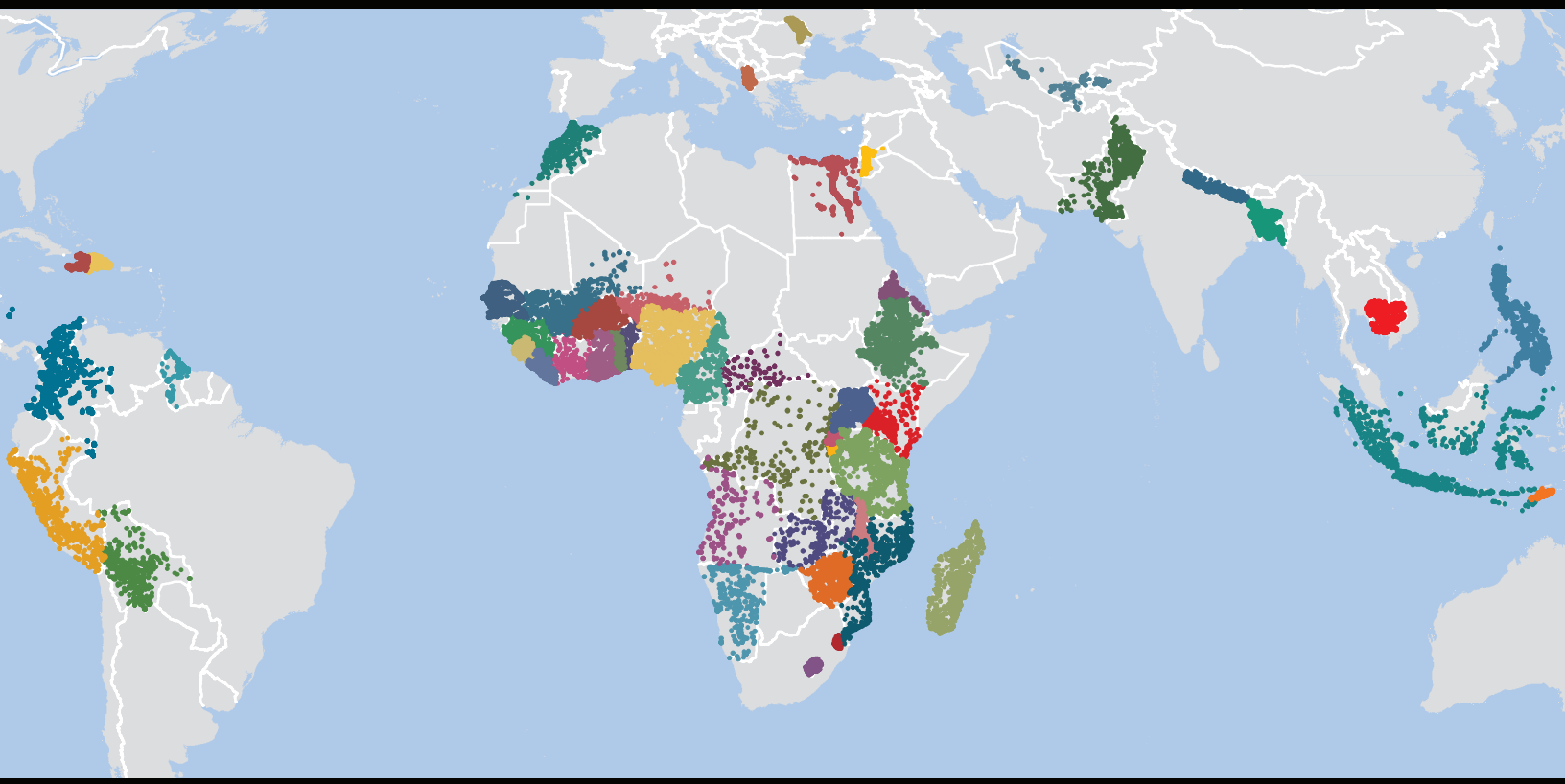
USAID
FROM THE AMERICAN PEOPLE

GUIDELINES ON THE USE OF DHS GPS DATA

DHS SPATIAL ANALYSIS REPORTS 8



SEPTEMBER 2013

This publication was produced for review by the United States Agency for International Development (USAID). The report was prepared by Carolina Perez-Heydrich of the University of North Carolina, Chapel Hill, NC, USA and Meredith College, Raleigh, NC USA; Joshua L. Warren of the University of North Carolina, Chapel Hill, NC, USA; Clara R. Burgert of ICF International, Calverton, MD, USA; and Michael E. Emch of the University of North Carolina, Chapel Hill, NC, USA.

MEASURE DHS assists countries worldwide in the collection and use of data to monitor and evaluate population, health, and nutrition programs. Additional information about the MEASURE DHS project can be obtained by contacting MEASURE DHS, ICF International, 11785 Beltsville Drive, Suite 300, Calverton, MD 20705 (telephone: 301-572-0200; fax: 301-572-0999; e-mail: reports@measuredhs.com; internet: www.measuredhs.com).

The main objectives of the MEASURE DHS project are:

- to provide decision makers in survey countries with information useful for informed policy choices;
- to expand the international population and health database;
- to advance survey methodology; and
- to develop in participating countries the skills and resources necessary to conduct high-quality demographic and health surveys.

Spatial Analysis Reports No. 8

Guidelines on the Use of DHS GPS Data

Carolina Perez-Heydrich[1,2]

Joshua L. Warren[1]

Clara R. Burgert[3]

Michael E. Emch[1]

[1]University of North Carolina, Chapel Hill, NC, USA
[2]Meredith College, Raleigh, NC, USA
[3]ICF International, Calverton, MD, USA

September 2013

Corresponding author: Clara R. Burgert, International Health and Development, ICF International, 11785 Beltsville Drive, Calverton, Maryland 20705, USA; Phone +1 301-572-0446; Fax +1 301-572-099; Email: Clara.Burgert@icfi.com

Acknowledgment: The authors would like to acknowledge the following people and organizations for their assistance: the Uganda DHS 2011 country data collect teams and partner organizations, Mark Janko, Jennifer Winston, Shuaiqing Liu, and Blake Zachary for their contributions to earlier drafts of this report, and Livia Montana, Thea Roy, and Mahmoud Elkasabi for reviewing earlier versions of this report.

Editor: Sidney Moore
Document Production: Chris Gramer

This study was carried out with support provided by the United States Agency for International Development (USAID) through the MEASURE DHS project (#GPO-C-00-08-00008-00). The views expressed are those of the authors and do not necessarily reflect the views of USAID or the United States Government.

Recommended citation:

Perez-Heydrich, Carolina, Joshua L. Warren, Clara R. Burgert, and Michael E. Emch. 2013. Guidelines on the Use of DHS GPS Data. Spatial Analysis Reports No. 8. Calverton, Maryland, USA: ICF International.

Contents

List of Tables

List of Figures

List of Appendices

Preface

One of the most significant contributions of the Demographic and Health Surveys (DHS) program since its initiation in 1984 is the creation of an internationally comparable body of data on the demographic and health characteristics of populations in developing countries. These data have been augmented in recent years by the addition of more spatial data in the datasets.

The *DHS Spatial Analysis* series joins the existing DHS comparative and analytical report series to meet the growing interest and use of demographic and health data in a spatial realm. The principal objectives of all DHS report series are to provide information for policy formulation at the international level and to examine individual country results in an international context.

Studies in the *DHS Spatial Analysis* series are based on a variable number of data sets, depending on the topic being examined. A range of methodologies are used in these studies, including geostatistical and multivariate statistical techniques. The topics covered are selected by DHS staff in consultation with the U.S. Agency for International Development.

It is anticipated that the DHS Spatial Analysis studies will enhance the understanding of analysts and policymakers regarding significant issues in the fields of international population and health and spatial analysis.

Sunita Kishor
Project Director

Executive Summary

Because the locations of Demographic and Health Survey (DHS) clusters are randomly displaced to protect the confidentiality of survey respondents, measurement error and misclassification bias are of major concern when spatially referenced DHS data is used to address health and population research questions. For instance, specific outcomes corresponding to DHS clusters may be related to predictors or covariates that are defined spatially. Common examples include exposure variables defined according to distance measures and exposure variables defined according to ancillary data. With this paper we explore the sensitivity of study results to spatial offsets associated with DHS data in three studies involving distance-based analyses, integration of ancillary raster data, and integration of areal or vector data. The paper provides guidelines on the use of DHS spatial data to reduce and/or account for measurement or misclassification errors.

We present three studies to frame potential problems associated with spatially offset data and develop guidelines for their appropriate usage. 1) The first study addresses how distance measures derived from offset data can result in biased effect estimates. We first quantify the bias across three scenarios: low, medium, and high density of destination points and/or lines to which distance measures from DHS clusters are derived, and then discuss how the bias differs across scenarios. Two case studies are presented for purposes of illustration: one on HIV testing and proximity to health centers, the other on the number of sexual partners and road access. 2) The second study addresses integration of ancillary data however the focus is on overlaying raster data, such as an environmental surface. We address how misspecification of overlaid covariates can bias resulting effect estimates. We first quantify the bias across three scenarios: low, medium, and high spatial autocorrelation of values within the ancillary dataset, and then discuss how the bias varies across scenarios. Additionally, we address how neighborhood definition affects results under each of these scenarios by comparing alternative methods for assigning raster values to DHS clusters (i.e., mean values across buffers of varying sizes). We then close the analysis with general guidelines on which neighborhood definition is most appropriate under the different raster scenarios. Two case studies are presented for purposes of illustration: one on anemia risk and helminth prevalence, the other on anemia risk and cropland cover. 3) The third study is similar to the second in that it addresses how integration of ancillary areal data can result in misspecification of overlaid covariates and consequently result in biased effect estimates. We first quantify the probability of misspecification across all points, and then provide probability-based methods to generate weighted covariate estimates that are expected to reduce the impact of misspecification when it occurs. We close the analysis of this study as we did the previous two studies with general guidelines on the most appropriate methods to use when linking areal data to spatially offset DHS cluster locations. One case study is presented for purposes of illustration: neighborhood determinants of HIV knowledge.

In each of the studies, we present a simulation analysis to motivate and justify the proposed guidelines. Following the simulation studies, we test the performance of the guidelines using non-displaced

and displaced DHS data. It is important to note that the application of these guidelines to DHS data is not meant to be rigorous analysis of particular research questions. Rather, the applications are meant to serve illustrative purposes, i.e., to demonstrate how well guidelines established according to empirical results of simulation studies perform in realistic applications of DHS GPS data.

Based on examination of these studies, we propose general guidelines to encourage appropriate use of DHS spatial data while considering how bias associated with DHS cluster offsets can impact study results. Our proposed guidelines include both statistical and non-statistical components. The non-statistical guidelines involve defining appropriate scales of analysis with which study questions can be addressed that will reduce the bias associated with using spatially offset data. These non-statistical guidelines are defined according to results from sensitivity analyses that compare the bias in effect estimates across multiple scales of analysis. The idea behind running sensitivity analyses across multiple scales is to show how reconsidering the spatial scale of the research question can circumvent potential problems associated with drawing inferences from analyses using spatially offset DHS data.

By refocusing the research question to a more appropriate scale of analysis, the bias associated with point displacement can be minimized. If investigators are reluctant to re-scale their research questions, we propose statistical alternatives that can be readily adopted by investigators to minimize the bias associated with point displacements. Our statistical guidelines involve the use of specific techniques, such as probability weights and/or alternative modeling strategies, for which we provide well-documented R code that can be readily reproduced.

Chapter 1

DHS GPS Data Displacement

The Demographic and Health Survey (DHS) data has geo-located survey locations dating to 1986. The early data were focused in West Africa and was gathered though the use of gazetteers. The first Global Positioning System (GPS) data was collected in Benin in 1996. The collection of GPS locations for surveys has become fairly standard practice since the early 2000s. Currently, there are over 120 surveys with GPS or geo-located data. To protect the confidentiality of respondents the geo-located data is displaced (Burgert et al., 2013). The displacement process moves the latitude and longitude to a new location under set parameters. Urban locations are displaced 0-2 kilometers while rural locations are displaced 0-5 kilometers with 1% (or every 100^{th} point) displaced 0-10 kilometers. The displacement is a random direction/random distance process. The steps in the displacement are: (1) A random direction (angle) between 0 and 360 degrees is chosen; (2) A random distance according to the urban and rural parameters is chosen; (3) The new location is created combining steps 1 and 2 to create a new latitude and longitude for the cluster; and (4) The new location is checked to make sure it falls within designated administrative boundaries. In surveys after 2008 this is usually administrative 2 boundaries while surveys before 2008 were restricted to DHS regional boundaries or national boundaries. The DHS GIS team uses a python script in ArcGIS to displace the data within the appropriate administrative boundaries.

For purposes of this analysis, we used the Uganda DHS 2011 data, including the GPS data for the 404 clusters covered in the survey. GPS points represented the approximate center of a cluster of households. The data were verified by MEASURE DHS staff to be in the proper administrative areas; 7 points were gazetted to the nearest village. A total of 400 clusters were verified and 4 were listed as missing GPS data and given {0,0} as the latitude/longitude coordinate. The verified cluster locations were displaced according to DHS protocol restricting the displacement to the first administrative level.

Multiple displaced datasets needed to be created for the simulations. To facilitate this process a displacement function was created in R that mirrored the DHS displacement python script in ArcGIS. The codes are listed in the Appendix (Section B.1). The R displacement script began by assigning maximum potential offset amounts to each original DHS location in the dataset with urban locations receiving a value of 2km and rural locations initially receiving a value of 5km. The value of every 100^{th} rural location was then changed to 10km based on the current DHS python script. Next, we created a buffer around each location using the specified maximum offset distance as the radius of the buffer. We then randomly generated 100,000 points within the buffer, thoroughly filling in the empty space. The process of creating the offset location was then completed by randomly selecting a single point from the generated points within the buffer.

However, selecting from each point with equal probability leads to a randomly selected location with respect to the angle of offset, but does not lead to random uniform distances. This is due to the fact that locations in the outer part of the buffer have a higher probability of being selected because the area of that region is greater than the area closer to the center. As a result, we were more likely to select larger distances under an equal probability sampling scheme. To account for this, we weighted the probability of selection based on distance from the buffer centroid. Selecting the offset location using these probabilities led to a randomly selected offset angle and a random uniform offset distance as desired. The process of selecting a random point from the randomly generated 100,000 points was repeated 100 times for each location in the dataset. In other words, for each location (i.e., DHS cluster) we simulated 100 displaced points, and these points were used in subsequent simulation case studies (Figure 1.1).

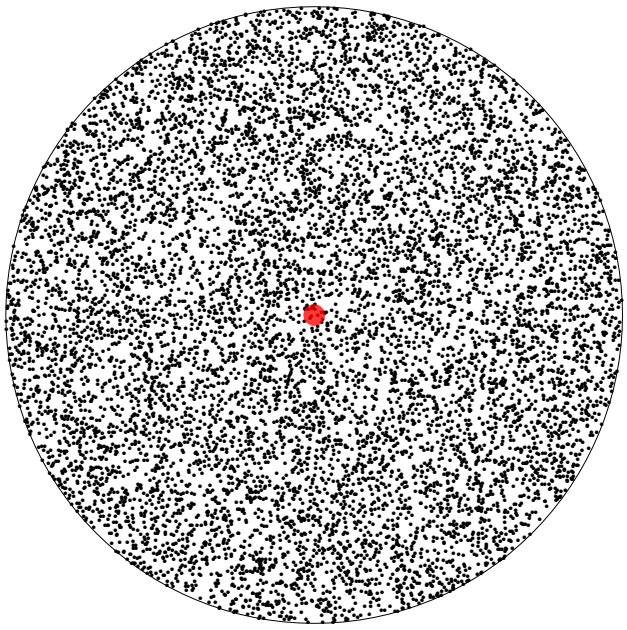


Figure 1.1: Schematic of randomly generated 100,000 points from which a sample of 100 displaced locations were used in subsequent simulation studies. The central point in red represents the true DHS cluster point, while the small black dots that comprehensively fill the circular radius around the true point represent randomly displaced locations.

Chapter 2

Background and Motivation

Theoretical concerns about error in spatial data are not new, and much work addressing them has emerged in the past few decades. As Goodchild, Guoqing, and Shiren (1992) note: "The contents of a geographical information system's database are almost without exception approximations to real geographical variation. To estimate the uncertainty associated with a given product (e.g. its confidence limits), it is necessary to (a) model the uncertainty existing in the database and (b) model the propagation of uncertainty through the operations performed on the data by the system. Neither of these requirements is trivial." This section, therefore, provides a summary of these issues, beginning with the broader spatial context, and then turns to ways in which researchers have sought to understand spatial errors and their effects on outcomes; particular attention is given to research related to our three case studies.

For geographic data associated with DHS clusters, spatial uncertainty arises from the random displacement of data to preserve the privacy of survey respondents. One outcome of this procedure, however, is the well-documented Modifiable Areal Unit Problem, in which statistical bias is introduced during the process of analyzing point-based spatial data at an aggregate areal level (Openshaw, 1984). Another problem that has particular relevance for our case studies, is the recently described as the Uncertain Geographical Context Problem (UGCoP), in which statistical bias is introduced when the method of delineating contextual areas/neighborhoods affects the results of an analysis (Kwan, 2012). For example, in the case of UGCoP, when buffers around clusters are used to generate an ecological variable (such as percentage of crop land or degree of exposure to traffic pollution), the buffer must be large enough to incorporate the true area associated with the cluster locations prior to displacement, which complicates the already difficult task of defining the proper contextual area/neighborhood associated with a cluster.

Concerns about these issues of spatial displacement and the recognition of specific sources of spatial error have been the subject of much work over the past few decades (see, for example, Goodchild and Gopal 1989). Most models have historically employed either an analytical approach or Monte Carlo simulation to address error propagation. The first employs mathematical techniques such as Taylor series expansions to characterize error, while the second models positional error based on a simulated sampling of inputs whose errors behave according to a known probability distribution (Foody and Atkinson, 2002; Hengl et al., 2010; Heuvelink et al., 1989). Heuvelink et al. (2007) note that Monte Carlo simulation techniques have essentially taken over the field of error propagation modeling. Nonetheless, other, less sophisticated techniques that involve model comparisons and the use of ancillary data have been incorporated into health studies. An example of concern over the latter was seen in discussion that arose when areal data with

centroid points were used in distance calculations (Hillsman and Rhoda, 1978). Efforts to address the bias problem have included weighting the centroid by the distribution of the population within each area, or using the centroids of smaller areal units, for example by using census blocks instead of tracts (Griffith, 1989). Such methods are either not possible or are inadvisable in the case of DHS data. First, finer spatial data often do not exist in DHS country settings, and second, while aggregation to smaller administrative units may be possible technically (via spatial join procedures), cluster displacement can result in clusters being assigned to the wrong administrative units.

Other efforts to characterize the bias introduced by positional error have used comparison as a tool for analysis. This method is seen particularly in studies attempting to understand the effects of environmental exposure on a variety of health outcomes. Zandbergen (2007) and Zandbergen and Green (2007), for example, compare methods of street geocoding with modeling of children's exposure to traffic pollution. Their results found bias and error in proximity analyses of distances less than 500 meters, with consistent overestimation of exposed children. Whitsel, Quibrera, Smith, Catellier, Liao, Henley, and Heiss (Whitsel et al.) also found exposure misclassifications in their study on the accuracy of commercial geocoding techniques. A study conducted by Ward et al. (2005) of non-Hodgkins lymphoma compared two geocoding methods to characterize the positional error and test the sensitivity and specificity of each to crop occurrence within 500m, 250m, and 100m of both sets of geocoded households. They found that geocoding errors affected crop exposure classification at 100 meters. Each of these studies shows that the spatial scale of the analysis is an important consideration, analogous to concerns related to the use of ancillary raster data with displaced DHS clusters.

A related concern involves studies that extract data from raster surfaces or polygons using map overlay techniques. Zandbergen et al. (2012) examine the effect of geocoding error on association with 30-meter resolution land cover by generating an error matrix to determine the agreement between the results for reference locations and geocoded locations in six US counties. They found that areas with relatively homogenous land cover resulted in fewer errors in matching points with the correct land cover type, whereas areas with heterogeneous land cover types were associated with larger error. One solution they offer is, if possible, to reclassify areas with heterogeneous land cover types into fewer categories. With regard to incorrect placement of points within polygons, Strickland et al. (2007) conducted a simulation study to estimate the quantity of addresses improperly assigned to counties as the result of geocoding error, by randomly displacing addresses according to the known distribution of the geocoding error. Their findings showed that approximately 5% of addresses were assigned to the wrong county. The study concluded that ancillary data (such as tax parcels) can be used to correct positional error. In DHS country settings, however, such high resolution and precision are rare for ancillary data.

While these studies provide a useful framework for understanding how positional error can affect analytical results, they do not provide a statistical method for correcting the problem. This is seen particularly where proposed solutions to the problem involve incorporating ancillary data to correct for the positional error. With DHS data, correction is often not possible because a) the ancillary data do not exist, and b) privacy issues preclude the use of ancillary data for this purpose. Recent work in environmental statistics, however, has made progress on this front, specifically in dealing with issues of spatially misaligned data. Gryparis et al. (2009) developed a framework for spatial error modeling that includes Bayesian models and out-of-sample regression calibrations to correct bias in exposure data extracted from interpolated surfaces. Berry et al. (2002) employed

Bayesian smoothing and regression splines to address similar problems, while Madsen et al. (2008) employed parametric bootstrapping techniques to acquire better exposure estimates. Because this approach is computationally intensive, Szpiro et al. (2010) developed an alternative parameter bootstrap to correct for the measurement error.

