Improving Survey Methodology Through Matrix Sampling Design, Integrating Statistical Review Into Data Collection, and Synthetic Estimation Evaluation

Mark Thomas Seiss

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
In Statistics

Eric A. Vance, Committee Chair
Ralph P. Hall
Leanna L. House
Inyoung Kim

March 28, 2014
Blacksburg, Virginia

Keywords: Census Coverage Measurement (CCM), Computer Assisted Personal Interviewing (CAPI), Data Editing, Household Survey, Impact Evaluation, Matrix Sampling, Multiple Imputation, On-the-Ground Statistician, Respondent Burden, Split Questionnaire Design, Synthetic Bias, Synthetic Estimation, Variance Component Estimation

Improving Survey Methodology Through Matrix Sampling Design, Integrating Statistical Review Into Data Collection, and Synthetic Estimation Evaluation

Mark Thomas Seiss

ABSTRACT

The research presented in this dissertation touches on all aspects of survey methodology, from questionnaire design to final estimation. We first approach the questionnaire development stage by proposing a method of developing matrix sampling designs, a design where a subset of questions are administered to a respondent in such a way that the administered questions are predictive of the omitted questions. The proposed methodology compares favorably to previous methods when applied to data collected from a household survey conducted in the Nampula province of Mozambique. We approach the data collection stage by proposing a structured procedure of implementing small-scale surveys in such a way that non-sampling error attributed to data collection is minimized. This proposed methodology requires the inclusion of the statistician in the data editing process during data collection. We implemented the structured procedure during the collection of household survey data in the city of Maputo, the capital of Mozambique. We found indications that the data resulting from the structured procedure is of higher quality than the data with no editing. Finally, we approach the estimation phase of sample surveys by proposing a model-based approach to the estimation of the mean squared error associated with synthetic (indirect) estimates. Previous methodology aggregates estimates for stability, while our proposed methodology allows area-specific estimates. We applied the proposed mean squared error estimation methodology and methods found during literature review to simulated data and estimates from 2010 Census Coverage Measurement (CCM). We

found that our proposed mean squared error estimation methodology compares favorably to the previous methods, while allowing for area-specific estimates.

DEDICATION

To my mother, who has always encouraged me to continue my education and has led by example, and my father and sister for their support.

ACKNOWLEDGEMENTS

I would first like to express my deepest gratitude to my adviser, Dr. Eric Vance, for his guidance in all aspects of my education and professional career. Over the past 6 years, his mentoring advice related to the importance of the communication of statistics has been invaluable to my professional development. Along with Dr. Ralph Hall, he gave me the opportunity to work and travel around the rural villages of Mozambique, an invaluable experience both personally and professionally. He has allowed me the freedom to incorporate my research interests from my work at the Census Bureau and my studies at Virginia Tech into my PhD research. My career in statistics will continue to benefit from all that I have learned from Dr. Vance for many years to come.

Many thanks to my committee members, Dr. Ralph Hall, Dr. Leanna House, and Dr. Inyoung Kim, for their tutelage and insightful discussions on my research. Dr. Hall gave me the opportunity to be involved in all aspects of the household survey in Mozambique, an experience I will remember forever. Dr. House and Dr. Kim were two of the most influential teachers during my PhD studies at Virginia Tech. I would like to thank the faculty of the Virginia Tech Statistics Department for providing me with substantial statistical knowledge and rigorous statistical training, the Laboratory of Interdisciplinary Statistical Analysis (LISA) for providing statistical collaboration training and the ability to collaborate with graduate students and faculty from a variety of backgrounds on 117 collaborative projects, my fellow students for the memories and insightful discussions over the years, the staff for all of their support, and Dr. Jean Gibbons for her financial support.

I would like to thank my co-workers at Census Bureau for all of their patience while I split my time between work and school. I would like to thank Inez Chen, Tom Mule, Pat Cantwell from the Decennial Statistical Studies Division (DSSD) and Bill Bell and Jerry Maples from the Center of Statistical Research Methodology (CSRM) for allowing me to include my synthetic bias research at the Census Bureau in my dissertation and their contributions to the research. In particular, I would like to thank Inez for reviewing all of my work and her support over the years.

The field work conducted in Mozambique was undertaken as part of a Stanford University research program on non-network water and sanitation in developing countries funded by the Woods Institute for the Environment. I would like to thank Dr. Jenna Davis, Dr. Valentina Zuin, and Maika Elizabeth Nicholson of Stanford University for the opportunity to collaborate on the project and their comments; Marcos Carzolio of Virginia Tech for his participation in the project; and Dr. J.P. Morgan of Virginia Tech and Dr. David Banks of Duke University for their comments.

Finally, I would like to thank my family and friends. Without their support, this achievement would not have been possible.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 2

CHAPTER 3

CHAPTER 4

LIST OF TABLES

CHAPTER 2

CHAPTER 3

CHAPTER 4

APPENDIX B

APPENDIX D

APPENDIX E

CHAPTER 1: INTRODUCTION

1. Overview

Sample surveys are widely used to provide estimates of population quantities such as totals and means. Many countries have set up centralized statistical agencies that collect statistical information about the state of the nation. This important statistical information includes national characteristics such as the demography, agriculture, labor force, health and living conditions, and trade. Government agencies increasingly use these results to formulate policies and allocate government funds. Sardanal et al. (1992) attributes much of the developments in sample survey methodology to government statistical offices.

The use of sample surveys has also extended to the academic and private business sectors. Private sector businesses have increasingly used the results of surveys to influence business decisions. Rao (2003) explains that this use by private sector businesses is attributed to their heavy reliance on local conditions. The academic sector uses sample surveys in many research areas. The statistical collaborators at the Laboratory of Interdisciplinary Statistical Analysis (LISA) at Virginia Tech have worked with sample survey data collected by graduate students and faculty from a large range of disciplines, including education, engineering, public and international affairs, and agriculture.

Due to this increased reliance on sample survey data, the quality of the data collected from these surveys and the estimates produced from them have increased in importance. The research presented in this dissertation proposes methodology that seeks to improve the quality of the

collected data and estimates produced. This proposed methodology relates to all facets of sample surveys, from questionnaire design to data collection to the evaluation of estimates.

2. Organization of Dissertation

The larger acceptance of survey methodology has led to increasingly larger surveys with the need for more detailed data. With this increase in length, increased burden due to the increased time of application is placed on the respondent and enumerator, which may lead to higher rates of non-response, higher rates of premature termination, and inaccurate responses. The increased length of the survey also results in higher financial costs. One solution to the increased length of surveys is called a matrix sampling design, also known as a split questionnaire design, where respondents are only administered a subset of the questions from the questionnaire. The idea behind these designs is that the questions are administered to the respondents in such a way that the collected information is predictive of the information not collected in the omitted questions. In Chapter 2, *Stochastic Search Variable Selection Application to Matrix Sampling*, we apply a variation of Stochastic Search Variable Selection (SSVS) to design matrix samples.

Developments in mobile computing have given researchers increased capabilities in the field of data collection. One of these capabilities is computer-assisted personal interviewing (CAPI) where the collected data is instantaneously accessible for analysis. De Waal (2011) states that the main advantage of CAPI is the ability to start the data editing process during data collection by informing the surveyor of possible errors as they are collected. In many situations though, the software necessary to implement this real-time editing is not available. In Chapter 3, *The Importance of Cleaning Data During Fieldwork: Evidence from Mozambique*, we propose data

editing procedures that seek to achieve many of the benefits of real-time editing in spite of the limitations of the software. These proposed procedures require the involvement of statisticians in all aspects of the survey process, rather than just the post processing phase. This total involvement has the added benefit of producing a statistician with subject matter expertise, rather than only a specialist in data analysis.

In previous years, policy makers and researchers were content with nationwide estimates of quantities of interest that result from surveys. In recent years, there has been an increased need for estimates of specific domains, or sub-populations defined by divisions such as geography or demographics (Rao 2003). In some cases, these sub-populations may contain very little or no sample. In keeping with common usage, we refer to these small domains with little or no sample as small areas. Synthetic estimation, also known as indirect estimation, is often used by researchers to provide reliable estimates for small areas. These reliable estimates are achieved by "borrowing" sample from other areas included in the data. Synthetic estimation is conducted under the synthetic assumption, or the assumption that there is no variation in the quantities of interest for areas with similar characteristics. This assumption is violated when areas exhibit strong area-specific effects not captured by the model that produced the synthetic estimates. The potential error that is introduced by the violation of the synthetic assumption is known as synthetic estimation error.

Previous methods have viewed synthetic estimation error as fixed and a bias. The mean squared error (MSE) of the synthetic estimate is then the sum of the sampling variance of the synthetic estimate and the squared synthetic bias. In cases where the synthetic estimates pool data across a

large number of areas, the sampling variance of the synthetic estimate may be small, and the squared synthetic bias component may dominate the MSE. As Rao (2003) notes, while the estimates of the sampling variance of synthetic estimates are readily obtained, the squared synthetic bias component is more difficult to estimate. Since estimates of synthetic estimation error are generally unstable, these previous methods provide aggregate estimates over groups of areas and assume constancy within the groups. In some cases, there is the need for area-specific measures of synthetic bias. One of these instances was the release of estimates within states, counties, and places of the U.S. during the 2010 Census Coverage Measurement (CCM) by the US Census Bureau. During previous coverage evaluations, the US Census Bureau only estimated the variance of these synthetic estimates, but area-specific estimates of synthetic bias were also released during the 2010 CCM. In Chapter 4, *Model Based Approach to Synthetic Bias Estimation*, we propose model-based methodology of producing area-specific measures of synthetic bias.

Chapter 5 provides a discussion of the important results of the preceding three chapters. Included in this chapter are recommendations for future work that build upon the presented research.

References

De Waal, T., Pannekoek, J., and Scholtus, S. (2011), *Handbook of Statistical Data Editing and Imputation*, Hoboken, NJ: John Wiley and Sons.

Rao, J.N.K. (2003), *Small Area Estimation,* Hoboken, NJ*:* John Wiley and Sons, Inc.

Sarndal, C. E., Swensson, B., and Wretman, J. (1992), *Model Assisted Survey Sampling,* New York: Springer-Verlag.

CHAPTER 2: IMPROVED MATRIX SAMPLING DESIGNS USING A BAYESIAN VARIABLE SELECTION APPROACH

Chapter 2 Abstract

In recent years, the demand for quantitative data from surveys has grown as researchers and policymakers require more information to make informed decisions. With the increased demand, the length of the survey and time to administer it have also increased, imposing increased burden on both the respondent and the enumerator, which leads to higher rates of non-response and premature termination. Matrix sampling designs are a solution to these problems by administering a subset of the survey questions to each respondent in a manner such that the applied questions are predictive of omitted questions. Imputation methods, such as multiple imputation, can then be used to impute values for the omitted questions. The imputation of missing values resulting from a matrix sampling design is preferable to that of non-response or premature termination, since the missing data mechanism is known. This manuscript proposes an application of Stochastic Search Variable Selection (SVSS) as an improved method of creating matrix sampling designs. Our method is applicable to all types of variables and requires less pre-specification than previous methods. We applied our methodology to simulations based on household survey data collected in the Nampula province of Mozambique. These simulations have led to two prominent conclusions. First, we find that our proposed method generally performs better than previous methods. Second, the multiply imputed data resulting from our matrix sampling designs generally produces higher quality data than the application of the full survey to a smaller number of respondents. These improvements come in addition to the reduction in surveyor and enumerator burden that results from shorter surveys.

1. Introduction

Researchers in many fields administer surveys to subsets of the population, identified through probability sampling, to collect data representative of the entire population. In recent years, the use of surveys has increased enormously as policy makers have increased their demand for quantitative data to influence policymaking (Rao 2003). As this demand for quantitative data has increased, the lengths of the questionnaires and the time and cost to administer them have increased. Adams and Gale (1982) and Berdie (1989) showed that longer questionnaires result in decreased data quality due to increased respondent and enumerator burden and higher rates of non-response and premature termination. A fatigued respondent or enumerator introduces non-sampling error in the collected data reducing the precision of the survey estimates (Thomas et al. 2006).

One remedy to the increasing costs and burden on the respondent and enumerator is to implement a matrix sampling design, also known as split questionnaire design. Under this design, the enumerators administer a subset of the questionnaire questions to the respondents. The decrease in the total number of questions administered to each respondent should decrease the burden on the respondent and enumerator and reduce the rates of non-response and premature termination that are associated with longer questionnaires. As Raghunathan and Grizzle (1995) and Thomas et al. (2006) note, the missingness assigned with matrix sampling designs is Missing at Random (MAR) or Missing Completely at Random (MCAR), since the missing data mechanism is known. The missing data mechanism of matrix sampling designs is preferable to that of either non-response or premature termination, in which the post-survey data analyses

require unverifiable assumptions about the missing data mechanism (Raghunathan and Grizzle 1995).

A good matrix sampling design selects questions to be included for a given respondent such that the collected data are predictive of the uncollected data associated with the omitted questions (Thomas et al. 2006). Using the collected data, researchers can impute values for the uncollected data, and minimize the efficiency lost from not asking the omitted questions. Multiple imputation (Rubin 2009) is well-suited for this situation (Thomas et al. 2006). We use a multiple imputation method called chained regression equations (Raghunathan et al. 2001). Other multiple imputation methods include Markov Chain Monte Carlo (Schafer 1997), propensity scoring (Rosenbaum and Rubin 1983), and predictive mean matching (Little 1988).

In the development of matrix sampling designs, we partition survey questions into core and split questions, adopting the terminology used in Thomas et al. (2006). Each matrix sampling design contains all core questions and a subset of split questions. The variables created from the core and split questions will be referred to as core and split variables. Researchers classify questions as core questions for two reasons. First, the researcher requires the information obtained by the question from every respondent. Second, the variable created from the information obtained by the question is a strong predictor of many other variables, and the inclusion of the question improves the prediction for many split questions. Split questions are important to the goals of the study, but the information collected by the question is not required from every respondent. The information not obtained from the respondent due to the omission of split questions can be regained by predicting the values using data from the administered core and split questions.

In practice, a training sample must be available to determine which set of variables are predictive of each other, since the matrix sampling designs are determined prior to data collection. One could use data from a survey that is similar to the current survey as the training sample. Another option would be to administer the complete survey to a subset of the sample during the initial implementation of the survey. For example, if one were to create a matrix sampling design for a survey related to water use in Mozambique, he or she could use previously obtained data from a similar study in Senegal or Kenya to create the matrix sampling design. One could also administer the whole survey for the first week and use the collected data to create the matrix sampling design for the remainder of the survey.

The following manuscript proposes new methodology that improves previous matrix sampling methodology by implementing a Bayesian variable selection technique. Previous methodology requires pre-specification of the way that we partition split variables for the assignment of questions to respondents. This pre-specification may not allow for the optimal assignment, which recovers the most information about the omitted questions when multiple imputation is used. Our methodology is less restrictive than these previous methods, allowing for the optimal assignment based on the training data set.

We will simulate survey implementations using data collected from a household survey administered in the Nampula province of Mozambique. These simulations will first show that our proposed methodology outperforms previous methodology. Second, the simulations will show that the multiply imputed data resulting from a matrix sampling design using our proposed methodology is better (based on root mean squared error) than administering the full

questionnaire to a smaller number of respondents. This improvement over the full questionnaire does not include the reduction in surveyor and enumerator burden that results from shorter surveys. Based on the research by Adams and Gale (1982) and Berdie (1989), we would expect further improvement in the outcomes associated with matrix sampling designs due to this reduction. These findings suggest that researchers would collect higher quality data by using matrix sampling designs, and our proposed methodology should be the methodology used for allocation of questions.

We outline the manuscript as follows. Section 2 describes multiple imputation methods and their application in this research. Section 3 describes previous research found during literature review. In Section 4, we propose a new method of assigning matrix sampling designs. Section 5 applies the proposed methodology and a method found during literature review to the simulated household data. Section 6 discusses conclusions that were drawn from the research.

2. Multiple Imputation

With survey data, missing values are common due to non-response (Rubin 1996). Many standard statistical procedures only perform analysis on complete case data, or observations without missing values. These procedures ignore observations with missing values, which may contain important information. In the case that there are systematic differences between complete cases and incomplete cases, the inferences resulting from analyzing only complete cases may not be valid for the complete dataset of both observed and missing values (Little 1988).

Another approach for handling data with missing values is to impute a single value. After imputation, a complete dataset is created with no missing values, and standard statistical procedures may be used to analyze all respondents. The missing data model accounts for possible systematic differences between complete and incomplete cases, but treats the missing values as known quantities. The imputed values from the missing data model have uncertainty associated with them, which is not taken into account by single imputation methods. Rubin (1987) shows that single imputation biases the variances associated with the estimated parameters toward zero.

Rubin (1987) proposed multiple imputation, an alternate procedure for dealing with missing values that is motivated by both Bayesian and Frequentist paradigms. Under the Bayesian paradigm, missing values are unknown quantities, which are estimated using posterior predictive distributions. By imputing a sampled value from the posterior predictive distribution of each missing value, we create a single complete dataset, or dataset with all observations and no missing values. Sampling from the posterior predictive distributions is repeated multiple times to produce multiple complete datasets. Each of the complete datasets is then analyzed using standard statistical procedures, which may include Frequentist methods. The results from the analysis of each complete dataset are combined to make valid inferences that properly reflect the possible differences between complete and missing observations and the uncertainty involved with the imputation of the missing data. Rubin (1987) showed that even in extreme cases of missing values in datasets, no more than 5 complete datasets are needed for efficient estimates and standard errors.

The research presented in this manuscript utilizes multiple imputation to deal with missing values resulting from matrix sampling designs using the R statistical software (RDC Team 2009) and its MICE package (Buuren and Groothuis-Oudshoorn 2011). Van Buuren and Groothuis-Oudshoorn (2011) provides examples of applying the functions in this package to survey data with missing values. The *mice* function in the MICE package performs multiple imputation for data containing missing values. The function uses a technique called chained equations, in which a univariate imputation model is specified for each variable in the dataset. The algorithm begins by imputing missing values with initial values, which may or may not be specified by the user. For each variable, the specified univariate model is fit to observed values of the given variable conditioned on the observed and imputed values of the other variables. The *mice* function contains multiple built-in univariate models, but also allows user-defined univariate models. The fitted model is then used to estimate the unobserved values and replace the initial imputed values. The algorithm sequentially and repeatedly imputes each variable in this manner for a user-specified number of iterations (default is 5). The user should analyze the output from the *mice* function to determine whether enough iterations were specified for convergence to be reached.

The *mice* function allows the user to specify the number of imputed data sets to be created (default is 5). The function executes a stream of chained equations for each imputed data set in parallel, and outputs the values of the final iteration for each stream. The MICE package contains functions that pool together the multiply imputed data sets in the manner suggested by Rubin (1987). See Raghunathan et al. (2001) for more detailed information on chained

regression equations and Appendix B for an example. See Buuren and Groothuis-Oudshoorn (2011) for more detailed information on the MICE package and its functions.

3. Previous Methodology

Dating back to the 1950's, the earliest applications of matrix sampling involve research in the education field. These early applications involved administering a subset of test items to each examinee. Hooke (1956) and Lord (1960) derived equations that approximate the distribution of examinee test scores that would have been generated had all examinees been administered all questions. Shoemaker (1973) summarized previous work that developed statistical procedures for estimation based on matrix sampling designs and outlined a procedure for implementation.

Navarro and Griffin (1993) conducted research leading up to the 2000 Decennial Census, determining whether the application of a matrix sampling design was a viable alternative to administering the full long form questionnaire to a sample of households. For the 2000 Decennial Census, the Census Bureau increased the length of the long form questionnaire to meet the demand for more detailed household information. The goal of the research was two-fold. The first goal was to maintain the level of reliability of the 1990 Census, while collecting additional information. The authors measured the reliability of sample estimates by the coefficient of variation. The second goal was the reduction of respondent burden from the 1990 Census, thus improving response rates and data quality. Respondent burden was crudely measured as the average number of questions administered to the population, reported as a percentage of the number of questions administered during the 1990 Census. For example, one of the matrix sampling designs specified three matrix sampling forms, each administered to 6.7%

of the population. Two of the forms contained 48 questions, while the third contained 36 questions. The respondent burden for this matrix sampling design was 8.84 (0.067(36+48+48)) and 9.35 (0.167(56)) for the 1990 Census (the long form sampling rate in 1990 was 16.7%). The respondent burden of the matrix sampling design was 95% of the 1990 Census burden.

The research compared five different matrix sampling designs, assigned based on varying the number of questions administered and the sampling fraction. The assignment of variables was not based on any type of analysis of the relationship between the variables; simply whether they were asked in the same section. The authors concluded that the data based on the five matrix sampling designs provided reliable estimates for small area estimation and, in all but one of the designs, reduced respondent burden. In the future work section, Navarro and Griffin (1993) included the use of correlation analysis for the creation of matrix sampling designs.

Raghunathan and Grizzle (1995) generated matrix sampling designs based on correlation analysis using data from the Cancer Risk Behaviour Survey questionnaire. The authors determined that multiple imputation methods only perform better than complete case methods in terms of efficiency when strong correlations between split variables are present. While there may not be an increase in efficiency, the authors suggested that multiple imputation methods may still be preferable, due to the limitations that using only observed data may have on data analysis.

For more information on the history of matrix sampling and a review of previous applications, see Gonzalez and Eltinge (2007).

Thomas et al. (2006) developed an index that measures the efficiency recovered by multiply imputing a given split variable using information from another split variable rather than using only observed values. The index compares this recovered efficiency to a baseline measure, the efficiency gained by using the complete dataset (dataset with all values observed) rather than the incomplete dataset (dataset with missing values). For core variables X and split variables Y_j and $Y_{j`}$, Thomas et al. (2006) defines the index for $Y_j$ given $X$ and $Y_{j`}$ as the following.

$$I(Y_j|X, Y_{j`}) = \frac{V_{NI} - V_{MI}}{V_{NI} - V_{COMP}} \quad (1)$$

where,

$V_{COMP} = \frac{V(Y_j)}{n_{total}}$ = Variance of Y_j using the complete dataset

$V_{NI} = \frac{V(Y_j)}{n_{obs}}$ = Variance of Y_j using the incomplete dataset

V_{IMP} = Variance between multiply imputed datasets

$V_{MI} = V_{COMP} + V_{IMP}$ = Variance of Y_j using the multiply imputed dataset

$n_{total} = n_{obs} + n_{mis}$ = Total number of records in the complete dataset; equal to the sum of the number of observed records and the number of missing records.

The index takes on values between 0 and 1, where 1 indicates that X and $Y_{j`}$ perfectly predict $Y_j$ ($V_{MI}$=$V_{COMP}$) and 0 when $X$ and $Y_{j`}$ are no improvement over the no imputation estimator (V_{NI}=V_{MI}). The higher the index value, the better the predictor that split item $Y_{j`}$ is of Y_j.

Schafer and Schenker (2000) derived an approximation of the variance between multiply imputed datasets for continuous and binary variables. Thomas et al. (2006) approximate the variance of Y_j ($V(Y_j)$) using a training dataset, such as data from a similar survey or an initial implementation of the survey.

Using their index, Thomas et al. (2006) develops an algorithm to assign split variables to blocks. Blocks, in this case, refers to groups of variables, which are selected as a group to be observed for a given respondent. The resulting questionnaire administered to the respondent is a random selection of blocks of variables. The authors define four goals for the assignment of split variables to these blocks.

1. Assign each split variable to a single block
2. Assign an approximately equal number of split variables to each block
3. Assign logically linked items to the same block (skip items)
4. Assign one or more items to each block that predicts items omitted from the block.

The authors define a criterion to evaluate how well a given assignment of variables achieves these four goals and a simple algorithm for achieving an approximately optimal assignment.

Thomas et al. (2006) applies this index and assignment algorithm to data from the National Health and Nutrition Examination Survey (NHANES). The authors used NHANES II data as the training dataset to calculate the index between split variables and assign variables to blocks. The resulting matrix sampling design was applied to NHANES III data with simulated missing data and imputation using multiple imputation. The findings include large losses in efficiency by using matrix sampling designs rather than administering the entire questionnaire. These findings were in contrast to the findings in publications up to that point, such as Raghunathan and Grizzle (1995), which recovered much of the information lost. Thomas et al. (2006) discusses how this emphasizes the importance of having variables that are strong predictors of each other.

4. Bayesian Variable Selection Approach to Matrix Sampling

Stochastic Search Variable Selection (SSVS) is a Bayesian model formulation originally developed in George and McCulloch (1993), used as a variable selection tool for covariates in a regression model. Unlike Bayesian model selection, which includes variables based on the model drawn from a probability distribution, SSVS includes all covariates in the model, but specifies the prior distribution of the regression coefficients to be a mixture of two normal distributions. If a variable is selected as a strong predictor of the response variable, the prior distribution of the corresponding regression coefficient is non-informative, assigning relatively equal weight to all possible values a priori. Otherwise, if the variable is not selected as a strong predictor, the prior distribution is peaked at zero, concentrating the prior weight around zero. The resulting posterior draws for the regression coefficients for variables not included in the model will therefore be relatively equal to zero. We will use a variation of SSVS to identify mutually exclusive groups of variables that are strong predictors of one another and create a matrix sampling design based on these groupings.

The following sections describe a modified version of the SSVS model formulation that will be applied to assign variables for matrix sampling designs. Section 4.1 introduces the notation that will be used in the description of the methodology. Section 4.2 describes the assumed sampling distribution of the data. Bayesian estimation was used to estimate unknown parameters in the sampling distribution. Section 4.3 describes the prior distributions specified for these unknown parameters. Section 4.4 outlines the sampling of these parameters using the Metropolis-Hastings algorithm. Section 4.5 describes the use of cluster analysis to assign variables to groups based on the posterior distributions produced by the Metropolis-Hastings algorithm. Section 4.6

describes how the results from the cluster analysis will be used to implement the matrix sampling design.

4.1 Parameters

Table 2.1 defines the notation used to represent the observed data, parameters, and indices in the model formulation proposed in this section.

Table 2.1: Definition and Notation of Statistical Quantities

Statistical Quantity	Notation	Dimension	Description
Indices			
Observation	I	N	Index indicating questionnaire respondent
Split Variable	J	J	Index indicating split variable number
Core Variable	K	K	Index indicating core variable number
Observed Data			
Matrix of Split Variables	Y	NxJ	Matrix containing the J split variables that may or may not be selected for a given questionnaire.
Matrix of Core Variables	X	$Nx(K+1)$	Matrix containing the K core variables selected for every questionnaire and a vector of ones.
Parameters			
Core Variable Regression Coefficients	θ	$(K+1)xJ$	Matrix containing J columns of $K+1$ regression coefficients corresponding to the intercept and each core variable. The j^{th} column contains the $K+1$ regression coefficients corresponding to the intercept and core variables for the model fitting the j^{th} split variable.
Split Variable Regression Coefficients	β	$(J-1)xJ$	Matrix containing J columns of $J-1$ regression coefficients corresponding to the split variables (omitting the split variable used as the response variable). The j^{th} column contains the $J-1$ regression coefficients corresponding to the split variables (omitting the j^{th} split variable) for the model fitting the j^{th} split variable.
Normal Likelihood Variance	σ^2	$J_{cont}x1$	Variance term of the normal likelihood for continuous split variables, where J_{cont} is the total number of continuous split variables.
Group Assignment Matrix	Δ	JxJ	Matrix of indicator variables, where the ij^{th} entry specifies that the i^{th} and j^{th} split variables are assigned to the same group.
Group Assignment Probability	α	JxJ	Matrix containing the probabilities of the inclusion of one variable in the same group as another, where the ij^{th} entry represents the probability that the i^{th} and j^{th} split variables are assigned to the same group.

4.2 Likelihood

Prior to any analysis, all continuous split variables were centered and scaled. We specify the following likelihood for continuous and binary variables. Nominal variables were factor coded into $L-1$ binary variables. Let Y_{-j} denote the matrix of all split variables with the exception of Y_j.

Continuous Variables: $$Y_j|\beta_j,\sigma_j \sim Normal\left(X\theta_j + Y_{-j}\beta_j, \sigma_j^2\right) \quad (2)$$

Binary Variables: $$Y_j|\beta_j \sim Bernoulli\left(\Phi^{-1}\left(X\theta_j + Y_{-j}\beta_j\right)\right) \quad (3)$$

The probit regression model specified for binary variables makes use of the following latent variable formulation of the Bernoulli likelihood proposed in Albert and Chib (1993). Augmenting the data with the latent variable allows the user to apply Gibb's Sampling to the regression coefficients in a manner similar to those used for normal linear models.

$$\pi\left(Y_j|\beta_j, z_j\right) = \prod_i \left[1_{z_{ij\geq 0}} Y_{ij} + 1_{z_{ij<0}}\left(1 - Y_{ij}\right)\right] Normal\left(z_{ij}|X\theta_j + Y_{-j}\beta_j, 1\right) \quad (4)$$

4.3 Prior Distributions

For Bayesian estimation, prior distributions were specified for unknown parameters. The following describes these prior distributions.

Split Variable Regression Coefficients

The prior specification of the regression coefficients corresponding to the split variables is the following mixture of multivariate normal variables.

$$\beta_{jj'}|\Delta_{j'j} \sim \left(1 - \Delta_{jj'}\right) Normal\left(0, \tau_{jj'}^2\right) + \Delta_{j'j} Normal\left(0, c_{jj'}^2 \tau_{jj'}^2\right) for\ j' \neq j \quad (5)$$

When the split variable j' is selected to be grouped with split variable j, the prior distribution of $\beta_{jj'}$ is assigned to be a diffuse prior with $c_{jj'}^2$ specified to be large. On the other hand, if split

variable j' and split variable j are assigned to different groups, the prior distribution is concentrated around zero with variance $\tau_{jj'}^2$ specified to be small.

George and McCulloch (1993) suggested looking at the ratio of the variance of the least squares regression parameter ($Var_{LS}[\beta_{jj'}]$) to the value of $\tau_{jj'}^2$. A large value of this ratio would possibly group variables together that are not strong predictors of one another. A small value of this ratio would possibility do the opposite, assigning variables that are strong predictors of each other to opposite groups. The parameter $c_{jj'}^2$ should be specified such that it is large enough to support values not equal to 0, but not so large that unrealistic values are supported. The authors suggest that these parameters be treated as tuning constants for the sampling algorithm.

Core Variable Regression Coefficients

The prior specification of the regression coefficients corresponding to the core variables is the same as the split variables, with the exception that all core variables will be assumed to be assigned to the same group as all split variables. Under this assumption, all core variables are assigned a non-informative prior distribution.

$$\theta_{kj} \sim Normal\left(0, c_{kj}^2 \tau_{kj}^2\right) \tag{6}$$

Normal Variance Term

The normal variance term σ_j^2 was specified with a diffuse prior, as follows.

$$\sigma_j^2 \sim Inverse - Gamma(0,0) \tag{7}$$

Group Assignment Matrix

Each element of the matrix Δ is a binary number determining which variables are assigned to the same group. A priori, each element of Δ is independently specified to have a Bernoulli distribution with probability parameter equal to the corresponding element of α.

$$\Delta_{jj'} \sim Bernoulli(\alpha_{jj'}) \tag{8}$$

The prior distribution of $\alpha_{jj'}$, the probability that variable *j* is assigned to the same group as variable *j'*, was specified to have a non-informative Beta distribution.

$$\alpha_{jj'} \sim Beta(1,1) \tag{9}$$

4.4 Estimation

We used the Metropolis algorithm to generate posterior distributions of the unknown parameters in the above model specification. Tables 2.2 and 2.3 describe the sampling of each unknown parameter for continuous and binary variables respectively. Recall that nominal variables are factor coded into *L-1* binary variables, where *L* is the number of categories.

