\star}_{qs})$ we propose an estimation method by using the equation in (3.16) through the above two conditional distributions (3.17) and (3.19) simultaneously. We want to clearly state that the boundaries of the EA geography at which the data has been aggregated should be available in order to use the method.

3.5.1.1 Estimation using the SEM algorithm under aggregation and displacement

The computational steps of the algorithm for the proposed method are described below:

- Step 1: For each EA $j = 1, 2, \ldots, R$ take the WorldPop grid of points $(T^{\star}_{jq1}, T^{\star}_{jq2})$, $(q = 1, 2, \cdots, Q_j)$ around (W_{j1}, W_{j2}) of width 2δ that covers the complete (T_{j1}, T_{j2}) space, where δ is the maximum displacement distance.

- Step 2: For each EA $j = 1, 2, \ldots, R$ pseudo-samples $T^{\star}_j = (T^{\star}_{j1}, T^{\star}_{j2})$ of the unknown points $T_j = (T_{j1}, T_{j2})$ are drawn from $f(T_j|W_j)$ in (3.17) with $\hat{f}\left(T_j = T^{\star}_{jq}\right) = \frac{N_{jq} \times I_{\text{DRA}}(T^{\star}_{jq}) \times I_{\text{UR}}(T^{\star}_{jq}) \times I_{\text{NS}}(T^{\star}_{jq})}{\sum_q [N_{jq} \times I_{\text{DRA}}(T^{\star}_{jq}) \times I_{\text{UR}}(T^{\star}_{jq}) \times I_{\text{NS}}(T^{\star}_{jq})]}, q = 1, 2, \cdots, Q_j$, in (3.18).

- Step 3: Obtain a pilot estimate $\hat{f}_H(T^{\star}_{qs})$ of $f_H(T^{\star}_{qs})$ in (3.15) using the sample aggregated data set of coordinates $T^{\star}_j$ $(j = 1, \cdots, R)$ from Step 2 with an initial bandwidth matrix

$$H^{(0)} = \begin{pmatrix} l & 0 \\ 0 & l \end{pmatrix},$$

  where l takes a value that is sufficiently large, such as twice the range of the observation coordinates and $T^{\star}_{qs} = (T^{\star}_{q1}, T^{\star}_{s2})$ for $q = 1, \cdots, L$ and $s = 1, \cdots, L$ denote the grid points that covers the complete $T_{jk} = (T_{jk1}, T_{jk2})$ space i.e., all possible true locations and L is the grid size. The generation process of these grid points is presented in sub-section 3.3.1.

- Step 4 (S-step): For $j = 1, 2, \ldots, R$ pseudo-samples $T^{\star}_{jk} = (T^{\star}_{jk1}, T^{\star}_{jk2})$ of the true unknown points $T_{jk} = (T_{jk1}, T_{jk2})$ are selected by sampling from the conditional distribution $f(T_{jk}|T_j = T^{\star}_j)$ in (3.19). This conditional distribution corresponds to the current density estimate restricted to the EA boundary e.g., the square grid around the centre of $T^{\star}_j$.

- Step 5 (M-step): Estimate the bandwidth matrix H by using the multivariate plug-in estimator proposed by Wand and Jones (1994) and re-compute the bivariate kernel density $\hat{f}_H(T^{\star}_{qs})$ using the pseudo-samples generated in Step 4. Alternatively, other bandwidth selection methods can also be applied.

- Step 6: Repeat Steps 4-5 for a total of C_1 (burn-in iterations) + C_2 (additional iterations).

- Step 7: Exclude the C_1 burn-in density estimates and calculate the density estimate of $f_H(T^{\star}_{qs})$ by taking the average of the remaining C_2 density estimates $\hat{f}_H(T^{\star}_{qs})$.
- Step 8: Repeat Steps 2-7, C_3 times and calculate the final density estimate by taking average over C_3 estimates. Larger values for C_1, C_2, and C_3 —for example, $C_1 = 5$, $C_2 = 20$, and $C_3 = 100$—enhance the realization of the unknown true coordinates of observations and subsequently improve the density estimates, accounting for both aggregation and displacement errors. The sensitivity of the results to different combinations of C_1, C_2, and C_3 can be assessed by evaluating the quality of the density estimates, such as the RIMSE of the density estimates defined in (3.27).

3.5.2 Domain estimation using aggregated and randomly displaced coordinates

Our next aim is to develop a method to obtain estimates of the domain (e.g., upazila) parameters of interest accounting for misplacement uncertainty due to aggregation and random displacement errors. Following the description in sub-section 3.3.2 and the equation (3.9), a parameter of interest defined for domain U_i e.g., the proportion of poor households is expressed as

$$\bar{Y}_i = \frac{\sum_U Y_{jk} I_i(T_{jk})}{\sum_U I_i(T_{jk})}, i = 1, \dots, M; k = 1, \dots, N_j; j = 1, \dots, R, \tag{3.22}$$

where Y_{jk} denotes a variable measured at the kth unit in the jth EA level and $I_i(T_{jk}) = 1$ if $T_{jk} \in U_i$, otherwise it takes value 0.

Using a similar expansion to (3.22) with W_j in place of T_{jk} may lead to bias in domain estimates whenever the aggregated and displaced W_j is misplaced i.e., belongs to a different domain than T_{jk}. However, we aim to estimate the parameter $\bar{Y}_i$ using the values W_k accounting for misplacement uncertainty via $f(T_{jk} \mid T_j, W_j)$ in (3.16) that corrects for misplacement error. In particular, we can generate pseudo samples of the unobserved true data given the observed data from $f(T_{jk} \mid T_j, W_j)$ but we need to estimate $f(T_j)$ and $f(T_{jk})$. In that case, we propose an estimator of $\bar{Y}_i$ using pseudo samples $T^{\star}_{jk}$ of the true unknown points T_{jk} drawn by using the proposed SEM algorithm that we presented in subsection 3.5.1.1. The proposed domain estimator is expressed as

$$\hat{\bar{Y}}_i = \frac{\sum_U Y_{jk} I_i(T^{\star}_{jk})}{\sum_U I_i(T^{\star}_{jk})}, i = 1, \dots, M; k = 1, \dots, N_j; j = 1, \dots, R, \tag{3.23}$$

where $I_i(T^{\star}_{jk}) = 1$ if $T^{\star}_{jk}$ belongs to the domain i, otherwise it takes value 0.

Iterative samples of true location data out of all possible true locations are needed for a better un-

derstanding of the true location. Therefore, for each complete set of sample coordinates $T^{\star}_{jk} = (T^{\star}_{jk1}, T^{\star}_{jk2})$, from Step 4 in the proposed SEM algorithm in 3.5.1.1, we obtain the domain parameter estimates $\hat{\bar{Y}}_i$ in (3.23) and obtain the final estimate by taking an average of estimates over $C_2 \times C_3$ samples.

3.6 Simulation study

This empirical study aims to evaluate the performance of the proposed methods. Specifically, in section 3.6.1, we focus on the effects of random displacement. Here, we assess the performance of two proposed methods: KDE and EDC (detailed in sections 3.3 and 3.4, respectively), the method by Warren et al. (2016b), and a naive method that disregards displacement errors when estimating the upazila proportions of poor households. We use the method by Warren et al. (2016b) to select the most probable domain (upazila) for each household. Hereafter, this is referred to as the Most Probable Domain (MPD) method. Subsequently, in section 3.6.2, we move our focus to both aggregation and displacement of household coordinates. We evaluate the performance of our proposed external data and kernel density-based (KDE-ED) method, described in section 3.5. This is compared against a naive method that ignores both aggregation and displacement errors when estimating 1) upazila proportions of poor households, and 2) the density of the poor households. These methods under 1) only displacement and 2) aggregation and displacement can be applied in the case of both population/register and sample data. In the case of register data, we assessed the impact of 1) displacement and 2) aggregation and displacement errors on density and upazila proportion estimates using repeated displaced data and aggregated and displaced data, respectively. In this simulation study, to compare the performance of the methods, we compute quality measures such as bias, RMSE of the estimates by using only the (simulated) register data under the geography of Bangladesh.

3.6.1 Analysis of simulated register data under the displacement of household coordinates

In this simulation study, we aim to assess the performance of the proposed KDE and EDC estimators for domain proportions compared with the MPD and the naive estimators under the displacement of household locations. We generate a population of household locations (geo-coordinates) from the $100m \times 100m$ gridded population raster from WorldPop that estimates the population of Bangladesh, adjusted using the 2011 census population, where each grid cell represents the number of people per pixel. We described the WorldPop data in detail in Chapter 1. The constrained

population raster in Bangladesh for 2020 can be downloaded from WorldPop (Bondarenko et al. (2020)). First, we calculate the number of households for each grid by dividing the grid cell population by 4.4, which is the average household size according to the 2011 Bangladesh population census. While average household sizes vary across divisions (admin 1) from 4.1 to 5.5, we use the overall country average for simplicity. Moreover, as this empirical study mainly evaluates the impact of random displacement of all household coordinates, little variations in household size likely have insignificant effects on the assessment.

Next, we select households' locations randomly from each WorldPop grid cell. Since the number of households in the country is large, instead of the whole geography of Bangladesh, we select the district Khulna of Bangladesh with 14 upazilas and generate 20k household coordinates randomly within WorldPop grid cells by determining the sample number of households for each grid proportionally to the population number of households. We note that the total number of people in the district of Khulna is about two million, according to the WorldPop. We determine whether a household falls in the rural or urban area by using the rural and urban boundaries from the Bangladesh 2011 population census. For each household in the population/register data, as a variable of interest, we generate whether a household is poor. We aim to estimate the proportion of poor households for each upazila. For this simulation study, we generate data (poor or non-poor) for the households within each upazila by sampling from the Binomial distribution with parameters equal to the estimated upazila proportions which have been produced using the Elbers,Lanjouw, and Lanjouw (ELL) poverty mapping method (Faizuddin et al., 2012). The summary statistics of the upazila proportions of poor households based on the simulated population is given in Table 3.1.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	0.180	0.344	0.383	0.376	0.412	0.499

Table 3.1: Summary statistics of the upazila proportions of poor households using the simulated population.

We consider the selected 20k household coordinates as the true household coordinates. Then, the displaced household coordinates are generated by applying the DHS displacement algorithm (Burgert et al., 2013) to the true household coordinates. In particular, data are displaced within the restricted district boundary by using the maximum displaced distance in line with the DHS displaced mechanism that is 2km for urban households, and 5km for rural households with 1% of rural households being randomly allocated 10km maximum distance. While the specific 1% of rural households is unknown to analysts, this small proportion introduces limited uncertainty. This potential impact can be addressed by replicating the DHS displacement mechanism, as incorpo-

rated in our proposed methods. However, only displaced household coordinates are provided to the data analysts. The true and displaced household locations are presented in Figure 3.6. Due to the random displacement, the original locations move to different locations which leads to change in the spatial pattern of the data.

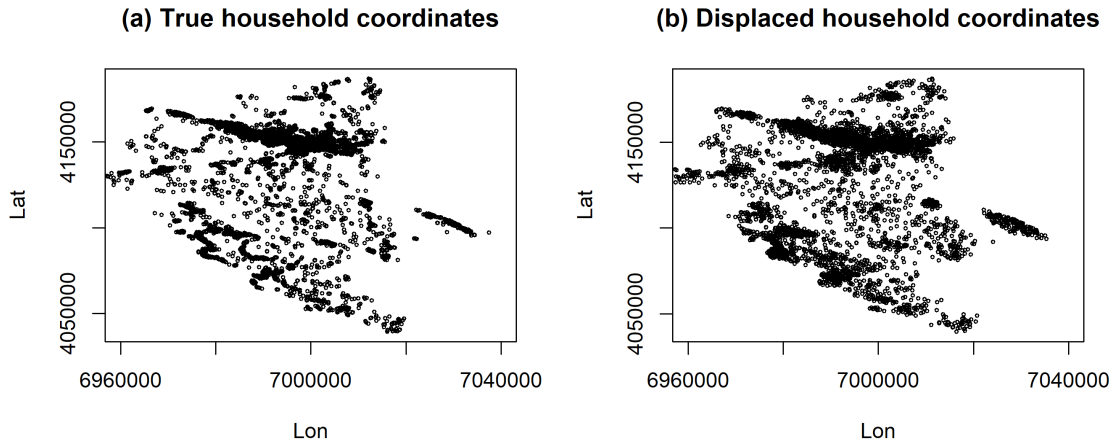

Figure 3.6: Spatial distribution of the (a) true household coordinates and (b) displaced household coordinates.

We obtain the estimates of upazila population proportions of poor households by using the proposed KDE and EDC estimators, the MPD estimator and the naive estimator that ignores the misplacement error due to the displacement process. In the case of upazila proportion estimates, the KDE based estimator uses the kernel density corresponding to the distribution of household (continuous) coordinates. For density estimation, we employ a Bivariate Gaussian Kernel along with the plug-in bandwidth selection method proposed by Wand and Jones (1994). This is implemented by using the R functions **Hpi** (bandwidth selector) and **kde** (kernel density estimation) provided by the **ks** package Duong et al. (2022). However, the impact of household locations misplacement on the estimates of upazila proportions is assessed by obtaining the estimates under the repeated displaced data. In particular, the following four estimators for upazila proportions are evaluated in this simulation study:

i) naive Estimator for upazila proportion that ignores the household misplacement error due to the displacement process.

ii) KDE based Estimator for upazila proportion, the proposed SEM estimator with $A_1 = 20$ burn-in iterations, $A_2 = 200$ sample iterations and grid size $L = 400$. Under this proposed method, to estimate the kernel density of the distribution of unknown true coordinates, we employ varying values of burn-in iterations A_1 and sample iterations A_2 to investigate their

impact on the results. Larger values of A_1 and A_2 enhance the precision in estimates, allowing for improved realization of the unknown true coordinates. However, employing a setting of $A_1 = 20$ and $A_2 = 200$ establishes a balance between achieving consistency in estimates and optimising computational time under random displacement.

iii) EDC based Estimator for upazila proportion, the proposed external data based estimator, iterated with $B_1 = 200$ pseudo-samples and $B_2 = 20$ additional sample steps using the WorldPop grid points and additional information from external data. Larger pseudo-samples, B_1, of unknown true locations yield better realizations of the true coordinates and, consequently, more accurate predicted domain membership probabilities. We have checked that the results with $B_1 = 200$ are reasonably close to those obtained using additional samples, $B_2 = 20$. Therefore, it is recommended to use at least $B_1 = 200$ pseudo-samples under random displacement.

iv) MPD Estimator for upazila proportion, the MPC method proposed by Warren et al. (2016b) which we described in Chapter 2. We apply this method for the most probable domain assignment to household locations and obtain the upazila proportion estimates.

The simulation steps (generating the displaced data and obtaining the estimates of upazila proportions) are independently repeated 200 times. We compare the performance of the naive, the MPD, the KDE and the EDC estimators for upazila proportion by computing the Root-Mean-Square-Error (RMSE), Bias, and Absolute Bias (AB) for the ith $(i = 1, \cdots, M)$ upazila proportion estimate over $B = 200$ replications as follows:

$$\text{RMSE}_i = \sqrt{\frac{\sum_{b=1}^{B}(\hat{\bar{Y}}_{bi} - \bar{Y}_i)^2}{B}} \tag{3.24}$$

$$\text{Bias}_i = \frac{\sum_{b=1}^{B}(\hat{\bar{Y}}_{bi} - \bar{Y}_i)}{B} \tag{3.25}$$

$$\text{AB}_i = \frac{\sum_{b=1}^{B} \mid \hat{\bar{Y}}_{bi} - \bar{Y}_i \mid}{B} \tag{3.26}$$

where $\hat{\bar{Y}}_{bi}$ represents the estimated proportion in upazila i using any of the aforementioned methods under random displacement, while $\bar{Y}_i$ represents the true proportion for upazila i in replication round b.

3.6.1.1 Results and discussion under random displacement using register data

We present and discuss the simulation study results for the MPD method, the proposed methods KDE and EDC and the naive method under random displacement. The 200 displaced data sets of household coordinates are used to obtain the estimated proportions of poor households for $M = 14$ upazilas and their Biases, Absolute Biases (ABs) and RMSEs.

The summary statistics of the absolute biases of the estimated upazila proportions of poor households over 200 replications for each estimator under displacement of household coordinates are given in Table 3.2. We observe that the two proposed methods, KDE and EDC, have lower absolute bias than the MPD and the naive method that ignores the misplacement error. For instance, the mean absolute biases for the estimates using the proposed KDE and EDC estimators are lower than those of the naive and MPD estimators across the 14 upazilas. Although the MPD estimator has a marginally larger mean absolute bias over the 14 upazilas than the naive estimator, it has a slightly smaller median absolute bias than the naive estimator. However, among the two proposed methods, the external data-based EDC estimator performs best in terms of lower absolute bias.

Estimator	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
naive	0.0037	0.0080	0.0142	0.0181	0.0208	0.0743
KDE	0.0023	0.0059	0.0082	0.0169	0.0194	0.0699
EDC	0.0017	0.0045	0.0068	0.0135	0.0153	0.0544
MPD	0.0034	0.0074	0.0137	0.0189	0.0202	0.0790

Table 3.2: Summary statistics of the absolute biases of the estimated upazila proportions of poor households over 200 replications for each estimator under displacement of household coordinates.

In order to assess the performance of the two proposed KDE and EDC estimators in terms of variability, the statistics for the distribution of the RMSEs of the estimated upazila proportions of poor households over 200 replications for each estimator are given in Table 3.3. We see that the proposed estimators have lower RMSEs of the estimates compared with the other two approaches. For example, the EDC estimator has the lowest mean RMSE of the estimated proportion over 14 upazilas, followed by the KDE, naive, and MPD estimators in that order. Also, among these estimators, the EDC estimator has the lowest 25th percentile of the RMSE values for the upazila estimates, followed by the KDE, MPD, and naive estimators. Moreover, the maximum and the minimum values are also improved using the EDC estimator compared with the naive estimates, which show a large RMSE. We notice that overall the MPD and the naive methods perform similarly. The EDC estimator, which uses additional information from external data, clearly outperforms the other estimators regarding lower RMSE. However, if we had more uncertainty measurements for the external data sources, the KDE might perform better than the EDC. Also, while RMSE incorporates both bias

and variability, our observations suggest that the patterns in RMSE might be influenced by the bias of the estimates.

In summary, these results in the considered simulation scenario indicate that the proposed KDE and EDC-based approaches give a considerable advantage, for lower RMSE, over the two other methods one of which is a simple naive approach that ignores the misplacement error. Specifically, the EDC method provides an improvement of approximately 30% in mean RMSE over the naive method (reduction from 0.0210 to 0.0147).

Estimator	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
naive	0.0047	0.0103	0.0176	0.0210	0.0238	0.0762
KDE	0.0028	0.0066	0.0103	0.0182	0.0207	0.0706
EDC	0.0020	0.0052	0.0087	0.0147	0.0167	0.0551
MPD	0.0040	0.0089	0.0159	0.0217	0.0241	0.0805

Table 3.3: Summary statistics of the RMSEs of the estimated upazila proportions of poor households over 200 replications for each estimator under displacement of household coordinates.

Nevertheless, as the proposed estimators may yield improved estimates in certain upazila proportions while not in others, upazila-level measures of bias and RMSE are also necessary to explore the accuracy and precision of the estimates within each upazila. Figures 3.7-3.8 show that the biases and absolute biases of the estimates of upazila proportions of poor households over 200 replications for each method: the naive, the MPD, the two proposed KDE and EDC methods. A closer inspection of the results reveals that the KDE estimates' absolute bias is less than the naive estimates' absolute bias in 9 upazilas out of 14 (64.3%). Also, for 12 upazilas out of 14 (85.7%) the EDC estimates' absolute bias is less than the naive estimates' absolute bias. On the other hand, the naive estimates' absolute bias is less than the MPD estimates' absolute bias in 8 upazilas out of 14 (57.1%).

Estimates for some upazilas using the naive method are more biased, primarily for two reasons. Firstly, these upazilas are more affected by misplacement errors, and secondly, they have smaller areas in terms of square kilometres. As shown in Figure 3.8(a-c), upazilas with smaller areas have higher misplacement errors, resulting in a greater absolute bias in the estimates when using the naive method that ignores misplacement errors. In contrast, the proposed EDC method corrects the bias of these estimates compared to other methods, particularly for upazilas with small areas. As an example, for upazila 1 in Figures 3.7-3.8(a) or upazila 10 in Figures 3.7-3.8(b), which has the smallest size in square kilometers, the proposed methods outperform the naive method by correcting the bias in estimate. When misplacement errors are minimal, all methods perform similarly, as expected. We note that misplacement errors depend not only on the area of the upazila but also

on the shape of the upazila and whether people live close to the boundary. However, we observe notably poor performance in the proposed methods for upazila 9, as indicated in Figures 3.7-3.8(a), or upazila 14 as shown in Figures 3.7-3.8(b). Although this upazila is moderate in area size and has the highest percentage of misplaced households, the proposed methods demonstrate a larger bias than the naive method. This exceptional performance could be due to the complex shape of this upazila and the intersections of the displacement buffer with multiple upazila boundaries. Specifically, the maximum buffer around the displaced household coordinates within this upazila intersects with many upazila boundaries, introducing additional uncertainty when the proposed methods attempt to correct bias in the estimates.

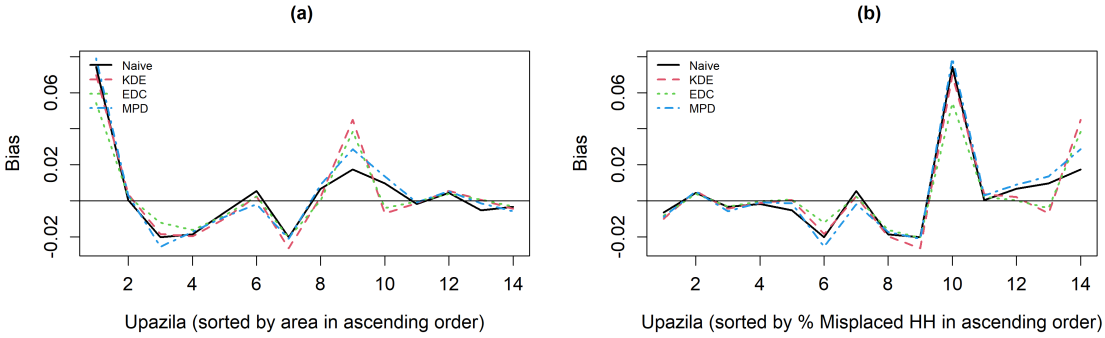

Figure 3.7: Plot of the biases of the estimated upazila proportions of poor households over 200 replications for each estimator under displacement of household coordinates. The areas are sorted by square kilometres.

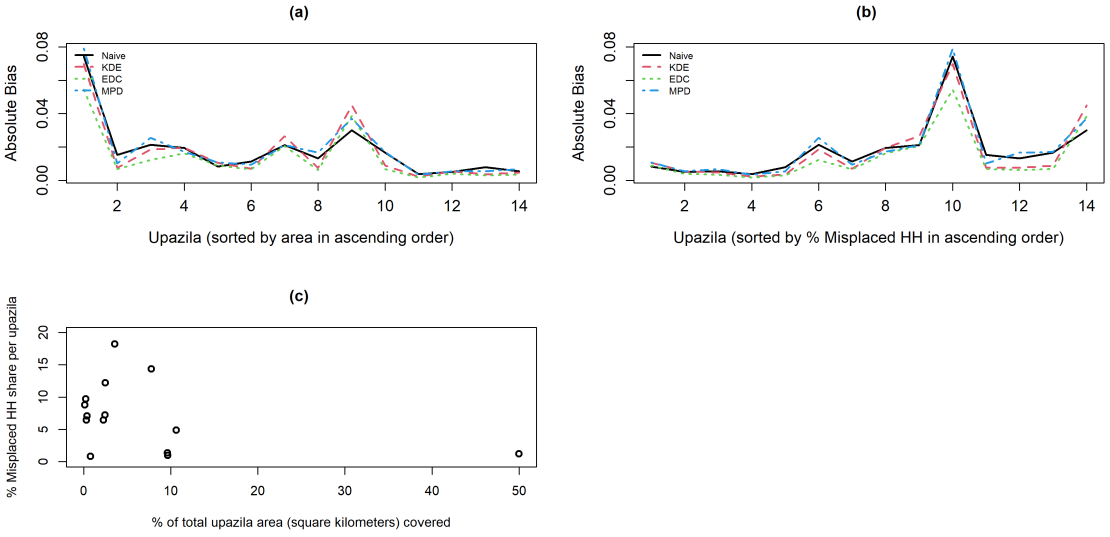

Figure 3.8: (a-b) Plot of the ABs of estimated upazila proportions of poor households over 200 replications, for each estimator, under displacement of household coordinates. These are sorted by area (in square kilometres) and misplacement level. (c) presents the relationship between the percentage of the total upazila area covered and the share of misplaced households for each upazila.

Figure 3.9 illustrates the upazila-level RMSE for the four estimates. The empirical RMSE of the KDE estimates decreases in 11 out of the 14 upazilas compared to the RMSE of the naive estimates. In comparison, the RMSE of the EDC estimates is lower than that of the naive estimates in 12 out of the 14 upazilas. In the case of the MPD method, 50% (7 out of 14) upazilas have lower RMSE than the naive method. Therefore, the proposed methods outperform the naive method for lower empirical RMSE of the estimates. Also, regarding the performance of the methods in terms of RMSE, taking into account the size of the upazila area and misplacement error, we have similar observations to those we observe in the case of absolute bias.

Overall, in the given simulation scenario, the empirical study indicates that the proposed KDE and EDC estimators improve the upazila estimates in terms of both bias and variability for most upazilas, especially when there is more misplacement of household locations. However, the naive estimator is preferable for upazilas with minimal misplacement of household locations.

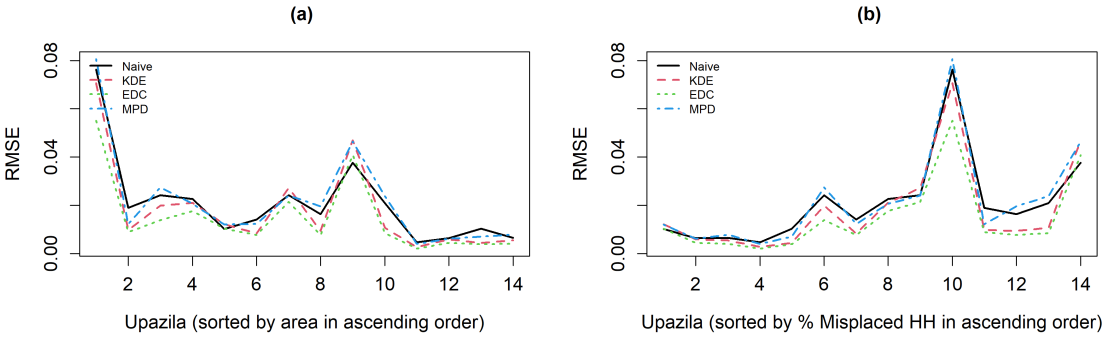

Figure 3.9: Plot of the RMSEs of the estimated upazila proportions of poor households over 200 replications for each estimator under displacement of household coordinates.

3.6.2 Analysis of simulated register data under aggregation and displacement of household coordinates

The purpose of this empirical study is to evaluate the performance of the proposed method (KDE-ED) over the naive method under both aggregation and displacement of household coordinates by computing quality measures for the estimated 1) upazila proportions of poor households and 2) densities of the poor households. Following the same process described in subsection 3.6.1, we generate the simulated register data. In particular, we calculate the number of households for each grid by dividing the grid cell population by 4.4 and select households' locations randomly from each WorldPop grid cell. In this section, we randomly selected approximately 525k household locations from 14 upazilas, in contrast to the 20k households chosen for only displacement in section 3.6.1. This increase is due to computational constraints related to the number of displaced locations. Under both aggregation and displacement, this number decreases to the number of EAs, since household locations are aggregated at the EA level prior to displacing EA points. The effect of aggregation on density estimation is influenced by the number of households: the greater the number of households, the more pronounced the aggregation effects (Groß et al., 2017). Therefore, in this empirical study, we use a sizable, constant number of households to evaluate the combined impact of aggregation and random displacement errors. For each of these 525k households, we generated binary values indicating whether a household is categorised as poor or non-poor. The summary statistics of upazila proportions of poor households based on the simulated population is given in Table 3.4.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	0.211	0.354	0.381	0.385	0.418	0.514

Table 3.4: Summary statistics of upazila proportions of poor households from the simulated population.

The generated about 525k household coordinates are treated as the true household coordinates. We determine whether a household falls in the rural or urban area by using the rural and urban boundaries from the Bangladesh 2011 population census. Also, we consider both irregular and simple shapes as the geography of EAs. In particular, in this study, we consider mauza (admin 5 of Bangladesh) and a square grid (with side length denoted by γ) as the geography of EAs. The extent of the aggregation error within a mauza boundary under a register data is fixed. However, in the case of a square grid, as the side length increases, the extent of aggregation error also increases.

In order to protect respondents' confidentiality, first, the household coordinates within an EA are aggregated to the EA centroid (i.e., the centroid of the mauza/square grid). Next, the EA centroid

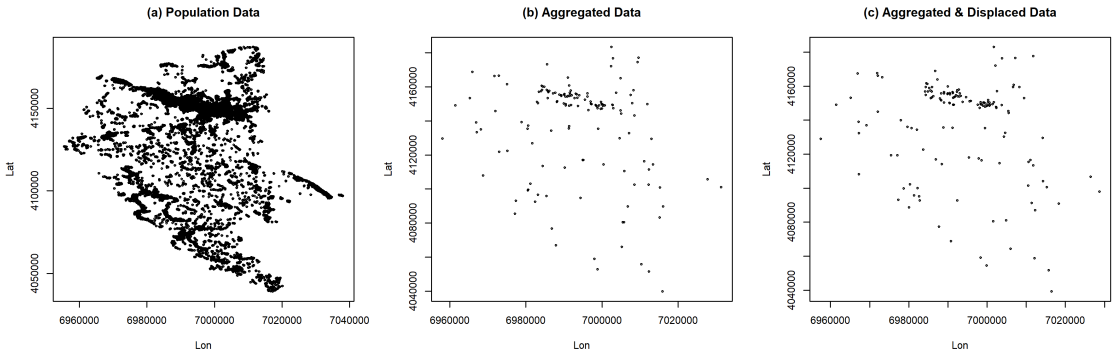

Figure 3.10: Spatial distribution of the (a) true household coordinates, (b) aggregated EA (mauza) centroid coordinates, and (c) displacement of aggregated EA centroid data by applying DHS displacement algorithm with 5000m maximum displaced distance.

coordinates are displaced using the DHS random displacement algorithm that uses a random direction and random distance which is denoted by $\delta = 2000m, 4000m$ etc. Therefore, only the displaced EA centroid coordinates are provided to researchers. In this study, we examine the ability of the proposed KDE-ED method to correct measurement error under various scenarios for the intensity of the aggregation and displacement error processes. The intensity of the aggregation error depends on the size of the EA geography, while the intensity of the displacement error depends on the maximum displacement. As described in Chapter 1, EAs are typically constructed with a focus on counting populations, usually encompassing between 100 to 300 households, rather than on geographic size (measured in square meters). Due to this, their specific sizes can vary based on population density; for instance, areas with higher populations tend to have smaller geography of EAs. Furthermore, households in adjacent areas, rather than distant ones, may be grouped to form an EA, facilitating easier access for enumerators. Considering these factors, this empirical study, for the sake of simplicity, assumes that the geography of the EA is smaller than the maximum displacement buffer. However, we do consider varying EA geography sizes and displacement distances to examine the intensification of the aggregation and displacement errors more thoroughly. In particular, we consider the following scenarios: (I) $\gamma = 1000m, \delta = 2000m$, (II) $\gamma = 1000m, \delta = 4000m$, (III) $\gamma = 3000m, \delta = 4000m$, (IV) $\gamma = 4000m, \delta = 5000m$. The true household coordinates, aggregated and displaced household coordinates under mauza and square grid are presented in Figures 3.10-3.11. We observe that the household coordinates are concentrated at the centroid of the mauza or the square grid and its original spatial pattern has been changed due to aggregation and displacement processes. Moreover, the household coordinates can be misplaced if they are moved at locations other than their original upazila geography. In our simulation setting, when a mauza is used as EA geography, misplacement can occur only due

to displacement, not the aggregation process. This is because aggregation is confined within upazila boundaries since the geography of a mauza is nested within the upazila boundary. On the other hand, when a square grid is used as an EA, the household coordinates may be misplaced at the upazila level due to both aggregation and displacement processes, as neither aggregation nor displacement is restricted within the upazila boundary. Therefore, we obtain upazila proportion estimates by using the proposed KDE-ED method under both aggregation and displacement, specifically in instances where the aggregation process is restricted within upazila boundaries, as occurs when a mauza is used as the EA geography.