These are just a few examples of error propagation models that have been developed recently. It is important to note, however, that most of these models were developed within a research environment, with minimal efforts expended to develop user-friendly statistical tools such that enable typical users of spatial data to incorporate spatial uncertainty into their models. Indeed, as Heuvelink et al. (2007) point out, the past three decades have seen considerable effort devoted to understanding the effects of positional error and its propagation through analyses. Nevertheless, incorporating error propagation analysis remains the exception rather than the rule. The authors offer four reasons for this situation. First, in order to model the effects of spatial error propagation, it is necessary to know the error in the spatial inputs, and this is often not available because of poor documentation of the accuracy of spatial data. Second, spatial error propagation analysis requires considerable computational time and efforts, particularly when implementing simulation techniques such as Monte Carlo simulations. This adds to the cost of an analysis. Third, error propagation analysis requires statistical knowledge that most mainstream spatial analysts do not have. Fourth, and perhaps most important for the authors of this report, mainstream users of GIS data lack interest in error propagation, whether due to greater interests in the subject matter (as opposed to analytical concerns), or to lack of familiarity with the underlying mathematics of uncertainty. The authors, therefore, are calling not only for further research in error propagation modeling, but also for the development of tools that will aid the broader community of spatial data users in incorporating error propagation into research studies. An initial effort was made with the development of the Data Uncertainty Engine, which allows users to incorporate spatial uncertainty into their analyses (Brown and Heuvelink, 2007).

A wealth of literature has used DHS spatial data to visualize disease patterns, assign environmental or contextual exposures, estimate access to health services, or validate prevalence estimates made by other surveillance systems. Table 2.1 summarizes and categorizes published studies that have made use of DHS GPS data. One of the most common uses of DHS spatial data involves linking clusters to environmental or contextual data in order to generate new covariates of interest. However, as noted in the previous section, uncertainty associated with displaced locations can have deleterious impacts on the quality of inferences drawn from a study; uncertainties in data will lead to uncertainties in analytic results. In this report, we address the importance of spatial uncertainty in drawing inferences from analyses involving spatial DHS data. Because geographic points need to be offset for ethical considerations, namely, to protect the confidentiality of residents, proper use of DHS spatial data should account for the baseline uncertainty surrounding survey cluster locations (see (Mansour et al., 2012)). Here, we use empirical studies to investigate the impacts of spatially offset point locations on inferences that are drawn from the data. Specifically, we focus on three main mechanisms of linking ancillary spatial data to DHS GPS data: 1) spatially linking DHS clusters to continuous and categorical raster data, 2) spatially linking DHS clusters to areal data, and 3) calculating distances to resources. Through implementation of simulation studies, followed by case studies using DHS data, we determine how point displacement affects the specification of spatially linked covariates; additionally,we look at how this specification can affect the interpretation of analytic results. Based on these empirical findings, we then develop general guidelines on the use of DHS GPS data in order to mitigate potential biases associated with point displacement. In the following sections, we explore the sensitivity of study results that do not use comparison

Spatial use of DHS data	**Studies**
Linking geographic location and environmental variables	Balk et al (2004), Simler (2006), Baschieri (2007), Pande et al (2008), Feldacker et al (2010), Jankowska et al (2010), Messina et al (2010), Mansour et al (2011), Messina et al (2011), Bandyopadhyay et al. (2012), de Castro and Fisher (2012)
Estimating distance to resources and/or health care services	Montana et al (2000), Hong et al (2006), Noor et al (2009), Walker and Vajjhala (2009), Gabrysch et al (2011), Blanford et al. (2012), Kashima et al (2012), Pickering and Davis (2012)
Validating population or health estimates from other sources	Montana et al (2008), Gonese et al (2010), Johnson et al (2010), Thuilliez (2010)
Identifying mechanisms driving the spatial distribution of a characteristic of interest	Kandala et al (2006), Subramanian et al. (2006), Kazembe et al (2007), Uthman (2008), Khatab and Fahmeir (2009), Cuadros et al (2010), Giardina et al. (2012), Gosoniu et al. (2012), Kandala et al (2010)
Spatial interpolation/imputation	Gosoniu et al (2010), Lamarange et al (2011), Taylor et al (2011), Messina et al (in press)
Detecting or identifying spatial clusters	Kandala et al (2005), Kandala et al (2007), Messina et al (2010), Chin et al (2011), Dake et al (2011), Go et al (2011), Magalhaes et al (2011), Pawloski et al. (2012), Cuadros et al. (2013)

Table 2.1: Review of literature pertaining to usage of DHS GPS data.

to account for spatial uncertainty in locations, that is comparing inferences drawn from analyses using offset points with inferences drawn from analyses using true locations. In a series of simulation studies, we illustrate how consideration of baseline uncertainty in point locations corresponding to DHS survey clusters can influence the validity of inferences drawn from studies. After demonstrating the impact of using spatially misaligned data, we provide recommendations on the use of spatial DHS data that should minimize the reporting of potentially biased results.

Chapter 3

Influence of Offsets on Distance-based Analyses

3.1 Goals of Simulation Study

We propose a simulation study to analyze how the random offset of DHS cluster locations affects the resulting statistical analysis and the conclusions drawn from the analysis. Specifically, we focus on the situation in which distance from the DHS cluster to the closest resource location is used as the main covariate of interest in the study. These resource locations include points such as health resource facilities and line segments such as road networks. To date, these distances have been calculated from the offset DHS cluster locations since the true DHS cluster locations are not available for analysis. This results in measurement error in the observed distance covariate (distance to closest resource location). In the presented simulation study, we investigate the impact this measurement error has on the statistical inference across different spatial density settings of the resources. These spatial density settings range from dense (resources located in numerous locations across the geographic domain) to sparse (very few resources located across the geographic domain), with five total settings selected.

We are also interested in determining how the random offset affects a cluster's linkage to the actual closest (based on Euclidean distance) resource location when distances are calculated using the offset cluster locations. It is common to use health resource level treatment information to determine the level of care residents of a particular cluster likely received. For offset clusters however, the observed closest resource location may not be the true closest resource location of interest, leading to incorrect treatment assignment information. We explore how the offset affects this assignment across the spatial density settings.

We carry out each of the proposed analyses for studies which consider only rural DHS locations, only urban DHS locations, and a mixture of location types. Separate analyses are required due to the different level of measurement error introduced to the different location types under the DHS guidelines. The true Uganda DHS cluster locations (represented by points), health resource facilities (points), and road networks (lines) are used in the simulation study to ensure that the results are realistic and useful for future research efforts.

We also note that while researchers are commonly calculating distances from the DHS cluster location point to a resource of interest and using the derived continuous measurement as a covariate in the analysis, categorical distances could also be used in subsequent analyses. Using categorical

distances could potentially reduce the effect of using the displaced DHS cluster location though some error would still remain. This categorical distance-based analysis was not considered here but may be of interest in future studies.

3.2 Distance-based Covariate Simulation Study

3.2.1 Methods

In the simulation study, we generate datasets for analysis, collecting information during each analysis which aid in answering the proposed questions of interest. In order to generate a single dataset, we begin by using the true (non-displaced) Uganda DHS cluster locations along with the resource locations of interest, in this case, health resource facilities (points) and road networks (lines) across Uganda. Next, we randomly offset these true DHS cluster locations based on DHS guidelines presented in Appendix B. These offset DHS cluster locations represent the information available to all researchers. We then calculate distances from the true DHS cluster locations to the nearest resource locations and calculate similar distances using the observed (displaced) DHS cluster locations. Next, we choose regression coefficient values for the Poisson statistical regression model considered in the study such that $\beta_0 = 2.50$ and $\beta_1 = -0.10$. These values represent similar values estimated using DHS data and distance-based covariates (Feldacker et al. 2010; Lohela et al. 2012). Using these $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ parameters, we are able to generate a dataset of Poisson distributed data, one data point for each DHS cluster. The proposed model is given as $Y_i|\lambda_i \overset{ind}{\sim} \text{Poisson}(\lambda_i)$ where $\ln(\lambda_i) = \beta_0 + \beta_1 * x_i^{(t)}$ and $x_i^{(t)}$ is the distance from the true location of DHS cluster i to its closest resource location. Recall that this distance covariate is unknown by the researchers. To create a new spatial density setting, we randomly remove a portion of the available resource locations. Spatial density 1 begins with 2,596 health resource facilities and 1,585 road segments. We then remove 25%, 50%, and 75% for spatial density settings 2, 3, and 4, respectively. Spatial density setting 5 has only 1% of the original resource locations. We repeat this data generation process while subsetting the true Uganda DHS cluster locations to rural locations alone and then again using urban locations alone. We generate 100 independent datasets by repeating the process under each spatial density setting for a total of 500 datasets.

We propose three methods to analyze the generated datasets of interest; they are the following:

- Method 1 (MLE): Maximum likelihood estimation of the data using the true distances to the closest health resource ($x_i^{(t)}$) (unknown to researcher),
- Method 2 (Naive): Maximum likelihood estimation of the data using the observed distances to the closest health resource ($x_i^{(0)}$) (available to researcher), and
- Method 3 (RC): Maximum likelihood estimation of the data using the estimated true distances to the closest health resource ($\widehat{x}_i^{(t)}$) (regression calibration).

Method 1 represents the optimal analysis, based on the true distances of interest. Use of Method 1 is not possible in the current studies because access to the true DHS locations that are needed to calculate the true distances is not available to the public. Method 2 represents the naive analysis which is most commonly used in previous studies. This analysis fits the correct statistical model, similar to Method 1, with the incorrect distance covariate $x_i^{(0)}$. Method 3 fits the same statistical model as methods 1 and 2, while using estimates of the true distances $\widehat{x}_i^{(t)}$. These estimates are

obtained using the regression calibration technique (Carroll and Stefanski, 1990) which is available in the Stata statistical software package (Hardin et al., 2003) and can also be implemented by hand in other statistical software programs (see Appendix C for more details). Method 3 can be used by all researchers and represents a compromise between methods 1 and 2.

We fit each of the three methods to a single generated dataset and collect information from each of the model fits. First, we collect the estimate of β_1, $\widehat{\beta_1}^{(j)}$, for each method $j = 1, 2, 3$. β_1 represents the main parameter of interest in the study because it describes the association between the distance to closest resource and the outcome. Collection of $\widehat{\beta_1}^{(j)}$ allows us to estimate the bias of the estimator from each method, $\mathrm{E}\left[\widehat{\beta_1}^{(j)}\right] - \beta_1$, and the mean squared error (MSE),

$$\mathrm{E}\left[\left(\widehat{\beta_1}^{(j)} - \beta_1\right)^2\right]$$

where E(.) represents the expected value of a random variable. We also collect the absolute difference in the observed distance and true distance for each DHS cluster and average these values such that

$$\frac{1}{n}\sum_{i=1}^{n}\left|x_i^{(0)} - x_i^{(t)}\right|$$

where n is the number of DHS clusters in the analysis. Along with this information we collect the relative absolute difference in the observed distance and true distance for each DHS cluster and average these values such that

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\left|x_i^{(0)} - x_i^{(t)}\right|}{\overline{x}^{(t)}}$$

where $\overline{x}^{(t)} = \frac{1}{n}\sum_{i=1}^{n} x_i^{(t)}$ is the average true distance to the closest resource location. The relative absolute difference metric gives us an idea if the difference in the distances is meaningful in magnitude or relatively negligible. We collect these data for the estimated distances as well, replacing $x_i^{(0)}$ with $\widehat{x}_i^{(t)}$ in the above equations. Lastly, for each cluster we determine if it was correctly assigned to the true closest resource location. We then average these binary responses to estimate the proportion of DHS clusters that are correctly assigned. This average is collected for each generated dataset.

We use standard mixed statistical models which account for the fact that each method is applied to the same dataset through use of random effects. We then test whether the bias of the estimator associated with each method is significantly different from zero and whether this bias changes across spatial density setting. We also formally compare the MSE values for each method across all spatial density settings. These analyses are repeated separately for the rural and urban location studies.

3.2.2 Results: Point Resource Locations

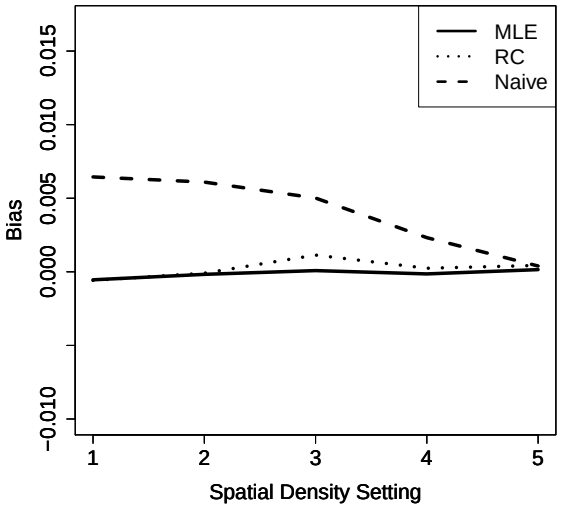
Figure 3.1 and Table 3.1 displays the bias analysis results for each of the methods and location specific analyses. As expected, Method 1 provides statistically significant unbiased estimates of the true β_1 parameter at each spatial density setting for each analysis. Method 3 is statistically significantly unbiased at more spatial density settings than Method 2 for all locations and rural only locations. For urban locations alone, Method 2 and Method 3 perform very similarly. As the

spatial density of the resources decreases, the bias decreases for methods 2 and 3. This result is fairly consistent across each of the location specific analyses. Method 3 is clearly preferred over Method 2 in terms of bias for all locations and rural only locations.

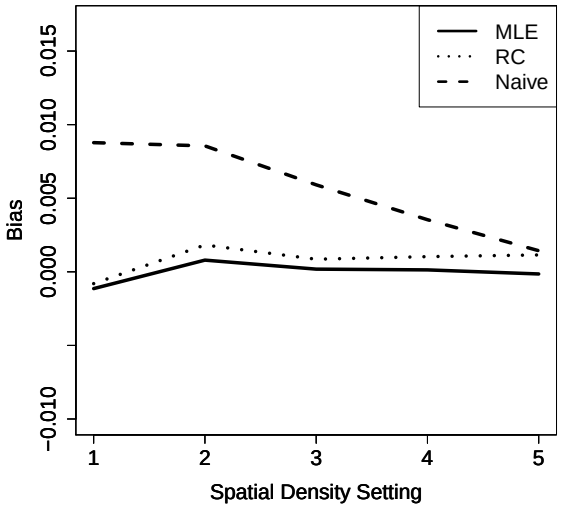
Figure 3.2 and Table 3.2 display the MSE analysis results. Method 2 produces MSEs which are statistically larger than methods 1 and 3 for spatial density settings 1-3 for all locations and rural only locations. However, for the urban only locations, methods 1 and 2 produce similar MSE estimates across spatial density settings 2-5. At spatial density setting 1, the MSE for Method 3 is significantly larger than Method 1 and Method 2. Once again, as the spatial density of the resources decreases, the MSE from methods 2 and 3 become more similar to the MSE of the Method 1 estimator. Overall, Method 3 performs as well or better than Method 2 at all settings other than urban locations only, spatial density setting 1.

Figure A.1 (in the appendix) displays the average absolute difference analysis results. As the spatial density of the health resources decreases the amount of measurement error in the distance covariate increases. Figure A.2 displays the average relative difference analysis results. As the spatial density of the health resources decreases the amount of relative measurement error actually decreases. Though the absolute change increases, this amount of change is less meaningful as the distances in general increase when the spatial density decreases. These results explain why the bias and MSE results seem to improve as the spatial density decreases.

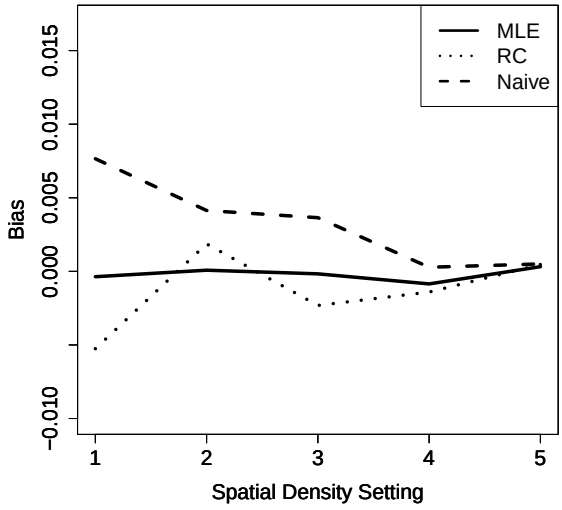
Figure A.3 displays the proportion of DHS clusters that were linked to their actual closest health resources (based on non-displaced locations) across each spatial density setting. Clearly, as the density of health resources decreases, the probability that a DHS cluster is linked to the correct closest health resource increases for all location types.



(a) All Locations.



(b) Rural Locations.

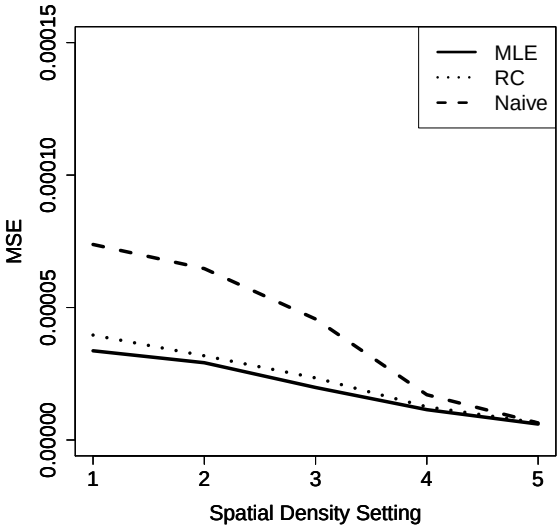


(c) Urban Locations.

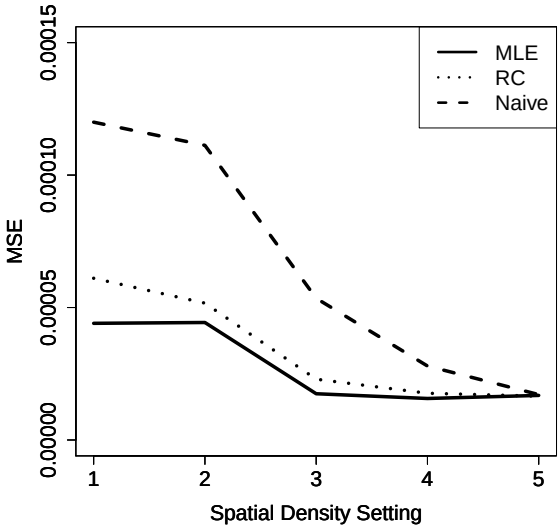
Figure 3.1: Bias Results for Point Resource Locations (Shown on Same Scale).

Table 3.1: Bias Results for Point Resource Locations: Not Statistically Different from Zero

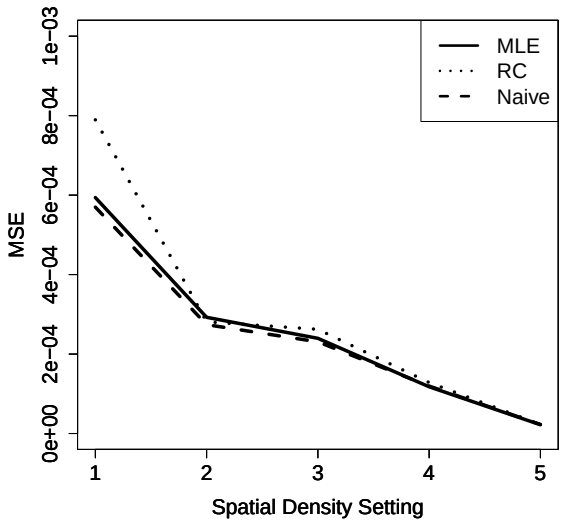
Method	Spatial Density Setting				
All Locations	1	2	3	4	5
1: MLE	X	X	X	X	X
2: Naive					X
3: RC	X	X	X	X	X
Rural Locations	1	2	3	4	5
1: MLE	X	X	X	X	X
2: Naive					X
3: RC	X		X	X	X
Urban Locations	1	2	3	4	5
1: MLE	X	X	X	X	X
2: Naive		X	X	X	X
3: RC		X	X	X	X



(a) All Locations.



(b) Rural Locations.



(c) Urban Locations.