Table 2.2: Conditional Posterior Distributions for Continuous Variables

Parameter	Sampling Method	Sampling Formula
Group Assignment (Δ)	Metropolis	Proposal: $F(\Delta)$ PDF: $\prod_{j'}\prod_{j}1_{j'\neq j}\left[(1-\Delta_{jj'})Normal(\beta_{jj'}\|0,\tau_{jj'}^2)+\Delta_{jj'}Normal(\beta_{jj'}\|0,c_{jj'}^2\tau_{jj'}^2)\right]Bernoulli(\Delta_{jj'}\|\alpha_{jj'})$
Core Variable Regression Coeffiients (θ)	Gibbs Sampling	$\theta_j \sim Normal\left(\frac{1}{\sigma_j}\left(\frac{1}{\sigma_j}X^TX+D^TD\right)^{-1}X^TY_j,\left(\frac{1}{\sigma_j}X^TX+D^TD\right)^{-1}\right)\ where\ D=diagonal(c_{kj}^2\tau_{kj}^2)$
Split Variable Regression Coeffiients (β)	Gibbs Sampling	$\beta_j \sim Normal\left(\frac{1}{\sigma_j}\left(\frac{1}{\sigma_j}Y_{-j}^TY_{-j}+D^TD\right)^{-1}Y_{-j}^TY_j,\left(\frac{1}{\sigma_j}Y_{-j}^TY_{-j}+D^TD\right)^{-1}\right)$ $where\ D=diagonal\left((1-\Delta_{jj'})\tau_{jj'}^2+\Delta_{jj'}c_{jj'}^2\tau_{jj'}^2\right)$
Normal Likelihood Variance (φ)	Gibbs Sampling	$\varphi_j \sim Inverse-Gamma\left(\alpha_0=\frac{N}{2},\beta_0=\frac{1}{2}(Y_j-X\theta_j-Y_{-j}\beta_j)^T(Y_j-X\theta_j-Y_{-j}\beta_j)\right)$
Group Assignment Probability(α)	Gibbs Sampling	$\alpha_{jj'} \sim Beta(\alpha_0=\Delta_{jj'}+1,\beta_0=(1-\Delta_{jj'})+1)$

Table 2.3: Conditional Posterior Distributions for Binary Variables

Parameter	Sampling Method	Sampling Formula
Group Assignment (Δ)	Metropolis	Proposal: $F(\Delta)$ PDF: $\prod_{j'}\prod_{j}1_{j'\neq j}\left[(1-\Delta_{jj'})Normal(\beta_{jj'}\|0,\tau_{jj'}^2)+\Delta_{jj'}Normal(\beta_{jj'}\|0,c_{jj'}^2\tau_{jj'}^2)\right]Bernoulli(\Delta_{jj'}\|\alpha_{jj'})$
Latent Variable (z)	Gibbs Sampling	$z_{ij}\|y_{ij}=1 \propto 1_{z_{ij}\geq 0}Normal(z_{ij}\|X\theta_j+Y_{-j}\beta_j,1)$ $z_{ij}\|y_{ij}=0 \propto 1_{z_{ij}<0}Normal(z_{ij}\|X\theta_j+Y_{-j}\beta_j,1)$
Core Variable Regression Coeffiients (θ)	Gibbs Sampling	$\theta_j \propto Normal((X^TX+D^TD)^{-1}X^Tz_j,(X^TX+D^TD)^{-1})\ where\ D=diagonal(c_{kj}^2\tau_{kj}^2)$
Split Variable Regression Coeffiients (β)	Gibbs Sampling	$\beta_j \propto Normal\left((Y_{-j}^TY_{-j}+D^TD)^{-1}Y_{-j}^Tz_j,(Y_{-j}^TY_{-j}+D^TD)^{-1}\right)$ $where\ D=diagonal\left((1-\Delta_{jj'})\tau_{jj'}^2+\Delta_{jj'}c_{jj'}^2\tau_{jj'}^2\right)$
Group Assignment Probability(α)	Gibbs Sampling	$\alpha_{jj'} \sim Beta(\alpha_0=\Delta_{jj'}+1,\beta_0=(1-\Delta_{jj'})+1)$

The function *F(Δ)* represents the proposed value for the matrix Δ given the previous value. *F* changes the previous value of Δ in two ways to obtain the new proposed value. First, the function will add, subtract, or leave the number of groups unchanged. The probability of each operation was tuned for each application to obtain an acceptance rate of approximately 23%, as suggested in Gelman et al. (2003). If adding a group is randomly selected, one of the previous groups of variables is randomly split into two groups, with no restriction on the size of the two new groups. If subtracting a group is randomly selected, two of the previous groups of variables are randomly selected to be combined. If the number of groups is randomly selected to remain unchanged, no change is made to the previous group assignments.

The second operation that *F* performs on the previous group assignment matrix is to randomly shuffle split variables. During this operation, each split variable is randomly selected to either remain in its current group, or be randomly assigned to a different group. All split variables selected to be reassigned are shuffled into a new order and placed in the corresponding group. For example, suppose variables 1, 5, and 10 are selected to be reassigned from the current group assignment of 1 and 5 to group 1 and 10 to group 2. Assume the order is reshuffled to 5, 10, and 1. The new assignment would now be variables 5 and 10 to group 1 and variable 1 to group 2.

4.5 Cluster Analysis

The goal of our proposed matrix sampling design methodology is to assign variables to homogenous groups, or groups of variables that are predictive of one another. The posterior distribution of the probability matrix α was used to assign split variables to these groups. The sampled values of each element of α represent the posterior distribution of the probability that the j^{th} and $j`^{th}$ variables are assigned to the same group. The means of these posterior distributions (*mean(α|X,Y)*) were used for the assignment of split variables to groups.

We used agglomerative hierarchical clustering algorithm to create groups of predictive split variables. Massert and Kaufman (1983) gives a summary of agglomerative hierarchical clustering. Being a square matrix with the number of rows and columns equal to the number of split variables that we would like to cluster, the matrix *1- mean(α|X,Y)* can be used as the distance matrix in this algorithm. If split variable *j* and split variable *j`* are good predictors of one another and should be clustered together, the $jj`^{th}$ element of *1- mean(α|X,Y)* should be close to zero (*mean(α|X,Y)* close to one). On the other hand, if split variable *j* and split variable *j`* are

poor predictors of one another and should not be clustered together, the $jj`^{th}$ element of *1- mean(α|X,Y)* should be close to one (*mean(α|X,Y)* close to zero). Thus, the elements of *1- mean(α|X,Y)* are small for split variables that should be clustered together, and large for split variables that should be in separate groups.

The matrix *1- mean(α|X,Y)* meets the three required properties to be used as a distance matrix.

> *Positivity*: Since the values of α are probabilities between zero and one, the elements of the matrix *1- mean(α|X,Y)* are positive values.
>
> *Symmetry*: By definition, the matrices Δ and α are symmetric, thus *1- mean(α|X,Y)* is symmetric.
>
> *Triangle Inequality*: See Appendix A for proof that this property holds for the matrix *1- mean(α|X,Y)*.

The hierarchical clustering algorithm starts with all variables as their own cluster and iteratively combines clusters based on their proximity to one another until only one cluster remains. There are options when determining the proximity between clusters. We specified that Ward's method be used, where the proximity of two clusters is defined by the increase in squared error that results from merging them. See Ward (1963) for more information about Ward's method and Tan et al. (2006) for more information related to hierarchical clustering algorithms in general.

The results of hierarchical clustering are commonly summarized in a graphical display called a dendrogram. In the case of agglomerative hierarchical clustering, the dendrogram summarizes

the order in which clusters are merged together. Tan et al. (2006) discusses different metrics used to determine the correct number of clusters (or groups) to specify. For this research, we selected the number of clusters by visual inspection of the dendrogram. Hierarchical clustering was implemented using the *hclust* function in the STATS package of the R programming language (RDC Team 2009).

4.6 Matrix Sampling Design Implementation

The SSVS matrix sampling method assigns variables to homogenous groups. As previously stated, a good matrix sampling design ensures that variables associated with questions administered to the respondent are predictive of the variables associated with the questions not administered. To achieve this objective, we create matrix sampling designs from the SSVS matrix sampling variable assignments by randomly sampling variables within each group and administering the associated questions.

For example, assume our questionnaire contained four questions and we would like to administer only two questions to each respondent. Also assume that the SSVS matrix sampling method assigned two groups: variables 1 and 2 in group 1 and variables 3 and 4 in group 2. Based on these results, we create matrix sampling designs by randomly selecting one variable from variables 1 and 2 and one variable from variables 3 and 4. This sampling method ensures that whenever the questions associated with question 2 are omitted from the questionnaire, the questions associated with question 1 are administered. Similarly, the values from either variable 3 or 4 are collected from every respondent.

5. Application to the Nampula Household Survey Data

The research presented in this manuscript seeks to develop an improved procedure for creating matrix sampling designs. While many of the methods mentioned in Section 3 are only applicable to continuous variables, we require methodology that handles both continuous and categorical variables. For this reason, we compare our proposed methodology to the method proposed in Thomas et al. (2006), a recently developed method that handles both types of variables.

We applied the methodology proposed in Thomas et al. (2006) and our proposed methodology to data from a household survey administered in the Nampula province of Mozambique. Section 5.1 describes the survey data file and the variables selected from it to be included in the research. Section 5.2 compares the proposed methodology to the method proposed in Thomas et al. (2006). Section 5.3 compares simulations of survey data collected using our proposed methodology to simulations of complete survey data with a smaller sample size.

5.1 Data File and Variables

The research file derived from the Nampula household survey contained 22 variables (described below in Table 2.4) and 765 observations. The 765 observations were selected from the approximately 1,600 household respondents. These selected observations were records with observed values for all 22 of the selected variables. The 22 selected variables were those that were prominent in the presentations of results in the months following the Nampula household survey. Of the 22 selected variables, we classify two as core variables and the other 20 as split variables. Both core variables were categorical variables and are represented as such in the regression models. The 20 split variables contain 16 continuous variables (assumed to have a

normal distribution), 3 binary variables (assumed to have a Bernoulli distribution), and 1 nominal variable. The nominal variable was factor-coded into multiple binary variables. We centered and scaled all continuous variables prior to analysis.

Table 2.4: Nampula Baseline Survey Research File Variables

Split Variable Number	Variable	Description
Core Variables		
	District	One of six districts sampled from in the province of Nampula. Either 6 or 12 communities were sampled from each district.
	Respondent Gender	Gender of the household respondent.
Continuous Split Variables		
1	Wet Liters Per Capita Per Day (LPCD)	Liters of water used by the household per person per day during the wet season.
2	Dry Liters Per Capita Per Day (LPCD)	Liters of water used by the household per person per day during the dry season.
3	Percent Wet Water LPCD Protected Sources	Percent of water during the wet season obtained from protected sources.
4	Percent Dry Water LPCD Protected Sources	Percent of water during the dry season obtained from protected sources.
5	Wet Typical Volume Carried by Persons	Typical liters of water carried by a household member during a trip to a water source during the wet season.
6	Dry Typical Volume Carried by Persons	Typical liters of water carried by a household member during a trip to a water source during the dry season.
7	Hungry Months	Number of months during a given year the household is not satisfied with the amount of food obtained.
8	Average Number of Times Washing Hands	Average number of times the respondent washes hands during a given week.
9	Expenditures Per Person	Household expenditures per person per month.
10	Remittances Per Person	Amount of money given to the household per person per month from either relatives or non-relatives.
11	Time Cost of Illness	Total time spent traveling to and from the medical facility during the most recent trip.
12	Monetary Cost of Illness	Total cost of visiting the medical facility during the most recent trip.
13	Average Number of Times Washing Hands with Soap	Average number of times the respondent washes hands with soap during a given week.
14	Walk Time to Primary Source	Total time to walk to and from the primary source of water for the household.
15	Wet Season Wait Time at Primary Source	Time waiting to obtain water at the primary source of the household during the wet season.
16	Dry Season Wait Time at Primary Source	Time waiting to obtain water at the primary source of the household during the dry season.
Binary Split Variables		
17	Community Participation	Indicator of attending at least 1 community meeting.
18	Latrine	Indicator of the presence of a latrine at the household.
19	Enough Water for Daily Activities	Indicator of the whether the household is satisfied with the amount of water fetched for daily activities.
Nominal Split Variables		
20	Water Interference with School (5 Categories)	Variable indicating how much water fetching interferes with children's school attendance.

5.2. Matrix Sampling Design Simulation Results

In the Section 1, we discussed two ways in practice to create matrix sampling designs: use data from a similar survey or administer the entire questionnaire during an initial implementation of the survey. For this simulation, we took the second approach and selected the first two districts to be the initial implementation. Of the 765 total observations on the research file, 258 were collected in these two districts. The Thomas et al. (2006) and our proposed matrix sampling methods were applied to complete data from these two districts to create matrix sampling designs. For the Thomas et al. (2006) method, variables were assigned to four blocks, each containing five variables. The group assignments for the SSVS matrix sampling method were determined by visual analysis of the cluster dendrogram, provided in Figure 2.1. This visual analysis yielded five groups: four with two variables each and one with the remaining twelve variables.

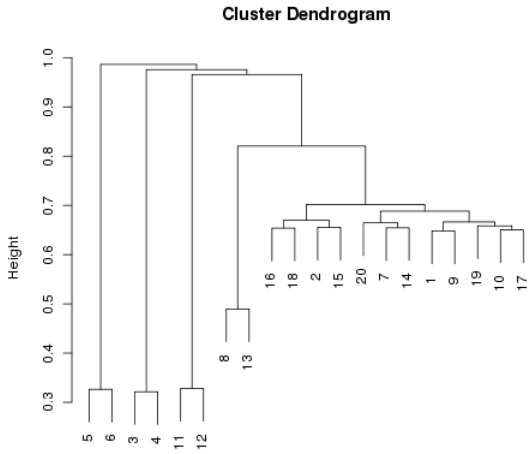


Figure 2.1: Matrix Sampling Clustering of Variables

Table 2.5 provides the assignment of variables to groups or blocks for the two methods.

Table 2.5: Variable Group Assignment

Matrix Sampling																					
Group 1												Group 2		Group 3		Group 4		Group 5	
1	2	7	9	10	14	15	16	17	18	19	20	3	4	5	6	8	13	11	12
Wet LPCD						Wet Season Wait Time						Percent Wet LPCD Protected Sources		Wet Volume Carried		Number Times Washing Hands		Time Cost of Illness	
Dry LPCD						Dry Season Wait Time													
Hungry Months						Community Participation						Percent Dry LPCD Protected Sources		Dry Volume Carried		Number Times Washing Hands with Soap		Monetary Cost of Illness	
Expenditure Per Person						Latrine													
Remittances Per Person						Enough Water for Daily Activities													
Walk Time to Primary Source						Water Interference with School													

Thomas et al. (2006)																					
Block 1					Block 2					Block 3					Block 4						
1	3	9	11	15	2	7	12	16	19	4	6	8	10	17	5	13	14	18	20		
Wet LPCD					Dry LPCD					Percent Dry LPCD Protected Sources					Wet Volume Carried						
Percent Wet LPCD Protected Sources					Hungry Months					Dry Volume Carried					Number Times Washing Hands with Soap						
Remittances Per Person					Monetary Cost of Illness					Number Times Washing Hands					Walk Time to Primary Source						
Time Cost of Illness					Dry Season Wait Time					Remittances Per Person					Latrine						
Wet Season Wait Time					Enough Water for Daily Activities					Community Participation					Water Interference with School						

Based on the two matrix sampling designs, we simulated missing data in the remaining four districts. The missing percentage for this simulation was set to be 50%, or ten of the twenty variables. For the Thomas et al. (2006) method, two of the four blocks were randomly selected for each questionnaire, and all variables contained in the two selected blocks were set to missing. For the SSVS matrix sampling method, half of the variables within each group were selected at random to be set to missing for each questionnaire. For comparison purposes, we also provide a third method, a random assignment of variables to be missing without group assignments.

We used multiple imputation with chained regression equations, described in Section 2, to impute values for the missing data. All 765 records on the Nampula household survey research file were input into the multiple imputation procedure. The 258 complete case records were included with the 507 records with missing values. The chained regression equations method specifies a univariate regression model for each variable with missing data (all split variables in this case) and iteratively imputes the missing values for a set number of iterations. Each of these

univariate imputation models includes all other variables as covariates, as suggested in Rubin (1996). The default methods identified by the *mice* function for each type of variable were used for the imputation. For continuous variables, predictive mean matching (Little 1988) was the default method. Logistic regression and multinomial logistic regression models were specified for categorical binary variables and nominal variables respectively. Multiple imputation requires that several complete datasets be created in the same manner. Rubin (1996) states, through theory and experience, that five imputations are adequate when the percentage of missing data is not large.

For each matrix sampling design, we repeated the above simulated missing values and imputation procedure for 1,000 replications of random missingness. The imputed values vary for each realization of random missingness created from the matrix sampling design. We compared the multiply imputed datasets resulting from the matrix sampling designs to the dataset of true values (the 765 complete records) in two ways.

The first way compares the variable means resulting from the multiply imputed datasets to the true variable means from the complete dataset. The variable means from the five multiply imputed datasets were averaged to produce the estimated variable means for the given replication of missingness. We compared the estimated variable means to the true variable means using the Root Mean Squared Error (RMSE) statistic, or the square root of the squared error loss averaged over the 1,000 replications. Table 2.6 provides the RMSE of the variable means estimated by each of the three methods. We also report the squared bias and variance of the variables mean estimates as percentages of the RMSE.

Table 2.6: RMSE of Variable Mean Estimates and Its Components

Variable Number	SSVS Matrix Sampling			Thomas et al. (2006)			Random Sampling		
	RMSE	MSE Components		RMSE	MSE Components		RMSE	MSE Components	
		% Bias^2	% Variance		% Bias^2	% Variance		% Bias^2	% Variance
	RMSE Values x10^{-2}								
1	2.68	60%	40%	2.85	62%	38%	2.51	50%	38%
2	1.83	6%	94%	1.82	9%	91%	1.86	3%	91%
3	1.87	0%	100%	1.87	12%	88%	1.95	7%	88%
4	2.11	9%	91%	1.76	0%	100%	1.88	1%	100%
5	0.97	59%	41%	1.77	75%	25%	2.42	79%	25%
6	1.00	55%	45%	1.13	45%	55%	1.86	65%	55%
7	3.71	48%	52%	3.70	46%	54%	3.80	50%	54%
8	3.44	25%	75%	3.67	41%	59%	3.52	33%	59%
9	2.81	1%	99%	2.82	4%	96%	2.81	1%	96%
10	5.69	78%	22%	5.87	79%	21%	5.64	77%	21%
11	2.31	82%	18%	2.61	82%	18%	2.84	82%	18%
12	2.12	19%	81%	2.49	38%	62%	2.85	46%	62%
13	2.84	7%	93%	3.05	23%	77%	2.59	0%	77%
14	2.73	7%	93%	2.70	5%	95%	2.51	5%	95%
15	2.45	1%	99%	2.42	0%	100%	2.46	2%	100%
16	2.63	11%	89%	2.60	12%	88%	2.74	14%	88%
Continuous Variable Average	2.57	27%	73%	2.68	31%	69%	2.78	31%	69%
17	2.03	65%	35%	2.06	65%	35%	1.97	62%	35%
18	2.55	81%	19%	2.77	84%	16%	2.60	83%	16%
19	1.19	6%	94%	1.15	0%	100%	1.23	3%	100%
20-1	0.96	65%	35%	1.05	64%	36%	1.03	66%	36%
20-2	0.94	6%	94%	0.91	1%	99%	0.94	4%	99%
20-3	1.24	0%	100%	1.34	9%	91%	1.25	0%	91%
20-4	1.32	25%	75%	1.44	39%	61%	1.26	28%	61%
20-5	0.98	1%	99%	0.95	7%	93%	0.95	0%	93%
Categorical Variable Average	1.40	31%	69%	1.46	34%	66%	1.40	31%	66%

Table 2.6 shows that the SSVS matrix sampling method produces the best overall results in terms of RMSE for continuous variables. The SSVS method produces equivalent results to random sampling for categorical variables, both of which perform slightly better than the Thomas et al. (2006) method. For all three methods, the variance component generally constitutes a larger percentage of the mean squared error than the squared bias component. Figure 2.2 provides boxplots that compare the squared bias and variance components of the variable mean estimates for the three methods over continuous and categorical variables.

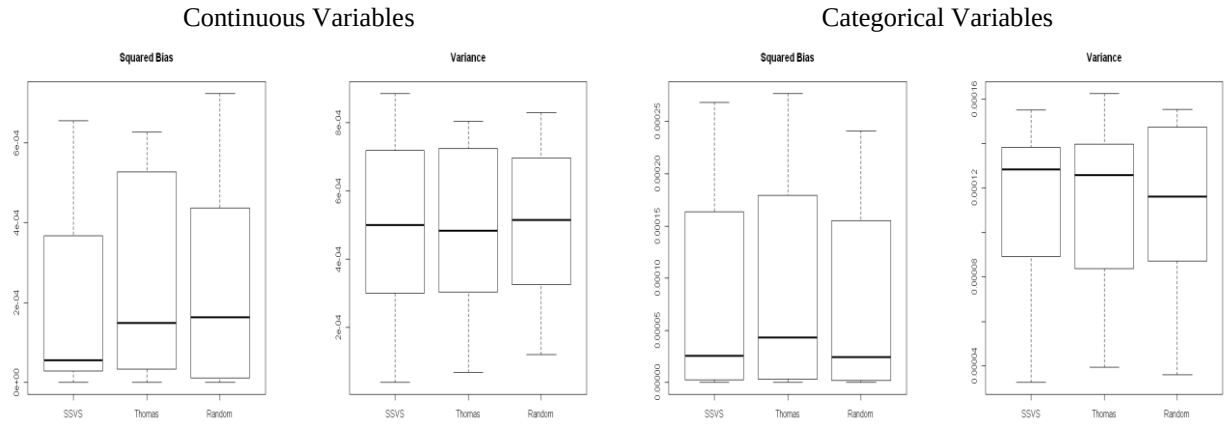


Figure 2.2: Comparison of the Squared Bias and Variance of the Three Methods

For continuous variables, Figure 2.2 shows that while the variance of the variable mean estimates are generally equivalent for the three methods, the SSVS matrix sampling method produces variable mean estimates with generally lower squared bias values in comparison to the Thomas et al. (2006) and random sampling methods, resulting in a lower RMSE. For categorical variables, Figure 2.2 shows that three methods produce variable mean estimates similar variances, but the SSVS and random sampling methods produce variable mean estimates with slightly lower squared bias values.

For the second evaluation, we compare the unit level means over the multiply imputed datasets to the true values of the complete data set. Each variable value for each respondent was averaged over the five multiply imputed datasets to produce an imputed data set for the given replication of missingness. In practice, this imputed data set would be the final data set in which researchers would conduct data analysis. This evaluation will quantify the difference between the final data produced by each matrix sampling design method and the true values of the complete data set.

We used different loss functions for the continuous variables and categorical variables. For continuous variables, the RMSE, or the squared root of the squared error loss averaged over respondents, was used. For categorical variables, we used the misclassification rate. Only variable values that were simulated to be missing were used in the calculation of these loss functions. Table 2.7 provides the average value of these loss functions over the 1,000 replications of missingness for each of the three matrix sampling design methods.

Table 2.7: Loss Function of Individual Variable Values

Variable Number	SSVS Matrix Sampling	Thomas et al. (2006)	Random Sampling
Continuous Variables (Mean Squared Error)			
1	0.81	0.80	0.86
2	0.74	0.75	0.89
3	0.57	0.79	0.99
4	0.60	0.82	1.01
5	0.05	0.29	0.48
6	0.05	0.28	0.48
7	1.80	1.79	1.81
8	1.67	1.77	1.81
9	1.78	1.74	1.80
10	2.28	2.29	2.28
11	0.57	0.74	0.85
12	1.25	1.43	1.61
13	1.56	1.74	1.70
14	1.74	1.77	1.76
15	1.74	1.70	1.75
16	1.74	1.71	1.76
MSE Average	1.19	1.28	1.37
Categorical Variables (Misclassification Rate)			
17	0.42	0.42	0.42
18	0.35	0.36	0.35
19	0.35	0.35	0.37
20	0.67	0.68	0.68
MR Average	0.45	0.45	0.45

Table 2.7 shows that, in terms of the loss functions defined, the SSVS matrix sampling method produces the best estimates of variables values at the respondent level for continuous variables, all three methods produce similar results for categorical variables. The improvement in the loss functions can be attributed to the variables that the SSVS matrix sampling method identified as

having another highly predictive variable in the dataset. These identified variables were variables 3, 4, 5, 6, 8, 11, 12, and 13.

5.3 Comparison of Matrix Sampling to Full Questionnaire

The following section seeks to determine the usefulness of our proposed matrix sampling design when applied to a household survey. One of the advantages of using the SSVS matrix sampling approach is the increased number of questionnaires that can be administered due to the decreased time to administer a questionnaire. If the SSVS matrix sampling approach was not used, all questions would be administered to a given respondent, but we would administer the questionnaire to a smaller number of households.

For the analysis provided in this section, we compare the results of the SSVS matrix sampling approach with all respondents observed to the results of a full questionnaire with a smaller sample size. Similar to the analysis done in Section 5.2, the first two districts (258 observations) were selected to be fully observed with no missing data. We applied the SSVS matrix sampling method to the data from these two districts to create a matrix sampling design.

We simulated the data in the remaining four districts (507 observations) to have missing data. We applied the SSVS matrix sampling design with half of the questions asked to all of the households in the four districts. These results were compared to a stratified random sampling of half of the households in each of the four districts, with the entire questionnaire administered to each sampled household. The resulting variable mean estimates were calculated and compared to true means using the RMSE statistic. Each method was applied to the data for 1,000

replications of missingness. Table 2.8 provides the RMSE of the mean estimates produced by the SSVS matrix sampling and half sample methods compared to the true values derived from the complete dataset. The table also provides the percentages of the mean squared error that its components, the squared bias and variance, constitute.

Table 2.8: RMSE of Variable Mean Estimates and Its Components

Variable Number	SSVS Matrix Sampling			Half Sample		
	RMSE	MSE Components		RMSE	MSE Components	
		% Bias^2	% Variance		% Bias^2	% Variance
	RMSE Values x10^{-2}					
1	2.68	60%	40%	3.22	0%	100%
2	1.83	6%	94%	2.98	0%	100%
3	1.87	0%	100%	2.91	0%	100%
4	2.11	9%	91%	2.93	0%	100%
5	0.97	59%	41%	3.01	0%	100%
6	1.00	55%	45%	3.03	0%	100%
7	3.71	48%	52%	2.95	0%	100%
8	3.44	25%	75%	2.97	0%	100%
9	2.81	1%	99%	2.55	0%	100%
10	5.69	78%	22%	2.69	0%	100%
11	2.31	82%	18%	3.36	0%	100%
12	2.12	19%	81%	3.64	0%	100%
13	2.84	7%	93%	3.02	0%	100%
14	2.73	7%	93%	2.84	0%	100%
15	2.45	1%	99%	2.83	0%	100%
16	2.63	11%	89%	2.71	0%	100%
Continuous Variable Average	2.57	27%	73%	2.98	0%	100%
17	2.03	65%	35%	1.40	0%	100%
18	2.55	81%	19%	1.24	0%	100%
19	1.19	6%	94%	1.37	0%	100%
20-1	0.96	65%	35%	0.64	0%	100%
20-2	0.94	6%	94%	0.90	0%	100%
20-3	1.24	0%	100%	1.40	0%	100%
20-4	1.32	25%	75%	1.22	0%	100%
20-5	0.98	1%	99%	1.13	0%	100%
Categorical Variable Average	1.40	31%	69%	1.16	0%	100%

As expected with stratified random sampling, the bias of the half sample method is nearly equal to zero. For continuous variables, the SSVS matrix sampling approach produces variable mean estimates with RMSE values that are generally lower than those from the half sample method. In particular, these lower RMSE values coincide with variables that the SSVS matrix sampling method identified as having another highly predictive variable in the data set (Variables 3, 4, 5, 6, 8, 11, 12, and 13). For categorical variables, the half sample method generally produced better

results than the SSVS matrix sampling approach. Figure 2.3 provides boxplots that compare the squared bias and variance components of the variable mean estimates for both methods over continuous and categorical variables.

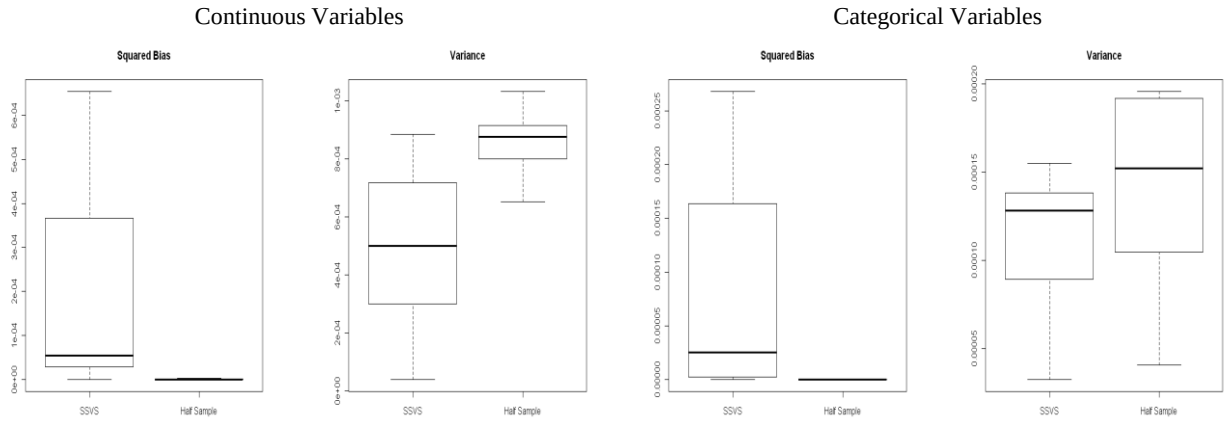


Figure 2.3: Comparison of the Squared Bias and Variance of the Two Methods

The half sample method produces mean estimates that are relatively unbiased. The larger sample size resulting from the SSVS matrix sampling method allows for mean estimates with smaller variability for both continuous and categorical variables. The difference in the variance is substantial for continuous variables, resulting in generally lower RMSE values associated with the SSVS method. The difference in variance is not as large for categorical variables, resulting in generally lower RMSE values associated with the half sample method.

To show the practical implications of these findings, look at variables eleven and twelve, time cost and monetary costs of illness, in terms of their unstandardized values (recall that all previous results are provided in terms of the standardized variables). The true mean of these variables are 1.34 hours and 5.53 Meticais, respectively. Figure 2.4 provides boxplots of the 1,000 replicate

variable means produced by the SSVS matrix sampling and half sample methods. Note that the horizontal lines plot the true variable mean values.

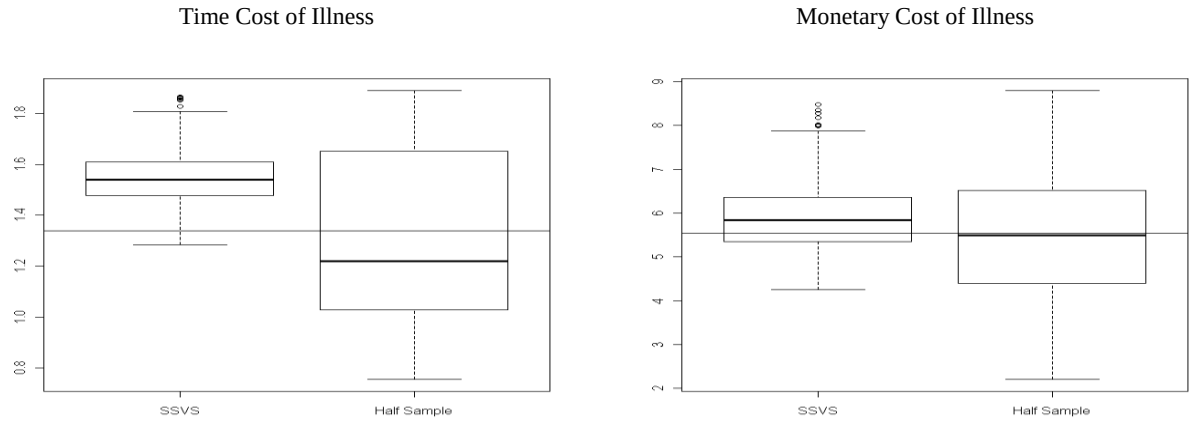


Figure 2.4: Unstandardized Mean Estimates of the Time Cost and Monetary Cost of Illness

Figure 2.4 shows that while the half sample method produces relatively unbiased variable mean estimates, they also have high variability. On the other hand, the SSVS method produced biased variable mean estimates, but they are substantially less variable. For example, the true mean value of the time cost of money variable is 1.34. Assume that we feel that a twenty percent margin of error is acceptable, or a mean value between 1.07 and 1.61. The SSVS method produces mean estimates in this acceptable range 75% of the time in the simulations, while the half sample method produced mean estimates in the acceptable range 39% of the time. Similarly, for the monetary cost of illness, the SSVS method produces mean estimates in the acceptable range in 83% of the simulations, while the half sample method produces means in the acceptable range in 52% of the simulations.

