Copula-based Predictions in Small Area Estimation

by

Kanika Grover

A Thesis submitted to the Faculty of Graduate Studies of
The University of Manitoba
in partial fulfilment of the requirements of the degree of

MASTER OF SCIENCE

Department of Statistics
University of Manitoba
Winnipeg

Abstract

In survey sampling, policymaking regarding the allocation of resources to subgroups (called small areas) or the determination of subgroups with specific properties in a population should be based on reliable estimates. Information, however, is often collected at a larger scale (e.g., surveys) than that of these subgroups. As a result, the sample sizes within subgroups (areas) are often too small to warrant the use of the traditional area-specific direct estimates. Mixed models (unit-level or area-level) are usually used to borrow strength from other resources to get reliable estimates for small areas.

The underlying assumptions of unit-level regression models in small area estimation are often not met in applications. For instance, in most business surveys, errors may have a skewed distributions, and the relationship of response and auxiliary variables often deviate from a linear form. In order to develop a strategy for modeling non-normal continuous data, a flexible small area estimation model has been recently proposed using the linear regression to estimate the error terms and a multivariate exchangeable copula model to characterize the error distribution within each small area.

In this thesis, a likelihood framework is proposed to estimate the intra-class dependence of the multivariate exchangeable copula for empirical best unbiased

prediction (EBUP) of small area means. We consider both parametric and semi-parametric approaches, and propose a bootstrap method under each approach to obtain a nearly unbiased estimate of the mean squared prediction error (MSPE) of the EBUP of small area means.

Our findings suggest that the parametric and semi-parametric approaches yield similar prediction results in respect of the EBUP of small area means; the proposed bootstrap method is capable of capturing complete information of the MSPE of the EBUP of small area means; low relative bias of MSPE estimation of the EBUP of small area means is observed; and in particular, the semi-parametric method consistently performs well in comparison with the parametric method.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisors Dr. Mahmoud Torabi and Dr. Elif Acar. Their cordial support, motivation and guidance boosted my confidence whenever I was discouraged. Their contribution to the direction of this research is innumerable. I will remain grateful to them for their kindness and support as a teacher in the real sense of the term. Thank you very much.

I would like to thank Dr. Liqun Wang and Dr. Depeng Jiang for serving as a member of my thesis examining committee and providing insightful comments. I would also like to thank Dr. Louis-Paul Rivest for providing their package materials which turn out to be an asset for this work. Thanks to University of Manitoba for providing necessary research facilities for completing this thesis. To my colleagues and friends, I say thanks for being there.

Finally, my deepest thanks and gratitude to my parents for their continuous love and concerns. I wish to give my heartfelt thanks to my sister and brother for their love and the emotional support.

This thesis is dedicated to my mother, my source of inspiration.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Small area estimation (SAE) has gained popularity in sample surveys due to growth in the demand of reliable small area statistics, in both private and public sectors. Sample surveys, whether conducted by policy makers or private organizations, aim to provide reliable estimates not only for the whole population but also for small areas/domains. SAE is a method to tackle with problems when sample size is not large enough in small areas/domains to provide estimates with adequate precision. The term "small area" refers to the sample size of the area rather than the actual size of the area where we assume that the sample size is small compared to the population size. In addition, the term “area” does not always define as geographical region or district; it may define a specific demographic group or domain (e.g., age-sex-race domains).

In the context of estimation, the traditionally designed direct domain estimators only use domain-specific sampled data that are often less reliable due to small sample size. To overcome this difficulty, it is common to borrow strength from

related sources. This is usually done in a modeling framework, where auxiliary data on covariates, typically obtained from administrative records and census, are used to share information across the small areas. The resulting estimators are referred to as indirect estimators, which are broadly classified into two categories: (i) traditional estimators such as synthetic estimators based on implicit linking model, and (ii) model-based estimators based on explicit linking model. The traditional estimators proceed with an assumption that small area will follow the same features as the large area. If the implicit linking model is close to the true model, the estimators are expected to have small mean squared prediction errors (MSPE).

To handle complex issues in traditional indirect estimators like a violation of homogeneity within a domain or changes in the population structure, the use of mixed models is proposed in the literature. The area-level model of Fay and Herriot (1979) and the unit-level model of Battese et al. (1988) have been extensively employed in SAE as a special cases of mixed models. In both, area-specific random effects are usually assumed to be normally distributed. In addition, the area-level covariates are used in the area-level model while the unit- and area-level covariates can be used in the unit-level model. Also, the design-based direct estimates in the area-level model often primes to unreliable estimates due to small sample sizes and not including survey weights (You and Rao, 2003; Torabi and Rao, 2010). Many popular small area models and methods have been developed under area-level and unit-level models (Rao and Molina, 2015). The commonly used model-based estimators are empirical best linear unbiased prediction (EBLUP)

(Prasad and Rao, 1990) and (Lahiri and Rao, 1995), James-Stein shrinkage estimation or the empirical Bayes (EB) procedures (Fay and Herriot, 1979) and (Ghosh and Lahiri, 1987), and the hierarchical Bayes (HB) (Datta and Ghosh, 1991) and (Ghosh and Lahiri, 1992). These estimators often provide reliable estimates for SAE obtained under assumed models with respect to the nature of the response variable.

However, when the underlying assumptions of area-level and unit-level models are not met in applications, the EBLUP may lead to biased estimates. The violations may be due to:

1. Errors may have a skewed distribution.

2. Relationship between auxiliary data and response variable is not linear.

In order to address these data aspects in the modeling strategy, Rivest et al. (2016) introduced a flexible small area model that is characterized by multivariate exchangeable copulas. Their work outlined a two-stage semi-parametric approach where the marginal distribution of regression errors is estimated using their empirical distribution and the intra-class dependence is quantified via the empirical Kendall's tau. Using this model, they derived the empirical best unbiased prediction (EBUP) of small area means and used the jackknife to estimate the MSPE of the EBUP.

In an effort to contribute to this new methodology, this thesis revisits the quantification of the dependence measure of errors within each small area and

the estimation of MSPE of the EBUP of small area means. Our first contribution is a maximum pseudo copula log-likelihood framework to estimate the intra-class dependence of the error distribution. The second contribution is a complete assessment of the MSPE of the EBUP, in particular the pattern of the cross-product term, via extensive simulations. The third contribution is a bootstrap approach to estimate the MSPE of the EBUP of small area means.

A major advantage of the proposed approach is that it can accommodate both the two-stage parametric and two-stage semi-parametric estimation. Along with this, the copula-based SAE gives a more flexibility in modeling the error distribution. Further, we compare the performances of the proposed methods to those in Rivest et al. (2016), which we refer to as the RVB method throughout the thesis.

The remaining of this chapter provides an overview of small area estimation and copula models.

1.1 A Brief Overview of Small Area Estimation

SAE is a solution to the problem of producing adequate estimates of the characteristics of interest, such as mean, counts and proportions for areas with small or no sample. While the main interest is in point estimators, evaluation of their precision and the estimation error are also important (Pfeffermann et al., 2013).

Before introducing small area estimation methods, we briefly describe small area models, which are broadly divided into two types: (i) Area-Level Model, and

(ii) Unit-Level Model. First, we introduce linear mixed models (LMM) which are regarded as a general case of the above mentioned models.

1.1.1 Linear Mixed Model

Suppose sample data obey the linear mixed model where the term "mixed" denotes the mixture of random and fixed effects. Then, the LMM can be expressed as

$$y = X\beta + Z\upsilon + \xi \tag{1.1}$$

where y is the $n \times 1$ vector of sample observations, X and Z are the known matrices $n \times p$ of full rank, β is the vector of unknown fixed effects, υ and ξ are distributed independently with mean 0 and covariance matrices G and R, respectively depending on some variance component parameters θ. If υ and ξ follow a normal distribution, then the distribution of y can be derived as

$$y \sim N(X\beta, V)$$

$$y|\upsilon \sim N(X\beta + Z\upsilon, R)$$

where V is the variance-covariance matrix of y, $\mathrm{Var}(y) = V = R + ZGZ^T$. To be more precise $V = V(\theta)$ is non-singular for all θ, belonging to a specified subset of Euclidean q-space, $\theta = (\sigma_1^2, \ldots, \sigma_q^2)$ but not necessarily positive. Under model (1.1), we are generally interested in estimating the linear combination $\mu = l^T\beta + h^T\upsilon$, of fixed parameter β and realized value υ, for the specified vectors of l and h, of

constants. Also, if they are known, i.e., v and ξ follow normal distributions then, the BLUP of μ is given as

$$\tilde{\mu} = l^T \tilde{\beta} + h^T \tilde{v}, \tag{1.2}$$

where $\tilde{\beta} = (X^T V^{-1} X)^{-1} (X^T V^{-1} Y)$ is generalized least square estimator of β, and the BLUP of v is defined as $\tilde{v} = GZ^T V^{-1}(y - X\tilde{\beta})$.

The BLUP (1.2) depends on the variance components. Generally, these values are unknown. If θ is replaced by $\hat{\theta}$, i.e., replacing variance components with their estimates is called EBLUP, obtained as

$$\hat{\mu} = l^T \hat{\beta} + h^T \hat{v}, \tag{1.3}$$

where $\hat{\beta} = \tilde{\beta}(\hat{\theta})$, $\hat{v} = \tilde{v}(\hat{\theta})$, and $\hat{\mu} = \tilde{\mu}(\hat{\theta})$. Commonly used approaches to estimate the variance components are maximum likelihood (ML) and restricted ML (REML).The details of various approaches to estimate the variance components can be found in Ghosh and Rao (1994) and Rao and Molina (2015).

1.1.2 Small Area Models

As discussed, small area models are classified into two types: (i) Area-level model, where area-specific covariates are modeled with the small-area response variable, and (ii) Unit-level model, where unit-specific covariates are modeled with the unit-level response variable.

Area-Level Model

Fay and Herriot (1979) introduced the basic area-level model to predict small areas using the census information. It assumes that the model linking area means is related to areas-specific auxiliary data. Such models use information on the area-level when the unit-level information is unavailable and are also useful to reduce computational complexities. The basic area-level model is represented as

$$y_i = x_i^T \beta + v_i + \xi_i \qquad i = 1, \ldots, m, \tag{1.4}$$

where y_i is a direct estimator, x_i is a p $\times$ 1 vector containing information for p variables, β is a p $\times$ 1 vector of regression coefficients, v_i is area-specific random effects assumed to be independent and identically distributed as $N(0, \sigma_v^2)$, and the sampling error ξ_i is assumed to be independent and identically distributed as $N(0, \sigma_{\xi_i}^2)$ where $\sigma_{\xi_i}^2$ is assumed to be known due to identifiability issue.

The above mentioned model is represented as a combination of the sampling model and the linking model. More specifically, it is assumed that the population area mean, and area-specific auxiliary data have the following form available for area $i (= 1, \ldots, m)$,

$$\mu_i = x_i^T \beta + v_i, \tag{1.5}$$

where μ_i is the small area parameter of interest. We assume that a direct survey estimator y_i is available as

$$y_i = \mu_i + \xi_i. \tag{1.6}$$

Furthermore, combining equation (1.5) and (1.6) leads to equation (1.4), where ξ_i (design-induced errors) and ν_i (model error) are assumed to be independent.

Unit-Level Model

The population model for unit-level data can be represented as

$$y_{ij} = x_{ij}^T \beta + \nu_i + \xi_{ij} \qquad i = 1,\ldots,m; j = 1,\ldots,N_i, \tag{1.7}$$

where y_{ij} is the variable of interest and x_{ij} is the vector of element-specific auxiliary variable, $i = 1,\ldots,m; j = 1,\ldots,N_i$, β is the vector of regression parameters, ν_i's are the area-specific random effects which are assumed to be independent with mean 0 and variance σ_ν^2, and ξ_{ij}'s are individual unit errors which are random variables with mean 0 and variance σ_ξ^2. Also, ν_i and ξ_{ij} are assumed to be independent. If N_i is large and the sampling fraction ($f_i = n_i/N_i$) is negligible, the mean of the i^{th} area can be written as

$$\mu_i = \bar{x}_i^T \beta + \nu_i, \qquad i = 1,\ldots,m, \tag{1.8}$$

where $\bar{x}_i$ is the known mean of covariates at the i^{th} area. Here, we assume that the sample is representative of the population, i.e., there is no selection bias. Thus, our model can be written as

$$y_{ij} = x_{ij}^T \beta + \nu_i + \xi_{ij}, \qquad i = 1\ldots,m; j = 1,\ldots,n_i, \tag{1.9}$$

where sample size in each area can be $n_i \geq 1$.

