The Canadian Journal of Statistics
Vol. xx, No. yy, 20??, Pages 1–24
La revue canadienne de statistique

Copula-based predictions in small area estimation

Kanika Grover[1], Elif Acar[2] and Mahmoud Torabi[3]*

[1] *Census Operations Division, Statistics Canada, 170 Tunney's Pasture Driveway, Ottawa, ON K1A 0T6, Canada*
[2] *Department of Statistics, University of Manitoba, Winnipeg, MB R3T 2N2, Canada &*
Hospital for Sick Children, Toronto, ON M5G 1X8, Canada
[3] *Departments of Community Health Sciences & Statistics, University of Manitoba, Winnipeg, MB R3E 0W3, Canada*

Key words and phrases: Best unbiased predictor; bootstrap; multivariate exchangeable copula; pseudo-copula likelihood; small area estimation.

MSC 2010: Primary 62D99; secondary 62H12

Abstract: Unit-level regression models are commonly used in small area estimation to obtain an empirical best linear unbiased prediction of small-area characteristics. The underlying assumptions of these models, however, may be unrealistic in some applications. Previous work developed a copula-based small area estimation model where the empirical Kendall's tau was used to estimate the dependence between two units from the same area. In this paper, we propose a likelihood framework to estimate the intra-class dependence of the multivariate exchangeable copula for the empirical best unbiased prediction (EBUP) of small area means. One appeal of the proposed approach lies in its accommodation of both parametric and semi-parametric estimation approaches. Under each estimation method, we further propose a bootstrap approach to obtain a nearly unbiased estimator of the mean squared prediction error (MSPE) of the EBUP of small area means. The performance of the proposed methods is evaluated through simulation studies and also by a real data application. *The Canadian Journal of Statistics* xx: 1–24; 20?? © 20?? Statistical Society of Canada

Résumé: Insérer votre résumé ici. *La revue canadienne de statistique* xx: 1–24; 20?? © 20?? Société statistique du Canada

1. INTRODUCTION

Small area estimation (SAE) plays an important role in survey sampling due to the increasing need for reliable small area statistics in both private and public sectors. In policy making regarding allocation of resources to subgroups (small areas), or determination of subgroups with specific characteristics (e.g., in health and medical studies) in a population, it is desirable that the decisions are based on reliable estimates. However, if samples in some areas are small relative to their corresponding populations, inference for these areas using direct estimates may result in very high coefficients of variation.

Direct estimates are the most desirable when each area has a sufficiently large sample size. However, for small sample sizes, one is bound to use mixed models to borrow information from other resources such as other (possibly bigger) surveys, census, and administrative data. In particular, to handle complex issues such as violation of homogeneity within a domain or changes in the population structure, the use of mixed models is proposed in the literature (Rao & Molina ,

* *Author to whom correspondence may be addressed.*
E-mail: Mahmoud.Torabi@umanitoba.ca

2015). The area-level model of Fay & Herriot (1979) and the unit-level model of Battese, Harter & Fuller (1988) have been extensively employed in SAE as special cases of linear mixed models. In both, area-specific random effects are 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.

Small area models have also been proposed for non-normal responses. In particular, there are many applications in small area estimation where responses are proportions or counts. MacGibbon & Tomberlin (1989) and Malec et al. (1997) used logistic regression models with area-specific random intercept and slope effects, respectively, to estimate small area proportions. Molina, Saei & Lombardía (2007) and López-Vizcaíno, Lombardía & Morales (2013, 2015) used multinomial logistic mixed models to estimate labor force characteristics in UK small areas. Mortality and disease rates for small areas are also often used to construct disease maps such as cancer atlases to display geographical variability of a disease and identify high-rate areas warranting intervention. Tsutakawa, Shoop & Marienfield (1985), Clayton & Kaldor (1987), Nandram, Sedransk & Pickle (1999), Langford et al. (1999), Datta, Ghosh & Waller (2000), Dean & MacNab (2001), among others, used Poisson mixed models to study rates of different diseases in small areas. Ghosh et al. (1998) and Torabi & Shokoohi (2015) proposed generalized linear models (GLMs) with random area effects to predict small area statistics. Ghosh et al. (1999) and Torabi (2019) extended the GLMs to handle spatial data and applied the model to disease mapping.