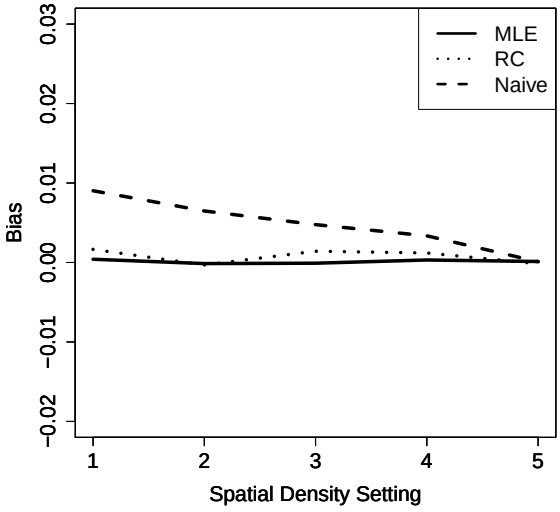
Figure 3.2: MSE Results for Point Resource Locations (NOT Shown on Same Scale).

Table 3.2: MSE for Point Resource Locations: Statistically Different Pairings (Spatial Density Settings)

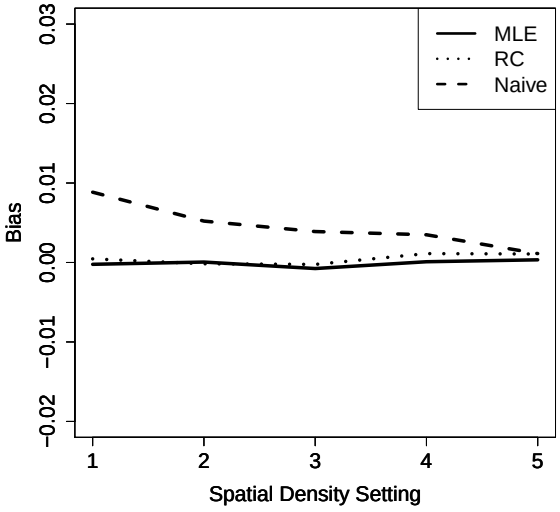
Method	Spatial Density Setting		
All Locations	1	2	3
1: MLE	—	1-3	None
2: Naive	—	—	1-3
3: RC	—	—	—
Rural Locations	1	2	3
1: MLE	—	1-3	None
2: Naive	—	—	1-3
3: RC	—	—	—
Urban Locations	1	2	3
1: MLE	—	None	1
2: Naive	—	—	1
3: RC	—	—	—

3.2.3 Results: Line Resource Locations

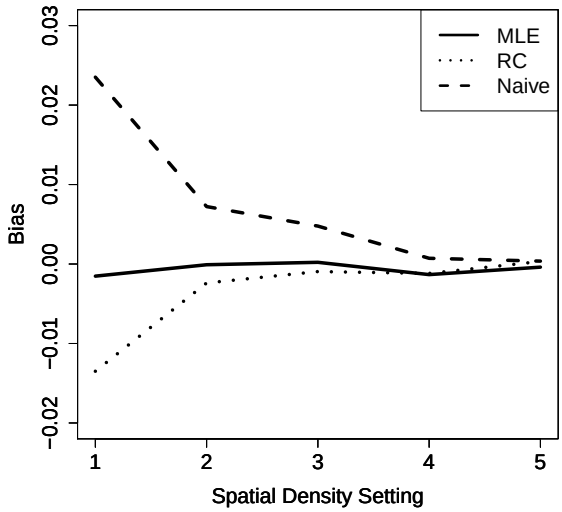
Figures 3.3 and 3.4 and Tables 3.3 and 3.4 show the bias and MSE results for the line resource locations. Figures A.4, A.5, and A.6 display the absolute difference, relative difference, and proportion of correctly linked resources results. These line resource results are very similar to the point resource results and overall suggest that Method 3 is preferred over Method 2 in most situations.



(a) All Locations.



(b) Rural Locations.

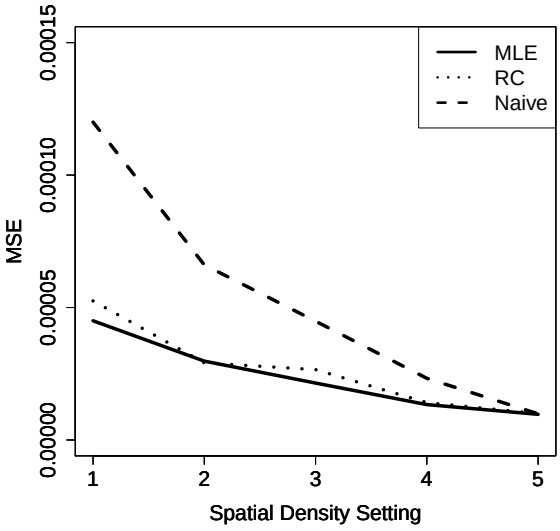


(c) Urban Locations.

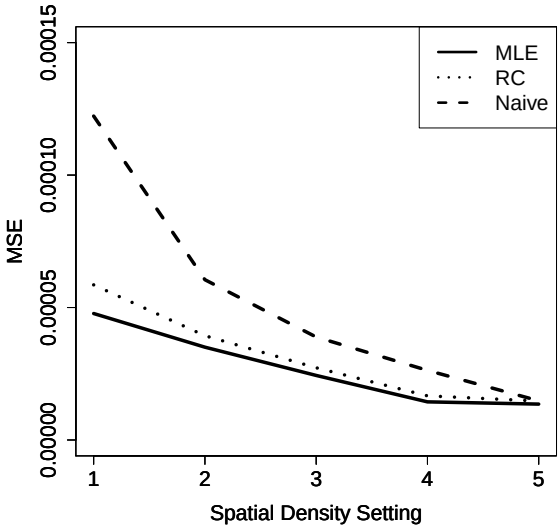
Figure 3.3: Bias Results for the Line Resource Locations (Shown on Same Scale).

Table 3.3: Bias Results for Line Resource Locations: Not Statistically Different from Zero

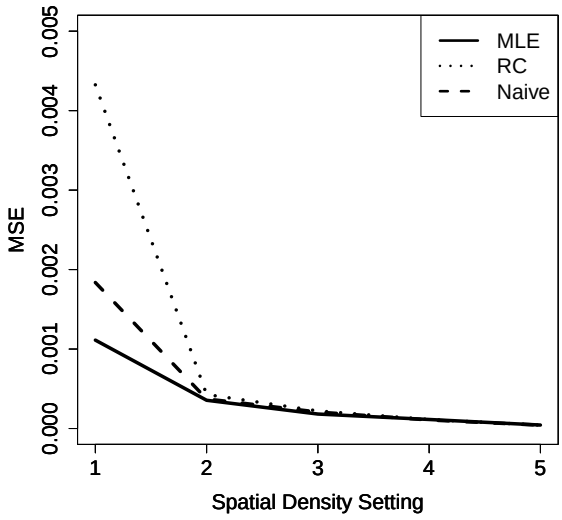
Method	Spatial Density Setting				
All Locations	1	2	3	4	5
1: MLE	X	X	X	X	X
2: Naive					X
3: RC		X	X	X	X
Rural Locations	1	2	3	4	5
1: MLE	X	X	X	X	X
2: Naive					X
3: RC	X	X	X	X	X
Urban Locations	1	2	3	4	5
1: MLE	X	X	X	X	X
2: Naive			X	X	X
3: RC		X	X	X	X



(a) All Locations.



(b) Rural Locations.



(c) Urban Locations.

Figure 3.4: MSE Results for Line Resource Locations (NOT Shown on Same Scale).

Table 3.4: MSE for Line Resource Locations: Statistically Different Pairings (Spatial Density Settings)

Method	Spatial Density Setting		
All Locations	1	2	3
1: MLE	—	1-3	None
2: Naive	—	—	1-3
3: RC	—	—	—
Rural Locations	1	2	3
1: MLE	—	1-2	None
2: Naive	—	—	1-2
3: RC	—	—	—
Urban Locations	1	2	3
1: MLE	—	1	1
2: Naive	—	—	1
3: RC	—	—	—

3.3 Proposed Guidelines

The results presented indicated that as the spatial density of health resources decreases, the amount of measurement error in the observed distance covariate increases. However, the relative amount of measurement error decreases, suggesting that the distances become much larger overall so that the increased amount of measurement error is negligible. This can be seen in the bias and MSE simulation results. The regression-calibration technique (Method 3) reduces the measurement error across all spatial density settings by attempting to estimate the true distance covariate of interest. However, when only urban locations are considered the naive (Method 2) and regression-calibration methods perform similarly due to the decreased level of measurement error in urban locations. The proportion of DHS clusters correctly assigned to their true closest health resource increases as the spatial density of health resources decreases. These results are consistent for point and line based analyses. Another potential approach investigators may consider is categorizing distance covariates into broader distance classes, which would be less precise than using an adjusted continuous covariate, but would also reduce measurement error in these covariates. Further work is needed to test this approach empirically. Overall, **we suggest that regression calibration should be used for distance based covariates**, particularly as the spatial density of health resources increases. Information from the closest resource facility location should only be used as a covariate in situations with less dense health resource locations. Tables 3.5, 3.6, and 3.7 display general guidelines for the point locations while Tables 3.8, 3.9, and 3.10 show similar guidelines for the line resource locations. These results can help determine the appropriate spatial density setting of an analysis and the suggested guideline. To use the presented tables, researchers simply calculate the distances from the available DHS clusters to the resource locations of interest. These distances are summarized using the sample mean, median, and standard deviation and compared with the appropriate tables based on the type of resources being considered (point or line).

Table 3.5: Point Location Guidelines: All Locations.

Observed Distances to Closest Health Resource (km)				
Spatial Density	Mean	Median	SD	Selected Method
1	4.15	3.01	3.97	3: RC
2	4.77	3.84	4.50	3: RC
3	5.76	4.47	5.58	3: RC
4	8.69	6.56	8.22	3: RC
5	64.27	48.96	59.72	2: Naive or 3: RC

Table 3.6: Point Location Guidelines: Rural Locations.

Observed Distances to Closest Health Resource (km)				
Spatial Density	Mean	Median	SD	Selected Method
1	5.13	4.04	4.10	3: RC
2	6.17	4.70	5.24	3: RC
3	7.83	6.30	6.32	3: RC
4	11.01	8.32	9.82	3: RC
5	47.73	39.61	31.34	2: Naive or 3: RC

Table 3.7: Point Location Guidelines: Urban Locations.

Observed Distances to Closest Health Resource (km)				
Spatial Density	Mean	Median	SD	Selected Method
1	1.58	1.15	1.42	2: Naive
2	2.24	1.35	2.25	2: Naive or 3: RC
3	2.23	1.46	2.13	2: Naive or 3: RC
4	3.43	1.65	4.34	2: Naive or 3: RC
5	26.50	11.89	32.16	2: Naive or 3: RC

Table 3.8: Line Location Guidelines: All Locations.

Observed Distances to Closest Health Resource (km)				
Spatial Density	Mean	Median	SD	Selected Method
1	2.73	1.46	3.90	3: RC
2	3.36	1.87	4.56	3: RC
3	4.66	2.40	6.44	3: RC
4	8.06	5.37	8.57	3: RC
5	57.56	58.93	30.13	2: Naive or 3: RC

Table 3.9: Line Location Guidelines: Rural Locations.

Observed Distances to Closest Health Resource (km)				
Spatial Density	Mean	Median	SD	Selected Method
1	3.68	2.32	4.66	3: RC
2	4.91	3.29	5.90	3: RC
3	6.29	4.03	7.22	3: RC
4	9.39	6.44	9.84	3: RC
5	89.26	60.92	73.08	2: Naive or 3: RC

Table 3.10: Line Locations Guidelines: Urban Locations.

Observed Distances to Closest Health Resource (km)				
Spatial Density	Mean	Median	SD	Selected Method
1	0.90	0.69	0.79	2:Naive
2	1.36	0.96	1.76	2: Naive or 3: RC
3	1.99	1.55	2.32	2: Naive or 3: RC
4	3.18	1.82	4.35	2: Naive or 3: RC
5	32.68	15.93	31.78	2: Naive or 3: RC

3.4 Case Study: HIV Testing and Proximity to Health Centers

The case studies presented here are for purposes of illustratiing the methods and guidelines discussed in previous sections. Therefore, the results should be used only to assess the performance of each of the methods in the given data setting.

3.4.1 Methods

The goal of this Uganda point resource case study was to determine whether the distance a woman travels to the closest available health resource center is associated with the likelihood of her ever being tested for HIV (Uganda DHS 2011, *V766B*). Health center resource locations from Uganda (World Health Organization, http://apps.who.int/geonetwork/srv/en/metadata.show?id=93& currTab=simple) were used to define the distance-based covariate used in the study. The outcome of interest, i.e., number of respondents in DHS clusters who had ever been tested for HIV, was obtained from the 2011 Uganda DHS.

Distances from each DHS cluster to the closest respective health center were calculated for the available clusters across Uganda based on actual cluster locations and DHS offset locations. The distances were then used in the statistical analysis that relates the DHS number of people who had ever been tested for HIV and the closest distance to health center. We hypothesized that DHS clusters located further from health centers would have lower number of people who had ever been tested for HIV on average, controlling for the population of the DHS cluster. A Poisson regression statistical model was used to analyze the association of interest while controlling for the population of each DHS cluster using a population offset term. All three methods considered in Section 3.2.1 were applied to the dataset and results are compared. Uganda DHS clusters from both urban and rural regions were included in the analysis.

Table 3.11: HIV testing and proximity to health center case study results.

Method	Estimate (SE)	Guidelines Choice
1: MLE	-0.007 (0.003)	NA
2: Naive	-0.009 (0.003)	
3: RC	-0.012 (0.004)	***

*** chosen method based on the presented guidelines.

3.4.2 Results

In the study, the observed distances to the closest health center were 4.03 km (mean) and 3.24 km (median) with a sample standard deviation of 3.97 km. These figures are similar to those of Spatial Density 1 in Table 3.5 (mean 4.15 km, median 3.01 km, and SD 3.97 km). Based on these results, we choose the regression calibration estimator (Method 3) due to the reduction in bias and MSE seen in this spatial density setting, suggesting that it is a superior estimator when compared with the naive version (Method 2). The results in Table 3.11 suggest that as the distance from a DHS cluster to the closest health center increases, the expected number of people within the cluster who have ever been tested for HIV decreases as expected. Access to health facilities appears to be an important factor in the probability of HIV testing among individuals within the Uganda DHS clusters. Overall, all methods performed similarly in terms of parameter estimates and standard errors. Note that we can not calculate bias and MSE in this setting because

we do not know the true value of the main parameter of interest. This parameter was not defined and can only be estimated in the case study.

3.4.3 Discussion

This analysis of HIV testing and proximity to health centers in Uganda was carried out for purposes of illustration; it is likely that the true effect of the distance to the closest available health resource center and the likelihood of ever being tested for HIV will be moderated by other unaccounted variables. The purpose of this case study was to demonstrate how well guidelines, which were established following empirical results of a simulation study, perform in a realistic application of DHS GPS data. The results suggest that the proposed guidelines performed well in practice, although all methods performed similarly in terms of statistical parameter estimation.

3.5 Case Study: Number of Sexual Partners and Road Access

3.5.1 Methods

The goal of this line segment resource case study was to determine whether the distance between a woman and the closest main road (dual lane highway) is associated with the likelihood of her having multiple sexual partners (including husband) within the past 12 months (Uganda DHS 2011, *V766B*). Main road maps from Uganda (United Nations Office for the Coordination of Humanitarian Affairs, http://cod.humanitarianresponse.info/dataset/uganda-roads) were used to define the distance-based covariate used in the study. The outcome of interest, i.e., number of women who had multiple sexual partners within the past month in DHS clusters, was obtained from the 2011 Uganda DHS.

Distances from each DHS cluster to the closest respective point on a main road were calculated for the available Uganda clusters across Uganda, based on actual cluster locations and DHS offset locations. These distances were then used in the statistical analysis which relates the number of women who had multiple sexual partners within the past 12 months and the closest distance to main road. We hypothesized that DHS clusters located further from main roads would have lower numbers of women who had multiple sexual partners within the past month, on average, controlling for the population of the DHS cluster. A Poisson regression statistical model was used to analyze the association of interest while controlling for the population of each DHS cluster using a population offset term. All three methods considered in 3.2.1 were applied to the dataset and the results were compared. Uganda DHS clusters from both urban and rural regions were included in the analysis.

Table 3.12: Number of sexual partners and road access case study results.

Method	Estimate (SE)	Guidelines Choice
1: MLE	-0.014 (0.004)	NA
2: Naive	-0.014 (0.004)	***
3: RC	-0.014 (0.003)	***

*** chosen method based on the presented guidelines

3.5.2 Results

The observed distances to the closest health resource were 27.91 km (mean) and 12.97 km (median) with a sample standard deviation of 38.28 km. This resembles Spatial Density 5 as detailed in Table 3.8. Based on these results, we can choose either the regression calibration estimator (Method 3) or the naive estimator (Method 2) due to their similar performance in bias and MSE seen in this spatial density setting. The results in Table 3.12 suggest that as the distance from a DHS cluster to the closest main road increases, the number of women in the cluster who had multiple sexual partners within the past month decreases, on average, as expected. Overall, all methods performed similarly in terms of parameter estimates and standard errors. Note that as with the point resource case study in Section 3.4, we can not calculate bias and MSE in this setting.

3.5.3 Discussion

This analysis number of sexual partners a woman has and road access in Uganda was carried out for purposes of illustration; it is likely that the true effect of a womans distance to the closest main road (highway) and the likelihood of her having multiple sexual partners (including husband) within the past month will be moderated by other unaccounted variables. The purpose of this case study was to demonstrate how well guidelines, which were established following empirical results of a simulation study, perform in a realistic application of DHS GPS data. The results suggest that the proposed guidelines performed well in practice, although all methods performed similarly in terms of statistical parameter estimation.

Chapter 4

Influence of Offsets on Raster-based Analyses

4.1 Goals of Simulation Studies

We propose simulation studies to analyze how the random offset of DHS cluster locations affect statistical inferences and conclusions drawn from analyses involving covariates generated from ancillary raster data. We address how covariates generated from continuous as well as categorical raster surfaces can be altered differentially by point displacement, and propose the use of buffer means to mitigate the potential bias associated with misspecification of covariates due to these random offsets. We evaluate the performance of these methods (i.e., buffer means) across raster surfaces with varying levels of spatial smoothness (i.e., spatial autocorrelation) and varying coverage of raster cell types (for categorical rasters). It is expected that raster characteristics such as these would likely influence the extent of bias brought about through point displacements, and thus the effectiveness of proposed neighborhood definitions (for obtaining buffer means) to reduce the bias associated with point displacement was evaluated across several simulated raster surfaces.

4.2 Generation of Raster Surfaces

4.2.1 Continuous Raster Surfaces

In both simulation studies (continuous and discrete raster), we require simulated raster surfaces in order to define the covariates and outcomes of interest. In both cases, continuous rasters are created first. These rasters are used directly in the continuous raster study and discretized for use in the categorical raster study.

Continuous raster surfaces were simulated to represent varying degrees of spatial smoothness (Figure 4.1). First, a regular grid of 65×65 points was generated to encompass the entire Uganda study area. For each point in the regular grid, neighboring points were identified within a 10 km radius of the point, and a row-standardized weights matrix ($\mathbf{W}$) was generated from the resulting neighbors list. Each grid point (i, j) for row i and column j was then assigned a value $Z_{i,j}$ which was dependent on the mean values of its neighbors. Specifically, the spatial autoregressive random vector $\mathbf{Z} = (Z_{1,1}, \ldots, Z_{65,65})^T$ was generated by: (1) constructing the 65×65 inverse matrix $\mathbf{V} = (\mathbf{I} - \rho\mathbf{W})^{-1}$ (), where ρ represents a predefined autoregressive parameter, and (2) defining the product $\mathbf{Z} = \mathbf{Vq}$, where $\mathbf{q}$ was a vector of independent standard normal random variables.

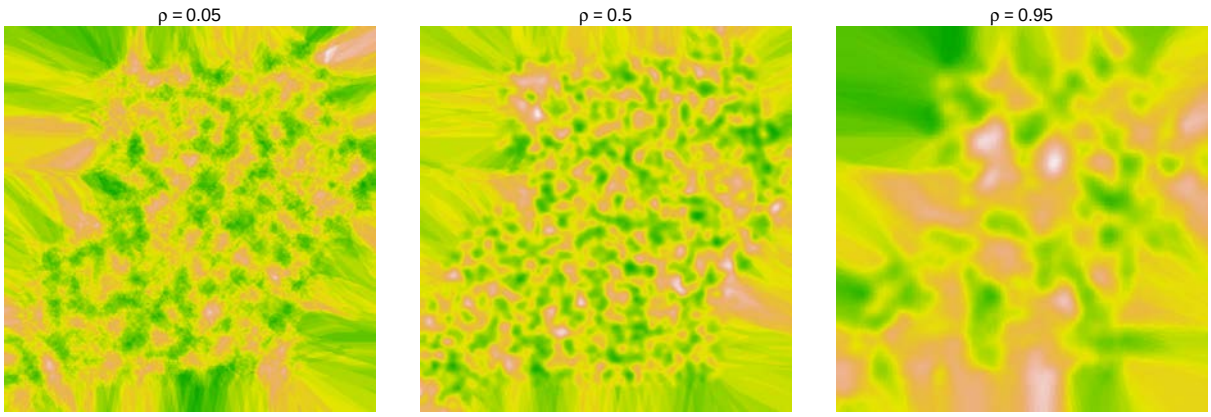


Figure 4.1: Continuous raster surfaces used in subsequent simulation studies. Each panel represents the surface generated assuming alternate definitions of ρ.