1.1.3 Model-based Estimation

MSPE of the EBLUP

In the LMM setting, the MSPE of the EBLUP of small area means with respect to the model (1.1) takes the form

$$\text{MSPE}(\hat{\mu}) = \text{MSPE}(\tilde{\mu}) + E(\hat{\mu} - \tilde{\mu})^2. \tag{1.10}$$

Also, under normality assumption it is clear that the MSPE of the EBLUP is always larger as compared to the BLUP of $\tilde{\mu}$. Now, the first term of MSPE($\hat{\mu}$) can be expressed as MSPE($\tilde{\mu}$) = $g_1(\theta) + g_2(\theta)$, where

$$g_1(\theta) = h^T(G - GZ^TV^{-1}ZG)h \tag{1.11}$$

and

$$g_2(\theta) = d^T(X^TV^{-1}X)^{-1}d \tag{1.12}$$

with $d^T = 1^T - b^TX$ and $b^T = h^TGZ^TV^{-1}$ (Rao and Molina, 2015), and $g_2(\theta)$ accounts for the variability of β.

The last term in the equation (1.10) can be approximated by Taylor series (Kackar and Harville, 1984). Prasad and Rao (1990) provided the following, for large m,

$$E(\hat{\mu} - \tilde{\mu})^2 = g_3(\theta) + o(m^{-1}), \tag{1.13}$$

where $g_3(\theta)$ is defined as

$$g_3(\theta) = \text{tr}\Big\{\frac{\partial b^T}{\partial \theta} V \Big(\frac{\partial b^T}{\partial \theta}\Big) V_a\Big\} \tag{1.14}$$

with the asymptotic covariance matrix of $\hat{\theta}$ as V_a. Noted that $g_2(\theta)$ and $g_3(\theta)$ are $O(m^{-1})$, and $g_1(\theta)$ is $O(1)$ for large m. Hence, the MSPE of the EBLUP of small area means lead to a second-order approximation as

$$\text{MSPE}(\hat{\mu}) = g_1(\theta) + g_2(\theta) + g_3(\theta), \tag{1.15}$$

where the approximation is accurate to the term $o(m^{-1})$, i.e., the terms which is neglected as m, the number of small areas, goes to infinity (Rao and Molina, 2015).

In general, the MSPE of the EBLUP estimator $\hat{\mu}$ (1.3), can be written as

$$\text{MSPE}(\hat{\mu}) = E(\hat{\mu} - \mu)^2 = \text{MSPE}(\tilde{\mu}) + E(\tilde{\mu} - \hat{\mu})^2 + 2E(\tilde{\mu} - \hat{\mu})(\tilde{\mu} - \mu), \tag{1.16}$$

where $\tilde{\mu} = \tilde{\mu}(\theta)$, BLUP of μ, and $\text{MSPE}(\tilde{\mu}) = E(\tilde{\mu} - \mu)^2$. Under the normality assumption of random effects v, and random errors ξ defined in model (1.1), the cross-product term is zero on condition that $\hat{\theta}$ is translation invariant i.e., $\hat{\theta}$ remains unchanged when y is changed to $-y$ or to $y - xa$ for all y and a.

Using the equations from (1.11) - (1.14), the MSPE of the EBLUP of $\hat{\mu}_i$ under the area-level model can be written as

$$\text{MSPE}(\hat{\mu}_i) = g_{1i}(\sigma_v^2) + g_{2i}(\sigma_v^2) + g_{3i}(\sigma_v^2) + o(m^{-1}),$$

here

$$g_{1i}(\sigma_v^2) = \gamma_{i1}\sigma_\xi^2 \quad \text{where} \quad \gamma_{i1} = \sigma_v^2/(\sigma_\xi^2 + \sigma_v^2),$$

$$g_{2i}(\sigma_v^2) = (1 - \gamma_{i1})^2 x_i^{'}[\sum_{i=1}^{m} x_i x_i^{'}/(\sigma_\xi^2 + \sigma_v^2)]^{-1} x_i,$$

$$g_{3i}(\sigma_v^2) = (\sigma_\xi^2)^2(\sigma_\xi^2 + \sigma_v^2)^{-3}V(\hat{\sigma}_v^2),$$

where $V(\hat{\sigma}_v^2)$ is asymptotic variance of $\hat{\sigma}_v^2$ which depends on the method of estimation used for σ_v^2.

Similarly, the MSPE of the EBLUP of $\hat{\mu}_i$ under the unit-level model is given by

$$\text{MSPE}(\hat{\mu}_i) = g_{1i}(\theta) + g_{2i}(\theta) + g_{3i}(\theta) + o(m^{-1}),$$

here

$$g_{1i}(\theta) = \gamma_{i2}\sigma_\xi^2/n_i \quad \text{where} \quad \gamma_{i2} = \sigma_v^2/(\sigma_v^2 + \sigma_\xi^2/n_i),$$

$$g_{2i}(\theta) = (\bar{x}_i - \gamma_{i2}\bar{x}_i)(\sum_{i=1}^{m} x_i^{'}V_i^{-1}x_i)^{-1}(\bar{x}_i - \gamma_{i2}\bar{x}_i),$$

$$g_{3i}(\theta) = n_i^{-2}(\sigma_v^2 + \sigma_\xi^2/n_i)^{-3}[\sigma_\xi^4 V_{vv}(\hat{\sigma}_v^2) + \sigma_v^4 V_{\xi\xi}(\hat{\sigma}_\xi^2) - 2\sigma_\xi^2\sigma_v^2 Cov_\theta(\hat{\sigma}_\xi^2, \hat{\sigma}_v^2)],$$

where V_{vv} and $V_{\xi\xi}$ are the asymptotic variance of $\hat{\sigma}_v^2$ and $\hat{\sigma}_\xi^2$, and $Cov_\theta(\hat{\sigma}_\xi^2, \hat{\sigma}_v^2)$ is the asymptotic covariance of $\hat{\sigma}_\xi^2$ and $\hat{\sigma}_v^2$.

Estimation of MSPE of the EBLUP

In reality, an estimator of MSPE is required as the MSPE depends on the unknown parameter vector θ. A naive approach (Rao and Molina, 2015) approximates MSPE$[\mu(\hat{\theta})]$ by MSPE$[\mu(\theta)]$ and then replace $\hat{\theta}$ for θ. Hence, the resulting MSPE estimator is stated as

$$\text{mspe}_{\text{N}}[\mu(\hat{\theta})] = g_1(\hat{\theta}) + g_2(\hat{\theta}),$$

which is the first-order unbiased. Another method for the MSPE estimation is obtained by substituting $\hat{\theta}$ for θ in the MSPE approximation equation defined in (1.15),

$$\text{mspe}_1(\hat{\mu}) = g_1(\hat{\theta}) + g_2(\hat{\theta}) + g_3(\hat{\theta}). \tag{1.17}$$

In this, $g_2(\hat{\theta})$ and $g_3(\hat{\theta})$ have desired order of approximation but $g_1(\hat{\theta})$ is not the correct estimator of $g_1(\theta)$ (Rao and Molina, 2015). In order to correct this, Prasad and Rao (1990) proposed another estimator of MSPE, given as

$$\text{mspe}(\hat{\mu}) = g_1(\hat{\theta}) + g_2(\hat{\theta}) + 2g_3(\hat{\theta}), \tag{1.18}$$

which is second-order unbiased provided that the variance components are estimated unbiasedly (e.g., REML); otherwise a correct term needs to be added in (1.18). Following Prasad and Rao (1990), which proposed mspe estimator in (1.18) for the linear mixed model (1.1) is also named as Prasad-Rao estimator under the assumption $E(\hat{\theta}) = \theta$. Further, the approach was extended by Datta and Lahiri (2000) including ML and REML estimators.

The second-order unbiased estimate of the MSPE under area-level model when $\hat{\sigma}_v^2$ is obtained by REML may be written as

$$\text{mspe}(\hat{\theta}_i) = g_{1i}(\hat{\sigma}_v^2) + g_{2i}(\hat{\sigma}_v^2) + 2g_{3i}(\hat{\sigma}_v^2). \tag{1.19}$$

Whereas, when $\hat{\sigma}_v^2$ is obtained using MLE method, an extra term of bias in $\hat{\sigma}_v^2$ must be added in the above equation (1.19).

Similarly, the second-oder unbiased estimate of the MSPE under unit-level model is defined as

$$\text{mspe}(\hat{\theta}_i) = g_{1i}(\hat{\sigma}_v^2, \hat{\sigma}_\xi^2) + g_{2i}(\hat{\sigma}_v^2, \hat{\sigma}_\xi^2) + 2g_{3i}(\hat{\sigma}_v^2, \hat{\sigma}_\xi^2), \tag{1.20}$$

where $\hat{\sigma}_v^2$ and $\hat{\sigma}_\xi^2$ are obtained using the REML. Similar to the area-level model, the bias terms should be added to the above equation (1.20) if one uses a bias estimation approach such as the MLE.

1.2 A Brief Overview of Copulas

Copulas are a popular multivariate modeling tool in many fields, such as finance and insurance, where the multivariate normality is questionable and the interest lies in multivariate dependence.

Copulas are functions that couple marginal distributions of random variables to form a multivariate distribution. Let $\mathbb{Z} = (Z_1, \ldots, Z_d)$ be a d-dimensional random vector. Then, by Sklar's theorem (Sklar, 1959), the joint distribution F of $\mathbb{Z}$ can be written as

$$F(z_1, \ldots, z_d) = C\{F_1(z_1), \ldots, F_d(z_d)\}, \tag{1.21}$$

where C is a copula function of $\mathbb{Z}$ and F_i is the marginal distribution functions of Z_i, for $i = 1, \ldots, d$. For the case when all the variables are continuous, C is unique and captures the complete dependence. In this case, i.e., when F and C are differentiable, the d-dimensional joint density function is given by

$$f(z_1,\ldots,z_d) = c(F_1(z_1),\ldots,F_d(z_d)) \text{ x } \prod_{i=1}^{d} f_i(z_i), \tag{1.22}$$

where c is the copula density defined as

$$c_\alpha(z_1,\ldots,z_d) = \frac{\partial^{i:d} C(z_1,\ldots,z_d)}{\partial_{z_1},\ldots,\partial_{z_d}}.$$

From Equations (1.21) and (1.22), the joint distribution is decomposed into marginal distributions and dependence characteristics, which can be modeled separately. This gives a considerable flexibility in model fitting as there is no imposed restriction on the marginal distributions. Furthermore, one can construct many new multivariate distributions by choosing different marginal distributions and a copula function.

Many parametric copula families have been proposed in the literature which exhibit very different dependence patterns. Most commonly used parametric copulas are elliptical (Gaussian and Student t-copulas) and Archimedean copulas (Clayton, Gumbel and Frank copulas). Below, we briefly introduce the scale-free dependence measure Kendall's tau (τ) and the parametric copulas used in this thesis.

Kendall's Tau

Consider two random variables X and Y with continuous marginal distributions F_1 and F_2, respectively, and joint distribution $F(x,y) = C(F_1(x), F_2(y))$. Kendall's tau of X and Y is defined as

$$\tau(X,Y) = \Pr[(X_1 - X_2)(Y_1 - Y_2) \geq 0] - \Pr[(X_1 - X_2)(Y_1 - Y_2) \leq 0], \tag{1.23}$$

where (X_1, Y_1) and (X_2, Y_2) are two independent pairs of random variables from F. The equation (1.23) gives the difference between the probability of concordance and the probability of discordance. It can also be expressed in terms of copula C as follows:

$$\tau(X,Y) = 4 \int_0^1 \int_0^1 C(z_1, z_2) dC(z_1, z_2) - 1.$$

See Nelsen (2006) for details.

Some Copula Families

In this thesis, we consider four copula families: Gaussian, Clayton, Frank and Gumbel. The contour plots of these copulas with standard normal margins are displayed in Figure 1.1 for $\tau = 0.2, 0.5$, and 0.8.