The analysis provided in this section has been conducted under the assumption that the time to conduct a full questionnaire is equal to the time it takes to conduct two questionnaires using a

matrix sampling design. This assumption may not be valid for some questionnaires, depending on the logistics required to implement them. For example, this assumption would not be valid if it takes 90 minutes to administer the questionnaire and 10 minutes to transition to the next respondent. Under this scenario, it would take 100 minutes to administer the full questionnaire to one respondent, but 110 minutes to administer the matrix sampling designed questionnaire to two respondents. Relating this additional time to the analysis in this section, we would be able to administer the full questionnaire to 55% of the households in the time it would take to administer the matrix sampling designed questionnaire to all households.

To determine the amount of additional time that may be allowed such that the SSVS matrix sampling design is equivalent to administering the full questionnaire to a smaller sample of households, we sampled varying percentages of households within each of the four districts (50% to 75% in intervals of 5%) with the entire questionnaire administered to each sampled household. We calculated the RMSE of the variable means for each of these percentages, and produced the plots shown in Figure 2.5. Each point represents the average RMSE over continuous variables or categorical variables.

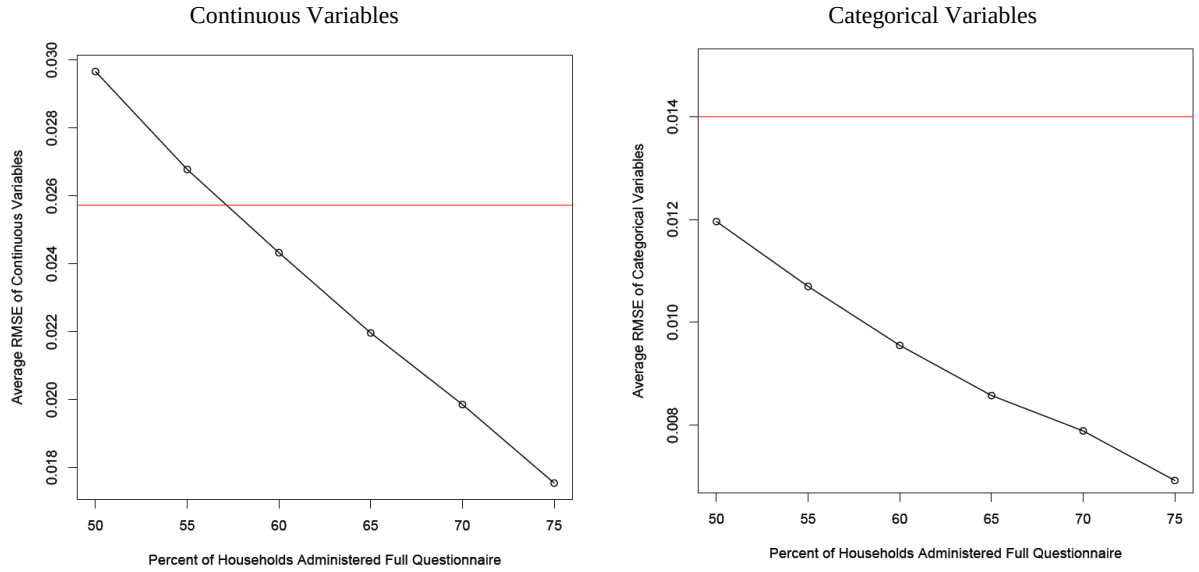


Figure 2.5: Administration of the Full Questionnaire to Smaller Samples of Households

The red line in Figure 2.5 represents the average RMSE of variable means produced by the matrix sampling designed questionnaire. For continuous variables, the matrix sampling designed questionnaire performs similar to the administration of the full questionnaire to between 55% and 60% of households. For categorical variables, the full questionnaire administered to 50% of households performed better than matrix sampling designed questionnaire. The 55% to 60% of households is equivalent to 10 to 22.5 minutes of transition time from one respondent to the next if it takes 90 minutes to administer the full questionnaire.

6. Conclusions

Matrix sampling designs are a potential solution to the respondent and enumerator burdens associated with long survey questionnaires. We can improve the quality of the data collected by administering a subset of the questions to each respondent in a way such that the administered

questions are predictive of the omitted questions. We developed the methodology proposed in this manuscript to offer researchers an improved method of creating matrix sampling designs. The proposed methodology can be applied to all types of variables and provides a specific procedure for assigning variables to homogenous groups, or groups of predictive variables. Unlike other methods previously developed, the number of groups does not need to be pre-specified and can be determined based on data from a similar survey or an initial implementation of the survey.

We applied the SSVS matrix sampling approach to complete case data from a household survey administered in the Nampula province of Mozambique with simulated missing values. The methodology was applied in a manner that could be used in practice. For continuous variables, the proposed methodology outperforms the method developed by Thomas et al. (2006) and a random sampling of variables in estimating the true variables means and the unit level variable values. For variable mean estimation, the variances of the mean estimates of the three methods are similar, but the variable means resulting from the SSVS matrix sampling method are less biased. For categorical variables, the variable mean estimates produced by three methods performed similarly. Note that SSVS method did not find any of these categorical variables to have another variable that was highly predictive of them, which may be a reason for the lack of differentiation of the SSVS method with the other two methods. Raghunathan and Grizzle (1995) had similar results, stating the matrix sampling method only works well when strong correlations are present among split variables.

Section 5.3 shows the advantages of using the SSVS matrix sampling approach in comparison to applying the full questionnaire to a smaller sample of households. For the continuous variables, the variable mean estimates from the SSVS matrix sampling approach were better estimates of the true mean values than those from the full questionnaire with a half sample in terms of RMSE. The difference in estimates was particularly apparent with those variables that the SSVS method identified as having strong predictors in the data set. For the categorical variables, the variable mean estimates from the half sample method produced better results. As noted above, this may be attributed to the fact that no other variables in the data set were highly predictive of these variables. For many of the continuous variables, the SSVS matrix sampling approach produced variable mean estimates that were slightly biased, but had substantially smaller variance estimates than the half sample method.

Section 5.3 also compares the use of the SSVS matrix sampling approach to administering the full questionnaire to varying percentages of households. Depending on the logistics of the survey, there will be time costs associated with the transition from one respondent to the next. For continuous variables, we found that the RMSE produced by the administering the SSVS matrix sampling approach to all households is equivalent to administering the entire questionnaire to between 55% and 60% of households. Equivalently, the SSVS matrix sampling approach produces better variable mean estimates if there is less than 10 to 22.5 minutes of transition time from one respondent to the next.

All improvements in data quality shown in this paper come in addition to the reduction in respondent burden, non-response, and pre-mature termination that comes with a shorter questionnaire, as noted in Berdie (1989) and Adams and Gale (1982).

References

Adams, L. L. M. and Gale, D. (1982). Solving the Quandary Between Questionnaire Length and Response Rate in Educational Research, *Research in Higher Education*, *17*(3), 231-240.

Albert, J. H. and Chib, S. (1993). Bayesian Analysis of Binary and Polychotomous Response Data, *Journal of the American statistical Association*, *88*(422), 669-679.

Berdie, D. R. (1989). Reassessing the Value of High Response Rates to Mail Surveys. *Marketing Research*, *1*(3), 52-64.

Buuren, S., and Groothuis-Oudshoorn, K. (2011). MICE: Multivariate Imputation by Chained Equations in R, *Journal of statistical software*, *45*(3).

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian data analysis*. CRC press.

George, E. I. and McCulloch, R. E. (1993). Variable Selection Via Gibbs Sampling, *Journal of the American Statistical Association*, *88*(423), 881-889.

Gonzalez, J. M., and Eltinge, J. L. (2007). Multiple Matrix Sampling: A Review, In *Proceedings of the Section on Survey Research Methods, American Statistical Association*,3069-3075.

Hooke, R. (1956). Some Applications of Bipolykays to the Estimation of Variance Components and Their Moments. *The Annals of Mathematical Statistics*, *27*(1), 80-98.

Little, R. J. (1988). Missing-Data Adjustments in Large Surveys, *Journal of Business and Economic Statistics*, *6*(3), 287-296.

Lord, F. M. (1962). Estimating Norms by Item-Sampling. *Educational and Psychological Measurement*.

Massart, D. L. and Kaufman, L. (1983). *The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis*, New York: John Wiley and Sons.

Navarro, A., and Griffin, R. A. (1993). Matrix Sampling Designs for the Year 2000 Census. In *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 480-485.

R Development Core Team (2009). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rao, J.N.K. (2003). *Small Area Estimation, Hoboken*, NJ*:* John Wiley and Sons, Inc.

Raghunathan, T. E., and Grizzle, J. E. (1995). A Split Questionnaire Survey Design, *Journal of the American Statistical Association*, *90*(429), 54-63.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., and Solenberger, P. (2001). A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models, *Survey methodology*, *27*(1), 85-96.

Rosenbaum, P. R. and Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects, *Biometrika*, 70, 41–55.

Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys, *New York, USA: John Willey and Sons*.

Rubin, D. B. (1996). Multiple Imputation After 18+ Years, *Journal of the American Statistical Association*, *91*(434), 473-489.

Rubin, D. B. (2009). *Multiple Imputation for Nonresponse in Surveys* (Vol. 307). Wiley.com.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*, New York: Chapman and Hall.

Schafer, J. L. and Schenker, N. (2000). Inference with Imputed Conditional Means. *Journal of the American Statistical Association*, *95*(449), 144-154.

Shoemaker, D. M. (1973). A Note on Allocating Items to Subtests in Multiple Matrix Sampling and Approximating Standard Errors of Estimate with the Jackknife. *Journal of Educational Measurement*, *10*(3), 211-219.

Tan, P., Steinbach, M., and Kumar, V. (2006). *Introduction to Data Mining*, New York: Addison Wesley.

Thomas, N., Raghunathan, T. E., Schenker, N., Katzoff, M. J., and Johnson, C. L. (2006). An Evaluation of Matrix Sampling Methods Using Data from the National Health and Nutrition Examination Survey, *Survey Methodology*, *32*(2), 217.

Ward Jr., Joe H. Hierarchical Grouping to Optimize an Objective Function, *Journal of the American statistical association*, 58.301 (1963): 236-244.

CHAPTER 3: THE IMPORTANCE OF CLEANING DATA DURING FIELDWORK: EVIDENCE FROM MOZAMBIQUE

Chapter 3 Abstract

In many small-scale surveys with limited resources, data editing is usually conducted by a statistician after data collection has concluded. Including the statistician and the data editing process in the data collection phase of the survey has many benefits. This paper describes a procedure for survey implementation of small-scale surveys in which the statistician identifies potential data entry errors and edits the data as they are collected. We implemented this procedure during a household survey conducted in Maputo, the capital of Mozambique, and detailed data on the editing process were recorded. This article analyzes these data to gain insight into the effects on the collected data. The results of the analysis indicate that the edited data may be of higher quality than data without edits. We also identify areas of improvement in the procedure for future household surveys.

1. Introduction

Researchers divide errors in survey data into two broad categories: sampling and non-sampling error (Bethlehem 1997). In an ideal world, survey data would be collected from the entire population of interest, and no sampling error would be present. Cost and time concerns make this approach unreasonable, so researchers sample a subset of the target population. Sampling introduces error into the data. Sampling error is the difference between statistical quantities calculated from the sample and what would have been calculated from the entire population. All errors not attributed to sampling error are known as non-sampling errors. Researchers place these errors in two categories: non-observation and observation (Bethlehem 1997). Non-

observation errors occur due to non-response and undercoverage. Observation errors include measurement errors and processing errors. The focus of this paper is to develop a procedure for survey implementation that identifies and edits data entry errors during data collection under the constraints of a small-scale survey.

Granquist (1984) defines three goals for the editing of data. The first goal is to gather information about significant differences in data for analytical purposes. The second is to provide feedback that improves the way in which data are collected and processed by reviewers. The final goal is to reduce the level of error in the data while maintaining consistency, integrity, and coherence. Hughes et al. (1990) identify two types of editing: micro-editing and macro-editing. Micro-editing aims at ensuring validity and consistency of individual data records. Macro-editing analyzes the data in aggregate.

In recent years, faculty and graduates in the Laboratory for Interdisciplinary Statistical Analysis (LISA)[1] and the Urban Affairs and Planning (UAP) program at Virginia Tech have collaborated with colleagues at Stanford University on survey projects in Senegal, Kenya, and Mozambique. The initial role of the statisticians from LISA was to assist with the editing and analysis of the survey data following data collection. Hartley (1980) stresses the importance of cooperation between the statistician and the research partners to gain enough insight into the data collection process to effectively perform these tasks. The statisticians from LISA found the process of working with data following the completion of fieldwork difficult and time consuming. Particularly challenging was the back and forth required with the research partners relating to background information on the contents of the survey, details related to the sample frame, and

[1] http://www.lisa.stat.vt.edu/

the research questions and goals for which the statistical analysis was intended to support. Deming (2000) discussed similar challenges and determined that cleaning data after it has been collected is too late, ineffective, and costly. Additionally, Deming (2000) noted that by waiting to identify and correct errors after data collection, it is not possible to locate the source of the errors, which means these errors will be repeated in future surveys. Bethlehem (1987) reinforces this point by observing that the transfer of data from one person/department to another is a source of error, misunderstanding, and delay.

Advances in portable computing technology have facilitated the integration of data collection and data editing. Computer-Assisted Personal Interviewing (CAPI) is the term given to the use of computing technology for data collection during personal interviews. At first, surveyors used laptops to assist data collection, but in recent years, the use of personal digital assistants (PDAs), tablets, and smart phones has grown significantly in the health and water sectors (Onono 2011). Gravlee et al. (2006) discusses experiences in using PDAs in observations studies, including observed advantages and disadvantages over the pen-and-paper method. The authors determine that the positives outweigh the negatives, and suggest ways to maximize the benefits. De Waal et al. (2011) discusses one of the main advantages of CAPI, the capability of researchers to start the data editing process during data collection by informing the surveyor of a possible error at the time the data value was recorded. The surveyor may confirm the data point was correctly entered, or correct an erroneously collected data point.

De Waal et al. (2011) defines data cleaning conducted by subject matter experts as interactive editing. The advantages to interactive editing include the ability to re-contact the respondents

and use of expert knowledge to edit the data. The authors state "Interactive editing is nowadays a standard way to edit data. … Generally, the quality of data editing in a computer assisted manner is considered high" (De Waal et al. 2011, p. 15-16). De Waal et al. (2011) also lists two potential problems with interactive editing. First, the use of computers permits the identification of large numbers of suspicious data values which may not require editing. De Waal et al. (2011) call this unnecessary editing of data "over editing". The second potential problem is what De Waal et al. (2011) call "creative editing", in which editors use subjective data editing procedures. The issue with "creative editing" is consistency, where a data editor may edit the same erroneous value in different ways.

Most literature found on the subject of data editing discuss procedures for surveys implemented on large-scale surveys. De Waal et al. (2011) outline a step-by-step process of a data editing; although, the procedure focuses on review methods by the statistician and does not discuss the interaction between the statisticians and surveyors. The authors also discuss balancing the use of manual editing and automatic editing by computers. The papers collected in Lyberg et al. (1997) refer to examples of data editing, but concentrate on procedures that combine macro-editing and automated editing by computers. Lyberg et al. (1997) develop the procedures for large-scale surveys, which have larger samples than the small-scale surveys in Senegal, Kenya, and Mozambique. The smaller samples allow for more extensive editing of the data over a shorter period of time.

Biemer and Caspar (1994) investigate a method of survey quality control called continuous quality improvement (CQI). The authors describe a general strategy for CQI that has many

similarities to the procedures presented in this paper. In particular, the authors similarly discuss the need for cooperation between team members such as the operators, data quality inspectors, and operation supervisor to ensure data quality and to make corrections if necessary. The authors focus on larger-scale surveys than those we discuss in this paper and only provide a general strategy, rather than outline a specific procedure.

The CAPI instrument used in the survey described in this paper allowed for the almost instantaneous access to collected data, but was not capable of flagging possible errors during interviews. The procedures proposed in this paper seek to achieve some of the benefits of interactive editing in spite of this limitation. The proposed procedures require statisticians to be involved with the cleaning and editing of data during the data collection process by reviewing the collected data and flagging any suspicious values. The small-scale of the survey mitigates concerns of a time-consuming data editing process due to the manageable number of surveys requiring review on a daily basis. In the proposed procedures, the statistician only edits potential errors after consulting the surveyor that collected the data point. This restriction minimizes the risk of "over editing". In many small-scale surveys, the surveyor is able to recall the correct data value, removing subjectivity from the editing process and the risk of "creative editing".

As an added benefit of this involvement, by engaging statisticians early in the design of the surveys, the surveys became more targeted on the key variables of interest. This process also provided the statisticians with important domain knowledge on the theories driving the research, which transformed their role from a provider of technical support to one in which the statistician can contribute to the design of the research. This involvement of the statistician in the entire

research process follows the suggestion by Bethlehem (1997), stating that the integration of data editing and data collection stages requires that statisticians cease to be specialists in data analysis. Instead, the statistician should have general knowledge and experience in all aspects of the survey process.

For many small-scale surveys, the survey software that implements real-time editing and the necessary technology to support it may not be available due to limited resources and other constraints. In this document, we offer an alternative approach to data collection for these small-scale surveys that offers many of the benefits of the real-time editing procedures. The description of this alternative data collection procedure consists of three main organizational levels. We implemented this procedure for a research project in Maputo, Mozambique, described in Section 2. Section 3 details implementation of the data editing arm of this project. Section 4 analyzes the data and the documentation of the errors. Section 5 provides a discussion of the results and methodology, including advantages, limitations, and improvements.

2. The Maputo Project

The Stanford University program on Water, Health, and Development funded by the Woods Institute for the Environment[2] has established a research program on non-network water and sanitation in developing countries. The Maputo project in Mozambique was one of three projects initiated under this program. In Maputo, the objective was to evaluate the impact of new regulations that legalize the resale of water by households with a water network connection to the population without access to this network. Prior to legalization, many residents of Maputo with access to the water network illegally resold water to their neighbors. The change in policy

[2]See http://woods.stanford.edu/research/centers-programs/water-health-development (Accessed 3/14)

went into effect in September 2010, which allowed a unique opportunity to conduct a baseline and follow-up study of households affected by the legislation. By visiting the same households before and after the policy change took effect, researchers can draw conclusions about the causal relationships between changes in behavior and the resale legalization. This paper discusses data collection in Maputo during the follow-up stage.

The Maputo project involved collaboration between Stanford University and students from a university located in Maputo. In addition, neighborhood guides provided assistance in the identification of households within the neighborhoods. The project selected eight peri-urban neighborhoods around Maputo in which to conduct interviews. Six of the neighborhoods were the same neighborhoods interviewed during the baseline survey. These neighborhoods, referred to as old neighborhoods, were interviewed from the last week of January to the end of March 2012. Prior to the follow-up survey, the project further funded data collection in two more neighborhoods that were not part of the baseline survey. These two neighborhoods, referred to as new neighborhoods, were interviewed from the last week of March to the second week of April 2012.

Due to limitations of resources, a random sampling of households within these eight neighborhoods was not a viable option. Instead, the survey team employed a stratified cluster random sampling design. Households were clustered based on proximity, as each neighborhood was subdivided into parcels approximately 1.5 to 2 hectares in size by divisions created by mapped roads. These parcels were assigned to four strata according to their distance to working standpipes and the municipal network. This stratified sampling allowed for a sample of

households with a wide range of water supply options. Within each sampled parcel, every fourth household was selected to be interviewed until the target of between four and eight sampled households was achieved for the parcel. If a household did not respond after three attempts, another household was sampled from within the parcel. In total, 1,864 households were interviewed, 1,369 household in the old neighborhoods and 495 in the new neighborhoods. Of the 1,396 households interviewed in the old neighborhoods, 1,289 (≈94%) were the same households visited during the baseline study, while the remaining households were new. All 495 households in the new neighborhoods were new since these neighborhoods were added after the baseline study was conducted.

The survey can be broken down into three main sections. The first section collected data on the location of the household and basic information about the people that lived within the household. This basic information included characteristics of the respondent and the people living within the household and whether or not the household had changed their main water source since the baseline survey. The second section collected specific data about the water source or sources used by the household. This section was split into smaller subsections, each associated with a specific type of water source. Based on the answers from the first section, the survey module either activated or skipped these subsections. The third and final section collected data on the effect of any change in water source and the social capital and the wealth of the household. The questions about the change in water source were activated based on the responses in the first section. If the household had changed their primary water source, the respondent was asked about the amount of water collected, the time spent collecting water, and the price of water. The questions related to the wealth of the household asked about the material composition of the

household, its level of access to sanitation, its ownership of assets, and its expenditure and income. The questions related to social capital asked about the relationships the household maintains with its neighbors and its involvement in community activities.

3. Data Editing Methodology

The following section provides a description of the procedure used to review the data collected during the Maputo follow-up survey. Section 3.1 describes the roles of all persons involved with the review process. Section 3.2 provides the daily procedure used to review the data and the duties of each person on that particular day.

3.1 Roles

For the Maputo follow-up survey, five graduate students from Stanford University and Virginia Tech and 23 household surveyors from Maputo implemented the survey. The three Stanford students were on-the-ground survey logistics coordinators, in charge of the day-to-day activities and supervising the surveyors. The two Virginia Tech students were statisticians providing support from the U.S. to the three on-the-ground survey logistics coordinators. The following provides a general description of the duties of the team members.

Surveyors

The 23 surveyors traveled from house to house within the neighborhoods collecting data. Initially, the survey logistics coordinators trained thirteen surveyors on the survey contents and the procedures for informed consent. On the final day of the training, the surveyors conducted pilot interviews on non-sampled households within one of the sampled neighborhoods.

Surveyors worked part-time as their schedules allowed. As the survey progressed in the six old neighborhoods from the baseline study, surveyors worked more and more sporadically as their schedules became more demanding. To ensure the two new neighborhoods were surveyed within the project timeframe, the surveyor logistics coordinators decided that ten more surveyors would be trained and included in the fieldwork team. While the new surveyors worked only in the new neighborhoods, some of the original surveyors also worked in the new neighborhoods once the old neighborhoods were completed.

The survey team used the Hewlett Packard iPAQ personal digital assistant (PDA) devices to collect data during the Maputo follow-up survey. These devices were loaded with The Survey System (TSS)[3] software running on the Microsoft Windows mobile platform. The use of PDA devices allowed for daily access to the data, rather than the delay in coding associated with pen and paper surveys.

Survey Logistics Coordinators

The survey logistics coordinators were the on-the-ground managers of the survey. Section 3.2 discusses their duties related to the data review and cleaning. The other duties performed by these coordinators were as follows: designing the survey questions, coding the survey in TSS, training the surveyors on the contents of the follow-up survey, training the surveyors on informed consent protocols, planning the schedule for survey implementation, travel arrangements for surveyors to and from interview sites, charging PDAs, and providing on-the-ground support to the surveyors while in the field. For the Maputo follow-up survey, two survey

[3] See http://www.surveysystem.com/ (Accessed 3/14)

logistics coordinators managed all field activities, while the third coordinator was a GPS specialist supporting logistics.

Statisticians

The two statisticians at Virginia Tech performed independent reviews of the collected data on a daily basis. The statisticians did not work in Maputo and conducted all data reviews in the U.S. In addition to the data review, the statisticians also provided statistical support to the survey logistics coordinators, such as providing tabulations for surveyor evaluation.

3.2 Data Cleaning Process

The following steps outline the procedure used by the survey logistics coordinators and statisticians to review and correct the data as it was collected. The statisticians reviewed the data on a daily basis, regardless of the number of interviews conducted on a given day. Figure 3.1 describes the flow of data and the duties performed by the survey team members over three days. Since it takes three days to process the data collected from a given day, the review of the data from one day overlaps with the data review from the following two days. The following provides a step-by-step description of the review process for one day of collected data.

Step 1: Data collection and download (day 1 – during the morning and day)

The first step in the data cleaning process was the collection of the data by the surveyors. Once the surveyors were finished with their interviews for the day, the survey logistics coordinators downloaded the data from the PDAs.

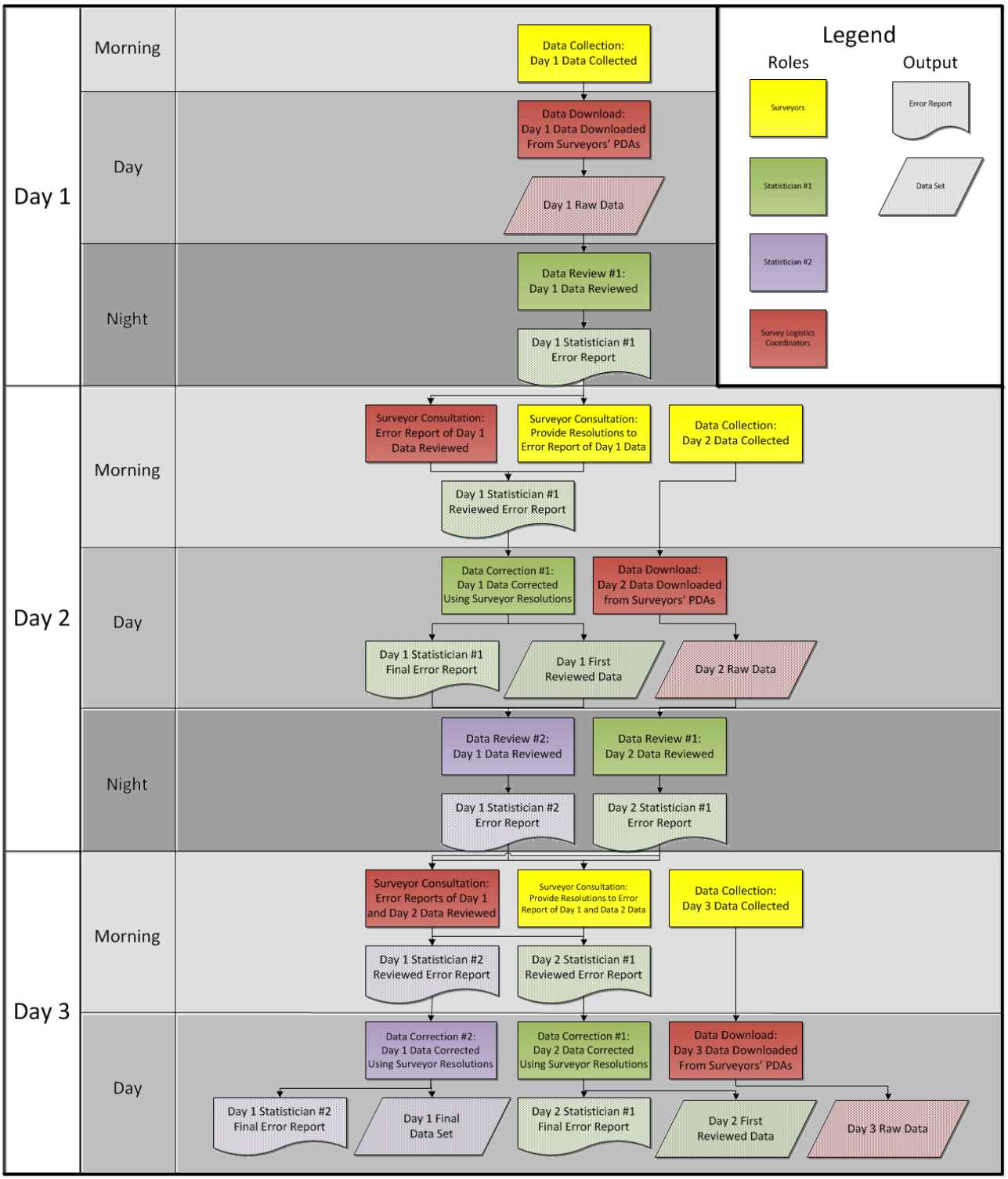


Figure 3.1: Data Cleaning Process

Step 2: Data review #1: statistician #1 (day 1 – during the night)

Once the surveyors collected the data, the survey logistics coordinators on the ground in Maputo transmitted the survey data to the first statistician for data review. The first statistician performed four tasks. The first three tasks were micro-edits, which reviewed individual records for consistency and validity. The fourth edit was a macro-edit, analyzing the data in aggregation.

Check the data for consistency. The data consistency check focused on critical variables and checked the structure of the data. First, several different questions were used to obtain the same data point. The statistician checked these values against one another to ensure consistency. Any inconsistencies in the values collected for the same data point lead to questions about the validity of either response. Additionally, these inconsistencies may propagate into the final summary tables, which may lead to questions about the quality of the data. For example, the Maputo survey collected the number of persons residing within each household and the number of persons in three age groups. Any difference between the total number of persons reported for the household and the sum of the number of persons in the age groups would lead to questions about whether one or both of the values were incorrect. The statisticians flagged any data points showing inconsistencies such as these.

Second, the answers to certain questions activated later sections of the survey. The statistician checked consistency between answers to the setup questions and the sections activated. Inconsistencies between the answers to setup questions and activated sections indicate that there was an error in the survey logic. A household may respond that they use a private tap in the first section of the survey, but the private tap questions in the second section may be left blank. The

identification of this error during the survey has two advantages. First, the logistics coordinator can correct the survey module before the problem extends to later households. Second, if the survey was still being conducted in the same neighborhood, the surveyor could revisit the household and collect the missing data. If the survey had moved on to a new neighborhood, it was not cost effective to revisit a neighborhood to correct the erroneous data point.

Flag suspicious data entries. The PDAs contained small screens that required styluses to input the data. Occasionally, the surveyor erred in recording the data. A surveyor may intend to enter two hours spent collecting water, but accidentally record 22 hours. Left unchecked, this potentially incorrect value would alter the reported values in the final summary tables for the community and other tabulations that include this data point. The statistician flagged as possible errors data entries that seemed suspicious and possibly incorrectly recorded.

Flag data omissions. There were cases in which the respondent did not know the answer to the question or refused to answer. The survey logistics coordinators instructed the surveyors to record these values as some variation of "9999", but sometimes the surveyors did not record any values and left the field blank. There were also cases in which a respondent answered zero for a question, but the surveyor left the field blank. It is impossible for analysts to determine whether the respondent did not know or refused to answer or the respondent responded with zero. When analyzing the data, treating these values as all zero may introduce a bias into any statistics calculated. Treating these values as all missing may reduce the sample size. As part of the data review, the statistician flagged any fields incorrectly left empty to determine whether the value should be missing or zero.

Tabulate important indices. In addition to the value-by-value review of the data, the statistician tabulated important indices such as liters of water per capita per day, household activities using water, and household expenditure. These tabulations provided an early indication of the results of important indices and ensured the surveyors as a whole were correctly collecting the data.[4] The first statistician created an error report for the data collected for each day, listing all errors found during any step of the error review. This error report contained the surveyor that conducted the interview, the household ID number, the first name of the respondent, a description of the error, the variable(s) in error along with the recorded value(s), and the question number(s) in which the error occurred. Table 3.1 provides an example of entries of the error report. Note that the first statistician fills in the first six columns of the table. Survey logistics coordinators fill the last column in Step 3.

Table 3.1: Sample Error Report

Surveyor Name	Household ID	Name of Respondent	Problem	Variable Name and Original Value	Question Number	Correction
Mike	00001	Teresa Smith	Should be April?	FL_MONTH=3	5	FL_MONTH=4
	00002	Cristina Jones	Goes on big monthly shopping trip but spends 0.	FL_MONTHLY_BIGSHOP_YN=1 FL_MONTHLY_BIGSHOP_AMT=0	791-792	FL_MONTHLY_BIGSHOP_YN=2 FL_MONTHLY_BIGSHOP_AMT= should be empty
Jane	00011	Joana Williams	New primary source but did not change baseline source.	FL_BL_SOURCE_OLDSOU_NO_CONT=2 FL_BL_SOURCE_PRIMARY_NOW=1 FL_BLSOURCE_CHANGE=2	46-48	FL_BLSOURCE_CHANGE=1

Additionally, the first statistician periodically generated tabulations of important indices for each surveyor to determine if there were any systematic differences between the data collected by a given surveyor and the other surveyors. The following describes reasons why a systematic difference may be found between surveyors in a CAPI study.