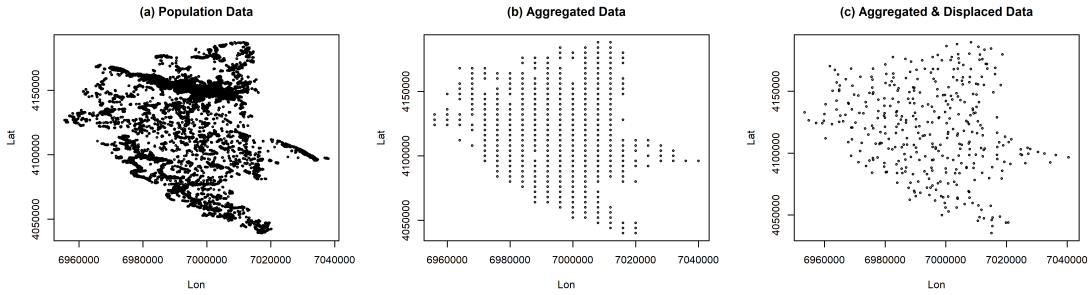

Figure 3.11: Spatial distribution of the (a) true household coordinates, (b) aggregated EA centroid coordinates (square grid: 4000m X 4000m), and (c) displacement of aggregated EA centroid data by applying DHS displacement algorithm with 5000m maximum displaced distance.

We estimate 1) upazila proportions of poor households and 2) the density of the population of poor households by using the proposed KDE-ED method and the naive method that ignores aggregation and displacement errors. In the algorithm for the proposed KDE-ED method, as outlined in Section 3.5.1.1, using larger values for C_1, C_2, and C_3 improves the realization of the unknown true coordinates of observations and subsequently enhances the consistency of estimates—such as upazila proportion—by accounting for both aggregation and displacement errors. We have examined that the setting $C_1 = 5, C_2 = 20$ and $C_3 = 100$ iterations offers a favorable balance between achieving consistent results and managing computational costs, accounting for both aggregation and displacement errors. The impact of aggregation and displacement errors on the density estimates and upazila proportion estimates is assessed by obtaining the estimates under the repeated aggregation and displacement of population household coordinates. Although detailed in Section 3.5, we restate here to prevent any confusion that, for the observational unit, we consider only poor households to obtain density estimates. In other words, we use the coordinates of poor households as input observations to estimate the population density of poor households.

We evaluate the performance of the density estimators: the naive or the proposed by using the

root mean integrated squared error (RMISE) of the density estimates. The RMISE is estimated by approximating it with a Riemann sum over a finely spaced grid that is evenly distributed, and it can be expressed as follows:

$$\text{RMISE}\{\hat{f}(T^\star)\} = \sqrt{E\left[\int \{f(T^\star) - \hat{f}(T^\star)\}^2 dT^\star\right]} \approx \sqrt{\left[\frac{1}{m_g}\sum_{l=1}^{m_g}\{f(g_l) - \hat{f}(g_l)\}^2 L_{g1}^2\right]}, \tag{3.27}$$

where m_g is the number of grid points g_l and L_{g1} is the grid width.

We employ a Bivariate Gaussian Kernel and the plug-in bandwidth selector approach suggested by Wand and Jones (1994) to estimate the density using the proposed estimator and the naive estimator that ignores aggregation and displacement errors. We derive the 'true' density from the original, undistorted data using a similar Bivariate Gaussian kernel and plug-in bandwidth selector for comparison and evaluation. This derived 'true' density serves as a benchmark in our evaluations, representing the true values against which the performances of the proposed and naive estimators are assessed.

The simulation steps (generating the aggregated and displaced data and obtaining the estimates) are independently repeated 200 times. We compare the performance of the naive and the KDE-ED estimator by computing the RMSE, Bias, and Absolute Bias (AB) for the upazila estimates and the mean RMISE of the density estimates over $B = 200$ replications under both aggregation and random displacement by using equations (3.24)-(3.27).

3.6.2.1 Results and discussion under aggregation and displacement using register data

We present and discuss the results of the empirical study for the proposed KDE-ED method and the naive method. Firstly, we present the results of the estimates for the proportion of poor households in each upazila, and secondly, we present the results of the density estimates for poor households.

Estimator	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
naive	0.0030	0.0072	0.0103	0.0128	0.0179	0.0325
KDE-ED	0.0017	0.0052	0.0072	0.0111	0.0132	0.0378

Table 3.5: Summary statistics of the absolute biases of the estimated upazila proportions of poor households over 200 replications for each estimator under aggregation and displacement of household coordinates with mauza geography.

The summary statistics of the absolute biases of the estimated upazila proportions of poor households over 200 replications for each estimator under aggregation and displacement of household coordinates with the mauza scenario are given in Table 3.5. We observe that the proposed method

(KDE-ED) has a lower absolute bias than that of the naive method that ignores aggregation and displacement errors. For example, the proposed KDE-ED estimator has the smallest mean and median absolute biases for the estimates of the 14 upazilas compared to the naive estimator. Also, Figures 3.12-3.13 show that most of the upazila estimates by using the proposed method have lower bias than the naive method. Therefore, in the simulation scenario under consideration, the presented KDE-ED method offers a significant advantage over the naive method that disregards misplacement errors due to aggregation and random displacement processes.

Estimator	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
naive	0.0048	0.0099	0.0137	0.0162	0.0222	0.0404
KDE-ED	0.0024	0.0062	0.0088	0.0123	0.0145	0.0395

Table 3.6: Summary statistics of the RMSEs of the estimated upazila proportions of poor households over 200 replications for each estimator under aggregation and displacement of household coordinates with mauza geography.

To evaluate the performance of the proposed KDE-ED estimator in terms of variability, summary statistics of the RMSEs of the estimated proportions of poor households over 200 replications for each estimator under aggregation and displacement of household coordinates with the mauza scenario are given in Table 3.6. As this table shows, the proposed estimator outperformed the naive approach, indicating lower RMSE values for the estimates. Additionally, the KDE-ED estimator demonstrated better performance than the naive estimator across various summary statistics such as mean, 75th percentile, maximum and minimum values of RMSEs. This indicates that the KDE-ED estimator produces more accurate and precise estimates than the naive estimator, thus representing a better choice for upazila estimates under aggregation and displacement.

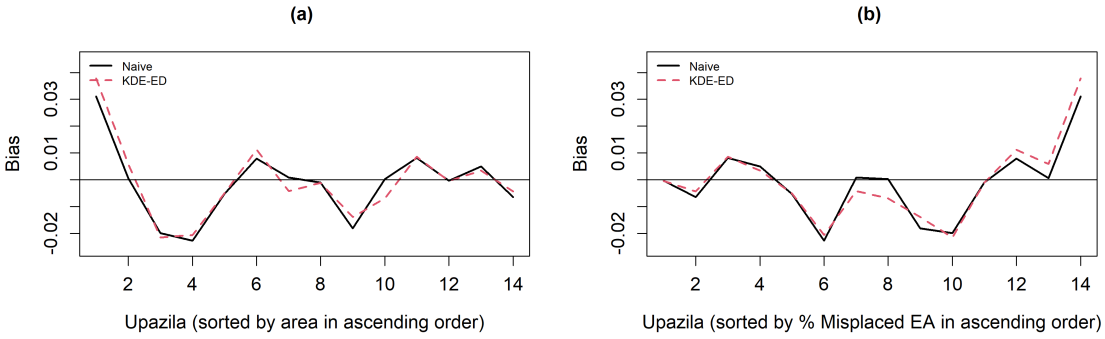

Figure 3.12: Plot of the biases of the estimated upazila proportions of poor households over 200 replications for each estimator under aggregation and displacement household coordinates with the mauza scenario.

We observe from Figure 3.14 that most of the upazila estimates using the proposed method have lower RMSE than the naive method. Therefore, these results in the simulation scenario under

consideration indicate that the proposed KDE-ED method gives a notable advantage, for lower RMSE, over the naive method that ignores the misplacement error due to aggregation and random displacement processes.

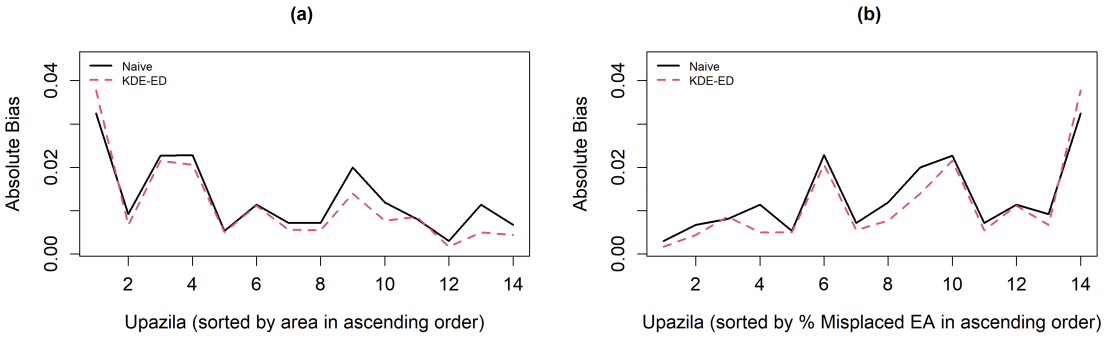

Figure 3.13: Plot of the ABs of the estimated upazila proportions of poor households over 200 replications for each estimator under aggregation and displacement household coordinates with the mauza scenario.

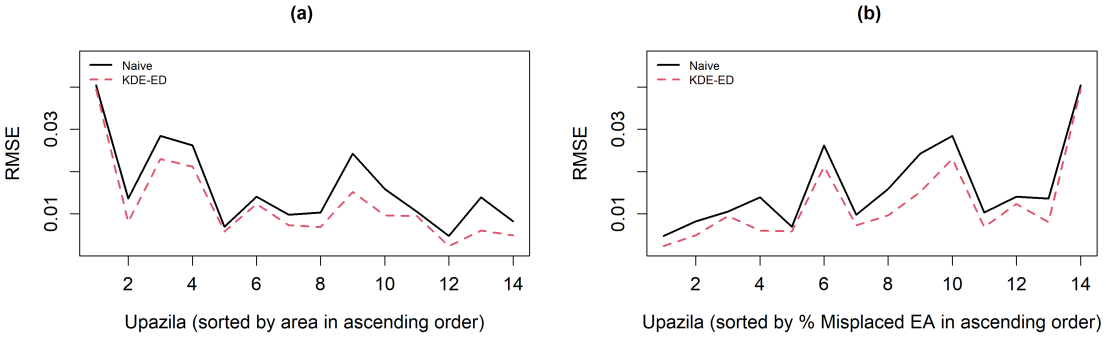

Figure 3.14: Plot of the RMSEs of the estimated upazila proportions of poor households over 200 replications for each estimator under aggregation and displacement household coordinates with the mauza scenario.

We now present and discuss the results of the density estimates of poor households from analysing the register data under aggregation and displacement using the naive and the proposed KDE-ED estimator. Under various scenarios with a grid for the intensity of the aggregation and displacement error processes, we investigate the ability of the proposed KDE-ED estimator to provide more precise estimates than the naive estimator that ignores the aggregation and displacement errors.

Table 3.7 presents mean RMISEs of the estimates of the density of the population of poor households over 200 replications by the naive and the proposed density estimators and for various scenarios of aggregation and displacement errors. It is apparent that the overall structure of the density estimates is well reflected by the KDE-ED method. For scenario I, which involved small aggregation and displacement errors, the mean RMISE of the density estimates is about two times higher

for the naive method compared to the KDE-ED method, averaged over the 200 replications. This indicates that in the case of small measurement errors, the proposed method performs slightly better than the naive method. As the measurement error increases, the proposed KDE-ED method outperforms the naive method in estimating the density of poor households. For example, in scenario IV with large aggregation and displacement errors, the mean RMISE value of the density estimates is about four times higher for the naive method than the KDE-ED method, averaged over the 200 replications. While the naive estimator ignores the aggregation and displacement errors, the proposed KDE-ED estimator corrects these errors and demonstrates better performance, particularly for large grid sizes and displacement distances.

Scenario	I	II	III	IV
Estimator	$\gamma = 1000, \delta = 2000$	$\gamma = 1000, \delta = 4000$	$\gamma = 3000, \delta = 4000$	$\gamma = 4000, \delta = 5000$
naive	3.19	5.44	12.24	25.04
KDE-ED	1.70	1.89	3.39	6.45

Table 3.7: Register data: Mean RMISE (Results $\times 10^{-8}$) for the naive and proposed multivariate kernel density estimators for various aggregation and displacement parameters with grids.

Figures 3.15-3.16 provide a visual comparison of the density of poor households and contour plots for each estimator under various levels of aggregation and displacement of household coordinates. The contour plots represent the relative density of poor households in each region, with higher numbers indicating higher density. For example, a '50' signifies a higher concentration of poor households compared to a '25'. Notably, the proposed KDE-ED estimator can accurately capture and preserve the underlying structure of the density for different aggregation and displacement levels. This contrasts the naive estimator, which fails to capture the same structure and exhibits a more scattered and less distinct density pattern. In particular, scenario IV, with large aggregation and displacement errors, presents a challenging case for density estimation where the data is strongly anonymised. In this scenario, the proposed estimator is able to generate a clear and recognisable density shape that reflects the underlying distribution of poor households. In contrast, the naive estimator produces density estimates that are less reliable and less accurate, failing to capture the same density structure that is present in the underlying population. Overall, these results indicate that the proposed KDE-ED estimator is a preferred method for estimating the density of poor households, especially under large aggregation and displacement errors.

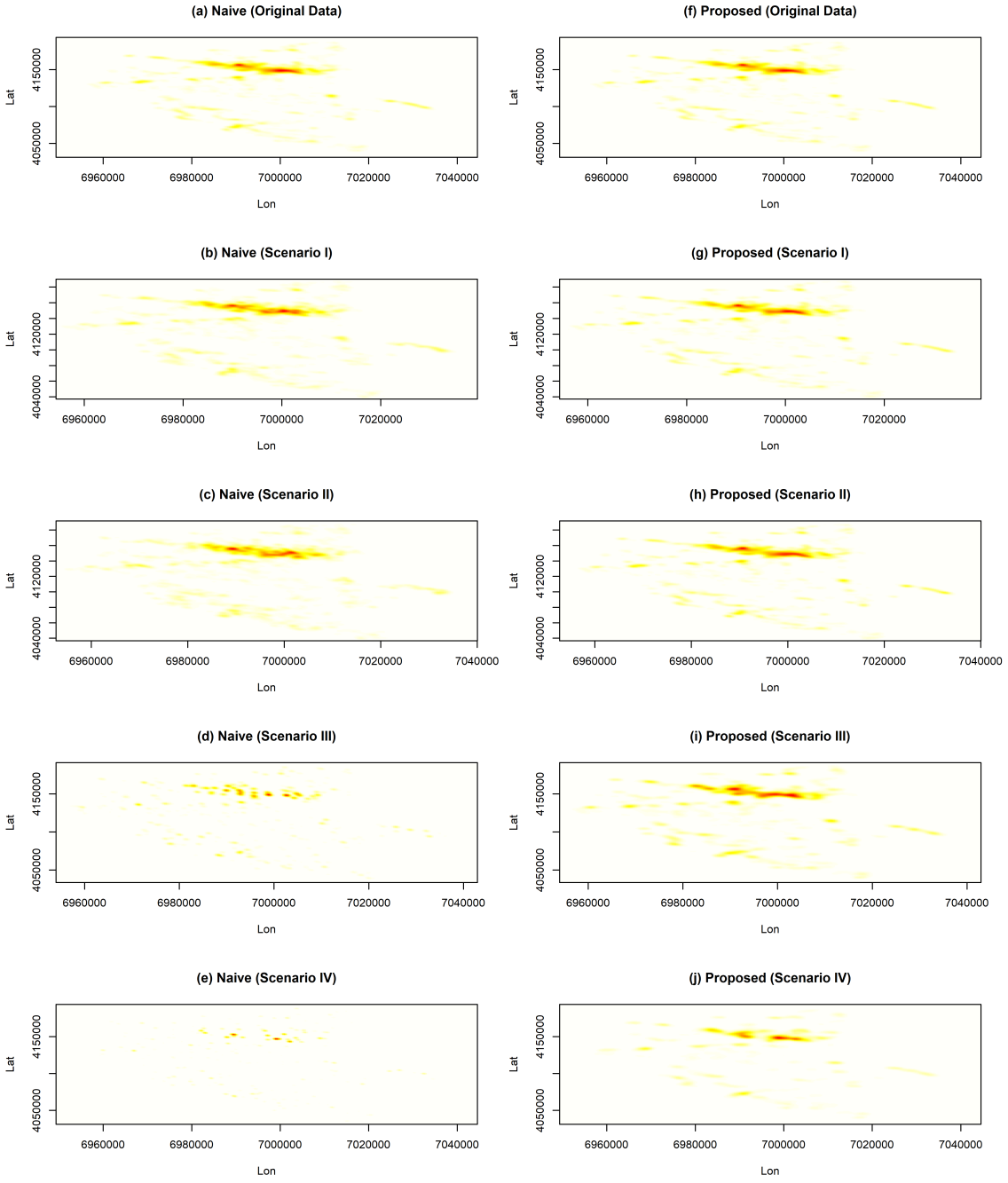

Figure 3.15: Density of poor households under both aggregation and displacement errors for each estimator with grids.

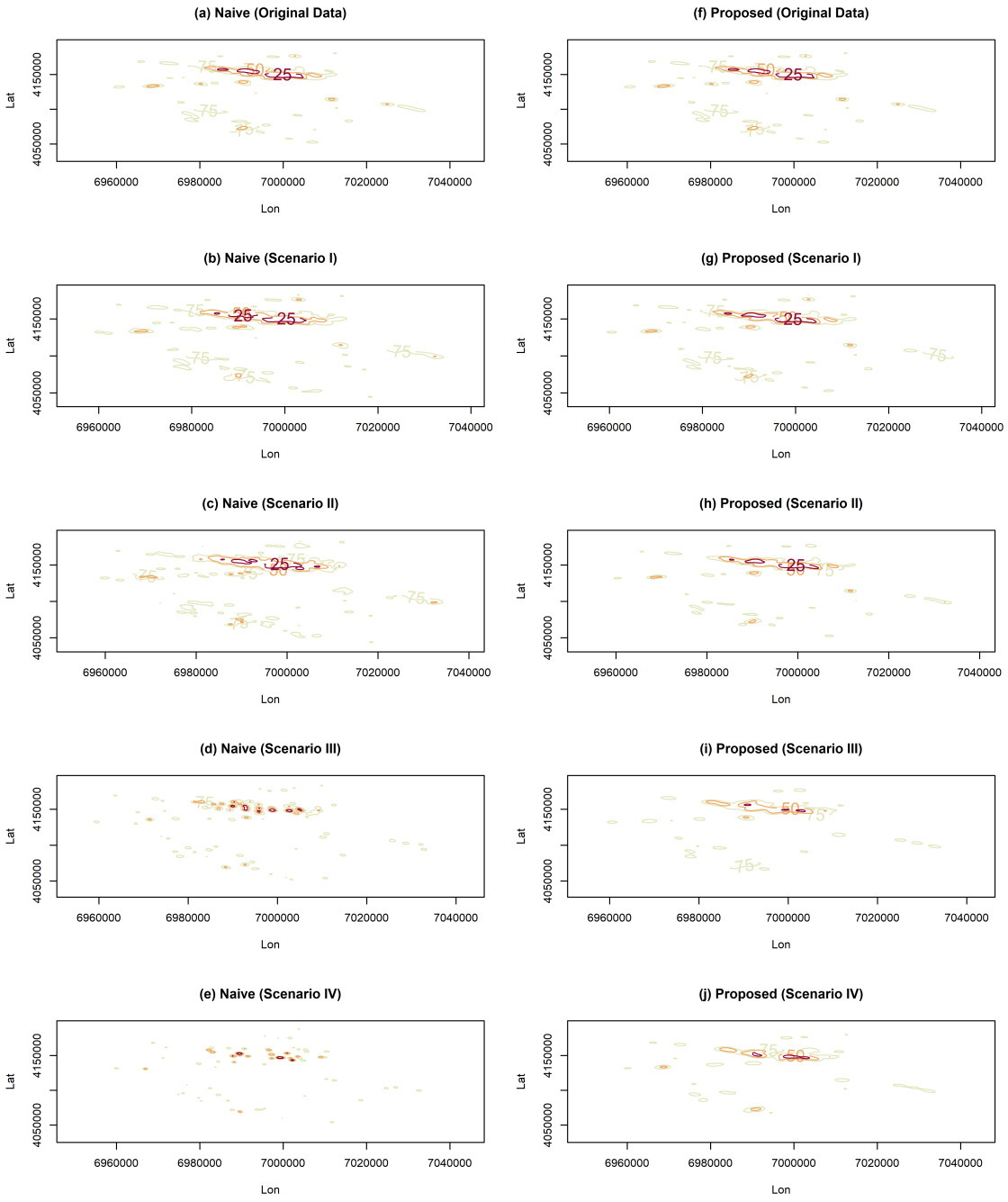

Figure 3.16: Contour plots of the density of poor households under both aggregation and displacement errors for each estimator with grids.

3.7 Application of the proposed method under EA displacement using the 2011 BDHS data

In order to apply the proposed method, we use the 2011 BDHS data at the household level and the GPS data for the 600 EAs, of which 393 are from rural areas, and 207 are from urban areas. To protect the confidentiality of respondents, the households' location data within the same EA were aggregated to a single coordinate, the EA centroid. Then, the EA centroids were displaced by applying the DHS random displacement algorithm. The boundary map of the 2011 BDHS EAs is not available to data analysts. As discussed in Section 1.2.1, EAs are generally designed to focus on population count (on average 120 households per EA) rather than geographic size; therefore, their specific sizes (measured in square meters) can vary based on population density and whether they are in an urban or rural area. Hence, the EA size varies but is unknown to data analysts. The boundary of an EA should be available to apply the proposed KDE-ED method, as discussed in Section 3.5, to address both aggregation and random displacement errors. However, in the 2011 BDHS, except for random displacement, the aggregation of EAs does not impact any upazila (admin 3 of Bangladesh) level estimation of indicators, such as poverty, since an EA is a lower geographical unit than upazila, and aggregation is constrained within upazila boundaries. Estimates, such as density estimates, at finer levels below the EA geography might be influenced by both aggregation and random displacement. In such cases, the proposed KDE-ED method can be applied by approximating EA size or geography; however, this requires further investigation into the effects of EA size approximation errors on the results. In this section, we aim to apply the proposed method to obtain upazila-level estimates of interest, discussed below, using the 2011 BDHS data, accounting for the EA random displacement errors. In particular, the centroids of the 600 EAs, selected with probability proportional to Bangladesh Census 2011 EA's household numbers, are available to data analysts but are displaced. Therefore, we apply the proposed EDC method, developed in Section 3.4, under only the displacement of EAs.

The true EA centroid coordinates were displaced within the designated admin 2 (district of Bangladesh) area by applying the DHS displacement algorithm that we described in detail in subsection 3.2.1. The urban EAs were displaced 0-2 kilometres, while rural EAs were displaced 0-5 kilometres with 1% of rural EAs being randomly displaced 0-10 kilometres. As a result, in the case of the BDHS 2011, the displaced EA coordinates may be misplaced at the lower admin level of Bangladesh e.g., at the upazila level (admin 3). Therefore, any estimates e.g., means or proportions at the upazila level may be biased using the displaced EA coordinates.

By mapping the displaced EA coordinates, the upazila identifiers corresponding to the EAs can be determined. These are subsequently referred to as the displaced upazila IDs. The 2011 BDHS data includes true upazila IDs corresponding to the EAs, even though the true EA coordinates are unknown. However, these true upazila IDs are not included in BDHS data sets after 2011, potentially due to concerns that releasing the true upazila ID might increase the risk of disclosure. These true upazila IDs can be used to validate any estimates at the upazila level. If the displaced upazila ID matches the true upazila ID for an EA, then the EA is considered a true non-misplaced EA. In contrast, if they do not match, the EA is considered a true misplaced EA. Additionally, if the displacement buffer around a displaced EA coordinate intersects with an upazila boundary, the EA is considered as a potentially misplaced EA.

We obtain upazila poverty estimates using the 2011 BDHS data by applying the proposed EDC method, the MPD method, and the naive method which ignores the misplacement error due to the EA displacement. Although the proposed KDE method could be used in this context, our assessment through an empirical study in Section 3.6.1 shows that the proposed EDC method performs better than the KDE method. Therefore, we do not use the proposed KDE method in this application. In particular, as the variable of our interest, we consider the number of poorer households for each EA. A household has been considered relatively poor if its wealth index is below the 40th percentile of the wealth indices of the survey households that we described in detail in the data description section of Chapter 1. In addition to the estimated upazila proportions of poor households, in this section, we also compare the performance of the methods under only EA displacement in terms of correct upazila classification probabilities for the 2011 BDHS EAs. In particular, by applying the following three methods, we obtain the estimated upazila poverty along with classification probabilities for each EA using the 2011 BDHS displaced EA coordinates:

a) The naive method that ignores the EA misplacement error due to the displacement process. With this method, true misplaced EAs are always classified into correct upazilas with probability 0, while correct upazila classification probabilities for true non-misplaced EAs are always equal to 1. Therefore, this approach always assigns true misplaced EAs to the wrong upazilas.

b) The proposed EDC method that accounts for the EA misplacement uncertainty to estimate upazila classification probabilities for each EA. The method iterated with $B_1 = 200$ Monte Carlo samples and $B_2 = 20$ additional sample steps using the WorldPop grid points and additional information from external data. We described this external data based method in the methodology section 3.4.

c) The MPD method proposed by Warren et al. (2016b). In order to take into account the misplacement uncertainty, this method obtains the estimated upazila classification probabilities for each EA by displacing the displaced EA coordinates and then selecting the most probable upazila as the correct one for each EA. However, this approach can produce biased estimates if the correct upazila is chosen with a probability of less than 1.

3.7.1 Results and discussion

This section first gives a summary of the EA misplacement statistics. Next, we present and discuss the results for the correct upazila classification probabilities for EAs that are obtained by applying the three methods: the MPD, the proposed EDC method and the naive method that ignores the misplacement error due to the displacement process. Finally, the estimated upazila proportions of relatively poor households for each method are presented and compared to the upazila estimates produced based on the 2011 BDHS data that are not affected by the displacement process.

We now present and discuss findings from an exploratory analysis of the 2011 BDHS data. By constructing displacement buffers around the centroid of the EA displaced coordinates, we find that about 44% (266 out of 600) of the EAs are potentially misplaced i.e., those EA displacement buffers intersect with an upazila boundary. Note that all three methods always perform equally well in classifying an EA into the correct upazila if the displacement buffer falls entirely within the boundaries of an upazila. To see how EA misplacement varies by geographic region (rural/urban), the distribution of only potentially misplaced EAs by true misplacement and geographic region is presented in Table 3.8. Among the potentially misplaced EAs, about 31% (83 out of 266) of EAs are true misplaced. To remind the reader, in this thesis, a true misplaced EA means that the displaced upazila ID and the true upazila ID for that EA are different. There were 393 rural EAs and 207 urban EAs in the 2011 BDHS, as described in 1.2.1. We observe that rural EAs are more likely to be potentially misplaced than urban EAs, for instance, 48.3% (190 out of 393) for rural and 36.7% (76 out of 207) for urban. This is because the maximum displacement distance was higher for rural EAs than for urban EAs. However, among the potentially misplaced EAs, the percentage of the true misplaced EAs is higher for the urban area (40.8%) than the rural area (27.6%).

	Non-misplaced	Misplaced	Total
Rural	138 (0.726)	52 (0.276)	190 (1.000)
Urban	45 (0.592)	31 (0.408)	76 (1.000)
Total	183 (0.688)	83 (0.312)	266 (1.000)

Table 3.8: Distribution of only potentially misplaced EAs by true misplacement and urban/rural. Proportions are given in parentheses.

Table 3.9 shows the distributions of EAs by the number of upazila intersects with the displacement buffer and a) misplacement, and b) rural-urban respectively. We observe that about 5% (12 out of 266) of the potentially misplaced EAs buffers intersect with more than four upazilas and most of them are true misplaced (8 out of the 12). For urban areas, 12 EAs whose buffers intersect with more than four upazilas compared with none for rural areas.

The DHS displacement process is not restricted within the rural/urban area boundaries. Therefore, it is possible that an EA that originated from a rural area can be placed to an urban area due to the EA displacement process and vice versa. Table 3.10 presents that 6% of EAs (22 out of 393) originally from rural areas are displaced to urban areas. On the other hand, 25% (52 out of 207) of urban EAs are moved to rural locations due to the displacement process.

Number of upazila intersects	1	2	3	4	5	6	7	8	9
Non-misplaced	334	140	36	3	2	1	1	0	0
Misplaced	0	54	15	6	3	3	0	1	1
Rural	203	155	33	2	0	0	0	0	0
Urban	131	39	18	7	5	4	1	1	1

Table 3.9: Distributions of EAs by the number of upazila intersects with the displacement buffer and a) misplacement, and b) rural-urban.

True vs. displaced	Rural	Urban	Total
Rural	371 (0.94)	22 (0.06)	393 (1.00)
Urban	52 (0.25)	155 (0.75)	207 (1.00)

Table 3.10: Distribution of true rural-urban EAs by displaced rural-urban EAs. The proportions are given in parentheses.

We obtain the estimated upazila classification probabilities for each EA by applying the MPD method, the proposed EDC method and the naive method. From these upazila classification probabilities for each EA, we can separate the correct upazila classification probability as the true upazila IDs for EAs are given in the 2011 BDHS data. As described in Section 1.2.1, the maximum displacement distance differs for rural/urban EAs (e.g., 2km for urban, 5km for rural), with 1% of rural locations randomly displaced up to 10km. Data analysts are not informed which rural EAs undergo this extended displacement, but our proposed methods account for this additional uncertainty. The summary statistics of the estimated correct upazila classification probabilities by using the EDC, MPD and naive methods for all potentially misplaced EAs and their different categories, such as true misplaced EAs and rural/urban EAs, are given in Table 3.11. We see that the proposed EDC method is better than the naive method for all categories except for the true non-misplaced category. This is because the true non-misplaced EAs are always classified to the correct upazilas with probability 1. However, it is unknown to the data analysts which EAs are true non-misplaced.

We also observe that the proposed EDC method outperforms the MPD method in terms of having higher correct upazila classification probabilities for all categories. For example, the proposed EDC method has a higher mean for correct upazila classification probabilities over true misplaced EAs than the MPD method. This indicates that the additional information for the underlying probability distribution of true locations from external data sources improves the classification of EAs into correct upazilas with higher probabilities. We also see that true non-misplaced EAs are classified correctly with higher probabilities compared to true misplaced EAs. Another observation is that the probability of correct upazila classification is higher for rural EAs than for urban EAs.

	Estimator	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
All	EDC	0.020	0.590	0.865	0.749	0.968	1.000
	MPD	0.001	0.556	0.804	0.716	0.944	1.000
	naive	0.000	0.000	1.000	0.688	1.000	1.000
Non-misplaced	EDC	0.044	0.748	0.922	0.841	0.976	0.998
	MPD	0.014	0.728	0.879	0.823	0.957	0.998
	naive	1.000	1.000	1.000	1.000	1.000	1.000
Misplaced	EDC	0.020	0.296	0.526	0.546	0.795	1.000
	MPD	0.001	0.216	0.400	0.443	0.624	1.000
	naive	0.000	0.000	0.000	0.000	0.000	0.000
Rural	EDC	0.028	0.630	0.877	0.776	0.974	1.000
	MPD	0.027	0.611	0.830	0.743	0.949	1.000
	naive	0.000	0.000	1.000	0.726	1.000	1.000
Urban	EDC	0.020	0.501	0.727	0.683	0.947	1.000
	MPD	0.001	0.405	0.678	0.638	0.912	1.000
	naive	0.000	0.000	1.000	0.592	1.000	1.000
Rural Misplaced	EDC	0.028	0.302	0.533	0.566	0.797	1.000
	MPD	0.027	0.217	0.404	0.437	0.611	1.000
	naive	0.000	0.000	0.000	0.000	0.000	0.000
Urban Misplaced	EDC	0.020	0.231	0.502	0.513	0.727	1.000
	MPD	0.001	0.216	0.384	0.455	0.681	1.000
	naive	0.000	0.000	0.000	0.000	0.000	0.000

Table 3.11: Summary statistics of the correct upazila classification probabilities over all potentially misplaced EAs using the proposed EDC, the MPD and the naive methods.