- The **Gaussian (Normal) copula** takes the form

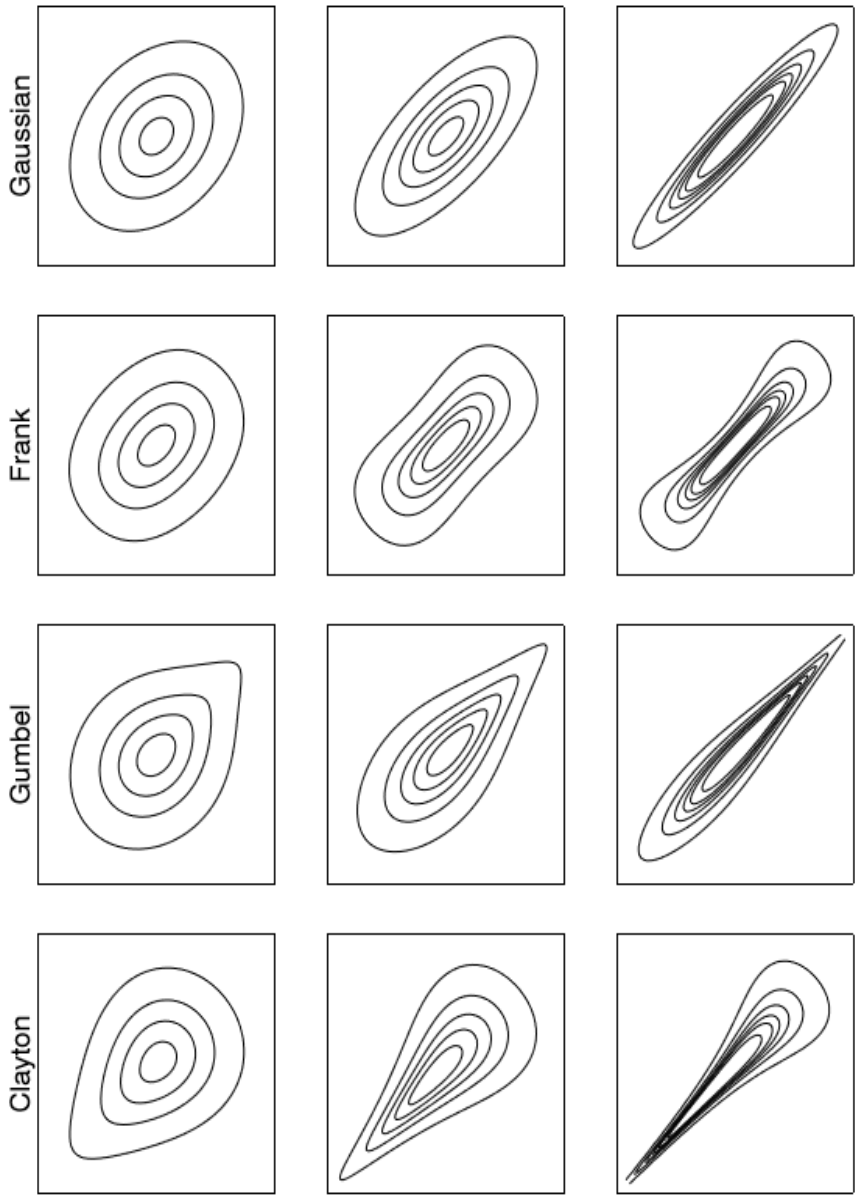
$$C_\rho(u,v) = \int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(v)} \frac{1}{2\pi(1-\rho^2)^{1/2}} exp\Big\{\frac{-(s^2 - 2\rho st + t^2)}{2(1-\rho^2)}\Big\} dsdt$$

$$= \Phi_G(\Phi^{-1}(u), \Phi^{-1}(v); \rho), \qquad -1 \leq \rho \leq 1,$$

  here Φ is the cdf of standard normal and $\Phi_G(u,v)$ is the standard bivariate normal distribution with correlation parameter ρ.

  For the Gaussian copula, Kendall's tau is given by

Figure 1.1: Contour plots for Gaussian, Frank, Gumbel and Clayton copula families with standard normal margins when $\tau = 0.2$ (left panel), 0.5 (middle panel), and 0.8 (right panel).



$$\tau = \frac{2}{\pi}\arcsin(\rho).$$

- The **Clayton copula** exhibits strong lower tail and relatively weak upper tail dependence. It takes the form

$$C_\alpha(u,v;\alpha) = (u^{-\alpha} + v^{\alpha} - 1)^{-1/\alpha}, \qquad \alpha \geq 0, \tag{1.24}$$

with the dependence parameter α. As α tends to 0, the copula approaches to the independent copula.

Kendall's tau for the Clayton copula is given by

$$\tau = \frac{\alpha}{\alpha + 2}.$$

- The **Frank copula** exhibits no tail dependence and can be used to represent both positive and negative dependence. It takes the form

$$C(u,v;\alpha) = -\frac{1}{\alpha}\ln\left\{1 + \frac{(e^{\alpha u} - 1)(e^{-\alpha v} - 1)}{e^{-\alpha} - 1}\right\}, \qquad \alpha \in \Re\backslash\{0\}. \tag{1.25}$$

Kendall's tau for the Frank copula is

$$\tau = 1 + \frac{4}{\alpha}\{D_1(\alpha) - D_2(\alpha)\}$$

where $D_j(\alpha) = \dfrac{j}{\alpha^j}\displaystyle\int_0^\alpha \frac{t^j}{e^t - 1}\,dt$ is the Debye function.

- The **Gumbel copula** takes the form

$$C(u,v;\alpha) = \exp\Big(-(u^{-\alpha} + v^{-\alpha})^{1/\alpha}\Big), \qquad \alpha \in [1,\infty). \tag{1.26}$$

In contrast to the Clayton copula, the Gumbel copula has strong upper tail and weak lower tail dependence.

Kendall's tau of the Gumbel copula is given by

$$\tau = 1 - \frac{1}{\alpha}.$$

Further details on copulas and their properties can be found in Nelsen (2006).

1.3 Thesis Outline

The structure of this thesis is as follows: Chapter 2 describes the SAE model based on multivariate exchangeable copulas and outlines the proposed methods. Chapter 3 provides extensive simulation studies to evaluate performance of the proposed methods. Chapter 4 contains a real data application on Landsat data. Chapter 5 summarizes our main findings and outlines future work.

Chapter 2

Small Area Estimation using Copulas

In this chapter, we briefly describe the small area model, where the joint error distribution is defined using the multivariate exchangeable copulas, and outline the estimation approaches used for small area predictors. After introducing the notations and the model in Section 2.1, small area predictors are defined in Section 2.2. Then, we describe the two-stage estimation procedure to estimate the model parameters in Section 2.3 and present the empirical BUP (EBUP) of small area means in Section 2.4. We derive the MSPE of the EBUP of small area means in Section 2.5. We also outline a bootstrap method under both parametric and semi-parametric methods for the estimation of MSPE of the EBUP of small area means in Section 2.6.

2.1 Multivariate Exchangeable Copula Model

Consider a population of m small areas of sizes $N_1, N_2, \ldots, N_m$. Let Y be the variable of interest and x be a corresponding p-dimensional vector of auxiliary variable. The population is defined by the linear model for Y given x,

$$Y_{ij} = x_{ij}^T \beta + \varepsilon_{ij}, \qquad i = 1, \ldots, m; j = 1, \ldots, N_i, \tag{2.1}$$

where the error terms ε_{ij} have the marginal distribution F_ε with zero mean and finite variance σ^2. Suppose the marginal error distribution F_ε is parametrized by δ, and the joint error distribution of each area is expressed in terms of a parametric exchangeable copula $C_{\alpha,1:N_i}$ as

$$F_{\alpha,N_i}(\varepsilon_{i,1}, \ldots, \varepsilon_{i,N_i}) = C_{\alpha,N_i}\{F_\varepsilon(\varepsilon_{i,1};\delta), \ldots, F_\varepsilon(\varepsilon_{i,N_i};\delta)\}. \tag{2.2}$$

Here α is the copula parameter, which quantifies the within-area dependence. Furthermore, it is connected with the intra-class correlation (ICC) of $\varepsilon_{ij} = \nu_i + \xi_{ij}$ given by $\rho = \sigma_\nu^2/(\sigma_\nu^2 + \sigma_\xi^2)$, where σ_ν^2 and σ_ξ^2 are the variances of ν_i and ξ_{ij}, respectively. Under the independence of these terms, the variance of ε_{ij} is given by $\sigma^2 = \sigma_\nu^2 + \sigma_\xi^2$.

We assume the following two conditions for the joint error distribution.

Assumption 1 *Exchangeable property: the value of copula is invariant under permutation of its arguments, i.e.,*

$$C(u_1, \ldots, u_d) = C(u_{\pi(1)}, \ldots, u_{\pi(d)}); u_1, \ldots, u_d \in [0,1].$$

Assumption 2 *Invariance property for dimensions: the joint cumulative function of say* $\varepsilon_1,\ldots,\varepsilon_{N_i}$ *is defined as* $F_{\alpha,\delta}(\varepsilon_1,\ldots,\varepsilon_{N_i}) = C_\alpha\{F_\varepsilon(\varepsilon_1),\ldots,F_\varepsilon(\varepsilon_{N_i})\}$ *which satisfies the invariance property, i.e.,*

$$F_{\alpha,\delta}(\infty,\ldots,\infty,\varepsilon_1,\ldots,\varepsilon_{N_i},\infty,\ldots,\infty) = F_{\alpha,\delta}(\varepsilon_1,\ldots,\varepsilon_{N_i}).$$

The model (2.1) where regression error is expressed in terms of an exchangeable copula was introduced by Rivest et al. (2016). This model provides flexibility to practitioners in modeling of the errors when their joint distribution deviates from a multivariate normal distribution.

2.1.1 Notation

Suppose that a random sample denoted by s_i of size n_i is drawn using simple random sample for each small area i for $i = 1,\ldots,m$ with known population size N_i. The sample $\{(Y_{ij}, x_{ij}); i = 1,\ldots,m, j = 1,\ldots,n_i\}$ is assumed to obey the model defined in equation (2.1), i.e., there is no sample selection bias. The non-sampled units within the i^{th} small area are denoted by r_i. Also, the mean of errors from the sampled data for each i^{th} small area is $\bar{\varepsilon}_{is} = \sum_{j=1}^{n_i} \varepsilon_{ij}/n_i$. The true small area mean is expressed as $\bar{Y}_{iU} = \sum_{j=1}^{N_i} Y_{ij}/N_i$ and the mean of the auxiliary variable for each area is $\bar{x}_{iU} = \sum_{j=1}^{N_i} x_{ij}/N_i$.

2.2 Small Area Predictors

The model (2.1) is a general case of the linear mixed model (1.1), where the multivariate exchangeable copula model is used to characterize the error distribution within each small area. Under this model with known parameters, one can obtain the best linear unbiased predictor (BLUP) of the small area means using a linear mixed model and the best unbiased predictor (BUP) using the conditional distribution of unobserved errors given the ones from the sampled. Here the "best" term stands for the lowest variance and "unbiased" refers to the true value of parameter being equivalent to the estimated value. In the following, we introduce the BLUP and BUP. We also define the MSPE under these estimators.

2.2.1 BLUP of Small Area Means

In the literature, Henderson (1975) developed the BLUP for mixed models. The most important property of this method is the ability to predict linear combination of fixed and random effects. Likewise, Rivest et al. (2016) derived the BLUP when the model is designed using exchangeable copula as

$$\bar{Y}_{iU}^{BLUP} = \bar{x}_{iU}^{T}\beta + \frac{n_i}{N_i}\bar{\varepsilon}_{is} + \frac{N_i - n_i}{N_i}\frac{n_i\rho\bar{\varepsilon}_{is}}{1+(n_i-1)\rho},$$

where $\bar{Y}_{iU}$ is the true small area mean, $\bar{x}_{iU}$ is the population mean of auxiliary variable, and $\bar{\varepsilon}_{is}$ is the sample mean of errors for the i^{th} small area. Also, ρ is the ICC as defined in Section 2.1.

The MSPE of the BLUP for the i^{th} small area is given by

$$\text{MSPE}(\bar{Y}_{iU}^{BLUP}) = \frac{\sigma^2\rho(1-\rho)}{1+(n_i-1)\rho} + O(n_i/N_i). \tag{2.3}$$

The $\text{MSPE}(\bar{Y}_{iU}^{BLUP})$ stands for the g_{1i} term defined in MSPE of the BLUP of small area means (Prasad and Rao, 1990; Rao and Molina, 2015).