The resulting vector $\mathbf{Z}$ represented a spatially correlated multivariate normal random vector with mean equal to the zero vector and covariance equal to $\mathbf{V}\mathbf{V}^T = (\mathbf{I} - \rho\mathbf{W})^{-1}(\mathbf{I} - \rho\mathbf{W})^{-T}$. $\mathbf{Z}$ was defined for three different values of ρ, i.e., 0.05, 0.50, and 0.95. Using different ρ values, the spatial autoregressive point process could take on varying levels of smoothness such that at $\rho = 0.05$ the point process would exhibit very low levels of spatial autocorrelation, whereas at $\rho = 0.95$ the point process would exhibit very high levels of spatial autocorrelation. The generated gridded points for each level of ρ were then used to interpolate a continuous surface by using the kriging tool in ArcMap 10 with an output cell size of 500 m.

4.2.2 Categorical Raster Surfaces

Continuous rasters were discretized into categorical rasters that represented rare, moderately prevalent, and prevalent cell types (approximated 15, 30, and 45% coverage, respectively). We focused on discretizing the surface associated with $\rho = 0.95$ for subsequent analyses. Continuous raster values were converted to binary values by setting all grid cells with values less than those pertaining to the 15^{th}, 30^{th}, and 45^{th} percentiles, respectively, to one, and all other values to zero for rare, moderately prevalent, and prevalent cell types. Because of the nature of a binary raster surface, analyses associated with discrete raster data will more resemble those discussed in point-in-polygon approaches (Section 5.1), whereby cell misclassification due to point displacement is a main factor contributing to bias in effect size estimates. The more frequently points lie along boundaries of binary grid cell clusters, the higher the probability or rate of misclassification for displaced points. In this way, guidelines associated with the integration of data from ancillary discrete raster surfaces will be based on misclassification rates and probabilities.

For a given binary raster surface, the probability of misclassification for each displaced DHS cluster point can be calculated as one minus the proportion of the cells within the maximum displacement buffer that differ in value to that in which the displaced point lies. The true point will be located within this maximum buffer with probability equal to 1.0. Thus, if the maximum buffer around a displaced point falls within an area comprised completely of 1's, then there is a 100% chance that the true point is also associated with a raster cell value of 1. If, however, the maximum buffer of a displaced point whose observed cell value is 1 intersects several patches of different valued cells (i.e., 1's and 0's), then the misclassification probability for that point will be 1 minus the proportion of cells within that displacement buffer that equal 1, or the complement

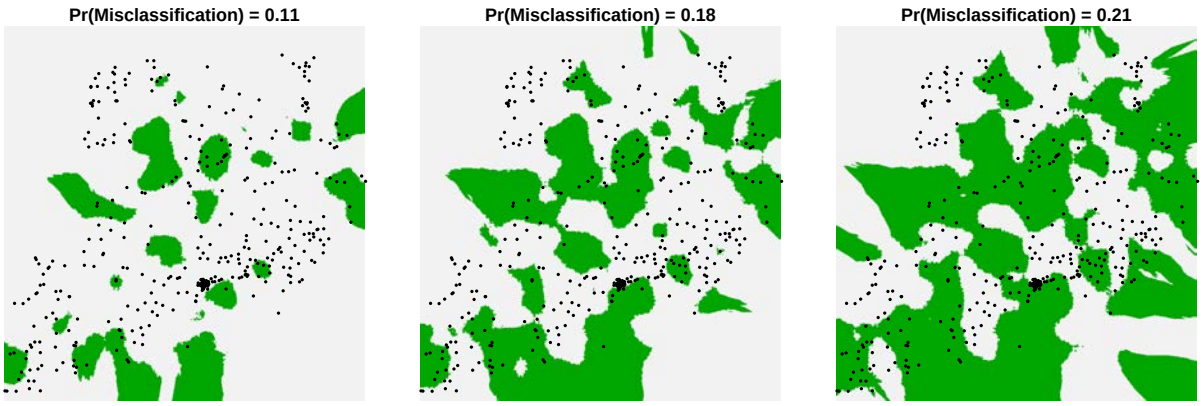


Figure 4.2: Categorical raster surfaces used in subsequent simulation studies. Each panel represents the surface generated assuming different misclassification rates.

of the proportion of like cells within the maximum displacement buffer. An `R` function to compute misclassification rates from discrete raster surfaces is provided in Appendix B.6. The resulting surfaces, along with the misclassification rates to which they pertain, are shown in Figure 4.2.

4.3 Continuous Raster Simulation Study

4.3.1 Methods

Generation of an Analysis Dataset

In the simulation study, we generate datasets for analysis, and collect covariate data during each analysis to aid in answering the proposed questions of interest. We define the true covariate of interest as the average of the continuous values within a 2 km buffer of the true DHS cluster location for urban clusters and 5 km for rural clusters. Next, we choose regression coefficient values for the Poisson statistical regression model considered in the study such that $\beta_0 = 1.00$ and $\beta_1 = -0.27$. Using these $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ parameters, we are able to generate a dataset of Poisson distributed data, one datapoint for each DHS cluster. The proposed model is given as $Y_i|\lambda_i \overset{ind}{\sim} \text{Poisson}(\lambda_i)$ where $\ln(\lambda_i) = \beta_0 + \beta_1 * x_i^{(t)}$ and $x_i^{(t)}$ is the average of the continuous values within the specified buffer created around the true location of DHS cluster i. Recall that this covariate is unknown by the researchers. We generate 100 independent datasets by repeating the process under each spatial smoothness setting for a total of 300 datasets. We repeat this data generation process for each level of considered raster cell type as well and analyze these results separately.

Noting that continuous raster surfaces will vary with regard to their respective scales of measurements for the data they represent, for this empirical study the covariate of interest was standardized to have a mean and variance of one in order to allow for the generation of more generalizable guidelines. Because the variability of a surface (i.e., how wide the spread of possible values spans), in addition to its smoothness can influence the magnitude of the estimated effect sizes and corresponding standard errors, standardization of covariates extracted from such surfaces allows for the sole consideration of spatial smoothness without loss of generalizability when developing guidelines. Thus, since guidelines here are developed based on standardized data, investigators should center and scale their covariate data accordingly when applying proposed guidelines from this study.

Analysis

We propose 14 methods to analyze the generated datasets of interest; they include the following:

- Method 1: Maximum likelihood estimation of the data using the true buffer average covariate based on the true DHS cluster location ($x_i^{(t)}$) (unknown to researcher),
- Method 2: Maximum likelihood estimation of the data using the point extracted cell value based on the offset DHS cluster location ($x_i^{(0)}$) (naive analysis), and
- Methods (u, r): Maximum likelihood estimation of the data using the estimated proportion covariate created using a combination of three urban (u=1 km, 2 km, 5 km) and four rural (r=1 km, 5 km, 10 km, and 20 km) buffer sizes created around the offset DHS location ($\widehat{x}_i^{(t)}(u, r)$).

Method 1 represents the optimal analysis, based on the true covariate of interest. Use of Method 1 is not possible in current studies because access to the true DHS locations needed to calculate these buffer averages is not available. Method 2 represents the naive analysis which has been used in previous studies. This analysis fits the correct statistical model, similar to Method 1, with the incorrect covariate $x_i^{(0)}$ based on cell extraction of the offset DHS location. Methods (u, r) fits the same statistical model as methods 1 and 2, while using estimates of the true buffer average covariate $\widehat{x}_i^{(t)}(u, r)$. These estimates are obtained using a combination of urban/rural buffer sizes and calculating the buffer averages accordingly. In addition to being able to provide urban and rural-specific guidelines, assessment of the different combinations of these two neighborhood definitions would also allow for the determination of optimal combinations of urban and rural buffer settings that minimize the bias associated with point offsets. Method (u, r) can be used by all researchers since it is based on the offset DHS cluster locations and represents a compromise between methods 1 and 2.

Simulation Study

We fit each of the fourteen methods to a single generated dataset and collect information from each of the model fits. We collect the estimate of β_1, $\widehat{\beta}_1^{(j)}$, for each method $j = 1, \ldots, 14$. β_1 represents the main parameter of interest in the study because it describes the association between the average of the continuous values surrounding a DHS cluster and the outcome. Collection of $\widehat{\beta}_1^{(j)}$ allows us to estimate the bias of the estimator from each method, $\mathrm{E}\left[\widehat{\beta}_1^{(j)}\right] - \beta_1$.

We use standard mixed statistical models which account for the fact that each method is applied to the same dataset through use of random effects. We then test whether the bias of the estimator associated with each method is significantly different from zero and whether this bias changes across spatial smoothness setting. These analyses are repeated separately for each of the three considered categorical cell type of interest. Determining what combinations of urban-rural buffer definitions and surface smoothness characteristics lead to unbiased effect size estimates will help develop guidelines pertaining to the use of DHS GPS data in studies linking data from ancillary continuous raster surfaces. Specifically, results here will define appropriate scales of analysis needed to minimize bias associated with effect estimates and covariate misspecification.

4.3.2 Results

Guidelines on the usage of DHS GPS data in the context of integration of ancillary continuous raster data should be based on minimizing bias in the effect estimates of interest. Based on this simulation study, for studies aimed at addressing the influence of contextual environmental data on DHS cluster-level outcomes, use of buffer sizes between 1 and 5 km and rural buffer sizes between 1 and 20 km provided unbiased estimates under surfaces of moderate to high autocorrelation. Under surfaces of low spatial autocorrelation, all neighborhood definitions failed to provide unbiased parameter estimates. Point extraction led to unbiased parameter estimates for surfaces with moderate to high spatial autocorrelation, but performed poorly for non-smooth surfaces.

One-sample t-tests addressing the bias observed in maximum likelihood estimates of regression parameters for all settings indicated that no significant bias was affecting effect size estimates. This indicated that any subsequent bias we observed from the simulation study could be attributed to point displacements rather than inherent variability in the effect size estimate (results not shown). One-sample t-tests addressing whether the bias in regression parameter estimates associated with neighborhood definitions (i.e., buffers around displaced points) was significantly different than zero indicated that across surfaces with moderate to high levels of spatial autocorrelation, estimates were unbiased when coverage was calculated for neighborhoods composed of rural buffer sizes between 1 and 10 km and urban buffer sizes between 1 and 5 km (Fig. 4.3B and 4.3C). The bias of estimates was dependent on the smoothness of the ancillary surface, with sensitivity to point displacements being higher within very low autocorrelated surfaces ($\rho = 0.05$; Fig. 4.3A). In other words, for a very noisy, unsmooth surface, point displacements can drastically alter observed raster values, and buffer averaging fails to help reduce any resulting bias. For an extremely smooth surface, however, any urban-rural buffer definition is fine (among those considered) because neighboring values will be similar up to very large distances away from the true DHS location. Interestingly, however, for a moderately smooth surface, if rural buffers are too large, the potential to capture data outside of a smooth region increases and neighborhood averages begin to deviate more significantly from values obtained at true DHS cluster locations.

One-sample t-tests addressing whether the bias in regression parameter estimates associated with point extraction was significantly different than zero indicated low levels of bias. Overall, point extraction provided unbiased results across most autocorrelation surfaces tested, and was thus shown to be relatively robust to point displacement (Table 4.1). Only point extractions from non-smooth surfaces with very low spatial autocorrelation were found to be associated significantly biased results.

ρ	t-statistic	p-value	Mean bias	Lower 95% CI	Upper 95% CI
0.05	9.6234	0.0000	0.0289	0.0230	0.0348
0.50	0.6381	0.5239	0.0019	-0.0040	0.0078
0.95	1.3989	0.1629	0.0042	-0.0017	0.0101

Table 4.1: Results of one-sample t-tests addressing the significance of bias associated with raster cell extraction.(A) Point extraction is not recommended for non-smooth surfaces ($\rho = 0.05$) because it yielded significantly biased effect size estimates. (B) Mean buffer values for moderately smooth surfaces ($\rho = 0.5$) provided unbiased effect estimates for all combinations of urban-rural buffers in which rural buffers were less than 20 km in radius. (C) Buffer means across all radii tested for urban and rural locations provided unbiased effect size estimates for surfaces with high spatial autocorrelation.

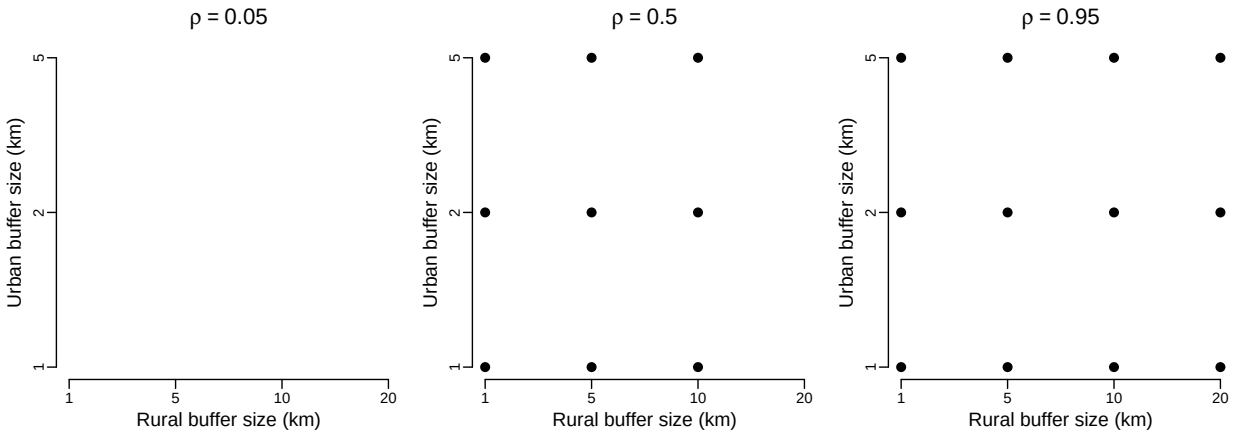


Figure 4.3: Circular buffer sizes associated with nonsignficant bias in effect estimates due to point displacement across multiple spatial smoothness levels of ancillary continuous raster data sets. Overall, use of urban buffer sizes between 1 - 5 km, and rural buffers between 1 - 10 km resulted in unbiased effect estimates across surfaces with moderate to high spatial autocorrelation ($\rho \geq 0.5$). Under surfaces of low spatial autocorrelation, no neighborhood combinations yielded unbiased effect size estimates.

4.4 Proposed Guidelines: Continuous Raster Data

According to results from this simulation study, in studies that integrate ancillary continuous raster data for analyses with DHS GPS data, the random offsets used to protect the privacy of DHS survey respondents could result in misclassified assignments of predictor variables at the DHS cluster-level depending on characteristics of the surface from which data is being linked. For relatively smooth surfaces, bias was low for both point extraction and most urban/rural neighborhood definitions. For consistency with results obtained from the simulation study using categorical raster data, **we suggest the use of 5 km buffers around both urban and rural DHS clusters**. Unlike with categorical raster data, **point extraction provided adequately unbiased estimates** for most surface types. Because highly non-smooth surfaces yielded highly biased estimates from both point extraction and neighborhood buffer approaches, we further recommend that if investigators plan on working with such surfaces, they attempt to smooth the surface in some way to mitigate the effects of such potential bias. Table 4.2 provides an overview of proposed guidelines. Note that these guidelines provide general rules to consider when linking continuous raster data to randomly displaced DHS point data. We have shown that the extent of bias in inferences drawn from linking ancillary continuous data to displaced DHS GPS data depends heavily on the smoothness of the ancillary raster surface.

ρ	Neighborhood definition	
	Urban	Rural
0.05	None	None
0.50	1-5 km	1-10 km
0.95	1-5 km	1-20 km

Table 4.2: Overview of general guidelines for neighborhood definitions for studies linking DHS GPS data to ancillary continuous raster data.

4.5 Case Study: Anemia Risk and Helminth Prevalence

4.5.1 Methods

The goal of this case study was to determine whether the predicted prevalence of malarial parasites (helminths) in a neighborhood is associated with the number of people who are anemic (among all respondents tested) in a DHS cluster. Raster data on helminth prevalence was obtained from the Malaria Atlas Project (http://www.map.ox.ac.uk/; Figure 4.4). The outcome of interest, i.e., number of respondents who are anemicin DHS clusters was obtained from the 2011 Uganda DHS.

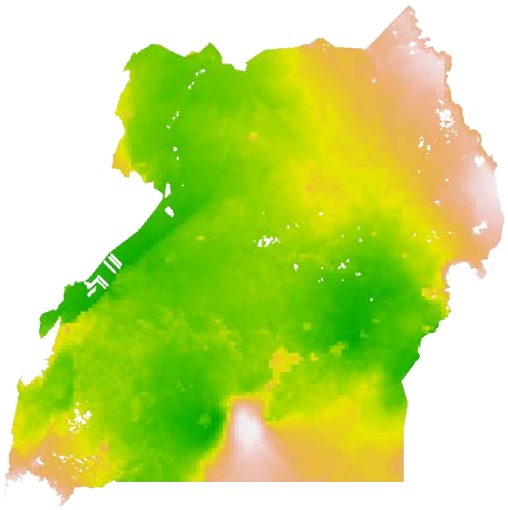


Figure 4.4: Malaria Atlas Project helminth data used in subsequent analyses. Data provided in this raster file pertained to helminth prevalence throughout the study area, with cell values corresponding to probabilities between zero and one. Prevalence ranged from 0 (white) to 0.60 (green), with the color scheme of orange to yellow representing prevalence values ranging approximately between 0.20 to 0.40.

To determine how smooth a raster surface is, in terms of proposed guidelines, investigators can convert a raster to a points shapefile that assumes cell values and fit a simultaneous autoregressive regression model on the points generated from the raster (Appendix B.4). This raster surface fit under the highly autocorrelated surface category explored in the simulation study ($\rho = 0.998$).Proposed guidelines suggest that to avoid bias in effect estimates when using continuous raster data from highly smooth surfaces investigators could either use point extraction or define neighborhoods around DHS clusters as having buffers of 5 km radius. For this case study, we present both approaches.

Average helminth prevalence was calculated using urban and rural buffer sizes of 5 km radii for each DHS cluster. Using both, true, non-displaced DHS cluster locations, along with the publicly available displaced locations, effect estimates associated with predictor variables generated for both true and displaced clusters were compared. Specifically, a Poisson regression model was fit to the data with anemia counts per cluster as the outcome variable, neighborhood helminth prevalence as the predictor variable, and an offset accounting for cluster-level population size. Slope parameters and standard errors were compared for the true and displaced datasets.

4.5.2 Results

Using neighborhood definitions from the proposed guidelines, the estimated effect sizes for the true and displaced datasets did not differ significantly. The estimated slope parameter, using the true dataset, was 0.137 (95% CI: 0.054, 0.219); for the displaced data, it was 0.132 (95% CI: 0.049 ,0.214). If direct cell extraction was used to generate covariates of interest, effect estimates obtained from the displaced and true DHS GPS data were also similar. The estimated slope parameter obtained using direct cell extraction with the true data was 0.978 (95% CI: 0.353, 1.604); for the displaced data, it was 0.995 (95% CI: 0.369, 1.620).while for the true data was .

4.5.3 Discussion

This analysis of anemia risk and helminth prevalence in Uganda was carried out for purposes of illustration; it is likely that the true effect of malaria prevalence on anemia incidence will be moderated by other unaccounted variables. The purpose of this case study was to demonstrate how well guidelines, which were established following empirical results of a simulation study, perform in a realistic application of DHS GPS data. The results showed that the proposed guidelines performed well in practice.

4.6 Categorical Raster Simulation Study

4.6.1 Methods

Generation of an Analysis Dataset

In the simulation study, we generate datasets for analysis, collecting information during each analysis which aid in answering the proposed questions of interest. In order to generate a single dataset, we begin by using the true (non-displaced) Uganda DHS cluster locations along with the created raster surfaces which have been reclassified to represent 9 categories of values. The analyses focused on only three levels of misclassification rates associated with raster surfaces that differed according to their representativeness (i.e., prevalence or percent cover) throughout the study area. Misclassification rates for these surfaces were approximately 11, 18, and 21 %. We then define the true covariate of interest as the proportion of the considered cell type within a 2 km buffer of the true DHS cluster location for urban clusters and 5 km for rural clusters. Next, we choose regression coefficient values for the Poisson statistical regression model considered in the study such that $\beta_0 = 1.00$ and $\beta_1 = -0.27$. Using these $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ parameters, we are able to generate a dataset of Poisson distributed data, one data point for each DHS cluster. The proposed model is given as $Y_i|\lambda_i \overset{ind}{\sim} \text{Poisson}(\lambda_i)$ where $\ln(\lambda_i) = \beta_0 + \beta_1 * x_i^{(t)}$ and $x_i^{(t)}$ is the proportion of considered cell type within the specified buffer created around the true location of DHS cluster i. Recall that this proportion covariate is unknown by the researchers. We generate 100 independent datasets by repeating the process under each spatial smoothness setting for a total of 300 datasets.