Figure 3.17 shows the performance of the two methods: the proposed EDC and the MPD for all potentially misplaced EAs and only for EAs whose buffers intersect with up to 4 upazilas under different categories A-G. For all scenarios, the proposed EDC method, which uses additional information from external data, clearly outperforms the MPD method in terms of higher classification probabilities. Also, the results reveal that the proposed method performs better for EAs whose buffers intersect with up to 4 upazilas. Also, excluding the EAs for those buffers that intersect with more than 4 upazilas the overall performance of the rural and urban EAs are approximately the same. This is because only the urban EAs buffers intersected with more than 4 upazilas. More-

over, in order to assess the proposed EDC method's performance from another aspect, Figure 3.18 represents the boxplot of the estimated correct upazila classification probabilities for a) true misplaced b) true non-misplaced and c) all potentially misplaced EAs by the number of intersected upazilas. For all scenarios a-c, we observe that the correct upazila classification probabilities are much smaller for EAs with more than 4 intersected upazilas. This indicates that the proposed EDC method performs better at accurately classifying EAs whose buffers intersect with up to 4 upazilas, but its performance reduces for intersections involving more than 4 upazilas. However, excluding the 12 EAs whose buffers intersect with more than 4 upazilas results in 10 additional false-out-of-sample upazilas (as defined in 3.3.2) from the 395 true in-sample upazilas in the 2011 BDHS. Furthermore, 8 of those 12 EAs are truly misplaced.

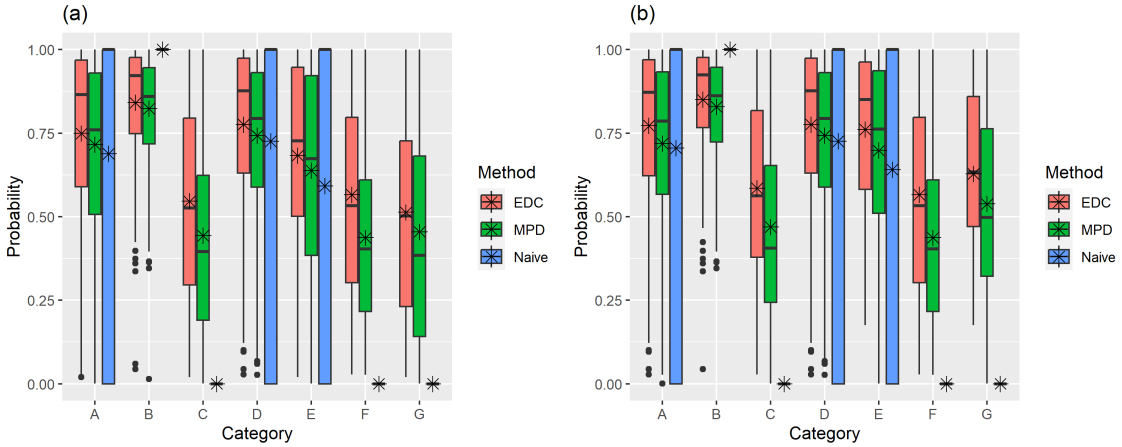

Figure 3.17: Boxplot of the correct upazila classification probabilities using the EDC, MPD and naive methods (left: (a) all potentially misplaced EAs, right: (b) only 2-4 intersected upazilas, Category: A: overall, B: non-misplaced, C: misplaced, D: rural, E: urban, F: rural misplaced, G: urban misplaced). The asterisk (*) is used as a marker to represent the mean probabilities for each method and category.

In summary, in the case of true misplaced EAs, the naive method has the correct upazila classification probabilities equal to 0 i.e., the EAs are consistently misclassified. On the other hand, for most of the EAs, the proposed EDC method has higher correct upazila classification probabilities than the MPD method. Therefore, the proposed EDC method is always preferred over the naive and MPD methods. Regarding the true non-misplaced EAs, the proposed EDC method performs better than the MPD method for most EAs in terms of higher correct classification probability. However, the naive method always performs perfectly by classifying true non-misplaced EAs into correct upazilas. Therefore, the proposed EDC method is always preferred in the case of true misplaced EAs, while the naive method should be used for true non-misplaced EAs. However, in reality, it is unknown to the data analysts which EAs are truly misplaced. It is possible to identify the EAs that are potentially misplaced. The naive method correctly classifies the potentially misplaced EAs

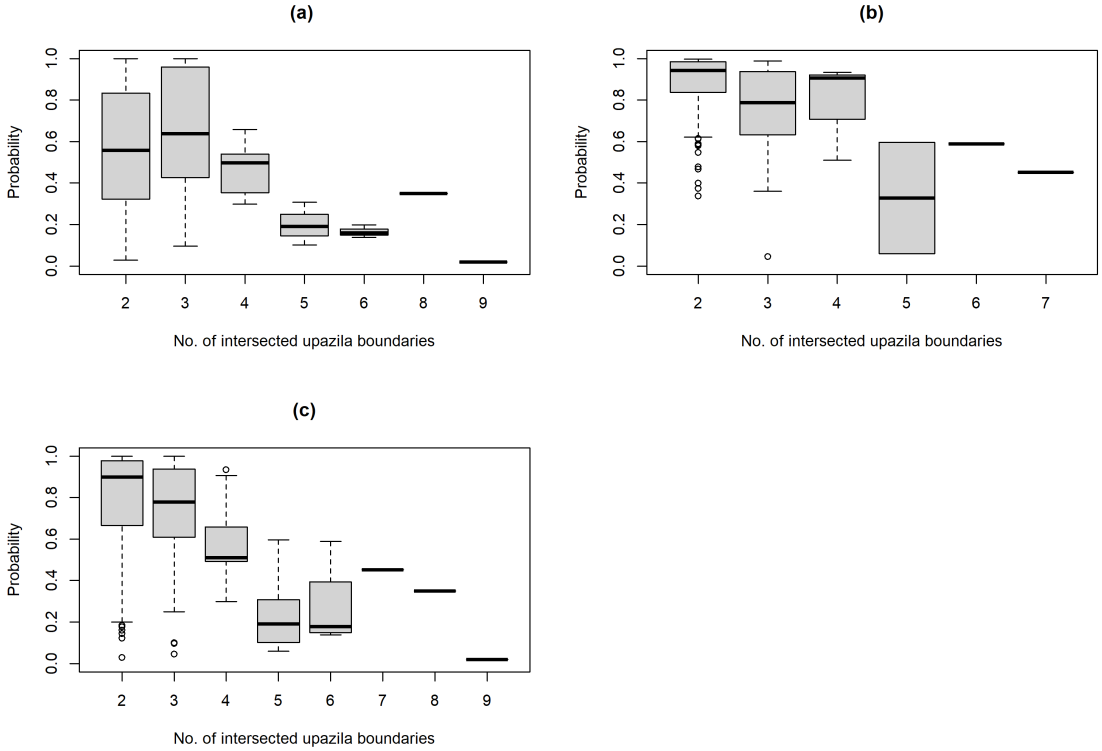

Figure 3.18: Boxplot of the correct upazila classification probabilities using the EDC method ((a) true misplaced, (b) true non-misplaced and (c) all potentially misplaced EAs).

with a probability of either 0 or 1. On the other hand, most of the EAs are classified with fairly high probabilities by the proposed EDC method. Also, some EAs, especially for those buffers that intersect with more than 4 upazilas, are correctly classified with very small probabilities. Note that EAs whose displacement buffers fall only within an upazila boundary are always classified into correct upazilas by all three methods: the naive, the MPD and the EDC. The average correct upazila classification probabilities over all potentially misplaced EAs is larger than the other two methods. In summary, the proposed EDC method should be preferred over the naive and MPD methods for correctly classifying the 2011 BDHS displaced EAs.

3.7.1.1 Poverty estimates at upazila level of Bangladesh

In this application, we obtain the upazila estimates of relatively poor households by using the MPD, the proposed EDC and the naive method that ignores the misplacement error due to the random displacement process. In addition, the upazila proportions estimates are obtained using the true upazila identifier for each EA that is available in the 2011 BDHS data. We use these true upazila estimates that are not affected by the displacement process as the validation estimates.

Estimated upazila proportions using the three methods are presented in Figure 3.19 against valida-

tion estimates for the 395 true in-sample upazilas, which are available only for the 2011 BDHS. We observe that the upazila estimates using the proposed EDC method are closer to the true upazila estimates than the upazila estimates obtained by the other two methods: the MPD, which picks the most likely upazila for each EA, and the naive that ignores the misplacement error. For example, the correlation coefficients between the upazila estimates generated by the naive, EDC, and MPD methods, and the validation (true) estimates are 0.867, 0.962, and 0.884, respectively. Moreover, it is apparent from Figure 3.19(a) that 36 false-out-of-sample upazilas are created by the DHS displacement process, and the naive method cannot provide estimates for those 36 upazilas. Although the MPD method can provide estimates for some false-out-of-sample upazilas by picking up the most likely upazila for each EA, it creates some new false-out-of-sample upazilas for which this method provides no estimates. Whilst, the proposed EDC method can produce estimates for all those 36 false-out-of-sample upazilas, and the estimates are reasonably close to the validation estimates (Figure 3.19(b)). We note that while the EDC method is capable of producing estimates for all out-of-sample upazilas that were falsely created by the displacement process, it may also introduce some additional false-in-sample upazilas beyond the 395 true sample upazilas in the 2011 survey due to the correction process involving the creation of a buffer around the displaced EA centroid. This result can be explained by the fact that in the case of the proposed EDC method, all upazilas, of which only one is correct but unknown to the data analysts, are selected within the displacement buffer according to the upazila classification probabilities for each EA. This implies that the higher the probability of correct upazila classification, the closer the upazila estimates will be to the validation upazila estimates. We have assessed that the proposed EDC method, which uses additional information from external data sources, has an overall higher correct upazila classification probability than the MPD and naive methods. Also, the correlation results presented above suggest that the EDC method has the highest degree of accuracy and precision in estimating upazila proportions, as it achieves the highest correlation coefficient compared to the Naive and MPD methods. In summary, the results of this application using the 2011 BDHS data suggest that the proposed EDC method is preferred to the MPD and naive methods.

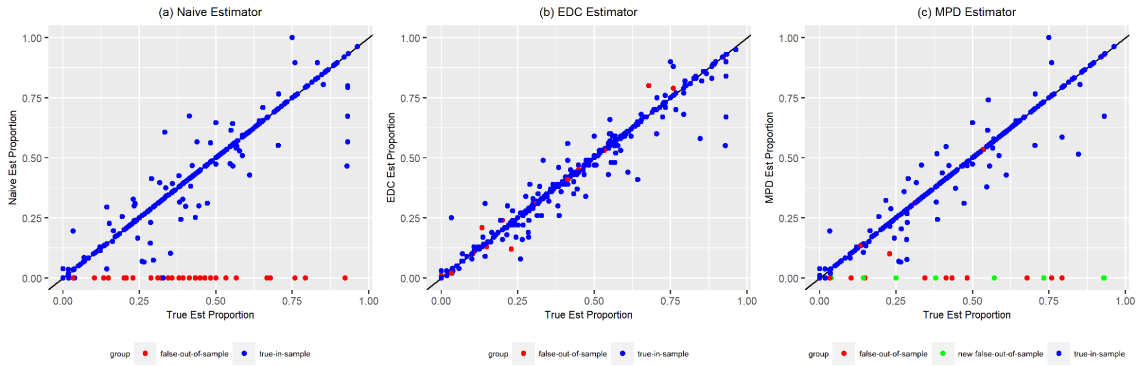

Figure 3.19: Plot of the estimated upazila proportions of poor households by the naive, the EDC and the MPD method against the upazila estimates obtained using the undisplaced (true) data (Blue dots: true-in-sample upazilas; Red dots: False-out-of-sample upazilas created due to the displacement process; Green dots: new false-out-of-sample upazilas created through the correcting process by using the MPD method). The figure demonstrates only the actual 395 sampled upazilas from the 2011 BDHS, and we exclude any false-in-sample upazilas originating from any methods.

3.8 Concluding remarks

This chapter presents the proposed estimators for obtaining density estimates of observations of coordinates and estimates of domain parameters by correcting measurement errors using data under two situations: a) only the DHS random displacement process and b) the aggregated and random displacement process. We first consider the random displacement issue, followed by the combined aggregation and random displacement problem.

Regarding the random displacement, the true location coordinates of observations are not available; they are randomly displaced before being released to data analysts to preserve respondent confidentiality. For instance, to protect respondent confidentiality and mitigate disclosure risk, DHS stopped releasing true EA centroids, reflecting a deliberate trade-off with the precision of location-dependent estimates. Therefore, in this chapter, our proposed methods attempt to address the analytical challenges arising from this lack of precise data, aiming to enhance the accuracy of estimates without compromising disclosure control risk. It is crucial to note that the DHS random displacement process remains consistent across all its surveys, with identical urban-rural distance parameters (Burgert et al., 2013). This uniformity ensures that our developed approach can be applied to any DHS survey. In particular, we introduce a measurement error model tailored to the DHS random displacement process, which can be potentially extended to other forms of random

displacement processes. Subsequently, we develop a new probability distribution for the DHS random displacement, representing a key contribution of this chapter.

We propose two estimation methods for the measurement error model to obtain density estimates and domain parameter estimates by drawing likely true location coordinates. The first method is based on the Kernel Density Estimates (KDE) as the marginal distribution of the true location data, and the estimation is implemented using the Stochastic Expectation-Maximization (SEM) algorithm. The proposed KDE-based method accounts for random displacement error while deriving its bandwidth as part of the estimation process, similar to the method used by Groß et al. (2017), which accounts for rounding error instead. As a result, the proposed method automatically selects an optimal bandwidth based on the data for each iteration without selecting an appropriate bandwidth beforehand. This makes it easy to implement, and the method can work with any bandwidth selection method, making it highly flexible. Also, we have observed that the grid size does not impact density estimation when dealing with random displacement. However, it is necessary to use a sufficiently large grid size (e.g., 400) to cover all possible true locations within the buffers around the displaced coordinates. Kernel density estimates are useful in approximating the marginal distribution of unknown true location coordinates with register/population data. It may not perform well when sampling is involved because the method cannot distinguish between the absence of sample data and the absence of population. Additionally, the random displacement process can create false in-sample and out-of-sample domains. Nevertheless, the method may still work under certain sampling designs that cover the entire population area, such as a simple random sample over the entire area.

The second method is the External Data Classification (EDC)-based method, which approximates the unknown underlying distribution of true location data as an alternative to the KDE-based method. The estimation of this EDC-based method is implemented through numerical integration. In the proposed estimation algorithm, it is recommended to use at least 200 pseudo-sample iterations of unknown true location coordinates to achieve better realizations in the estimates, accounting for random displacement errors. The EDC method is expected to improve the estimates over the KDE-based method as it is developed using additional information from external data sources, which are more likely to be associated with the underlying distribution of true location data. The EDC method works better than the KDE method by approximating the marginal distribution of the true location coordinates more accurately, using external data sources. However, this is subject to the availability and up-to-date-ness of the external data sources. Besides register data, the EDC method can better approximate the underlying distribution of true locations than the

KDE method in the case of survey data, such as the 2011 BDHS, where there is a clear connection between the true EA and population density of household distribution. Specifically, the sampled DHS EAs were selected with a probability proportional to the census EA size, i.e., the number of households.

We compare the performance of the two proposed KDE and EDC methods, the Most Probable Domain (MPD) method by Warren et al. (2016b) and the naive method that ignores the displacement error using simulated register data. The simulated register data for household location coordinates are generated from the gridded population raster of 100m $\times$ 100m size from WorldPop, which estimates the population of Bangladesh adjusted using the 2011 census population. In this raster, each grid cell represents the number of people per pixel. It is worth noting that in our simulation study, we use external data sources, such as the WorldPop gridded population count, to generate the simulated true location data. Therefore, there is no issue with the quality or representativeness of the external data source in the empirical study. However, in real situations, the quality and up-to-dateness of such external data sources should be assessed. We demonstrate the usefulness of external data sources, such as WorldPop, to approximate the underlying distribution of true EA locations using only a single set of displaced EA coordinates from the 2011 BDHS. Moreover, a sensitivity analysis could be conducted in future research by introducing slight errors to the simulated true household location data.

In the empirical study with simulated register data, the proposed KDE and EDC methods can correct to some extent the bias in the domain (upazila) estimates of the proportion of poor households under misplacement errors due to the random displacement process. Moreover, the proposed estimators have lower RMSEs of the upazila estimates than the naive and MPD estimators. Notably, upazilas with smaller areas in terms of square kilometres have higher misplacement errors, resulting in greater bias and variability in the estimates when using the naive method that ignores misplacement errors. In contrast, the proposed EDC method outperforms other methods by correcting misplacement errors, particularly for upazilas with small areas. However, when the misplacement errors are low, all methods perform similarly. In such cases, the naive method should be chosen for its simplicity. It is worth mentioning that misplacement errors depend not only on the area of the upazila but also on the shape of the upazila and whether people live closer to the boundary. Since the data analysts do not know which location units are true misplaced, all the corrected methods - the MPD, KDE, and EDC - introduce additional errors for true non-misplaced units. Therefore, future research could be done to estimate the expected number of misplaced units within each upazila using the displaced coordinates. This would be helpful to identify whether the proposed methods

work for obtaining precise upazila estimates by correcting the misplacement error. Although the MPD, KDE, and EDC methods introduce some bias in upazila estimates for true non-misplaced EAs, which are unknown to the data analysts, in the simulation study, the proposed EDC method produces very high correct upazila classification probabilities for non-misplaced units. Furthermore, the naive method always classifies misplaced units into the wrong upazilas, whereas the proposed EDC method can accurately classify them into the correct upazilas to some extent. We also observe that the correct upazila classification probabilities are much higher for displaced location units whose buffers intersect up to 4 upazilas. Therefore, overall, the proposed EDC method is preferred over the other methods - KDE, MPD, and the naive that ignores the misplacement error due to the displacement process.

We apply the proposed method under displacement using the real survey data. In particular, we obtain upazila poverty estimates using the 2011 BDHS data by applying the proposed EDC method, the MPD method, and the naive method that ignores the misplacement error due to the DHS random displacement process. Based on our empirical assessment in Section 3.6.1, which showed superior performance of the proposed EDC method over the KDE method, we do not use the KDE method in subsequent applications. We observe that the upazila estimates using the proposed EDC method are closer to the true upazila estimates than the upazila estimates obtained by the other two methods: the MPD, which picks the most likely upazila for each EA, and the naive that ignores the misplacement error. Furthermore, the random displacement process creates some false out-of-sample upazilas. While the MPD method can correct the false out-of-sample problem for a few upazilas by picking the most likely upazila for each EA, it creates new false out-of-sample and in-sample upazilas. On the other hand, the proposed EDC and KDE methods can produce estimates for all out-of-sample upazilas that are falsely created by the displacement process. This result can be explained by the fact that in the case of the proposed EDC and KDE methods, all upazilas, of which only one is correct but unknown to the data analysts, are selected within the displacement buffer according to the upazila classification probabilities for each EA. However, the EDC and KDE methods may also introduce additional false in-sample upazilas beyond the true sample upazilas in the survey due to the correction process involving the creation of a buffer around the displaced EA centroid.

Regarding the aggregation and random displacement process, the aggregation concentrates data at the centroid of the enumeration area (EA), while the random displacement moves the original location from one place to another. Previous studies, to the best of our knowledge, did not consider the uncertainty arising from both aggregation and random displacement. Therefore, we propose

a novel method, called the external data and kernel density-based (KDE-ED) method, to obtain density and domain parameter estimates that correct for both aggregation and random displacement errors under a measurement error model. This is another important contribution in this chapter.

The simulation study results suggest that the proposed KDE-ED method outperforms the naive method that ignores the aggregation and displacement errors, with lower bias and RMSE of the upazila estimates. Additionally, for the density estimates of poor households, the proposed KDE-ED method has a lower RMISE than the naive method. The proposed method can also preserve the underlying structure of the density for various aggregation and displacement levels. Therefore, the KDE-ED estimator is a preferred method for estimating the density of poor households, particularly under large aggregation and displacement errors.

Chapter 4

Linear mixed models under random displacement

4.1 Introduction

Multilevel models are statistical models designed to handle clustered or nested data, which are relevant in the context of surveys, for example, the Demographic and Health Surveys (DHS). The DHS is a two-stage stratified survey with a hierarchical data structure. Individuals are nested within Enumeration Areas (EAs), also known as clusters or primary sampling units (PSUs), and EAs are nested within the administrative boundaries of the country. The presence of this hierarchical structure highlights the necessity of using multilevel regression modelling (Islam, 2005; Alom et al., 2012; Imam et al., 2018; Bhuiyan et al., 2020).

Clustering occurs when observations within the same group or cluster demonstrate greater similarity than in other groups. Ignoring clustering can lead to incorrect standard errors, resulting in inaccurate inferences (Fielding et al., 2003). Multilevel models can account for the potential influence of cluster-level factors on individual-level outcomes, making them a useful tool for analysing observations which belong to groups or in cases where a hierarchical data structure exists in the survey.

Multilevel models can be used for both analytical inference and group prediction. Analytical inference involves testing hypotheses and making statistical inferences about the relationships between outcome and predictor variables at different levels of hierarchy. In contrast, group prediction involves using the model to predict outcomes for groups of individuals and obtain group means.

In multilevel models, random effects are specified at the group level to capture the inherent grouping structure of observations. For precise estimates of group-level variations, it is essential to have accurate information about the groups of observations, ideally based on their true location coordinates. However, in many situations, such as with DHS data, these actual coordinates may remain undisclosed to analysts and are randomly displaced to protect respondent confidentiality. Specifically, in the DHS data, true EA centroids are displaced but restricted to admin 2 boundaries (e.g., corresponding to a district in a country) using the DHS random displacement mechanism. Data users are provided only with these displaced EA centroid coordinates and the associated survey EA identifiers (Burgert et al., 2013). As discussed in detail in Section 2.6, while the displacement is not a problem when using multilevel models with random effects specified at the EA level or a displacement-restricted higher administrative level, challenges arise when random effects are specified at levels below the displacement-restricted boundaries, such as the admin 3 (sub-district) level. In such cases, the displacement process can lead to misplacement of observations, causing mixing among groups. This can result in biased estimates of random variance components in multilevel models and associated model-based estimates of finite population parameters, such as group/sub-district means. As reviewed in Chapter 2, some studies (e.g., Das et al., 2019a,b, 2020b) have fitted multilevel models with random effects specified at the sub-district level using DHS data, disregarding the uncertainty due to EA misplacement, which could lead to misleading inferences.

The impact of EA misplacement errors due to displacement intensifies as we move to lower administrative levels of a country. For example, a unit at the admin 4 level might experience more misplaced EAs in comparison to a unit at the admin 3 level. Consequently, specifying random effects at the admin 4 level can lead to more biased estimates of random variance components in multilevel models than if they were specified at the admin 3 level. The magnitude of such biases may directly correlate with the degree of misplacement error introduced by the random displacement process. In scenarios with a high degree of misplacement error, the proposed methodologies can significantly reduce bias in estimates, offering improved results for users. Therefore, it is essential to research the effect of misplacement error due to random displacement on multilevel model estimation, such as the linear mixed model (LMM), and develop methods accounting for misplacement error in model fitting and uncertainty estimation of the estimates of the model parameters. This chapter aims to contribute to addressing this gap in the literature.

Improved estimates of the parameters of the linear mixed model and associated model-based group means by correcting the misplacement error can be obtained when the DHS EA is assigned to the

correct group with a high probability. From this perspective, the proposed methods under displacement, for example, the external data-based classification (EDC) method, can be used to select likely true EA locations that belong to the correct groups with higher probabilities. Following an extensive literature review in Chapter 2, we developed the EDC method under a functional measurement error model in section 3.4 and assessed its performance over the candidate methods through a simulation study. In particular, the purpose of the EDC method is to obtain the likely true unknown EA locations given the displaced locations by using the developed conditional distribution of displaced coordinates and the underlying unknown marginal distribution of the true locations. We proposed a flexible framework for approximating the true location distribution using additional information from multiple external sources that improve the probabilities for assigning EAs into the correct groups. Therefore, we expect the EDC-based linear mixed model estimator to perform better than the naive estimator by drawing more likely true unknown locations to obtain improved estimates of the model parameters and group means under displacement. In this chapter, we discuss the linear mixed model fitting under the proposed EDC method and the variance estimation of model parameter estimates. We evaluate the performance of the proposed methods for both point and variance estimates under the EDC using a model-based simulation study.

When the estimates of the linear mixed model parameters, especially the variance components, are biased under the misplacement due to a random displacement process, an attempt can also be made to obtain the estimates of the bias and hence, to develop a bias-corrected estimator of the model parameters. In particular, one could potentially think about bootstrap bias correction methods for the linear mixed model parameters under displacement. To the best of our knowledge, none of the previous studies attempted to use a bootstrap bias correction method under the random displacement process. The general idea of the bootstrap bias estimation technique is that it estimates the bias of a biased estimator under repeated bootstrap samples from the original representative sample of the population (for details, see Efron and Tibshirani, 1992, section 10.2). Regarding the bias of the linear mixed model parameters under displacement, the bootstrap bias correction method should work to satisfy the two crucial assumptions. The first assumption is that the functional form of the underlying model, e.g., the linear mixed model is appropriate to the data. The second but most important assumption is that the magnitude of the bias of the naive estimates of the model parameters under displacement is the same as the magnitude of the bias of the estimates under bootstrap samples and repeated displacement process to the displaced locations. The bootstrap samples are generated under the model using the model parameters' estimates based on the given (observed) displaced data. Given these two assumptions are satisfied, the bootstrap bias correction method

can obtain unbiased estimates of the model parameters under displacement. Therefore, this chapter also aims to develop a method for the parametric bootstrap bias correction for correcting the bias of the linear mixed model parameters, especially the variance components under the misplacement due to the DHS displacement process. Hereafter, this is referred to as the Bootstrap bias correction (BC) method. We assess the performance of the BC method over the EDC and naive methods through a model-based simulation study.

The BC method should effectively correct the bias of global (model) parameters, and it may not work for the associated model-based finite population parameters, e.g., group means under random displacement. This is because the assumption of bootstrapping regarding the magnitude of the bias of predicted group means under random displacement may not be satisfied. One potential reason for this is that the bootstrap samples are generated under the model estimates and repeated displacement process to the displaced locations by creating buffers around them. This results in additional bias in the estimates for groups with displaced locations that are truly non-misplaced, which is unknown to the data analysts. In contrast, the random displacement process does not affect naive estimates of means for groups with non-misplaced locations. Therefore, when bias correction is applied to the naive estimates that involve non-misplaced units, it brings additional uncertainty. Additionally, due to random displacement, there may be a mixing of groups within the buffer around the true and displaced locations, potentially affecting the estimates of group means. In that case, the EDC method should be preferred to the BC method for estimating group means under misplacement due to displacement. We explore this in this chapter.

The chapter is organised as follows. Section 4.2 describes the linear mixed model under the EA displacement process and proposes methods for a) model fitting and b) variance estimation of the model parameters estimates under the EDC method accounting for the uncertainty due to the displacement error. In Section 4.3, we develop the Bootstrap bias correction (BC) method for obtaining bias-corrected estimates of the linear mixed model parameters under displacement and variance estimation method of model parameters estimates under the BC. Model-based simulation studies by bringing some characteristics of the DHS EA are conducted in Section 4.4 to assess the impact of the DHS EA displacement on the linear mixed model estimation and the performance of the proposed EDC and BC methods. We claim the BC method works well for correcting the bias of the model parameters, while the EDC method is for the group means. In Section 4.5, we discuss the simulation study results. Finally, we conclude with a discussion in Section 4.6.

4.2 Development of a framework of LMMs under random displacement

The linear mixed model (in a matrix notation) is described as follows

$$y = X\beta + Z(T)u + \varepsilon; \quad u \sim N\left(0, \sigma_u^2 I_m\right) \quad \text{and} \quad \varepsilon \sim N\left(0, \sigma_\varepsilon^2 I_n\right), \tag{4.1}$$

where y is a $n \times 1$ vector of responses, X is a $n \times p$ known design matrix for the fixed effects, β is a $p \times 1$ parameter vector of fixed effects, $Z\left(T\right) = [Z_1(T), \cdots, Z_m(T)]$, where $Z_i(T)$ is a $n \times q_i$ design matrix for the random effects specified to group $i \quad (i = 1, \cdots, m)$ based on $n \times 2$ true location coordinates vector T where the responses are measured and it can be household or EA centroid locations, $u = [u_1, \cdots, u_m]$ is a $q \times 1$ parameter vector of random effects where u_i is a $q_i \times 1$ vector such that $q = \sum_{i=1}^{m} q_i$, and ε is a $n \times 1$ vector of random errors. The unknown true parameters of the linear mixed model (4.1) include $\theta = (\beta, \sigma_u^2, \sigma_\varepsilon^2)$.

For the above linear mixed effect model (4.1), the vector of true location coordinates T is unknown and can be treated as missing while only the vector of displaced location coordinates W is observed. Also, the random effects u are treated as missing data. Therefore, we take the complete dataset to be $y^\star = \{y, u, Z(T), W\}$ where $\{y, W\}$ is the observed dataset.

Under the proposed model, the joint distribution of the data is factorised as follows,

$$\begin{aligned} f(y^\star; \theta) &= f(y, u \mid Z(T), W) f(T, W) \\ &= f(y, u \mid Z(T)) f(W \mid T) f(T), \end{aligned} \tag{4.2}$$

where, conditionally on T, the joint distribution of (y, u) and $Z\left(T\right)$ do not depend on W. Although the displacement mechanism is random and independent of the outcome y, it is possible to use the displaced coordinates W to inform about true location of observations given the relationship between W and T.

The first component on the right hand side (second line) of (4.2) defines the outcome model. The second component defines the error model for W given the true coordinates and the third term defines the marginal distribution of the unknown true coordinates. The last two terms together define the conditional distribution of the true locations T given the displaced locations W. We developed this conditional distribution $f(T \mid W)$ in Section 3.2.4. To recall, $f(T = T^\star|W) \propto f(W|T = T^\star) f(T = T^\star)$ where $T^\star$ is the vector of possible true locations. Utilizing this formulation we

can draw pseudo-samples (imputations) of T from $f(T = T^{\star}|W)$ using the estimation methods, e.g., the external data-based classification (EDC) under displacement to construct the design matrix $Z(T = T^{\star})$ in (4.1). The proposed EDC method, which we developed in Section 3.4 under a functional measurement error model by using additional information (e.g., population density, designated administrative boundary restriction, rural urban boundaries, boundary corrections for non-settlement places) from external sources. Given the superior performance of the proposed EDC method over the proposed KDE method, as demonstrated in our empirical assessment in Section 3.6.1, we do not use the KDE method in subsequent applications.

4.2.1 Estimation

As the vector of true locations T is unknown, we can draw pseudo-samples $T^{\star}$ of T using the EDC method and construct the unknown design matrix for the random effect u which is denoted by $Z(T = T^{\star})$. Therefore, fixing $T = T^{\star}$, one can fit the linear mixed model described by the first part on the right hand side of (4.2) and obtain estimated model parameters $\hat{\theta} = \left(\hat{\beta}, \hat{\sigma}_u^2, \hat{\sigma}_{\varepsilon}^2\right)$. Also, we obtain $E(y \mid u, T = T^{\star}) = X\hat{\beta} + Z(T = T^{\star})\hat{u}$, where $\hat{u} = \hat{\sigma}_u^2 Z'(T = T^{\star})\hat{\Sigma}_y^{-1}(y - X\hat{\beta})$ is the empirical best linear unbiased predictor (EBLUP) of the random effect u and $\hat{\Sigma}_y = Z(T = T^{\star})Z'(T = T^{\star})\hat{\sigma}_u^2 + \hat{\sigma}_{\varepsilon}^2 I_n$ is the estimated covariance matrix of y.

In particular, we use the Stochastic Expectation-Maximisation (SEM) algorithm that works by replacing the un-observed true point T in the complete data likelihood by drawing pseudo-samples of the un-observed true point given the observed data using the proposed EDC method under displacement (S-step) and then maximises the (complete) data likelihood for the updated (new) location samples and obtains the model parameter estimates and group means in the M-step. The process is repeated many times and we take the average of the estimates. Studies by Warren et al. (2016b) and Wilson et al. (2020) recommended drawing a large number of samples for better realizations of the true location. The computational steps of the algorithm are given as follows:

- Step 1: Generate pseudo-samples $T^{\star}$ of the true location coordinates T from the conditional distribution $f(T = T^{\star}|W)$ using the proposed EDC method under displacement described in Section 3.4.

- Step 2: Construct the design matrix $Z(T = T^{\star})$ using the pseudo-samples $T^{\star}$ of the true location coordinates T from Step 1 and fit the linear mixed model (4.1) to obtain the estimates $\hat{\theta} = \left(\hat{\beta}, \hat{\sigma}_u^2, \hat{\sigma}_{\varepsilon}^2\right)$ and the random effects $\hat{u}$ for each group i. Hence, obtain $E(y \mid u, T = T^{\star}) = X\hat{\beta} + Z(T = T^{\star})\hat{u}$ for each group i. The parameters are estimated using restricted maximum likelihood (REML) and using the empirical best linear unbiased

predictor (EBLUP) for u.