2.2.2 BUP of Small Area Means

The BUP of small area means for the model (2.1) is given as

$$\bar{Y}_{iU}^{BUP} = \bar{x}_{iU}^T\beta + \frac{n_i}{N_i}\bar{\varepsilon}_{is} + \frac{N_i-n_i}{N_i}E(\varepsilon_{ia}|\varepsilon_{ij}; j \in s_i), \tag{2.4}$$

where a is an un-sampled unit of i^{th} small area. The MSPE of the BUP is given by

$$\begin{aligned}\text{MSPE}(\bar{Y}_{iU}^{BUP}) &= E\left(\text{Cov}\{\varepsilon_{ia},\varepsilon_{ib}|\varepsilon_{ij}, j \in s_i\}\right) + O(n_i/N_i)\\ &= \sigma^2\rho - E\{E(\varepsilon_{ia}|\varepsilon_{ij}, j \in s_i)^2\} + O(n_i/N_i),\end{aligned} \tag{2.5}$$

where (a,b) stands for the units from r_i.

The BUP is derived in the case where the model parameters are known. In practice, the parameters are unknown and have to be estimated using the sample data.

2.3 Model Parameters Estimation

Given data $\{(Y_{ij}, x_{ij}); i = 1,\ldots,m, j = 1,\ldots,n_i\}$, one can fit the model in (2.1) through a two-stage estimation procedure, by first estimating the marginal error distribution and then the copula parameter.

2.3.1 Model Fitting and Conversion of Residuals to Copula Data

Assuming a simple linear regression in model (2.1), to estimate F_ε, first the regression vector β is estimated under the linear regression model and the residuals $e_{ij}(= Y_{ij} - x_{ij}^T\hat{\beta})$ are obtained. The residuals e_{ij} are then converted to the uniform scale using the marginal error distribution.

Parametric method

If a parametric model $F_\varepsilon(\cdot;\delta)$ is available, the marginal error distribution is estimated parametrically using $\hat{F}_\varepsilon = \hat{F}_\varepsilon(e_{ij};\hat{\delta})$, where $\hat{\delta}$ is the maximum likelihood estimate of δ. The estimated marginal error distribution is then used to obtain pseudo observations, i.e., the copula data

$$(\hat{u}_{i,1},\ldots,\hat{u}_{i,n_i}) = \left(\hat{F}_\varepsilon(e_{i,1}),\ldots,\hat{F}_\varepsilon(e_{i,ni})\right).$$

Semi-parametric method

In the absence of a suitable parametric model, the empirical cumulative distribution function

$$\tilde{F}(e) = \frac{1}{n+1}\sum_{i=1}^{m}\sum_{j\in s_i} 1_{\{\varepsilon_{ij}\le e\}}$$

is used to estimate the marginal error distribution where $1/(n+1)$ is used instead of $1/n$ to avoid evaluation of the copula density at boundary $[0,1]^n$ (Genest et al., 1995). The pseudo-observations are defined as

$$(\tilde{u}_{i,1},\ldots,\tilde{u}_{i,n_i}) = \left(\tilde{F}_\varepsilon(e_{i,1}),\ldots,\tilde{F}_\varepsilon(e_{i,n_i})\right), \tag{2.6}$$

This approach is often referred to as rank transformation.

2.3.2 Copula Parameter Estimation

After fitting the marginal error distribution, one can estimate the copula parameter α of the multivariate exchangeable copula C_α using the maximum pseudo copula log-likelihood.

Parametric method

For the fully parametric two-stage estimation, the pseudo copula log-likelihood is given by

$$\ell(\alpha) = \sum_i^m \ln[c_\alpha(\hat{u}_{i1},\ldots,\hat{u}_{in_i}|\alpha)], \tag{2.7}$$

where c_α is the copula density. Then, the estimated copula parameter $\hat{\alpha}$ is the one that maximizes the pseudo log-likelihood (2.7).

Semi-parametric method

For the semi-parametric approach, the margins are estimated by the empirical cdfs, which yields the pseudo copula log-likelihood as $\ell(\alpha)$ defined as

$$\ell(\alpha) = \sum_i^m \ln[c_\alpha(\tilde{u}_{i1},\ldots,\tilde{u}_{in_i}|\alpha)]. \tag{2.8}$$

The estimated copula parameter $\tilde{\alpha}$ is obtained by maximizing the pseudo log-likelihood (2.8).

2.4 Empirical BUP of Small Area Means

The BUP is constructed with a conditional distribution of un-sampled error term given the sampled error terms and defined as

$$E(\varepsilon_{ia} \mid \varepsilon_{ij}, j \in s_i) = \int_{-\infty}^{\infty} z w_{1i}\{F_\varepsilon(z), F_\varepsilon(\varepsilon_{ij}) : j \in s_i, \alpha\} dF_\varepsilon(z), \tag{2.9}$$

where z is the unobserved error term with marginal distribution defined as $F_\varepsilon(z)$. The weight function w_{1i} is the conditional density, when the copula family and margins are known, which is defined as

$$w_{1i}\{v, F_\varepsilon(\varepsilon_{ij}) : j \in s_i, \alpha\} = \frac{c_{\alpha, n_i+1}\{v, F_\varepsilon(\varepsilon_{i,j_1}), ..., F_\varepsilon(\varepsilon_{i,j_{n_i}})\}}{c_{\alpha, n_i}\{F_\varepsilon(\varepsilon_{i,j_1}), ..., F_\varepsilon(\varepsilon_{i,j_{n_i}})\}}. \tag{2.10}$$

Since the model parameters are unknown in practice, the estimates of the regression coefficient β, marginal error distribution F_ε, and the copula parameter α are used to obtain an empirical predictor of (2.9), i.e.,

$$\hat{e}_i = \frac{\sum_{i_1=1}^{m} \sum_{j_1 \in s_{i_1}} e_{i_1 j,1} w_{1i}\{\hat{F}_\varepsilon(e_{i_1 j,1}), \hat{F}_\varepsilon(e_{i,j}) : j \in s_i, \hat{\alpha}\}}{\sum_{i_1=1}^{m} \sum_{j_1 \in s_{i_1}} w_{1i}\{\hat{F}_\varepsilon(e_{i_1 j,1}), \hat{F}_\varepsilon(e_{i,j}) : j \in s_i, \hat{\alpha}\}}. \tag{2.11}$$

Then, the EBUP of $\bar{Y}_{iU}$ is defined as

$$\hat{\bar{Y}}_{iU} = \bar{x}_{iU}^T \hat{\beta} + \frac{n_i}{N_i} \bar{e}_{is} + \frac{N_i - n_i}{N_i} \hat{e}_i. \tag{2.12}$$

For the semi-parametric case, to get the EBUP of small area means (called $\tilde{\bar{Y}}_{iU}$), one can replace the $\hat{e}_i$ and $\hat{F}_\varepsilon(e_{i,j})$ in (2.12) by $\tilde{e}_i$ and $\tilde{F}_\varepsilon(e_{i,j})$.

2.5 Mean Square Prediction Error

The prediction error of the EBUP of small area means can be decomposed as

$$\hat{\bar{Y}}_{iU} - \bar{Y}_{iU} = (\bar{Y}_{iU}^{BUP} - \bar{Y}_{iU}) + (\hat{\bar{Y}}_{iU} - \bar{Y}_{iU}^{BUP}). \tag{2.13}$$

We can then write

$$\begin{aligned} \text{MSPE}(\hat{\bar{Y}}_{iU}) &= E\{(\hat{\bar{Y}}_{iU} - \bar{Y}_{iU})^2\} \\ &= E\{(\bar{Y}_{iU}^{BUP} - \bar{Y}_{iU})^2\} + E\{(\bar{Y}_{iU}^{BUP} - \hat{\bar{Y}}_{iU})^2\} \\ &\qquad + 2E\{(\bar{Y}_{iU}^{BUP} - \bar{Y}_{iU})(\bar{Y}_{iU}^{BUP} - \hat{\bar{Y}}_{iU})\} \\ &= M_{1i} \quad + \quad M_{2i} \quad + \quad M_{3i}, \end{aligned} \tag{2.14}$$

where M_{1i} is $\text{MSPE}(\bar{Y}_{iU}^{BUP})$ given in (2.5). Note that the $\text{MSPE}(\hat{\bar{Y}}_{iU})$ is a function of unknown parameters. It is then necessary to provide a nearly-unbiased estimator of the $\text{MSPE}(\hat{\bar{Y}}_{iU})$.

Under the jackknife approach (Rivest et al., 2016), the estimation of MSPE of the EBUP of small area means can only capture the variations of the first two terms of (2.14). Hence, the estimation of MSPE can lead to significant bias by ignoring the variation of the cross-product term in (2.14). In light of this, the cross-product term cannot be ignored and hence it is necessary to account for this term in the estimation procedure of MSPE of the EBUP of small area means.

2.6 MSPE Estimation using Bootstrap Method

For the complex small area models such as (2.1), the closed analytical expression of MSPE is not possible to obtain. In a situation when explicit formula for MSPE cannot be obtained, it is even a harder task to estimate the MSPE. One way to tackle this issue is to use re-sampling methods for the estimation of the MSPE of the EBUP of small area means. In order to properly capture the all variations in the MSPE of the EBUP of small area means including the cross-product term i.e., M_{3i} in (2.14), we propose a bootstrap approach.

Bootstrap methods have been previously employed in SAE (Hall and Maiti, 2006; Torabi, 2012) but these are not directly applicable under the exchangeable copula model. Furthermore, a careful treatment is needed when the two-stage semi-parametric estimation is employed. Below, we first outline the parametric bootstrap method under the exchangeable copula model in Algorithm 1 and then present a corresponding semi-parametric bootstrap approach in Algorithm 2.

Both algorithms start with the generation of copula data using the fitted copula $C_{\hat{\alpha}}$ to define bootstrap populations. In particular, under the parametric approach, the inverse cdf of the fitted marginal distribution $\hat{F}_{\varepsilon}(\cdot;\hat{\delta})$ is used to obtain the error terms of the bootstrap population. On the other hand, under the semi-parametric approach, obtaining the error terms from the generated copula data is not straightforward due to the absence of a generative marginal error distribution. We address this challenge using a quantile mapping approach that maps the generated copula data to the empirical quantiles of the residuals in the sample.

Using the fitted linear models with the corresponding bootstrap error terms, we define bootstrap populations under the proposed parametric and semi-parametric approaches. The remaining steps of the algorithms are very similar and involve simple random samples from the bootstrap populations, performing the estimation the same way it is done and getting the EBUP of the small area means.

Algorithm 1 Parametric bootstrap method for the estimation of MSPE of the EBUP of small area means

Given the estimated model parameters $(\hat{\beta}, \hat{\delta}, \hat{\alpha})$ from the dataset $\{(Y_{ij}, x_{ij}); i = 1,\ldots,m; j = 1,\ldots,n_i\}$:

Bootstrap population:

1: Generate copula data $u_{ij}^{(b)}$, $(i = 1,\ldots,m; j = 1,\ldots,N_i; b = 1,\ldots,B)$ from $C_{\hat{\alpha}}$, where B is the number of bootstrap runs.

2: Use the inverse cdf method to obtain bootstrap error terms $\varepsilon_{ij}^{(b)} = F_{\varepsilon}^{-1}(\hat{u}_{ij}^{(b)}; \hat{\delta})$, $(i = 1,\ldots,m; j = 1,\ldots,N_i; b = 1,\ldots,B)$.

3: Obtain bootstrap population $Y_{ij}^{(b)} = x_{ij}^T\hat{\beta} + \varepsilon_{ij}^{(b)}, (i = 1,\ldots,m; j = 1,\ldots,N_i; b = 1,\ldots,B)$.

4: Compute the mean of the bootstrap population as $\bar{Y}_{iU}^{(b)} = N_i^{-1}\sum_{j=1}^{N_i} Y_{ij}^{(b)}, (i = 1,\ldots,m)$.

Bootstrap sample:

5: Get a bootstrap sample from the bootstrap population using simple random sample without replacement.