Analysis

We propose 14 methods to analyze the generated datasets of interest. These methods include

- Method 1: Maximum likelihood estimation of the data using the true proportion covariate based on the true DHS cluster location ($x_i^{(t)}$) (unknown to researcher),

- Method 2: Maximum likelihood estimation of the data using the point extracted cell value (indicator variable taking value one if the extracted point is the considered cell type, zero otherwise) based on the offset DHS cluster location ($x_i^{(0)}$) (naive analysis), and
- Methods (u, r): Maximum likelihood estimation of the data using the estimated proportion covariate created using a combination of three urban (u=1 km, 2 km, 5 km) and four rural (r=1 km, 5 km, 10 km, and 20 km) buffer sizes created around the offset DHS location ($\widehat{x}_i^{(t)}(u, r)$).

Method 1 represents the optimal analysis, based on the true proportion covariate of interest. Use of Method 1 is not possible in current studies because access to the true DHS locations needed to calculate these proportions is not available. Method 2 represents the naive analysis which has been used in previous studies. This analysis fits the correct statistical model, similar to Method 1, with the incorrect covariate $x_i^{(0)}$ based on cell extraction of the offset DHS location. Methods (u, r) fit the same statistical model as methods 1 and 2, while using estimates of the true proportion covariate $\widehat{x}_i^{(t)}(u, r)$. These estimates are obtained using a combination of urban/rural buffer sizes and calculating the buffer proportions accordingly. In addition to being able to provide urban and rural-specific guidelines, assessment of the different combinations of these two neighborhood definitions allows for determination of optimal combinations of urban and rural buffer settings that can minimize the bias associated with point offsets. Method (u, r) can be used by all researchers because it is based on the offset DHS cluster locations and represents a compromise between methods 1 and 2.

Simulation Study

We fit each of the fourteen methods to a single generated dataset and collect information from each of the model fits. We collect the estimate of β_1, $\widehat{\beta}_1^{(j)}$, for each method $j = 1, \ldots, 14$. β_1 represents the main parameter of interest in the study because it describes the association between the proportion of the selected cell type surrounding a DHS cluster and the outcome. Collection of $\widehat{\beta}_1^{(j)}$ allows us to estimate the bias of the estimator from each method, $\mathrm{E}\left[\widehat{\beta}_1^{(j)}\right] - \beta_1$.

We use standard mixed statistical models which account for the fact that each method is applied to the same dataset through use of random effects. We then test whether the bias of the estimator associated with each method is significantly different from zero and whether this bias changes across spatial smoothness setting. These analyses are repeated separately for each of the three surfaces pertaining to different misclassification rates. Determining what combinations of urban-rural buffer definitions and surface smoothness characteristics lead to unbiased effect size estimates will help in developing guidelines for the use of DHS GPS data in studies linking data from ancillary categorical raster surfaces. Specifically, the results here define the appropriate scales of analysis needed to minimize bias associated with effect estimates and covariate misspecification.

4.6.2 Results

Guidelines on the usage of DHS GPS data in the context of integration of ancillary categorical raster data should focus on minimizing bias in the effect estimates of interest. Based on this simulation study, for studies aimed at addressing the influence of contextual environmental data on DHS cluster-level outcomes, it is appropriate to use a buffer between **1 and 5 km to define**

urban neighborhoods. The definition of rural neighborhoods will depend on specific characteristics of the surface at hand; however, generally, **rural buffers of 5 km** will provide unbiased estimates for surfaces that are associated with misclassification rates less than 20%. These are general and conservative guidelines because we note that prevalence of cell types along with surface smoothness can influence the bias of an estimate. Overall, higher levels of significant bias were observed for the moderately correlated surface, than for the low or highly autocorrelated surfaces, and effect estimates associated with rare cell values were the least biased among the three different effect estimates considered.

One-sample t-tests addressing the bias observed in maximum likelihood estimates of regression parameters for the three scenarios determined that there was no significant bias at the 0.05 significance level. Thus, any subsequent bias observed from the simulation study could be attributed to point displacements rather than inherent variability in the effect size estimate (results not shown). One-sample t-tests addressing whether the bias in regression parameter estimates associated with neighborhood definitions (i.e., buffers around displaced points) was significantly different than zero indicated that misclassification rates, which are a function of both surface smoothness and cell type prevalence, could affect bias in effect size estimates. In other words, depending on the overall probability of misclassification, different neighborhood definitions would yield unbiased effect size estimates. Thus, guidelines on the usage of DHS GPS data will depend on these factors. Overall, estimates were unbiased when coverage was calculated for neighborhoods composed of rural buffer sizes between 5 and 10 km and urban buffers between 1 and 5 km. The bias of estimates was dependent on misclassification rates, with estimates corresponding to surfaces with low misclassification rates ($< 20\%$) showing less sensitivity to point displacement than surfaces associated with high misclassification rates ($> 20\%$; Figure 4.5).

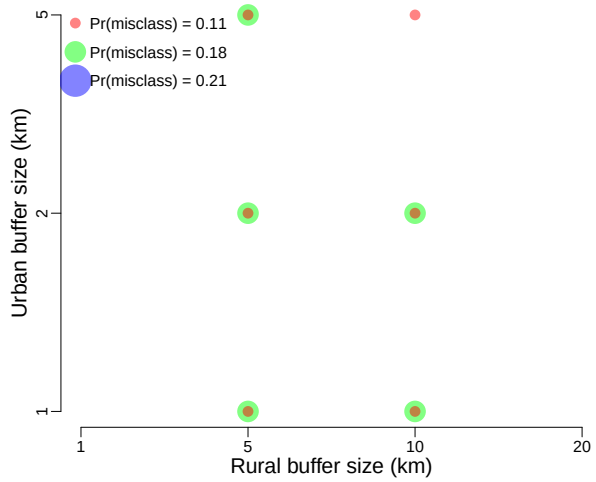


Figure 4.5: Circular buffer sizes associated with nonsignficant bias in effect estimates due to point displacement across multiple misclassification levels for ancillary categorical raster data sets. When misclassification rates are less than 20%, use of urban buffer sizes between 1 and 5 km, and rural buffer sizes at or exceeding 5 km resulted in unbiased effect estimates; however when the probability of misclassification is greater than 0.20, investigators should probably refrain from conducting analyses in which data from such surfaces is integrated with DHS data.

One-sample t-tests of least square mean bias addressing whether the bias in regression parameter

estimates associated with point extraction was significantly different than zero indicated strong bias. Overall, point extraction failed to provide unbiased results across the three surfaces tested, and was thus shown to be highly sensitive to point displacement (Table 4.3).

Misclassification rate	t-statistic	p-value	Mean bias	Lower 95% CI	Upper 95% CI
0.11	4.69	<0.0001	0.019	0.011	0.026
0.18	4.53	<0.0001	0.018	0.010	0.026
0.21	5.38	<0.0001	0.018	0.012	0.025

Table 4.3: Results of one-sample t-tests addressing the significance of bias associated with raster cell extraction. Point extraction is not recommended because it yielded significantly biased effect size estimates for all tested scenarios.

4.7 Proposed Guidelines: Categorical Raster Data

According to results from this simulation, in studies that integrate ancillary categorical raster data for analyses with DHS GPS data, the random offsets used to protect the privacy of DHS survey respondents could result in misclassified assignment of predictor variables at the DHS cluster level, depending on how the ancillary data is linked to DHS data. The magnitude of this bias, however, could be made negligible by **use of mean cell values across urban and rural neighborhoods of 5 km radius. Direct cell extraction is not recommended**, because this sort of data and subsequent inferences from analyses are adversely affected by random point displacements. Given that most displacements for DHS GPS data occur between 0 and 5 km, the proposed minimum buffer size of 5 km for urban and rural is reasonable. We found that surface misclassification rateswhich are a function of both the smoothness of the ancillary raster surface and the prevalence of cell types of interest could also influence proper specification of covariate data and bias in regression estimates, with surfaces associated with low misclassification rates resulting in lower bias than surfaces associated with high misclassification rates. Thus, investigators who plan to link environmental surface data to DHS GPS data should also consider the nature of the raster surface in question; for example, the prevalence of boundaries or edges along which points may lie could contribute to high misclassification rates. Table 4.4 presents an overview of proposed guidelines. Note that these guidelines provide general rules to consider when linking categorical raster data to randomly displaced DHS point data. We have shown that the extent of bias in inferences drawn from linking ancillary categorical data to displaced DHS GPS data will depend on the scale of neighborhoods used to define the process of interest and characteristics of the ancillary surface that could result in higher rates of covariate misspecification. Single point extraction is highly discouraged, because our empirical results suggest that using a neighborhood average will best mitigate the potential bias associated with systematic geographic displacements.

Pr(misclass) $\leq$ 0.11%)		0.11 < Pr(misclass) $\leq$ 0.18%		Pr(misclass) $\geq$ 0.20	
Urban	Rural	Urban	Rural	Urban	Rural
1-5 km	5-10 km	1-2 km	5-10 km	1, Not recommended	Not recommended

Table 4.4: Overview of general guidelines for neighborhood definitions for studies linking DHS GPS data to ancillary categorical raster data.

4.8 Case Study: Anemia Risk and Cropland Cover

4.8.1 Methods

The motivating research question for this case study was to determine whether the amount of cropland cover in a neighborhood is associated with the number of people who are anemic (i.e., anemia of any severity) within a DHS cluster. Raster data on land cover was acquired from LP DAAC (https://lpdaac.usgs.gov/; Figure 4.6). The outcome of interest, i.e., number of people who are anemic in DHS clusters, was obtained from the 2011 Uganda DHS (*V457*).

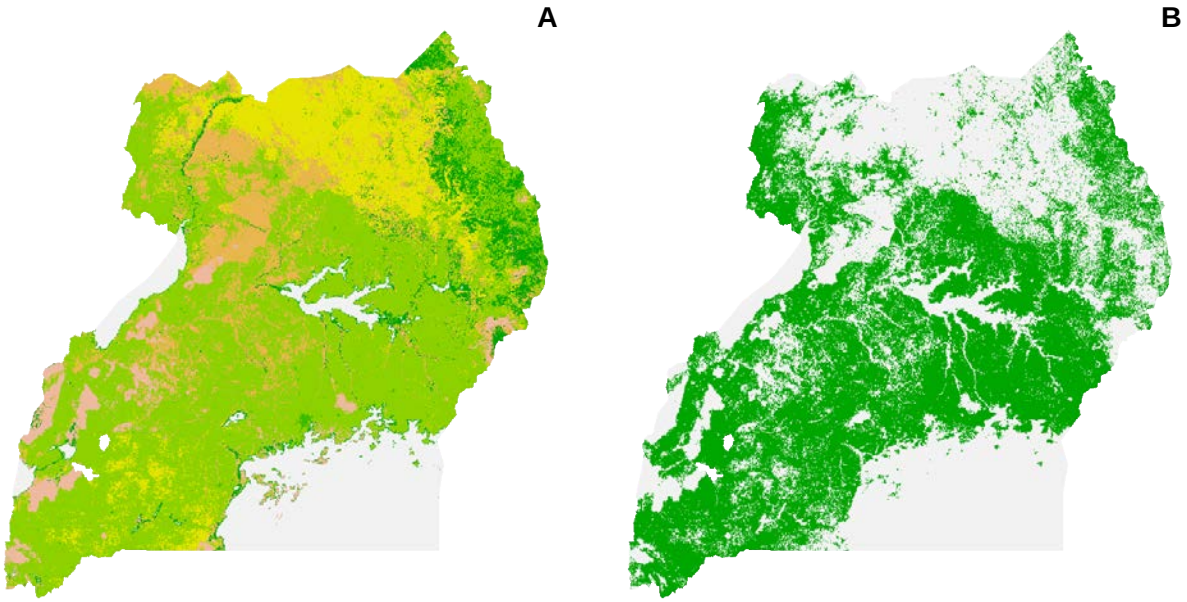


Figure 4.6: LP DAAC Land Cover data used in subsequent analyses. Map A: The original raster file was reclassified to include seven categories of land cover types ranging in values from 0 to 6: water, forest, woody savanna, savanna, cropland, urban, and other. Map B: With this dataset, cropland cover (green) accounted for roughly 50% of all land cover in Uganda.

Cropland cell types comprised roughly 50% of raster cell values in this dataset, and the surface displayed a high level of spatial autocorrelation in values ($\rho \approx 0.98$). Median misclassification rate under this binary raster surface was 0.157. To determine misclassification rates, in terms of proposed guidelines, investigators can use the function provided in Appendix B.6. Briefly, for each point in the dataset compute the probability that the true point was assigned a raster cell value different than that observed. In this way, misclassification rates can be computed similarly to those in point-in-polygon analyses (Section 5.1). Proposed guidelines suggest that in order to avoid bias in effect estimates when using categorical raster surfaces with misclassification rates less than 0.20, one could define 5 km circular neighborhoods around DHS clusters. For this case study, we also present results associated with simple point extraction.

Proposed guidelines suggest that to avoid bias in effect estimates, investigators should define neighborhoods around DHS clusters as having buffers with at least 5 km radius. The guidelines further suggest calculating the percentage of cover (i.e., proportion of raster cells of a certain

type) to generate covariates/predictor variables of interest. Point extraction from raster surfaces is strongly discouraged because effect estimates associated with this type of covariate data were found to be highly sensitive to random DHS cluster displacements.

For this case study, the percentage of cropland cover for the displaced data was calculated using urban and rural buffer sizes of 5 km radii for each DHS cluster. True percentage of cropland cover for DHS cluster i was calculated as the proportion of cells within the corresponding circular radius of 2 km for urban clusters and 5 km for rural clusters, as in the simulation study. Using both true, non-displaced DHS cluster locations along with the publicly available displaced locations, effect estimates associated with predictor variables generated for both true and displaced clusters were compared. Specifically, a Poission regression model was fit to the data with anemia counts per cluster as the outcome variable and neighborhood percent cropland cover as the predictor variable; slope parameters and standard errors were compared for the true and displaced datasets. Additionally, for illustrative purposes, the analysis was repeated using the method of exact cell extraction, which was shown to be biased in the simulation study.

4.8.2 Results

Using neighborhood definitions from the proposed guidelines, the estimated effect sizes for the true and displaced datasets did not differ significantly. The estimated slope parameter for the analysis using the true data was 0.111 (95% CI: 0.029, 0.194); for the displaced dataset, the estimated slope parameter was 0.097 (95% CI:0.014, 0.179). If the guidelines were disregarded, however, and direct cell extraction was used to generate covariates of interest, the effect estimates obtained from the true and displaced DHS GPS data were different, and yielded different conclusions based on p-values. The estimated slope parameter obtained using direct cell extraction with the true data was 0.202 (95% CI: 0.021, 0.383); for the displaced data, it was 0.161 (95% CI: -0.012, 0.334).

4.8.3 Discussion

The analysis of anemia risk and cropland cover in Uganda was carried out for purposes of illustration; it is likely that the true effect of cropland cover on anemia incidence will be moderated by other unaccounted variables. The purpose of this case study was to demonstrate how well guidelines, which were established following empirical results of a simulation study, perform in a realistic application of DHS GPS data. The results showed that the proposed guidelines performed well in practice and that violation of the proposed guidelines via direct cell extract resulted in biased estimates of the effect size.

Chapter 5

Influence of Offsets in Point-in-Polygon Analyses

5.1 Quantifying Misclassification Rates

When integrating ancillary areal data with georeferenced DHS data, point displacement could lead to misspecification of areal-level covariates in subsequent analyses. The probability that a point is misclassified to a particular areal (i.e. polygon) feature will depend on the size and shape of the feature to which the displaced point is overlaid. Figure 5.1 provides a diagram to help illustrate the potential misclassification problem associated with integrating ancillary areal data to geographically displaced DHS cluster data.

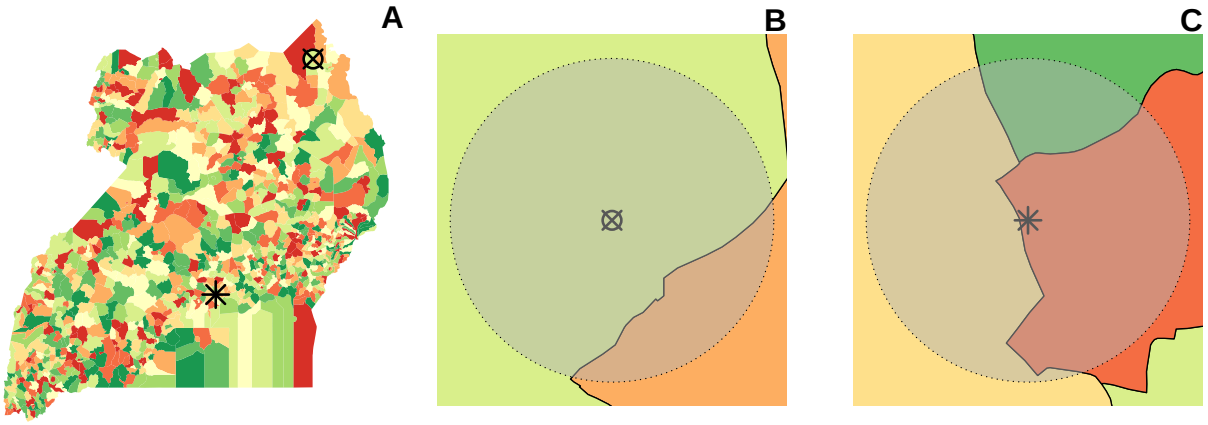


Figure 5.1: Schematic of hypothetical scenario where data from areal units is to be linked to geographically displaced DHS cluster points. In scenario (A) we demonstrate two displaced points that are represented by red characters. In scenario (B) the crossed circle represents the location of a rural cluster, and in scenario (C) the asterisk represents an urban location. For each of these cluster points, the maximum displacement buffers are overlaid around the displaced points. The true point locations for both of scenarios B and C lie within one of the respective overlapping features.

Specifically, the probability of misclassification can be calculated as one minus the proportion of the area within the maximum displacement buffer that intersects with the polygon feature in which the displaced point lies. The true point will be located within this maximum buffer with probability equal to 1.0. Thus, if the maximum buffer around a displaced point falls completely

within a polygon feature, there is a 100% chance that the true point also lies within this polygon feature. If, however, the maximum buffer intersects six polygon features, each with different intersection areas as in Figure 5.1, then we can say that the probability that the true point lies within one of these neighboring polygon features will be proportional to the intersection areas of these features. Using Figure 5.1C as an example, the probability that the true point location lies within either the yellow or the orange feature is highest because these two areas encompass the largest percentages of the maximum displacement buffer. Likewise, the probability that the true point location lies within the green feature is lowest because this area encompasses the smallest percentage of the maximum displacement buffer.

Note that the probability of misclassification is therefore highly dependent on the sizes/areas of ancillary areal features to which DHS data is to be linked. In the following sections we demonstrate how weighted estimates could be used to mitigate potential biases associated with polygon misspecification when linking DHS data to continuous areal data. We also extend this to the case of linking DHS data to categorical areal data, whereby covariate assignment is based on determining the areal unit with the highest probability of capturing the true DHS cluster location.

5.2 Mitigating Effects of Misclassification

To moderate the effect of assigning randomly displaced DHS cluster points to the incorrect areal feature, we propose using methods that explicitly account for the probability that the true point lies within a specific polygon within the displacement buffer radius. The two main steps involved in calculating this probability are listed as follows.

1. For each observed point, generate a buffer whose radius is equal to the maximum displacement distance for that point. That is, for points designated as urban by the DHS use a radius of 2 km, and for points designated as rural by the DHS use a radius of 5 km (or 10 km).

2. For each buffer region, calculate the proportional area covered by each intersecting polygon feature within the buffer region. Specifically, let a_i be the area of polygon i within a single buffer region. The proportional area of this feature within the buffer region is then given by $a_i/(\pi r^2)$, where r_i is the maximum displacement radius defining that buffer region.

If working with areal data that is continuous, these proportions can be used to generate a weighted average of values corresponding to the areal units that intersect the displacement buffer. In this case, the weight would be equal to the proportion of the given areal unit within the displacement buffer. If working with areal data that is categorical, then one can use these proportions to identify the areal unit with the highest probability of containing the true DHS point location. In this case, the covariate value associated with the most probable areal unit within a DHS cluster displacement buffer would be assigned to the cluster point. *R* code to reproduce these methods is provided in Appendix B.6. In the next section, we show how these methods perform with regard to moderating the bias associated with point-in-polygon misspecification.

5.3 Case Study: Neighborhood Determinants of HIV Knowledge

5.3.1 Methods

The goal of this case study was to determine whether comprehensive knowledge of HIV among DHS respondents in Uganda is associated with the economic characteristics of neighborhoods

(i.e., area-level poverty measures). Area-level poverty data obtained from the Uganda Bureau of Statistics (Emwanu et al., 2007) were linked to georeferenced administrative unit data. The outcome of interest, i.e., number of respondents interviewed in DHS clusters that had comprehensive knowledge of HIV, was obtained from the 2011 Uganda DHS. This indicator variable was constructed according to correct responses to all of the following DHS survey questions:

1. To reduce the risk of getting HIV one should always use condoms (*V754CP*)
2. To reduce the risk of getting HIV one should have only one sex partner (*V754DP*)
3. One can get HIV from mosquito bites (*V754JP*)
4. One can get HIV by witchcraft or supernatural means (*V823*)
5. A healthy looking person can have HIV (*V756*)

Area-level poverty was defined as the percentage of the population below the poverty line within an administrative unit. A separate analysis was also conducted to illustrate proposed methods using categorical areal data. In this case, area-level poverty was redefined as an ordinal variable corresponding to low ($< 10\%$ below poverty line), moderate ($10 - 50\%$ below poverty line), and high poverty levels ($> 50\%$ below poverty line). After assigning poverty data to administrative units, DHS cluster points were overlaid, displacement buffers were generated, and proportional areas were calculated as described in Section 5.2. For the analysis associated with continuous areal data, weighted averages of poverty measures were calculated for each DHS cluster. For the analysis associated with categorical areal data, the poverty level value associated with the most probable areal unit within a displacement buffer was assigned to each DHS cluster.