- Step 3: Iterate Steps 1-2 $B_1 + B_2$ times, with $B_1 = 300$ (large number of iterations) and $B_2 = 20$ (additional iterations).
- Step 4: Obtain the estimates $\widehat{\theta}^{(B_1)}$ and $\widehat{\theta}^{(B_1+B_2)}$ by averaging the derived B_1 and (B_1+B_2) estimates respectively. We then do trace plots for the average estimates to see if the two estimates, $\widehat{\theta}^{(B_1+B_2)}$ and $\widehat{\theta}^{(B_1)}$, are reasonably close. If they are, the algorithm converges. Otherwise, return to Step 1 and proceed through the subsequent steps, incrementing B_1 by a predetermined amount (for instance, $B_1 = B_1 + 20$), while maintaining the constant value $B_2 = 20$, until the algorithm converges.

By using the design matrix $Z\,(W)$ directly instead of $Z\,(T)$, based on the displaced location coordinates vector W in the model (4.1), the estimates of $\theta = \left(\beta, \sigma_u^2, \sigma_\varepsilon^2\right)$ are considered as the naive estimates that ignore the DHS displacement uncertainty.

4.2.2 Variance estimation of parameter estimates under the EDC method

In addition to point estimates, we are also interested in estimating the variances of the mixed model parameter estimates under displacement. To do so, we propose to use the law of iterated variances or total variance (see for details, Blitzstein and Hwang, 2019, section 9.5), conditional on the design matrix $Z(T = T^\star)$. The variance decomposition equation is expressed as:

$$V(\hat{\theta}) = E_{Z(T^\star)}[V(\hat{\theta} \mid Z(T^\star))] + V_{Z(T^\star)}[E(\hat{\theta} \mid Z(T^\star))]. \tag{4.3}$$

The first component on the right hand side of (4.3) defines the expected value of the conditional variances of the estimated parameter given the design matrix $Z(T = T^\star)$, where $T^\star$ is the vector of possible true locations obtained using the proposed EDC method under displacement described in Section 3.4. The second component defines the variance of the conditional expected values of the estimated parameter given the design matrix $Z(T = T^\star)$.

In order to compute the total variance of the model parameters estimates, for each set of pseudo-samples of true locations $T^\star$, we construct the design matrix $Z(T = T^\star)$ and hence, fit the linear mixed model (4.1) to obtain the estimates of the model parameters and their variances. The process is repeated a large number of times and we calculate the average of the variances of the estimated parameter and the variance of the parameter estimates, which gives the first and second components on the right hand side of (4.3), respectively.

4.3 Development of Bootstrap bias correction method under random displacement

The vector of true location coordinates T is unknown to the data analysts, and only the vector of displaced coordinates W is available. Therefore, instead of $Z(T)$ we use directly the design matrix $Z(W)$ based on the displaced EA coordinate vector W in the linear mixed model (4.1). This is written as

$$y = X\beta + Z(W)u + \varepsilon. \tag{4.4}$$

If we fit the model (4.4) to the displaced data vector (y, X, W) and obtain the estimated parameters $\hat{\theta} = \left(\hat{\beta}, \hat{\sigma}_u^2, \hat{\sigma}_\varepsilon^2\right)$ by using the restricted maximum likelihood (REML), these naive estimates of the linear mixed model parameters, especially the variance components, and the model-based finite population parameter, e.g., mean, are biased under the misplacement due to the random displacement of EA locations. So, we can attempt to obtain the bias estimates under displacement and develop a bias-corrected estimator of the parameters. In particular, we can use a bootstrapping technique to estimate the bias. The general idea of the simple but powerful bootstrap bias estimation method is that it can estimate the bias of the (biased) estimator through bootstrap samples generated by repeating the sampling and displacement processes to the displaced data. For details about the bootstrap bias estimate, see Efron and Tibshirani (1992), section 10.2.

Regarding the bias of the linear mixed model parameters under displacement, the bootstrap bias correction method should work to satisfy the two important assumptions. We first assume that the functional form of the linear mixed model (4.1) is appropriate to the data. Next, we assume that the magnitude of the bias of the naive estimates of the model parameters under displacement is the same as the magnitude of the bias of the estimates under bootstrap samples and repeated displacement process to the displaced locations. We generate the bootstrap samples under the model (4.4) using the naive estimates $\hat{\theta}$. If these assumptions are satisfied, the bootstrap bias correction (BC) method can obtain unbiased estimates of the model parameters under displacement. We claim that the BC method should work for correcting the bias of the global (model) parameters, and it may not work for the associated model-based finite population parameters, such as group means under displacement. This is because our second assumption of the bootstrapping regarding the magnitude of the bias of the predicted group means under displacement may not be satisfied. We have discussed some potential reasons for this in Section 4.1. In that case, the EDC method should be preferred to the BC method for estimating group means under misplacement. We explore this in Sections 4.4-4.5 under model-based simulation. We describe the proposed parametric bootstrap

bias correction (BC) method for correcting the bias of the linear mixed model parameters under displacement as follows: The bias of the naive estimates ($\hat{\theta}$) of the model parameters (θ) is given by Bias $\left(\hat{\theta}, \theta\right) = E\left(\hat{\theta}\right) - \theta$. This bias is unknown to us, as we do not know the true model parameters (θ). We obtain estimates of the bias by using the $B\,(b = 1, 2, \cdots, B)$ model-based bootstrap samples and displaced sets of EA coordinates which are obtained by applying the displacement process to the displaced (observed) EA coordinates. These model-based bootstrap samples using the naive estimates $\hat{\theta} = \left(\hat{\beta}, \hat{\sigma}_u^2, \hat{\sigma}_\varepsilon^2\right)$ are obtained from the following model

$$y^{*b} = X\hat{\beta} + Z\left(W\right)u^{*b} + \varepsilon^{*b}, \tag{4.5}$$

where, $u^{*b} \sim N\left(0, \hat{\sigma}_u^2 I_m\right)$ and $\varepsilon^{*b} \sim N\left(0, \hat{\sigma}_\varepsilon^2 I_n\right)$. Also, let the W^{*b} denote the set of B displaced EA coordinates which are obtained by displacing the given displaced (observed) EA coordinates W using the DHS displacement algorithm. Next, we obtain the bootstrap estimates of the parameters denoted by $\hat{\theta}^{*b}; b = 1, 2, \cdots, B$ using the bootstrap samples data vector $\left(y^{*b}, X, W^{*b}\right)$. Then, the bootstrap estimate of the bias is obtained by computing

$$\widehat{\text{Bias}}\left(\hat{\theta}, \theta\right) = \sum_{b=1}^{B} \hat{\theta}^{*b}/B - \hat{\theta}. \tag{4.6}$$

Finally, the bootstrap bias corrected (BC) estimator is defined as

$$\hat{\theta}_{\text{cor}} = \hat{\theta} - \widehat{\text{Bias}}\left(\hat{\theta}, \theta\right). \tag{4.7}$$

4.3.1 Variance estimation of parameter estimates under the BC method

We want to obtain the estimated variance of the bootstrap bias corrected estimates in (4.7) of the linear mixed model parameters accounting for random displacement error. The variance of the BC estimator $(\hat{\theta}_{\text{cor}})$ can be expressed in terms of the variance and covariances of the naive estimator $(\hat{\theta})$ and the bootstrap bias estimator, $\widehat{\text{Bias}}\left(\hat{\theta}, \theta\right)$ in (4.6) as follows:

$$V(\hat{\theta}_{\text{cor}}) = V(\hat{\theta}) + V\left(\widehat{\text{Bias}}\left(\hat{\theta}, \theta\right)\right) - 2 \times \text{Cov}\left(\hat{\theta}, \widehat{\text{Bias}}\left(\hat{\theta}, \theta\right)\right). \tag{4.8}$$

It would be challenging to obtain the covariance between the naive and bootstrap bias estimates based on a single level of B bootstrap samples described in Section 4.3. In that case, a double bootstrapping can be used to obtain the estimated variance of the BC estimator. In double boot-

strapping, we use two rounds of bootstrap sampling. To describe the process in general terms, in the first round, we take $B \quad (b = 1, 2, \cdots, B)$ bootstrap samples from the original sample. In the second round, we take $C \quad (c = 1, 2, \cdots, C)$ bootstrap samples from each of the B bootstrap samples obtained in the first round and obtain a bias-corrected estimate based on the C bootstrap samples. This results in a total of $B \times C$ bootstrap samples. So, it can be computationally intensive, but it would be a useful technique to obtain a more accurate estimate of the variance of the bootstrap bias-corrected estimator. However, we use model-based bootstrap samples under repeated displacement to account for displacement error, as described in Section 4.3. The computational details for obtaining the variance estimates of the BC estimator under a double bootstrapping are given below:

Round 1: Take $B \quad (b = 1, 2, \cdots, B)$ bootstrap samples under the model (4.5) using naive estimates $\hat{\theta} = \left(\hat{\beta}, \hat{\sigma}_u^2, \hat{\sigma}_\varepsilon^2\right)$, which are obtained by fitting the model (4.4) to the displaced data (y, X, W). Next, obtain B displaced sets of EA coordinates, denoted as $W^{*b}; b = 1, 2, \cdots, B$, by displacing the given displaced EA coordinates W using the DHS displacement algorithm. Then, the first round B bootstrap samples data are represented as (y^{*b}, X, W^{*b}).

Round 2: Take $C \quad (c = 1, 2, \cdots, C)$ bootstrap samples from each of the $B \quad (b = 1, 2, \cdots, B)$ bootstrap samples obtained in the first round and obtain the bias-corrected model parameter estimates based on the C bootstrap samples. We describe this below:

1) for each $b = 1, 2, \cdots, B$, we obtain estimates of the model parameters denoted by $\hat{\theta}^{*b} = \left(\hat{\beta}^{*b}, \hat{\sigma}_u^{2*b}, \hat{\sigma}_\varepsilon^{2*b}\right)$ using the first round bootstrap samples data (y^{*b}, X, W^{*b}).

2) we obtain C bootstrap samples using estimates $\hat{\theta}^{*b}$ under the following model:

$$y^{*c} = X\hat{\beta}^{*b} + Z\left(W^{*b}\right)u^{*c} + \varepsilon^{*c}, \tag{4.9}$$

where, $u^{*c} \sim N\left(0, \hat{\sigma}_u^{2*b} I_m\right)$ and $\varepsilon^{*c} \sim N\left(0, \hat{\sigma}_\varepsilon^{2*b} I_n\right)$.

3) we obtain C displaced sets of EA coordinates, denoted as $W^{*c}; c = 1, 2, \cdots, C$, by displacing the given displaced EA coordinates W^{*b} using the DHS displacement algorithm. Therefore, the second round C bootstrap samples data are represented as (y^{*c}, X, W^{*c}).

4) by using the second round bootstrap samples data (y^{*c}, X, W^{*c}), we obtain the estimates of the model parameters denoted by $\hat{\theta}^{*c}; c = 1, 2, \cdots, C$. We use these second round bootstrap samples-

based model parameter estimates to obtain the estimate of the bias of $\hat{\theta}^{*b}$ defined as

$$\widehat{\text{Bias}}\left(\hat{\theta}^{*b}, \theta\right) = \sum_{c=1}^{C} \hat{\theta}^{*c}/C - \hat{\theta}^{*b}. \tag{4.10}$$

5) for each $b = 1, 2, \cdots, B$, we obtain the bootstrap bias corrected estimates as

$$\hat{\theta}^{*b}_{\text{cor}} = \hat{\theta}^{*b} - \widehat{\text{Bias}}\left(\hat{\theta}^{*b}, \theta\right). \tag{4.11}$$

6) by using the above B bias-corrected estimates, we calculate the variance of the BC estimator as follows:

$$\hat{V}(\hat{\theta}_{\text{cor}}) = \sum_{b=1}^{B} \left(\hat{\theta}^{*b}_{\text{cor}} - \bar{\hat{\theta}}^{*}_{\text{cor}}\right)^2 /B, \tag{4.12}$$

where $\bar{\hat{\theta}}^{*}_{\text{cor}}$ is the average of the estimates $\hat{\theta}^{*b}_{\text{cor}}$ over B first round bootstrap samples. It can be noted that the estimated variance $\hat{V}(\hat{\theta}_{\text{cor}})$ is the sample variance of the bootstrap distribution of $\hat{\theta}_{\text{cor}}$ over B samples.

4.4 Model-based simulation

This section aims to assess the ability of the proposed EDC and BC methods to account for misplacement error uncertainty due to the DHS displacement process under a model-based simulation using the geography of Bangladesh with 544 upazilas (admin 3). Hence, to provide improved estimates of the linear mixed model parameters and upazila means compared to naive estimates that ignore the misplacement error. In particular, we evaluate the EDC, BC and naive methods using simulated data generated under the linear mixed model. The model-based simulation enables a controlled empirical evaluation of our proposed methods for both point and variance estimates. The naive method that ignores the misplacement error produces biased estimates of the linear mixed model parameters, especially the variance components, and upazila means under misplacement due to the displacement process. We investigate the effect of the misplacement error on the naive estimates under two distinct scenarios for the intensity of the displacement process. The proposed EDC method is developed under the functional measurement error model by bringing additional information from different external sources to approximate the marginal distribution of the true locations. Therefore, we aim to show that the proposed EDC method for point estimates of model parameters and upazila means performs better than the naive method. We also aim to show that the

BC method produces unbiased estimates of the global (model) parameters satisfying the assumptions on the model form and the magnitude of the bias under displacement and performs better than the EDC and naive methods. In contrast, for estimating upazila means, the EDC method is preferred to the other two methods.

We generate the model-based simulated data by incorporating certain characteristics from the 2011 Bangladesh Demographic and Health Survey (BDHS) EAs. The true coordinates of the 2011 BDHS EAs are not available. According to the documentation of the BDHS 2011 (NIPORT, 2013), as a sampling frame, the survey used the list of EAs prepared for the Bangladesh Population and Housing Census 2011. The list of the 2011 census EAs with their number of households is also not available. Thus, to generate true coordinates, we create a sampling frame, the list of EAs with their number of households, as close as possible to the 2011 BDHS sampling frame by using the WorldPop population data. In particular, for this simulation study, we generate the true EA coordinates using the WorldPop gridded population following the same principles of the 2011 BDHS EAs selected from the 2011 census EAs. We discussed the principles in Chapter 1.

We create EAs by aggregating the WorldPop gridded population. In particular, we use the $100m \times 100m$ gridded population raster from WorldPop that estimates the population of Bangladesh, adjusted using the 2011 census population, where each grid cell represents the population count per pixel. The constrained population raster in Bangladesh for 2020 can be downloaded from WorldPop (Bondarenko et al., 2020).

At first, we identify whether the worldPop grid cells fall into rural or urban regions by using the rural-urban boundaries from the Bangladesh 2011 census. According to the 2011 census, there were 296,718 EAs, and their location coordinates and the information on the proportion of rural/urban EAs are not available. We find about 16% of the WorldPop grid cells were from urban areas. Therefore, for this simulation study, we generate 16% of the total EAs from the urban grid cells and the rest of the EAs from the rural grid cells. Next, the number of rural and urban EAs for each upazila is determined proportionally to the number of rural and urban grid cells for each upazila in the WorldPop population raster. Then, following the 2011 census EA generation principles (BBS, 2011), we use the mauza geography to calculate population per mauza using the WorldPop grid cells population. In order to find the number of households, we divide the mauza's total population by 4.4, which is the average number of household members according to the 2011 census of Bangladesh. After that, we merge smaller mauzas or divide larger mauzas to build the EAs, making sure to respect the upazila geography, i.e., not merging mauzas from different upazilas. In some cases, a whole EA is a mauza. Mauzas having more than 200 households are divided into

two or more EAs. Also, mauzas having less than 70 households were merged with other adjacent smaller mauzas/EAs. Following this way, the boundary of the EAs are created with irregular shapes and we place the EA's total population approximately at the centre of the EA. Finally, we create 274,150 EAs. For this empirical study, we consider the list of these created EAs along with their population as the sampling frame of the survey EAs, the primary sampling units (PSUs). The sampling frame contains information about the coordinates of the centroid of the EAs, type of residence (rural or urban), upazila IDs and the number of households. An upazila is divided into rural and urban strata. The summary statistics of the EAs' number of households are given in Table 4.1. Notice that the average number of households per EA is about 130 with a standard deviation equal to 38.6 households which is consistent with that of the 2011 census.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	SD
	57.0	94.0	138.0	130.5	157.0	250.0	38.6

Table 4.1: Summary statistics of the EAs number of households.

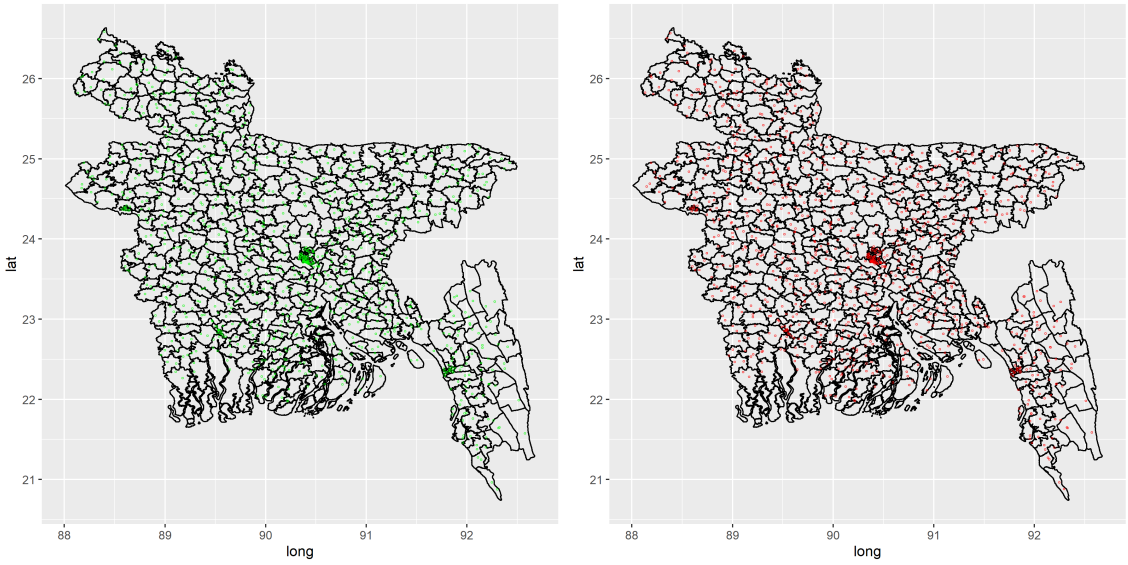

Figure 4.1: Spatial distribution of the true EA coordinates (left) and displaced EA coordinates (right) at upazila level of Bangladesh.

Following the 2011 BDHS EA selection process, we draw a sample of 1050 EAs with probability proportional to the EA population stratified by the upazila and by the rural urban areas within the upazila. In particular, we use the R package *sampling* (Tillé and Matei, 2022) and the systematic sampling method for PPS sampling without replacement to draw samples of EAs with probability proportional to the number of households. We note that one EA is selected from each stratum. Therefore, each upazila has two sampled EAs unless it has only a rural/urban region. Under this simulation design, the 1050 selected EA coordinates are considered as the true EA coordinates and the EA sample is fixed, see Figure 4.1 (left) for true EA locations. We consider 30 households for

each EA, and the households within an EA have the same coordinates as the EA centroid coordinate as the 2011 BDHS, i.e., households are georeferenced at the centre of the EA. It should be noted that in this simulation study, the issue of aggregating households is being ignored. Therefore, the total number of households is $n = 31,500$ which is spread over $m = 544$ upazilas of Bangladesh. The upazila-specific number of households ranges from 30 to 60. We generate outcomes for each household under the linear mixed model (4.1) with an explanatory variable defined as:

$$y = 1\beta_0 + \beta_1 X_1 + Z(T)u + \varepsilon; \quad u \sim N\left(0, \sigma_u^2 I_m\right) \text{ and } \varepsilon \sim N\left(0, \sigma_\varepsilon^2 I_n\right),$$

where $Z\left(T\right)$ is $31,500 \times 544$ known design matrix for the random effects using the 1050 true EA coordinates T, the vector of 1's is represented by '1' and X_1 is the fixed and known covariate that we generate from a uniform distribution with a minimum of 1 and maximum of 8, i.e., $X_1 \sim$ Uniform$(n, \quad \text{min} = 1, \quad \text{max} = 8)$.

The simulated data generation steps are as follows:

1) Generate $u \sim N(m, \quad \text{mean} = 0, \quad \sigma_u^2 = 5)$
2) Generate $\varepsilon \sim N(n, \quad \text{mean} = 0, \quad \sigma_\varepsilon^2 = 25)$
3) Generate $y = 1\beta_0 + \beta_1 X_1 + Z(T)u + \varepsilon$, where $\beta_0 = 35$ and $\beta_1 = 2.85$. It is noted that the coefficient of determination $R^2 = 0.53$ for the simple linear regression of y on X_1. Also, the intra-class correlation coefficient is 0.167, which is consistent with the 2011 BDHS data.
4) Create the true dataset $(y, X_1, Z(T))$ and the displaced dataset $(y, X_1, Z(W))$,

where $Z(W)$ is constructed based on the displaced coordinates generated by applying the DHS displacement algorithm to the true EA coordinates under two different displacement scenarios. Under scenario A (DHS scenario) data are displaced by using the maximum displaced distance in line with the DHS displaced distance parameters, which is 2km for urban EA points, and 5km for rural points with 1% of rural points being randomly allocated 10km maximum distance, see Figure 4.1 (right) for displaced EA locations. Further, to assess the influence of a more intensive displacement (higher displacement distance) which will intensify the misplacement issue, under scenario B, data are displaced by using the maximum displaced distance of 5km for urban points, and 10km for rural points with 1% of rural points being allocated 15km maximum distance.

We calculate the misplacement statistics for both Scenarios A and B using a single set of displacement coordinates, creating displacement buffers around the centroid of the EA displaced coordinates. We check whether the EA displacement buffers intersect with an upazila boundary, indicative of potential misplacements. Additionally, we identify the EAs that are truly misplaced, wherein the displaced upazila ID varies from the actual upazila ID. This simulation study deter-

mines the displaced and actual upazila ID for each EA based on their respective displaced and true EA coordinates. In Scenario A, of the 1050 EAs investigated, approximately 41% are potentially misplaced, and around 30% of these are demonstrated as truly misplaced. Conversely, Scenario B shows a higher misplacement rate, with approximately 49% of the EAs potentially misplaced and around 33% of these confirmed as truly misplaced.

The simulation steps (generation of the true outcomes under the model and displaced datasets, and the estimates) are independently repeated 300 times. Under the repeated displacement, we compare the effectiveness of the naive, the EDC and the BC estimators in the linear mixed model based estimates by computing the Root-Mean-Square-Error (RMSE), Relative Bias in Percent (RB) and the Absolute Relative Bias in Percent (ARB) for the ith $(i = 1, \cdots, m)$ upazila mean estimate over $L = 300$ simulations as follows:

$$\text{RMSE}_i = \sqrt{\frac{\sum_{l=1}^{L}(\hat{\bar{Y}}_{li} - \bar{Y}_{li})^2}{L}} \tag{4.13}$$

$$\text{RB}_i = \frac{\sum_{l=1}^{L}(\frac{\hat{\bar{Y}}_{li} - \bar{Y}_{li}}{\bar{Y}_{li}})}{L} \times 100 \tag{4.14}$$

$$\text{ARB}_i = \frac{\sum_{l=1}^{L}(|\ \frac{\hat{\bar{Y}}_{li} - \bar{Y}_{li}}{\bar{Y}_{li}}\ |)}{L} \times 100, \tag{4.15}$$

where $\hat{\bar{Y}}_{li}$ represents the estimated mean in upazila i using any of the aforementioned methods, while $\bar{Y}_{li}$ represents the true mean for upazila i in simulation round l.

Also, under the repeated displacement, we compare the usefulness of the naive, the EDC and the BC methods by computing the RMSE and the Bias for the estimates of the linear mixed model parameters over $L = 300$ simulations as follows:

$$\text{RMSE}(\hat{\theta}) = \sqrt{\frac{\sum_{l=1}^{L}(\hat{\theta}_l - \theta)^2}{L}} \tag{4.16}$$

$$\text{Bias}(\hat{\theta}) = \frac{\sum_{l=1}^{L}(\hat{\theta}_l - \theta)}{L} \tag{4.17}$$

where $\hat{\theta}_l$ is the estimates of the linear mixed model parameters using any of the aforementioned

methods in simulation round l and θ defines the true value of the parameter.

To implement the model-based simulation, we use **R** (R Core Team, 2022). We fit the linear mixed model to the data and obtained the parameters estimates using the package **lme4** (Bates et al., 2015).

4.5 Results and discussion

In this section, we present and discuss the model-based simulation study results. Section 4.5.1 reports and discusses the results of the effect of the displacement process on the estimates of the linear mixed model parameters, and upazila means under two displacement scenarios, A and B. The results on the performance of the proposed EDC and BC methods for estimating the linear mixed model parameters and upazila means under displacement are presented and discussed in Section 4.5.2. Finally, Section 4.5.3 reports the results of the estimated variance of the linear mixed model parameters estimates using the proposed estimators under the EDC and BC methods.

4.5.1 Assessing the effect of displacement on the LMM estimation

The 300 simulated data sets and the EA coordinates under the displacement scenarios of no misplacement, low (A) and high (B) displacement intensities are used to fit the linear mixed model and to obtain the upazila level mean estimates and their RMSEs, RBs and ARBs. Also, we have calculated the summary statistics of the distribution of the estimated model parameters. The results are presented in Figures 4.2-4.4.

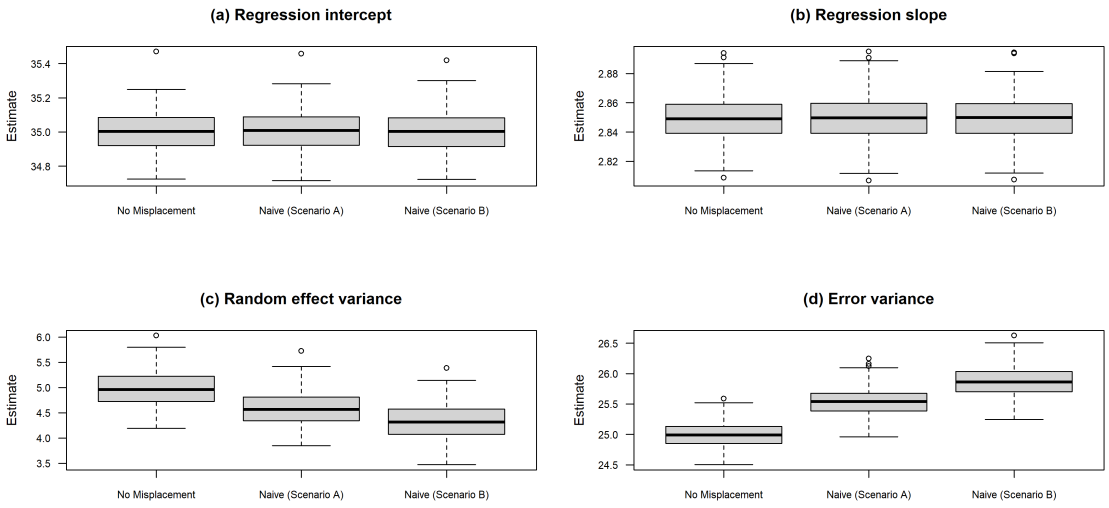

Figure 4.2: Plot of the distribution of the estimated parameters a) regression intercept, b) regression slope, c) random effect variance and d) error variance of the LMM using the naive estimator over the 300 simulation runs for the displacement scenarios A and B.

We start with a focus on the empirical assessment of the displacement effects on the estimates of the LMM parameters using the naive method that ignores the displacement error. Figure 4.2 shows distribution of the estimated parameters of the linear mixed model using the naive estimator over the 300 simulation runs for the displacement scenarios A and B. In this empirical study, the true values of the parameters considered are $\beta_0 = 35, \beta_1 = 2.85, \sigma_u^2 = 5$ and $\sigma_\varepsilon^2 = 25$ respectively. We observe that the naive estimates of the variance component parameters, i.e., the random effect variance and the error variance are biased under displacement scenarios A and B, and the magnitude of the bias increases as the maximum displacement distance increases. In contrast, as expected, the estimates of the random effect variance and the error variance are not biased under the scenario of no misplacement (true data) of EAs due to the displacement process. For example, the expected values of the random effect variance estimator over 300 simulation runs for the no, low, and high misplacement error due to the displacement process are 4.983, 4.566 and 4.345 respectively, while the true parameter value is 5.0. The bias in the variance components can be explained by a wrong classification of EAs, which leads to mixing observations from different groups. This makes the groups less homogeneous and less distinctive, resulting in lower random effect variance and higher error variance. However, the total variance remains unchanged. The results in the considered simulation scenario indicate that the fixed effect regression coefficients are not affected by the displacement process and are relatively close to their parameter values. This is because the measure of the fixed covariate values is not dependent on the EA coordinates. Moreover, although the estimates of the variance components are involved in the computation process to estimate fixed effect parameters using generalised least squares (GLS), their presence in both numerator and denominator may neutralise the effect of the biased estimates of the variance components under displacement.

We also observe that the random displacement process affects the estimated standard errors of the fixed effect and random effect parameters of the LMM (Figure 4.3). When the maximum displacement distance increases, the variation of the regression intercept and the random effect variance parameter estimates reduces compared to the variance estimates under no misplacement (true data) due to the displacement process. In contrast, the variation increases for the regression slope and error variance estimates when the maximum displacement distance increases. This is because the point estimates of the variance components are affected by the displacement process; thus, as expected, their variance estimates are also affected. In contrast, although the point estimates of the fixed effect parameters are not affected by the displacement process, their variance estimates are affected because the variance estimation involves the point estimates of the variance components

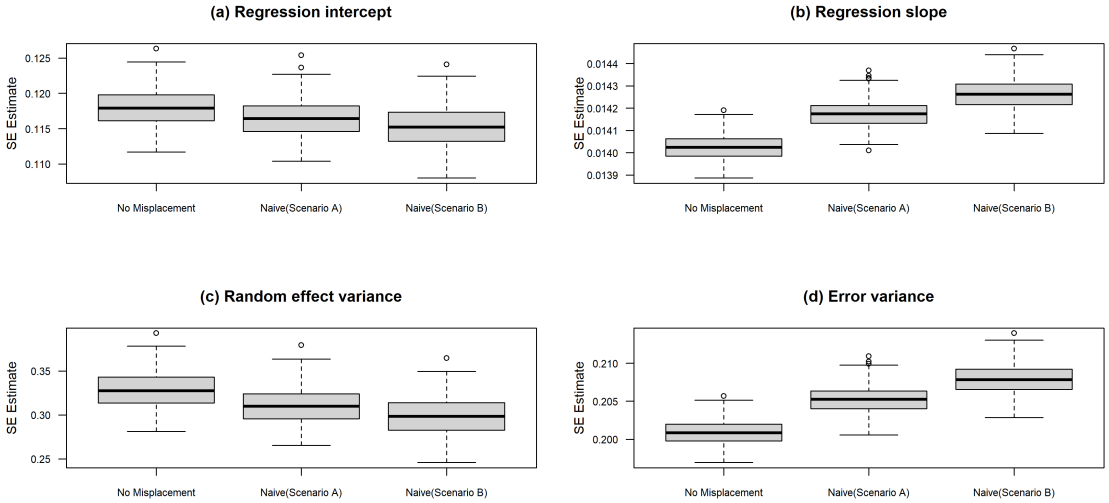

Figure 4.3: Plot of the distribution of the estimated standard errors (SE) of the LMM parameters a) regression intercept, b) regression slope, c) random effect variance and d) error variance using the naive estimator over the 300 simulation runs for the displacement scenarios A and B.

that are biased under displacement.

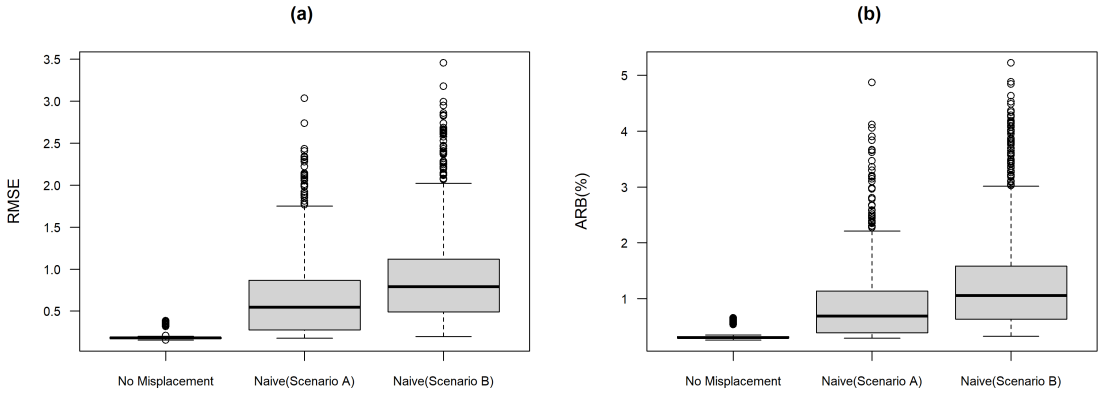

Figure 4.4: Plot of the distributions of the (a) RMSEs and (b) ARBs of the estimated upazila means using the naive estimator over the 300 simulation runs for the displacement scenarios A and B.