[4] In a previous impact evaluation, the research team shared summary statistics generated by the statistician from the field with all research partners via weekly update. This action proved extremely valuable, since it provided some transparency to the research and helped the research team gain the trust of those entities under evaluation. These entities, in turn, were more forthcoming with information about the project.

PDA error. There may be a problem with a surveyor's PDA that neither the survey logistics coordinators nor the surveyor realizes. This occurred in Maputo when one of the surveyor's PDAs contained an error in the logic of the survey. One of the survey questions asked which sources of water were available to the household. Based on the answer to these questions, the PDA activated certain subsections of the water source section. For this particular surveyor, the PDA activated the wrong sections. The statistician identified this error early using these tabulations, and the survey logistics coordinator corrected the survey module in that PDA.

Surveyor misunderstanding. Despite the extensive training given to the surveyors prior to survey implementation, the surveyor may not understand exactly what data should be collected for a given question. Questions that dealt with monthly expenditures are one common example of this. One sequence of questions asked how much the household spends on individual monthly expenditures such as charcoal, medicine, and meat. Another question asked how much money the household spends on other expenses during their monthly shopping trip to the market. Some surveyors included all monthly expenses in the monthly shopping trip and incorrectly reported zero for the individual monthly expense categories. The statistician identified this error early in some of the surveyor reports. The survey logistics coordinators clarified the question and retrained the surveyors, after which these errors stopped occuring.

Interview miscommunication. Miscommunication between the surveyor and the respondent may result in the collection of incorrect data. Respondents may interpret terminology in the survey as referring to one thing, while the survey calls for another. An example of this miscommunication in Maputo was the interpretation of what sources constitute a borewell and a standpipe. Some

respondents reported using a borewell, when in fact, the household used a standpipe. Using the tabulations provided by the statisticians, the survey logistics coordinators, using their knowledge of the water sources used the neighborhood, identified households where this miscommunication may have occurred and resolved the potential error with the surveyor.

Intentional survey error. Due to the logic of the survey, some answers entered for particular questions may cause the PDA to skip one or many questions in the survey. Surveyors may falsify data to intentionally shorten the length of the interview. Surveyors can significantly reduce the amount of time spent during an interview by reporting no water source, thus skipping the entire second section of the survey. The surveyor report included tabulations of the total amount of water reported, which would indicate zero estimates of total water from sources. The survey logistics coordinators fired surveyors who repeatedly reported households using zero water.

Step 3: Surveyor consultation #1 (day 2 – during the morning)

The survey logistics coordinators consulted the surveyors on the errors found by the first statistician in interviews from the previous day. Due to the short turnaround period, the surveyors were able to resolve errors in most cases, and the survey logistics coordinators added the resolution to the error report (shown in the last column of Table 3.1).

Along with the PDA, surveyors carried notebooks in the field, which enabled them to write down any relevant information not recorded in the survey module. In a few cases, the surveyor

identified a value recorded in error or instances of PDA malfunction. The survey logistics coordinators added these errors and their resolutions to the error report.

Step 4: Data correction #1: statistician #1 (day 2 – during the day)

After consulting the surveyors, the field logistics coordinators returned the error report with resolutions to the first statistician. The statistician then made the appropriate corrections to the data file and indicated that the error was resolved on the error report.

The first statistician relayed any outstanding errors from the error report to the second statistician and survey logistics coordinators. These outstanding errors included any errors that the first statistician could not resolve based on the corrections entered in the error report by the logistics coordinators. The second statistician was responsible for resolving these outstanding errors the next day (Step 5), once the survey logistics coordinators obtained the corrections from the surveyors.

Once all appropriate corrections were made, the first statistician transmitted the data and updated error report to the second statistician.

Step 5: Data review #2: statistician #2 (day 2 – during the night)

The second statistician acted as a second level of review, performing another thorough review of the data independent of the first statistician. While not generating the tabulations of important indices or the surveyor reports, the second statistician performed a similar role as the first

statistician. The second statistician added any errors identified to the error report created by the first statistician.

Step 6: Surveyor consultation #2 (day 3 – during the morning)

With the additions to the error report from the second statistician, the survey logistics coordinators consulted the surveyors for a second time. The surveyors provided corrections on the error report to any unresolved errors, including those outstanding from the review of the first statistician.

Step 7: Data correction #2: statistician #2 (day 3 – during the day)

After the second round of consultations with the surveyors, the survey logistics coordinator returned the error report to the second statistician with the resolutions to all outstanding errors. The second statistician corrected the entries on the data file according to these resolutions.

Step 8: Data storage: field logistics coordinators (day 3 – during the night)

Once the second statistician corrected all outstanding errors, the data file for a given day was ready for storage. For the Maputo follow-up survey, the second statistician transmitted the final reviewed data file back to the survey logistics coordinators, who added the data to the previously reviewed data. The second statistician may also act as the custodian of the data during the survey and add the newly corrected data files to the previously reviewed data.

4. Analysis

This section provides an analysis of the data collected during the error correction process. Section 4.1 provides general descriptive statistics. Section 4.2 discusses possible effects of the data cleaning procedure.

4.1 General Descriptive Statistics of Errors

Of the 1,864 households, 900 (≈48%) contained at least one data value that required editing by the statistician after the consultation of the surveyor by the field logistics coordinator. Due to the short turnaround period of the data review, the surveyor was able to resolve the errors in almost all cases, leaving very few instances in which the surveyor was required to revisit a household. The following lists some of the common types of errors edited by the statistician:

Incorrect Missing Values. The surveyors did not record a value for a question that required a value based on the logic of the survey. For example, a household indicated that it uses a neighbor's connection for water, but no value was entered for time waiting in line. Without consultation of the surveyor, it would not be known whether the household did not know the wait time (correct entry should be 99) or the household spent no time waiting in line.

Inconsistencies in the collected data. The surveyor collected data that was inconsistent with other data points. For example, the questionnaire asked the total number of hours each water source used by the household was available for use and the times that they were available. The statistician flagged these data values if the total number of hours reported did not equal the total number of hours calculated from the times the source was reported to be available. In some

cases, the total number of hours was the correct value, while in others, the times reported were correct. The surveyor was consulted to determine which data value was correct.

Suspicious Values. During the data review, the statisticians flagged data values that were suspicious and may have been collected in error. These values were identified by their relationship to the other data collected from households within the same neighborhood. For example, one surveyor entered a value of 2.5 for the amount spent per month at the local market. The statisticians flagged this value as suspicious since the other data collected from the same neighborhood were generally larger. The field logistics coordinators consulted the surveyor and found that the value was incorrectly entered into the PDA and the true value was 2,500.

Table 3.2 provides a summary of the number of errors corrected by the review stage. The first statistician identified and corrected a large percentage (88%) of the errors. Of the errors corrected by the second statistician, only a few were new errors. The majority of the corrections made by the second statistician were unresolved errors identified by the first statistician. The corrections made by the second statistician allowed the first statistician to focus on the data collected one day at a time, rather than correct data from multiple days. The inclusion of a second statistician spread the data management burden between the two statisticians, while enabling each of them to focus on new analysis.

Table 3.2: Errors Corrected by Review Stage

Review Stage	Number of Errors Corrected	Percentage of Total Errors
Statistician #1	1,275	88%
Statistician #2	173	12%
Total	1,448	100%

Table 3.3 provides the tabulations of the number of errors for each surveyor and the two surveyor groups. Between the two groups of surveyors, the second surveyor group committed a larger number of errors per survey than the first surveyor group. The first group on average committed approximately 0.7 errors per household, with a range from 0.46 to 0.99. The second group on average committed approximately 1.1 errors per household, with a range of 0.61 to 2.8. This discrepancy may be attributed to limitations of the second survey group and their training. There were fewer people available to train the surveyors in the second group due to on-going management of the first group of surveyors in the old neighborhoods. Half of the first group conducted a similar survey in 2010, while all surveyors in the second group were new to this type of survey. While the second group contained experienced surveyors, they had not administered a survey of the type used in Maputo before or used a PDA.

Table 3.3: Errors Corrected by Surveyor

Surveyor	All Surveys		Old Neighborhoods		New Neighborhoods	
	Total Households	Errors/ Household	Total Households	Errors/ Household	Total Households	Errors/ Household
1	3	0.67	3	0.67	0	0
2	86	0.80	86	0.80	0	0
3	153	0.67	132	0.71	21	0.38
4	111	0.71	99	0.63	12	1.42
5	151	0.81	133	0.85	18	0.50
6	163	0.60	146	0.58	17	0.82
7	30	0.93	30	0.93	0	0
8	138	0.99	120	1.03	18	0.78
9	182	0.73	161	0.80	21	0.19
10	165	0.81	146	0.88	19	0.26
11	83	0.65	83	0.65	0	0
12	163	0.49	133	0.53	30	0.33
13	102	0.46	97	0.45	5	0.60
14	41	0.71	0	0	41	0.71
15	34	0.91			34	0.91
16	19	2.05			19	2.05
17	32	1.06			32	1.06
18	39	1.05			39	1.05
19	12	2.75			12	2.75
20	44	0.61			44	0.61
21	30	1.63			30	1.63
22	41	1.10			41	1.10
23	42	0.76			42	0.76
Surveyor Group 1	1,530	0.71	1,369	0.73	161	0.52
Surveyor Group 2	334	1.08	0	0	334	1.08

4.2 Data Cleaning Effect

Tables 3.4 and 3.5 provide a comparison of the number of non-missing responses and the coefficient of variation of important indices in the Maputo data before and after data editing. The coefficient of variation is widely used in surveys as an indicator of the precision of the mean estimate. The data cleaning procedures generally had a non-trivial effect as the coefficient of variation of these variable decreased, particularly in the data collected in the new neighborhoods. Note that no imputation was performed for missing data. Only observed values were used for the calculation of the coefficient of variation.

Table 3.4: Variable Summaries from Old Neighborhoods

Variable	Before		After		Relative % Difference $\left(100 \times \frac{(After-Before)}{Before}\right)$
	Number of Non-Missing Responses	Coefficient of Variation (σ/μ)	Number of Non-Missing Responses	Coefficient of Variation (σ/μ)	
Water Use Yesterday (LPCD)	1365	0.63	1369	0.57	-8.57%
Total Water Activities (LPCD)	1363	0.53	1365	0.52	-0.66%
Total Expenditures (MZN/Month)	1348	0.66	1360	0.67	1.49%
Water Consumption (LPCD)	1322	0.71	1331	0.68	-4.16%
Water Consumption Cold (LPCD)	1287	0.78	1296	0.73	-5.68%
Water Consumption Hot (LPCD)	1322	0.69	1331	0.63	-8.68%
Number of Rooms in Household	1354	0.37	1366	0.37	-0.39%
Number of Households Using Standpipe	180	1.40	181	1.40	-0.03%
Public Well Trips Per Day Cold	34	0.74	37	0.72	-3.10%
Public Well Trips Per Day Hot	35	0.69	38	0.63	-9.05%
Number of Households Using Neighbor's Tap	267	0.50	265	0.42	-15.95%
Standpipe Total Containers Collected Per Day Cold	167	0.76	171	0.52	-31.81%
Standpipe Total Containers Collected Per Day Hot	171	0.51	175	0.52	0.95%

Table 3.5: Variable Summaries from New Neighborhoods

Variable	Before		After		Relative % Difference $\left(100 \times \frac{(After-Before)}{Before}\right)$
	Number of Non-Missing Responses	Coefficient of Variation (σ/μ)	Number of Non-Missing Responses	Coefficient of Variation (σ/μ)	
Water Use Yesterday (LPCD)	474	0.80	480	0.64	-19.80%
Total Water Activities (LPCD)	473	0.56	478	0.55	-0.74%
Total Expenditures (MZN/Month)	466	0.70	477	0.67	-4.72%
Water Consumption (LPCD)	463	0.89	465	0.61	-30.90%
Water Consumption Cold (LPCD)	448	0.97	449	0.64	-33.76%
Water Consumption Hot (LPCD)	463	0.83	465	0.60	-27.25%
Number of Rooms in Household	468	0.66	479	0.38	-41.75%
Number of Households Using Standpipe	29	0.93	33	0.72	-22.71%
Public Well Trips Per Day Cold	21	0.82	23	0.60	-27.32%
Public Well Trips Per Day Hot	23	0.63	25	0.42	-33.37%
Number of Households Using Neighbor's Tap	31	0.43	29	0.24	-45.53%
Standpipe Total Containers Collected Per Day Cold	36	0.58	35	0.51	-10.94%
Standpipe Total Containers Collected Per Day Hot	37	0.55	39	0.57	3.30%

Figure 3.2 provides a scatterplot of the relationship between the amount of experience a surveyor had and the number of errors he or she committed. Each data point provides the average number of errors committed by surveyors in the listed survey group and the amount of experience (in days worked) the surveyor had at the time of the survey. The scatterplot also provides regression line fit to the data points for each survey group, along with the 95% confidence limits around the regression line. The regression line for the original thirteen surveyors is flat, indicating there was no relationship between the amount of experience the original surveyors had at the time of the interview and the number of errors they committed. The regression line for the additional ten surveyors shows a negative slope, indicating a decrease in the number of errors a surveyor committed as he or she gathered more experience implementing the survey. These two findings suggest that the additional ten surveyors benefitted more from the review process than the original thirteen surveyors.

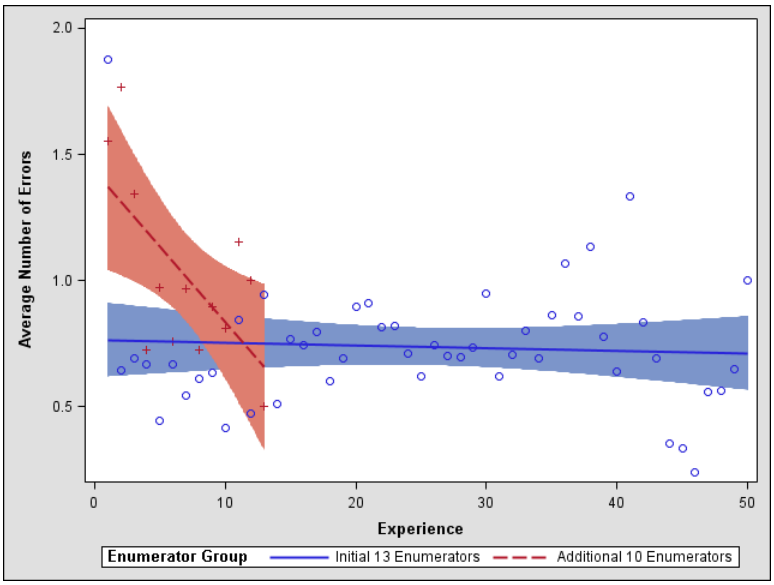


Figure 3.2: Relationship between Experience and Errors Committed

5. Methodology Discussion

The data cleaning procedures outlined in this paper provide an alternative to fully automated data editing for small-scale surveys in which resources are limited. In cases such as the Maputo survey, time constraints and resources made it impossible to implement a fully automated data editing procedure. The TSS program run on the PDAs only allowed for hard coded limits to values entered. For many data points, predetermining these limits was not reasonable because some outliers turned out to be legitimate data points and not data entry errors. The TSS program did not allow automated flags to ask the surveyor if data points were legitimate. The only recourse in this situation was to consult the surveyor. The field logistics coordinators took into account the surveyor feedback during training and the results of the pilot study while constructing the survey. The final version of the survey was therefore not available until a short time before the first day of survey implementation. This timeline would have made it difficult to program the automated edits into the survey instrument prior to the start of the survey.

Given these limitations in resources and timing, these data cleaning procedures provide many of the benefits associated with fully automated data editing. The following discusses the benefits of the data cleaning procedures outlined in this paper, as well as a limitation. We begin with the benefits.

Rapid feedback (response) to (from) surveyors. Data entry errors can be corrected by consulting the surveyors. For most errors in data entry, surveyors can recall the interview given the information listed on the error report and provide the intended response. For traditional paper-based surveys, consulting the surveyors after the data entry process is complete would either not be possible since the surveyors may not be available or not productive since the surveyors would not be able to recollect the intended response. The only recourse in these situations would be to either delete the data or estimate the value. This communication process among the entire team means that there is a much greater ability to identify and correct problems with the survey.

Improved Survey Management. Survey logistics coordinators can concentrate on the management of the data collection process. In some surveys, the logistics coordinators analyze the data during collection. Due to the time constraints, the logistics coordinators are limited to selective editing and macro-editing. With statisticians as part of the team, the survey logistics coordinators can concentrate on implementing the survey, while the statisticians provide micro-editing and macro-editing of all incoming data.

Real-time monitoring of surveyors and survey results. As a consequence of more thoroughly reviewed data, tabulations can be made evaluating the performance of the surveyors as a whole

and for each individual surveyor. If it is determined that certain indices are suspicious for all surveyors, the surveyors may have misinterpreted the question during training. In this case, the survey logistics coordinators can retrain the surveyors early in the survey implementation. The tabulations evaluating individual surveyors can show whether one or more surveyors are collecting systematically different data in some cases than the other surveyors. This may be a result of misinterpretation during training, which the survey logistics coordinator can correct early in the survey implementation. This may also be a result of intentionally falsified data, in which case the survey coordinators can take appropriate disciplinary action. With these advantages, the data cleaning procedure presented above achieves the second goal of data editing proposed by Granquist (1984).

Statisticians as subject matter experts. The statisticians for the Maputo follow-up survey were graduate level students, who had experience with similar types of household surveys undertaken in Africa. These two statisticians were able to use this past survey experience, experience with statistical software packages, and general statistical knowledge to analyze and present the data in ways scientists from other disciplines would not be able to. These findings reinforce the assertion by Bethlehem (1997) that data editing improves when the statistician becomes a subject matter expert rather than just a data analyst.

Extended training of surveyors. The overall data review and cleaning process described in this paper led to several benefits that emerged from the constant level of communication required among members of the research team. Within the early days of the fieldwork, the surveyors became acutely aware of the importance of completing accurate surveys. Further, the regular

feedback they received through the error review process effectively continued their training through the entire fieldwork. This approach stands in stark contrast to the more traditional method of training surveyors in one intensive period before the fieldwork, which we argue is an inadequate process based on our experience. Without this level of data review, a number of significant data collection errors would have persisted throughout the fieldwork, which would ultimately have affected the research findings.

A limitation in the implementation of the data cleaning procedure was the statisticians provided data analysis support from the U.S., away from the survey implementation site in Maputo. This resulted in a lack of familiarity with the day-to-day operations and the surveyors and limitations in the communication between the statisticians and the survey logistics coordinators. This lack of familiarity resulted in an insufficient understanding of the characteristics of the surveyed neighborhoods and surveyors' personalities. Knowledge of both improves the data review process for the statisticians. Communication between the statisticians and survey logistics coordinators was limited to email exchange. The communication could have been improved with increased discussion had at least one statistician been on the ground, rather than both located remotely. In a previous research project, LISA was able to provide an on-the-ground statistician who was able to review the data while traveling with the fieldwork teams. This level of connectedness removed the communication problems that exist with emails or explanations of complex problems in a written error report.

References

Bethlehem, J.G. (1987). The Data Editing Research Project of the Netherlands Central Bureau of Statistics, *Proceedings of the Third Annual Research Conference of the Bureau of the Census*, 194-203.

Bethlehem, J.G. (1997). Integrated Control Systems for Survey Processing. In: Lyberg, L., Biemer, P., Collins, M., De Leeuw, E., Dippo, C., Schwarz, N. and Trewin, D. (Eds.), *Survey Measurement and Process Quality*, New York, NY: John Wiley and Sons Inc., 371-392.

Biemer, P. and Caspar, R. (1994). Continuous Quality Improvement for Survey Operations: Some General Principles and Applications, *Journal of Official Statistics*, 10, 307-326.

Deming, W.E. (2000). *Out of the Crisis*, Cambridge, Mass.: MIT, Center for Advanced Educational Services.

De Waal, T., Pannekoek, J., and Scholtus, S. (2011). *Handbook of Statistical Data Editing and Imputation*, Hoboken, NJ: John Wiley and Sons.

Granquist, L. (1984). On the Role of Editing, *Statistical Review*, 105-118.

Gravlee, C.C., Zenk, S.N., Woods, S., Rowe, Z., and Schulz, A.J. (2006). Handheld Computers for Direct Observation of the Social and Physical Environment, *Field Methods* 18 (4): 382-97.

Hartley, H.O. (1980). Statistics as a Science and as a Profession, *Journal of the American Statistical Association*, Vol. 75, No. 369, 1-7.

Hughes, P.J., McDermid, I., and Linacre, S. (1990). The Use of Graphical Methods in Editing, *Proceedings of the Annual Research Conference*, U.S. Bureau of the Census, Washington, DC, 538-550.

Lyberg, L., Biemer, P., Collins, M. (1997). *Survey Measurement and Process Quality*, New York: John Wiley and Sons.

Onono, M., Carraher, N., Cohen, R., Bukusi, E., and Turan, J. (2011). Use of Personal Digital Assistants for Data Collection in a Multi-Site AIDS Stigma Study in Rural South Nyanza, Kenya, *African Health Services*, Volume 11, No. 3, September 2011.

CHAPTER 4: A MODELING APPROACH TO ESTIMATING THE MEAN SQUARED ERROR OF SYNTHETIC SMALL AREA ESTIMATORS

Chapter 4 Abstract

Synthetic estimators for population quantities of interest are often used for small areas, where direct survey estimates are either not possible or are unreliable. While these synthetic estimators have smaller sampling variability than do the direct estimators, the pooling of data across areas of estimates may introduce additional potential error known as synthetic estimation error, which can easily dominate the mean squared error (MSE) of the synthetic estimates. Estimates of synthetic estimation error for a single area are very unstable, so previous methods proposed in the literature average such estimates over groups of areas with the assumption of constancy of squared synthetic bias within the groups. In this paper, we propose a generic model that facilitates MSE estimation for synthetic estimators by assuming a parametric model for the synthetic error variance that depends on some covariates available for each area. This includes constant synthetic error variance within groups of areas as a special case. An extension of our basic model incorporates multiplicative random effects, which effectively achieves shrinkage estimation of the synthetic error variances. We compare our proposed modeling approach to methods proposed in the literature in a simulation study. We then apply several of the methods to a simulation motivated by synthetic estimates of census correct enumeration rates in the 2010 Census Coverage Measurement (CCM), and then also to the actual CCM data. Our proposed methods, particularly the model with random effects, compare favorably in the simulations to the methods proposed in the literature.

1. Introduction

Sample surveys are widely used to investigate various characteristics of populations of interest. Based on the data collected from a subset of the population, researchers produce estimates and make inferences about the population as a whole. These estimates and inferences are also made about subsets of the population, referred to generically as domains, such as geographic areas or socio-economic groups. Direct survey estimates for domains, those based entirely on data collected from respondents within each domain, will be unreliable for domains with small samples. For domains with no units sampled, direct survey estimation is not possible. For simplicity, and in keeping with common usage, we shall henceforth refer to domains with little or no sample as *small areas* rather than small domains. The ideas presented here apply equally to domains other than geographic areas.

One approach to dealing with unreliable direct survey estimates for small areas is to pool data across areas of estimates, also called "borrowing strength". The resulting estimates, known as *synthetic estimates*, have smaller variability due to the additional sample size "borrowed" from other areas. This pooling of data across areas introduces additional potential error known as synthetic estimation error. If synthetic error is regarded as fixed (nonrandom), it can be viewed as bias. The Mean Squared Error (MSE) of the synthetic estimate is then the sum of its sampling variance and the squared synthetic bias. With substantial pooling of data across areas, the sampling variance of the synthetic estimate may be small, but the synthetic bias is potentially large and can easily dominate the MSE. In these cases, reporting sampling variances of synthetic estimators as measures of precision is misleading. Unfortunately, Rao (2003 pg. 52) notes that

while the variance of these estimates are readily obtained through standard design-based methods, it is more difficult to estimate the squared bias component of the MSE.

A small number of methods (discussed in Section 2) have been proposed in the literature for estimating the MSE of synthetic estimators. Equivalently, since the sampling variances of synthetic estimates can be readily estimated, these can be viewed as methods for estimating the squared synthetic bias. This is a difficult problem if the synthetic bias is viewed as unique to each area, since the estimates of squared synthetic bias for a single small area are very unstable. To address this issue, proposed methods have generally averaged such estimates of squared synthetic bias across groups of similar areas. These grouped estimates can provide an average measure of precision of the synthetic estimators, which may be adequate for some purposes, such as comparing the performance of one synthetic estimator to another. However, the grouped estimates do not provide a measure of precision for specific individual areas, except under an implicit assumption that the synthetic bias is equal across areas within a group. This assumption may not be valid when small areas exhibit strong area-specific effects (Rao 2003 pg. 53).

An alternative perspective, adopted in this paper, views the synthetic error as random with mean zero and some variance. This variance may be assumed to be constant across groups of small areas, or follow a parametric function dependent on some covariate(s) available for each area. We pose estimation of synthetic error variance as a modeling problem with two parts: (*i*) specification of a parametric function for the synthetic error variance, and (*ii*) estimation of the parameters of that function using the data. Proposed methods have tended to blend these two aspects together, obscuring their distinct characteristics. This paper focuses mostly on the

second aspect of the modeling approach, which we study by simulation. We also illustrate the modeling approach on a real application involving estimation of census correct enumerations using data from a post-enumeration follow-up survey.

More specifically, the paper proceeds as follows. Section 2 reviews previously proposed methods for MSE estimation of synthetic estimators, and compares several of these in a simulation study. Section 3 proposes a generic model relating direct survey and synthetic estimates, which facilitates estimation of simple parametric models for the synthetic error variance. We consider some alternative models and estimation approaches, including a variance model with random effects that permits empirical Bayes shrinkage estimation of the synthetic error variance. These alternatives are compared in several sets of simulations that include cases where the correct model is used and cases with model misspecification. Section 4 applies several of the methods to an application taken from the 2010 Census Coverage Measurement (CCM). The application involves estimating the MSE of synthetic estimates of census correct enumeration rates. We present results both for a simulation motivated by this application and for the actual CCM data. Section 5 summarizes the paper's results and discusses areas for future research.

2. Review of Design-Based Methods

We use the following general notation provided in Table 4.1 throughout the paper. We will introduce additional notation as needed.

Table 4.1: General Notation

Notation	Statistical Quantity
Y_i	Population characteristic of interest for area *i*
y_i	Direct survey estimate of Y_i, assumed to be unbiased
$e_i = y_i - Y_i$	Sampling error of y_i, which has $var(e_i) = \sigma_{e_i}^2$
y_i^s	Synthetic estimate of Y_i
$Bias(y_i^s) = E[y_i^s] - Y_i$	Synthetic bias of y_i^s from a design-based perspective

In general, an estimator is considered synthetic for a small area if it is derived from a reliable direct estimate of a larger area that contains the small area. Synthetic estimation requires the assumption (called the synthetic assumption) that the small areas have the same characteristics as the larger area. Rao (2003 pg. 46 – 51) discusses various types of synthetic estimators such as ratio estimators and regression synthetic estimators. Additional synthetic estimators can be obtained as the mean function from any fixed effects model for y_i.

Violation of the synthetic assumption introduces synthetic error into the synthetic estimates. If we consider the synthetic error as a fixed effect, then it can be viewed as bias and the MSE of the synthetic estimator is the sum of the sampling variance of the synthetic estimate and the squared synthetic bias. Rearranging the terms of the MSE equation, we obtain the following expression for squared synthetic bias in terms of the MSE and the sampling variance of the synthetic estimate.

$$Bias^2(y_i^s) = MSE_i - Var\left[y_i^s\right] \tag{1}$$

Rao (2003 pg. 52) derived the following approximately unbiased estimate of the MSE of a synthetic estimate under the assumption that the sampling variance of the synthetic estimate is small.

$$MSE_i \approx (y_i^s - y_i)^2 - \sigma_{e_i}^2 \quad (2)$$

Combining Equations 1 and 2, we derive an expression for the squared synthetic bias, which we will also refer to as the Method of Moments (MOM) estimate.

$$Bias^2(y_i^s) = (y_i^s - y_i)^2 - \sigma_{e_i}^2 - Var\left[y_i^s\right] \quad (3)$$

This unbiased estimator of the MSE and squared synthetic bias tends to be unstable. Gonzalez-Waksberg (1973) proposed averaging the MSE values over groups of similar areas to obtain a more stable estimator, as shown below. The index *D* references the group and m_D is the number of areas assigned to group *D*.

$$MSE_D^{GW} = \frac{1}{m_D}\sum_{i\in D}(y_i^s - y_i)^2 - \frac{1}{m_D}\sum_{i\in D}\sigma_{e_i}^2 = F_{1,D} - F_{2,D} \quad (4)$$

$$\text{where } F_{1,D} = \frac{1}{m_D}\sum_{i\in D}(y_i^s - y_i)^2 \text{ and } F_{2,D} = \frac{1}{m_D}\sum_{i\in D}\sigma_{e_i}^2$$

Subtracting the estimated variance of the synthetic estimate, we derive an estimate of the squared synthetic bias.

$$\widehat{Bias}^2(y_i^s)^{GW} = MSE_D^{GW} - Var\left[y_i^s\right]$$

If used to provide MSE estimates for individual areas, the Gonzalez-Waksberg estimator implicitly assumes that the MSE within the group of areas is approximately constant among the areas. Marker (1995) proposed a method similar to that of Gonzalez-Waksberg that allows for

area-specific estimates of MSE. Instead of assuming that the MSE is constant within groups, Marker (1995) assumed that the squared synthetic bias is approximately constant for areas assigned to the same group, and estimated it by averaging over groups of similar areas.

$$\widehat{Bias}^2(y_i^s)^M = F_{1,D} - F_{2,D} - F_{3,D}$$

$$\text{where } F_{3,D} = \frac{1}{m_D}\sum_{i \in D} Var\left[y_i^s\right]$$

Marker (1995) suggested using the corresponding area specific MSE measure.

$$MSE_i^M = Var\left[y_i^s\right] + \widehat{Bias}^2(y_i^s)^M \tag{5}$$

In cases where the sampling variance of the direct survey estimate is large relative to the squared synthetic bias, both the Gonzalez-Waksberg method and Marker method can provide negative estimates. Lahiri and Pramanik (2012) thus proposed adjustments to these two methods that provide strictly positive estimates, while still being design consistent. The following provides their adjustment to the Gonzalez-Waksberg method.

$$MSE_D^{GW-Adj} = \frac{2F_{1,D}}{1+exp\left(\frac{2F_{2,D}}{F_{1,D}}\right)} \tag{6}$$

Subtracting the estimated sampling variance of the synthetic estimate from Equation 6, we derive a corresponding estimate of the squared synthetic bias.

$$\widehat{Bias}^2(y_i^s)^{GW-Adj} = MSE_D^{GW-Adj} - Var\left[y_i^s\right]$$

Using the following two approximations, the adjusted estimator is approximately equal to the Gonzalez-Waksberg estimator.

$$exp\left(\frac{2F_{2,D}}{F_{1,D}}\right) \approx 1 + \frac{2F_{2,D}}{F_{1,D}} \qquad\qquad \left(1 + \frac{F_{2,D}}{F_{1,D}}\right)^{-1} \approx 1 - \frac{F_{2,D}}{F_{1,D}}$$

Lahiri and Pramanik (2012) discuss that, in some cases, this method produces MSE estimates that are less than the naïve estimator, or the estimator under the assumption of no synthetic bias. In these situations, the estimated squared synthetic bias is negative. Lahiri and Pramanik (2012) introduced an adjustment to the Marker estimator, which does not have this drawback and guarantees that the estimated squared synthetic bias will be strictly positive.

$$\widehat{Bias}^2(y_i^s)^{M-Adj} = \frac{2F_{1,D}}{1 + exp\left(\frac{2\left(F_{2,D} + F_{3,D}\right)}{F_{1,D}}\right)}$$

The corresponding MSE estimator for the adjustment to the Marker estimator is then:

$$MSE_i^{M-Adj} = Var\left[y_i^s\right] + \widehat{Bias}^2(y_i^s)^{M-Adj}$$

2.1 Simulations with Constant Synthetic Error Variance (Over Groups of Areas)

We defined the following simulation study to mimic data collected from a sample survey. We will vary the specifications of the synthetic error variance and evaluate the methods of MSE

estimation of synthetic estimators. The simulation data contained 1,000 replicate data sets. Each replicate data set was split into *m* areas, with each area *i* containing N_i =*500* units.