Paired t-tests were used to compare the performance of the naive and proposed point-in-polygon procedures. In other words, we tested whether the bias associated with simple point extraction was significantly different from the bias associated with either the most probable value approach or the weighted average approach. For this analysis, the squared bias obtained from non-displaced and displaced points was compared between the two procedures (i.e., naive and proposed) to determine how different the obtained value was from the true value. For the categorical areal data analysis the proportion of misclassified points is presented from both the naive and proposed methods, and misclassification rates are compared between the two methods through the use of a McNemar's test.

In addition to quantifying the bias associated with misclassification in point-in-polygon analyses, we also addressed how effect estimates would be influenced. Specifically, linear regression models of the form: $y_i = \beta_0 + \beta x_i + \varepsilon_i$ were fit to the data, where y_i represents the count of respondents interviewed within DHS clusters that had comprehensive knowledge of HIV transmission, x_i represents either the true, weighted average, or observed (i.e., naive) percent of the population below the poverty line within the administrative unit assigned to DHS cluster i (for continuous areal data), or the true, most probable, or observed (i.e., naive) poverty level within the administrative unit assigned to DHS cluster i (for categorical areal data). The regression model also included an offset for population size calculated as the total number of individuals in a sampled cluster. We also repeated this analysis using another areal dataset comprised of smaller administrative units (i.e., third-level administrative units)

5.3.2 Results

First-level Administrative Units

For this dataset, misclassification rates were fairly low on average (4.5% ± 10.15%), with the highest rates ($\geq$ 50%) consistently observed along the boundaries of areal units (Figure 5.2). Results from paired t-tests indicated no significant differences between covariate values obtained from non-displaced (i.e., true) DHS cluster points and covariate values obtained by either extraction from displaced points, weighted estimates (for continuous poverty measures), or most probable values (for discrete poverty measures; results not shown); however the squared bias associated with observed continuous values was significantly lower than that associated with either probability-based methods for continuous data (Table 5.1). McNemar's tests indicated that the rates of misclassification did not differ between approaches that used naive (i.e., observed) and most probable classification schemes (p-value > 0.45). Parameter estimates obtained via Poisson regression models were comparable

	Covariate	t-statistic	p-value
Continuous	Weighted Average (rural 5 km)	-2.8359	0.0050
	Weighted average (rural 10 km)	-5.0372	0.0000
Categorical	Observed	-1.0000	0.3179
	Most probable value (rural 5 km)	-1.7366	0.0833
	Most probable value (rural 10 km)	-0.5769	0.5644

Table 5.1: Results from paired t-tests comparing squared bias of observed (i.e., naive point extraction) covariate values to squared bias of covariates obtained through other approaches (i.e., probability-based methods) indicated significant differences between the squared bias of observed continuous values and that obtained through respective approaches. For this dataset, classification of DHS clusters to level 1 administrative units was relatively robust to point displacement, because the squared bias of observed discrete values was not significantly different from the squared bias of discrete values obtained through probability-based methods. This was to be expected for this case study because point displacements for the 2011 Uganda DHS data were constrained to first-level administrative units.

among models using the true, weighted average (for continuous poverty measures), most probable values (for the categorical poverty measures), and observed covariate values (Table 5.2).

	Covariate	$\widehat{\beta}$	SE($\widehat{\beta}$)
Continuous	True	-0.0117	0.0010
	Observed	-0.0115	0.0010
	Weighted average (rural 5 km)	-0.0119	0.0010
	Weighted average (rural 10 km)	-0.0117	0.0010
Categorical	True	-0.2914	0.0302
	Observed	-0.2834	0.0304
	Most probable value (rural 5 km)	-0.3001	0.0303
	Most probable value (rural 10 km)	-0.2911	0.0300

Table 5.2: Parameter estimates associated with Poisson regression model were comparable across all covariate values used and true covariate values.

Third-level Administrative Units

For this dataset,the probability of misclassification was relatively high on average ($32 \pm 22\%$ using 5 km rural buffers), with the highest rates ($\geq 50\%$) consistently observed along the boundaries of areal units as in Section 5.3.2 (Figure 5.3). Results from paired t-tests indicated no significant differences between covariate values obtained from non-displaced (i.e., true) DHS cluster points and covariate values obtained by either extraction from displaced points, weighted estimates (for continuous poverty measures), or most probable values (for discrete poverty measures; results not shown). Likewise, the squared bias associated with observed continuous values was not significantly different from the squared bias associated with either of the probability-based methods for continuous data (Table 5.3). McNemar's tests indicated that the rates of misclassification did not differ between approaches that used naive (i.e., observed) and most probable classification schemes (p-value > 0.45). Parameter estimates obtained via Poisson regression models were comparable among

	Covariate	t-statistic	p-value
Continuous	Weighted average (rural 5 km)	0.4994	0.6179
	Weighted average (rural 10 km)	-0.7556	0.4505
Categorical	Most probable value (rural 5 km)	1.0936	0.2751
	Most probable value (rural 10 km)	0.1371	0.8910

Table 5.3: Results from paired t-tests comparing squared bias of observed (i.e., naive point extraction) covariate values to squared bias of covariates obtained through other approaches (i.e., probability-based methods) indicated no significant differences between the squared bias of observed continuous values and that obtained through respective approaches. For this dataset, classification of DHS clusters to level 3 administrative units was relatively robust to point displacement, and the squared bias of observed discrete values was not significantly different from the squared bias of discrete values obtained through probability-based methods.

models using the true, weighted average (for continuous poverty measures), most probable values (for the categorical poverty measures), and observed covariate values (Table 5.4).

	Covariate	$\widehat{\beta}$	SE($\widehat{\beta}$)
Continuous	True	-0.0015	0.0004
	Observed	-0.0019	0.0004
	Weighted average (rural 5 km)	-0.0019	0.0005
	Weighted average (rural 10 km)	-0.0023	0.0005
Categorical	True	-0.0837	0.0385
	Observed	-0.0907	0.0378
	Most probable value (rural 5 km)	-0.0833	0.0398
	Most probable value (rural 10 km)	-0.0869	0.0399

Table 5.4: Parameter estimates associated with Poisson regression model were comparable across all covariate values used and true covariate values obtained from third level administrative areal units.

Discussion

This analysis of neighborhood determinants of HIV knowledge in Uganda was carried out for purposes of illustration; it is likely that the true effects of neighborhood characteristics on comprehensive knowledge of HIV by DHS respondents will be moderated by other unaccounted variables. As

with the ancillary polygon file used in the analysis pertaining to first-level administrative units, the areal data from third-level administrative units yielded discrete covariate values and regression parameter estimates that were robust to misclassification bias associated with point displacement. However, results obtained here are not necessarily generalizable because misclassification bias is also likely to be highly contingent on the magnitude of spatial autocorrelation in areal unit data as well as the average area/size of areal units overlaid. In this case study, average misclassification rates were relatively high (0.32 using rural buffers of 5 km, and 0.55 using rural buffers of 10 km). Despite the high probability of point-in-polygon misclassification, very little bias was observed from naive analyses. This could be the result of strong spatial dependencies inherent in the poverty data used for the case study. However, had more spatially independent areal data been used, the effect of misclassification could have been more severe with regard to resulting bias. Adopting naive values obtained via simple point extraction may adequately represent true patterns when data from areal units exhibit strong spatial dependencies; however, when areal units are relatively small and spatial structure of variables of interest is weak, a weighted average or most probable value approach should mitigate any potential bias associated with misclassification due to random point displacement.

5.4 Proposed Guidelines

The problem of potential misclassification bias observed with point-in-polygon analyses is akin to that observed with integration of ancillary categorical raster data. Namely, when average distances to borders increase, the average probability of misclassification also increases. Although the naive approach to linking areal data to DHS data may work in some circumstances, it would certainly not be a generalizable solution for all areal datasets. Therefore, because the probability-based approaches presented above will reduce the effects of misclassification under most circumstances, **we recommend that investigators implement a weighted average or most probable value approach when linking DHS cluster data to ancillary areal data**. Misclassification rates are highly dependent on areal unit size. In Appendix B.6 we include an R function to help investigators quantify misclassification rates, so that they may determine whether rates are low enough (according to their standards) to justify simple point extraction. Additionally, the function provides probability-based covariate values along with observed point extraction values associated with a polygon shapefile. See Appendix B.3 for details on its implementation.

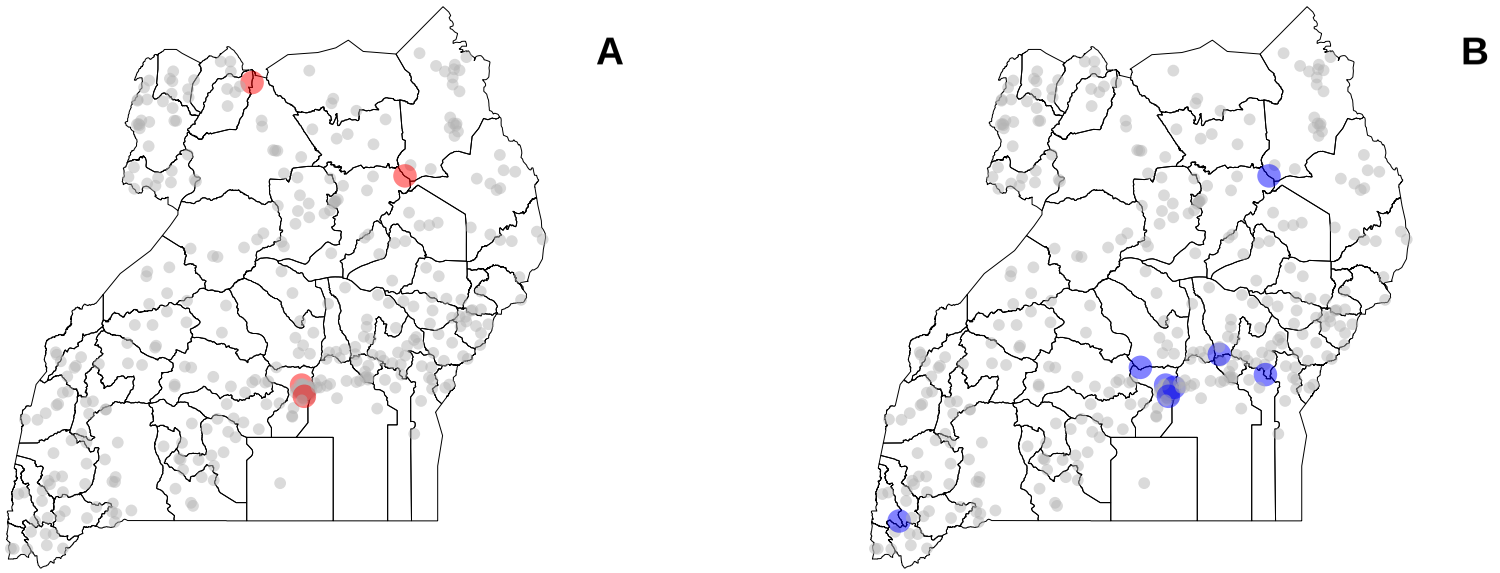


Figure 5.2: Observed cluster locations with high misclassification rates ($\geq$ 50%) from the 2011 UgandaDHS overlaid with level 1 administrative boundaries. Maximum displacement buffer radii were determined by the DHS displacement protocol whereby points in urban areas were displaced up to 2 km and points in rural areas were displaced up to 5 km (or 10 km for 1% of rural points). The red points in map (A) correspond to observed DHS cluster locations with high misclassification rates assuming a maximum displacement buffer of 5 km for rural points; the blue points in map (B) correspond to analogous cluster locations assuming a maximum displacement buffer of 10 km for rural points.

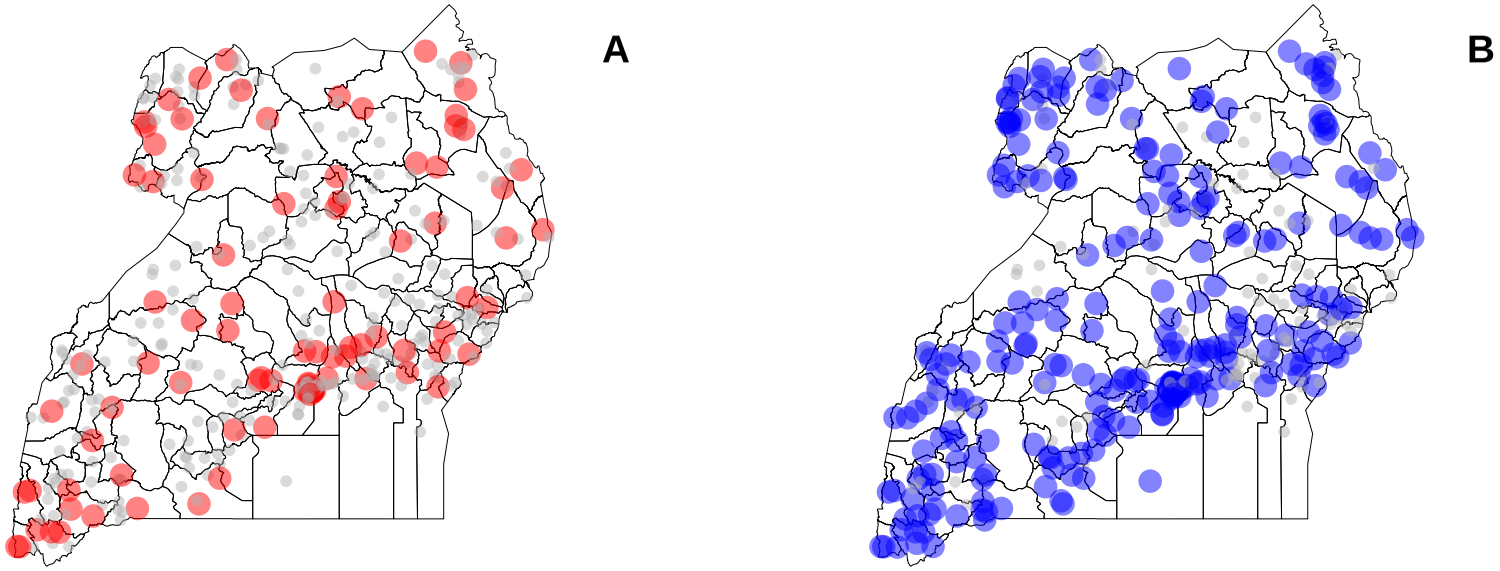


Figure 5.3: Observed cluster locations with high misclassification rates ($\geq$ 50%) from the 2011 Uganda DHS overlaid with level 3 administrative boundaries. Maximum displacement buffer radii were determined by the DHS displacement protocol whereby points in urban areas were displaced up to 2 km and points in rural areas were displaced up to 5 km (or 10 km for 1% of rural points). The red points in map (A) correspond to observed DHS cluster locations with high misclassification rates assuming a maximum displacement buffer of 5 km for rural points; the blue points in map (B) correspond to analogous cluster locations assuming a maximum displacement buffer of 10 km for rural points.

Chapter 6

Summary

This report highlights general considerations regarding the use of randomly displaced DHS GPS data. The results indicate that the effect of displacement varies with respect to the scale of analysis and the specific ancillary data to which investigators are linking DHS data. For instance, with regard to generating distance-based covariates, the density of destination points will likely influence the magnitude of the bias associated with point offsets. Additionally, the smoothness of a raster surface will variably impact the extent of bias in a covariate associated with random displacements. Linking displaced data to very smooth surfaces will likely have little impact on analysis results because covariate values obtained from displaced data will be very similar to those associated with the true, non-displaced location. However, if displaced data is to be linked to highly variable surfaces, point displacement may cause problems when linking to this ancillary surface data. The issue of scale presents itself again when linking to ancillary areal data because the sizes of these areal units could lead to higher or lower rates of misclassification for displaced points for small and large areal units, respectively. Finally, although the guidelines presented here are intended to be general, circumstances associated with a specific project may depart from those addressed in this report. Special care should be taken to pursue research aims that can be addressed by an appropriate scale of analysis with respect to point displacement.

6.1 Distance-based Analyses

When ignored, the DHS offset of locations can inflate the bias and MSE of the statistical model estimators for analyses that use distance to closest resources as a covariate. However, the impact of the offset changes based on how densely the resource locations are distributed in the domain of interest. In general, we recommend the use of regression calibration which attempts to unbias the covariate and can help to reduce the bias and MSE of the resulting estimator. Ignoring the offset and using the observed distances to closest resource without adjustment most often leads to poor estimation of the effect of interest. However, when considering only urban locations with a very dense set of resource locations, this naive approach is preferred over regression calibration though still not recommended. In general, using the described distance based covariate is preferable in settings where the resources of interest are spatially less dense across the domain. No major differences are seen between results where the resources are points as opposed to line segments of interest.

6.2 Integration of Ancillary Raster Data

The impacts of point displacements on misspecification of covariates and interpretation of analytic results are affected by the smoothness of the raster surface to which DHS GPS data is linked. Overall, empirical results obtained using simulated surfaces indicated that the impact of this displacement could be moderated through the generation of average values using neighborhood buffers. Note that guidelines here were developed based on standardized data, thus investigators should center and scale their covariate data accordingly (i.e., convert values to z-scores) when applying proposed guidelines from this study. Point extraction is generally not recommended with categorical raster data because this most often leads to biased results; however, it may be an adequate approach with continuous raster data. Empirical results suggest taking averages across circular rural and urban buffers of around 5 km. We note, however, that other continuous and/or categorical raster surfaces may yield different results due to differences in smoothness and/or grid cell size.

6.3 Integration of Ancillary Areal Data

Misclassification rates of covariate data linked from ancillary areal units are consistently highest when points are located near areal unit boundaries. Thus, as the area of these units decreases, point distances to boundaries will also decrease, and misclassification rates will be high. We also expect that the nature of spatial dependence in data from areal units will also impact the effects of misclassification on bias. Adopting naive values obtained via simple point extraction may adequately represent true patterns when areal units are relatively large with respect to the DHS point displacement protocol and spatial dependence of data is high, however when areal units are relatively small and respective data exhibits low levels of spatial dependence, a weighted average or most probable value approach should mitigate any potential bias associated with misclassification due to random point displacement. Thus, probability-based approaches are recommended over simple point extractions when spatially linking areal data to DHS GPS data.

6.4 Additional considerations when using DHS GPS data

The guidelines presented in this report are ultimately dependent on the scale of true processes of interest. With the raster-based simulations we demonstrated how buffer means could be used to generate covariates of interest. There we assumed that the true process of interest was occurring at a particular neighborhood scale, i.e., 2 km in urban areas and 5 km in rural areas. If, however, interest lay in understanding mechanisms associated processes occurring at a larger spatial scale, say 10 km neighborhoods, then covariates generated at this scale or something slightly larger would likely be more appropriate than those generated using a 5 km buffer. In other words, based on the raster simulation results, neighborhood-level covariates should be defined with regard to the spatial scale of underlying processes under investigation.

Guidelines provided here also assume that ancillary data are of good quality and relevant to DHS GPS data with regard to temporal overlap. Failure to uphold these assumptions will likely lead to further problems in generating interpretable and relevant study results. For example, linking DHS data to an interpolated surface with high levels of prediction error will result in misspecification of covariates due to problems with the ancillary data file, rather than issues associated with random DHS point displacement. Likewise, if linking DHS data to temporally

varying data such as census-based data or land cover data, special care should be given to ensuring that the time periods represented by the ancillary datasets correspond to the time periods associated with the DHS surveys to which data will be linked. Otherwise, any associations identified in subsequent analyses are likely to be confounded by temporally disjunct processes.

Point displacements for most countries are constrained to administrative boundaries, which helps reduce potential biases associated with displacement. Empirical results presented in this report correspond to worst-case scenario displacements, which are unconstrained by these boundaries. Guidelines presented here are therefore likely to be conservative in nature. Thus, investigators should note if they plan to use data from surveys with constrained displacements, some of these guidelines may be overly conservative.

References

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

Baschieri, A. (2007). Effects of modernisation on desired fertility in Egypt. *Population, Space and Place 13*, 353–376.

Berry, S. M., R. J. Carroll, and D. Ruppert (2002). Approximate quasi-likelihood estimation in models with surrogate predictors. *Journal of the American Statistical Association 97*, 160–169.

Brown, J. and G. Heuvelink (2007). The Data Uncertainty Engine (DUE): A software tool for assessing and simulating uncertain environmental variables. *Computational Geosciences 33*, 172–190.

Burgert, C. R., J. Colston, T. Roy, and B. Zachary (2013). *Geographic displacement procedure and georeferenced data release policy for the Demographic and Health Surveys.* DHS Spatial Analysis Report No. 7. Calverton, Maryland, USA: ICF International.