We now focus on the results for the assessment of the displacement effects on the upazila means estimates. Figure 4.4(a) presents the distribution of the RMSEs of the estimated upazila means using the naive estimator over the 300 simulation runs for the displacement scenarios A and B. These results indicate larger RMSEs of the estimates as the intensity of displacement increases. For example, the averages of RMSE's of the estimated mean over 544 upazilas for the no, low and high displacement errors are 0.1911, 0.6724 and 0.9382 respectively. Also, to assess the impact of the EA misplacement on the bias of the upazila mean estimates, the distribution of the ARBs of the estimated 544 upazila means using the naive estimator for each displacement scenario are

presented in Figure 4.4(b). Like the RMSE's, the results show that the naive upazila mean estimates have higher biases as the displacement error increases. For example, the averages of ARBs of the estimated mean over 544 upazilas for the no, low and high displacement error are 0.3207, 0.9287 and 1.3409 respectively.

4.5.2 Evaluation of the performance of the proposed methods

This section focuses on the simulation study results for evaluating the performance of the proposed EDC and BC methods for the point estimates of the LMM parameters and upazila means correcting for displacement error over the naive method that ignores the displacement error. Overall, we aim to see that the proposed BC method outperforms the EDC and naive methods for estimating the LMM parameters under displacement. We also expect that for the estimates of upazila means, the proposed EDC method is preferred to the two competing BC and naive methods, as we have explained some potential reasons in Section 4.1. To remind the reader, the BC method may not work for correcting the bias of the estimates of upazila means due to possible issues related to the bootstrapping assumptions presented in the methodology Section 4.3. Two likely reasons are: first, creating a buffer around truly non-misplaced EAs unknown to the data analysts may introduce additional uncertainty in the estimates of upazila means. Second, due to random displacement, there may be a mixing of groups within the buffer around the true and displaced locations, potentially affecting the estimates of group means. In this section, the results for each method are produced based on 300 simulation runs. Moreover, for each simulation, the results of the EDC based method are produced based on 300 pseudo-samples (imputations) of the true EA coordinates.

We first report and discuss the results of the point estimates of the LMM parameters using the BC, EDC and naive methods. Also, to validate the results by these methods, we present the results of the point estimates using non-misplaced (true) data. To see the complete picture of the distribution of the point estimates, the density curves of the estimated regression parameters and variance component parameters of the linear mixed model for the naive, EDC and BC estimators over 300 simulations for displacement scenarios A and B are presented in Figure 4.5. We observe from Figure 4.5(a-d) that the overall shape of the distributions of fixed effect parameters estimates using the naive, EDC and BC methods is very similar to the distribution using non-misplaced data for each displacement scenario. Also, the summary statistics, e.g., median, of the estimates of regression coefficients over 300 simulation runs using the naive, EDC and BC methods are approximately equal to their validated estimates (Figure 4.6(a-d)). For each method, the expected value of the regression parameters estimates over 300 simulation runs is approximately equal to the parameter

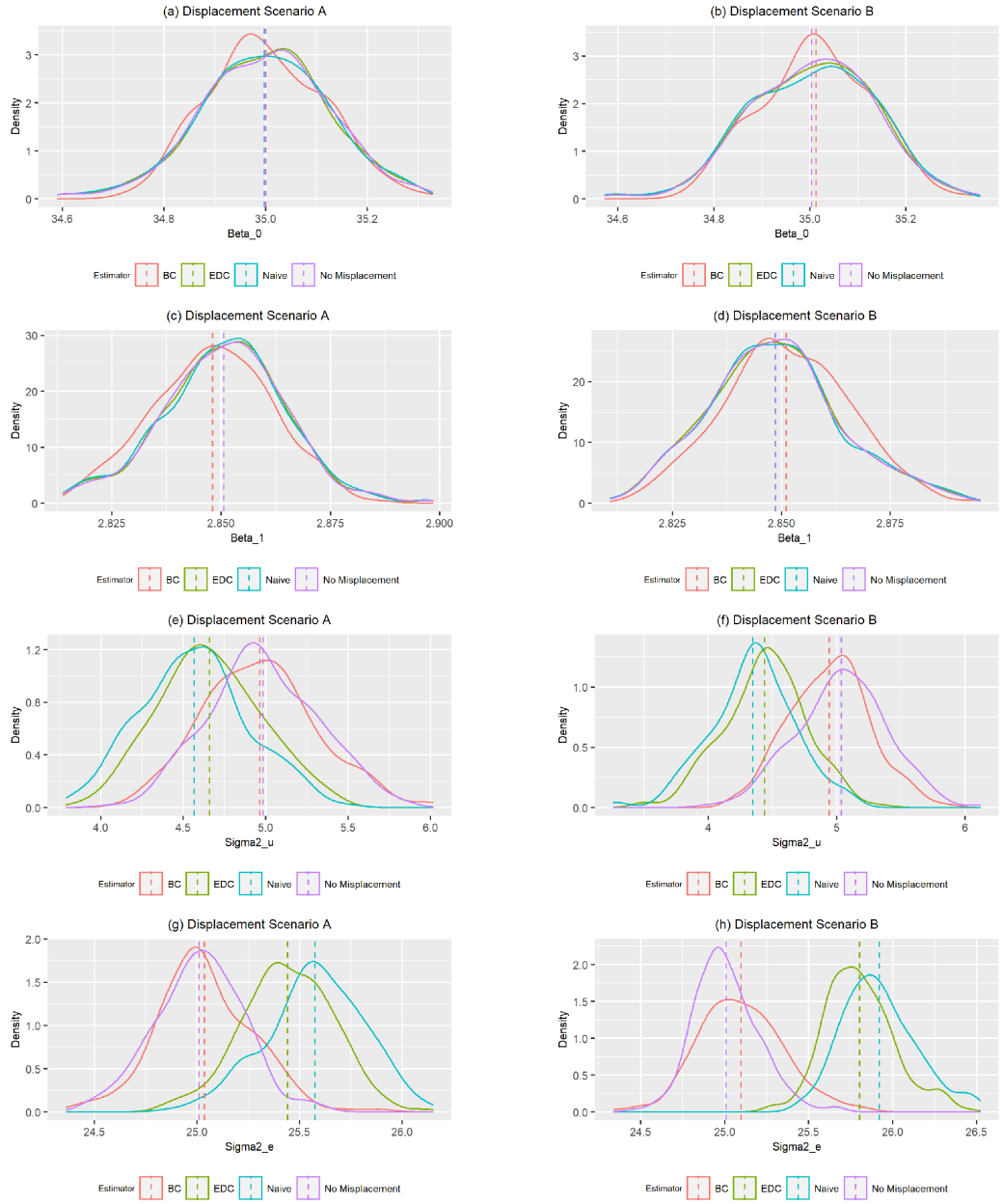

Figure 4.5: Density curves of the estimated parameters of the linear mixed model for each estimator over the 300 simulation runs for the displacement scenario A and B. For EDC based estimator, the parameter estimates are averaged over 300 pseudo-samples of the true EA coordinates for each simulation. The vertical dotted lines indicate the mean estimates for each estimator.

value. Therefore, as anticipated, all three methods perform on the same level for the point estimates of the fixed effect coefficients for each displacement scenario, as the estimates are not affected by the displacement process.

However, the estimates of the variance component parameters, especially the random effect variance, are affected by the displacement. Figure 4.5(e-h) shows the density curves of the estimates of the variance components for three methods over 300 simulation runs for displacement scenarios A and B. As expected, the expected value of the estimates of the random effect variance using non-misplaced data is very close to the true value of 5.0. Also, the distribution of the estimates is symmetrical around the true value (Figure 4.6(e-h)). The estimate of random effect variance based on the displaced location data using the naive method is biased downwards. The mean of the estimates over 300 simulations is well below the true value, and this difference increases as the intensity of the displacement error increases. For each displacement scenario, the proposed EDC-based estimate of the random effect variance has a lower bias than the naive estimates. For example, under displacement scenario A, the average random effect variances are 4.566 and 4.660 for the naive and EDC estimators. Thus, the EDC-based estimator can correct the bias of the random effect variance estimates to some extent and is preferred to the naive estimator. Finally, the mean of the estimates of the random effect variance applying the BC method is again close to the true value of 5.0 for each displacement scenario. Also, the distribution of the estimates of the random effect variance is moderately symmetrical around the parameter (true) value. Therefore, the BC method outperforms the EDC and naive methods and can correct the bias of the estimates under displacement very well.

The estimate of the error variance using the naive method is biased upwards for each displacement scenario. The BC method can produce an approximately unbiased estimate by correcting the bias. For example, the mean of the error variance estimates over 300 simulations using the naive, EDC and BC methods are 25.58, 25.44 and 25.04. This indicates that the mean of the estimates using the BC method is very close to the true value of 25.0. Thus, in terms of lower bias, the BC method outperforms both the EDC and naive methods.

We have also examined the variability of the estimates of the linear mixed model parameters using the naive, EDC, and BC methods (Table 4.2). For displacement scenarios A and B, the RMSEs of the fixed effect parameter estimates (regression intercept and slope) are about the same for each method. This is because the random displacement process does not affect the fixed effect parameters, as we observed earlier. On the other hand, regarding the estimates of variance components, the naive method has the largest RMSEs of the estimates. In contrast, the proposed EDC and

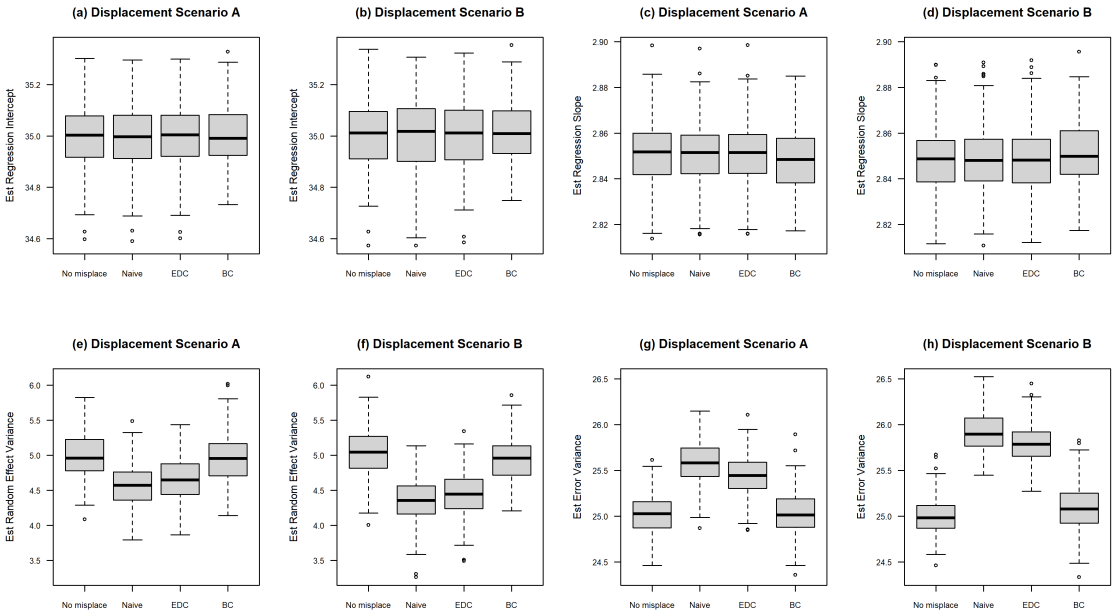

Figure 4.6: Plot of the distribution of the estimated parameters (a-b: regression intercept, c-d: regression slope, e-f: random effect variance, g-h: error variance) of the linear mixed model for each method over the 300 simulation runs for the displacement scenario A and B. For EDC based estimator, these parameter estimates are averaged over 300 pseudo-samples of true EA coordinates for each simulation.

BC methods have much smaller RMSEs of the estimates under both displacement scenarios. Furthermore, the BC method has a smaller RMSE of the estimates of the variance components than the EDC method. For example, under displacement scenario A, the RMSEs of the random effect variance parameter estimates are 0.543, 0.462, and 0.348 using the naive, EDC, and BC methods, respectively. Similarly, under displacement scenario B, the RMSEs of the estimates of the random effect variance parameter are 0.732, 0.645, and 0.317 using the naive, EDC, and BC methods, respectively. Therefore, the proposed methods, especially the BC method, outperform the naive method in terms of a lower RMSE for estimating the variance components of the linear mixed model under random displacement.

	Scenario A				Scenario B			
Parameter	No misplace	naive	EDC	BC	No misplace	naive	EDC	BC
β_0	0.126	0.127	0.126	0.114	0.127	0.131	0.128	0.116
β_1	0.014	0.014	0.014	0.014	0.015	0.015	0.015	0.014
σ_u^2	0.330	0.543	0.462	0.348	0.355	0.732	0.645	0.317
σ_ε^2	0.213	0.622	0.493	0.237	0.192	0.946	0.829	0.268

Table 4.2: RMSE of the linear mixed model parameter estimates over the 300 simulation runs for each estimator for scenarios A and B.

We report the results of the upazila mean estimates using the naive, EDC and BC methods and

evaluate the performance of the three methods in terms of bias and RMSE of the point estimates. In order to compare the methods with respect to the bias of the upazila mean estimates under the linear mixed model, the upazila true means are plotted against the expected estimates for each estimator under the displacement scenarios A and B in Figure 4.7. The EDC-based estimator under the DHS displacement process has a lower bias than the naive and BC estimators.

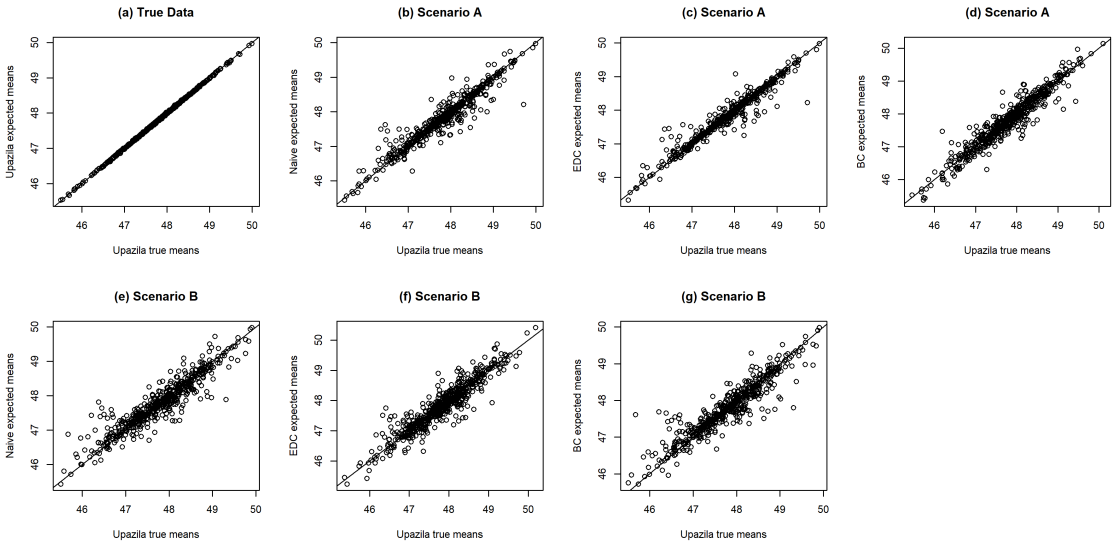

Figure 4.7: Plot of the upazila true means against the expected estimates for each estimator: under displacement scenario-A [b) naive, c) EDC and d) BC] and under displacement scenario-B [e) naive, f) EDC and g) BC].

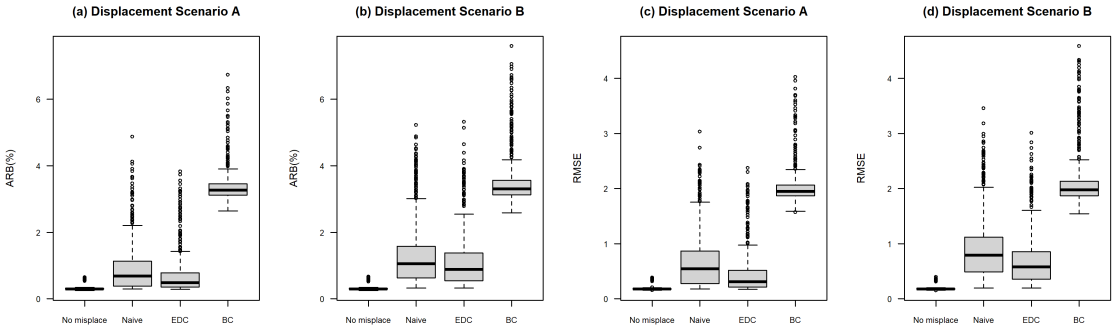

Figure 4.8: Plot of the distribution of the ARBs and the RMSEs of the upazila mean estimates for each method based on the linear mixed model over the 300 simulation runs for the displacement scenarios A and B.

Figure 4.8(a-b) shows the distribution of the Absolute Relative Biases (ARBs) of the upazila mean estimates for each method based on the linear mixed model over the 300 simulation runs for the displacement scenario A and B. We observe that the EDC estimator has lower bias than the naive and BC estimators. For example, under displacement Scenario A, the average ARBs are 0.9287, 0.7369 and 3.386 for the naive, EDC and BC estimators respectively. For scenario B, the average ARBs

are 1.3409, 1.1720 and 3.556 for the naive, EDC and BC estimators respectively. These results indicate that the proposed EDC method under the DHS displacement process gives a considerable advantage over the BC and the simple naive approach that ignores the DHS EA displacement error. The BC method performs worse than the naive method. Figure 4.9 represents ARBs of the estimated upazila means using the linear mixed model for each estimator under displacement scenarios A and B. Closer inspection of the results as shown in Figure 4.9 reveals that for 84% of upazilas, the ARB of the EDC estimator is lower than that of the naive estimator. However, we observe that specific upazilas, notably those between 160-190, show higher ARBs of the estimated means for each method, even with non-misplaced data. This elevated behaviour occurs because these upazilas have fewer households. This situation arises when only one EA is selected, yielding only 30 households from upazilas that are exclusively rural or urban. In contrast, upazilas comprising both rural and urban areas contain 60 households when two EAs are selected.

In order to compare the three methods in terms of variability, the distribution of the RMSEs of the estimated upazila means using the linear mixed model for each method over 300 simulations for two displacement Scenarios A and B is presented in Figure 4.8(c-d). For Scenario A, averaged over the 300 simulations, the mean RMSEs over the 544 upazilas for the naive, the EDC and the BC estimators are 0.6724, 0.4642 and 2.041 respectively. This shows that the EDC estimator has a lower average RMSE than the naive and BC methods. However, the BC method performs worse than the naive method. Also, to visually inspect the results over 544 upazilas, RMSEs of the naive, EDC and BC estimators based on the linear mixed model are presented in Figure 4.10. We observe that for 92% upazilas, the RMSE of the EDC estimator is lower than that of the naive estimator. Thus, the EDC based method outperforms the naive method in terms of lower variability. In the case of large-displacement error with scenario B, averaged over the 300 simulations, the mean RMSEs over the 544 upazilas for the naive, the EDC and the BC estimators are 0.9382, 0.7327 and 2.153 respectively. Therefore, for such an extreme displacement scenario, the EDC based estimator under the DHS displacement process outperforms the naive estimator. In contrast, the BC method performs worse than the naive and EDC methods.

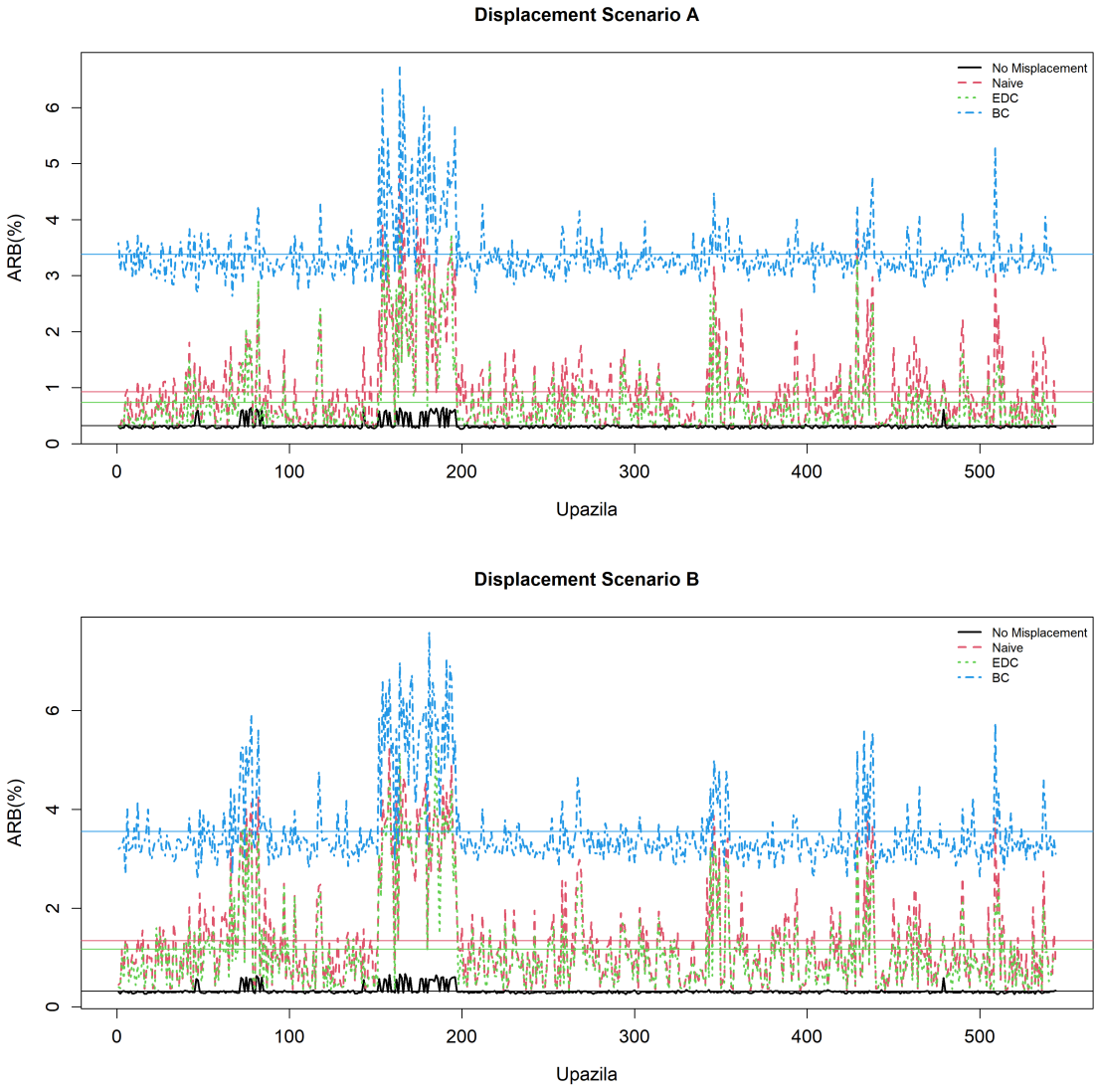

Figure 4.9: Plot of the ARBs of the estimated upazila means using the linear mixed model for each estimator under displacement scenarios A and B. Horizontal lines indicate the average ARB for each estimator.

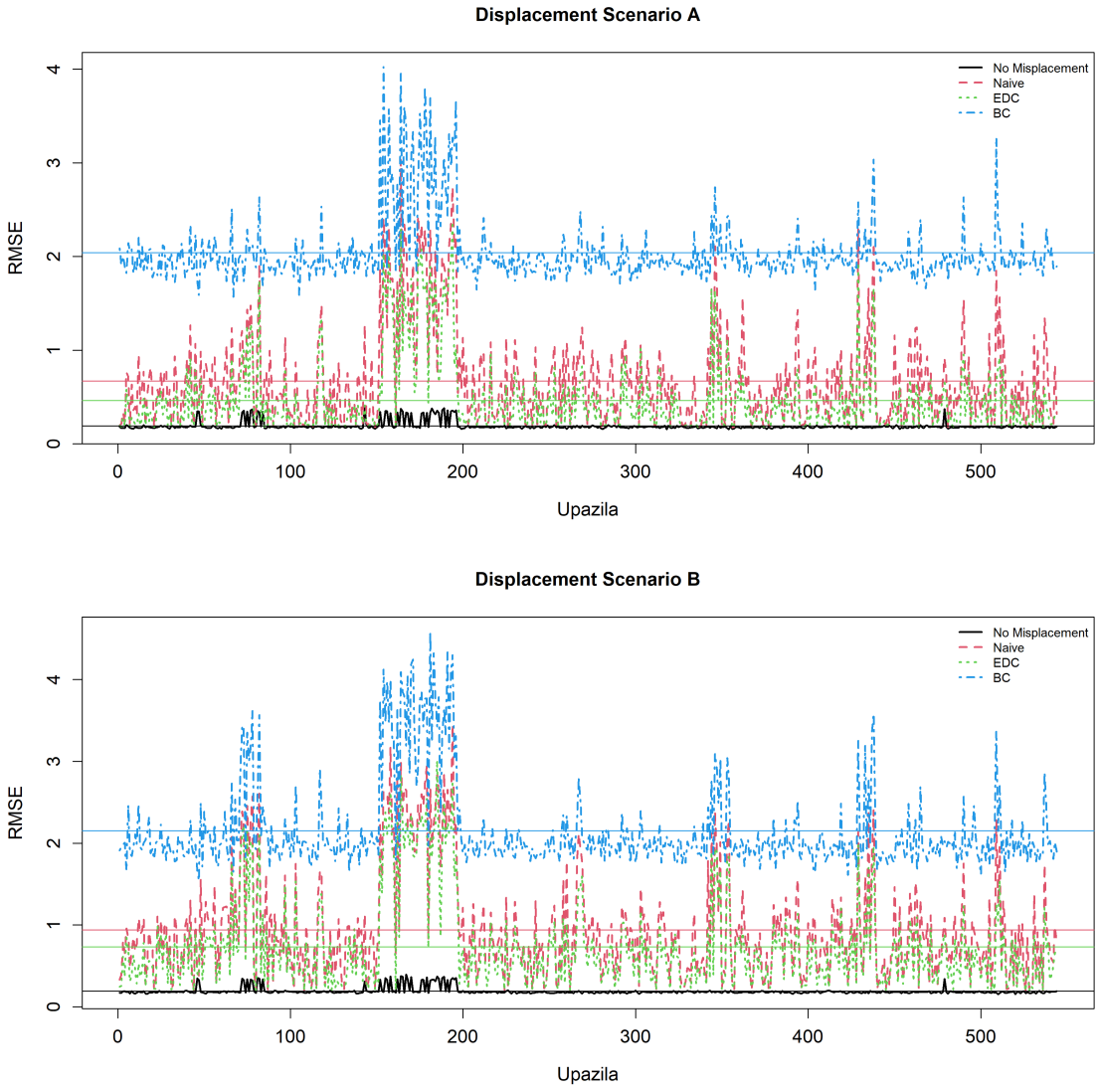

Figure 4.10: Plot of the RMSEs of the estimated upazila means using the linear mixed model for each estimator under displacement scenarios A and B. Horizontal lines indicate the average RMSE for each estimator.

4.5.3 Variance estimation of the model parameters estimates under the proposed methods

In this section, we present the simulation study results of the variance estimates for the linear mixed model parameter using the proposed variance estimation methods under the EDC and BC estimators, which we have developed in Sections 4.2.2 - 4.3.1. To obtain the estimated variance of the model parameter estimates using the EDC, we propose using the law of total variance formula conditional on the design matrix for the random effects under random displacement. Regarding estimating the variance of the model parameter estimates under the BC method, we use a model-based double bootstrapping technique under repeated random displacement.

We evaluate the effectiveness of the proposed variance estimation methods by comparing the average estimated variances and the empirical variance of the point estimates of the model parameters over 300 simulation runs. The empirical variance is obtained by calculating the variance of the point estimates of the model parameters over 300 simulation runs using the proposed EDC and BC methods. We also report the results of the estimated variance and the empirical variance of the point estimates obtained using the naive method, which ignores the misplacement error due to the random displacement process.

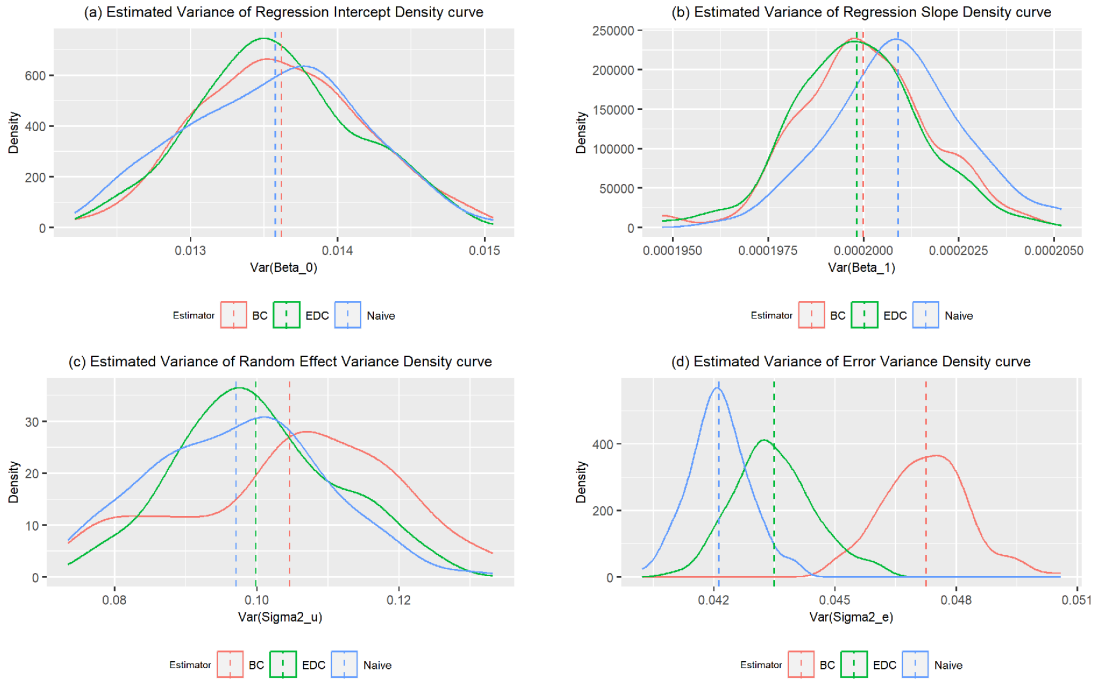

Figure 4.11: Density curves of the estimated variance of the parameter estimates (regression intercept (a), regression slope (b), random effect variance (c), and error variance (d)) of the LMM for the BC, EDC and naive estimators over 300 simulation runs. The vertical dotted line indicates the mean of the variance estimates for each estimator.

Figure 4.11 shows the density curves of the estimated variances of the point estimates of the LMM parameters for the BC, EDC and naive estimators over 300 simulations. The shape of the distribution of variance estimates using the proposed BC and EDC-based estimator under the random displacement process looks symmetric. Table 4.3 presents the average estimated variances and the empirical variance of the linear mixed model parameters estimates over 300 simulations for the BC, EDC and naive methods. The average estimated variances are reported for the regression intercept ($\hat{V}(\hat{\beta}_0)$), the slope ($\hat{V}(\hat{\beta}_1)$), the variance of the random intercept ($\hat{V}(\hat{\sigma}_u^2)$), and the variance of the error term ($\hat{V}(\hat{\sigma}_\varepsilon^2)$). The empirical variances are also given in brackets. Ratios of estimated to empirical variances are provided below each set of estimators.

Estimator	$\hat{V}(\hat{\beta}_0)$	$\hat{V}(\hat{\beta}_1)$	$\hat{V}(\hat{\sigma}_u^2)$	$\hat{V}(\hat{\sigma}_\varepsilon^2)$
BC	0.01362 (0.01403) 0.971	0.000200 (0.000185) 1.081	0.10462 (0.11018) 0.950	0.04726 (0.05120) 0.923
EDC	0.01358 (0.01604) 0.847	0.000200 (0.000196) 1.020	0.09985 (0.09864) 1.012	0.04349 (0.04879) 0.891
naive	0.01358 (0.01624) 0.836	0.000201 (0.000199) 1.010	0.09705 (0.10675) 0.909	0.04213 (0.05661) 0.744

Table 4.3: The average of the estimated variances and the empirical variance of the linear mixed model parameters estimates using the BC, EDC and naive methods over 300 simulation runs for the displacement scenario A. Empirical variances are given within the brackets. Ratios of estimated to empirical are provided below each set of estimators.