For area *i* and unit *j*, we simulated data from the following distributions.

$$y_{ij}|\theta_i \sim Normal(\theta_i, \sigma_e^2)\ for\ j = 1, \dots, N_i$$

$$\theta_i|\mu, \sigma_{v\,i}^2 \sim Normal\left(\mu, \sigma_{v\,i}^2\right)\ for\ i = 1, \dots, m$$

The values of μ and σ_e^2 were set to 10 and 1, respectively, and remained constant for all simulated data. We varied the values of $\sigma_{v\,i}^2$, as several specifications corresponding to various assumptions were investigated. We investigated two values of *m*, 100 and 250. In both cases, areas were assigned to five mutually exclusive groups, with $m_D = 20$ (for $m = 100$) and $m_D = 50$ (for $m = 250$) areas assigned to each group *D*. θ_i is the superpopulation mean for area *i*.

From the population of N_i =*500* units in area *i*, n_i=8 units were subsampled. We researched various numbers of sampled units, but the relative magnitudes of σ_e^2/n_i and $\sigma_{v\,i}^2$ are what really matters for the relative error measures examined here. We choose to fix σ_e^2 at 1 and n_i at 8, while varying $\sigma_{v\,i}^2$.

Lahiri and Pramanik (2010) simulated data in the same manner, although their simulation contained *m*=20 areas and all were assigned to the one group. Lahiri and Pramanik researched two combinations of σ_v^2 ($\sigma_{v\,i}^2 = \sigma_v^2$ for all *i*) and σ_e^2: $(\sigma_e^2, \sigma_v^2) = \{(50,1); (1,10)\}$. These two

combinations represent the two extremes where the σ_e^2 is large in comparison to σ_v^2, and vice versa.

For a given population and subsample, we define the following estimates and their variances. Note that the synthetic estimates and their sampling variances are constant across areas for a given replicate.

Direct Survey Estimate: $$\hat{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$$

Synthetic Estimate: $$\hat{y}_i^s = \sum_{i=1}^{m_D} \frac{N_i}{N}\hat{y}_i$$

Direct Sample Variance Estimate: $$\hat{\sigma}_{e_i}^2 = \left(1 - \frac{n_i}{N_i}\right)\frac{s_i^2}{n_i} \; where \; s_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}\left(y_{ij} - \hat{y}_i\right)^2$$

Sampling Variance of the Synthetic Estimate: $$\widehat{Var}\left[y_i^s\right] = \sum_{i=1}^{m_D}\left(\frac{N_i}{N}\right)^2 \hat{\sigma}_{e_i}^2$$

We first simulated data under the assumption that the variance component $\sigma_{v\,i}^2$ was constant within each of the five mutually exclusive groups. Three specifications of the magnitude of $\sigma_{v\,i}^2$ were investigated. The specifications range from simulations where the magnitude of $\sigma_{v\,i}^2$ is large in comparison to the variance component σ_e^2 to simulations where $\sigma_{v\,i}^2$ is small in comparison to the σ_e^2. In Section 3, we provide additional simulations under differing assumptions about $\sigma_{v\,i}^2$.

Table 4.2 provides the Relative Root Mean Squared Error (RRMSE) of the estimated squared synthetic bias from each design-based method to the true value specified in the simulation. We calculate the RRMSE for this table and all others that follow using Equation 7, where the index *D* references groups of areas (or all areas) and m_D is the number of areas assigned to group *D*. We present the RRMSE values as percentages.

$$RRMSE_D = \sqrt{\frac{1}{m_D}\sum_{i\in D}\left\{\frac{1}{1{,}000}\sum_{i=1}^{1{,}000}\left[\frac{\left(\hat{\sigma}_{v\,i}^2-\sigma_{v\,i}^2\right)}{\sigma_{v\,i}^2}\right]^2\right\}} \times 100\% \qquad (7)$$

We also considered using the Root Mean Square Error (RMSE) to evaluate the estimates produced by each method, but selected the RRMSE due to its comparability across specifications of $\sigma_{v\,i}^2$.

Table 4.2 RRMSE of Designed Based Methods Under the Assumption of Constant Synthetic Error Variance Within Group Assignments

Group	$\sigma_{v\,i}^2 / \sigma_{e\,i}^2$	m_D =20 Areas Per Group			m_D =50 Areas Per Group		
		Gonzalez-Waksberg/ Marker	Adjusted Gonzalez-Waksberg	Adjusted Marker	Gonzalez-Waksberg/ Marker	Adjusted Gonzalez-Waksberg	Adjusted Marker
1	0.08	463% (36.3%)	436% (0.8%)	439%	281% (35.9%)	344%	346%
2	0.12	309% (35.5%)	290% (0.4%)	292%	199% (31.2%)	226%	227%
3	0.16	245% (36.3%)	223% (0.6%)	224%	157% (24.9%)	176%	177%
4	0.20	196% (37.4%)	176% (1.0%)	176%	127% (21.9%)	134%	134%
5	0.24	174% (41.5%)	159% (0.8%)	160%	110% (17.2%)	114%	115%
All		296% (37.4%)	276% (0.8%)	277%	185% (26.2%)	215%	216%
1	0.8	73% (6.4%)	66%	66%	46% (0.2%)	43%	43%
2	1.2	59% (2.3%)	55%	55%	37%	35%	35%
3	1.6	52% (1.1%)	49%	49%	33%	32%	32%
4	2.0	45% (0.3%)	43%	43%	30%	29%	29%
5	2.4	45% (0.1%)	44%	44%	28%	28%	28%
All		56% (2.0%)	52%	52%	35% (0.0%)	34%	34%
1	4	40%	40%	40%	25%	25%	25%
2	6	37%	37%	37%	23%	23%	23%
3	8	34%	34%	34%	23%	23%	23%
4	10	33%	33%	33%	21%	21%	21%
5	12	35%	35%	35%	21%	21%	21%
All		36%	36%	36%	23%	23%	23%

Note: The numbers listed in parentheses are the non-zero percentages of areas that were estimated to have negative synthetic error variance.

Note that the Gonzalez-Waksberg and Marker estimators are equal for our simulation since all areas have the same synthetic estimate and thus the same sampling variance of the synthetic estimate.

The Lahiri-Pramanik adjustments improve the performance of the design-based estimators when the values of $\sigma_{v\,i}^2$ are small in comparison to σ_e^2 and $m_D = 20$. The adjustments do not improve the design-based estimators for small values of $\sigma_{v\,i}^2$ and $m_D = 50$ as the unadjusted estimators provide better estimates. The unadjusted estimators produce a number of negative estimates for these small values of $\sigma_{v\,i}^2$. These negative estimates may not be bad in an absolute sense, but they may be awkward since they are negative estimates for variance terms. The adjusted estimators generally avoid negative estimates.

For medium values of $\sigma_{v\,i}^2$, the adjustments have a small effect, as the RRMSE values are slightly better for the adjusted methods. For large values of $\sigma_{v\,i}^2$, the adjustments have no effect as the RRMSE values are equivalent for the adjusted and unadjusted methods. All four estimators avoid zero estimates for both medium and large values of $\sigma_{v\,i}^2$. These findings are consistent with the results presented in Lahiri and Pramanik (2012).

In general, the estimates are poor for small $\sigma_{v\,i}^2$, which is an indication that estimation is difficult in the presence of a large amount of noise from the sampling variance of y_i. On the other hand, an estimate of a small variance that is four times too large may still be small and not too inaccurate in an absolute sense. The results are better for larger values $\sigma_{v\,i}^2$ and larger number of areas assigned to groups ($m_D = 50$ versus $m_D = 20$). The results of the largest values of $\sigma_{v\,i}^2$ start to approach the theoretical limit of the RRMSE that would occur for the usual variance

estimates from data not obscured by the sampling error (i.e., if the v_i were observed). In this case, the variance estimates are proportional to a chi-square distribution with m_D degrees of freedom. The RRMSE is then the Coefficient of Variation (CV), which is $\sqrt{2/m_D}$ (32% for $m_D = 20$ and 20% for $m_D = 50$). Simulations done for even larger values of $\sigma^2_{v\,i}$ (not presented here) essentially produced these CVs.

For the remainder of this manuscript, the only design-based results that we present will be those generated from the adjusted Marker estimator.

3. A Modeling Approach

Fay and Herriot (1979) proposed the following small area estimation model, where v_i are area random effects.

$$y_i = Y_i + e_i \quad e_i \sim ind\ N\left(0, \sigma^2_{e_i}\right) \tag{7}$$

$$Y_i = \mu_i + v_i \quad v_i \sim ind\ N\left(0, \sigma^2_{v_i}\right) \tag{8}$$

The standard formulation of the Fay-Herriot model replaces μ_i with a linear regression function $X_i\beta$. Song (2007) proposed empirical Bayes estimators, where area specific estimates of the unknown parameters in this model formulation were obtained using the plug-in method. Song then used numerical integration to derive a posterior distribution for $\sigma^2_{v_i}$ and investigated a number of estimators based on this posterior distribution.

We assume that the synthetic estimate is an unbiased estimate of μ_i, $E[y_i^s] = \mu_i$, and thus:

$$y_i^s = \mu_i + \varepsilon_i \quad \varepsilon_i \sim ind\ N\left(0, \sigma_{\varepsilon_i}^2\right) \tag{9}$$

where $\sigma_{\varepsilon_i}^2$ represents the sampling variance of the synthetic estimate. This is a bivariate extension of the Fay-Herriot model from which we derive the following as the MSE of the synthetic estimate under the assumption of independence between the residual terms (ε_i and v_i).

$$y_i^s - Y_i = \varepsilon_i - v_i \ \rightarrow \ MSE(y_i^s) = \sigma_{\varepsilon_i}^2 + \sigma_{v_i}^2 \tag{10}$$

Comparing Equation 5 to Equation 10, we see that under this model formulation, the synthetic error variance, $\sigma_{v_i}^2 = Var[Y_i - \mu_i]$, is analogous to the squared synthetic bias ($Bias^2(y_i^s)$) used in Section 2 for the design-based methods.

Combining Equations 7, 8, and 9, we derive the following distribution of d_i, the difference between the synthetic estimate and the direct survey estimate, under the assumption of independence between the residual terms (ε_i, v_i, and e_i).

$$d_i = y_i^s - y_i = \varepsilon_i - v_i - e_i \sim N\left(0, \sigma_{\varepsilon_i}^2 + \sigma_{v_i}^2 + \sigma_{e_i}^2\right)$$

We estimated the values of the variance terms $\sigma_{\varepsilon_i}^2$ and $\sigma_{e_i}^2$ using standard design based methods and treat them as fixed quantities. We use the values of d_i to fit a parametric model for $\sigma_{v_i}^2$.

We considered two simple model formulations for $\sigma_{v_i}^2$. The first assumes that $\sigma_{v_i}^2$ is constant for areas i within groups D ($\sigma_{v_i}^2 = \sigma_{v_D}^2\ for\ i \in D$), which is a typical assumption for the design based approaches of estimating the MSE of synthetic estimators. The second model formulation

assumes that $\sigma_{v_i}^2$ depends on some covariate(s) Z_i via the function $\sigma_{v_i}^2 = (Z_i\gamma)^2$, analogous to a regression equation.

We extended these two model formulations for $\sigma_{v_i}^2$ by adding random effects via the following:

$$for\ i \in D, \quad \sigma_{v_i}^2 = {\sigma_v^2}/{\varphi_D} \quad or \quad \sigma_{v_i}^2 = {(Z_i\gamma)^2}/{\varphi_D}$$

$$where\ \varphi_D \sim Gamma(\delta, \delta)\ and\ \delta \sim Gamma(1,1)$$

This specification implies that the random effects are centered around 1, $E[\varphi_D] = 1$, with $Var[\varphi_D] = {1}/{\delta}$. Large values of δ suggest that there is little variation between groups, while small values suggest large variation between groups. When applied to estimate $\sigma_{v_i}^2$, this model also allows for Bayesian or empirical Bayes shrinkage estimation of $\sigma_{v_i}^2$ across groups of *D*. The model thus provides a compromise between the fixed effects model $\sigma_{v_i}^2 = \sigma_{v_D}^2$ and the simpler model $\sigma_{v_i}^2 = \sigma_v^2$ that assumes $\sigma_{v_i}^2$ is constant across all areas. Like other shrinkage estimators, more weight is placed on the fixed effects model when more data are available. Appendix C derives the shrinkage estimates of $\sigma_{v_i}^2$ across groups.

This model specification with random effects was inspired by a model that Otto and Bell (1995) formulated for estimated covariance matrices. The univariate version of this model was used by Maples (2011). Arora and Lahiri (1997), Gershunskaya and Lahiri (2005), and You and Chapman (2006) used similar models.

We examine both Maximum Likelihood (ML) and Hierarchical Bayes (HB) methods of estimation for the two model formulations for $\sigma_{v_i}^2$. The first method of estimation applies ML to the first model formulation in which $\sigma_{v_i}^2$ is constant for areas *i* within group *D* ($\sigma_{v_i}^2 = \sigma_{v_D}^2$ for $i \in D$). The Fixed Group ML estimate of $\sigma_{v_D}^2$ is the value corresponding to the mode of the likelihood of d_i.

$$\arg\max_{\sigma_{v_D}^2} \sum_{i \in D} \ln\left[N\left(d_i | 0, \sigma_{\varepsilon_i}^2 + \sigma_{v_D}^2 + \sigma_{e_i}^2\right)\right]$$

In some cases, Fixed Group ML produces zero estimates of $\sigma_{v_D}^2$, which may be awkward in situations where the researcher has a strong belief that synthetic estimation error is prevalent. Li and Lahiri (2010) proposed the following adjusted likelihood method:

$$\arg\max_{\sigma_{v_D}^2} \sum_{i \in D} \ln\left[h\left(\sigma_{v_D}^2\right) \times L\left(\sigma_{v_D}^2\right)\right]$$

where $L\left(\sigma_{v_D}^2\right)$ is the profile likelihood or residual (marginal) likelihood for $\sigma_{v_D}^2$ and $h\left(\sigma_{v_D}^2\right)$ is an adjustment factor designed to force the estimator to be strictly positive. Li and Lahiri (2010) considered the adjustment factor $h\left(\sigma_{v_D}^2\right) = \sigma_{v_D}^2$. Yoshimori and Lahiri (2013) proposed the adjustment factor $\left(tan^{-1}\left\{tr\left[I - B\left(\sigma_{v_D}^2\right)\right]\right\}\right)^{\frac{1}{m_D}}$, where $B\left(\sigma_{v_D}^2\right)$ is an $m_D \times m_D$ diagonal matrix whose entries are the "shrinkage factors" in the empirical Bayes predictor, $B_i\left(\sigma_{v_D}^2\right) = \sigma_{e_i}^2 / \left(\sigma_{v_D}^2 + \sigma_{e_i}^2\right)$. Yoshimori and Lahiri used simulations to compare the performance of their

proposed adjusted likelihood estimator to other variance estimators, including the Li-Lahiri adjusted likelihood, ML, and residual maximum likelihood (REML).

These adjusted likelihood approaches can be viewed loosely as Bayesian where $h\left(\sigma_{v_D}^2\right)$ is the prior, if one takes the posterior mode as the estimate of $\sigma_{v_D}^2$. We investigated HB estimation methods of this first model formulation, for which a flat prior ($h\left(\sigma_{v_D}^2\right) \propto 1$) and the prior proposed in Gelman (1995) ($h\left(\sigma_{v_D}^2\right) \propto \left(\sigma_{v_D}^2\right)^{\frac{-1}{2}}$) were considered in addition to the prior corresponding to the Li-Lahiri adjusted likelihood. Using the Metropolis-Hastings MCMC algorithm (Metropolis et al. 1953), we derived posterior distributions for $\sigma_{v_D}^2$. For this research, we estimated $\sigma_{v_D}^2$ using the posterior means. While we produced results using the prior proposed by Li and Lahiri (2010), we did not include these results here since they underperformed the other estimation methods in all simulations due to significant overestimation of the true values. We did not consider a prior corresponding to the adjusted likelihood proposed by Yoshimori and Lahiri (2013) since we became aware of their research after our simulations were essentially complete.

HB estimation was also applied for the second model formulation, where $\sigma_{v_i}^2$ depends on some covariate(s) Z_i. We specified the prior distributions of the regression parameters γ to be uninformative, as shown below.

$$d_i \sim N\left(0, \sigma_{\varepsilon_i}^2 + (Z_i\gamma)^2 + \sigma_{e_i}^2\right)$$

$$\gamma \sim Normal\left(\underline{0}, 100I\right)$$

The matrix Z may contain categorical (analogous to group assignments) or continuous covariates. Estimates of $\sigma_{v_i}^2$ were generated for each area from the posterior distribution $p((Z_i\gamma)^2|d_i)$ using either the posterior mean or mode. These estimates will be referred to as Fixed Covariate estimates.

HB estimation was applied to the two model formulations augmented by random effects. The first model formulation with random effects, referred to as the Random Effects model, was specified as follows:

$$d_i \sim N\left(0, \sigma_{\varepsilon_i}^2 + {\sigma_v^2}/{\varphi_D} + \sigma_{e_i}^2\right)$$

$$\sigma_v \sim Normal(0,10)$$

$$\phi_D \sim Gamma(shape = \delta, rate = \delta)$$

$$\delta \sim Gamma(1,1)$$

Estimates of $\sigma_{v_i}^2$ were generated for each group from the posterior distribution $p\left({\sigma_v^2}/{\varphi_D}\,|d_i\right)$ using either the posterior mean or posterior mode.

The second model formulation with random effects, referred to as the Fixed Covariate with Random Effects model, was specified as follows.

$$d_i \sim N\left(0, \sigma_{\varepsilon_i}^2 + {(Z_i\gamma)^2}/{\varphi_D} + \sigma_{e_i}^2\right)$$

$$\gamma \sim Normal(\underline{0}, 100I)$$

$$\phi_D \sim Gamma(shape = \delta, rate = \delta)$$

$$\delta \sim Gamma(1,1)$$

Estimates of $\sigma_{v_i}^2$ were generated for each area from the posterior distribution $p\left({(Z_i\gamma)^2}/{\varphi_D} \,|d_i\right)$ using either the posterior mean or posterior mode. This Fixed Covariate with Random Effects model is only applied in Section 4.

The following three sections provide the results of the application of the estimation methods described above to simulation data. We also provide the results from the application of the design-based adjusted Marker method to the simulation data, treating it as estimating the $\sigma_{v_i}^2$ (synthetic error variances). Section 3.1 applies the methods to data simulated under the assumption of constant variance within groups. Sections 3.2 and 3.3 present results under the assumption that $\sigma_{v_i}^2$ varies across areas as a function of some covariate. In Section 3.2, we assume that this covariate is observed and can be used for synthetic error variance estimation. In Section 3.3, we assume that this covariate is unknown, and we use an incorrect variable for synthetic error variance estimation.

Table 4.3 provides a summary of the model-based methods, including the parametric function specified for the synthetic error variance and the estimation of the parameters of the function using the data.

Table 4.3: Summary Model-Based Methods

Reference	Parametric Function	Estimation of Parameters
Fixed Group Effect – Maximum Likelihood	$d_i \sim N\left(0, \sigma_{\varepsilon_i}^2 + \sigma_{v_D}^2 + \sigma_{e_i}^2\right)$	$\arg\max_{\sigma_{v_D}^2} \sum_{i \in D} \ln\left[N\left(d_i \mid 0, \sigma_{\varepsilon_i}^2 + \sigma_{v_D}^2 + \sigma_{e_i}^2\right)\right]$
Fixed Group Effect – Hierarchical Bayes	$d_i \sim N\left(0, \sigma_{\varepsilon_i}^2 + \sigma_{v_D}^2 + \sigma_{e_i}^2\right)$ Prior Distributions Flat Prior: $h\left(\sigma_{v_D}^2\right) \propto 1$ Gelman Prior: $h\left(\sigma_{v_D}^2\right) \propto \left(\sigma_{v_D}^2\right)^{\frac{-1}{2}}$	$E\left[\sigma_{v_D}^2 \mid d_i \; for \; i \in D\right]$
Fixed Covariate – Hierarchical Bayes	$d_i \sim N\left(0, \sigma_{\varepsilon_i}^2 + (Z_i\gamma)^2 + \sigma_{e_i}^2\right)$ Prior Distribution: $\gamma \sim Normal\left(\underline{0}, 100I\right)$	Posterior Mean: $E\left[\sigma_{v_i}^2 \mid \underline{d}\right]$ Posterior Mode: $\arg\max_{(Z_i\gamma)^2} \; p\left((Z_i\gamma)^2 \mid \underline{d}\right)$
Random Effects – Hierarchical Bayes	$d_i \sim N\left(0, \sigma_{\varepsilon_i}^2 + {\sigma_v^2}/{\varphi_D} + \sigma_{e_i}^2\right)$ Prior Distributions: $\sigma_v \sim Normal(0,10)$ $\phi_D \sim Gamma(shape = \delta, rate = \delta)$ $\delta \sim Gamma(1,1)$	Posterior Mean: $E\left[{\sigma_v^2}/{\varphi_D} \mid \underline{d}\right]$ Posterior Mode: $\arg\max_{\sigma_v^2/\varphi_D} \; p\left({\sigma_v^2}/{\varphi_D} \mid \underline{d}\right)$
Fixed Covariate with Random Effects – Hierarchical Bayes	$d_i \sim N\left(0, \sigma_{\varepsilon_i}^2 + {(Z_i\gamma)^2}/{\varphi_D} + \sigma_{e_i}^2\right)$ Prior Distributions: $\gamma \sim Normal\left(\underline{0}, 100I\right)$ $\phi_D \sim Gamma(shape = \delta, rate = \delta)$ $\delta \sim Gamma(1,1)$	Posterior Mean: $E\left[{(Z_i\gamma)^2}/{\varphi_D} \mid \underline{d}\right]$ Posterior Mode: $\arg\max_{(Z_i\gamma)^2} \; p\left({(Z_i\gamma)^2}/{\varphi_D} \mid \underline{d}\right)$

3.1 Simulations with Constant Synthetic Error Variance (Over Groups of Areas)

Tables 4.4 and 4.5 provide the RRMSE of the estimated synthetic error variance from each model-based method to the true value specified in the simulation. The simulation specifies that the synthetic error variance is constant within the five assigned groups, for varying ratios of the synthetic error variance to the sampling variance. Tables 4.4 and 4.5 provide the results for m_D =20 and m_D =50, respectively.

Table 4.4: RRMSE Under the Assumption of Constant Synthetic Error Variance Within Group Assignments – $\boldsymbol{m_D = 20}$ Areas Per Group

Group	$\sigma_{v_i}^2 / \sigma_{e_i}^2$	Adjusted Marker	Fixed Group ML	Fixed Group HB		Fixed Covariate		Random Effects	
				Flat Prior	Gelman Prior	Posterior Mode	Posterior Mean	Posterior Mode	Posterior Mean
1	0.08	439%	554% (11.0%)	914%	722%	453%	722%	348%	534%
2	0.12	292%	370% (12.1%)	616%	486%	305%	485%	218%	342%
3	0.16	224%	285% (11.7%)	470%	373%	241%	374%	164%	256%
4	0.20	176%	225% (13.2%)	375%	297%	190%	297%	126%	197%
5	0.24	160%	199% (12.9%)	329%	263%	171%	263%	113%	171%
All		277%	351% (12.2%)	580%	459%	291%	459%	212%	328%
1	0.8	66%	75% (1.3%)	118%	99%	73%	99%	56%	84%
2	1.2	55%	59% (0.4%)	94%	80%	57%	80%	42%	58%
3	1.6	49%	52% (0.4%)	79%	68%	51%	68%	39%	46%
4	2.0	43%	45% (0.1%)	69%	59%	44%	59%	37%	40%
5	2.4	44%	45%	69%	60%	44%	60%	39%	41%
All		52%	56% (0.4%)	88%	75%	55%	75%	43%	56%
1	4	40%	40%	59%	52%	39%	52%	34%	48%
2	6	37%	37%	55%	48%	36%	48%	30%	37%
3	8	34%	34%	49%	43%	34%	43%	30%	32%
4	10	33%	33%	49%	43%	33%	43%	30%	31%
5	12	35%	35%	50%	44%	34%	44%	32%	32%
All		36%	36%	53%	46%	35%	46%	31%	37%

Note: The numbers listed in parentheses are the non-zero percentages of areas that were estimated to have zero synthetic error variance. Zero estimates were defined to be less than 0.00001, 0.0001, and 0.0005 for the first second and third simulations.

Table 4.5: RRMSE Under the Assumption of Constant Synthetic Error Variance Within Group Assignments – $\boldsymbol{m_D = 50}$ Areas Per Group

Group	$\sigma_{v_i}^2 / \sigma_{e_i}^2$	Adjusted Marker	Fixed Group ML	Fixed Group HB		Fixed Covariate		Random Effects	
				Flat Prior	Gelman Prior	Posterior Mode	Posterior Mean	Posterior Mode	Posterior Mean
1	0.08	346%	474% (4.1%)	597%	529%	427%	529%	383%	476%
2	0.12	227%	317% (2.4%)	400%	356%	286%	355%	247%	310%
3	0.16	177%	247% (1.1%)	311%	279%	226%	279%	187%	235%
4	0.20	134%	189% (1.3%)	240%	214%	173%	214%	139%	176%
5	0.24	115%	157% (1.3%)	200%	179%	145%	179%	113%	144%
All		216%	299% (2.0%)	377%	335%	270%	335%	234%	293%
1	0.8	43%	53%	69%	63%	51%	63%	49%	62%
2	1.2	35%	40%	51%	46%	38%	46%	33%	40%
3	1.6	32%	35%	44%	41%	34%	41%	29%	33%
4	2.0	29%	30%	37%	35%	30%	35%	26%	28%
5	2.4	28%	29%	36%	33%	29%	33%	26%	27%
All		34%	38%	49%	45%	37%	45%	34%	40%
1	4	25%	25%	30%	28%	25%	28%	24%	29%
2	6	23%	23%	28%	26%	23%	26%	21%	24%
3	8	23%	23%	27%	26%	23%	26%	21%	22%
4	10	21%	21%	25%	24%	21%	24%	20%	21%
5	12	21%	21%	25%	24%	21%	24%	20%	21%
All		23%	23%	27%	25%	23%	25%	21%	23%

Note: The numbers listed in parentheses are the non-zero percentages of areas that were estimated to have zero synthetic error variance. Zero estimates were defined to be less than 0.00001, 0.0001, and 0.0005 for the first second and third simulations.

The posterior mode estimates from the Random Effects model produce the best results overall. This indicates that shrinkage to the synthetic error variance estimate over all areas improves estimation. This is true even in comparison to the Fixed Covariate posterior mode and Fixed Group ML estimates, which use the correct model. This is reminiscent of Stein's result on shrinkage estimates of means (Stein 1956). The improvement in RRMSE (particularly the two simulations with larger synthetic error variance) can be attributed to the smaller variance of the estimator. Although the posterior mode estimates from the Random Effects model generally provided the best results overall, the differences with estimates from the other methods are small for large values of $\sigma_{v_i}^2$ and larger number of areas assigned to groups.

The posterior mean estimators of the Fixed Covariate model and Random Effects model do poorly compared to the posterior modes. In general, the posterior mean estimates from Fixed Group HB methods, Fixed Covariate model, and Random Effects model tend to overestimate the true synthetic error variance, even for large values, and thus do worse than the mode estimates. This is attributed to the right-skew of the synthetic error variance posterior distributions, which was more severe as the true synthetic error variance decreased. Note that the Fixed Group HB method with the Gelman prior and the mean estimator of the Fixed Covariate model provide very similar results. Also note that while the mode estimator of the Fixed Covariate model and the Fixed Group ML estimator provide similar results for larger values of synthetic error variance, the mode estimator performs better for smaller values. For these reasons, we will avoid presenting the Fixed Group HB methods, mean estimates from the Fixed Covariate model and Random Effects model, and the Fixed Group ML estimates in subsequent tables.

While the mode estimator of the Random Effects model generally produces the best results, the adjusted Marker and mode of the Fixed Covariate model provide similar results except for small values of $\sigma_{v_i}^2$, where the adjusted Marker estimator provides somewhat better results.

As noted earlier, Yoshimori and Lahiri (2013) provides results from a simulation study comparing the performance of various estimators of σ_v^2. Much of their focus involves the estimation of the empirical Bayes shrinkage factors $B_i(\sigma_v^2)$ and the empirical Bayes predictors, two topics not considered here. Yoshimori and Lahiri also provide results on the estimation of σ_v^2, noting that, for small values of $\sigma_v^2/\sigma_{e_i}^2$, all methods that they considered substantially overestimated σ_v^2. This finding is consistent with our results, with the exception of the Gonzalez-Waksberg and Marker methods, the two methods that allow for negative estimates.

We cannot make specific comparisons between the results of Yoshimori and Lahiri and our results, since the variance estimation methods they consider mostly differ from ours. As noted earlier, we did not evaluate the new methods proposed by Yoshimori and Lahiri in our study since we became aware of these after our simulations were essentially complete. The closest comparison that we can make between our results and those of Yoshimori and Lahiri involves the Fixed Group ML estimates of σ_v^2. For the Fixed Group ML, we find far fewer zero estimates and larger biases towards overestimation for roughly comparable situations. These roughly comparable situations differ some in regard to the values of m_D(we use 20 and 50, they use 15 and 45) and of $\sigma_v^2/\sigma_{e_i}^2$ (we have 0.08 and 0.12 versus their 0.05 and 0.10, though we both have 10), though the effect of these differences is not large. More substantial differences in the Fixed Group ML results come from the fact that we estimate the sampling variances $\sigma_{e_i}^2$, whereas they

take the true value of $\sigma_{e_i}^2$ as known, and they estimate a mean for the Fay-Herriot model, while we have $E[d_i] = 0$ by assumption. When we reapplied the Fixed Group ML method to our simulation data in the same manner as Yoshimori and Lahiri, we found that the results were consistent with what they reported.

3.2 Simulations with Non-constant Synthetic Error Variance with Correct Fixed Covariate Model Specification

This simulation specifies that the synthetic error variances $\sigma_{v\,i}^2$ are unique for each of 100 areas. We assume that we are able to correctly model the synthetic error variance with the covariate $Z_i = \sqrt{\sigma_{v\,i}^2}$. The estimators resulting from the adjusted Marker and Random Effects model are biased as a result of the failure of the constant bias within groups assumption. The estimators from the Fixed Covariate model are based on the correctly specified model.

The methods that group areas together assigned 100 areas to either two or five groups, with each group containing fifty or twenty areas. Areas were assigned to groups based on the values of the synthetic error variance. For example, in the case where areas are assigned to five groups, the twenty areas assigned to the first group were those with the twenty smallest values of synthetic error variance ($\sigma_{v\,i}^2$). The twenty areas assigned to the second were the twenty areas with the next smallest values of synthetic error variance ($\sigma_{v\,i}^2$), and so on. The group assignments for the case where areas were assigned to two groups were made in a similar way. The Fixed Covariate method used the covariate Z_i in the model.

Table 4.6 provides the RRMSE between the estimated synthetic error variance and the true synthetic error variance within each group and over all areas. Table D.3 in Appendix D provides

boxplots for each area of the errors of each estimate (difference between the estimated synthetic error variance and the true synthetic error variance).

Table 4.6: RRMSE Under the Assumption of Non-constant Synthetic Error Variance With Correct Model Specification

$\sigma_{v\,i}^2 / \sigma_{e\,i}^2$	Group	5 Groups			2 Groups		
		Adjusted Marker	Fixed Covariate Posterior Mode	Random Effects Posterior Mode	Adjusted Marker	Fixed Covariate Posterior Mode	Random Effects Posterior Mode
$0.0816\ to\ 0.2400$ $in\ steps\ of\ 0.0016$	1	370%	436%	290%	242%	324%	275%
	2	275%	234%	203%	141%	182%	156%
	3	225%	180%	165%			
	4	184%	169%	131%			
	5	170%	198%	119%			
	All	255%	263%	192%	198%	263%	223%
$0.816\ to\ 2.4$ $in\ steps\ of\ 0.016$	1	62%	59%	49%	43%	46%	44%
	2	55%	37%	41%	32%	32%	30%
	3	49%	28%	39%			
	4	44%	28%	36%			
	5	45%	37%	38%			
	All	52%	39%	41%	38%	39%	38%
$4.08\ to\ 12$ $in\ steps\ of\ 0.08$	1	41%	33%	34%	32%	25%	31%
	2	38%	20%	31%	25%	21%	24%
	3	35%	17%	30%			
	4	34%	20%	30%			
	5	35%	25%	31%			
	All	36%	23%	31%	29%	23%	28%

The posterior mode from the Random Effects model provides the best results for small values of synthetic error variance and $m_D = 20$. The adjusted Marker method provides the best results for these small values when $m_D = 50$. While the Fixed Covariate model performs best for medium values and $m_D = 20$, all three methods produce similar results for medium values and $m_D = 50$. The posterior mode from the Fixed Covariate model does best for the simulation with the largest values of synthetic error variance.