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

Chin, B., L. Montana, and X. Basagana (2011). Spatial modeling of geographic inequalities in infant and child mortality across Nepal. *Health and Place 17*, 929–936.

De Castro, M. C. and M. Fisher (2012). Is malaria illness among young children a cause or a consequence of low socioeconomic status? Evidence from the United Republic of Tanzania. *Malaria Journal 11*.

Emwanu, T., P. O. Okwi, J. G. Hoogeveen, P. Kristjanson, and N. Henninger (2007). Nature, distribution and evolution of poverty and inequality in uganda. Technical report, Uganda Bureau of Statistics and the International Livestock Research Institute (ILRI).

Foody, G. and P. Atkinson (Eds.) (2002). *Uncertainty in Remote Sensing and GIS*. Chichester, UK: Wiley.

Gabrysch, S., S. Cousens, J. Cox, and O. M. R. Campbell (2011). The influence of distance and level of care on delivery place in rural Zambia: A study of linked national data in a geographic information system. *PLoS Medicine 8*.

Giardina, F., L. Gosoniu, L. Konate, M. Diouf, and R. Perry (2012). Estimating the burden of Malaria in Senegal: Bayesian zero-inflated binomial geostatistical modeling of the MIS 2008 data. *PLoS ONE 7*.

Go, M., B. J. Cobum, J. T. Okano, and S. Blower (2011). Construction of geospatial health policy maps for Lesotho. In *18th Conference on Retroviruses and Opportunistic Infections, Boston*.

Gonese, E., J. Dzangare, S. Gregson, N. Jonga, O. Mugurungi, and V. Mishra (2010). Comparison of HIV prevalence estimates for Zimbabwe from antenatal clinic surveillance (2006) and the 2005-06 Zimbabwe demographic and health survey. *PloS One 5*.

Goodchild, M. and S. Gopal (Eds.) (1989). *Accuracy of Spatial Databases*. London and New York: Taylor & Francis.

Gosoniu, L., A. Msengwa, C. Lengeler, and P. Vounatsou (2012). Spatially explicit burden estimates of Malaria in Tanzania: Bayesian geostatistical modeling of the Malaria Indicator Survey Data. *PLoS ONE 7*.

Griffith, D. A. (1989). *Accuracy of Spatial Databases*, Chapter Distance calculations and errors in geographic databases, pp. 81–90. London and New York: Taylor and Francis.

Gryparis, A., C. J. Paciorek, A. Zeka, J. Schwartz, and B. A. Coull (2009). Measurement error caused by spatial misalignment in environmental epidemiology. *Biostatistics 10*, 258–274.

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

Hengl, T., G. B. M. Heuvelink, and E. E. Van Loon (2010). On the uncertainty of stream networks derived from elevation data: the error propagation approach. *Hydrology and Earth System Sciences 14*, 1153–1165.

Heuvelink, G., B. P.A., and A. Stein (1989). Propagation of errors in spatial modelling with GIS. *International Journal of Geographical Information Systems 3*, 303–322.

Heuvelink, G. B., P. A. Burrough, and A. Stein (2007). *Developments in analysis of spatial uncertainty since 1989. Classics from IJGIS: twenty years of the international journal of geographical science and systems*. London: Taylor and Francis.

Hillsman, E. L. and R. Rhoda (1978). Errors in measuring distances from populations to service centers. *The Annals of Regional Science 12*, 74–88.

Kandala, N., C. Ji, N. Stallard, S. Stranges, and F. P. Cappuccio (2007). Spatial analysis of risk factors for childhood morbidity in Nigeria. *The American Journal of Tropical Medicine and Hygiene 77*, 770.

Kandala, N., M. A. Magadi, and N. J. Madise (2006). An investigation of district spatial variations of childhood diarrhoea and fever morbidity in Malawi. *Social Science & Medicine 62*, 1138–1152.

Kashima, S., i. E. Suzuk, T. Okayasu, R. Jean Louis, and A. Eboshida (2012). Association between proximity to a health center and early childhood mortality in Madagascar. *PLoS ONE 7*.

Kazembe, L. and J. Namangale (2007). A Bayesian mulinomial model to analyse spatial patterns of childhood co-morbidity in Malawi. *European Journal of Epidemiology 22*, 545–556.

Kwan, M. (2012). The uncertain geographic context problem. *Annals of the Association of American Geographers 102*, 958–968.

Madsen, L., D. Ruppert, and N. Altman (2008). Regression with spatially misaligned data. *Environmetrics 19*, 453–467.

Magalhaes, R. J. S. and A. C. A. Clements (2011). Mapping the risk of anaemia in preschool-age children: The contribution of malnutrition, malaria, and helminth infections in West Africa. *PLoS Medicine 8*.

Mansour, S., D. Martin, and J. Wright (2012). Problems of spatial linkage of a geo-referenced demographic and health survey (DHS) dataset to a population census: A case study of Egypt. *Computers, Environment and Urban Systems 36*, 350.

Messina, J. P., M. Emch, J. Muwonga, K. Mwandagalirwa, S. B. Edidi, and N. Mama (2010). Spatial and socio-behavioral patterns of HIV prevalence in the Democratic Republic of Congo. *Social Science & Medicine 71*, 1428–1435.

Messina, J. P., S. M. Taylor, S. R. Meshnick, A. M. Linke, A. K. Tshefu, and B. Atua (2011). Population, behavioural and environmental drivers of malaria prevalence in the Democratic Republic of Congo. *Malaria Journal 10*, 161.

Montana, L., R. Bessinger, and S. Curtis (2000). Linking health facility and population level data in Kenya using geographic information systems.

Montana, L. S., V. Mishra, and R. Hong (2008). Comparison of HIV prevalence estimates from antenatal care surveillance and population-based surveys in sub-Saharan Africa. *Sexually Transmitted Infections 84 Suppl 1*, i78–i84.

Openshaw, S. (1984). *The modifiable areal unit problem*. Norwich, UK: Geo Books.

Pickering, A. J. and J. Davis (2012). Freshwater availability and water fetching distance affect child health in sub-Saharan Africa. *Environmental Science and Technology*.

Simler, K. R. (2006). *Nutrition mapping in Tanzania: An exploratory analysis. International Food Policy Research Institute (IFPRI)* (204 ed.).

Strickland, M. J., C. Siffel, B. R. Gardner, A. K. Berzen, and A. Correa (2007). Quantifying geocode location error using gis methods. *Environmental Health 6*, 1–8.

Szpiro, A. A., L. Sheppard, and T. Lumley (2010). Efficient measurement error correction with spatially misaligned data.

Taylor, S. M., J. P. Messina, C. C. Hand, J. J. Juliano, J. Muwonga, and A. K. Tshefu (2011). Molecular malaria epidemiology: Mapping and burden estimates for the Democratic Republic of the Congo, 2007. *PloS One 6*, e16420.

Ward, M. H., J. R. Nuckols, J. Giglierano, M. R. Bonner, C. Wolter, M. Airola, and P. Hartge (2005). Positional accuracy of two methods of geocoding. *Epidemiology 16*, 542–547.

Whitsel, E. A., P. M. Quibrera, R. L. Smith, D. J. Catellier, D. Liao, A. C. Henley, and G. Heiss. accuracy of commercial geocoding: Assessment and implications. *Epidemiologic Perspectives & Innovations* (1), 8.

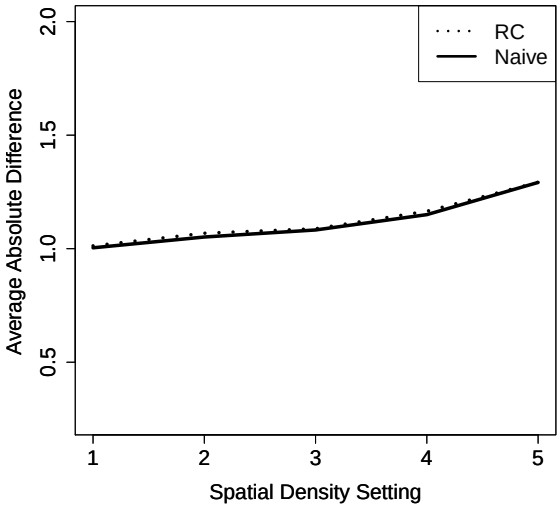
Zandbergen, P. A. (2007). Influence of geocoding quality on environmental exposure assessment of children living near high traffic roads. *BMC Public Health 7*, 37.

Zandbergen, P. A. and J. W. Green (2007). Error and bias in determining exposure potential of children at school locations using proximity-based GIS techniques. *Environmental Health Perspectives 115*, 1363.

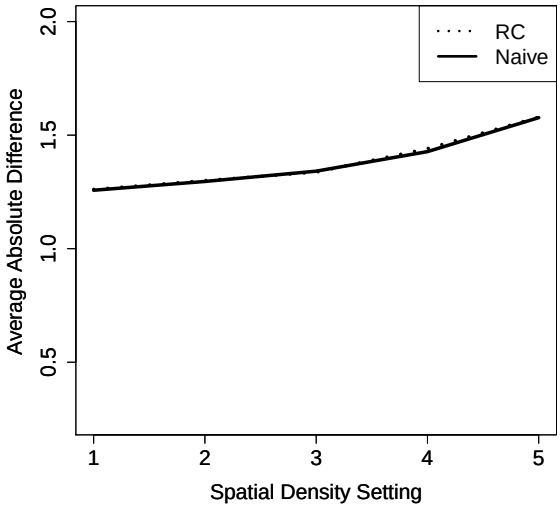
Zandbergen, P. A., T. C. Hart, K. E. Lenzer, and M. E. Camponovo (2012). Error propagation models to examine the effects of geocoding quality on spatial analysis of individual-level datasets. *Spatial and spatio-temporal epidemiology 3*, 69–82.

Appendix A

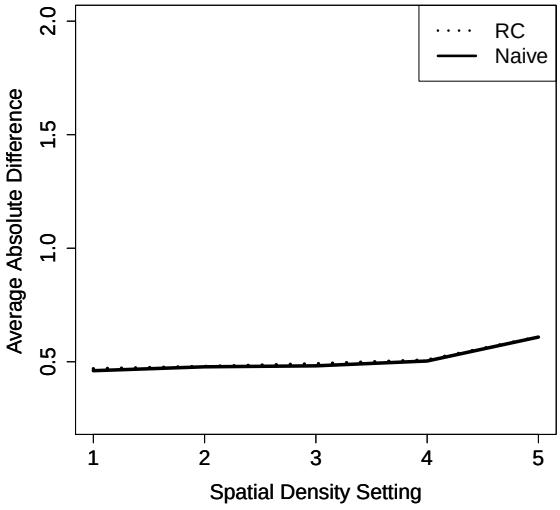
Supplementary Figures



(a) All Locations.

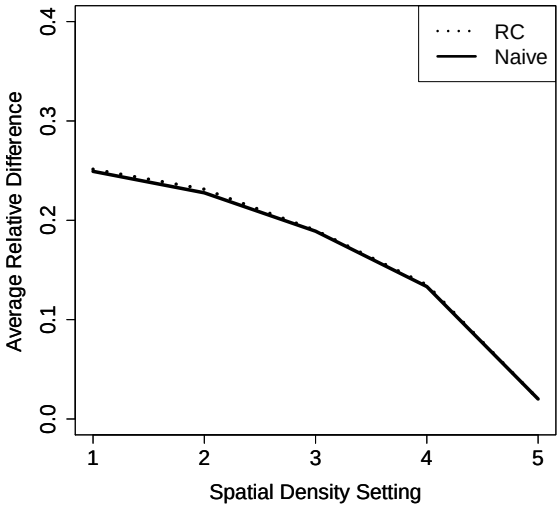


(b) Rural Locations.

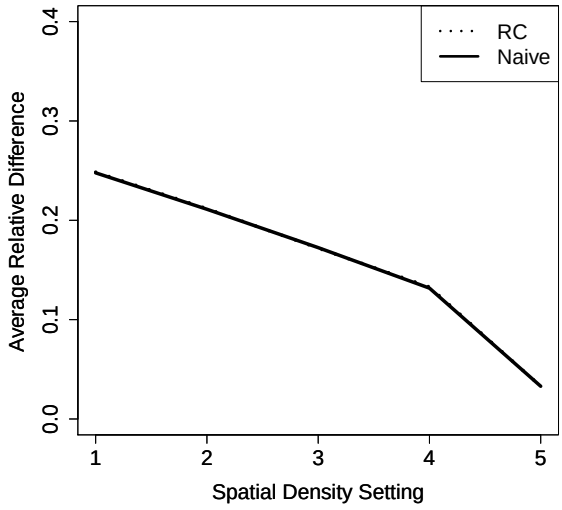


(c) Urban Locations.

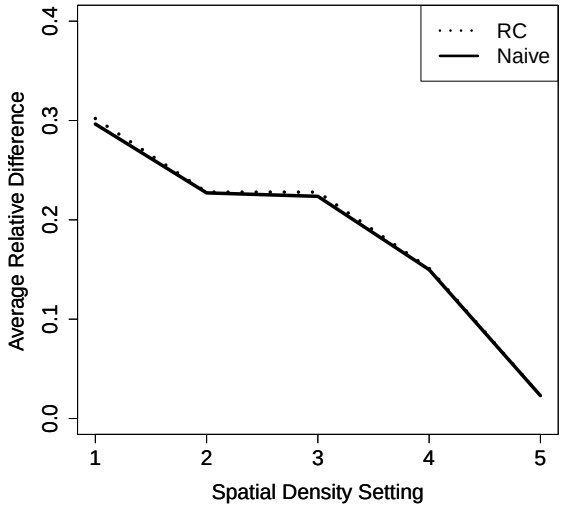
Figure A.1: Absolute Difference Results for Point Resource Locations (Shown on Same Scale).



(a) All Locations.

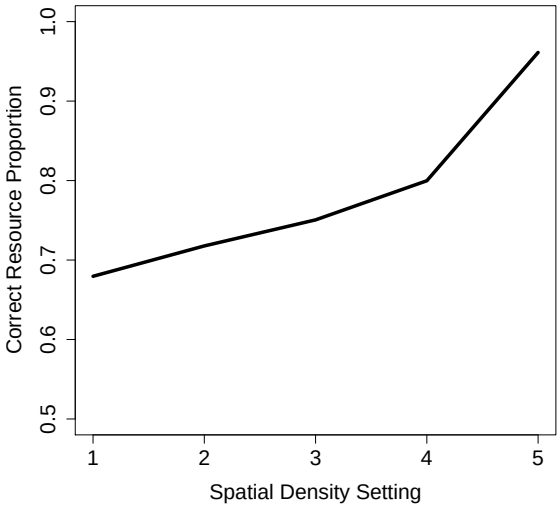


(b) Rural Locations.

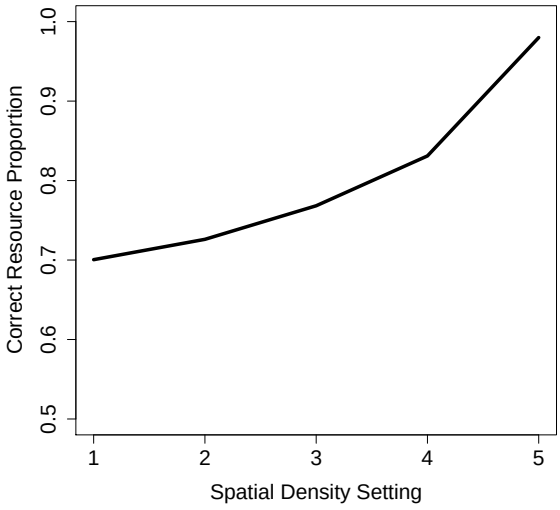


(c) Urban Locations.

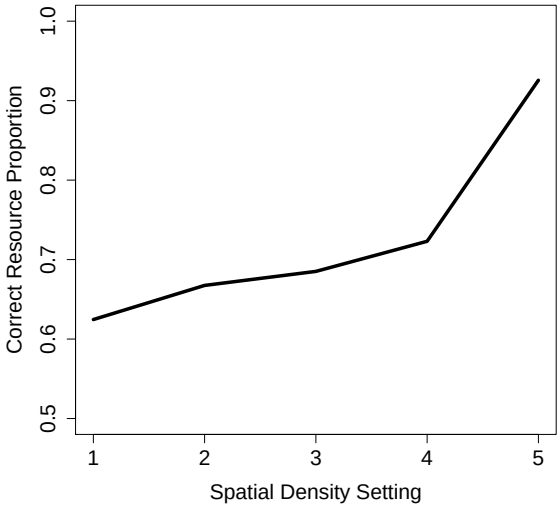
Figure A.2: Relative Difference Results for Point Resource Locations (Shown on Same Scale).



(a) All Locations.

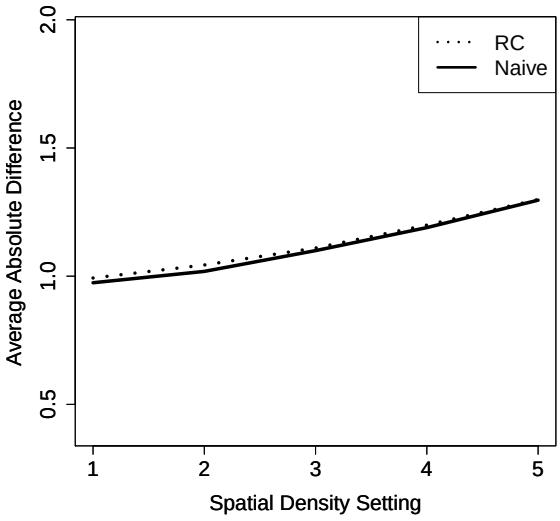


(b) Rural Locations.

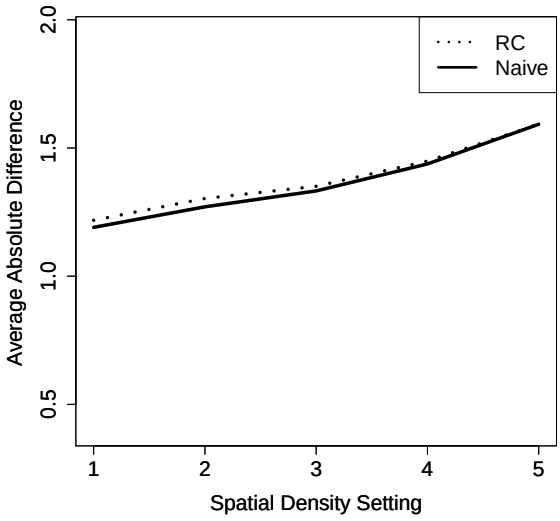


(c) Urban Locations.

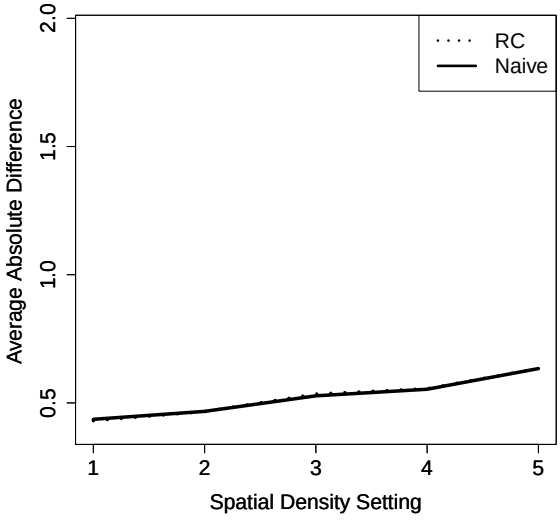
Figure A.3: Correct Resoure Proportions for Point Resource Locations.



(a) All Locations.

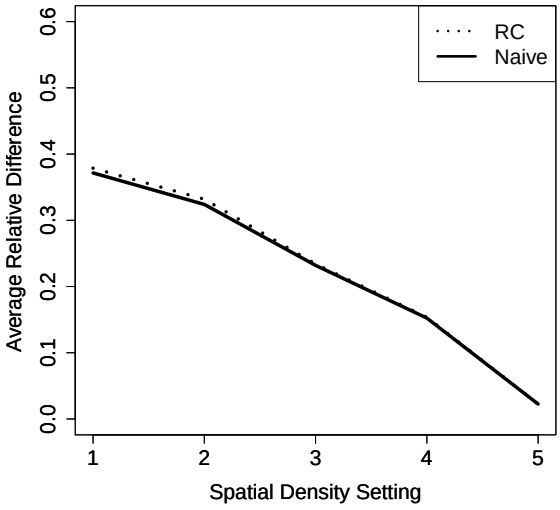


(b) Rural Locations.

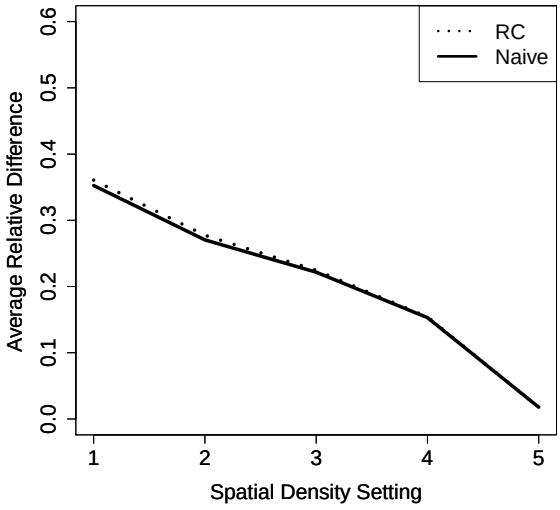


(c) Urban Locations.