Regarding the variance estimation, the average estimated variances of the fixed effect parameters (regression intercept and slope) are similar for all three methods. However, for the variance components (random effect variance and error variance), which are affected by misplacement errors, the proposed BC and EDC methods have slightly larger estimated variances than the naive method. Furthermore, the BC method has slightly larger estimated variances than the EDC method for the variance components.

We also compare the average estimated variances with the empirical variances for each method. The results indicate that the average estimated variances of the BC and EDC methods align fairly closely with the empirical variances for linear mixed model parameter estimates, particularly for the fixed effects. This suggests a fairly accurate variance approximation by these methods. However, closer and more nuanced consideration, focusing on the ratio between estimated and empirical variances, reveals instances of underestimation by the proposed methods. For instance, the EDC method underestimates the variance of error variance parameter estimates, highlighting potential issues as it does not account for the bias of the estimator. On the other hand, the naive method, which ignores the misplacement errors, has larger empirical variances than the average estimated variances. The naive method has larger empirical variances for the parameter estimates than the

proposed methods.

Overall, the variance estimators proposed under the BC and EDC methods perform moderately well in estimating the variances of model parameter estimates under the random displacement process, albeit with some instances of underestimation. The approximation of variance is notably accurate for fixed effect parameters. However, the estimation of variance components, which are affected by misplacement errors due to random displacement, may not be as reliable as it fails to account for the bias of the estimator.

4.6 Conclusion

In this chapter, we address the issue that arises when random effects in multilevel models are specified at levels below the DHS displacement-restricted boundaries, such as admin 3. In particular, misplacement or mixing of observations among groups (for instance, admin 3 units) can lead to biased estimates of random variance components in multilevel models and associated model-based estimates of finite population parameters, e.g., group means. When random effects are specified at progressively lower administrative levels, the potential for misplaced EAs increases, leading to greater bias in estimates. This implies that a DHS user specifying random effects at the admin 4 level might obtain more biased estimates of random variance components in multilevel models than if specified at the admin 3 level. Therefore, in situations with high misplacement errors, failing to account for this error can result in misleading inferences, making our proposed methodologies even more invaluable. By applying our proposed methods, users can significantly mitigate the bias in estimates, providing more accurate results and ensuring the reliability of their conclusions.

In this chapter, we explore the potential methods for estimating the linear mixed model and obtaining the estimates of the model parameters and the model-based group means under displacement. In particular, we develop the framework of the linear mixed model under displacement. The proposed model is fitted by an algorithm under the EDC method, which is developed under the measurement error model by bringing additional information from different external sources. We also develop a methodology for estimating the global (model) parameters of the linear mixed model using a parametric bootstrap bias correction (BC) under displacement. The BC method is developed and expected to produce unbiased estimates of the model parameters satisfying the two crucial assumptions on the model form and the magnitude of the bias of the estimates under displacement. In addition to the point estimates, we develop the variance estimation of the model parameters under the proposed EDC and BC methods. The variance approximation is obtained by using the law

of total variance for the EDC-based estimator and model-based bootstrapping for the BC-based estimator under random displacement.

We compare the performance of the two proposed methods EDC and BC and the naive method that ignores the displacement error using simulated data. The simulated data are generated under the linear mixed model by bringing some characteristics of the 2011 BDHS EAs. The naive estimates of the variance component parameters and upazila means are biased under displacement. The proposed EDC method can correct the bias of the model parameters and upazila means to some extent and is preferred to the naive method. Also, the EDC estimator has a lower RMSE of the upazila mean estimates compared to the BC and the naive estimators. The BC method performs better than the EDC and the naive method by producing unbiased estimates of the linear mixed model parameters under displacement. However, it does not work for estimating upazila means and has a higher RMSE of the upazila estimates than the EDC and the naive methods. Thus, the BC method should be used for obtaining the estimates of the global parameters under displacement, while for the estimates of the finite population parameters, the EDC method is preferred. Finally, the proposed variance estimation methods under the EDC and BC estimators provide moderately accurate estimates of the variance for the linear mixed model parameters estimates under random displacement. In comparison, the naive method tends to underestimate the variance of the model parameter estimates ignoring the misplacement errors due to the random displacement process.

Chapter 5

Regression models with spatial covariates under random displacement

5.1 Introduction

Researchers are often interested in examining how spatial covariates, e.g., the distance from the Enumeration Area (EA) centroid to the nearest health facility of the country, affect the outcome of interest in regression models. For example, Feldacker et al. (2010) studied whether HIV infection in rural Malawi is related to the distance from the Demographic and Health Survey (DHS) EA centroid to major roads and health clinics. The DHS true EA coordinates are not available to the data analysts and are displaced using the DHS displacement algorithm to protect the respondents' confidentiality. Therefore, the statistical analyses with a spatial covariate that is constructed using the randomly displaced location can potentially be inaccurate, and the random displacement process induces a measurement error in covariate problem in regression models, e.g., a linear regression. The measurement error in covariate in a regression model leads to a biased estimate of its coefficient (Aigner, 1973; Carroll et al., 2006, section 3.2).

Various methods have been proposed in the literature to address the issue of measurement error in covariates. These include maximum likelihood methods (Rabe-Hesketh et al., 2003), Bayesian Markov chain Monte Carlo algorithms (Goldstein and Shlomo, 2020), and regression calibration (Hardin et al., 2003; Spiegelman et al., 1997). The regression calibration (RC) approach corrects the measurement error by replacing spatial covariate values, measured with errors, with their expected values given the displaced locations. This ensures the derivation of unbiased and consistent estimates (Carroll et al., 2006; Karra et al., 2020). We have described the RC approach in detail

in Section 2.7. The term "calibration" in this context refers to the adjustment of the erroneous measurements towards their expected true values. Given the displaced coordinates, spatial covariate values are available at the individual level but are measured with error. The RC approach attempts to obtain expected true values by correcting this error. The process might not be directly related to the ecological fallacy, wherein inferences about individuals are erroneously derived from grouped data (Wakefield and Shaddick, 2006). Previous studies (e.g., Warren et al., 2016a,b; Karra et al., 2020) that focused on estimates of spatial covariate effects accounting for displacement uncertainty have been discussed, presented mathematically and reviewed critically along with the limitations and further methodological developments in the literature review, Chapter 2 of this thesis. Following the literature review, in Chapter 3, we have developed the external data-based classification (EDC) method under a functional measurement error model by using additional information from external sources and assessed its performance over the candidate methods through a simulation study. In particular, the EDC method is used to draw likely true EA locations given the displaced locations by using the developed conditional distribution of displaced coordinates and the underlying marginal distribution of the true locations. As the distribution of the true locations is unknown, we have developed a flexible framework in Section 3.4 for approximating the distribution using information from multiple external sources that improve the selection probabilities for likely true locations. This is one of the key differences between the proposed EDC method and other competing methods in the literature, for example, the methods by Warren et al. (2016a) and Karra et al. (2020). We discussed this in detail in Chapters 2 and 3. In this chapter, we aim to show that the proposed EDC method can be used to obtain unbiased estimates of the effects of the spatial covariates under displacement. In particular, following the regression calibration under a numerical integration by Karra et al. (2020), we can use the proposed EDC method to calculate the expected value of the unknown true spatial covariate of interest given the displaced location data. More specifically, we can obtain this expected value using a numerical integration by repeatedly drawing pseudo-samples of the true location coordinates using the proposed EDC method under displacement given the displaced location coordinates. We refer to this as the external data-based classification under regression calibration (EDC-RC) method hereafter. In this chapter, we aim to describe the methodology for calculating the expected covariate values using the proposed EDC method in order to obtain unbiased estimates of the spatial covariate parameters of regression models such as the linear regression model.

Estimates of the parameters of spatial covariates in regression models are biased due to the random displacement process. Consequently, efforts can be made to estimate the bias estimates and

develop a bias-corrected estimator for the model parameters. Specifically, bootstrap bias correction methods for linear regression model parameters under displacement could be considered. To the best of our knowledge, no prior studies have attempted to use a bootstrap bias correction method to correct the bias of the estimates of spatial covariate parameters arising from the random displacement process. The bootstrap bias estimation technique estimates an estimator's bias by drawing multiple bootstrap samples from the representative population sample. For more details, refer to Efron and Tibshirani (1992), section 10.2. When addressing spatial covariate parameter estimates bias due to displacement, the bootstrap bias correction method should meet two key assumptions. The first assumption is that the functional form of the underlying model, such as the linear regression model, is appropriate for the data. The second, and perhaps more vital, assumption is that the magnitude of the bias of naive estimates of model parameters under displacement reflects the magnitude of the bias of estimates derived from bootstrap samples and repeated displacement processes applied to displaced locations. Bootstrap samples are generated from the model using parameter estimates derived from observed displaced data. When these assumptions are met, the bootstrap bias correction method can produce unbiased parameter estimates under displacement. This chapter introduces a non-parametric bootstrap bias correction for the spatial covariate parameter estimates in the linear regression model under the DHS displacement process, termed the Bootstrap Bias Correction (BC) method. Using a model-based simulation study, we evaluate the BC method's performance against the EDC-RC and naive methods. We expect the BC and EDC-RC methods to perform better than the naive method. Also, the BC method should be preferred to the EDC-RC method, as it does not require additional information from external sources. We explore this in this chapter.

The proposed methodological framework for the EDC-RC and BC methods can be used for various types of spatial covariates in regression models. For example, one type of covariate is the distance from the DHS EA centroid to the nearest health facility of the country. Hereafter, it is referred to as a distance covariate. Another type is when a covariate is available at the disaggregated level of the geography from a different survey. For example, the proportion of poor people at the sub-district level obtained from the Household Income Expenditure Survey (HIES) of the country. We refer to the later type of covariate as an area-level covariate. The DHS EA locations are used to link the area-level covariate values to the outcome of interest measured at the DHS EA points. In past studies, the effects of only distance type covariates on the outcome of interest under displacement were studied (for example, see Warren et al., 2016a; Karra et al., 2020). The effect of linking the area-level covariate to the DHS EA locations is yet to be explored. Therefore, in this chapter,

we consider both types of covariates to assess the performance of the proposed EDC-RC and BC methods to correct the bias of the estimates of the spatial covariate effect under displacement over the naive method that ignores the displacement error.

The chapter is organised as follows. Section 5.2 describes the theory and methods for estimating the effects of spatial covariates in the linear regression model under displacement. In Section 5.2.1, we discuss the proposed methodology under regression calibration for calculating the expected covariate values using the EDC method in order to obtain unbiased estimates of the spatial covariate parameters of the linear regression model. Section 5.2.2 develops a bootstrap bias correction method for correcting the bias of the estimates of the spatial covariate parameters of the linear regression model under the displacement process. In Section 5.3, we develop the variance estimation methods for the point estimates of the model parameters under the proposed EDC-RC and BC methods in the presence of random displacement errors. A model-based simulation study is conducted in Section 5.4 to assess the performance of the proposed EDC-RC and BC methods over the naive method that ignores the displacement error. We claim the EDC-RC and BC methods work well for correcting the bias of the spatial covariate parameters. In Section 5.5, we present and discuss the simulation study results. Finally, some concluding remarks are given in Section 5.6.

5.2 Theory and methods for estimating spatial covariates effects under random displacement

5.2.1 Development of the EDC-RC method under random displacement

In this section, we aim to describe the EDC under regression calibration (EDC-RC) method for estimating the spatial covariates (e.g., distance, environmental, area level variable etc) effects in regression models accounting for the observation locations displacement uncertainty. The method we propose here should be applicable for estimating the effects of different types of spatial covariates, e.g., distance and area-level covariate in regression models such as the linear regression model. We describe the theory in the context of the linear regression model.

Let T_k denote the true location coordinates of the kth $(k = 1, \cdots, n)$ unit. The unit of analysis can be the individual, household or Enumeration Area (EA). The true value T_k can be represented as $T_k = (T_{k1}, T_{k2})$, where T_{k1} and T_{k2} denote the exact geographical longitude and latitude coordinates of the kth unit. Let $X_{1k}(T_k)$ be the value of a spatial covariate which is constructed or measured based on the true location coordinates T_k. Also, we denote a non-spatial covariate value

by X_{2k}. We want to estimate a relationship described by the linear regression model formulated as follows:

$$y_k = \beta_0 + \beta_1 X_{1k}(T_k) + \beta_2 X_{2k} + \varepsilon_k, \tag{5.1}$$

where, y_k is the outcome, β_0 is the unknown intercept, β_1 is the unknown coefficient of the spatial covariate $X_{1k}(T_k)$, β_2 is the unknown coefficient of the non-spatial covariate X_{2k} and ε_k is a random-error term which is independent of $X_{1k}(T_k)$ and X_{2k}. For simplicity, we make the assumption that the random errors ε_k in (5.1) are independent and identically distributed with mean 0 and constant variance σ_ε^2. The unknown true parameters of the linear regression model (5.1) are $\theta = (\beta_0, \beta_1, \beta_2, \sigma_\varepsilon^2)$.

As the true coordinates T_k are unknown, the true spatial covariate values $X_{1k}(T_k)$ in (5.1) are also unknown to the data analysts. It is well known in the literature that if we simply replace T_k with the available displaced coordinates W_K, the estimates of the model parameters, especially β_1 in (5.1) will be biased (Warren et al., 2016a; Karra et al., 2020). In that case, following the regression calibration (RC) approach by Carroll et al. (2006), section 4.2, we can replace the unknown true spatial covariate values $X_{1k}(T_k)$ in (5.1) by its expected values given the displaced coordinates W_K in order to obtain unbiased and consistent estimates. In particular, we can calculate the expected value of the $X_{1k}(T_k)$ given W_k and defined as $\hat{X}_{1k}(T_k) = E\{X_{1k}(T_k) \mid W_k\}$. By replacing $X_{1k}(T_k)$ with the expected value $\hat{X}_{1k}(T_k)$, the linear regression model (5.1) can be expressed as

$$y_k = \beta_0 + \beta_1 \hat{X}_{1k}(T_k) + \beta_2 X_{2k} + \varepsilon'_k, \tag{5.2}$$

where ε'_k are the random errors, with mean 0 and uncorrelated with the covariates in model (5.2). Hence, the standard ordinary least squares method can be used to fit the model (5.2) and obtain the estimates (for details, see Karra et al., 2020).

We can calculate the expected value $E\{X_{1k}(T_k) \mid W_k\}$ by using the proposed external data-based classification (EDC) method under a functional measurement error model given the displaced coordinates. We have developed the measurement error model in Section 3.2. Under the measurement error model, the expected value $E\{X_{1k}(T_k) \mid W_k\}$ can be defined as

$$\hat{X}_{1k}(T_k) = E\{X_{1k}(T_k) \mid W_k\} = \int X_{1k}(T_k) f(T_k \mid W_k) dT_k, \tag{5.3}$$

where, the expected value is obtained by integrating over all possible true locations, $f(T_k \mid W_k)$ is the conditional distribution of the true locations T_k given the displaced locations W_k. We de-

veloped this conditional distribution $f(T_k \mid W_k)$ in Section 3.2.4. To recall, $f(T_k = T_k^\star|W_k) \propto f(W_k|T_k = T_k^\star)f(T_k = T_k^\star)$ where $T_k^\star$ is a likely true locations. Utilizing this formulation we can draw pseudo-samples (imputations) of T_k from $f(T_k = T_k^\star|W_k)$ using the proposed external data-based classification (EDC) method under displacement to reconstruct the spatial covariate values $X_{1k}(T_k^\star)$. The proposed EDC method, which we developed in Section 3.4 under a functional measurement error model by using additional information (e.g., population density, designated administrative boundary restriction, rural urban boundaries, boundary corrections for non-settlement places) from external sources. In particular, as the marginal distribution $f(T_k)$ is unknown, we have developed a flexible framework in Section 3.4 for approximating the distribution using information from multiple external sources that improve the selection probabilities for likely true locations. This is one of the key differences between the proposed EDC method and other competing methods in the literature, for example, the methods by Warren et al. (2016a) and Karra et al. (2020). We discussed this in detail in Chapters 2 and 3.

5.2.1.1 Estimation

It is not possible to obtain the expected values in (5.3) analytically. As the true location T_k is unobserved, we can use numerical integration for estimation. In particular, we draw pseudo-samples (imputations) of T_k from $f(T_k = T_k^\star|W_k)$ using the proposed EDC method given the displaced locations. This works by replacing the unobserved true point T_k in the target distribution by generating pseudo-samples of the unobserved true point given the observed data and then, obtaining spatial covariate values $X_{1k}(T_k^\star)$ with the updated (new) location samples $T_k^\star$. The process is repeated a large number of times and we take the average of the estimates. The computational steps for obtaining the estimates of the expected values $E\{X_{1k}(T_k) \mid W_k\}$ under the proposed EDC-RC method are given as follows:

- Step 1: For each unit $k = 1, \cdots, n$, generate a pseudo-sample $\{T_k^\star; k = 1, \cdots, n\}$ of the unknown true location coordinates T_k from the conditional distribution $f(T_k = T_k^\star|W_k)$ using the proposed EDC method under displacement described in Section 3.4.
- Step 2: Obtain the spatial covariate values $X_{1k}(T_k^\star)$ using the pseudo-samples $\{T_k^\star; k = 1, \cdots, n\}$ of the true location coordinates T_k from Step 1.
- Step 3: Iterate Steps 1-2 $B_1 + B_2$ times with $B_1 = 300$ (large number of iterations) and $B_2 = 20$ (additional iterations).
- Step 4: Compute the estimates of the expected values $\widehat{X}_{1k}(T_k) = E\{X_{1k}(T_k) \mid W_k\}$ by

averaging over the derived B_1 and $B_1 + B_2$ estimates, respectively. Plot these estimated expected values to visually assess their closeness. If they appear reasonably close on the plot, it suggests convergence of the algorithm. Otherwise, return to Step 1 and proceed through the subsequent steps, incrementing B_1 by a predetermined amount (for instance, $B_1 = B_1 + 20$), while keeping the constant value $B_2 = 20$, until the algorithm meets the convergence criteria.

If we now replace $X_{1k}(T_k)$ in (5.1) with the expected values $\hat{X}_{1k}(T_k) = E\{X_{1k}(T_k) \mid W_k\}$, we can fit the model (5.1) using a standard method, for example, the least squares method and obtain the estimates of the model parameters, denoted by $\hat{\theta}_{\text{RC}}$.

5.2.2 Development of a bootstrap bias correction method under random displacement

The true location coordinates T_k are unknown to data analysts; only the displaced coordinates W_k are available. Thus, in the linear regression model (5.1), instead of $X_{1k}(T_k)$, we directly use the spatial covariate values $X_{1k}(W_k)$ based on W_k. This can be expressed as:

$$y_k = \beta_0 + \beta_1 X_{1k}(W_k) + \beta_2 X_{2k} + \varepsilon_k. \tag{5.4}$$

When we fit the model (5.4) to the displaced dataset $\{y_k, X_{1k}(W_k), X_{2k}\}$ and use the least squares method, we obtain the estimated parameters $\hat{\theta} = \left(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}^2_\varepsilon\right)$. However, these naive estimates of the linear regression model parameters, particularly the spatial covariate parameter, are biased due to the random displacement of observation locations. Consequently, we aim to obtain the bias estimates under displacement and develop a bias-corrected estimator for the parameters. Specifically, a bootstrapping technique can be used to estimate this bias. The bootstrap bias estimation method estimates the bias of an estimator using bootstrap samples derived from repeated sampling and displacement processes on the displaced location data. For further details on bootstrap bias estimation, refer to Efron and Tibshirani (1992), section 10.2.

Concerning the bias of the linear regression model parameter estimates due to displacement, the bootstrap bias correction method must satisfy two critical assumptions. Firstly, we assume that the functional form of the linear regression model (5.1) is well-suited to the data. Secondly, we assume that the bias magnitude of the naive estimates of the model parameters under displacement is consistent with the bias observed in estimates from bootstrap samples with repeated displacement processes applied to the displaced locations. We create bootstrap samples based on model (5.4)

using the naive estimates, $\hat{\theta}$. Given these two assumptions are met, the bootstrap bias correction (BC) method can produce unbiased estimates of the model parameters despite displacement. We assert that the BC method effectively corrects the bias in model parameter estimates. Below, we outline the proposed bootstrap bias correction (BC) method designed to correct the bias of the linear regression model parameter estimates caused by displacement:

The bias of the naive estimates ($\hat{\theta}$) of the model parameters (θ) is given by $\text{Bias}\left(\hat{\theta}, \theta\right) = E\left(\hat{\theta}\right) - \theta$. This bias is unknown to us, as we do not know the true model parameters (θ). We obtain the estimates of the bias by using the $B\,(b = 1, 2, \cdots, B)$ model-based bootstrap samples and the displaced set of observation coordinates which are obtained by applying the displacement process to the displaced (observed) observation coordinates. These model-based bootstrap samples are generated by adding bootstrapped residuals to the predicted values of y_k, which are obtained using naive estimates $\hat{\theta}$, and can be expressed as follows:

$$y_k^{*b} = \hat{\beta}_0 + \hat{\beta}_1 X_{1k}(W_k) + \hat{\beta}_2 X_{2k} + \varepsilon_k^{*b}, \tag{5.5}$$

where ε_k^{*b} is the bootstrapped random error. These bootstrapped errors are obtained by sampling with replacement from the estimated random errors of model (5.4). If the estimated random errors are influenced by displacement, adjustments are made by rescaling and centering to ensure that the re-sampled errors accurately reflect the underlying variation of the original sample. Unlike the linear mixed model in Chapter 4, where the use of a parametric assumption (e.g., normal distribution) for the error terms was informed by the nature of the model, in the linear regression model (5.1), we do not explicitly assume a parametric, such as normal distribution of the error term. Consequently, we use a non-parametric bootstrap approach, which allows for greater flexibility. Let the W_k^{*b} denote the B displaced set of observation coordinates which are obtained by displacing the given displaced coordinates W_k using the DHS random displacement algorithm. Next, we obtain the bootstrap estimates of the parameters denoted by $\hat{\theta}^{*b}; b = 1, 2, \cdots, B$ using the bootstrap samples data set $\{y_k^{*b}, X_{1k}(W_k^{*b}), X_{2k}\}$. Then, the bootstrap estimate of the bias is obtained by calculating

$$\widehat{\text{Bias}}\left(\hat{\theta}, \theta\right) = \sum_{b=1}^{B} \hat{\theta}^{*b}/B - \hat{\theta}. \tag{5.6}$$

Finally, the bootstrap bias corrected estimator is defined as

$$\hat{\theta}_{\text{cor}} = \hat{\theta} - \widehat{\text{Bias}}\left(\hat{\theta}, \theta\right). \tag{5.7}$$

5.3 Variance estimation of the point estimates of model parameters under random displacement

5.3.1 Variance estimation under the EDC-RC method

We want to obtain the estimated variance of the linear regression model parameter estimates, $\hat{\theta}_{\text{RC}}$ under the proposed EDC-RC method. In this regard, we can use a bootstrap resampling technique. In particular, we take $B \quad (b = 1, 2, \cdots, B)$ model-based bootstrap samples under repeated displacement to account for displacement error from the original sample. For each bootstrap sample, we obtain the estimate of the model parameters by using the proposed EDC-RC method. The computational details for obtaining the variance estimates of the EDC-RC estimator under a bootstrapping are given below:

Step 1: Draw $B \quad (b = 1, 2, \cdots, B)$ bootstrap samples under the model (5.5) using naive estimates $\hat{\theta} = \left(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}^2_\varepsilon\right)$, which are obtained by fitting the model (5.4) to the displaced data $\{y_k, X_{1k}(W_k), X_{2k}\}$. Then, generate B displaced sets of observation coordinates, denoted as $W^{*b}_k; b = 1, 2, \cdots, B$, by displacing the given displaced coordinates W_k using the random displacement algorithm. Finally, the B bootstrap samples data are expressed as $\{y^{*b}_k, X_{1k}(W^{*b}_k), X_{2k}\}$.

Step 2: For each $b = 1, 2, \cdots, B$ bootstrap sample data $\{y^{*b}_k, X_{1k}(W^{*b}_k), X_{2k}\}$ from Step 1, we obtain the estimated expected spatial covariate values conditional on W^{*b}_k applying the proposed EDC-RC method under the estimation algorithm in Section 5.2.1.1. The estimated expected spatial covariate values are denoted by $\hat{X}_{1k}(T^{*b}_k)$. Using data $\left\{y^{*b}_k, \hat{X}_{1k}(T^{*b}_k), X_{2k}\right\}$ we obtain the estimates of the model parameters denoted by $\hat{\theta}^{*b}_{\text{RC}}; b = 1, 2, \cdots, B$.

Step 3: Finally, using the $B \quad (b = 1, 2, \cdots, B)$ estimates of the model parameters in Step 1, we calculate the estimated variance of the model parameter estimates under the EDC-RC method as follows:

$$\hat{V}(\hat{\theta}_{\text{RC}}) = \sum_{b=1}^{B} \left(\hat{\theta}^{*b}_{\text{RC}} - \bar{\hat{\theta}}^{*}_{\text{RC}}\right)^2 / B, \tag{5.8}$$

where $\bar{\hat{\theta}}^{*}_{\text{RC}}$ is the average of the estimates $\hat{\theta}^{*b}_{\text{RC}}$ over B bootstrap samples. Also, the estimated standard error of the model parameters estimates under the EDC-RC method is given by

$$\hat{SE}(\hat{\theta}_{\text{RC}}) = \sqrt{\sum_{b=1}^{B} \left(\hat{\theta}^{*b}_{\text{RC}} - \bar{\hat{\theta}}^{*}_{\text{RC}}\right)^2 / B}. \tag{5.9}$$

5.3.2 Variance estimation under the BC method

In this section, we aim to calculate the estimated variance of the bootstrap bias-corrected estimates in equation (5.7) for the parameters of the linear regression model, while taking into account the random displacement error. The variance estimation of the bias-corrected (BC) estimator ($\hat{\theta}_{\text{cor}}$) is similar to the one discussed in Section 4.3.1, where we described the variance estimation of the BC estimator for the parameters of the linear mixed model. To recall that the variance of the BC estimator ($\hat{\theta}_{\text{cor}}$) can be expressed using the variance and covariances of the naive estimator ($\hat{\theta}$) and the bootstrap bias estimator, $\widehat{\text{Bias}}\left(\hat{\theta}, \theta\right)$ in (5.6) as $V(\hat{\theta}_{\text{cor}}) = V(\hat{\theta}) + V\left(\widehat{\text{Bias}}\left(\hat{\theta}, \theta\right)\right) - 2 \times \text{Cov}\left(\hat{\theta}, \widehat{\text{Bias}}\left(\hat{\theta}, \theta\right)\right)$.

As mentioned in Section 4.3.1, obtaining the covariance between the naive and the bootstrap bias estimator using a single level of bootstrap samples is challenging. To address this, a double bootstrapping approach can be employed, involving two rounds of bootstrap sampling. In the first round, we generate B model-based bootstrap samples under repeated displacement, as described in Section 5.2.2, and in the second round, C model-based bootstrap samples are obtained from each of the B samples from the first round. We then compute a bias-corrected estimate based on these C bootstrap samples. The detailed computational steps for obtaining variance estimates of the BC estimator using double bootstrapping are provided below:

In the first round, generate B bootstrap samples (where $b = 1, 2, \cdots, B$) under the model specified in equation (5.5) using the naive estimates $\hat{\theta} = \left(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}_\varepsilon^2\right)$, obtained by fitting the model described in equation (5.4) to the displaced data $\{y_k, X_{1k}(W_k), X_{2k}\}$.

To create each bootstrap sample, apply the displacement algorithm to displace the given coordinates W_k, resulting in a set of displaced observation coordinates denoted as W_k^{*b}. The data for the first round of bootstrap samples, representing B samples, can be expressed as $\{y_k^{*b}, X_{1k}(W_k^{*b}), X_{2k}\}$.

In the second round, draw C bootstrap samples (where $c = 1, 2, \cdots, C$) from each of the B bootstrap samples obtained in the first round. Subsequently, compute the bias-corrected estimates for the model parameters based on these C bootstrap samples. The procedure is described in detail below:

Step 1: For each $b = 1, 2, \cdots, B$, using the first round bootstrap samples data $\{y_k^{*b}, X_{1k}(W_k^{*b}), X_{2k}\}$, we obtain estimates of the model parameters denoted by $\hat{\theta}^{*b} = \left(\hat{\beta}_0^{*b}, \hat{\beta}_1^{*b}, \hat{\beta}_2^{*b}, \hat{\sigma}_\varepsilon^{2*b}\right)$ and the estimated random errors of the model denoted by $\hat{\varepsilon}_k^{*'b}$.

Step 2: We generate C bootstrap samples using estimates $\hat{\theta}^{*b}$ under the following model:

$$y_k^{*c} = \hat{\beta}_0^{*b} + \hat{\beta}_1^{*b} X_{1k}(W_k^{*b}) + \hat{\beta}_2^{*b} X_{2k} + \varepsilon_k^{*c}, \tag{5.10}$$

where ε_k^{*c} are the bootstrapped random errors which are obtained by sampling with replacement from the estimated random errors $\hat{\varepsilon}_k^{*'b}$. These bootstrapped random errors are added to the predicted outcomes to generate the bootstrap samples, as indicated in (5.10).

Step 3: We create C displaced set of observation coordinates, denoted as $W_k^{*c}; c = 1, 2, \cdots, C$, by displacing the given coordinates W_k^{*b} using the random displacement mechanism. Therefore, the second round C bootstrap samples data are represented as $\{y_k^{*c}, X_{1k}(W_k^{*c}), X_{2k}\}$.

Step 4: We obtain the estimates of the model parameters denoted by $\hat{\theta}^{*c}$ using the second round bootstrap sample data from Step 3. We use these second round bootstrap sample-based model parameter estimates to obtain the estimate of the bias of $\hat{\theta}^{*b}$ defined as

$$\widehat{\text{Bias}}\left(\hat{\theta}^{*b}, \theta\right) = \sum_{c=1}^{C} \hat{\theta}^{*c}/C - \hat{\theta}^{*b}. \tag{5.11}$$

Step 5: For each $b = 1, 2, \cdots, B$, we compute the bootstrap bias corrected estimates as

$$\hat{\theta}_{\text{cor}}^{*b} = \hat{\theta}^{*b} - \widehat{\text{Bias}}\left(\hat{\theta}^{*b}, \theta\right). \tag{5.12}$$

Step 6: We calculate the variance of the model parameter estimates under the bias-corrected (BC) method using the B bias-corrected estimates obtained in Step 5. The calculation is performed as follows:

$$\hat{V}(\hat{\theta}_{\text{cor}}) = \sum_{b=1}^{B} \left(\hat{\theta}_{\text{cor}}^{*b} - \bar{\hat{\theta}}_{\text{cor}}^{*}\right)^2 /B, \tag{5.13}$$

where $\bar{\hat{\theta}}_{\text{cor}}^{*}$ is the average of the estimates $\hat{\theta}_{\text{cor}}^{*b}$ over B first round bootstrap samples. Also, the estimated standard error of the model parameters estimates under the BC method is given by

$$\hat{SE}(\hat{\theta}_{\text{cor}}) = \sqrt{\sum_{b=1}^{B} \left(\hat{\theta}_{\text{cor}}^{*b} - \bar{\hat{\theta}}_{\text{cor}}^{*}\right)^2 /B}. \tag{5.14}$$

5.4 Model-based simulation

The model-based simulation enables a controlled empirical evaluation of the proposed EDC-RC and BC methods for estimating the spatial covariate parameters in the linear regression model under displacement. In particular, this section aims to assess the ability of the proposed EDC-RC and BC methods to correcting the bias of the spatial covariate parameters under displacement and hence, to provide unbiased estimates of the model parameters compared to the naive estimates that ignore the displacement error. We evaluate the EDC-RC, BC and naive methods using simulated data generated under the linear regression model, as described in the following paragraph. The naive method that ignores the displacement error produces biased estimates of the spatial covariate parameters. The proposed EDC-RC method is developed under the functional measurement error model by bringing additional information from different external sources to approximate the marginal distribution of the true locations. We also developed the BC method to correct the bias in the estimates of the model parameters, subject to meeting assumptions regarding the model form and the magnitude of the bias due to displacement. Therefore, we aim to show that the proposed EDC-RC and BC methods perform better than the naive method by obtaining unbiased estimates of the linear regression model parameters.

We generate the model-based simulated data by incorporating some characteristics from the 2011 Bangladesh Demographic and Health Survey (BDHS) EAs and using the geography of Bangladesh with 544 upazilas (admin 3). The true coordinates of the 2011 BDHS EAs and also the survey's sampling frame, the list of the 2011 census EAs with their number of households are not available. Therefore, we have generated the true coordinates for 274,150 EAs with their number of households using the WorldPop $100m \times 100m$ gridded population of Bangladesh for 2020 following the same principles of the 2011 BDHS EAs selected from the 2011 census EAs. The constrained population raster in Bangladesh for 2020 can be downloaded from WorldPop (Bondarenko et al., 2020). Finally, following the 2011 BDHS EA selection process, we have drawn a sample of 1050 EAs with probability proportional to the EA population stratified by the upazila and by the rural urban areas within the upazila. We have discussed this in Section 4.4.