The assumption of normality in SAE is often not met in applications. For instance, in most business surveys, response variables (outcomes) have skewed distributions and their relationships to auxiliary variables often deviate from a linear form (Chandra & Chambers, 2011). Diallo & Rao (2018) considered small area predictors for a skewed-normal model. Another concern is the issue of outliers that has been highlighted in the recent articles such as Chambers & Tzavidis (2006), Sinha & Rao (2009), Jiongo, Haziza & Duchesne (2013), and Chambers et al. (2014). In the case of unit-level models, Jiang & Nguyen (2012) also considered a unit-level model where the residual variance varies between small areas. Rao & Molina (2015) gave an extensive review of model-based SAE models.

In order to develop a strategy for modelling non-normal continuous outcomes, a more flexible unit-level model in terms of the distribution of the error terms within each small area was proposed by Rivest, Verret & Baillargeon (2016) using multivariate exchangeable copulas. Their work outlined a two-stage semi-parametric approach where the marginal distribution of errors is estimated using the empirical distribution function and the intra-class dependence is quantified via the empirical Kendall's tau. Using the model, they derived the empirical best unbiased prediction (EBUP) of small area means and employed the jackknife method to estimate the mean squared prediction error (MSPE) of the EBUP of small area means.

The contributions of this paper are twofolds. Under the same setting as in Rivest, Verret & Baillargeon (2016), we propose a maximum pseudo-copula log-likelihood framework to estimate the intra-class dependence of the multivariate exchangeable copula and obtain the EBUP of small area means. Our approach can accommodate both two-stage parametric and two-stage semi-parametric estimation methods as alternatives to that of Rivest, Verret & Baillargeon (2016), which is referred to as the RVB method throughout the paper. In addition, we propose a bootstrap approach to obtain a nearly-unbiased estimate of the MSPE of the EBUP under each estimation method.

The outline of the paper is as follows. In Section 2, we review the small area model where the joint error distribution is specified using a multivariate exchangeable copula, and define the best unbiased prediction (BUP) of small area means and the corresponding MSPE under this model. In Section 3, we present the proposed parametric and semi-parametric estimation methods, where the copula parameter is estimated using the maximum pseudo-copula log-likelihood, and define the EBUP of small area means and its MSPE. This section also presents the proposed bootstrap methods for the estimation of MSPE of the EBUP of small area means under both parametric and semi-parametric approaches. Section 4 contains simulation studies which compare the finite sample performance of the proposed approaches with that of the RVB method under various scenarios, including misspecification of the copula family and marginal error distribution. Section 5 illustrates the appeal of the proposed methods with an application of the Landsat data. Concluding remarks are given in Section 6. Technical details are deferred to Appendix. Supplementary Materials contain R codes and files relating to the simulations and analysis conducted in this article.

2. COPULA MODEL FOR SMALL AREA ESTIMATION

2.1. Multivariate Exchangeable Copula Model

Consider a population of m small areas of sizes $N_1, N_2, \ldots, N_m$, respectively. Let Y be the response variable and x be a p-dimensional vector of auxiliary variables. Suppose the relationship between Y and x follows the linear model

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

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

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

Here the parameter α quantifies the within-area dependence. Note that the above distribution is connected to the standard normal mixed model set-up. Battese, Harter & Fuller (1988) considered a unit-level regression model (1a) with random intercept assuming the errors are $\varepsilon_{ij} = \nu_i + \xi_{ij}$ where ν_i and ξ_{ij} are independent random variables with respective distributions $N(0, \sigma_\nu^2)$ and $N(0, \sigma_\xi^2)$. The joint error distribution of each area is then a $N(0, \sigma^2\Sigma(\rho, N_i))$, where $\sigma^2 = \sigma_\nu^2 + \sigma_\xi^2$ and $\rho = \sigma_\nu^2/(\sigma_\nu^2 + \sigma_\xi^2)$, with ρ called the intra-class correlation (ICC).

We assume that the joint error distribution (1b) satisfies the following two conditions:

1. *Exchangeable property*: The joint error distribution is invariant under permutation of its arguments, i.e.,

$$F(\varepsilon_1, \ldots, \varepsilon_d) = F(\varepsilon_{\pi(1)}, \ldots, \varepsilon_{\pi(d)}).$$

2. *Invariance property*: The joint error distribution satisfies the invariance property for dimensions, i.e.,

$$F(\infty, \ldots, \infty, \varepsilon_1, \ldots, \varepsilon_d, \infty, \ldots, \infty) = F(\varepsilon_1, \ldots, \varepsilon_d).$$

These assumptions are key to the use of copulas in small area estimation as they facilitate inference of the model in (1a) and (1b), called (1) in the following development.