Note that the correct model is specified for Fixed Covariate model, but the Random Effects and adjusted Marker methods used an incorrect model. The incorrect models used groupings of areas

based on the continuous covariate rather than the continuous covariate itself. Since the groups have the correct order of variances (smallest to largest), the misspecification of the model is limited.

3.3 Simulations with Non-constant Synthetic Error Variance with Incorrect Fixed Covariate Model Specification

In this section, we use the same simulated data that was used in Section 3.2, but we assume that we do not observe the covariate Z_i. Instead, we assume we have a different covariate that is a uniform random number ($Z_i = Uniform(0,1)$) drawn for each area. The estimators resulting from all methods are biased due to using incorrect covariate(s). The incorrect covariate is Z_i for the Fixed Covariate model, and they are used to assign the group indicators for the adjusted Marker and Random Effects model.

The methods that group areas together assigned the 100 areas to either two or five groups based on the covariate Z_i, with each group containing fifty or twenty areas. For example, in the case where areas are assigned to five groups, the twenty areas assigned to the first group were those with the twenty smallest values of Z_i. The twenty areas assigned to the second group were the twenty areas with the next smallest values of Z_i, and so on. The group assignments for the case where areas were assigned to two groups were made in a similar way. The Fixed Covariate method used the Z_i as the lone covariate in the model.

Table 4.7 provides the RRMSE within each group and over all areas. Table D.4 in Appendix D provides the boxplots for each area of the errors of each estimate (difference between the estimated synthetic error variance and the true synthetic error variance).

Table 4.7: RRMSE Under the Assumption of Non-constant Synthetic Error Variance With Incorrect Model Specification

$\sigma_{v_i}^2 / \sigma_{e_i}^2$	Group	5 Groups			2 Groups		
		Adjusted Marker	Fixed Covariate Posterior Mode	Random Effects Posterior Mode	Adjusted Marker	Fixed Covariate Posterior Mode	Random Effects Posterior Mode
$0.0816\ to\ 0.2400$ $in\ steps\ of\ 0.0016$	1	421%	416%	314%	272%	339%	299%
	2	295%	296%	218%	126%	170%	143%
	3	220%	219%	161%			
	4	175%	175%	128%			
	5	145%	144%	106%			
	All	270%	268%	200%	212%	268%	235%
$0.816\ to\ 2.4$ $in\ steps\ of\ 0.016$	1	138%	127%	107%	111%	123%	111%
	2	131%	121%	101%	96%	106%	95%
	3	124%	115%	96%			
	4	120%	107%	91%			
	5	115%	102%	87%			
	All	126%	115%	97%	104%	115%	103%
$4.08\ to\ 12$ $in\ steps\ of\ 0.08$	1	88%	76%	70%	55%	55%	51%
	2	51%	38%	38%	29%	29%	30%
	3	37%	24%	30%			
	4	35%	27%	34%			
	5	39%	34%	40%			
	All	54%	44%	45%	44%	44%	42%

The posterior mode of the Random Effects model generally provides the best estimates for small values $\sigma_{v_i}^2$ and $m_D = 20$, while the adjusted Marker method provides slightly better estimates for these small values and $m_D = 50$. For medium values of $\sigma_{v_i}^2$, the Random Effects model performed best for $m_D = 20$, while the adjusted Marker and Random Effects methods performed best for $m_D = 50$. For large values of $\sigma_{v_i}^2$, the Fixed Covariate and Random Effects models performed similarly, slightly outperforming the adjusted Marker method for $m_D = 20$ and performing similarly for $m_D = 50$. The shrinkage estimator helps compensate for incorrect model specification when the group sizes are smaller.

The posterior mode estimates of the Fixed Covariate model perform slightly better than the adjusted Marker estimates for large and medium $\sigma_{v_i}^2$ and $m_D = 20$, essentially the same for large

$\sigma_{v\,i}^2$ and $m_D = 50$ or small $\sigma_{v\,i}^2$ and $m_D = 20$, and slightly worse for small and medium $\sigma_{v\,i}^2$ and $m_D = 50$. The adjusted Marker method was used such that five parameters were estimated, one for each group. We assigned areas to these five groups based on a covariate that was unrelated to the true synthetic variance ($\sigma_{v\,i}^2$), and thus the five parameters were not truly different. The Fixed Covariate method was used such that two parameters were estimated, one for the overall mean and one for the covariate that is unrelated to the true synthetic variance ($\sigma_{v\,i}^2$). Both of these two methods use the incorrect model, but the adjusted Marker method estimates more unnecessary parameters than the Fixed Covariate model.

4. Synthetic Estimation of Correct Enumerations (CEs) in the 2010 U.S. Census Coverage Measurement (CCM)

After the 2010 Census, the Census Bureau conducted an evaluation of the census coverage of persons and housing units called Census Coverage Measurement (CCM), with the intention of providing information for the improvement of future censuses. As part of this evaluation, the Census Bureau used dual system estimation (U.S. Census Bureau 2004) to produce estimates of the true population of persons and number of housing units. As reviewed in Seber (1982), this estimation of the number of persons in a closed area is a capture-recapture method widely used by biologists and ecologists to estimate biological populations.

Dual system estimation requires two independent enumerations to generate estimates of the true population. The first enumeration was the Census itself. The second enumeration was the Population Sample (P-sample), consisting of all persons and housing units enumerated during an independent post enumeration survey conducted in a probability sample of block clusters a short

time after the census. Using clerical and computational methods, a matching operation compared information (age, sex, race, etc.) from the census and the post enumeration survey to match records between the P-sample and the census for the sample block clusters. The Census Bureau effectively estimated the proportion of P-sample persons also found in the census (match proportion).

The census records within the sample block clusters (called the E-sample or enumeration sample) were examined to determine which of these really should have been enumerated at their location in the census (correct enumerations or CEs) and which should not (erroneous enumerations or EEs). For the CCM, a person is correctly enumerated in the census if the person was correctly counted in the block cluster containing his/her Census Day residence, or in a block cluster adjacent to the block cluster containing his/her Census Day residence (Mule 2010). The census counts for areas could then be deflated by the estimated CE proportions to estimate the number of persons who were CEs. The estimated numbers of CEs were then inflated by dividing them by the match proportions to estimate the true populations. This is a form of Dual System Estimate (DSE) and is approximately unbiased assuming independence between the census and the P-sample.

In reality, the CE and match proportions were estimated by fitting logistic regression models to the E-sample and P-sample data. Covariates in these models were data collected during the census, which included age groups, race groups, sex, tenure (owner versus renter), etc. See Mule and Griffin (2010) and Olson and Griffin (2012) for more information. The fitted logistic regression was then applied to all census records to estimate a match probability and CE

probability for each record. These probabilities were aggregated to form synthetic estimates of CEs and population, via a form of DSE.

During the 2010 CCM, the Census Bureau produced synthetic estimates of CEs and populations within geographic areas of the United States, such as states, counties, and metropolitan regions. The MSE estimates of the synthetic population estimates reported in the 2010 Census Coverage Measurement Estimation Report (Mule 2012, Davis and Mulligan 2012, and Keller and Fox 2012) incorporated estimates of both the variance and bias components. We'll focus instead on estimating the MSEs of the synthetic CE estimates by applying the modeling assumptions discussed in Section 3 to the CE proportion estimates. Section 4.1 provides results of applying MSE estimation methods to simulations motivated by CE data from the 2010 CCM. Section 4.2 provides the results of the application of the MSE estimation methods to the actual CE data from the 2010 CCM.

4.1 Simulations Motivated by Synthetic Estimation of CEs in the 2010 CCM

We created simulated data sets using CE data from the state and county person data sets from the 2010 CCM. The simulated state datasets have 51 records (including the District of Columbia), while the simulated county data sets have 245 records. The 245 counties in the county data sets were those counties with census populations greater than 100,000. Each simulated data set contains the state (or county), census division (nine collections of states based on geographic proximity), census count (C_i), synthetic CE proportion estimate (ce_i^s), direct sample CE proportion variance estimate $\left(\sigma_{e_i}^2\right)$, and sampling variance of the synthetic CE proportion estimate $(\sigma_{\varepsilon_i}^2)$ from the 2010 CCM files. We set the value of the synthetic error variance $(\sigma_{v_i}^2)$ for each

observation to be proportional to the sum of the direct sample variance estimate and the sampling variance of the synthetic estimate. This proportion was set to one of three different values: 0.02, 0.1, and 1.0. Since the direct sample variance estimate and sampling variances of the synthetic estimate are distinct for each area, the synthetic error variance is distinct for each area and constant across simulations.

$$\sigma^2_{v_{i,\alpha}} = \alpha * \left(\sigma^2_{\varepsilon_i} + \sigma^2_{e_i}\right) for\ \alpha = 0.02, 0.1, and\ 1.0$$

Varying the proportion α for the state and county data sets resulted in 6 simulated data sets. For each simulated data set, the difference between the direct survey estimate and synthetic estimate was a random draw based on the values of the direct sample variance estimate, sampling variance of the synthetic estimate, and the synthetic error variance.

$$d_{i,\alpha} \sim Normal\left(0, \sigma^2_{v_{i,\alpha}} + \sigma^2_{\varepsilon_i} + \sigma^2_{e_i}\right) for\ \alpha = 0.02, 0.1, and\ 1.0$$

We replicated each of the 6 simulated data sets 1,000 times. Among replicates, all variable values for corresponding observations are equal, with the exception of the random draw of the difference between the direct survey estimate and synthetic estimate.

Many of the synthetic MSE estimation methods required that states and counties be assigned to groups of similar states and counties. For the state file, the 51 states were partitioned into three groups, each with 17 states, based on the census count of persons within the state. The first group contained the 17 states with the smallest census count of persons and the third group

contained the 17 states with the largest census count of persons. For the county file, the 245 counties were partitioned into five groups, each with 49 counties. The assignment of groups for counties was based on the census count of persons, similar to the assignment of states.

The following tables provide the results of applying the synthetic error variance estimation methods to the 2010 CCM state simulated data and county simulated data. Tables 4.8 and 4.9 provide the RRMSE between the estimated synthetic error variance and the true synthetic error variance within each group and over all areas for the state and county data sets, respectively. Tables D.5 and D.6 in Appendix D provide the distribution of the error of the estimates (estimate minus true value) generated by each MSE estimation method for the state and county data sets.

In addition to the methods studied in previous sections, here we also applied the Fixed Covariate with Random Effects model to the 2010 CCM data. The covariate matrix in this model formulation only contained the natural log of the census count as a continuous covariate, the same covariate that was used in the Fixed Covariate model. The random effects corresponded to the census division in which the state or county is found.

Table 4.8: RRMSE for 2010 CCM State Simulation Data

$\sigma_{v\,i}^2$	Group	Adjusted Marker	Fixed Covariate Posterior Mode	Random Effects Posterior Mode	Fixed Covariate with Random Effects Posterior Mode
$0.02 * (\sigma_{e\,i}^2 + \sigma_{\varepsilon\,i}^2)$	1	3,511%	515%	124%	285%
	2	3,317%	237%	183%	136%
	3	3,370%	881%	438%	843%
	All	3,400%	605%	283%	520%
$0.2 * (\sigma_{e\,i}^2 + \sigma_{\varepsilon\,i}^2)$	1	422%	120%	90%	89%
	2	402%	95%	91%	88%
	3	410%	144%	111%	125%
	All	412%	121%	98%	102%
$1.0 * (\sigma_{e\,i}^2 + \sigma_{\varepsilon\,i}^2)$	1	160%	93%	74%	84%
	2	158%	70%	75%	78%
	3	162%	101%	89%	92%
	All	160%	89%	80%	85%

The proposed methods outperform the adjusted Marker method in all three state simulations. The improvement of the proposed methods over the adjusted Marker method is largest for the smallest specification of $\sigma_{v\,i}^2$, and the improvement decreases as the value of $\sigma_{v\,i}^2$ increases.

Of the proposed methods, the Random Effects model generally performs best overall specifications of $\sigma_{v\,i}^2$, particularly for small values of $\sigma_{v\,i}^2$. For the medium and large values of $\sigma_{v\,i}^2$, the Random Effects and Fixed Covariate with Random Effects models perform similarly and generally better than the Fixed Covariate model. Shrinkage seems to be helpful for almost all states for these values of $\sigma_{v\,i}^2$.

Table 4.9: RRMSE for 2010 CCM County Simulation Data

$\sigma_{v_i}^2$	Group	Adjusted Marker	Fixed Covariate Posterior Mode	Random Effects Posterior Mode	Fixed Covariate with Random Effects Posterior Mode
$0.02 * (\sigma_{e_i}^2 + \sigma_{\varepsilon_i}^2)$	1	6,442%	278%	95%	161%
	2	8,896%	341%	132%	148%
	3	16,149%	279%	148%	147%
	4	4,686%	107%	117%	89%
	5	6,367%	789%	165%	484%
	All	9,423%	426%	134%	250%
$0.2 * (\sigma_{e_i}^2 + \sigma_{\varepsilon_i}^2)$	1	857%	112%	98%	99%
	2	1,197%	132%	102%	98%
	3	2,170%	130%	102%	106%
	4	605%	91%	95%	94%
	5	843%	199%	102%	142%
	All	1,261%	138%	100%	109%
$1.0 * (\sigma_{e_i}^2 + \sigma_{\varepsilon_i}^2)$	1	394%	117%	110%	111%
	2	555%	191%	137%	150%
	3	996%	209%	160%	178%
	4	267%	70%	83%	79%
	5	385%	169%	115%	131%
	All	579%	160%	124%	134%

Like the state simulations, the proposed methods outperform the adjusted Marker method in all three county simulations. The improvement of the proposed methods over the adjusted Marker method is largest for the smallest specification of $\sigma_{v_i}^2$, and the improvement decreases as the specification of $\sigma_{v_i}^2$ increases. Unlike the state simulations, the proposed methods perform best for the medium specification of $\sigma_{v_i}^2$. The Random Effects and Fixed Covariate with Random Effects models perform worse in the largest specifications of $\sigma_{v_i}^2$ than the medium, which is in contrast to the other simulations previously presented in this paper. We speculate that this anomaly is attributable to large amount of variation found in the direct sample CE proportion estimates for the counties. More research is required to confirm this speculation.

Of the proposed methods, the Random Effects model generally performs best, although the performance within groups differs. Like the state simulations, the Random Effects and Fixed

Covariate with Random Effects models generally outperform the Fixed Covariate model, indicating that shrinkage is helpful regardless of the size of the counties. The Random Effects and Fixed Covariate with Random Effects models performed similarly, with the exception of the largest group of states, in which the Random Effects model performed best. We suspect that the that the poorer fit of the Fixed Covariate with Random Effects model when modeling the largest groups of states may be due to the skewed county distribution of the census count variable. We used the natural log transformed census count for modeling, but the largest states may still show up as outliers that were difficult to fit with a linear model.

4.2 Application of Various Methods to Correct Enumerations from the 2010 CCM

Table 4.10 provides average estimates from the synthetic error variance estimation methods when applied to actual data from the 2010 CCM. Note that for states, the Random Effects model tends to produce the lowest estimates of $\sigma_{v}^{2}{}_{i}$ and the adjusted Marker method tends to produce the highest. For counties the adjusted Marker estimates are the lowest in most cases, while the Fixed Covariate estimates tend to be the highest.

Table 4.10: Summarized Average Estimated Synthetic Error Variance

File	Group	Average Sampling Variance of the Synthetic Estimate	Average Direct Sample Variance Estimate	Adjusted Marker's	Fixed Covariate Posterior Mode	Random Effects Posterior Mode	Fixed Covariate with Random Effects Posterior Mode
		All Values x 10^{-4}					
State Person	1	0.169	1.393	0.130	0.052	0.011	0.022
	2	0.069	1.071	0.179	0.021	0.019	0.016
	3	0.061	0.451	0.085	0.013	0.009	0.024
	All	0.099	0.972	0.131	0.029	0.013	0.020
County Person	1	0.343	7.240	2.572	3.692	2.378	3.143
	2	0.195	4.763	2.082	3.123	3.441	2.729
	3	0.111	7.134	0.916	2.576	2.687	2.397
	4	0.301	5.048	1.537	2.242	1.860	1.934
	5	0.165	3.754	1.280	1.654	1.505	1.390
	All	0.223	5.588	1.677	2.657	2.374	2.318

The estimated values $\sigma_{v}^{2}{}_{i}$ are much higher for counties, suggesting that the aggregation to states washes out a considerable amount of variation. This is consistent with the intuitive thinking that counties have stronger area-specific effects that are not captured by the logistic regression model used to estimate the CE proportion. The true squared synthetic bias or synthetic error variance is unknown, thus no evaluation statistic can be calculated to compare the estimates from the various methods to the true values. We can compare the 2010 CCM data to the simulated values to determine which simulation the CCM most closely resembles. Figure 4.1 provides this comparison by providing boxplots of the method of moments estimate ($\hat{\sigma}_{v_i}^2 = (y_i^s - y_i)^2 - \sigma_{e_i}^2 - \sigma_{\varepsilon_i}^2$) of the squared synthetic bias or synthetic error variance for each area.

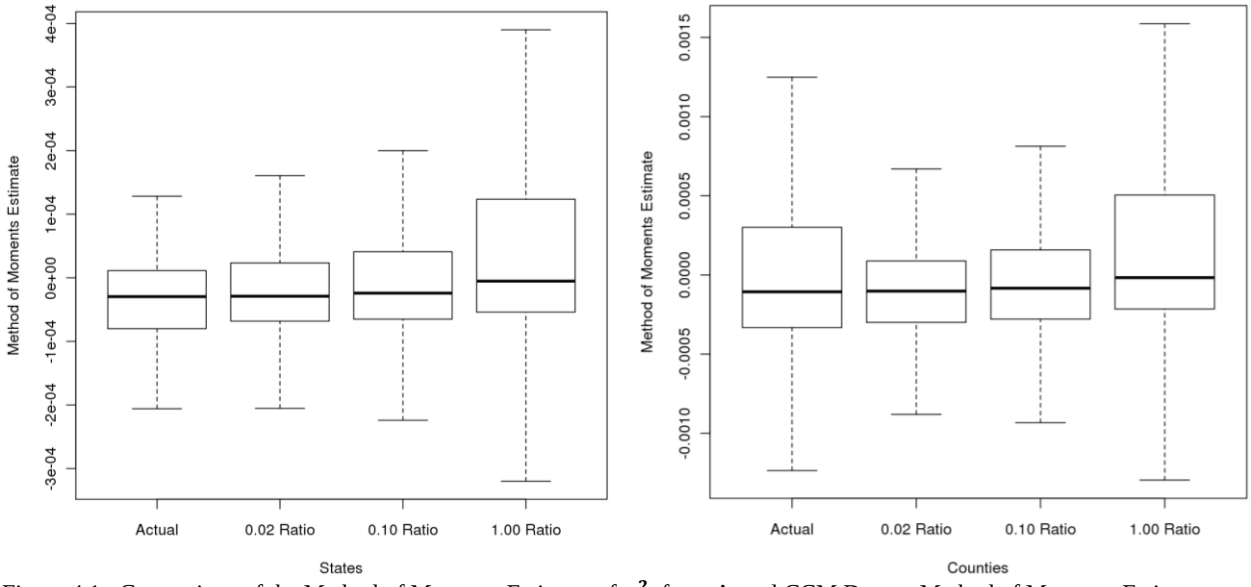


Figure 4.1: Comparison of the Method of Moments Estimate of $\boldsymbol{\sigma_{v_i}^2}$ from Actual CCM Data to Method of Moments Estimates from the Simulated Data

For states, the method of moments estimates of $\sigma_{v_i}^2$ from the 2010 CCM data most closely resembles the estimates from the 0.02 ratio simulation. For counties, the method of moments estimates of $\sigma_{v_i}^2$ from the 2010 CCM data most closely resembles the estimates from the 1.0

ratio simulation. In both cases, the Random Effects model performed best in the corresponding simulations, though the Fixed Covariate with Random Effects model did about as well for the county simulations.

Figure 4.2 compares the 90% confidence intervals that result from the naïve MSE estimator (no bias) and to the Fixed Covariate with Random Effects MSE model estimator for both states and counties. We selected the Fixed Covariate with Random Effects model estimator since it performs nearly as well as the Random Effects model, but also provides area specific estimates of synthetic estimation error. The states and counties are sorted according to their synthetic CE proportion estimate, which is represented by the black circles. The red lines give the upper and lower 90% confidence limits for the naïve estimator, and the blue lines give the limits for the Fixed Covariate with Random Effects model estimator.

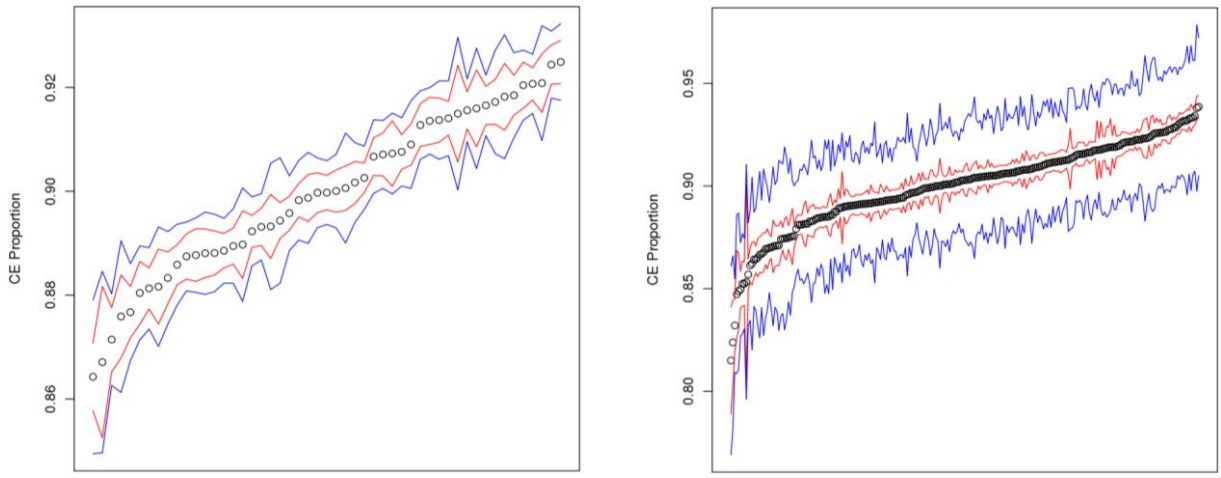


Figure 4.2: Comparison of 90% Confidence Intervals Resulting from the Naïve MSE Estimator (Red) and the Fixed Covariate with Random Effects Model MSE Estimator (Blue)

The addition of the synthetic error bias substantially widens the confidence intervals for both states and counties. While the naïve confidence intervals are too narrow for states, they are

much too narrow for counties, suggesting that there is more synthetic error at the county level than the state level. The confidence intervals from the Fixed Covariate with Random Effects model seem to reflect a reasonable amount of uncertainty in the CE proportion estimates. The differences in the state CE rates are not overwhelmed by statistical uncertainty, as there seems to be real differences across states. The uncertainty of the county CE proportion estimates are larger, making it difficult to find real differences across counties.

5. Discussion and Future Research

The Fixed Covariate and Random Effects models compared favorably with the other MSE estimation methods when applied to simulated data. Throughout the simulations, one of these two methods generally produced the best estimates in terms of RRMSE, while not producing negative or zero estimates. Overall, the best performing estimator was the posterior mode from the Random Effects model. In particular, the benefits of shrinkage estimation showed for $m_D = 20$, even when the underlying model was incorrect. The posterior mode estimates did better than the corresponding posterior mean estimates in almost all cases. The posterior means generally overestimated the true values, substantially in some cases.

The Lahiri and Pramanik (2010) adjustments improved the Gonzalez-Waksberg and Marker estimators, effectively avoiding zero estimates for small $\sigma_{v_i}^2$ without deterioration in performance. These findings are consistent with the results presented in Lahiri and Pramanik (2010).

Yoshimori and Lahiri (2013) proposed adjusted likelihood estimates of σ_v^2, refining an approach proposed by Li and Lahiri (2010) that produced strictly positive estimates. The simulation results in Yoshimori and Lahiri (2013) are broadly consistent with ours in that they show substantial overestimation of σ_v^2 when $\sigma_v^2/\sigma_{e_i}^2$ is small, and little bias when $\sigma_v^2/\sigma_{e_i}^2$ is large. We cannot directly compare our simulation results with those from Yoshimori and Lahiri (2013) though, due to differences in the simulation set ups, assumptions, and in the variance estimation methods considered.

We applied the MSE estimation methods to simulated data motivated by estimation of CEs in the 2010 CCM as well as to actual CE data from the 2010 CCM. The proposed methods generally produced better estimates than the design-based methods when applied to the simulated CE data. In particular, the Random Effects model outperformed all other methods for the 2010 CCM simulated data. We also compared $\sigma_{v_i}^2$ estimates from the true CE data to those from the simulated data. For states and counties, the proposed methods substantially outperformed the design-based methods for the simulated data sets that most closely resembled the CCM data. The MSE estimates resulting from the proposed methods widened (in comparison to the naïve MSE estimates) the confidence intervals of the synthetic CE proportion estimates, substantially for counties. There seemed to be differences in state CE rates, as they were not overwhelmed by the additional statistical uncertainty. The additional uncertainty added to the county CE rates made it difficult to find any real differences.

Note that in the simulations in Section 3, apart from the unadjusted Gonzalez-Waksberg/Marker approach (which produce some negative variance or squared bias estimates), the other methods

all tend to substantially overestimate small values of $\sigma_{v_i}^2$. The bias is a substantial contributor to the RRMSE of the estimates of $\sigma_{v_i}^2$. There is thus a cost to avoiding negative estimates of $\sigma_{v_i}^2$ when it is small. On the other hand, while the estimates for small $\sigma_{v_i}^2$ had high RRMSEs, this may not reflect poor estimation in the absolute sense. An estimate that is several multiples of a small variance may still be small. For large $\sigma_{v_i}^2$, estimates based on a correct model acted like direct variance estimates with a $\chi_{m_D}^2$ distribution.

The Fixed Covariate and Fixed Covariate with Random Effects models have the advantage of flexibility. All other methods require the assignment of areas to groups and the assumption of equal squared bias or synthetic error variance for all areas assigned to the group. The Fixed Covariate and Fixed Covariate with Random Effects models can specify the model for this case by using the group indicators as covariates in the linear predictor. When the squared bias or synthetic area error variance is unique to each area, the Fixed Covariate and Fixed Covariate with Random Effects models can provide area-specific estimates by using continuous covariates. A similar flexibility can be found when comparing logistic regression to post-stratification.

The methods proposed in this manuscript required the assumption that the differences between the synthetic and design-based estimates were normally distributed. Future work may look into the robustness of the results to violations of this normality assumption. Future research may also look into extending the Fixed Covariate with Random Effects model with additional random effects. We specified random effects for the mean estimate within each group. Future models may specify random effects for regression parameters associated with the covariates.

References

Arora, V., and Lahiri, P. (1997). On the Superiority of the Bayesian Method Over the BLUP in Small Area Estimation Problems, *Statistica Sinica*, *7*(4), 1053-1063.

Bell, W. R. (2012). A General Overview of Estimating the Means Squared Error of Synthetic Estimators with Comments on the Application to Census Coverage Measurement, An unpublished paper developed at the U.S. Census Bureau.

Davis, P. and Mulligan, J. (2012). 2010 Census Coverage Measurement Estimation Report: Net Coverage for Household Population in the United States.

Fay, R.E. and Herriot, R.A. (1979). Estimates of Income of Small Places: An Application of James-Stein Procedures to Census Data, *Journal of the American Statistical Association*, 74, 269-277.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*, CRC press.

Gelman, A. (2005). Prior Distributions for Variance Parameters in Hierarchical Models, *Bayesian Analysis*, 1, 1-19.

Gonzalez, M.E., and Waksberg, J. (1973). Estimation of the Error of Synthetic Estimates, In *First Meeting of the International Association of Survey Statisticians*, Vienna, Austria, vol. 18, no. 25, pp. 57-60

Gershunskaya, J. B., and Lahiri, P. (2005). Variance Estimation for Domains in the US Current Employment Statistics Program, In *Proceedings of the American Statistical Association, Survey Research Methods Section*, 3044-3051.

Keller, A. and Fox, T. (2012). 2010 Census Coverage Measurement Estimation Report: Components of Census Coverage for Household Population in the United States.

Lahiri, L. and Pramanik, S. (2010). Estimation of Average Design-based Mean Squared Error of Synthetic Small Area Estimators, *Statistics Canada International Symposium Series: Proceedings*, Ottawa, Ontario: Statistics Canada.

Li, H. and Lahiri, P. (2010). An Adjusted Maximum Likelihood Method for Solving Small Area Estimation Problems, *Journal of Multivariate Analysis*, 101, No. 4, 882-892.

Marker, D.A. (1995). Small Area Estimation: A Bayesian Perspective, Unpublished Ph.D. Dissertation, University of Michigan, Ann Arbor.

Maples, J. J. (2011). Using Small Area Modeling to Improve Design-Based Estimates of Variance for County Level Poverty Rate Estimates in the American Community Survey. *Statistics*, 02.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics, 21*, 1087.

Mule, T. (2012). 2010 Census Coverage Estimation Measurement Report: Summary of Estimates of Coverage for Persons in the United States.

Mule, T. and Griffin, R. (2010). 2010 Census Coverage Measurement Estimation Methodology.

Olson, D. and Griffin, R. (2012). Census Coverage Measurement Estimation Report: Aspects of Modeling.

Otto, M. C. and Bell, W. R. (1995). Sampling Error Modeling of Poverty and Income Statistics for States, In *American Statistical Association, Proceedings of the Section on Government Statistics*, 160-165.

Rao, J.N.K. (2003). *Small Area Estimation, Hoboken*, NJ*:* John Wiley and Sons, Inc.

Seber, G.A.F. (1982). *The Estimation of Animal Abundance, and Related Parameters*, New York: MacMillan.

Song, Haoliang (2007). Synthetic Bias Estimation in Small Area Estimation. *Dissertation Abstracts International*, 68, No. 03.

Stein, C. (1956). Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution, In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, No. 399, 197-206.

You, Y. and Chapman, B. (2006). Small Area Estimation Using Area Level Models and Estimated Sampling Variances, *Survey Methodology*, *32*(1), 97.

Yoshimori, M. and Lahiri, L. (2010). A New Adjusted Maximum Likelihood Method for the Fay-Herriot Small Area Model, Preprint submitted to *Journal of Multivariate Analysis.*

CHAPTER 5: CONCLUSIONS AND DIRECTIONS FOR FUTURE RESEARCH

In recent years, researchers have increasingly administered surveys to subsets of the population to collect information about populations of interest. The increased use of sample surveys was brought on by the increased demand by policy makers for information to make informed decisions (Rao 2003). The work in this dissertation seeks to improve a wide range of survey methodology topics, such that higher quality data are collected and the inferences we make from the estimates generated from the data are more accurate. These improvements would result in improved evidence for sound policy decisions.

The increased need for survey data requires researchers to expand the scope of their surveys. As Thomas et al (2006) notes, matrix sampling designs allow researchers to include additional questions or respondents in the surveys without additional burden on the respondent, enumerator, or subsequent data analyst. Chapter 2, *Stochastic Search Variable Selection Application to Matrix Sampling*, proposed the Stochastic Search Variable Selection (SSVS) matrix sampling approach, an improved method of assigning matrix sampling designs. Our proposed methodology is applicable to all types of variables and requires less pre-specification than previous methods. The simulation results presented in Chapter 2 showed that the proposed methodology administers questions in such a way that multiple imputation generally recovers more of the information lost from the omitted questions than previous methods. The results in Chapter 2 also showed that implementing the SSVS matrix sampling method provides better data than applying the entire questionnaire to a smaller sample of respondents. The simulations did not include the additional improvements in data quality that result from the reduced respondent

burden, enumerator burden, non-response, and premature termination associated with shorter questionnaires (Berdie 1989 and Adams and Gale 1982).