6: Perform the parametric estimation, using the bootstrap sample $\{(Y_{ij}^{(b)}, x_{ij}); i = 1,...,m; j = 1,...,n_i\}$ to get the bootstrap estimates $(\hat{\beta}^{(b)}, \hat{\delta}^{(b)}, \hat{\alpha}^{(b)}, \hat{e}_i^{(b)}), (b = 1,\ldots,B)$.

7: The bootstrap EBUP of small area means is calculated as

$$\hat{\bar{Y}}_{iU}^{(b)} = \bar{x}_{iU}^T\,\hat{\beta}^{(b)} + \frac{n_i}{N_i}\bar{e}_{is}^{(b)} + \frac{N_i - n_i}{N_i}\hat{e}_i{}^{(b)}.$$

8: After obtaining the bootstrap EBUP for a large number of bootstrap samples, B, compute the parametric bootstrap MSPE estimation by

$$\text{mspe}_{\text{boot}}(\hat{\bar{Y}}_{iU}) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\bar{Y}}_{iU}^{(b)} - \bar{Y}_{iU}^{(b)})^2. \tag{2.15}$$

Algorithm 2 Semi-parametric bootstrap method for the estimation of MSPE of the EBUP of small area means

Given the estimated model parameters $(\tilde{\beta}, \tilde{\delta}, \tilde{\alpha})$ from the dataset $\{(Y_{ij}, x_{ij}); i = 1,\ldots,m; j = 1,\ldots,n_i\}$:

Bootstrap population:

1: Generate copula data $u_{ij}^{(b)}$, $(i = 1,\ldots,m; j = 1,\ldots,N_i; b = 1,\ldots,B)$ from $C_{\tilde{\alpha}}$.

2: Proceed with quantile mapping of $u_{ij}^{(b)}$ to $\tilde{u}_{ij}$ and get $\varepsilon_{ij}^{*(b)}$, $(i = 1,\ldots,m; j = 1,\ldots,N_i; b = 1,\ldots,B)$.

3: Obtain bootstrap population $Y_{ij}^{*(b)} = x_{ij}^T\tilde{\beta} + \varepsilon_{ij}^{*(b)}, (i = 1,\ldots,m; j = 1,\ldots,N_i; b = 1,\ldots,B)$.

4: Compute the bootstrap population mean as $\bar{Y}_{iU}^{*(b)} = N_i^{-1}\sum_{j=1}^{N_i} Y_{ij}^{*(b)}$.

Bootstrap sample:

5: Get a bootstrap sample from the bootstrap population using simple random sample without replacement.

6: Perform the semi-parametric estimation, using the bootstrap sample $\{(Y_{ij}^{*(b)}, x_{ij}); i = 1,...,m; j = 1,...,n_i\}$ to get the bootstrap estimates $(\tilde{\beta}^{*(b)}, \tilde{\delta}^{*(b)}, \tilde{\alpha}^{*(b)}, \tilde{e}_i^{*(b)})$.

7: The bootstrap EBUP of small area means is calculated as

$$\tilde{\bar{Y}}_{iU}^{*(b)} = \bar{x}_{iU}^T\,\tilde{\beta}^{*(b)} + \frac{n_i}{N_i}\tilde{e}_{is}^{*(b)} + \frac{N_i - n_i}{N_i}\tilde{e}_i{}^{*(b)}.$$

8: After obtaining the bootstrap EBUP for a large number of bootstrap samples, B, compute the semi-parametric bootstrap MSPE estimation by

$$\text{mspe}_{\text{boot}}(\tilde{\bar{Y}}_{iU}) = \frac{1}{B}\sum_{b=1}^{B}(\tilde{\bar{Y}}_{iU}^{*(b)} - \bar{Y}_{iU}^{*(b)})^2.$$

Chapter 3

Simulation Study

This chapter evaluates the performance of the proposed parametric and semi-parametric methods to that of the RVB method under various settings using Gaussian, Clayton, Frank, and Gumbel copulas.

3.1 Simulation Setting

Following Rivest et al. (2016), we generated the responses Y_{ij} for the population units from the simple linear regression model (2.1) with β_0 and β_1 equal to 1. We considered a population with $m = 20$ and 40 small areas, where each area consists of $N_i = 200$ units. The population data were generated as follows: we first generated the copula data using the exchangeable copula in (2.1). For the copula family, we considered the Gaussian, Clayton, Frank, and Gumbel copulas with ICC of $\rho = 0.5$. For the marginal error distribution, we considered standard normal and skewed normal distributions with $\mu_\varepsilon = 0$ and $\sigma^2_\varepsilon = 1$. Here, the skewed

normal distribution, $SKN(\zeta, \omega, \gamma)$, is defined in terms of the location (ζ), scale (ω), and skewness (γ) parameters, with conversions from the mean μ_ε, variance σ^2_ε, and skewness parameter γ by

$$\mu_\varepsilon = \zeta + \omega\kappa\sqrt{\frac{2}{\pi}}, \tag{3.1}$$

where $\kappa = \frac{\gamma}{\sqrt{1+\gamma^2}}$ and ω can be deduced as

$$\sigma^2_\varepsilon = \omega^2(1 - \frac{2\kappa^2}{\pi}).$$

Then other parameter, ζ, can be obtained using (3.1), after plug-in κ and ω. For a given $\mu_\varepsilon = 0$, $\sigma^2_\varepsilon = 1$, and $\gamma = 10$, we obtained $\zeta = 1.31$ and $\omega = 1.65$. We generated R = 500 independent sets of $\{e^{(r)}_{ij}; i = 1,\ldots,m; j = 1,\ldots,N_i; r = 1,\ldots,R\}$, $u^{(b)}_{ij}$ to $\tilde{u}_{ij}$ were generated from $N(1,1)$ and kept fixed across the Monte-Carlo replicates. Consequently, R = 500 population datasets $\{Y^{(r)}_{ij}; i = 1,\ldots,m; j = 1,\ldots,N_i; r = 1,\ldots,R\}$ were obtained using the model (2.1): $Y_{ij} = \beta_0 + \beta_1 x_{ij} + e_{ij}$. The corresponding population mean for the r^{th} simulation run ($r = 1,\ldots,R$) and i^{th} area was given by

$$\bar{Y}^{(r)}_{iU} = N_i^{-1} \sum_{j=1}^{N_i} Y^{(r)}_{ij}.$$

Using simple random sample, sampled data $\{Y^{(r)}_{ij}; i = 1,\ldots,m; j = 1,\ldots,n_i; r = 1,\ldots,R\}$ were drawn for each area of size $n_i = 4$.

Using the sampled data $\{(Y_{ij}^{(r)}, x_{ij}); i = 1,\ldots,m; j = 1,\ldots,n_i\}$, the model parameters $\beta = (\beta_0, \beta_1)$ were estimated under linear regression, δ and α were then estimated using the log-likelihood framework (see Section 2.3) and the known copula family for the parametric approach. Afterwards, the EBUP of small area means $\hat{\bar{Y}}_{iU}^{(r)}$ (2.12) was obtained using the parametric approach, and similarly the EBUP of small area means $\tilde{\bar{Y}}_{iU}^{(r)}$ was obtained using the semi-parametric approach. Noted that δ was also estimated using log-likelihood framework under the semi-parametric approach, and the estimation of α was proceeded as defined in Section 2.3.

We employed the Algorithm 1 and Algorithm 2 for the estimation of MSPE of small area mean predictors using a bootstrap approach with $B = 100$.

3.1.1 Evaluation of Estimators Performance

The performances of the parametric, semi-parametric, and RVB approaches were assessed using the bias and the empirical MSPE (EMSPE) of small area mean predictors; In the case of parametric approach, these are calculated using

$$\text{Bias}(\hat{\bar{Y}}_{iU}) = \frac{1}{R}\sum_{r=1}^{R}\{\hat{\bar{Y}}_{iU}^{(r)} - \bar{Y}_{iU}^{(r)}\}, \tag{3.2}$$

and the EMSPE of $\hat{\bar{Y}}_{iU}$ as

$$\begin{aligned}
\text{EMSPE}(\hat{\bar{Y}}_{iU}) &= \frac{1}{R}\sum_{r=1}^{R}\{\hat{\bar{Y}}_{iU}^{(r)} - \bar{Y}_{iU}^{(r)}\}^2 \\
&= \frac{1}{R}\sum_{r=1}^{R}\{\bar{Y}_{iU}^{BUP(r)} - \bar{Y}_{iU}^{(r)}\}^2 + \frac{1}{R}\sum_{r=1}^{R}\{\bar{Y}_{iU}^{BUP(r)} - \hat{\bar{Y}}_{iU}^{(r)}\}^2 \\
&\qquad + 2\frac{1}{R}\sum_{r=1}^{R}\{\bar{Y}_{iU}^{BUP(r)} - \hat{\bar{Y}}_{iU}^{(r)}\}\{\bar{Y}_{iU}^{BUP(r)} - \bar{Y}_{iU}^{(r)}\} \\
&= \tilde{M}_{1i} + \tilde{M}_{2i} + \tilde{M}_{3i},
\end{aligned} \tag{3.3}$$

where $\bar{Y}_{iU}^{BUP(r)}$ is the BUP of small area mean for r^{th} simulation run at area i. To evaluate the performance of MSPE estimation of the EBUP of small area means using the three approaches, we define the relative bias (RB) of MSPE estimator, say mspe, where RB of mspe is given by

$$\text{RB}(\text{mspe}_{\text{boot}}(\hat{\bar{Y}}_{iU})) = \frac{\frac{1}{R}\sum_{r=1}^{R}\text{mspe}_{\text{boot}}(\hat{\bar{Y}}_{iU}^{(r)})}{\text{EMSPE}(\hat{\bar{Y}}_{iU})} - 1. \tag{3.4}$$

Note that the bootstrap approach was used for the parametric and semi-parametric methods while the jackknife method was used for the RVB method to get the estimation of MSPE of the EBUP of small area means.

The performance of semi-parametric approach was assessed similarly by replacing $\hat{\bar{Y}}_{iU}$ with $\tilde{\bar{Y}}_{iU}$.

3.1.2 Simulation Experiments

Our objective is to evaluate the estimation performance under four different scenarios:

(i) Correctly specified joint model: the copula and error margins are correctly specified during estimation procedure.

(ii) Misspecified copula: the Gaussian copula is misspecified by other copula families in the estimation process.

(iii) Misspecified margins: the skewed normal distribution is misspecified as normal under the parametric method.

(iv) Different sample size: the sample drawn from each area ranges from 1 to 4.

Under setting (i)-(iii), the true variance of the BLUP of small area means is 0.1, calculated using (2.3).

3.2 Simulation Results

The results under each simulation experiment are summarized in Sections 3.2.1 - 3.2.4. Additional simulation results are deferred to Appendix.

3.2.1 Results under Correctly Specified Joint Model

In this simulation experiment, the true regression model and error margins were correctly specified in the estimation. Tables 3.1, 3.2 and 3.3 summarize the results for different copula family (Clayton, Gaussian, Frank, and Gumbel), number of small areas $m(= 20, 40)$, and error margins (normal and skewed normal). The output of Table 3.1 consists of Bias and MSE of τ. In respect to the estimation of copula parameter, the RVB method uses the empirical Kendall's tau obtained from the pairs of residuals, see Rivest et al. (2016).

In Table 3.2, $\tilde{M}_1$ is $\tilde{M}_{1i}$ averaged over small areas where $\tilde{M}_{1i}$ is calculated when all parameters are known and thus, it is expected to be approximately equal in all cases. Similarly, $\tilde{M}_2$ and $\tilde{M}_3$ are $\tilde{M}_{2i}$ and $\tilde{M}_{3i}$ averaged over small areas, respectively. Notably, the cross-product term, $\tilde{M}_3$, is not ignorable for m=20. The RVB method uses jackknife for the MSPE estimation which constitutes only $\tilde{M}_1$ and $\tilde{M}_2$. As seen in Table 3.2,the magnitude of the cross-product term $\tilde{M}_3$ is the same as the $\tilde{M}_2$ in most cases, and has a negative sign in all cases; noting that $\tilde{M}_2$ captures the variation due to the estimation of the model parameters. As a result of ignoring the cross-product term, the MSPE estimator in the RVB method is always higher than the true EMSPE which leads RB to be positive and larger than the parametric and semi-parametric bootstrap approaches as shown in Table 3.3. Note that the biases of small area mean predictors are very similar for the three estimation methods. To give an illustration, there is 1.3 units absolute difference between the RB of parametric and the RVB method and 2.4

Table 3.1: Average bias and mean squared error of copula parameter expressed in terms of τ when the true $\tau = 0.33$ for three different methods (parametric, semi-parametric, and RVB), the copula family (Clayton, Gaussian, Frank, and Gumbel), and normal error margins are correctly specified.