In this simulation design, the 1050 selected EAs coordinates are considered as the true EA coordinates, and they are fixed throughout the simulation. We consider 30 households for each EA, and the households within an EA have the same coordinates as the EA centroid coordinate in the 2011 BDHS, i.e., households are georeferenced at the centre of the EA. Note that in this simulation study, we ignore the aggregation of households issue. Therefore, the total number of households is

$n = 31,500$ which is spread over $m = 544$ upazilas of Bangladesh. The upazila-specific number of households ranges from 30 to 60. We generate outcomes for each household under the linear regression model (5.1) with only spatial covariates for simplicity and defined as:

$$y_k = \beta_0 + \beta_1 X_{1k}(T_k) + \varepsilon_k. \tag{5.15}$$

We consider a spatial covariate, e.g., the nearest distance between the true location and the health facilities, and separately, a covariate available at the area (upazila) level in Bangladesh. For simplicity in this empirical study, we consider only one type of spatial covariate and exclude other non-spatial covariates in (5.15). The calculation of nearest distance and the linkage of the upazila level covariate to the household based on the displaced location coordinates may not be accurate that leads the biased estimates of the model parameters. In order to assess the impact of the displacement, we consider the linear regression model with two types of spatial covariates, distance and area-level, as described below:

Let $X_{1k.d}(T_k)$ denote the true distance between the kth household location coordinates and the nearest health facility. For this simulation study, as health facility location data, the location coordinates of family welfare centers and hospitals of Bangladesh is used. The source of the spatial data is the Local Government Engineering Department (LGED) of Bangladesh (Rahman and Szabó, 2020). First, we calculate the distance for each household from the nearest health facility $X_{1k.d}(T_k)$. Next, we generate an outcome variable y_k based on the equation (5.15), by the following relationship:

$$y_k = -0.75 - 0.10 X_{1k.d}(T_k) + \varepsilon_k, \tag{5.16}$$

where $\varepsilon_k \sim N(0, 0.15)$ is a randomly generated error term. We generate the outcome data by setting the true values of the parameters to $\beta_0 = -0.75$ and $\beta_1 = -0.10$ using the 2011 BDHS data, as detailed below. We calculate the distance to the nearest health facility for each DHS cluster. We then fit a linear regression model using the z-score (which represents the level of underweight in children from the 2011 BDHS) as the outcome, with distance as a covariate, to estimate the model parameters. These estimates serve as the true parameter values for this empirical study. The z-score is a standardized measure representing the weight-for-age (WAZ) of a child, which provides an assessment of a child's nutritional status relative to a reference population (WHO, 1986). This results in the coefficient of determination $R^2 = 0.55$ for the regression of y_k on $X_{1k.d}(T_k)$.

Regarding the linked (area-level) covariate, let, $X_{1k.a}(T_k)$ indicate the covariate values available

at the lower administrative level e.g., upazila level of Bangladesh. In this simulation study, as an area covariate, we use the estimated upazila proportions of poor people which have been produced using the ELL poverty mapping method by combining 2011 census and 2010 HIES data of Bangladesh. Then, we assign the upazila proportions to households by using the true location coordinates T_k and obtain a covariate value for each household $X_{1k.a}(T_k)$. Finally, we generate an outcome variable y_k by the following relationship:

$$y_k = -0.65 - 0.50X_{1k.a}(T_k) + \varepsilon_k, \tag{5.17}$$

where $\varepsilon_k \sim N(0, 0.10)$ is a randomly drawn error term. To generate the simulated outcome data from (5.17), we set the true values of the parameters to $\beta_0 = -0.65$ and $\beta_1 = -0.50$, which we get from fitting a regression model using the 2011 BDHS data, as discussed below. We assign upazila proportions to households based on the 2011 BDHS displaced coordinates, thereby obtaining a covariate value for each household. Next, we fit a linear regression model of z-scores (a measure of underweight in children, as previously defined) on upazila proportions associated with household children. The parameter estimates from this model act as the 'true' parameter values for this empirical study. The coefficient of determination for the regression of y_k on $X_{1k.a}(T_k)$ is $R^2 = 0.11$.

To obtain the displaced EA coordinates, the true EA coordinates are randomly displaced in line with the DHS displacement parameters: urban locations are displaced between 0-2 kilometres, while rural locations are displaced between 0-5 kilometres. Additionally, 1% of rural locations are randomly displaced by a range of 0-10 kilometres.

We calculate the misplacement statistics using a single set of displaced coordinates by creating buffers with maximum displacement distance around the displaced coordinates of the EA. About 42% (442 out of 1050) of the EAs are potentially misplaced, meaning their EA displacement buffers intersect with an upazila boundary. Additionally, of the potentially misplaced EAs, approximately 29% (128 out of 442) are truly misplaced. In this thesis, a truly misplaced EA refers to an EA where the displaced upazila ID differs from the actual upazila ID for that EA. In this simulation study, we know the displaced upazila ID and the actual upazila ID for each EA using the displaced and true EA coordinates. Area-level covariate assignment to EA/households will be wrong when the EA is misplaced due to displacement. Also, for the distance covariate, the mean distance from the EA centroid to the nearest health facility, based on true and displaced coordinates, is 5.42km and 5.65km, respectively.

The simulation steps (generation of the outcome variable under the model, parameter estimates and displacement process) are independently repeated 300 times. Under the repeated displacement, we compare the performance of the naive, the EDC-RC and the BC methods by computing the Root-Mean-Square-Error (RMSE) and the Bias for the estimates of the linear regression model parameters over $L = 300$ simulations as follows:

$$\text{RMSE}(\hat{\theta}) = \sqrt{\frac{\sum_{l=1}^{L}(\hat{\theta}_l - \theta)^2}{L}} \tag{5.18}$$

$$\text{Bias}(\hat{\theta}) = \frac{\sum_{l=1}^{L}(\hat{\theta}_l - \theta)}{L} \tag{5.19}$$

where $\hat{\theta}_l$ are the estimates of the linear regression model parameters using any of the methods aforementioned in simulation round l and θ defines the true value of the parameter.

To implement the model-based simulation, we use **R** (R Core Team, 2022). We fit the linear regression model to the data and obtained the parameters estimates using the function **lm**.

5.5 Results and discussion

In this section, we present and discuss the model-based simulation study results. We first discuss the results using the distance-based covariate. Then, the upazila-level linked covariate based results are presented and discussed. Section 5.5.1 reports and discusses the results on the performance of the proposed EDC-RC and BC methods for estimating the linear regression model parameters under random displacement. Also, the results of the estimated standard error of the linear regression model parameter estimates using the proposed estimators under the EDC-RC and BC methods are reported and discussed in Section 5.5.2.

5.5.1 Evaluation of the performance of the proposed methods for point estimates

In the case of the distance covariate based linear regression model, we run 300 simulations to obtain the empirical distributions for the estimates of the linear regression model parameters using the naive, EDC-RC and BC methods. The summary statistics of the estimates of the parameters of the linear regression model over 300 simulation runs are presented in Table 5.1. Figure 5.1(a) shows the empirical distribution of the estimates of the distance covariate parameter for each method over 300 simulation runs. As expected, the mean estimates of the distance covariate parameter over 300 simulation runs using true data (non-displaced) is approximately equal to the true value of the parameter $\beta_1 = -0.10$. The range of the estimates is small and is moderately symmetric around the parameter (true) value. The naive estimate of β_1 using displaced location data is biased. The proposed EDC-RC and BC methods can correct the bias very well and produce an approximately unbiased estimate of β_1. For example, the mean estimates of β_1 over 300 simulation runs using the proposed EDC-RC and BC methods are closer to the true value than the estimate obtained using the naive method. Both EDC-RC and BC methods performed similarly and outperformed the naive method in terms of reducing the bias in the estimates under random displacement.

The estimates of β_1 using the non-displaced data gives the smallest RMSE whereas the proposed BC and EDC-RC methods have a somewhat larger RMSE of the estimates of the distance covariate parameter. Also, the EDC-RC method has a slightly smaller RMSE of the estimates than the BC method. Furthermore, estimates with the naive method that ignores the displacement error have a much larger RMSE compared to the BC and EDC-RC methods (Table 5.1). Therefore, the proposed BC and EDC-RC methods perform better than the naive method in terms of a lower bias and RMSE for estimating the distance covariate parameter under random displacement.

Regarding the linked upazila-level covariate parameter estimates of the linear regression model, the

Parameter	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	RMSE
$\hat{\beta}_{0T}$	-0.7582	-0.7524	-0.7503	-0.7502	-0.7479	-0.7409	0.0033
$\hat{\beta}_{0W}$	-0.7806	-0.7690	-0.7634	-0.7641	-0.7592	-0.7463	0.0156
$\hat{\beta}_{0BC}$	-0.7726	-0.7567	-0.7499	-0.7502	-0.7437	-0.7292	0.0086
$\hat{\beta}_{0EDC-RC}$	-0.7645	-0.7561	-0.7537	-0.7533	-0.7507	-0.7403	0.0055
$\hat{\beta}_{1T}$	-0.1011	-0.1003	-0.1001	-0.1000	-0.0996	-0.0984	0.0005
$\hat{\beta}_{1W}$	-0.0990	-0.0971	-0.0963	-0.0964	-0.0957	-0.0939	0.0038
$\hat{\beta}_{1BC}$	-0.1031	-0.1009	-0.0999	-0.0999	-0.0990	-0.0969	0.0012
$\hat{\beta}_{1EDC-RC}$	-0.1022	-0.1003	-0.0997	-0.0998	-0.0993	-0.0982	0.0007
$\hat{\sigma}^2_{\varepsilon\ T}$	0.1475	0.1493	0.1499	0.1500	0.1507	0.1533	0.0011
$\hat{\sigma}^2_{\varepsilon\ W}$	0.1628	0.1659	0.1670	0.1670	0.1680	0.1709	0.0170
$\hat{\sigma}^2_{\varepsilon\ BC}$	0.1475	0.1507	0.1518	0.1518	0.1529	0.1560	0.0025
$\hat{\sigma}^2_{\varepsilon\ EDC-RC}$	0.1560	0.1599	0.1610	0.1609	0.1620	0.1637	0.0110

Table 5.1: Summary statistics for the distribution of the linear regression model parameter estimates with a distance covariate for each method over the 300 simulation runs. T and W refer to estimates based on non-displaced and displaced data, while BC and EDC-RC indicate the estimates based on the two proposed methods.

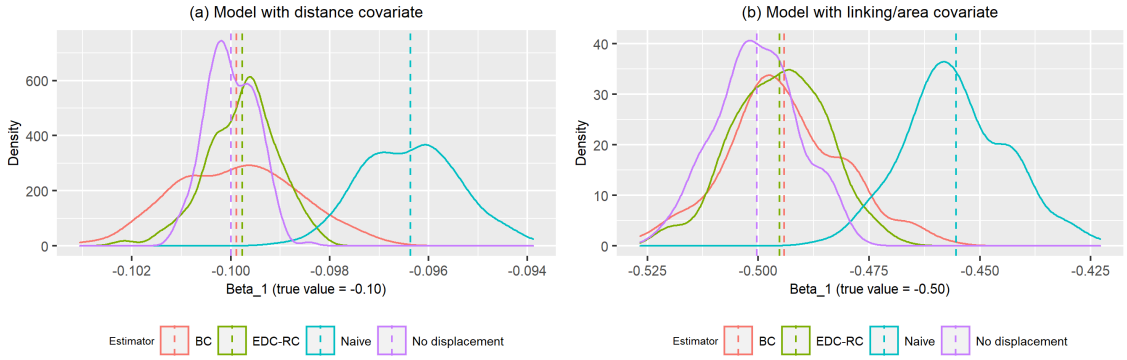

Figure 5.1: Density curves of the estimated (a) distance covariate and (b) area/linked covariate parameters of the linear regression model for each estimator over the 300 simulation runs. The vertical dotted lines indicate the mean estimates for each estimator. The true values of the parameters for distance and area covariates are -0.10 and -0.50 respectively.

summary statistics for the distribution of the model (5.17) parameter estimates for each method over the 300 simulation runs with a linked upazila-level covariate are given in Table 5.2. Figure 5.1(b) shows the density plot of the estimates of the linked covariate parameter over 300 simulation runs for the naive, EDC-RC and BC methods. The mean estimates of β_1 using true (non-displaced) data over 300 simulation runs is very close to the true value of $\beta_1 = -0.50$, as we expected. Also, the estimates using true data have a small range and are symmetric around the true value of $\beta_1 = -0.50$. The estimate of the linked covariate parameter using the naive method that ignores the displacement error is biased. For example, the mean naive estimates over 300 simulation runs is -0.4554 whereas the true value is $\beta_1 = -0.50$. The proposed biased corrected BC and EDC-RC methods correct the bias of the estimates very well and obtain an approximately unbiased estimate

of the linked covariate parameter under displacement. For instance, the mean estimates of β_1 over 300 simulation runs using the BC and EDC-RC methods are -0.4942 and -0.4951 respectively, which are very close to the true parameter value of $\beta_1 = -0.50$. Therefore, the proposed bias corrected methods work very well to correct the bias and outperform the naive methods. The last column of Table 5.2 presents the RMSE of the estimates of the linear regression model parameters for each of the methods. As expected, the estimates of the linked upazila covariate parameter using true (non-displaced) data have the smallest average root mean squared error (RMSE), whereas the naive estimates that ignore the displacement error have the highest RMSE. The proposed BC and EDC-RC estimators have much lower RMSE than the naive estimator. For example, the average RMSE of the estimates of the linked covariate parameter is about 3.2 times lower when using the proposed BC method than the naive method. Similarly, the proposed EDC-RC method produced an average RMSE about 4 times lower than the naive method. Thus, BC and EDC-RC methods produced more precise estimates than the naive method. Additionally, the EDC-RC method performed slightly better than the BC method in terms of the variability of the estimates.

Parameter	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	RMSE
$\hat{\beta}_{0T}$	-0.6581	-0.6527	-0.6500	-0.6500	-0.6477	-0.6380	0.0038
$\hat{\beta}_{0W}$	-0.6797	-0.6693	-0.6659	-0.6662	-0.6634	-0.6535	0.0169
$\hat{\beta}_{0BC}$	-0.6674	-0.6563	-0.6522	-0.6527	-0.6495	-0.6379	0.0058
$\hat{\beta}_{0EDC-RC}$	-0.6665	-0.6553	-0.6530	-0.6525	-0.6497	-0.6404	0.0049
$\hat{\beta}_{1T}$	-0.5244	-0.5062	-0.5005	-0.5003	-0.4944	-0.4778	0.0094
$\hat{\beta}_{1W}$	-0.4865	-0.4625	-0.4562	-0.4554	-0.4473	-0.4229	0.0461
$\hat{\beta}_{1BC}$	-0.5267	-0.5021	-0.4953	-0.4942	-0.4855	-0.4576	0.0144
$\hat{\beta}_{1EDC-RC}$	-0.5223	-0.5020	-0.4946	-0.4951	-0.4875	-0.4684	0.0115
$\hat{\sigma}^2_{\varepsilon\ T}$	0.0981	0.0996	0.1000	0.1000	0.1005	0.1019	0.0008
$\hat{\sigma}^2_{\varepsilon\ W}$	0.0996	0.1013	0.1017	0.1018	0.1023	0.1036	0.0019
$\hat{\sigma}^2_{\varepsilon\ BC}$	0.0983	0.1000	0.1004	0.1004	0.1010	0.1023	0.0009
$\hat{\sigma}^2_{\varepsilon\ EDC-RC}$	0.0989	0.1005	0.1010	0.1010	0.1015	0.1034	0.0013

Table 5.2: Summary statistics for the distribution of the linear regression model parameter estimates with a linked/upazila level covariate for each method over the 300 simulation runs. T and W refer to estimates based on non-displaced and displaced data, while BC and EDC-RC indicate the estimates based on the two proposed methods.

5.5.2 Estimated standard error of the model parameters under the proposed methods

In this section, we present the results of the simulation study on estimating the standard errors of the linear regression model parameters. Specifically, we focus on models with distance-based and upazila (or area)-level linked covariates, and we use the proposed standard error estimators under

the EDC-RC and BC methods. To estimate the standard error of the model parameters under these methods, we use single and double bootstrapping techniques, respectively, under the model and repeated random displacement, as described in Section 5.3. We evaluate the effectiveness of the proposed methods by comparing the average estimated standard errors and empirical standard errors of the model parameters over 300 simulation runs. The empirical standard error is obtained by calculating the standard deviation of the point estimates of the model parameters over the 300 simulation runs under the proposed EDC-RC and BC methods. Additionally, we report the estimated standard error and empirical standard error of the point estimates obtained using the naive method, which disregards displacement errors.

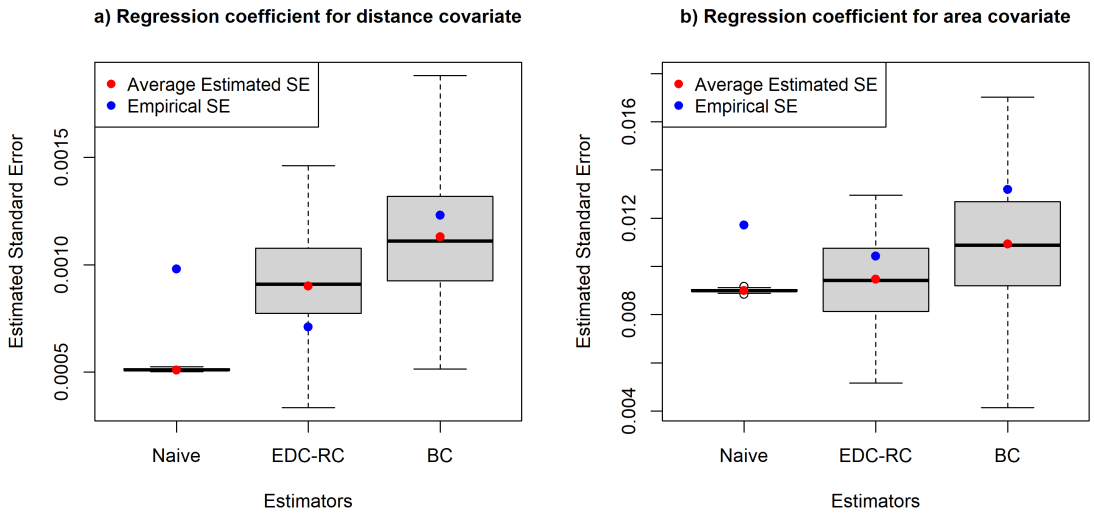

Figure 5.2: Boxplot of the estimated standard error (SE) of the regression coefficient for a) distance covariate and b) area (upazila) linked covariate under the naive, EDC-RC and BC estimators over 300 simulation runs.

Figure 5.2 shows the boxplot of the estimated standard errors of the point estimates of the regression coefficient for a) the distance covariate and b) the upazila linked covariate under the naive, EDC-RC and BC estimators over 300 simulations. The shape of the distribution of standard error estimates using the proposed EDC-RC and BC estimators under the DHS displacement process looks symmetric. Table 5.3 compares the average estimated standard errors and the empirical standard errors of the linear regression model parameters for three different estimators: naive, EDC-RC, and BC methods over 300 simulations under random displacement. It presents the results separately for the model with a distance covariate and the model with an area/linked covariate. For each model, it reports the estimated standard errors of the intercept and slope coefficients.

The estimated standard errors of the parameter estimates obtained under the proposed EDC-RC and BC methods are generally higher than those obtained under the naive method, which ignores the

displacement errors. The differences in estimated standard errors between the proposed methods and the naive method are relatively large in the case of the distance-based covariate. In contrast, the empirical standard errors of the point estimates under the naive method is somewhat larger than the proposed EDC-RC method. For example, the empirical standard error of the estimate for the distance coefficient is 0.00098 using the naive method and 0.00071 using the proposed EDC-RC method. The results indicate that the proposed methods provide a moderately accurate estimation of the true standard errors under random displacement. Regarding comparing the EDC-RC and BC methods, the EDC-RC-based method generally provides slightly lower estimated standard errors than the BC-based method.

Estimator	Model with distance covariate		Model with area/linked covariate	
	$\widehat{SE}(\hat{\beta}_0)$	$\widehat{SE}(\hat{\beta}_1)$	$\widehat{SE}(\hat{\beta}_0)$	$\widehat{SE}(\hat{\beta}_1)$
naive	0.00363 (0.00667)	0.00051 (0.00098)	0.00361 (0.00456)	0.00900 (0.01172)
EDC-RC	0.00564 (0.00443)	0.00090 (0.00071)	0.00383 (0.00418)	0.00947 (0.01042)
BC	0.00771 (0.00862)	0.00113 (0.00123)	0.00441 (0.00514)	0.01093 (0.01319)

Table 5.3: The average of the estimated standard errors and the empirical standard error of the linear regression model parameters using the naive, EDC-RC and BC methods over 300 simulation runs. Empirical standard errors are given within the brackets.

We compare the average estimated standard errors to the empirical standard errors of the model parameter estimates across the naive, EDC-RC, and BC methods. A closer alignment is observed between the average estimated and empirical standard errors for the distance-based covariate using the BC method, and for the upazila-level covariate using the EDC-RC method. Nonetheless, a noticeable overestimation is apparent in the model with the distance-based covariate using the EDC-RC approach. This implies that while the proposed methods generally offer reasonably accurate approximations of standard error estimates for model parameters, there are some occurrences of both overestimation and underestimation. On the other hand, the difference between the average estimated standard errors and the empirical standard errors under the naive method, which ignores the random displacement errors, is relatively large. The naive method tends to underestimate the standard errors. One possible explanation for the difference is that the estimates under the naive method do not account for displacement error. Another possible reason is that the naive estimator of the spatial covariate parameter is biased under random displacement, while the proposed EDC-RC and BC estimators are approximately unbiased.

In summary, the methods proposed under EDC-RC and BC estimators are effective in estimating standard errors of linear regression model parameters, yielding moderately accurate estimates with instances of over- and underestimation, in contrast to the naive method, which tends to underestimate the standard errors consistently. The empirical standard errors align somewhat closely with

the average estimated standard errors under the proposed methods, suggesting that the standard error estimates of the model parameter estimates are moderately well approximated in the presence of random displacement errors. Furthermore, the method based on EDC-RC produces slightly lower estimated standard errors compared to the one based on BC.

5.6 Conclusion

In this chapter, we examine the potential methods for estimating regression models, e.g., a linear regression model with spatial covariates under displacement. We propose two methods to fit the model. The proposed EDC under regression calibration (EDC-RC) method, which is developed under a measurement error model by bringing additional information from different sources to approximate the marginal distribution of the true locations. We also develop a methodology for estimating the model parameters of the linear regression model with spatial covariates using a bootstrap bias correction (BC) under displacement. The BC method is developed and expected to produce unbiased estimates of the model parameters satisfying the two crucial assumptions on the model form and the magnitude of the bias of the estimates under displacement. In addition to the point estimates, we develop the variance estimation of the estimates of the model parameters under the proposed EDC-RC and BC methods using model based bootstrapping under repeated random displacement.

We compare the usefulness of the two proposed EDC-RC and BC methods and the naive method that ignores the displacement error using simulated data under two types of spatial covariates, distance-based and linked area-level covariates. The simulated data are generated under the linear regression model, incorporating certain characteristics from the 2011 BDHS EAs. The naive estimates of the spatial covariate parameters are biased under displacement. The proposed EDC-RC and BC methods can produce approximately unbiased estimates of the spatial covariate parameters by correcting the bias of the estimates under displacement and outperform the naive method. Estimates using the proposed methods have a much lower RMSE of the estimates than the naive method. Finally, the proposed model-based bootstrap methods effectively estimate the variance of linear regression model parameter estimates under EDC-RC and BC estimators, exhibiting instances of over- and underestimation in the presence of random displacement errors. In comparison, the naive method tends to under-estimate the variance of the model parameter estimates ignoring the random displacement errors.

In this chapter, we developed two methods: EDC-RC and BC, in the context of linear models, to

correct the bias of spatial covariate effects under random displacement. However, there is often interest in estimating the effect of a spatial covariate on a binary outcome. In such cases, our methods can be directly extended to generalized linear models (e.g., logistic regression for a binary outcome) to mitigate the bias in estimates of spatial covariates arising from the random displacement of location coordinates. While RC is expected to provide improved estimates for non-linear models, such estimates might not maintain the same consistency as in linear regressions; however, any bias in estimates of a covariate measured with error in non-linear models is typically minor (for further details, see Carroll et al. (2006)). Thus, future research could extend the proposed methods to non-linear settings with spatial covariates to assess their performance under random displacement.

Chapter 6

Summary and research plan outline for future studies

6.1 Introduction

The thesis addresses the issue of obtaining precise statistical estimates and inferences using georeferenced data that has been aggregated and randomly displaced to protect respondents' confidentiality. The thesis develops methods for estimating density and domain parameters under measurement error models using georeferenced data that is randomly displaced, as well as aggregated and randomly displaced data. Furthermore, methods to correct bias in the estimates of spatial covariate parameters of the linear regression model and variance-component parameters of the linear mixed model due to the random displacement process are proposed. This chapter summarises the developed methods, their effects on inferences, their limitations, and promising directions for further research in the field of statistical use of aggregated and randomly displaced georeferenced data.

6.2 Thesis summary

One key contribution of the thesis is to have proposed estimators for obtaining density estimates and domain parameter estimates by correcting measurement errors due to the random displacement process, and the combined aggregation and random displacement process. This is shown in Chapter 3. In particular, in the case of displacement, we propose a measurement error model under the DHS random displacement process, followed by the development of a new conditional probability distribution of the displaced coordinates given the true coordinates. The measurement error model proposed can be potentially extended to other types of random displacement processes, for

example, the donut random displacement process. We then propose two estimation approaches for the measurement error model to obtain density and domain parameter estimates by drawing likely true location coordinates. The first method is based on Kernel Density Estimates (KDE) as the marginal distribution of the true location data. The estimation is implemented using the Stochastic Expectation-Maximisation (SEM) algorithm. The KDE-based method incorporates random displacement error and derives the bandwidth during estimation. This allows for automatic selection of an optimal bandwidth based on the data for each iteration, eliminating the need for pre-selection. As a result, implementation is simplified, and the method can be used with any bandwidth selection method, offering a high degree of flexibility. Additionally, since the grid size is unrelated to the displacement error, it does not affect kernel density estimation with randomly displaced data. However, when the proposed KDE-based method is used to correct the displacement error, a large grid size (e.g., 400) is required to cover all potential true locations within the buffers surrounding the displaced coordinates. Kernel density estimates effectively approximate the marginal distribution of unknown true location coordinates using register data. However, their performance may be limited in sampling scenarios where distinguishing between the absence of sample data and the absence of population units is challenging. Additionally, the random displacement process can create false in-sample and out-of-sample domains. Nonetheless, the method can still yield satisfactory results in sampling designs that cover the entire population area, such as a simple random sample across the entire area.

The second method is the External Data Classification (EDC)-based method, which is an alternative to the KDE-based method by approximating the unknown underlying distribution of true location data using additional information from external data sources. This is described in Chapter 3. Estimation in this EDC-based method is implemented through numerical integration. The EDC method is expected to improve the estimates over the KDE-based method as it utilises additional information from external data sources, which are more likely to be associated with the underlying distribution of the true location data. The EDC method performs better than the KDE method by more accurately approximating the marginal distribution of the true location coordinates using external data sources. However, the effectiveness of the EDC method relies on the availability and up-to-date nature of external data sources. Apart from the register data, in scenarios with survey data, such as the 2011 BDHS, where a clear link exists between the true enumeration area (EA) and the population density of household distribution, the EDC method outperforms the KDE method. This is particularly evident when the sampled DHS EAs are selected with a probability proportional to the census EA size, corresponding to the number of households.

In Chapter 3, we compare the effectiveness of the proposed KDE and EDC methods with the Most Probable Domain (MPD) method by Warren et al. (2016b), which selects the most likely domain for each EA, and the naive method that ignores displacement errors. The comparison is conducted using simulated register data generated from the WorldPop $100m \times 100m$ gridded population raster, which estimates the population of Bangladesh based on the 2011 census. Each grid cell in the raster represents the number of people per pixel. It is important to note that in the simulation study, external data sources, e.g., the WorldPop gridded population count, are used to generate the simulated true location data, ensuring the quality and representativeness of the data. However, in real situations, the quality and up-to-dateness of such external data sources should be assessed. The study demonstrates the utility of external data sources, such as WorldPop gridded population datasets, in approximating the underlying distribution of the true enumeration area (EA) locations using only a set of displaced EA coordinates from the 2011 BDHS. We note that to apply the proposed method, we consider the WorldPop gridded population counts as fixed, assuming they are measured without error. Therefore, future research could focus on extending the proposed method by accounting for uncertainty in the estimates of the WorldPop gridded population counts. The proposed KDE and EDC methods partially correct the bias in upazila (the admin 3 level of Bangladesh) estimates of the proportion of poor households under misplacement errors due to the random displacement process. These proposed estimators show lower root mean squared errors (RMSEs) than the naive and MPD estimators. Upazilas with smaller areas in terms of square kilometres tend to have higher misplacement errors, resulting in greater bias and variability when using the naive method. However, the proposed EDC method performs better by correcting misplacement errors, especially for upazilas with smaller areas. When the misplacement errors are low, all methods perform similarly. In such cases, the naive method should be selected for its simplicity. We observe that all correction methods may introduce additional errors for non-misplaced units since, in practice, it is not possible to determine which units are truly non-misplaced definitively. While the proposed methods demonstrate high correct upazila classification probabilities for non-misplaced units, the naive method has a perfect correct upazila classification probability of 1 for these units. However, regarding misplaced units, the correct upazila classification probability using the proposed method consistently exceeds that of the naive method, which is 0. Furthermore, the proposed methods generate higher correct upazila classification probabilities for units when their buffers around displaced coordinates intersect up to four upazilas. In summary, results based on the considered simulation scenario, we find that the proposed EDC method outperforms other methods, including KDE, MPD, and the naive method, which disregards misplacement errors, in terms of achieving higher correct upazila classification probabilities across all study units. In situations with high

misplacement errors, our proposed methodologies become even more invaluable than the naive method, which ignores misplacement errors. The proposed method should be applied to obtain estimates for upazilas with a higher number of misplaced units, which remain unknown to data analysts. Therefore, future research could focus on estimating the expected number of misplaced units within each domain (upazila) based on the displaced coordinates and other associated characteristics. These could include the domain size (measured in square km), its shape, and whether the unit is in a rural or urban location.

The proposed method is applied to real survey data under the DHS random displacement in Chapter 3. It is important to note that the DHS random displacement process remains consistent across all its surveys for every country, using identical urban-rural distance parameters. This consistency ensures that our developed approach can be applied to any DHS survey. However, upazila poverty estimates are obtained using the 2011 BDHS data by applying the proposed EDC method, the MPD method, and the naive method that disregards misplacement errors due to the DHS random displacement process. It is observed that the upazila estimates obtained from the proposed EDC method are closer to the true upazila estimates than those obtained from the MPD and naive methods. The random displacement process generates false out-of-sample upazilas. While the MPD method can address this issue for a few upazilas by selecting the most likely upazila for each enumeration area (EA), it introduces new false out-of-sample and in-sample upazilas. The proposed EDC and KDE methods can provide estimates for all falsely created out-of-sample upazilas resulting from the displacement process. This can be explained by the fact that, in the case of the proposed EDC and KDE methods, all upazilas within the displacement buffer, including the correct but unknown one, are selected based on the upazila classification probabilities for each EA. However, due to the correction process that involves creating a buffer around the displaced EA centroid, the EDC and KDE methods might introduce additional false in-sample upazilas, though they do not create false out-of-sample upazilas.

Under the aggregation and random displacement process, the aggregation concentrates data in one point, e.g., the centroid of the enumeration area (EA), while the random displacement moves the original location from one place to another. To the best of our knowledge, previous studies did not consider the uncertainty arising from aggregation and random displacement. Therefore, we propose a novel method called the external data and kernel density-based (KDE-ED) method to obtain density and domain parameter estimates by correcting aggregation and random displacement errors using a measurement error model. This is one of the important contributions in Chapter 3. The simulation study results indicate that the proposed KDE-ED method outperforms the naive

method, which ignores the aggregation and displacement errors, with lower bias and RMSE for the upazila estimates of the proportion of poor households. Furthermore, for the density estimates of poor households, the proposed KDE-ED method has a lower root mean integrated squared error (RMISE) compared to the naive method. Additionally, the proposed KDE-ED estimator can capture and preserve the underlying structure of the density, derived from the true coordinates of observation, more accurately across different aggregation and displacement levels. Therefore, the KDE-ED estimator is considered a preferred method for estimating the density of poor households, particularly when dealing with considerable aggregation and displacement errors.