Chapter 3, *The Importance of Cleaning Data During Fieldwork: Evidence from Mozambique*, developed a structured procedure for implementing a household survey in situations with limited resources and other constraints. Even with small questionnaires, enumerators make data entry errors, despite the increased use of Computer–Assisted Personal Interviewing (CAPI) in recent years. The main advantage of CAPI is the ability of the survey collection software on the CAPI device to flag suspicious values as soon as they are collected and the surveyor to correct errors during the interview (De Waal et al 2011). In many small-scale surveys, the technology to implement CAPI is available, but the software necessary for real-time editing is unavailable.

The procedures proposed in Chapter 3 achieve many of the benefits of real-time editing by including an on-the-ground statistician in the survey team to review the data as they are collected. We implemented the proposed procedures during the implementation of a household survey in the Maputo, the capital of Mozambique, and collected detailed data related to the data cleaning process. We found that the data processed using the proposed data cleaning procedures was generally of higher quality than the original data collected from the surveyors.

In addition to collecting higher quality data, we also found additional benefits to the proposed procedures.

- Extended training of the surveyors throughout the survey implementation

- Improved survey management with the survey logistics coordinators focusing on survey implementation
- Real-time monitoring of the surveyors and survey results
- Statisticians gaining general knowledge and experience in all aspects of the survey processes, as suggested in Bethlehem (1997)

In recent years, policy makers increasingly require researchers to use the collected survey data to provide estimates of small domains, or subpopulations defined by geography or some other characteristic (Rao 2003). Researchers often use synthetic (or indirect) estimates for these small domains, which "borrow" strength from data collected in other areas. While indirect estimates have the advantage of lower variability estimates, the pooling of data across areas also introduces synthetic estimation error into the estimates, which may dominate the mean squared error of the synthetic estimates. Estimates of mean squared error and synthetic estimation error are often unstable, so previous methodology averaged estimates of the mean squared error or synthetic estimation error assuming constancy within the averaged groups. In Chapter 4, *Model Based Approach to Synthetic Bias Estimation*, we proposed a parametric model for synthetic estimation error that uses covariates available for the area. This parametric model can incorporate the constancy within groups assumption as a special case, but also provides area-specific estimates. We also extended the model to incorporate multiplicative random effects to achieve shrinkage estimation.

In simulation studies, we showed that our proposed methodology compares favorably to other mean squared error estimation methods, while providing area specific estimates of synthetic

estimation error. We applied our proposed methodology to the estimates of correct enumerations for states and counties during 2010 Census Coverage Measurement, a case where area-specific estimates of synthetic estimation error are required. We showed that the naïve confidence interval estimates of the correct enumeration estimates are far too narrow, compared to the estimated confidence intervals from our propose methodology. We are hopeful that our proposed methodology will be implemented for various estimates produced during 2020 Census Coverage Measurement.

Future Work

We recommend that future research related to *Stochastic Search Variable Selection Application to Matrix Sampling* continue to develop the SSVS matrix sampling method and matrix sampling methodology in general. Only a small number of methods have been proposed on this topic in the literature despite the associated benefits. The research that has been published mainly involves simulation studies, without implementation in fields other than education. Further development of the matrix sampling methodology and simulation studies may convince researchers to implement matrix sampling designs for surveys in field such as public health.

We conducted the analysis in this chapter in a way such that it is generalizable to all household surveys. The next step in the research would be to apply the methodology to a specific survey. This application would incorporate information specific to the survey, such as the selection of data to create the SSVS matrix sampling design, classification of variables into core and split variables, the varying times to collect different variables, and the amount of time to transition from one respondent to another. One potential application would be the follow-up questionnaire

of the Nampula household survey. The data used in our research was the baseline questionnaire of the Nampula household survey, but the follow-up questionnaire was recently administered. The baseline questionnaire can be used to create the SSVS matrix sampling design, and missing data could then be simulated in the follow-up data based on the SSVS matrix sampling design. This application can incorporate the specifics of the survey discussed above, making a strong case for future use in a household survey.

We recommend that future research related to the *The Importance of Cleaning Data During Fieldwork: Evidence from Mozambique* continues to develop data editing methodology directed at small-scale surveys with limited resources and lack of infrastructure. The procedures presented in Chapter 3 were developed for a survey conducted in an area with a reliable internet connection, allowing for the statisticians to be involved remotely. Many surveys in developing countries are not administered in areas where an internet connection is available, requiring the statistician to be on the ground. In previous research, the Laboratory for Interdisciplinary Statistical Analysis (LISA) at Virginia Tech was able to provide this on-the-ground statistician as a member of the survey team and implement similar procedures. Future surveys in developing countries should incorporate the on-the-ground statistician and continue to develop data editing procedures similar to our proposed methods. The increased attention on data quality will hopefully encourage researchers conducting surveys in developing areas to include an on-the-ground statistician to improve the overall quality of data collected from areas where the data is often noisy due to non-sampling errors.

We recommend that future work related to the *Model Based Approach to Synthetic Bias Estimation* continues with our proposed parametric model to improve the accuracy of synthetic estimation error estimates, particularly when the true synthetic estimation error is small compared to the sampling variance of the synthetic estimate and the direct sample variance estimate. Our proposed parametric model and previous methods did not perform well in these simulation studies, overestimating the true synthetic estimation error. Alternative prior specification may improve estimation in these cases. The methods proposed in this manuscript required the assumption that the differences between the synthetic and design-based estimates were normally distributed. Future work may look into the robustness of the results to violations of this normality assumption.

The next step in this research is the application of the Fixed Covariate with Random Effects model to state and county population estimates from the 2020 Census Coverage Measurement (CCM). Our model will provide the Census Bureau with area-specific estimates of the mean squared error of these population estimates that are more precise and accurate than those produced during the 2010 CCM.

References

Adams, L. L. M. and Gale, D. (1982). Solving the Quandary Between Questionnaire Length and Response Rate in Educational Research, *Research in Higher Education*, *17*(3), 231-240.

Berdie, D. R. (1989). Reassessing the Value of High Response Rates to Mail Surveys, *Marketing Research, 1*(3), 52-64.

Bethlehem, J.G. (1987). The Data Editing Research Project of the Netherlands Central Bureau of Statistics, *Proceedings of the Third Annual Research Conference of the Bureau of the Census*, 194-203.

De Waal, T., Pannekoek, J., and Scholtus, S. (2011). *Handbook of Statistical Data Editing and Imputation*, Hoboken, NJ: John Wiley and Sons.

Rao, J.N.K. (2003). *Small Area Estimation, Hoboken*, NJ*:* John Wiley and Sons, Inc.

Thomas, N., Raghunathan, T.E., Schenker, N., Katzoff, M.J., and Johnson, C.L. (2006). An Evaluation of Matrix Sampling Methods Using Data from the National Health and Nutrition Examination Survey, *Survey Methodology*, Vol. 32, No. 2, 217-231.

APPENDIX A: PROOF OF THE TRIANGLE INEQUALITY PROPERTY - SUPPLEMENTARY MATERIAL FOR CHAPTER 2

Let X, Y, and Z be split variables. Then the distance between X and Y is the following

$$d(X,Y) = 1 - mean(\alpha|X,Y)$$

$$= 1 - \frac{1}{n}\sum_{i=1}^{n} Bernoulli\left(\Delta_{XY,i}|\alpha_{XY,i}\right) Beta\left(\alpha_{XY,i}|1,1\right)$$

$$= 1 - \frac{1}{n}\sum_{i=1}^{n} Beta\left(\alpha_{XY,i}|\Delta_{XY,i}+1, 2-\Delta_{XY,i}\right)$$

$$= 1 - \frac{1}{n}\sum_{i=1}^{n}\left\{I_{\Delta_{XY,i}=1} Beta\left(\alpha_{XY,i}|2,1\right) + I_{\Delta_{XY,i}=0} Beta\left(\alpha_{XY,i}|1,2\right)\right\}$$

$$= 1 - \frac{1}{n}\sum_{i=1}^{n}\left\{I_{\Delta_{XY,i}=1} Beta\left(\alpha_{XY,i}|2,1\right)\right\} - \frac{1}{n}\sum_{i=1}^{n}\left\{I_{\Delta_{XY,i}=0} Beta\left(\alpha_{XY,i}|1,2\right)\right\}$$

By the Central Limit Theorem, as n→∞

$$d(X,Y) = 1 - \frac{1}{n}\left\{n_1\left(\frac{2}{3}\right) + n_0\left(\frac{1}{3}\right)\right\} \quad where \sum_{i=1}^{n} I_{\Delta_{XY,i}=1} = n_1 \; and \; \sum_{i=1}^{n} I_{\Delta_{XY,i}=0} = n_0$$

$$= 1 - \frac{1}{3n}\{2n_1 + n_0\}$$

Since $n = n_1 + n_0$,

$$d(X,Y) = 1 - \frac{\{2n_1 + n_0\}}{3(n_1 + n_0)} = \frac{3(n_1 + n_0) - \{2n_1 + n_0\}}{3(n_1 + n_0)}$$

$$= \frac{3(n_1 + n_0) - \{2n_1 + n_0\}}{3(n_1 + n_0)}$$

$$= \frac{n_1 + 2n_0}{3(n_1 + n_0)} = \frac{n_1 + 2n_0}{3n}$$

$$= \frac{\left(\sum_{i=1}^{n} I_{\Delta_{XY,i}=1}\right) + 2\left(\sum_{i=1}^{n} I_{\Delta_{XY,i}=0}\right)}{3n}$$

$$= \frac{\sum_{i=1}^{n}\left(I_{\Delta_{XY,i}=1} + 2I_{\Delta_{XY,i}=0}\right)}{3n}$$

Suppose $d(X,Z) > d(X,Y) + d(Y,Z)$. Then

$$\frac{\sum_{i=1}^{n}\left(I_{\Delta_{XZ,i}=1} + 2I_{\Delta_{XZ,i}=0}\right)}{3n} > \frac{\sum_{i=1}^{n}\left(I_{\Delta_{XY,i}=1} + 2I_{\Delta_{XY,i}=0}\right)}{3n} + \frac{\sum_{i=1}^{n}\left(I_{\Delta_{YZ,i}=1} + 2I_{\Delta_{YZ,i}=0}\right)}{3n}$$

$$\rightarrow 0 > \sum_{i=1}^{n}\left(I_{\Delta_{XY,i}=1} + 2I_{\Delta_{XY,i}=0}\right) + \sum_{i=1}^{n}\left(I_{\Delta_{YZ,i}=1} + 2I_{\Delta_{YZ,i}=0}\right) - \sum_{i=1}^{n}\left(I_{\Delta_{XZ,i}=1} + 2I_{\Delta_{XZ,i}=0}\right)$$

$$\rightarrow 0 > \sum_{i=1}^{n}\left(I_{\Delta_{XY,i}=1} + 2I_{\Delta_{XY,i}=0} + I_{\Delta_{YZ,i}=1} + 2I_{\Delta_{YZ,i}=0} - I_{\Delta_{XZ,i}=1} - 2I_{\Delta_{XZ,i}=0}\right)$$

Let $A_i = I_{\Delta_{XY,i}=1} + 2I_{\Delta_{XY,i}=0} + I_{\Delta_{YZ,i}=1} + 2I_{\Delta_{YZ,i}=0} - I_{\Delta_{XZ,i}=1} - 2I_{\Delta_{XZ,i}=0}$. Consider the possible cases:

Case	$\Delta_{XZ,i}$	$\Delta_{XY,i}$	$\Delta_{YZ,i}$	A_i
1	1	1	1	1
2	1	0	1	Impossible
3	1	1	0	Impossible
4	1	0	0	3
5	0	1	1	Impossible
6	0	0	1	1
7	0	1	0	1
8	0	0	0	2

Thus $min(A_i) = 1$ and

$$\rightarrow 0 > \sum_{i=1}^{n} A_i > \sum_{i=1}^{n} 1 = n$$

$$\rightarrow 0 > n$$

By contradiction, $d(X,Z) > d(X,Y) + d(Y,Z)$ and the triangle inequality holds for the matrix $1 - mean(\alpha|X,Y)$.

APPENDIX B: CHAINED REGRESSION EQUATIONS EXAMPLE - SUPPLEMENTARY MATERIAL FOR CHAPTER 2

For this example, assume that there are three split variables ($Y_1, Y_2,$ and Y_3) with missing values and a single core variable (X) observed for all respondents. Let $Y_1^{obs}, Y_2^{obs},$ and Y_3^{obs} be the observed values of the split variables, and $Y_1^{mis}, Y_2^{mis},$ and Y_3^{mis} be the missing values.

Assume that Y_1, Y_2 and Y_3 are continuous variables. We will use chained regression equations to impute the missing values of these three split variables. For each split variable, we specify a univariate imputation model. For simplicity, let the univariate imputation models be non-Bayesian linear regression.

The initial step of the chained regression equations algorithm imputes initial values for all missing values ($Y_1^{mis(0)}, Y_2^{mis(0)},$ and $Y_3^{mis(0)}$). The MICE function allows the user to specify the initial imputed values, but the default setting is a random selection from the observed values for each variable.

Let $\theta_1, \theta_2,$ and θ_3 be generic parameter vectors associated with the imputation models. For the linear regression models, θ includes the regression parameter vector β and the error variance σ^2. The values of these two parameters are a deterministic function for non-Bayesian linear regression, but these parameters may also be draws from a random distribution, which is the case for Bayesian linear regression. The t^{th} iteration of the chained regression equation is successive draws from the distributions given in Table B.1.

Table B.1: Chained Regression Equation Steps

Step	Generic Equations	Example Specific Equations
1	$\theta_1^{(t)} \sim P\left(\theta_1 \mid Y_1^{obs}, Y_2^{(t-1)}, Y_3^{(t-1)}, X\right)$	$\beta_1 = (Z_1^T Z_1) Z_1^T Y_1^{obs}$ where $Z_1 = \left[\underline{1} \quad X \quad Y_2^{(t-1)} \quad Y_3^{(t-1)}\right]$ $\sigma_1^2 = (Y_1^{obs} - Z_1^T \beta_1)^T (Y_1^{obs} - Z_1^T \beta_1) \Big/ n_{Y_1}^{obs} - 4$
2	$Y_1^{mis(t)} \sim P\left(Y_1 \mid Y_1^{obs}, Y_2^{(t-1)}, Y_3^{(t-1)}, \theta_1^{*(t)}, X\right)$	$Y_1^{mis(t)} \sim Normal(Z^T \beta_1, \sigma_1^2)$
3	$\theta_2^{(t)} \sim P\left(\theta_2 \mid Y_2^{obs}, Y_1^{(t)}, Y_3^{(t-1)}, X\right)$	$\beta_2 = (Z_2^T Z_2) Z_2^T Y_2^{obs}$ where $Z_2 = \left[\underline{1} \quad X \quad Y_1^{(t)} \quad Y_3^{(t-1)}\right]$ $\sigma_2^2 = (Y_2^{obs} - Z_2^T \beta_2)^T (Y_2^{obs} - Z_2^T \beta_2) \Big/ n_{Y_2}^{obs} - 4$
4	$Y_2^{mis(t)} \sim P\left(Y_2 \mid Y_2^{obs}, Y_1^{(t)}, Y_3^{(t-1)}, \theta_2^{*(t)}, X\right)$	$Y_2^{mis(t)} \sim Normal(Z^T \beta_2, \sigma_2^2)$
5	$\theta_3^{(t)} \sim P\left(\theta_3 \mid Y_3^{obs}, Y_1^{(t)}, Y_2^{(t)}, X\right)$	$\beta_3 = (Z_3^T Z_3) Z_3^T Y_3^{obs}$ where $Z_3 = \left[\underline{1} \quad X \quad Y_1^{(t)} \quad Y_2^{(t)}\right]$ $\sigma_3^2 = (Y_3^{obs} - Z_3^T \beta_3)^T (Y_3^{obs} - Z_3^T \beta_3) \Big/ n_{Y_3}^{obs} - 4$
6	$Y_3^{mis(t)} \sim P\left(Y_3 \mid Y_3^{obs}, Y_1^{(t)}, Y_2^{(1)}, \theta_3^{*(t)}, X\right)$	$Y_3^{mis(t)} \sim Normal(Z^T \beta_3, \sigma_3^2)$

The rows included in the Z matrices are only those that coincide with observed values of the split variable that is imputed. For example, the rows of Z_1 coincide with the observed values of split variable Y_1. The values of n^{obs} are the number of observed values of the given split variable. The steps in Table B.1 are iterated for a pre-specified number of iterations *T*. Buuren and Groothuis-Oudshoorn (2011) state that convergence of the algorithm often occurs within 10 to 20 iterations. The draws from the final iteration ($Y_1^{*(T)}$, $Y_2^{*(T)}$, and $Y_3^{*(T)}$) are combined with the observed values (Y_1^{obs}, Y_2^{obs}, and Y_3^{obs}) to form a data set with no missing values.

APPENDIX C: SHRINKAGE ESTIMATE DERIVATION - SUPPLEMENTARY MATERIAL FOR CHAPTER 4

The models used in Otto and Bell (1995), Maples (2011) and the other cited references in Section 3 applied the random effects variance models to direct estimates of the sampling error variances. These applications differ from our use of modeling the synthetic error variance $\sigma_{v_i}^2$ since the residual terms v_i are unobserved, and hence direct estimates are unavailable.

To see how this random effects variance model achieves empirical Bayes shrinkage estimates, suppose we observe the residual term v_i. Since the random effects variance model assumes that the synthetic error variance $\sigma_{v_i}^2$ is constant within groups, the direct estimate of $\sigma_{v_D}^2$ would be:

$$s_D^2 = \frac{1}{m_D} \sum_{i \in D} v_i^2$$

Since the residual terms v_i are distributed $N\left(0, \sigma_{v_D}^2\right)$ with $\sigma_{v_D}^2 = \sigma_v^2 / \varphi_D$ under the random effects model, the distribution of $\varphi_D v_i^2 / \sigma_{v_D}^2$ conditional on φ_D, $\sigma_{v_D}^2$, and δ is distributed χ_1^2. The distribution of s_D^2 conditional on φ_D, $\sigma_{v_D}^2$, and δ is thus the following:

$$s_D^2 | \varphi_D, \sigma_{v_D}^2, \delta \sim \frac{m_D \sigma_{v_D}^2}{\varphi_D} \chi_{m_D}^2$$

Using Bayes' theorem, it is easy to show that:

$$p(\varphi_D|s_D^2,\sigma_v^2,\delta) \propto p(s_D^2|\varphi_D,\sigma_v^2,\delta)p(\varphi_D|\sigma_v^2,\delta) \propto \varphi_D^{\frac{m_D}{2}+\delta-1} e^{-\varphi_D\left[\delta+\frac{m_D s_D^2}{2\sigma_v^2}\right]}$$

which is the kernel of a $Gamma\left(\frac{m_D}{2}+\delta, \delta+\frac{m_D s_D^2}{2\sigma_v^2}\right)$ density. This implies that $1/\varphi_D$ has an inverse gamma distribution, and we have (Gelman et al. 2004, pg. 575) the derivation given in Equation C.1.

$$E\left[\sigma_{v_i}^2|s_D^2,\sigma_v^2,\delta\right] = E\left[\frac{\sigma_v^2}{\varphi_D}|s_D^2,\sigma_v^2,\delta\right] = \frac{\sigma_v^2\delta+\frac{m_D s_D^2}{2}}{\frac{m_D}{2}+\delta-1} = \left[\frac{\delta}{\frac{m_D}{2}+\delta}\sigma_v^2 + \frac{m_D/_2}{\frac{m_D}{2}+\delta}s_D^2\right]\left[\frac{\frac{m_D}{2}+\delta}{\frac{m_D}{2}+\delta-1}\right] \quad \text{(C.1)}$$

For large values of m_D (e.g., $m_D = 50$), the second factor is approximately equal to 1 and the expectation is the weighted average of a common variance parameter σ_v^2 and the "estimated variance" for group *D*, s_D^2. Equation C.1 shrinks the s_D^2 toward the common σ_v^2 around which the $\sigma_{v_D}^2$ are centered. Since we do not observe v_i and s_D^2, this shrinkage occurs within the Metropolis-Hastings iteration used in fitting the model to the d_i.

APPENDIX D: DISTRIBUTION OF PREDICTION ERRORS - SUPPLEMENTARY MATERIAL FOR CHAPTER 4

Table D.1: Distribution of Errors Under the Assumption of Constant Synthetic Error Variance Within Group Assignments – Design Based Methods

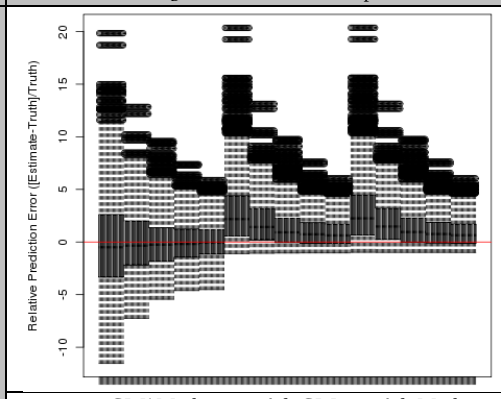
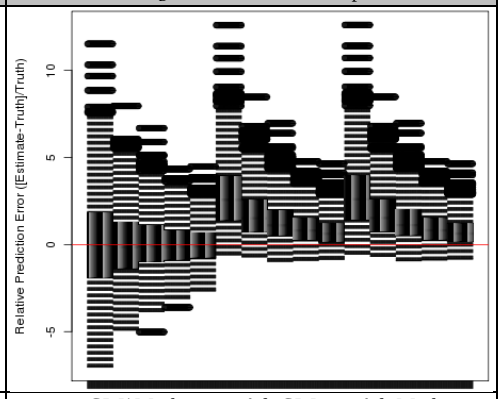
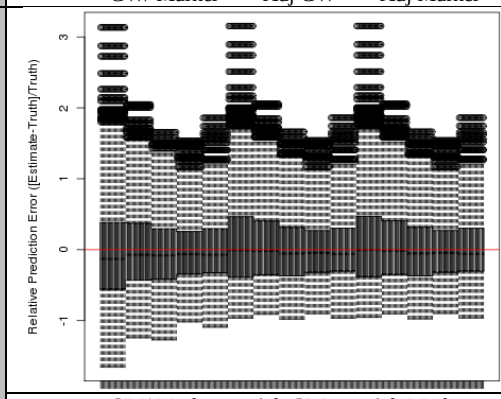
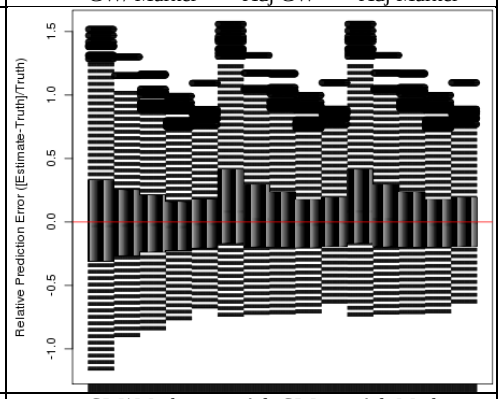
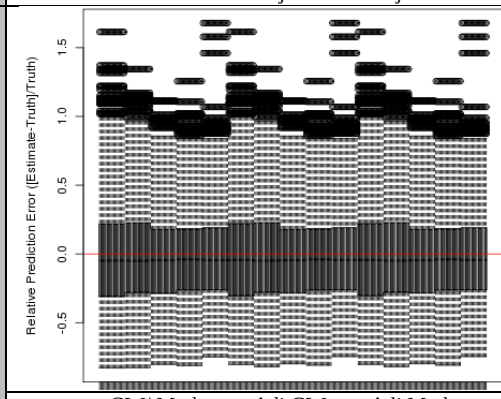
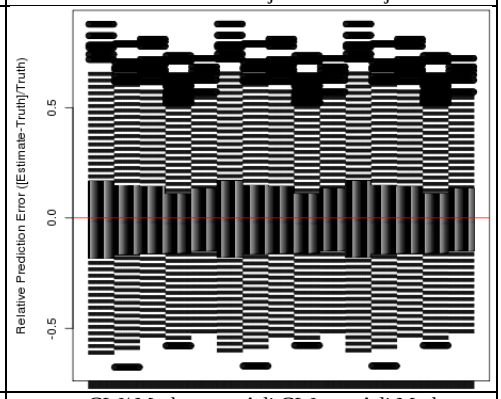
$\sigma_{v_i}^2$	m_D =20 Areas Per Group	m_D =50 Areas Per Group
0.010 $for\ i \in Group\ 1$ 0.015 $for\ i \in Group\ 2$ 0.020 $for\ i \in Group\ 3$ 0.025 $for\ i \in Group\ 4$ 0.030 $for\ i \in Group\ 5$		
0.10 $for\ i \in Group\ 1$ 0.15 $for\ i \in Group\ 2$ 0.20 $for\ i \in Group\ 3$ 0.25 $for\ i \in Group\ 4$ 0.30 $for\ i \in Group\ 5$		
0.50 $for\ i \in Group\ 1$ 0.75 $for\ i \in Group\ 2$ 1.00 $for\ i \in Group\ 3$ 1.25 $for\ i \in Group\ 4$ 1.50 $for\ i \in Group\ 5$		

Table D.2: Distribution of Errors Under the Assumption of Constant Synthetic Error Variance Within Group Assignments – Model Based Method

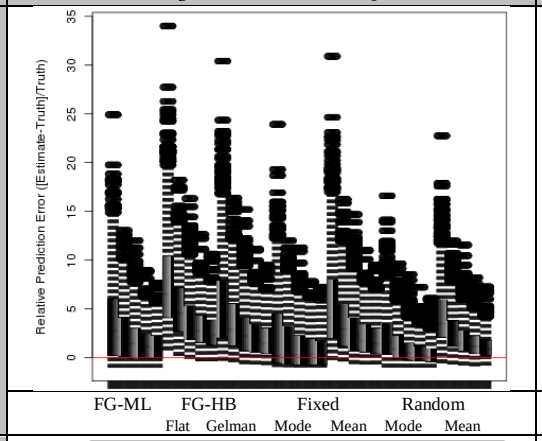
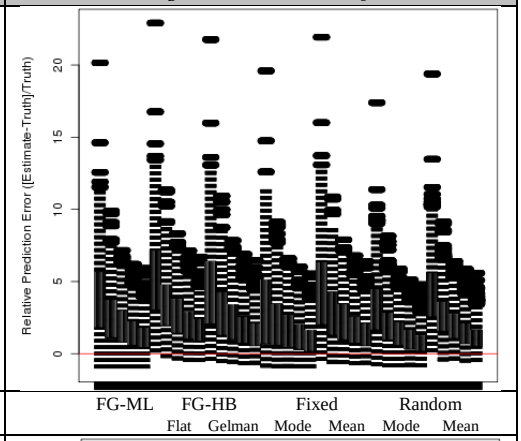
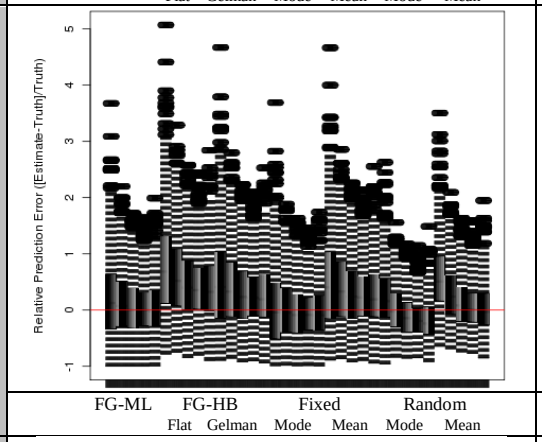
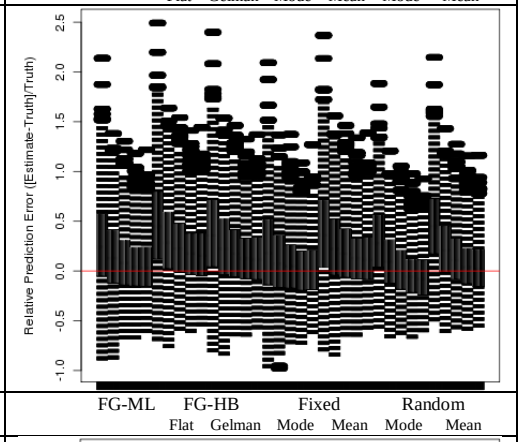
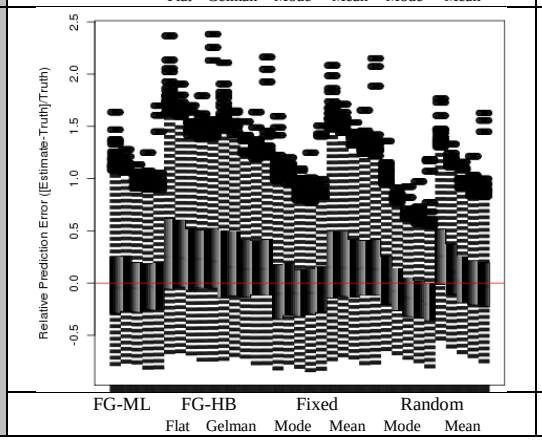
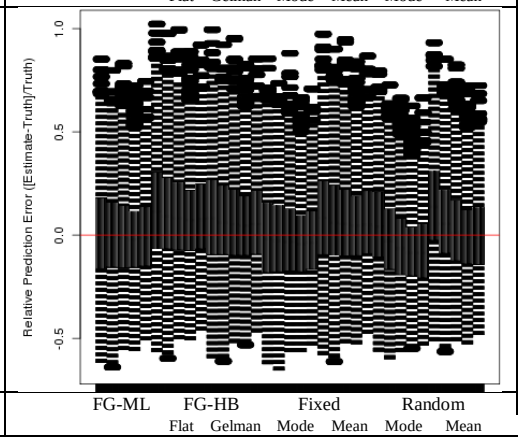
$\sigma^2_{v_i}$	m_D =20 Areas Per Group	m_D =50 Areas Per Group
0.010 $for\ i \in Group\ 1$ 0.015 $for\ i \in Group\ 2$ 0.020 $for\ i \in Group\ 3$ 0.025 $for\ i \in Group\ 4$ 0.030 $for\ i \in Group\ 5$		
0.10 $for\ i \in Group\ 1$ 0.15 $for\ i \in Group\ 2$ 0.20 $for\ i \in Group\ 3$ 0.25 $for\ i \in Group\ 4$ 0.30 $for\ i \in Group\ 5$		
0.50 $for\ all\ i\ in\ Group\ 1$ 0.75 $for\ all\ i\ in\ Group\ 2$ 1.00 $for\ all\ i\ in\ Group\ 3$ 1.25 $for\ all\ i\ in\ Group\ 4$ 1.50 $for\ all\ i\ in\ Group\ 5$		

Table D.3: Distribution of Errors Under the Assumption of Non-constant Synthetic Error Variance With Correct Model Specification