		Parametric		Semi-parametric		RVB Method	
m	γ	Bias_τ	MSE_τ	Bias_τ	MSE_τ	Bias_τ	MSE_τ
Clayton							
20	0	-0.016	0.006	-0.024	0.006	0.006	0.009
	10	-0.061	0.010	-0.036	0.010	0.005	0.009
40	0	-0.011	0.003	-0.019	0.003	0.000	0.005
	10	-0.039	0.006	-0.028	0.006	-0.003	0.005
Gaussian							
20	0	-0.027	0.008	-0.014	0.008	-0.005	0.009
	10	-0.046	0.009	-0.026	0.009	-0.006	0.009
40	0	-0.015	0.003	-0.007	0.003	-0.004	0.004
	10	-0.026	0.004	-0.015	0.004	-0.005	0.004
Frank							
20	0	-0.023	0.006	-0.024	0.006	0.001	0.007
	10	-0.035	0.007	-0.033	0.007	0.000	0.007
40	0	-0.008	0.003	-0.008	0.003	0.000	0.004
	10	-0.018	0.004	-0.017	0.004	0.002	0.004
Gumbel							
20	0	-0.019	0.010	-0.015	0.010	0.008	0.011
	10	-0.035	0.011	-0.024	0.011	0.009	0.011
40	0	-0.009	0.005	-0.007	0.005	0.007	0.006
	10	-0.026	0.006	-0.021	0.006	0.008	0.006

units absolute difference between semi-parametric and the RVB method in case of the Clayton copula with skewed normal error margins for $m = 40$.

In the case of correctly specified copula family and $m = 20$, Figure 3.1 shows that the variability of absolute RB is relatively small under the parametric and

Table 3.2: Decomposition of EMSPE for the three small area mean predictors (parametric, semi-parametric, and RVB) under each copula family (Clayton, Gaussian, Frank, and Gumbel) with correctly specified error distribution, and error margins (normal, skewed-normal with skewness parameter 10) for $m(=20,40)$.

		Parametric			Semi-parametric			RVB Method		
m	γ	$\widetilde{M}_1$	$\widetilde{M}_2$	$\widetilde{M}_3$	$\widetilde{M}_1$	$\widetilde{M}_2$	$\widetilde{M}_3$	$\widetilde{M}_1$	$\widetilde{M}_2$	$\widetilde{M}_3$
Clayton										
20	0	0.085	0.011	**-0.012**	0.085	0.021	**-0.020**	0.085	0.022	**-0.017**
	10	0.086	0.018	**-0.013**	0.086	0.019	**-0.015**	0.084	0.018	**-0.013**
40	0	0.079	0.005	-0.006	0.079	0.012	-0.011	0.079	0.013	-0.010
	10	0.079	0.010	-0.007	0.079	0.011	-0.008	0.079	0.010	-0.007
Gaussian										
20	0	0.118	0.011	**-0.020**	0.118	0.020	**-0.028**	0.118	0.019	**-0.028**
	10	0.120	0.012	**-0.020**	0.120	0.021	**-0.029**	0.120	0.021	**-0.032**
40	0	0.108	0.005	-0.009	0.108	0.010	-0.015	0.108	0.010	-0.015
	10	0.109	0.006	-0.009	0.109	0.010	-0.014	0.109	0.011	-0.016
Frank										
20	0	0.090	0.013	**-0.013**	0.090	0.016	**-0.015**	0.090	0.015	**-0.015**
	10	0.085	0.012	**-0.011**	0.085	0.016	**-0.013**	0.085	0.015	**-0.015**
40	0	0.081	0.006	-0.006	0.081	0.008	-0.008	0.081	0.008	-0.007
	10	0.078	0.007	-0.007	0.078	0.009	-0.008	0.078	0.008	-0.007
Gumbel										
20	0	0.091	0.012	**-0.015**	0.091	0.018	**-0.021**	0.091	0.016	**-0.021**
	10	0.079	0.013	**-0.012**	0.079	0.017	**-0.015**	0.079	0.015	**-0.018**
40	0	0.082	0.006	-0.007	0.082	0.009	-0.010	0.082	0.008	-0.009
	10	0.070	0.006	-0.006	0.070	0.009	-0.008	0.070	0.007	-0.007

semi-parametric approaches in comparison with the RVB method.

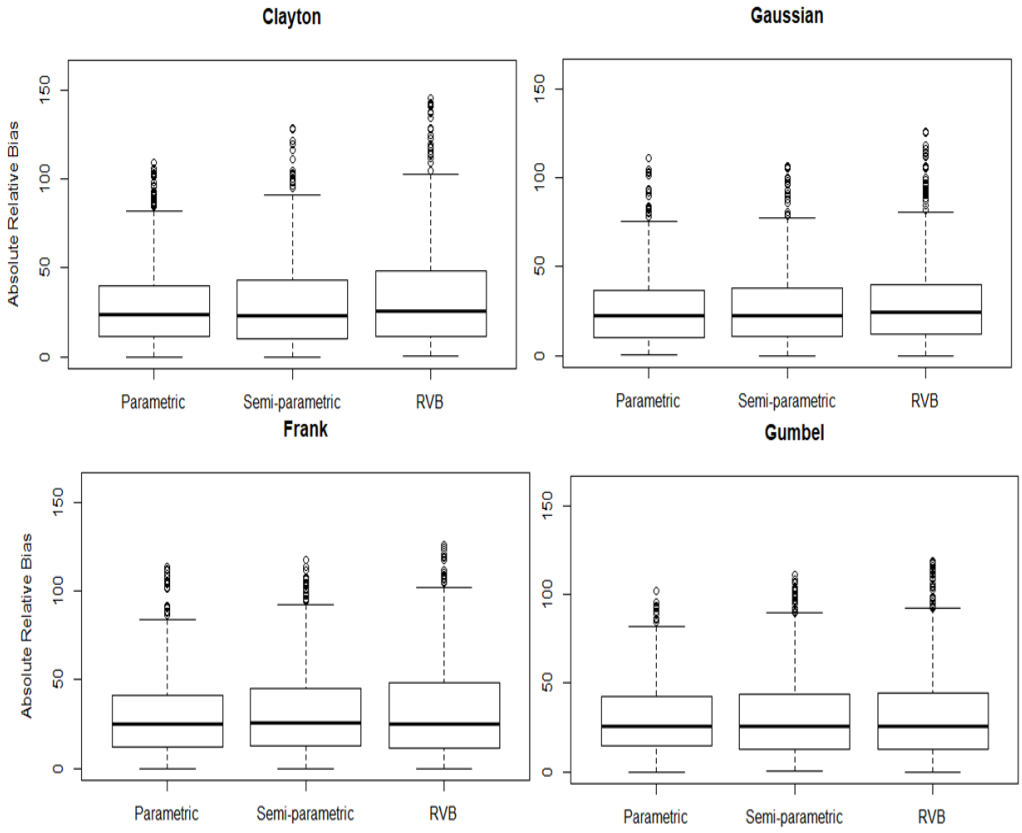
To study the effect of cross-product term in the MSPE estimation for smaller m, we considered the case when m = 10 and marginal error distribution assumed to be normal (Appendix A.1). The percent RB of MSPE estimator (averaged over small

Table 3.3: Average response bias, empirical MSPE, and percent relative bias of MSPE estimate of small area mean predictors for three different methods (parametric, semi-parametric, and RVB) when copula family (Clayton, Gaussian, Frank, and Gumbel) and margins (normal, skewed-normal with skewness parameter 10) are correctly specified for $m(=20,40)$.

		Parametric			Semi-parametric			RVB Method		
m	γ	Bias	EMSPE	RB (%)	Bias	EMSPE	RB (%)	Bias	EMSPE	RB (%)
Clayton										
20	0	-0.012	0.084	-2.4	-0.002	0.087	-2.2	-0.016	0.090	**10.0**
	10	-0.002	0.091	-4.4	-0.004	0.089	-0.2	-0.019	0.089	**12.3**
40	0	-0.006	0.078	-1.3	0.000	0.080	2.5	-0.009	0.082	**3.7**
	10	-0.002	0.082	-3.6	-0.001	0.079	2.5	-0.009	0.081	**4.9**
Gaussian										
20	0	0.003	0.109	-3.7	0.003	0.109	-1.8	0.003	0.108	**5.5**
	10	0.001	0.112	-8.0	-0.004	0.111	-4.5	0.000	0.109	**6.4**
40	0	-0.001	0.104	-1.9	-0.001	0.104	0.2	-0.001	0.103	**3.9**
	10	0.002	0.105	-2.8	-0.002	0.105	-1.9	-0.003	0.104	**3.8**
Frank										
20	0	-0.005	0.089	-1.1	-0.002	0.090	3.3	-0.012	0.090	**6.7**
	10	-0.014	0.087	-6.9	-0.011	0.087	-1.2	-0.016	0.085	**7.1**
40	0	-0.006	0.081	0.1	-0.005	0.082	2.4	-0.012	0.082	**4.9**
	10	-0.005	0.079	-3.3	-0.005	0.079	-1.3	-0.010	0.078	**2.6**
Gumbel										
20	0	-0.001	0.088	-6.8	0.006	0.088	-1.1	0.000	0.086	**5.8**
	10	0.008	0.079	-10.1	0.009	0.080	-6.2	-0.004	0.074	**8.1**
40	0	-0.001	0.080	-2.5	0.004	0.081	0.4	-0.001	0.081	**2.5**
	10	0.005	0.070	-5.7	0.005	0.071	-4.2	0.001	0.068	**4.4**

areas) under the Clayton copula family is -2.1, 4.9, and 16.1 for the parametric, semi-parametric and RVB methods, respectively, which shows that the RB is comparatively high in the RVB method.

Figure 3.1: Boxplots of absolute RB (averaged over small areas) for each method when m=20 and model is correctly specified.



3.2.2 Results under Copula Misspecification

Under this setting, the error were drawn from a standard normal distribution only. The EBUP and MSPE estimation of the small area mean predictors were calculated when true copula family, the Gaussian copula, was misspecified as the Clayton, Gumbel and Frank copulas. Table 3.4 summarizes the results under the copula family misspecification. As expected, $\tilde{M}_2$, which captures the variation due to the

estimation of model parameters, was observed to be at least tripled in magnitude in comparison with $\tilde{M}_2$ when the copula family is correctly specified irrespective of the method used for the inference. Thus, in practice, a wrong assumption regarding the copula family will lead to significant increase in MSPE of small area mean predictors.

Table 3.4: Decomposition of EMSPE for the three small area mean predictors (parametric, semi-parametric, and RVB) under a situation where the Gaussian copula is misspecified by other copula family (Clayton, Frank, and Gumbel) in the cases of error normal margins for different $m(=20,40)$.

		Parametric			Semi-parametric			RVB Method		
m	γ	$\widetilde{M}_1$	$\widetilde{M}_2$	$\widetilde{M}_3$	$\widetilde{M}_1$	$\widetilde{M}_2$	$\widetilde{M}_3$	$\widetilde{M}_1$	$\widetilde{M}_2$	$\widetilde{M}_3$
True Gaussian										
20	0	0.118	0.011	-0.020	0.118	0.020	-0.028	0.118	0.019	-0.028
40		0.108	0.005	-0.009	0.108	0.010	-0.015	0.108	0.010	-0.015
Misspecified Clayton										
20	0	0.118	**0.051**	-0.019	0.118	**0.052**	-0.025	0.118	**0.057**	-0.022
40		0.110	**0.047**	-0.013	0.110	**0.047**	-0.018	0.110	**0.057**	-0.015
Frank										
20	0	0.118	**0.046**	-0.024	0.118	**0.050**	-0.026	0.118	**0.045**	-0.026
40		0.110	**0.045**	-0.016	0.110	**0.047**	-0.018	0.110	**0.046**	-0.018
Gumbel										
20	0	0.118	**0.048**	-0.022	0.118	**0.054**	-0.027	0.118	**0.044**	-0.025
40		0.110	**0.043**	-0.014	0.110	**0.048**	-0.017	0.109	**0.042**	-0.016

3.2.3 Results under Misspecified Marginal Error Distribution

Another simulation experiment was conducted when the population was drawn with error margins of skewed normal distribution. In this setting, we evaluated the performance specifically for the parametric case when error margin is mis-

specified as the standard normal distribution. As a known fact, distributional assumption is required in parametric case, that is to say, error marginal distribution needs to be specified while doing estimation. The results under this setting are reported in Table 3.5.