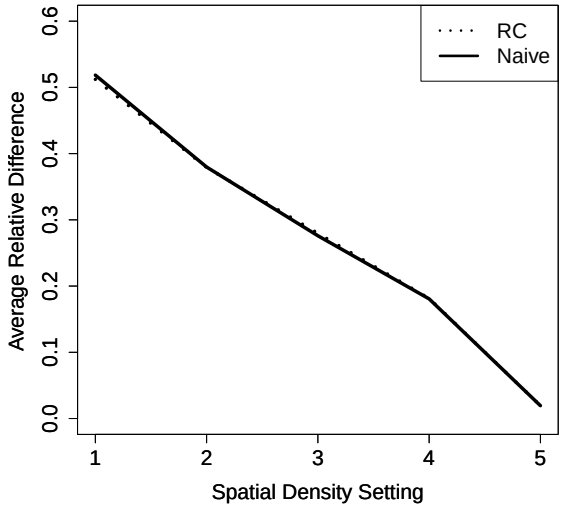
Figure A.4: Absolute Difference Results for Line Resource Locations (Shown on Same Scale).



(a) All Locations.

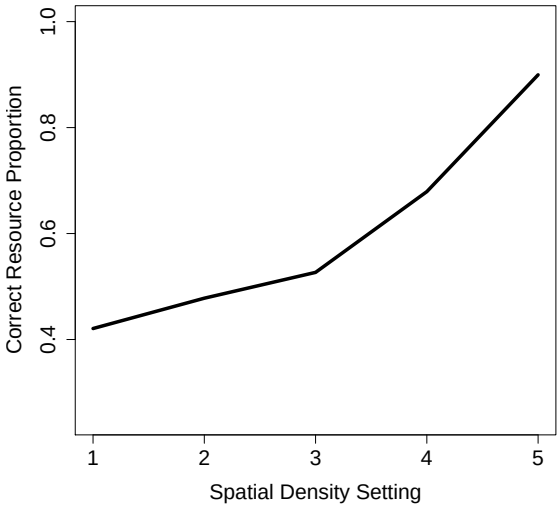


(b) Rural Locations.

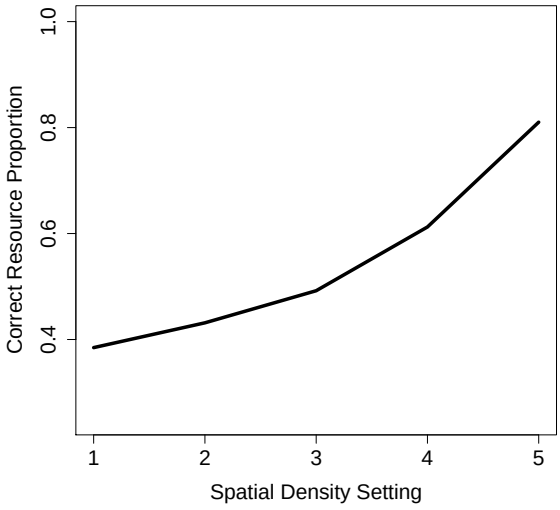


(c) Urban Locations.

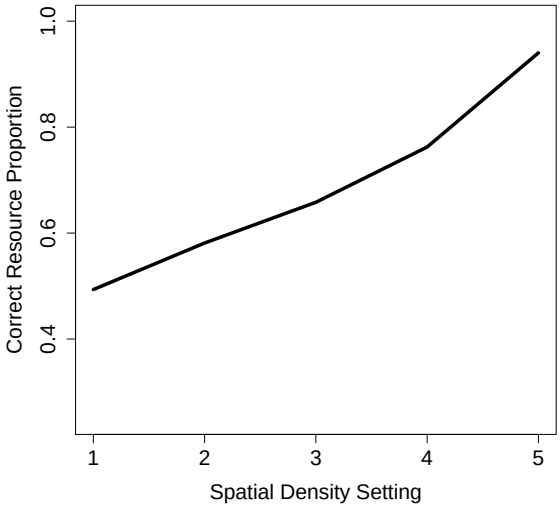
Figure A.5: Relative Difference Results for Line Resource Locations (Shown on Same Scale).



(a) All Locations.



(b) Rural Locations.



(c) Urban Locations.

Figure A.6: Correct Resoure Proportions for Line Resource Locations (Shown on Same Scale).

Appendix B

R code

B.1 Point Displacement Code

We begin by assigning maximum potential offset amounts to each original DHS location in the dataset with urban locations receiving a value of 2km and rural locations initially receiving a value of 5km. The value of every 100th rural location is then changed to 10km based on the current DHS offset program. Next, we create a buffer around each location using the specified maximum offset distance as the radius of the buffer. We then randomly generate 100,000 points within the buffer, thoroughly filling in the empty space. The process of creating the offset location is then completed by randomly selecting a single point from the generated points within the buffer. However, selecting from each point with equal probability leads to a randomly selected location with respect to the angle of offset, but does not lead to random uniform distances. This is due to the fact that locations in the outer part of the buffer have a higher probability of being selected because the area of that region is greater than the area closer to the center. As a result, we are more likely to select larger distances under an equal probability sampling scheme. To account for this, we weight the probability of selection based on distance from the buffer centroid. Selecting the offset location using these probabilities lead to a randomly selected offset angle and a random uniform offset distance as desired. The process of selecting a random point from the randomly generated 100,000 points is repeated 100 times for each location in the dataset. In other words, for each location (i.e., DHS cluster) we simulate 100 displaced points.

```
#pts:        The points to be offset.
#admin:      The administrative boundary of the region of interest.
             Points will not be offset outside of these boundaries.
#samp_num:   The number of times the points should be offset.
#other_num:  The number of buffer filling points, should be as large
             as possible (considering time constraints).  We recommend
             100,000 or larger.

#The function returns a list of length "samp_num".  Each entry is a matrix
#(n x 2) of offset locations, where n is the number of pts.

displace <- function(pts, admin, samp_num, other_num){

#Required Packages
require(rgdal)
require(maptools)
require(rgeos)
require(spatstat)
```

```
require(splancs)
require(fields)

n <- length(pts)
offset.dist <- ifelse(pts$URBAN_RURA == "U", 2000, 5000)
rural <- which(pts$URBAN_RURA == "R")
rur.n <- floor(0.01*length(rural))
offset.dist[sample(rural, rur.n, replace = FALSE)] <- 10000

r.pts0 <- list(0)
for(i in 1:nrow(pts)){
   r.pts0[[i]]<-matrix(0,nrow=samp_num,ncol=2)

   #-- Buffer around point --#
   pdsc <- disc(radius = offset.dist[i], centre = c(coordinates(pts)[i,1],
           coordinates(pts)[i,2]))
   pdsc <- as(pdsc, "SpatialPolygons")
   proj4string(pdsc) <- CRS("+proj=utm +zone=36 +datum=WGS84")

   #-- Intersection with admin --#
   ov <- over(admin, pts[i,])
   ov<-c(1:length(ov[[1]]))[is.na(ov[[1]])==0]
   poly <- admin[which(admin@data$OBJECTID == ov),]
   int <- gIntersection(pdsc, poly)

   #-- Generating random point
   if(!is.null(int)){
     rpt <- csr(int@polygons[[1]]@Polygons[[1]]@coords, other_num)
     probs<-1/rdist(coordinates(pts[i,]),rpt)
     rpt<-rpt[sample(c(1:other_num),size=samp_num,prob=(probs/sum(probs))),]
   r.pts0[[i]] <- rpt
     }
   if(is.null(int)){
     rpt <- csr(pdsc@polygons[[1]]@Polygons[[1]]@coords, other_num)
     probs<-1/rdist(coordinates(pts[i,]),rpt)
     rpt<-rpt[sample(c(1:other_num),size=samp_num,prob=(probs/sum(probs))),]
     r.pts0[[i]] <- rpt
     }
   }

#Arranging the Output
if(samp_num==1){
  r.pts<-list(0)
  r.pts[[1]]<-matrix(0,nrow=n,ncol=2)
  for(k in 1:n){
     r.pts[[1]][k,]<-c(r.pts0[[k]])
     }
  r.pts[[1]]<- SpatialPoints(r.pts[[1]], CRS("+proj=utm +zone=36 +datum=WGS84"))
  }

if(samp_num>1){
  r.pts<-list(0)
  for(j in 1:samp_num){
     r.pts[[j]]<-matrix(0,nrow=n,ncol=2)
     for(k in 1:n){
        r.pts[[j]][k,]<-r.pts0[[k]][j,]
        }
     r.pts[[j]]<- SpatialPoints(r.pts[[j]], CRS("+proj=utm +zone=36 +datum=WGS84"))
     }
```

```
  }
return(r.pts)
}
```

B.2 Regression Calibration

Regression calibration for linear and generalized linear models can be carried out using Stata Statistical Software. Replicates of the distance covariate are required to complete the analysis. To obtain replicates of the distance covariate, the offset function can be used numerous times and the distance covariate can be calculated for each set of offset locations. These replicate covariates can then be used by Stata to estimate the measurement error variance. More information is available at: `http://www.stata.com/merror/`.

B.3 Point-in-Polygon Analysis

The function returns a matrix object with rows pertaining to elements of the points object (in that order), and three columns corresponding to the weighted average (`weighted`), most probable value (`most.probable`), and misclassification rate (`misclass`).

```
#Output from Function Call:
#Matrix with one entry for each location:
#1)  Weighted Average of Polygon Values
#2)  Most Probable Polygon Value
#3)  Missclassification Error

#Notes:
#1)  Rural_Code={0,1};
     0: Maximum Rural Distance is 5,000m
     1: Maximum Rural Distance is 10,000m
#2)  n_Approximation;  We suggest 10,000 or larger.  Larger values will lead to
     improved approximations but will take more time.
#3)  NA_Option={0,1};
     0: Leaves missing polygon values as missing in weighted means
     1: Removes the missing polygon values and reweights the remaining polygon
     values to calculate the weighted means

#Function
point_in_polygon_fun<-function(Observed_DHS_Points,Polygon,Polygon_Values,
                               Rural_Code,n_Approximation,NA_Option){

#Packages
require(rgdal)
require(maptools)
require(rgeos)
require(spatstat)
require(splancs)
require(fields)

Observed_DHS_Points<-
spTransform(Observed_DHS_Points, CRS("+proj=utm +zone=36 +datum=WGS84"))
Polygon<-spTransform(Polygon, CRS("+proj=utm +zone=36 +datum=WGS84"))

#Assigning Offset Distances
n<-length(Observed_DHS_Points)
```

```
#Typical Case
if(Rural_Code==0){
  offset.dist<-ifelse(Observed_DHS_Points$URBAN_RURA=="U", 2000, 5000)
  }

#Worst Case Scenario
if(Rural_Code==1){
  offset.dist<-ifelse(Observed_DHS_Points$URBAN_RURA=="U", 2000, 10000)
  }

final<-matrix(0,nrow=n,ncol=3)
for(i in 1:n){

   #Creating Buffer Around Point with Maximum Offset as Radius
   pdsc<-disc(radius = offset.dist[i], centre = c(coordinates(Observed_DHS_Points)[i,1],
         coordinates(Observed_DHS_Points)[i,2]))
   pdsc<-as(pdsc, "SpatialPolygons")
   proj4string(pdsc) <- CRS("+proj=utm +zone=36 +datum=WGS84")

   #Filling in the Buffer with Points
   rpt<-csr(pdsc@polygons[[1]]@Polygons[[1]]@coords, n_Approximation)
   rpt<-SpatialPoints(rpt)
   proj4string(rpt)<-CRS("+proj=utm +zone=36 +datum=WGS84")

   #Which Region is each Point in?
   ov<-over(rpt, Polygon)
   ov<-ov[is.na(ov[,1])==0,]   #Removing Points Outside of the Polygon
   proportions<-matrix(0,nrow=length(unique(ov[,1])),ncol=2)
   for(j in 1:length(unique(ov[,1]))){

      #Proportion of Points in this Region
      proportions[j,1]<-mean(ov[,1]==unique(ov[,1])[j])

      #Polygon Value for the Proportion
      proportions[j,2]<-Polygon_Values[unique(ov[,1])[j]]
      }

   #Determining Polygon of Observed Point
   ov_point<-over(Observed_DHS_Points[i,], Polygon)

   #Leave Missing Polygon Values as Missing
   if(NA_Option==0){

     #Continous Polygon Value
     final[i,1]<-proportions[,1]%*%proportions[,2]

     #Discrete Polygon Value
     final[i,2]<-proportions[proportions[,1]==max(proportions[,1]),2]
     }

   #Reweighting the Non-Missing Polygon Values
   if(NA_Option==1){

     #Removing the Missing Observations
     proportions_1<-proportions[is.na(proportions[,2])==0,1]
     proportions_2<-proportions[is.na(proportions[,2])==0,2]

     if(length(proportions_1)>0){
```

```
        #Reweighting the Proportions
        proportions_1<-proportions_1/sum(proportions_1)

        #Continous Polygon Value
        final[i,1]<-proportions_1%*%proportions_2

        #Discrete Polygon Value
        final[i,2]<-proportions_2[proportions_1==max(proportions_1)]
        }

      if(length(proportions_1)==0){

        #Continous Polygon Value
        final[i,1]<-NA

        #Discrete Polygon Value
        final[i,2]<-NA
        }
      }

    #Probability of Missclassification
    final[i,3]<-1-proportions[(unique(ov[,1])==ov_point[,1]),1]

    #Completion Percentage
    print(c("Percent Complete", 100*round(i/n,2)))
    }

return(final)
}
```

B.4 Determining Spatial Autocorrelation Coefficient from Raster Data

In order to determine the level of smoothness, i.e. spatial autocorrelation coefficient, of an ancillary raster dataset, investigators can run the following R function, which calls the raster dataset as its sole argument. The function first converts raster grid cells to points, then extracts cell values from a subsample of those points/grids. Using a distance-based weights matrix, a simulataneous autoregressive model is fit to the data, and the estimated autocorrelation coefficient is extracted The function is given below.

```
rho.from.data <- function(ras){

require(spdep)
require(maptools)

#-- Generate Spatial Pixels data frame from raster file --#
r2r <- as(ras, "SpatialPixelsDataFrame")
r2p <- SpatialPixelsDataFrame(points=coordinates(r2r), data=r2r@data)

#-- Convert grid cells to points --#
rp <- SpatialPointsDataFrame(coordinates(r2p),data=r2p@data)

#-- Take a random subsample of points --#
rpsamp <- spsample(r2p, 1000, type = "regular")
```

```
#-- Extract cell values --#
test <- extract(ras, rpsamp, method = "simple")

#-- Generate projected Spatial Points Data Frame object from subsample --#
rpsamp.df <-SpatialPointsDataFrame(coordinates(rpsamp), data = as.data.frame(test))
proj4string(rpsamp.df) <- CRS("+proj=utm +zone=36 +datum=WGS84")

#-- Generate neighborhood adjacency, row-standardized distance-based(100km)weights matrix --#
dnb <- dnearneigh(coordinates(rpsamp.df), 0, 100000)
dlist <- nbdists(dnb, coordinates(rpsamp.df))
idlist <- lapply(dlist, function(x){1/x})
w100k <- nb2listw(dnb, glist = idlist, style = "W", zero.policy = TRUE)

#-- Obtain autocorrelation coefficient via a SAR model --#
sar <- lagsarlm(test ~ 1, data = rpsamp.df@data, listw = w100k)

sar$rho

}
```

B.5 Calculating Percentage of Cover

The percentage of cover (percent cover) within a prespecified neighborhood buffer can be calculated via the `extract` function from the `raster` package, which takes the following as arguments: a binary or continuous raster object (`ras`), points object (`pts`), and neighborhood buffer radius for each of the points in the dataset (`radius`).

```
extract(raster, pts, buffer = radius, fun = function(x){mean(x, na.rm = TRUE)})
```

B.6 Misclassification Rates in Discrete Rasters

The following function will return a vector of misclassification rates, i.e., probability that the binary cell value associated with true point location is not equal to the observed value of the raster grid cell within which the observed point resides, for each point provided in the `pts` object specified. The arguments of this function are:

- `ras`: binary raster file (as a `*Raster` class object)
- `pts`: displaced points object
- `urban`: binary vector indicating whether a DHS cluster point was designated as urban

```
misclass <- function(ras, pts, ru){
ptval <- extract(ras, pts, method="simple")
ptval[which(is.na(ptval))] <- 0
x <-  extract(ras, pts, buffer=ifelse(ru=="u", 2000, 5000), fun = function(x){mean(x, na.rm = TRUE)})
mc <- numeric(length(ptval))
mc[which(ptval==1)] <- 1-x[which(ptval==1)]
mc[which(ptval==0)] <- x[which(ptval==0)]
mc
}
```

The function returns a vector of misclassification rates. In the case of zero-inflated misclassification rates, investigators could summarize these rates as the proportion of non-zero rates (as was done

in 4.6.1); however, when misclassification rates are not zero-inflated, investigators could summarize these rates in terms of medians (as was done in the case study in 4.8.1).

Appendix C

Regression Calibration

Regression calibration for linear and generalized linear models can be carried out using Stata Statistical Software. Replicates of the offset distance covariate ($x_i^{(0)}$) are required to complete the analysis. To obtain replicates of $x_i^{(0)}$, the displacement function can be used numerous times and $x_i^{(0)}$ can be calculated for each set of offset locations. These replicate covariates can then be used by Stata to estimate the required model parameters. More information about creating these replicates can be seen below and more information regarding the Stata regression calibration function is available at: `http://www.stata.com/merror/`.

In regression calibration, we want to replace $x_i^{(t)}$ in the first stage regression model with

$$\mathrm{E}\left(x_i^{(t)}|x_i^{(0)}\right),$$

the conditional expected value of the true distance covariate given the displaced distance covariate, $x_i^{(0)}$. We define the model for the observed distance at DHS cluster i such that $x_i^{(0)} = x_i^{(t)} + u_i$ where u_i are independent and identically distributed errors with $\mathrm{E}(u_i) = 0$ and $\mathrm{Var}(u_i) = \sigma_u^2$. This model was selected after using the displacement function to simulate both variables and analyzing their relationship over numerous simulations. We assume that the $x_i^{(t)}$ variables are independent and identically distributed with $\mathrm{E}\left(x_i^{(t)}\right) = \mu_x$ and $\mathrm{Var}\left(x_i^{(t)}\right) = \sigma_x^2$. Jointly we have

$$\mathrm{E}\begin{pmatrix} x_i^{(t)} \\ x_i^{(0)} \end{pmatrix} = \begin{pmatrix} \mu_x \\ \mu_x \end{pmatrix} \text{ and } \mathrm{Cov}\begin{pmatrix} x_i^{(t)} \\ x_i^{(0)} \end{pmatrix} = \begin{bmatrix} \sigma_x^2 & \sigma_x^2 \\ \sigma_x^2 & \sigma_u^2 + \sigma_x^2 \end{bmatrix}.$$

The best linear unbiased predictor (BLUP) of $x_i^{(t)}$ is then given by

$$\mathrm{E}\left(x_i^{(t)}|x_i^{(0)}\right) = \mu_x + \frac{\sigma_x^2}{\sigma_u^2 + \sigma_x^2}\left(x_i^{(0)} - \mu_x\right).$$

We work with the empirical BLUP since μ_x, σ_x^2, and σ_u^2 are unknown and must be estimated.

We begin by working with the available DHS cluster locations, which have already been displaced, and treating these locations as the true DHS locations (non-displaced). We then offset these locations m times and for each displacement we calculate $x_{ij}^{(0)}$ $j = 1, \ldots, m$ for each cluster location $i = 1, \ldots, n$. Recall that $x_i^{(t)}$ is not changing with each displacement since we displace

the same locations with each run of the displacement function. In order to estimate σ_u^2, we calculate the sample variance of the

$$\left(x_{1j}^{(0)} - x_1^{(t)}, \ldots, x_{nj}^{(0)} - x_n^{(t)}\right)$$

values for each of the m displacements and take the average of these m values. To calculate σ_x^2 we calculate the sample covariance of

$$\left(x_{1j}^{(0)}, \ldots, x_{nj}^{(0)}\right) \text{ and } \left(x_1^{(t)}, \ldots, x_n^{(t)}\right)$$

for each of the m displacements and take the average of these m values. Assuming these parameters are now known, we then estimate μ_x using the generalized least squares estimator. Finally, we plug in these estimates to obtain

$$\widehat{x}_i^{(t)} = \widehat{\mu}_x + \frac{\widehat{\sigma}_x^2}{\widehat{\sigma}_u^2 + \widehat{\sigma}_x^2}\left(x_i^{(0)} - \widehat{\mu}_x\right).$$

These estimated distances are then used in the first stage regression model and the analysis is carried out as usual. In order to obtain standard errors for the estimated regression coefficients, we rely on bootstrapping techniques which sample with replacement from the original data and repeat the described analysis numerous times to obtain multiple regression coefficient estimates. The standard error is estimated by taking the sample standard deviation of the obtained values.