Chapter 4 deals with the issues that come up when random effects in multilevel models are set at levels below the DHS displacement-restricted boundaries, such as sub-district or admin 3 of a country. This can cause observations to mix among groups or sub-districts, leading to biased estimates of random variance components in multilevel models and related model-based estimates for finite population parameters, for example, group means. When random effects are set at lower administrative levels, there is a bigger chance of misplaced EAs, leading to more bias in the estimates. This suggests that a DHS user defining random effects at the admin 4 level might obtain more biased estimates of random variance components than if defined at the admin 3 level. This, in situations with high misplacement errors, failing to account for these errors can lead to misleading inferences. In such cases, by applying our proposed methods, users can significantly reduce the bias in estimates, ensuring the reliability of their conclusions. In Chapter 4, we explore methods for estimating linear mixed models, obtaining estimates of the model parameters, and model-based group means under displacement. The framework of the linear mixed model under displacement is developed. The proposed model is fitted using an algorithm under the EDC method, which incorporates additional information from external sources through a measurement error model. Apart from the point estimates, variance estimation of the model parameter estimates under the EDC method is also developed using the law of total variance. A methodology for estimating the global parameters of the linear mixed model using a parametric bootstrap bias correction (BC) under displacement is also proposed to reduce the bias of the estimates of the model parameters. The proposed BC method is expected to perform well in reducing the bias in estimates, provided the assumptions regarding the model form and magnitude of bias under random displacement are met. Furthermore, variance estimation for the estimates of the linear mixed model parameters is developed under the proposed BC method, accounting for random displacement errors. This is conducted using model-based bootstrapping under repeated random displacement. The BC method should effectively correct the bias of global (model) parameters. However, it may not work for the associated model-based

finite population parameters, such as group means under random displacement. This is because the assumption of bootstrapping regarding the magnitude of the bias of predicted group means under random displacement may not be satisfied. There are two likely reasons for this. First, creating a buffer around truly non-misplaced EAs unknown to the data analysts may introduce additional uncertainty in the estimates of upazila means. Second, due to random displacement, there may be a mixing of groups within the buffer around the true and displaced locations, potentially affecting the estimates of group means. We have explored this through a simulation study.

The performance of the proposed EDC and BC methods and the naive method that ignores displacement error are compared using simulated data in Chapter 4. The data are generated under the linear mixed model with characteristics similar to the 2011 BDHS EAs. The naive estimates of variance component parameters and upazila means are biased under displacement, while the proposed EDC method can correct the bias to some extent and is preferred over the naive method. Additionally, the EDC estimator has a lower RMSE of upazila mean estimates compared to the BC and naive estimators. The BC method produces unbiased estimates of the linear mixed model parameters under displacement, outperforming the EDC and naive methods. However, it is not suitable for estimating upazila means and has a higher RMSE of upazila estimates than the EDC and naive methods. As a result, the BC method is recommended for obtaining estimates of global parameters under displacement. In contrast, the EDC method is preferred for estimating finite population parameters such as upazila means. Finally, the proposed variance estimators under the BC and EDC methods perform moderately well in estimating the variances of the model parameters estimates under the random displacement process. Specifically, the variance approximation is good for the fixed effect parameters. However, for the variance components, which are affected by misplacement errors due to random displacement, a variance estimation may not be as appropriate as it does not account for the bias of the estimator.

The potential methods for estimating a linear regression model with spatial covariates under displacement are investigated in Chapter 5. A framework of the linear regression model with spatial covariates under displacement is developed, and two methods are proposed for fitting the model. The first is the external data based on the regression calibration (EDC-RC) method, which uses a measurement error model to incorporate additional information from different sources and approximate the marginal distribution of the true locations. The second method is a bootstrap bias correction (BC) for estimating the model parameters of the linear regression model with spatial covariates under displacement. The BC method can obtain unbiased estimates of the model parameters when the assumptions are satisfied regarding the model's suitability to the data and the extent

of bias arising from random displacement. In addition to the point estimates, the variance estimation of the model parameter estimates under the proposed EDC-RC and BC methods is developed using model-based bootstrapping under repeated random displacement.

In Chapter 5, we compare the performance of the two proposed methods, EDC-RC and BC, with the naive method that ignores displacement error using simulated data. The simulated data are generated using a linear regression model with two types of spatial covariates: distance-based and linking area-level covariates. The simulated data incorporate characteristics from the 2011 BDHS. For instance, the true values of the parameters for the linear regression model are chosen based on estimates derived using outcome and covariate values from the 2011 BDHS. The spatial covariate parameters estimated by the naive method are biased under displacement. However, the proposed EDC-RC and BC methods can produce approximately unbiased estimates of the spatial covariate parameters by correcting the bias of the estimates under displacement, and they outperform the naive method. The RMSE of the estimates using the proposed methods is much lower than that of the naive method. Finally, the proposed model-based bootstrap methods effectively achieve variance estimation of linear regression model parameter estimates under EDC-RC and BC estimators in the presence of random displacement errors, albeit with instances of over- and underestimation. In comparison, the variance of the model parameter estimates tends to be underestimated by the naive method, which ignores the random displacement errors.

6.3 Future research

This thesis proposes methods for obtaining precise statistical estimates and inferences with aggregated and displaced georeferenced data. However, theoretical and practical research questions still need to be addressed regarding the aggregation and displacement process, modifications of the methods, development of quality measures for point estimates, and other related issues which could be explored in future research. The thesis concludes with some ideas for future research in this context as follows:

6.3.1 Implementation of the proposed methods under other random displacement processes

One potential area for future research is the implementation of the proposed methods under different random displacement processes. In this thesis, while the measurement error model and the performance of the proposed methods have been developed and illustrated for the commonly used DHS random displacement processes, other displacement processes are described in the literature

review Chapter 2, such as the Donut random displacement process. To extend the measurement error model to other displacement processes, developing the probability distribution of the displaced coordinates under that specific displacement process would be necessary. For example, in the Donut random displacement process, a displaced location is randomly chosen within a minimum and maximum displacement distance range, forming a "donut-shaped" region around the original location (Hampton et al., 2010). The minimum and maximum displaced distances are denoted as δ_1 and δ_2, respectively. To assess the performance of the proposed methods under the Donut random displacement process, the probability distribution of the Donut random displaced coordinates would need to be developed, similar to how the probability distribution was developed for the DHS displaced coordinates in Section 3.2.3. The domain of the probability distribution function would be a bounded set of all points between the two smaller and larger circles with radii δ_1 and δ_2.

Under the Donut random displacement process, the likelihood of preserving respondent confidentiality should be higher, as no displaced points are selected below the minimum displaced distance δ_1. However, this process may introduce larger misplacement errors compared to the DHS random displacement process. In this case, the naive method that ignores the displacement error may perform worse in obtaining statistical estimates. Nevertheless, the proposed methods are expected to outperform the naive methods by correcting misplacement errors under the Donut random displacement process. Future research could focus on evaluating the performance of the proposed methods under the Donut random displacement process and comparing them to the naive method. This analysis would provide insights into the effectiveness of the methods in correcting misplacement errors and obtaining accurate statistical estimates.

6.3.2 Estimation of misplacement probability under random displacement process

In this thesis, the proposed EDC and KDE methods introduce additional uncertainty for unit (household or EA) locations that are truly non-misplaced but unknown to the data analysts. Consequently, domain (upazila) estimates for potentially but not truly misplaced units will be biased under the proposed methods, while the naive estimates that disregard misplacement errors are expected to be unbiased. However, the naive method produces biased upazila estimates when truly misplaced units are present. Moreover, upazilas with smaller areas in terms of square kilometres tend to have higher misplacement errors, leading to increased bias and variability when using the naive method. The proposed methods outperform the naive method in such cases by effectively correcting misplacement errors, particularly for upazilas with smaller areas. Overall, the performance of the naive method is better when misplacement errors arising from the random displacement

process are minimal. Hence, considering the expected number of misplaced units within each upazila becomes crucial in determining the effectiveness of the proposed method and selecting the appropriate method to use without intending to increase the disclosure risk of the respondents.

All the corrected methods discussed in this thesis introduce additional errors for truly non-misplaced units, as the analysts are unaware of which units are misplaced. Therefore, future research can focus on estimating the expected number of misplaced units within each upazila using displaced coordinates. This estimation can help identify which units are more likely to be misplaced, allowing the method to focus on the most probable misplacement units. This also helps avoid disturbing potential but not truly misplaced units. Factors such as the area and shape of the upazila, the proximity of households to the boundary, and whether the locations are in urban or rural areas are potential factors associated with misplacement errors. Therefore, using these potential factors, a statistical model could be developed to estimate the probabilities of misplacement for location units due to the random displacement process. This model can be used to estimate the number of misplaced units within each upazila, and the proposed method is best applied to upazilas with higher such estimates.

6.3.3 Dealing with random displacement uncertainty in spatial estimates through Kriging predictions

In the proposed Kernel Density Estimation (KDE) based method, density estimates of the location data are used to approximate the distribution of the unknown true location coordinates, which is a component of the measurement error model for the random displacement process. The corrected density estimates are obtained by using the Stochastic Expectation-Maximization (SEM) algorithm, allowing for estimation that accounts for random displacement error, as detailed in Section 3.3.

An alternative approach to estimating the distribution of unknown true locations can be using Kriging predictions, a geostatistical interpolation method. Kriging predictions allow us to estimate values at unobserved locations based on available spatial data and the underlying spatial correlation structure of the random field (Matheron, 1963; Goovaerts, 2019). For an overview of Kriging predictions, see Myers (1994). However, the observed location data is randomly displaced, which can affect the spatial correlation of the data and, consequently, impact the accuracy of Kriging predictions. Therefore, similar to kernel density estimates, an iterative estimation algorithm needs to be developed to obtain displacement error-corrected Kriging predictions for potential true location coordinates.

Displacement error-corrected Kriging predictions have the potential to provide better estimates of

the distribution of unknown true locations compared to kernel density estimates. This is because Kriging predictions consider the spatial correlation and variability of a continuous spatial process, enabling more accurate predictions at unsampled locations (Myers, 1994; Goovaerts, 2019). In light of this, conducting future research to examine the effectiveness of using Kriging predictions to address the uncertainty due to random displacement could greatly improve statistical estimates.

6.3.4 Development of an R package

In the future, I plan to develop an R package based on the methodologies presented in this thesis. This will make it easier for other researchers to apply and build upon the methods we have proposed. An R package will also promote reproducibility and replicability, which are crucial in academic research. While the development of the R package is still in its planning stages, all codes and simulation setups from this thesis are currently available upon request to facilitate further research.

Bibliography

Aigner, D. J. (1973). Regression with a binary independent variable subject to errors of observation. *Journal of Econometrics*, 1(1):49–59.

Ajayakumar, J., Curtis, A. J., and Curtis, J. (2019). Addressing the data guardian and geospatial scientist collaborator dilemma: how to share health records for spatial analysis while maintaining patient confidentiality. *International Journal of Health Geographics*, 18:1–12.

Allshouse, W. B., Fitch, M. K., Hampton, K. H., Gesink, D. C., Doherty, I. A., Leone, P. A., Serre, M. L., and Miller, W. C. (2010). Geomasking sensitive health data and privacy protection: an evaluation using an e911 database. *Geocarto international*, 25(6):443–452.

Alom, J., Quddus, M. A., and Islam, M. A. (2012). Nutritional status of under-five children in Bangladesh: a multilevel analysis. *Journal of Biosocial Science*, 44(5):525–535.

Arbia, G., Espa, G., and Giuliani, D. (2015). Measurement errors arising when using distances in microeconometric modelling and the individuals' position is geo-masked for confidentiality. *Econometrics*, 3(4):709–718.

Arima, S. and Polettini, S. (2019). A unit level small area model with misclassified covariates. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182(4):1439–1462.

Balk, D., Pullum, T., Storeygard, A., Greenwell, F., and Neuman, M. (2004). A spatial analysis of childhood mortality in west africa. *Population, Space and Place*, 10(3):175–216.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

BBS (2011). Population and housing census 2011: preliminary results. Published by the Bangladesh Bureau of Statistics (BBS), Ministry of Planning, People's Republic of Bangladesh. Dhaka, Bangladesh. Retrieved from https://bbs.portal.gov.bd/sites/default/files/files/bbs.portal.gov.bd/page/7b7b171a_731a_4854_8e0a_f8f7dede4a4a/PHC2011PreliminaryReport.pdf.

BBS (2014). Bangladesh population and housing census-2011. national report volume-2: Union statistics. Published by the Bangladesh Bureau of Statistics (BBS), Ministry of Planning, People's Republic of Bangladesh. Dhaka, Bangladesh. Retrieved from http://203.112.218.65:8008/WebTestApplication/userfiles/Image/National%20Reports/Union%20Statistics.pdf.

Bhuiyan, M. K. J., Hossain, M. J., Islam, M. A., Imam, M. F., and Quddus, M. A. (2020). Small area estimation of nutritional status of under-five children in sylhet division: An M-Quantile approach. *The Bangladesh Journal of Agricultural Economics*, 41(1):59–71.

Blitzstein, J. K. and Hwang, J. (2019). *Introduction to Probability*. Chapman & Hall, Boca Raton, FL, second edition.

Bondarenko, M., Kerr, D., Sorichetta, A., Tatem, A., et al. (2020). Census/projection-disaggregated gridded population datasets, adjusted to match the corresponding unpd 2020 estimates, for 51 countries across sub-saharan africa using building footprints. Published by the WorldPop, University of Southampton, UK. Retrieved on 17 October 2020 from https://hub.worldpop.org/doi/10.5258/SOTON/WP00683.

Boulos, M. N. K., Curtis, A. J., and AbdelMalik, P. (2009). Musings on privacy issues in health research involving disaggregate geographic data about individuals. *International Journal of Health Geographics*, 8(1):1–8.

Bruzelius, C. and Shutes, I. (2022). Towards an understanding of mobility in social policy research. *Global Social Policy*, 22(3):503–520.

Buonaccorsi, J. P. (2010). *Measurement error: models, methods, and applications*. Chapman and Hall/CRC, New York, 1st edition.

Burgert, C. R., Colston, J., Roy, T., and Zachary, B. (2013). *Geographic Displacement Procedure and Georeferenced Data Release Policy for the Demographic and Health Surveys*. DHS Spatial Analysis Reports No. 7. Calverton, Maryland, USA: ICF International. Available at http://dhsprogram.com/pubs/pdf/SAR7/SAR7.pdf.

Burgert, C. R., Dontamsetti, T., and Gething, P. W. (2018). The dhs program's modeled surfaces spatial datasets. *Studies in Family Planning*, 49(1):87–92.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: a modern perspective*. Chapman and Hall/CRC, New York, 2nd edition.

Carroll, R. J. and Stefanski, L. A. (1990). Approximate quasi-likelihood estimation in models with surrogate predictors. *Journal of the American Statistical Association*, 85(411):652–663.

Cassa, C. A., Grannis, S. J., Overhage, J. M., and Mandl, K. D. (2006). A context-sensitive approach to anonymizing spatial surveillance data: impact on outbreak detection. *Journal of the American Medical Informatics Association*, 13(2):160–165.

Chao, A. and Colwell, R. K. (2017). Thirty years of progeny from chao's inequality: Estimating and comparing richness with incidence data and incomplete sampling. *Sort-statistics and Operations Research Transactions*, 1:3–54.

Chen, D. (2018). *A Comparison of Alternative Bias-Corrections in the Bias-Corrected Bootstrap Test of Mediation*. Doctoral dissertation, University of Nebraska-Lincoln, URL: http://digitalcommons.unl.edu/cehsdiss/318.

Das, S., Chandra, H., and Saha, U. R. (2019a). District level estimates and mapping of prevalence of diarrhoea among under-five children in Bangladesh by combining survey and census data. *PloS One*, 14(2):e0211062. Available at https://doi.org/10.1371/journal.pone.0211062.

Das, S., Kumar, B., Hossain, M. Z., Rahman, S. T., and Rahman, A. (2020a). Estimation of child undernutrition at disaggregated administrative tiers of a north-eastern district of Bangladesh: An application of small area estimation method. In Rahman, A., editor, *Statistics for Data Science and Policy Analysis*, pages 267–281, Singapore. Springer. doi:10.1007/978-981-15-1735-8_20.

Das, S., Kumar, B., and Kawsar, L. A. (2020b). Disaggregated level child morbidity in Bangladesh: An application of small area estimation method. *PloS One*, 15(5):e0220164. doi: 10.1371/journal.pone.0220164.

Das, S., Rahman, A., Ahamed, A., and Rahman, S. T. (2019b). Multi-level models can benefit from minimizing higher-order variations: an illustration using child malnutrition data. *Journal of Statistical Computation and Simulation*, 89(6):1090–1110.

Deffner, V., Küchenhoff, H., Breitner, S., Schneider, A., Cyrys, J., and Peters, A. (2018). Mixtures of berkson and classical covariate measurement error in the linear mixed model: Bias analysis and application to a study on ultrafine particles. *Biometrical Journal*, 60(3):480–497.

Dehnad, K. (1987). Density estimation for statistics and data analysis. *Technometrics*, 29(4):495.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete

data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Duong, T. (2007). ks: Kernel density estimation and kernel discriminant analysis for multivariate data in r. *Journal of Statistical Software*, 21(7):1–15. URL: https://www.jstatsoft.org/article/view/v021i07.

Duong, T., Wand, M., Chacon, J., and Gramacki, A. (2022). ks: Kernel smoothing. r package version 1.14.0. URL: https://cran.r-project.org/package=ks.

Dupriez, O. and Boyko, E. (2010). *Dissemination of Microdata Files: Principles, Procedures, and Practices*. International Household Survey Network. Working Paper No. 005. Available at http://ihsn.org/sites/default/files/resources/IHSN-WP005.pdf.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Stat.*, 7:1–26.

Efron, B. and Tibshirani, R. (1992). *An Introduction to the Bootstrap*. Wiley & Sons, New York, 1st edition.

Elkies, N., Fink, G., and Bärnighausen, T. (2015). "scrambling" geo-referenced data to protect privacy induces bias in distance estimation. *Population and Environment*, 37(1):83–98.

Faizuddin, A., Naomi, A., Mehar, A., Dean, J., Mehrin, M., Iffath, S., Nobuo, Y., and Salman, Z. (2012). *Poverty Maps of Bangladesh-2010 Technical Report*. Washington, D.C. : World Bank Group. Working Paper No. 90487. Available at http://documents.worldbank.org/curated/en/160611468014459434/Technical-report.

Falkingham, J. and Namazie, C. (2002). *Measuring health and poverty: a review of approaches to identifying the poor*. The DFID Health Systems Resource Centre, London.

Fanshawe, T. and Diggle, P. (2011). Spatial prediction in the presence of positional error. *Environmetrics*, 22(2):109–122.

Feldacker, C., Emch, M., and Ennett, S. (2010). The who and where of hiv in rural Malawi: Exploring the effects of person and place on individual hiv status. *Health & place*, 16(5):996–1006.

Fielding, A., Yang, M., and Goldstein, H. (2003). Multilevel ordinal models for examination grades. *Statistical Modelling*, 3(2):127–153.

Fotheringham, A. S. and Zhan, F. B. (1996). A comparison of three exploratory methods for cluster detection in spatial point patterns. *Geographical Analysis*, 28(3):200–218.

Gabrosek, J. and Cressie, N. (2002). The effect on attribute prediction of location uncertainty in spatial data. *Geographical Analysis*, 34(3):262–285.

Galea, S. and Vaughan, R. D. (2019). Public health, politics, and the creation of meaning: A public health of consequence, july 2019. *American Journal of Public Health*, 109(7):966–968.

Gelman, A. and Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, Cambridge, 1st edition.

Gething, P., Tatem, A., Bird, T., and Burgert, C. (2015). *Creating spatial interpolation surfaces with DHS data*. DHS Spatial Analysis Reports No. 11. Rockville, Maryland, USA: ICF International. Available at http://dhsprogram.com/pubs/pdf/SAR11/SAR11.pdf.

Gidado, O. (2012). *Community and individual level factors influencing modern contraceptive use among married women of reproductive age group (15-49) in Nigeria*. Phd thesis, The African Digital Health Library (ADHL), University of Ibadan, Kenya. Retrieved from https://library.adhl.africa/handle/123456789/12176.

Gleser, L. J. (1991). Measurement error models. *Chemometrics and Intelligent Laboratory Systems*, 10(1):45–57. Proceedings of the Mathematics in Chemistry Conference.

Goldstein, H. and Shlomo, N. (2020). A probabilistic procedure for anonymisation, for assessing the risk of re-identification and for the analysis of perturbed data sets. *Journal of Official Statistics*, 36(1):89–115.

Goovaerts, P. (2019). Kriging interpolation. In Wilson, J. P., editor, *The Geographic Information Science & Technology Body of Knowledge*. 4th quarter edition. doi: 10.22224/gistbok/2019.4.4.

Grace, K., Nagle, N. N., Burgert-Brucker, C. R., Rutzick, S., Van Riper, D. C., Dontamsetti, T., and Croft, T. (2019). Integrating environmental context into dhs analysis while protecting participant confidentiality: A new remote sensing method. *Population and Development Review*, 45(1):197–218.

Grace, Y. Y. (2017). *Statistical analysis with measurement error or misclassification: strategy, method and application*. Springer, New York, NY, 1st edition.

Groß, M., Kreutzmann, A.-K., Rendtel, U., Schmid, T., and Tzavidis, N. (2020). Switching between different non-hierachical administrative areas via simulated geo-coordinates: A case study for student residents in berlin. *Journal of Official Statistics*, 36(2):297–314.

Groß, M., Rendtel, U., Schmid, T., Schmon, S., and Tzavidis, N. (2017). Estimating the density

of ethnic minorities and aged people in berlin: multivariate kernel density estimation applied to sensitive georeferenced administrative data protected via measurement error. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(1):161–183.

Hampton, K. H., Fitch, M. K., Allshouse, W. B., Doherty, I. A., Gesink, D. C., Leone, P. A., Serre, M. L., and Miller, W. C. (2010). Mapping health data: improved privacy protection with donut method geomasking. *American Journal of Epidemiology*, 172(9):1062–1069.

Hardin, J. W., Schmiediche, H., and Carroll, R. J. (2003). The regression-calibration method for fitting generalized linear models with additive measurement error. *The Stata Journal*, 3(4):361–372.

Higgins, J. J. (2004). *An introduction to modern nonparametric statistics*. Brooks/Cole Pacific Grove, CA, 1st edition.

Härdle, W. and Scott, D. (1990). *Smoothing by weighted averaging of rounded points*. Université catholique de Louvain, Center for Operations Research and Econometrics (CORE). LIDAM Discussion Papers, No. CORE 1990040.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Longhurst, J., Nordholt, E. S., Seri, G., and Wolf, P. (2010). *Handbook on statistical disclosure control*. Version 1.2, ESSnet on Statistical Disclosure Control. Rtrieved from https://cros-legacy.ec.europa.eu/system/files/SDC_Handbook.pdf.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., and De Wolf, P.-P. (2012). *Statistical disclosure control*. Wiley, New York.

Imam, M. F., Islam, M. A., and Hossain, M. (2018). Factors affecting poverty in rural Bangladesh: An analysis using multilevel modelling. *Journal of the Bangladesh Agricultural University*, 16(1):123–130.

Islam, M. (2005). Consistency in reporting contraception between spouses in bangladesh: a multilevel data analysis. *Fertility and Sterility*, 84(1):S169.

Karra, M., Canning, D., and Sato, R. (2020). Adding measurement error to location data to protect subject confidentiality while allowing for consistent estimation of exposure effects. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 69(5):1251–1268.

Kasaye, H. K., Bobo, F. T., Yilma, M. T., and Woldie, M. (2019). Poor nutrition for under-five

children from poor households in ethiopia: Evidence from 2016 demographic and health survey. *PloS One*, 14(12):e0225996.

Kwan, M.-P., Casas, I., and Schmitz, B. (2004). Protection of geoprivacy and accuracy of spatial information: How effective are geographical masks? *Cartographica: The International Journal for Geographic Information and Geovisualization*, 39(2):15–28.

Lazar, A., Bondarenko, M., Chamberlain, H., Jochem, W., Darin, E., Qader, S., and Tatem, A. (2021). Small area population estimates using random forest top-down disaggregation (part 2): the popRF 'r' package. WorldPop, University of Southampton. doi:10.5258/SOTON/WP00727.

Leasure, D., Darin, E., Tatem, A., and Bondarenko, M. (2021). Small area population estimates using random forest top-down disaggregation: An r tutorial. WorldPop, University of Southampton, UK. Available at https://eprints.soton.ac.uk/446189/.

Leyland, A. H. and Goldstein, H. (2001). *Multilevel modelling of health statistics*. John Wiley and Sons Ltd.

Leyland, A. H. and Groenewegen, P. P. (2020). *Multilevel Modelling for Public Health and Health Services Research*. Springer Open, Switzerland.

Lohela, T. J., Campbell, O. M., and Gabrysch, S. (2012). Distance to care, facility delivery and early neonatal mortality in Malawi and Zambia. *PloS One*, 7(12):e52110.

Matheron, G. (1963). Principles of geostatistics. *Economic Geology*, 58(8):1246–1266.

Melamed, D. and Vuolo, M. (2019). Assessing differences between nested and cross-classified hierarchical models. *Sociological Methodology*, 49(1):220–257.

Minnotte, M. C. (1998). Achieving higher-order convergence rates for density estimation with binned data. *Journal of the American Statistical Association*, 93(442):663–672.

Muller, M. E. and Muller, M. (1958). A note on the generation of random normal deviates. *Annals of Mathematical Statistics*, 29(2):610–11.

Myers, D. E. (1994). Spatial interpolation: an overview. *Geoderma*, 62(1-3):17–28.

NIPORT (2013). Bangladesh demographic and health survey 2011. National Institute of Population Research and Training Dhaka, Bangladesh. MEASURE DHS ICF International Calverton, Maryland, U.S.A. Retrieved from https://dhsprogram.com/pubs/pdf/fr265/fr265.pdf.

Ogbo, F. A., Page, A., Idoko, J., and Agho, K. E. (2018). Population attributable risk of key

modifiable risk factors associated with non-exclusive breastfeeding in Nigeria. *BMC Public Health*, 18(1):1–9.

Pande, S., Keyzer, M. A., Arouna, A., and Sonneveld, B. G. (2008). Addressing diarrhea prevalence in the west african middle belt: social and geographic dimensions in a case study for Benin. *International Journal of Health Geographics*, 7(1):1–17.

Perez-Heydrich, C., Joshua L, W., Clara R, B., and Michael, E. E. (2013). *Guidelines on the Use of DHS GPS Data.* DHS Spatial Analysis Report No. 8, Calverton, Maryland, USA: ICF International. Available at https://www.dhsprogram.com/pubs/pdf/SAR8/SAR8.pdf.

Perez-Heydrich, C., Warren, J. L., Burgert, C. R., and Emch, M. E. (2016). Influence of demographic and health survey point displacements on raster-based analyses. *Spatial Demography*, 4(2):135–153.

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL:http://www.R-project.org/.

Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2003). Maximum likelihood estimation of generalized linear models with covariate measurement error. *The Stata Journal*, 3(4):386–411.

Rahman, M. M. and Szabó, G. (2020). National spatial data infrastructure (nsdi) of BANGLADESH development, progress and way forward. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, 5(4):131–138.

Rushton, G. and Lolonis, P. (1996). Exploratory spatial analysis of birth defect rates in an urban population. *Statistics in medicine*, 15(7-9):717–726.

Scott, D. W. (1985). Averaged shifted histograms: effective nonparametric density estimators in several dimensions. *The Annals of Statistics*, 13(3):1024–1040.

Scott, D. W. (2015). *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley, 2nd edition.

Scott, D. W. and Sheather, S. J. (1985). Kernel density estimation with binned data. *Communications in Statistics-Theory and Methods*, 14(6):1353–1359.

Seidl, D. E., Jankowski, P., and Tsou, M.-H. (2016). Privacy and spatial pattern preservation in masked gps trajectory data. *International Journal of Geographical Information Science*, 30(4):785–800.

Shi, X., Alford-Teaster, J., and Onega, T. (2009). Kernel density estimation with geographically

masked points. In *2009 17th International Conference on Geoinformatics*, pages 1–4. IEEE, USA. Conference paper. doi:10.1109/GEOINFORMATICS.2009.5292881.

Silverman, B. (1982). Kernel density estimation using the fast fourier transform applied statistics. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 31(1):93–99.

Silverman, B. W. (2018). *Density estimation for statistics and data analysis*. Springer, New York, NY, 1st edition.

Simonoff, J. S. and Simonoff, J. S. (1996). Further applications of smoothing. In Simonoff, J. S., editor, *Smoothing Methods in Statistics*, pages 252–274. Springer-Verlag, New York.

Smith, T. and Shively, G. (2019). Multilevel analysis of individual, household, and community factors influencing child growth in Nepal. *BMC Pediatrics*, 19(1):1–14.

Sorichetta, A., Hornby, G. M., Stevens, F. R., Gaughan, A. E., Linard, C., and Tatem, A. J. (2015). High-resolution gridded population datasets for latin america and the caribbean in 2010, 2015, and 2020. *Scientific Data*, 2(1):1–12.

Spiegelman, D., McDermott, A., and Rosner, B. (1997). Regression calibration method for correcting measurement-error bias in nutritional epidemiology. *The American Journal of Clinical Nutrition*, 65(4):1179S–1186S.

Statistics Canada (2011). Canadian community health survey (cchs) annual component: user guide 2012 and 2011-2012 microdata files. Statistics Canada, Ottawa. Retrieved on 7 October 2021 from https://www23.statcan.gc.ca/imdb-bmdi/pub/document/3226_D7_T9_V8-eng.pdf.

Stevens, F. R., Gaughan, A. E., Linard, C., and Tatem, A. J. (2015). Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. *PloS One*, 10(2):e0107042.

Tillé, Y. and Matei, A. (2022). 'sampling' r package version 2.9. URL: https://cran.r-project.org/package=sampling.

Vayena, E., Salathé, M., Madoff, L. C., and Brownstein, J. S. (2015). Ethical challenges of big data in public health. *PLoS Computational Biology*, 11(2):e1003904.

Wakefield, J. and Shaddick, G. (2006). Health-exposure modeling and the ecological fallacy. *Biostatistics*, 7(3):438–455.

Walter, P., Groß, M., Schmid, T., and Weimer, K. (2022). Iterative kernel density estimation applied

to grouped data: estimating poverty and inequality indicators from the German microcensus. *Journal of Official Statistics*, 38(2):599–635.

Wand, M. (1994). Fast computation of multivariate kernel estimators. *Journal of Computational and Graphical Statistics*, 3(4):433–445.

Wand, M. P. and Jones, M. C. (1994). Multivariate plug-in bandwidth selection. *Computational Statistics*, 9(2):97–116.

Wang, B. and Wertelecki, W. (2013). Density estimation for data with rounding errors. *Computational Statistics & Data Analysis*, 65:4–12.

Wang, L., Ardern, C. I., and Chen, D. (2020). Geographic variation in cardiovascular disease mortality: A study of linking risk factors and built environment at a local health unit in Canada. In Delmelle, L. Y., editor, *Geospatial Technologies for Urban Health*, pages 31–51. Springer.

Warren, J. L., Perez-Heydrich, C., Burgert, C. R., and Emch, M. E. (2016a). Influence of demographic and health survey point displacements on distance-based analyses. *Spatial Demography*, 4(2):155–173.

Warren, J. L., Perez-Heydrich, C., Burgert, C. R., and Emch, M. E. (2016b). Influence of demographic and health survey point displacements on point-in-polygon analyses. *Spatial Demography*, 4(2):117–133.

Wasfi, R. A., Ross, N. A., and El-Geneidy, A. M. (2013). Achieving recommended daily physical activity levels through commuting by public transportation: Unpacking individual and contextual influences. *Health & Place*, 23:18–25.

WHO, W. G. (1986). Use and interpretation of anthropometric indicators of nutritional status. *Bulletin of the World health organization*, 64(6):929–941.

Wilson, E., Hazel, E., Park, L., Carter, E., Moulton, L. H., Heidkamp, R., and Perin, J. (2020). Obtaining district-level health estimates using geographically masked location from demographic and health survey data. *International Journal of Health Geographics*, 19(1):1–14.

Wilson, K. and Wakefield, J. (2021). Estimation of health and demographic indicators with incomplete geographic information. *Spatial and Spatio-temporal Epidemiology*, 37:100–421.

Ybarra, L. M. and Lohr, S. L. (2008). Small area estimation when auxiliary information is measured with error. *Biometrika*, 95(4):919–931.

Zandbergen, P. A. (2014). Ensuring confidentiality of geocoded health data: Assessing geographic

masking strategies for individual-level data. *Advances in Medicine*, 2014:Article ID 567049, 14 pages. URL: https://doi.org/10.1155/2014/567049.

Zhang, S., Freundschuh, S. M., Lenzer, K., and Zandbergen, P. A. (2017). The location swapping method for geomasking. *Cartography and Geographic Information Science*, 44(1):22–34.