As expected, the EMSPE of the small area means remains the same in the semi-parametric and RVB methods. Whereas, in the parametric method there is a significant increase in $\tilde{M}_2$ of the EMSPE results under misspecification of the marginal distribution. In practice, depending upon the provided data information, one should move forward with either parametric or semi-parametric method in order to achieve good performance in terms of the EMSPE of small area mean predictors.

So far, in Sections 3.2.1 to 3.2.3, we discussed results can change when one assumes true or false information in the MSPE estimation of small area mean predictors. For simplicity, we assumed that the sample sizes are the same in each small area. Next, we evaluate the performance of all three methods when the number of sample size is different in small areas.

3.2.4 Different Sample Size

Another simulation experiment was conducted with different sample sizes ranging from 1 to 4 in the case of $m = 20$. In this simulation experiment, we considered the Frank copula and normal margins. In particular, we assumed that areas 1-5 have sample sizes one, areas 6-10 have sample sizes two, areas 11-15 have sample

Table 3.5: Decomposition of EMSPE for the three small area mean predictors (parametric, semi-parametric, and RVB) under each copula family (Clayton, Gaussian, Frank, and Gumbel) when error margins (with skewness parameter 10) are misspecified as normal distribution for different $m(=20,40)$.

		Parametric			Semi-parametric			RVB Method		
m	γ	$\widetilde{M}_1$	$\widetilde{M}_2$	$\widetilde{M}_3$	$\widetilde{M}_1$	$\widetilde{M}_2$	$\widetilde{M}_3$	$\widetilde{M}_1$	$\widetilde{M}_2$	$\widetilde{M}_3$
Clayton										
True										
20	10	0.086	0.018	-0.013	0.086	0.019	-0.015	0.086	0.018	-0.015
40		0.079	0.010	-0.007	0.079	0.011	-0.008	0.079	0.010	-0.007
Misspecified										
20	0	0.084	**0.037**	-0.012	0.084	0.019	-0.015	0.084	0.018	-0.013
40		0.078	**0.035**	-0.004	0.078	0.011	-0.008	0.078	0.010	-0.007
Gaussian										
True										
20	10	0.120	0.012	-0.020	0.120	0.021	-0.029	0.120	0.021	-0.032
40		0.109	0.006	-0.009	0.109	0.010	-0.014	0.109	0.011	-0.016
Misspecified										
20	0	0.120	**0.032**	-0.018	0.120	0.022	-0.033	0.120	0.021	-0.032
40		0.109	**0.030**	-0.006	0.109	0.011	-0.016	0.109	0.011	-0.016
Frank										
True										
20	10	0.085	0.012	-0.011	0.085	0.016	-0.013	0.085	0.015	-0.015
40		0.085	0.012	-0.011	0.085	0.016	-0.013	0.085	0.015	-0.015
Misspecified										
20	0	0.086	**0.022**	-0.011	0.086	0.016	-0.015	0.086	0.015	-0.015
40		0.077	**0.018**	-0.005	0.077	0.008	-0.008	0.077	0.008	-0.007
Gumbel										
True										
20	10	0.079	0.013	-0.012	0.079	0.017	-0.015	0.078	0.015	-0.017
40		0.070	0.006	-0.006	0.070	0.009	-0.008	0.070	0.007	-0.008
Misspecified										
20	0	0.077	**0.026**	-0.015	0.077	0.017	-0.017	0.077	0.015	-0.018
40		0.069	**0.020**	-0.008	0.069	0.008	-0.008	0.069	0.007	-0.007

sizes three, and areas 16-20 have sample sizes four. The true MSPE of the BLUP (see (2.3)) is 0.250, 0.167, 0.125, and 0.100 for small areas with sample size n_i = 1, 2, 3, and 4, respectively. As expected, the true MSPE of the BLUP of small area

means decreases with increasing sample sizes.

Table 3.6: Decomposition of EMSPE for the three estimators under Frank copula family for different sample sizes with standard normal error margins.

		Parametric			Semi-parametric			RVB Method		
Area i	n_i	$\widetilde{M}_1$	$\widetilde{M}_2$	$\widetilde{M}_3$	$\widetilde{M}_1$	$\widetilde{M}_2$	$\widetilde{M}_3$	$\widetilde{M}_1$	$\widetilde{M}_2$	$\widetilde{M}_3$
1-5	1	0.267	0.027	-0.011	0.267	0.031	-0.014	0.267	0.035	-0.013
6-10	2	0.169	0.026	-0.019	0.169	0.033	-0.023	0.169	0.034	-0.022
11-15	3	0.117	0.024	-0.020	0.117	0.030	-0.023	0.116	0.032	-0.024
16-20	4	0.096	0.021	-0.018	0.096	0.026	-0.022	0.096	0.026	-0.022
1-20		0.162	0.024	-0.017	0.162	0.030	-0.020	0162	0.032	-0.020

The results based on this experiment are summarized in Tables 3.6 and 3.7. Turning on to the evaluation of three approaches, the EMSPE of small area mean predictors decreases with increasing sample sizes. Also, we observed that the EMSPE of small area mean predictors in this scenario (sample sizes vary from 1 to 4) are consistently larger than the corresponding values from Table 3.3 where the sample sizes were 4 in all areas. Moreover, in terms of RB of MSPE estimation, in general, the parametric approach shows smaller RB with increasing sample sizes (Table 3.7). However, totally opposite trend is noticed in the RVB method i.e., firstly the RB increases with an increase in sample size from 1 to 2 which is almost 89.8%, and then 23.3% increment when sample size is 3. Furthermore, there is a sudden downfall by 49.9% in RB when sample size is 4. This demonstration illustrates how good the parametric method captures variations of the true EMSPE. On the other hand, considering empirical version of the estimation procedure, semi-parametric approach performs better than the RVB method in terms of RB of the MSPE estimation of small area mean predictors.

Table 3.7: Average response bias, EMSPE, and percent relative bias of MSPE estimate of small area mean predictors for different sample sizes under Frank copula family with error margins drawn from standard normal distribution.

		Parametric			Semi-parametric			RVB Method		
Area_i	n_i	Bias	EMSPE	RB (%)	Bias	EMSPE	RB (%)	Bias	EMSPE	RB (%)
1-5	1	0.027	0.282	**-10.3**	0.026	0.284	**-4.9**	0.027	0.289	**7.9**
6-10	2	0.012	0.176	**-6.2**	0.011	0.179	**-1.1**	0.000	0.180	**15.0**
11-15	3	-0.002	0.121	**0.8**	0.001	0.123	**6.5**	-0.011	0.124	**18.5**
16-20	4	-0.002	0.099	**-1.0**	0.002	0.100	**5.0**	-0.009	0.100	**10.0**
1-20		0.009	0.169	**-5.9**	0.010	0.172	**-0.6**	0.002	0.174	**11.5**

3.3 Summary

For all the scenarios considered in our simulations, the EMSPE was found to be very similar in all three estimators when sample size was equal for each small area. We observed that the semi-parametric method consistently performed well even when the copula family is misspecified. As expected, the parametric method had the best performance among the three methods when the joint model was correctly specified. However, a slight variation was observed in the parametric method in the context of variation in model parameters estimation, i.e., $\tilde{M}_2$, when the copula family or the marginal error distribution were misspecified.

For the case when sample sizes were different, unexpectedly the EMSPE was comparatively high in the RVB method, specifically when sample size in each area was either 1 or 2. Not only that, RB showed variation in its trend with an increase in the sample sizes. However, the impact of difference in sample sizes on RB was found to be almost negligible over both parametric and semi-parametric methods.

As expected, the proposed bootstrap method was capable to capture complete information in the MSPE of small area mean predictors. Overall, assuming the cross-product term to be negligible was found to be inappropriate for non-normal data due to the significant contribution of the cross-product term to the MSPE.

Chapter 4

Data Application

In this chapter, we apply our proposed methodology to the real data using land observatory satellite (LANDSAT) on county crop data. We then investigate the estimation performance under all three methods for comparison.

4.1 LANDSAT - County Crop Data

The county crop data considered here is based on LANDSAT readings obtained in 1978 during growing season of corn and soybean by the U.S Department of Agriculture (USDA) in the counties (area) of north central Iowa (Battese et al., 1988). The auxiliary data has been employed in form of LANDSAT data, where "segment" is the primary sampling unit and "pixel" is a picture element for which satellite information is recorded. Approximately 250 hectares, where each pixel is about 0.45 hectares, are represented by each segment. Further, it is estimated from satellite photographs by counting the number of individual pixels. Whereas,

the area of corn and soybean crop is determined by interviewing farm operators. Based on these information, the USDA procedures were used to classify 12 counties in 37 segments.

The aim of this study is to predict mean hectare of corn crop in each county based on 1978 June Enumerative Survey and satellite data. Table 4.1 presents (i) number of segments in each county, (ii) number of hectares for corn crop and pixels for corn and soybeans in each sample segment and (iii) mean of pixels per segments for corn crop. As per the report of preliminary analysis of the corn data, second segment of county named Hardin deviated from other observations. Therefore, data regarding that second segment is deleted for further analysis leaving 36 total segments.

The residuals display asymmetric distribution in small areas. Also, the maximum log-likelihood is obtained under the Frank copula and thus, estimation is carried forward by comparing Frank's EBUP of small area means in each method. For the parametric approach, skewed normal distribution is used for error margins. The log-likelihood values were -4.21 and -3.75; average bootstrap residuals were 145.3 and 160.1; and the residual exchangeable Kendall's tau values were 0.27 and 0.28, respectively for parametric and semi-parametric approaches. The small area mean predictors are presented in Table 4.2 along with root mean squared prediction error (Rmspe) where B=1000 bootstrap samples were used in the bootstrap method.

Table 4.1: Survey and satellite data for corn and soybeans in 12 north central Iowa counties.

	No. of Segments		Reported Hectares	No. of pixel in sample segments		Mean no. of pixel per segment	
County	Sample(n_i)	County (N_i)	Corn (Y_1)	Corn (X_1)	Soybeans (X_2)	Corn ($\bar{X}_1$)	Soybeans ($\bar{X}_2$)
Cerro Gordo	1	545	165.76	374	55	295.29	189.70
Hamilton	1	566	96.32	209	218	300.40	196.65
Worth	1	394	76.08	253	250	289.60	205.28
Humboldt	2	424	185.35	432	96	290.74	220.22
			116.43	367	178		
			162.08	361	137		
Franklin	3	564	152.04	288	206	318.21	188.06
			161.75	369	165		
			92.88	206	218		
Pocahontas	3	570	149.94	316	221	257.17	247.13
			64.75	145	338		
			127.07	355	128	291.77	185.37
Winnebago	3	402	133.55	295	147		
			77.70	223	204		
			206.97	459	77	301.26	221.36
Wright	3	567	108.33	290	217		
			118.17	307	258		
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
			88.59	220	262		
			165.35	355	160		
Hardin	5	556	104.00	261	221	325.99	177.05
			88.63	187	345		
			153.70	350	190		

Table 4.2: EBUP of small area means and corresponding Rmspe for LANDSAT data under Frank copula.