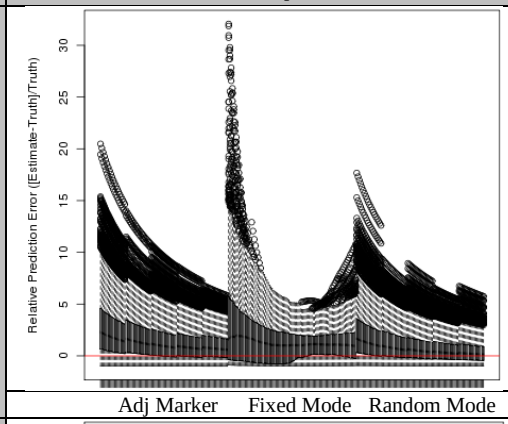
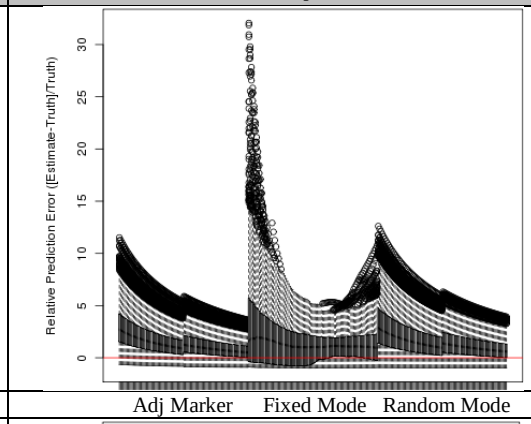
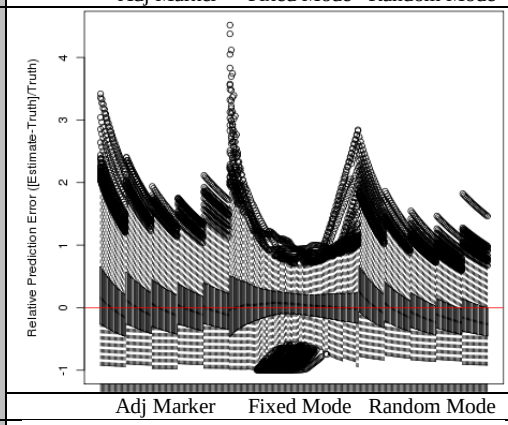
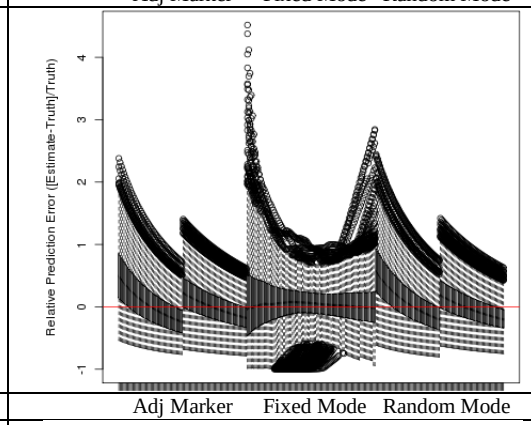
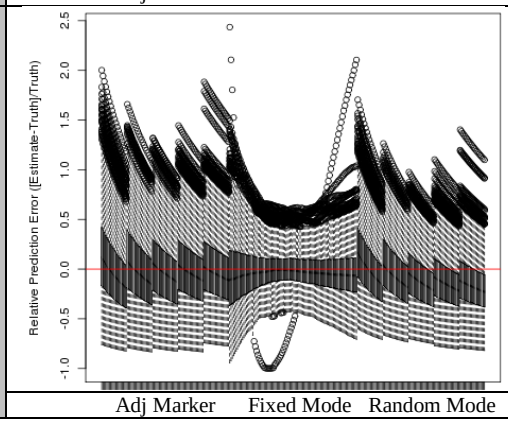
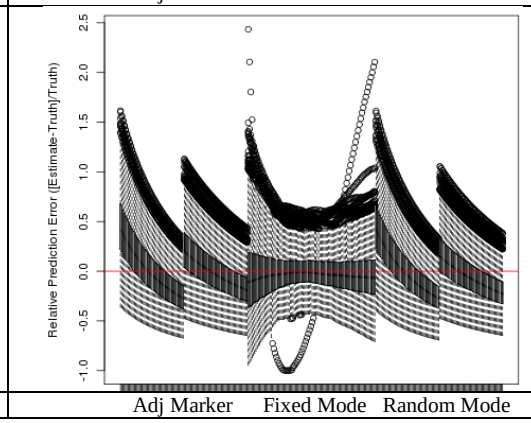
$\sigma_{v\,i}^2$	5 Groups	2 Groups
0.0102 *to* 0.0300 *in steps of* 0.0002		
0.102 *to* 0.300 *in steps of* 0.002		
0.51 *to* 1.5 *in steps of* 0.01		

Table D.4: Distribution of Errors Under the Assumption of Non-constant Synthetic Error Variance With Incorrect Model Specification

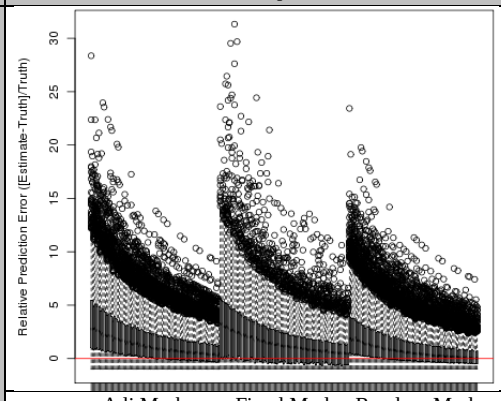
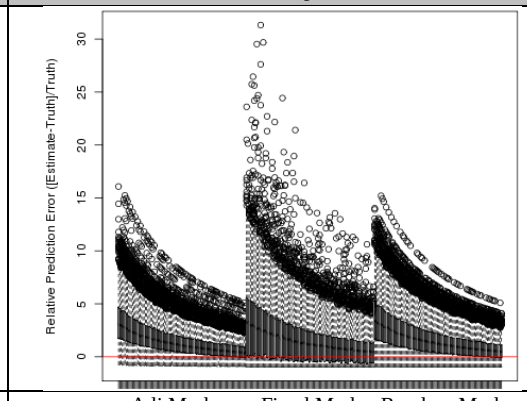
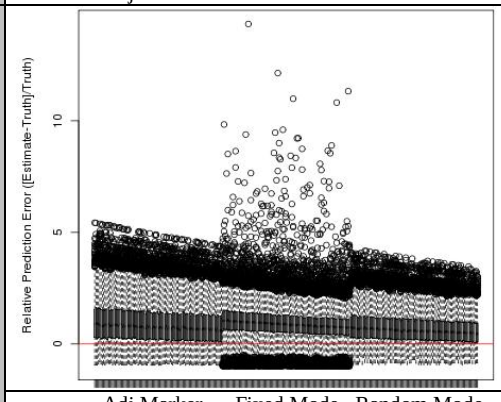
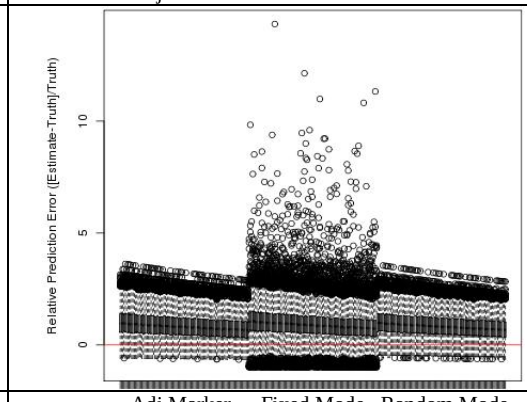
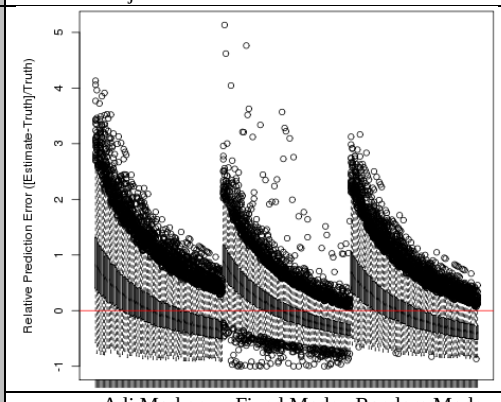
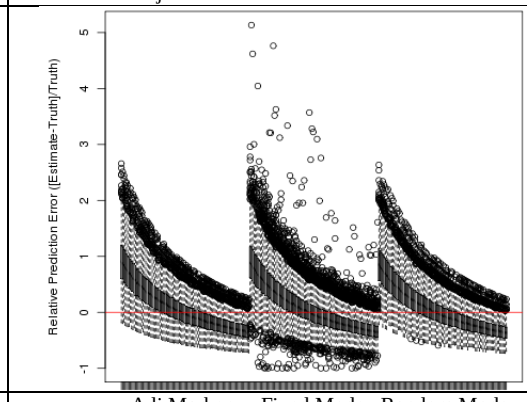
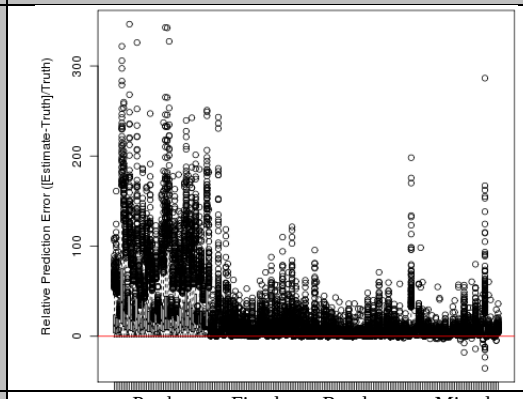
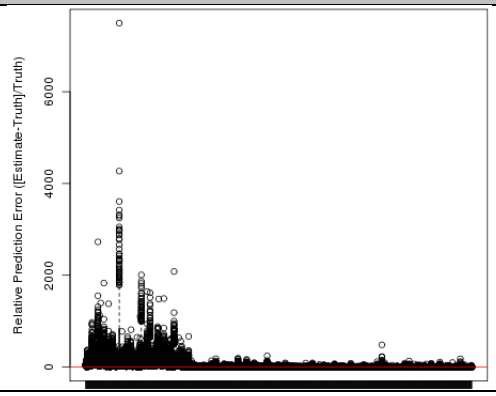
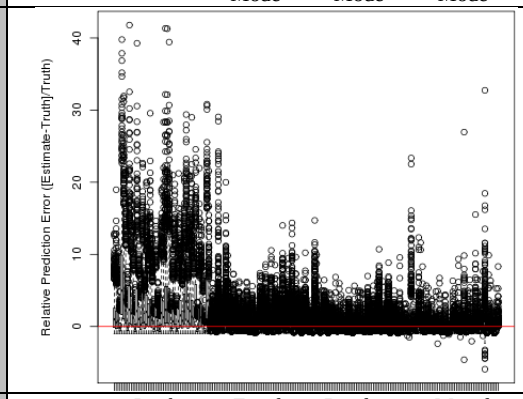
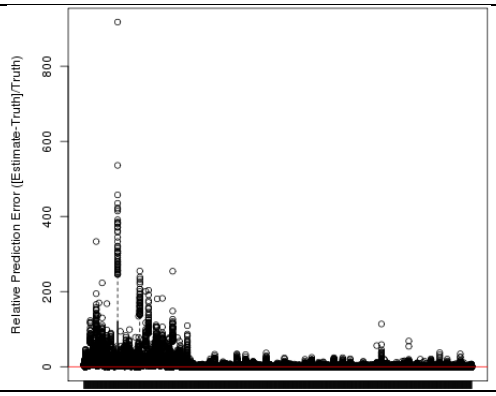
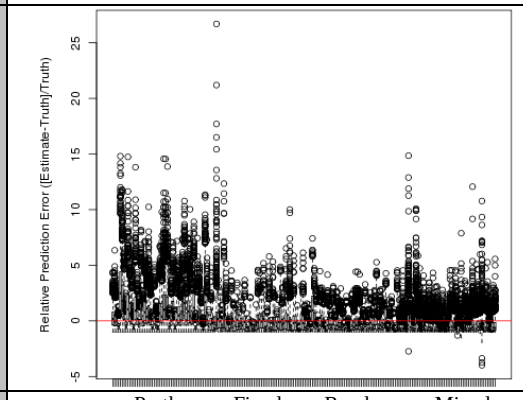
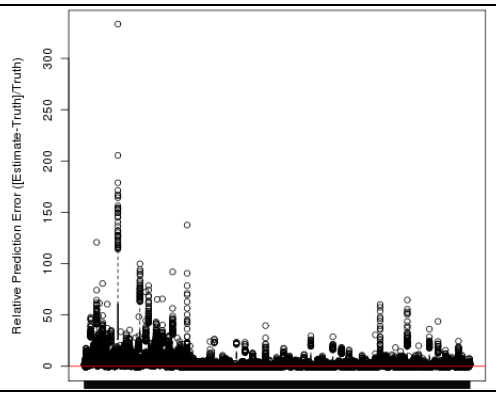
$\sigma^2_{v_i}$	5 Groups	2 Groups
0.0102 *to* 0.0300 *in steps of* 0.0002		
0.102 *to* 0.300 *in steps of* 0.002		
0.51 *to* 1.5 *in steps of* 0.01		

Table D.5: Distribution of Errors for 2010 CCM Simulation Data

$\sigma^2_{v_i}$	States	Counties
$0.02 * (\sigma^2_{e_i} + \sigma^2_{\varepsilon_i})$		
	Partha Fixed Mode Random Mode Mixed Mode	Partha Fixed Mode Random Mode Mixed Mode
$0.1 * (\sigma^2_{e_i} + \sigma^2_{\varepsilon_i})$		
	Partha Fixed Mode Random Mode Mixed Mode	Partha Fixed Mode Random Mode Mixed Mode
$1.0 * (\sigma^2_{e_i} + \sigma^2_{\varepsilon_i})$		
	Partha Fixed Mode Random Mode Mixed Mode	Partha Fixed Mode Random Mode Mixed Mode

APPENDIX E: RMSE COMPONENTS OF THE PREDICTED SYNTHETIC ERROR BIAS– SUPPLEMENTARY MATERIAL FOR CHAPTER 4

E.1 Constant Squared Bias/Synthetic Error Variance Within Groups

Table E.1: Relative Standard Deviation of Designed Based Methods Under the Assumption of Constant Synthetic Error Variance Within Group Assignments

Group	$\sigma_{v\,i}^2$	m_D =20 Areas Per Group			m_D =50 Areas Per Group		
		Gonzalez-Waksberg/ Marker	Adjusted Gonzalez-Waksberg	Adjusted Marker	Gonzalez-Waksberg/ Marker	Adjusted Gonzalez-Waksberg	Adjusted Marker
1	0.010	463%	323%	321%	281%	195%	195%
2	0.015	309%	220%	219%	199%	142%	141%
3	0.020	245%	181%	179%	156%	116%	116%
4	0.025	196%	145%	144%	127%	94%	94%
5	0.030	174%	133%	132%	110%	85%	85%
All		277%	200%	199%	175%	127%	126%
1	0.10	73%	65%	65%	46%	41%	41%
2	0.15	59%	55%	55%	37%	35%	35%
3	0.20	52%	49%	49%	33%	32%	32%
4	0.25	45%	43%	43%	30%	29%	29%
5	0.30	45%	44%	44%	28%	28%	28%
All		55%	51%	51%	35%	33%	33%
1	0.50	40%	40%	40%	25%	25%	25%
2	0.75	37%	37%	37%	23%	23%	23%
3	1.00	34%	34%	34%	23%	23%	23%
4	1.25	33%	33%	33%	21%	21%	21%
5	1.50	35%	35%	35%	21%	21%	21%
All		36%	36%	36%	23%	23%	23%

Table E.2: Relative Bias of Designed Based Methods Under the Assumption of Constant Synthetic Error Variance Within Group Assignments

Group	$\sigma_{v\,i}^2$	m_D =20 Areas Per Group			m_D =50 Areas Per Group		
		Gonzalez-Waksberg/ Marker	Adjusted Gonzalez-Waksberg	Adjusted Marker	Gonzalez-Waksberg/ Marker	Adjusted Gonzalez-Waksberg	Adjusted Marker
1	0.010	-13%	293%	300%	9%	283%	286%
2	0.015	-3%	189%	193%	1%	176%	178%
3	0.020	-8%	132%	135%	9%	132%	133%
4	0.025	-6%	99%	102%	-1%	95%	96%
5	0.030	7%	88%	90%	2%	77%	77%
All		-5%	160%	164%	4%	153%	154%
1	0.10	-3%	11%	12%	1%	13%	13%
2	0.15	0%	6%	6%	0%	5%	5%
3	0.20	-4%	0%	0%	1%	4%	4%
4	0.25	-2%	0%	0%	-1%	1%	1%
5	0.30	0%	1%	1%	0%	1%	2%
All		-2%	4%	4%	0%	5%	5%
1	0.50	-1%	-1%	-1%	0%	0%	0%
2	0.75	0%	0%	0%	-1%	0%	0%
3	1.00	-3%	-3%	-3%	0%	0%	0%
4	1.25	-2%	-2%	-2%	-1%	-1%	-1%
5	1.50	-1%	-1%	-1%	0%	0%	0%
All		-2%	-1%	-1%	0%	0%	0%

Table E.3: Relative Standard Deviation of Model Based Methods Under the Assumption of Constant Synthetic Error Variance Within Group Assignments– m_D=20 Areas Per Group

Group	$\sigma_{v_i}^2$	Fixed Group - ML	Fixed Group - HB		Fixed Covariate		Random Effects	
			Flat Prior	Gelman Prior	Posterior Mode	Posterior Mean	Posterior Mode	Posterior Mean
1	0.010	420%	488%	470%	391%	470%	264%	316%
2	0.015	281%	327%	315%	265%	314%	175%	210%
3	0.020	227%	266%	256%	217%	256%	143%	172%
4	0.025	179%	209%	202%	173%	202%	114%	137%
5	0.030	161%	190%	184%	157%	184%	107%	127%
All		254%	296%	285%	241%	285%	161%	192%
1	0.10	72%	87%	85%	73%	85%	51%	59%
2	0.15	58%	71%	69%	57%	69%	42%	49%
3	0.20	52%	64%	61%	51%	61%	37%	44%
4	0.25	45%	55%	53%	44%	53%	33%	39%
5	0.30	45%	56%	54%	44%	54%	35%	41%
All		54%	67%	64%	54%	64%	40%	46%
1	0.50	40%	50%	47%	39%	47%	34%	39%
2	0.75	37%	46%	44%	36%	44%	30%	35%
3	1.00	34%	43%	40%	33%	40%	27%	32%
4	1.25	33%	42%	39%	32%	39%	26%	31%
5	1.50	35%	43%	41%	33%	41%	27%	32%
All		36%	45%	42%	35%	42%	29%	34%

Table E.4: Relative Bias of Model Based Methods Under the Assumption of Constant Synthetic Error Variance Within Group Assignments– m_D=20 Areas Per Group

Group	$\sigma_{v_i}^2$	Fixed Group - ML	Fixed Group - HB		Fixed Covariate		Random Effects	
			Flat Prior	Gelman Prior	Posterior Mode	Posterior Mean	Posterior Mode	Posterior Mean
1	0.010	361%	773%	549%	229%	548%	227%	431%
2	0.015	241%	522%	370%	151%	369%	129%	271%
3	0.020	173%	388%	272%	105%	272%	81%	190%
4	0.025	136%	312%	218%	79%	218%	52%	142%
5	0.030	118%	268%	188%	67%	189%	37%	115%
All		206%	452%	319%	126%	319%	105%	230%
1	0.10	21%	80%	52%	3%	52%	23%	60%
2	0.15	14%	61%	41%	2%	41%	2%	31%
3	0.20	5%	46%	29%	-5%	29%	-11%	13%
4	0.25	4%	42%	27%	-4%	27%	-16%	6%
5	0.30	4%	40%	27%	-3%	27%	-17%	4%
All		10%	54%	35%	-2%	35%	-4%	23%
1	0.50	1%	32%	21%	-5%	21%	6%	28%
2	0.75	1%	30%	20%	-4%	20%	-5%	14%
3	1.00	-3%	25%	16%	-7%	16%	-13%	5%
4	1.25	-1%	26%	17%	-6%	17%	-15%	2%
5	1.50	-1%	26%	18%	-5%	18%	-16%	0%
All		-1%	28%	18%	-6%	18%	-9%	10%

Table E.5: Relative Standard Deviation of Model Based Methods Under the Assumption of Constant Synthetic Error Variance Within Group Assignments– m_D=50 Areas Per Group

Group	$\sigma_{v_i}^2$	Fixed Group - ML	Fixed Group - HB		Fixed Covariate		Random Effects	
			Flat Prior	Gelman Prior	Posterior Mode	Posterior Mean	Posterior Mode	Posterior Mean
1	0.010	279%	285%	293%	294%	293%	201%	215%
2	0.015	192%	199%	204%	201%	204%	143%	152%
3	0.020	153%	160%	163%	160%	163%	117%	124%
4	0.025	122%	127%	130%	129%	130%	93%	99%
5	0.030	107%	112%	114%	113%	114%	83%	88%
All		170%	177%	181%	179%	181%	127%	136%
1	0.10	46%	50%	50%	47%	50%	38%	41%
2	0.15	37%	40%	39%	37%	39%	31%	33%
3	0.20	33%	36%	36%	33%	36%	29%	31%
4	0.25	30%	32%	32%	30%	32%	26%	28%
5	0.30	28%	31%	30%	28%	30%	25%	27%
All		35%	38%	37%	35%	37%	30%	32%
1	0.50	25%	27%	26%	25%	26%	23%	24%
2	0.75	23%	25%	25%	23%	25%	21%	23%
3	1.00	23%	25%	25%	23%	25%	21%	22%
4	1.25	21%	23%	23%	21%	23%	19%	21%
5	1.50	21%	23%	23%	21%	23%	19%	21%
All		23%	25%	24%	23%	24%	21%	22%

Table E.6: Relative Bias of Model Based Methods Under the Assumption of Constant Synthetic Error Variance Within Group Assignments– m_D=50 Areas Per Group

Group	$\sigma_{v_i}^2$	Fixed Group - ML	Fixed Group - HB		Fixed Covariate		Random Effects	
			Flat Prior	Gelman Prior	Posterior Mode	Posterior Mean	Posterior Mode	Posterior Mean
1	0.010	383%	524%	440%	310%	440%	326%	424%
2	0.015	252%	346%	291%	204%	291%	202%	270%
3	0.020	195%	268%	226%	159%	226%	147%	199%
4	0.025	144%	204%	170%	116%	170%	103%	146%
5	0.030	115%	166%	138%	91%	138%	77%	113%
All		218%	301%	253%	176%	253%	171%	231%
1	0.10	27%	47%	38%	20%	38%	31%	47%
2	0.15	15%	31%	25%	10%	25%	10%	23%
3	0.20	10%	25%	20%	7%	20%	2%	13%
4	0.25	6%	19%	15%	3%	15%	-3%	7%
5	0.30	5%	18%	14%	2%	14%	-5%	4%
All		13%	28%	22%	8%	22%	7%	19%
1	0.50	2%	13%	10%	0%	10%	6%	15%
2	0.75	1%	11%	8%	-2%	8%	-2%	7%
3	1.00	0%	10%	7%	-1%	7%	-4%	3%
4	1.25	-1%	9%	6%	-3%	6%	-7%	1%
5	1.50	0%	10%	7%	-2%	7%	-7%	0%
All		1%	11%	8%	-1%	8%	-3%	5%

E.2 Nonconstant Squared Bias/Synthetic Error Variance – Correct Model Specification

Table E.7: Relative Standard Deviation Under the Assumption of Non-constant Synthetic Error Variance With Correct Model Specification

$\sigma_{v_i}^2$	Group	20 Areas Per Group			50 Areas Per Group		
		Adjusted Marker	Fixed Covariate Posterior Mode	Random Effects Posterior Mode	Adjusted Marker	Fixed Covariate Posterior Mode	Random Effects Posterior Mode
$\sigma_{v_i}^2 = 0.0102\ to\ 0.0300$ $in\ steps\ of\ 0.0002$	1	274%	342%	225%	145%	245%	180%
	2	207%	191%	165%	97%	143%	113%
	3	179%	148%	143%			
	4	148%	131%	118%			
	5	139%	157%	111%			
	All	190%	194%	152%	121%	194%	146%
$\sigma_{v_i}^2 = 0.102\ to\ 0.300$ $in\ steps\ of\ 0.01$	1	61%	57%	46%	36%	43%	35%
	2	54%	36%	40%	29%	31%	28%
	3	49%	28%	37%			
	4	44%	28%	33%			
	5	45%	37%	35%			
	All	51%	37%	38%	33%	37%	32%
$\sigma_{v_i}^2 = 0.51\ to\ 1.5$ $in\ steps\ of\ 0.01$	1	39%	32%	32%	24%	24%	23%
	2	37%	20%	29%	22%	21%	21%
	3	34%	17%	27%			
	4	34%	19%	26%			
	5	35%	24%	27%			
	All	36%	22%	28%	23%	22%	22%

Table E.8: Relative Bias Under the Assumption of Non-constant Synthetic Error Variance With Correct Model Specification

$\sigma_{v_i}^2$	Group	20 Areas Per Group			50 Areas Per Group		
		Adjusted Marker	Fixed Covariate Posterior Mode	Random Effects Posterior Mode	Adjusted Marker	Fixed Covariate Posterior Mode	Random Effects Posterior Mode
$\sigma_{v_i}^2 = 0.0102\ to\ 0.0300$ $in\ steps\ of\ 0.0002$	1	245%	251%	180%	183%	174%	196%
	2	179%	132%	117%	99%	111%	104%
	3	135%	102%	82%			
	4	107%	108%	57%			
	5	98%	119%	44%			
	All	153%	142%	96%	141%	142%	150%
$\sigma_{v_i}^2 = 0.102\ to\ 0.300$ $in\ steps\ of\ 0.01$	1	8%	8%	13%	7%	6%	13%
	2	6%	6%	0%	2%	4%	-1%
	3	0%	5%	-10%			
	4	0%	4%	-14%			
	5	2%	3%	-16%			
	All	3%	5%	-5%	5%	5%	6%
$\sigma_{v_i}^2 = 0.51\ to\ 1.5$ $in\ steps\ of\ 0.01$	1	-1%	-4%	2%	2%	-2%	2%
	2	0%	-1%	-5%	0%	-3%	-5%
	3	-3%	-1%	-12%			
	4	-2%	-3%	-13%			
	5	-1%	-4%	-15%			
	All	-1%	-2%	-9%	1%	-2%	-1%

E.3 Nonconstant Squared Bias/Synthetic Error Variance – Incorrect Model Specification

Table E.9: Relative Standard Deviation Under the Assumption of Non-constant Synthetic Error Variance With Incorrect Model Specification

$\sigma_{v_i}^2$	Group	20 Areas Per Group			50 Areas Per Group		
		Adjusted Marker	Fixed Covariate Posterior Mode	Random Effects Posterior Mode	Adjusted Marker	Fixed Covariate Posterior Mode	Random Effects Posterior Mode
$\sigma_{v_i}^2 = 0.0102\ to\ 0.0300$ $in\ steps\ of\ 0.0002$	1	295%	301%	235%	157%	248%	186%
	2	218%	227%	175%	92%	144%	109%
	3	173%	176%	138%			
	4	145%	149%	116%			
	5	126%	128%	101%			
	All	191%	196%	153%	125%	196%	147%
$\sigma_{v_i}^2 = 0.102\ to\ 0.300$ $in\ steps\ of\ 0.01$	1	97%	72%	73%	60%	71%	58%
	2	94%	70%	70%	56%	66%	53%
	3	90%	70%	67%			
	4	88%	66%	66%			
	5	86%	64%	64%			
	All	91%	68%	68%	58%	68%	58%
$\sigma_{v_i}^2 = 0.51\ to\ 1.5$ $in\ steps\ of\ 0.01$	1	60%	39%	47%	31%	32%	30%
	2	45%	29%	35%	18%	19%	17%
	3	36%	24%	28%			
	4	31%	19%	24%			
	5	27%	17%	21%			
	All	40%	25%	31%	25%	25%	24%

Table E.10: Relative Bias Under the Assumption of Non-constant Synthetic Error Variance With Incorrect Model Specification

$\sigma_{v_i}^2$	Group	20 Areas Per Group			50 Areas Per Group		
		Adjusted Marker	Fixed Covariate Posterior Mode	Random Effects Posterior Mode	Adjusted Marker	Fixed Covariate Posterior Mode	Random Effects Posterior Mode
$\sigma_{v_i}^2 = 0.0102\ to\ 0.0300$ $in\ steps\ of\ 0.0002$	1	296%	284%	205%	210%	217%	223%
	2	197%	189%	129%	82%	86%	90%
	3	135%	128%	81%			
	4	97%	92%	52%			
	5	71%	65%	31%			
	All	159%	152%	100%	146%	152%	156%
$\sigma_{v_i}^2 = 0.102\ to\ 0.300$ $in\ steps\ of\ 0.01$	1	98%	104%	79%	93%	100%	95%
	2	92%	98%	73%	78%	83%	79%
	3	86%	91%	68%			
	4	81%	84%	63%			
	5	76%	79%	58%			
	All	87%	91%	68%	85%	91%	87%
$\sigma_{v_i}^2 = 0.51\ to\ 1.5$ $in\ steps\ of\ 0.01$	1	62%	63%	49%	36%	35%	32%
	2	23%	22%	13%	-20%	-21%	-22%
	3	-2%	-2%	-10%			
	4	-17%	-18%	-24%			
	5	-28%	-30%	-34%			
	All	8%	7%	-1%	8%	7%	5%

E.4 2010 CCM Simulations

Table E.11: Components of RRMSE for 2010 CCM State Simulation Data

$\sigma_{v\,i}^2$	Group	Relative Standard Deviation				Relative Bias			
		Adjusted Marker	Fixed Covariate Posterior Mode	Random Effects Posterior Mode	Fixed Covariate with Random Effects Posterior Mode	Adjusted Marker	Fixed Covariate Posterior Mode	Random Effects Posterior Mode	Fixed Covariate with Random Effects Posterior Mode
$0.02 * (\sigma_{e\,i}^2 + \sigma_{\varepsilon\,i}^2)$	1	2,463%	425%	106%	226%	2,056%	150%	-31%	83%
	2	2,116%	209%	158%	117%	2,043%	14%	-8%	-10%
	3	2,074%	564%	362%	451%	2,053%	239%	90%	223%
	All	2,218%	399%	209%	265%	2,051%	134%	17%	99%
$0.1 * (\sigma_{e\,i}^2 + \sigma_{\varepsilon\,i}^2)$	1	314%	98%	32%	52%	207%	-43%	-82%	-65%
	2	274%	56%	43%	26%	207%	-70%	-77%	-82%
	3	268%	105%	82%	75%	208%	-38%	-53%	-50%
	All	285%	86%	52%	51%	207%	-50%	-71%	-66%
$1.0 * (\sigma_{e\,i}^2 + \sigma_{\varepsilon\,i}^2)$	1	122%	76%	46%	60%	58%	-9%	-49%	-45%
	2	108%	48%	50%	39%	60%	-34%	-40%	-64%
	3	106%	79%	64%	63%	58%	-22%	-12%	-50%
	All	112%	68%	53%	54%	59%	-22%	-34%	-53%

Table E.12: Components of RRMSE for 2010 CCM County Simulation Data

$\sigma_{v\,i}^2$	Group	Relative Standard Deviation				Relative Bias			
		Adjusted Marker	Fixed Covariate Posterior Mode	Random Effects Posterior Mode	Fixed Covariate with Random Effects Posterior Mode	Adjusted Marker	Fixed Covariate Posterior Mode	Random Effects Posterior Mode	Fixed Covariate with Random Effects Posterior Mode
$0.02 * (\sigma_{e\,i}^2 + \sigma_{\varepsilon\,i}^2)$	1	2,789%	163%	32%	88%	3,085%	-24%	-77%	-49%
	2	3,474%	188%	64%	78%	4,079%	-14%	-60%	-49%
	3	6,131%	126%	64%	55%	6,056%	-39%	-61%	-62%
	4	2,320%	62%	71%	33%	2,791%	-67%	-70%	-76%
	5	2,850%	320%	103%	190%	3,355%	58%	-53%	13%
	All	3,513%	172%	67%	89%	3,873%	-17%	-64%	-45%
$0.1 * (\sigma_{e\,i}^2 + \sigma_{\varepsilon\,i}^2)$	1	361%	51%	27%	30%	366%	-68%	-87%	-83%
	2	450%	62%	36%	33%	506%	-57%	-78%	-80%
	3	784%	50%	30%	28%	789%	-66%	-80%	-84%
	4	301%	30%	34%	17%	321%	-81%	-83%	-91%
	5	369%	87%	46%	58%	404%	-54%	-74%	-71%
	All	453%	56%	34%	33%	477%	-65%	-80%	-82%
$1.0 * (\sigma_{e\,i}^2 + \sigma_{\varepsilon\,i}^2)$	1	142%	39%	43%	43%	151%	-17%	-28%	-39%
	2	177%	44%	43%	57%	220%	31%	-10%	-4%
	3	303%	38%	45%	55%	367%	21%	-14%	-9%
	4	119%	24%	39%	34%	125%	-30%	-23%	-45%
	5	145%	69%	45%	64%	169%	-11%	-8%	-33%
	All	177%	43%	43%	51%	206%	-1%	-17%	-26%