		Parametric		Semi-parametric		RVB Method	
County	n_i	EBUP	Rmspe	EBUP	Rmspe	EBUP	Rmspe_{jk}
Cerro Gordo	1	124.2	11.9	123.1	12.3	120.2	11.4
Hamilton	1	125.2	12.0	124.6	12.1	127.5	11.8
Worth	1	108.2	11.1	108.1	11.4	104.9	10.7
Humboldt	2	106.2	**8.6**	106.2	**8.8**	102.3	**12.2**
Franklin	3	142.4	5.7	141.3	5.9	145.2	4.3
Pocahontas	3	110.5	6.1	109.1	6.1	111.8	6.0
Winnebago	3	110.0	**6.0**	111.3	**6.2**	111.0	**11.2**
Wright	3	118.7	6.3	119.4	6.4	120.8	7.5
Webster	4	114.6	4.4	113.8	4.6	117.4	3.8
Hancock	5	118.8	4.2	121.4	4.4	125.2	6.4
Kossuth	5	109.4	4.1	109.9	4.2	107.8	4.1
Hardin	5	137.2	4.8	136.4	4.9	140.8	4.7

As shown in Table 4.2, in the cases of the parametric and semi-parametric approaches, Rmspe values significantly decrease with an increase in the number of sample segments. Furthermore, the improvement in Rmspe is modest when sample size is greater than 3 or more. However, Rmspe of counties 4 and 7 namely, Humboldt and Winnebago, are at least 1.5 times higher in the RVB method compared to the corresponding values in the parametric and semi-parametric approaches. Note that in the MSPE estimation of the RVB method for the counties 4 and 7, the Rmspe values do not decrease with increasing sample sizes unlike the parametric and semi-parametric approaches. One possible explanation for the higher magnitude of Rmspe for counties 4 and 7 is the usage of jackknife, which does not capture the all variations of the EBUP of small area means, as also shown in the simulation study.

Chapter 5

Conclusion

In this thesis, we have studied a small area model using the unit-level regression model (Rivest et al., 2016) where the joint distribution of error terms belongs to the family of the multivariate exchangeable copulas. For this model, we proposed a more flexible and general framework for predicting small area means where the copula parameter is estimated via the maximum pseudo copula log-likelihood method. The proposed approach can accommodate both parametric and semi-parametric methods. We also investigated the impact of the variation of the EBUP of small area means due to misspecification of the copula family or marginal error distribution in order to assess performance of the proposed methods, and to also compare with the RVB method.

Unlike the normal model, the cross-product term involved in the MSPE of the EBUP of small area means is not negligible. We addressed this aspect of cross-product term by doing inspection on the contribution of each term under the MSPE of the EBUP of small area means. Our results indicated that each

decomposed term of the MSPE has its own contribution. Hence, the proposed bootstrap method was used to capture all variations in the MSPE of EBUP of small area means including the possible cross-product term.

Overall, the results from the simulation experiments for the proposed approach showed various phases in different settings like misspecified copula family, misspecified margins, and difference in sample size selection. As expected, the parametric method performed relatively better when the underlying distribution is correctly specified in context of the variation of model parameter estimation of small area mean predictors. In almost all scenarios, our semi-parametric approach performed relatively well in comparison to the RVB method. Nevertheless, in context of RB, semi-parametric approach performed relatively well in almost all scenarios in terms of the MSPE of the EBUP of small area means. In addition, when the simulation was performed under different sample sizes, the parametric approach was more consistent in the decrement of RB as compared to the RVB method. As a result, the proposed bootstrap approach performed better for the MSPE estimation than the jackknife estimation used in the RVB method. Also, we investigated the results based on double bootstrap method however, the gained in the RB was not significant, see Appendix A.2. Thus, we proceeded with single phase bootstrap method for the estimation of MSPE of the EBUP of small area means as we got relatively small RB in most scenarios.

We also applied our proposed methodology on the real data example: LANDSAT on county crop data. In practice, one needs to decide which copula family best suits the data. So, we proceeded our analysis with the decision of selection of

the copula family based on the log-likelihood value. We observed that Rmspe of parametric and semi-parametric methods decrease with an increase in sample size. Whereas, the RVB method showed dislocation from the track at some points.

In conclusion, a major advantage of proposed method is that it provides more flexible approach for small area estimation under both parametric and semi-parametric methods. The proposed bootstrap procedure can capture all terms of MSPE of small area mean predictors to obtain its corresponding estimate. The proposed approach can be further extended to discrete outcomes and also for a situation when the covariates are measured with error. These are some of the topics for future study.

Appendix A

A.1 Results for correctly specified joint model in the case of m=10

We also considered the situation when number of small area m =10 under similar setting as of Section 3.2.1, where error margins are from standard normal distribution. The results are summarized in Table A.1. As expected, the cross-product term is relatively high in all cases in comparison with a situation when m=20. Also, the percent RB of MSPE estimator is almost high in most of the cases especially in the RVB method. Thus, when number of small area is small, the cross-product term cannot be ignored while doing estimation of MSPE of small area mean predictors.

A.2 Results while using double bootstrap method

We also studied the double bootstrap method when number of bootstrap samples in second-phase B2 (= 1) under the same setting of the simulation study in Section 3.2.1, for the Clayton copula family where error margins are from standard

Table A.1: Decomposition of EMSPE, average bias, EMSPE, and percent relative bias of MSPE estimate of small area mean predictors for three different methods (parametric, semi-parametric, and RVB) when copula family (Clayton, Gaussian, Frank, and Gumbel) and margins (normal) are correctly specified for $m = 10$.

	Parametric			Semi-parametric			RVB Method		
Copula	$\tilde{M}_1$	$\tilde{M}_2$	$\tilde{M}_3$	$\tilde{M}_1$	$\tilde{M}_2$	$\tilde{M}_3$	$\tilde{M}_1$	$\tilde{M}_2$	$\tilde{M}_3$
Clayton	0.101	0.024	-0.029	0.101	0.041	-0.041	0.101	0.041	-0.031
Gaussian	0.130	0.023	-0.035	0.130	0.039	-0.052	0.129	0032	-0.045
Frank	0.108	0.027	-0.029	0.108	0.032	-0.033	0.108	0.029	-0.031
Gumbel	0.107	0.028	-0.039	0.107	0.037	-0.048	0.104	0.034	-0.042
	Bias	EMSPE	RB (%)	Bias	EMSPE	RB (%)	Bias	EMSPE	RB (%)
Clayton	-0.008	0.095	-2.1	0.006	0.101	4.9	-0.011	0.112	16.1
Gaussian	0.002	0.118	-10.2	0.001	0.118	-4.3	-0.009	0.115	13.9
Frank	0.001	0.106	-3.8	0.006	0.107	3.8	-0.016	0.105	10.5
Gumbel	-0.008	0.096	-7.3	0.003	0.096	2.1	-0.003	0.096	15.6

normal distribution. The results are summarized in Table A2, where RB_{boot2} is calculated using the MSPE estimator defined in Hall and Maiti (2006) and Torabi (2012) described as: using Algorithm 1, we can obtain $\text{mspe}_{boot}(\hat{\bar{Y}}_{iU}) \equiv \hat{k}_i$, averaged over simulation runs, for first-phase bootstrap method. Again, from the estimated parameters obtained in first phase, generate population dataset for second-phase and repeat steps from 1 to 8 in Algorithm 1 to get the EBUP of small area means ($\hat{\bar{Y}}_{iU}^{**(b2)}$) and average $\text{mspe}_{boot}^{**}(\hat{\bar{Y}}_{iU}) = \hat{l}_i$ for R simulation runs, where $\text{mspe}_{boot}^{**}(\hat{\bar{Y}}_{iU}) = \frac{1}{R}\sum_{r=1}^{R}\left[\frac{1}{B}\sum_{b=1}^{B}\left\{\frac{1}{B2}\sum_{b2=1}^{B2}(\hat{\bar{Y}}_{iU}^{**(b2)} - \bar{Y}_{iU}^{**(b2)})^2\right\}\right]$. Then, we have following MSPE estimators proposed by Hall and Maiti (2006):

$$\text{mspe}_{boot2}(\hat{\bar{Y}}_{iU}) \approx \begin{cases} 2\hat{k}_i - \hat{l}_i, & \hat{k}_i \geq \hat{l}_i \\ \hat{k}_i \exp\{-(\hat{l}_i - \hat{k}_i)/\hat{l}_i\}, & \hat{k}_i \leq \hat{l}_i \end{cases}$$

and

$$\text{mspe}_{boot3}(\hat{\bar{Y}}_{iU}) \approx \hat{k}_i^2/\hat{l}_i.$$

The RB (for example in the case of mspe_{boot2}) can be written as

$$\text{RB}(\text{mspe}_{boot2}(\hat{\bar{Y}}_{iU})) = \frac{\text{mspe}_{boot2}(\hat{\bar{Y}}_{iU})}{\text{EMSPE}(\hat{\bar{Y}}_{iU})} - 1.$$

Similarly, we can obtain for semi-parametric approach by employing Algorithm 2. As a result, we observed that using double bootstrap method, we gain in terms of RB with an increase in sample size but not significant. Also, this procedure is computational intensive and thus, in this paper we follow only single phase bootstrap method as we also got relatively small RB.

Table A2: Decomposition fo EMSPE, average bias, EMSPE, and percent relative bias of MSPE estimate of small area mean predictors in case of the double bootstrap method for three different methods (parametric, semi-parametric) when the copula family (Clayton) and margins (normal) are correctly specified for $m(=20,40)$.

Clayton		Parametric			Semi-parametric	
m	$\tilde{M}_1$	$\tilde{M}_2$	$\tilde{M}_3$	$\tilde{M}_1$	$\tilde{M}_2$	$\tilde{M}_3$
20	0.084	0.011	-0.012	0.084	0.022	-0.021
40	0.080	0.006	-0.006	0.080	0.013	-0.012
	Bias	EMSPE	RB_{boot3} (%)	Bias	EMSPE	RB_{boot3} (%)
20	-0.013	0.082	3.6	-0.003	0.084	4.1
40	-0.006	0.079	-1.2	0.000	0.080	0.8

Bibliography

Battese, G. E., Harter, R. M., and Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83(401):28–36. 9, 55

Datta, G. S. and Ghosh, M. (1991). Bayesian prediction in linear models: Applications to small area estimation. *The Annals of Statistics*, pages 1748–1770. 10

Datta, G. S. and Lahiri, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica*, pages 613–627. 19

Fay, R. E. and Herriot, R. A. (1979). Estimates of income for small places: an application of james-stein procedures to census data. *Journal of the American Statistical Association*, 74(366a):269–277. 9, 10, 14

Genest, C., Ghoudi, K., and Rivest, L.-P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82(3):543–552. 31

Ghosh, M. and Lahiri, P. (1987). Robust empirical bayes estimation of means from stratified samples. *Journal of the American Statistical Association*, 82(400):1153–1162. 10

Ghosh, M. and Lahiri, P. (1992). A hierarchical bayes approach to small area estimation with auxiliary information. In *Bayesian Analysis in Statistics and Econometrics*, pages 107–125. Springer. 10

Ghosh, M. and Rao, J. (1994). Small area estimation: an appraisal. *Statistical Science*, pages 55–76. 13

Hall, P. and Maiti, T. (2006). On parametric bootstrap methods for small area prediction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(2):221–238. 35, 63

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, pages 423–447. 29

Kackar, R. N. and Harville, D. A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association*, 79(388):853–862. 16

Lahiri, P. and Rao, J. (1995). Robust estimation of mean squared error of small area estimators. *Journal of the American Statistical Association*, 90(430):758–766. 10

Nelsen, R. B. (2006). An introduction to copulas, 2nd. *New York: Springer Science Business Media*. 22, 25

Pfeffermann, D. et al. (2013). New important developments in small area estimation. *Statistical Science*, 28(1):40–68. 11

Prasad, N. and Rao, J. N. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, 85(409):163–171. 10, 16, 19, 30

Rao, J. N. and Molina, I. (2015). *Small-Area Estimation, 2nd Edition*. Wiley Online Library. 9, 13, 16, 17, 18, 19, 30

Rivest, L.-P., Verret, F., and Baillargeon, S. (2016). Unit level small area estimation with copulas. *Canadian Journal of Statistics*, 44(4):397–415. 10, 11, 28, 29, 34, 39, 44, 59

Sklar, M. (1959). Fonctions de repartition an dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8:229–231. 20

Torabi, M. (2012). Small area estimation using survey weights under a nested error linear regression model with structural measurement error. *Journal of Multivariate Analysis*, 109:52–60. 35, 63

Torabi, M. and Rao, J. N. (2010). Mean squared error estimators of small area means using survey weights. *Canadian Journal of Statistics*, 38(4):598–608. 9

You, Y. and Rao, J. (2003). Pseudo hierarchical bayes small area estimation combining unit level models and survey weights. *Journal of Statistical Planning and Inference*, 111(1-2):197–208. 9