2.2. Prediction of Small Area Means

One of the main interests in SAE is to predict small area means $\gamma_i = \sum_{j=1}^{N_i} Y_{ij}/N_i$. Suppose a random sample s_i of size n_i is drawn from each small area i ($i = 1, \dots, m$) with known population size N_i under the model (1), assuming no sample selection bias. Denote by r_i the set of un-sampled units within the i^{th} small area, and $\{(Y_{ij}, x_{ij}); i = 1, \dots, m; j = 1, \dots, n_i\}$ as the sampled data. Define the mean of errors from the sampled data for the i^{th} small area as $\bar{\varepsilon}_{is} = \sum_{j=1}^{n_i} \varepsilon_{ij}/n_i$.

Under the true model with known parameters, one can obtain the BUP of the small area means by predicting the population of error terms in each area. This amounts to using the mean of errors for the sampled units and the conditional expectation for the un-sampled units. Hence, we get

$$\hat{\gamma}_i^{BUP} = \bar{X}_i^\top \beta + \frac{n_i}{N_i}\bar{\varepsilon}_{is} + \frac{N_i - n_i}{N_i}\,\mathbb{E}(\varepsilon_{ia}|\varepsilon_{ij}; j \in s_i), \tag{2}$$

where $\bar{X}_i$ is the population means of covariates for the ith small area, $a \in r_i$ is an un-sampled unit of the i^{th} small area, and the conditional expectation, which is the same for all $a \in r_i$, is defined as

$$\begin{aligned}\mathbb{E}(\varepsilon_{ia} \mid \varepsilon_{ij}; j \in s_i) &= \int_{-\infty}^{\infty} z\,\frac{f_{1:n_i+1}(z, \varepsilon_{i,j_1}, \dots, \varepsilon_{i,j_{n_i}})}{f_{1:n_i}(\varepsilon_{i,j_1}, \dots, \varepsilon_{i,j_{n_i}})}\,dz \\ &= \int_{-\infty}^{\infty} z\,\frac{c_{1:n_i+1}\{F_\varepsilon(z), F_\varepsilon(\varepsilon_{i,j_1}), \dots, F_\varepsilon(\varepsilon_{i,j_{n_i}}); \alpha\}}{c_{1:n_i}\{F_\varepsilon(\varepsilon_{i,j_1}), \dots, F_\varepsilon(\varepsilon_{i,j_{n_i}}); \alpha\}}\,f_\varepsilon(z)\,dz \\ &= \int_{-\infty}^{\infty} z\,w_{1i}\{F_\varepsilon(z), F_\varepsilon(\varepsilon_{ij}); j \in s_i, \alpha\}\,dF_\varepsilon(z). \end{aligned} \tag{3}$$

Here, the weight function

$$w_{1i}\{v, F_\varepsilon(\varepsilon_{ij}); j \in s_i, \alpha\} = \frac{c_{1:n_i+1}\{v, F_\varepsilon(\varepsilon_{i,j_1}), \dots, F_\varepsilon(\varepsilon_{i,j_{n_i}}); \alpha\}}{c_{1:n_i}\{F_\varepsilon(\varepsilon_{i,j_1}), \dots, F_\varepsilon(\varepsilon_{i,j_{n_i}}); \alpha\}} \tag{4}$$

comes from the copula representation of the conditional density, where $c_{1:n_i}(u_{i1}, \dots, u_{in_i}; \alpha) = \partial^{n_i} C_{1:n_i}(u_{i,1}, \dots, u_{i,n_i}; \alpha)/\partial_{u_{i,1}} \dots \partial_{u_{i,n_i}}$ is the copula density for the i^{th} small area ($i = 1, \dots, m$).

Provided that n_i is negligible with respect to N_i, the MSPE of the BUP of small area means can be approximated as

$$\begin{aligned}\text{MSPE}(\hat{\gamma}_i^{BUP}) &\approx \mathbb{E}\left\{\text{Cov}(\varepsilon_{ia}, \varepsilon_{ib} \mid \varepsilon_{ij}; j \in s_i)\right\} \\ &= \mathbb{E}[\mathbb{E}\{\varepsilon_{ia}\varepsilon_{ib} \mid \varepsilon_{ij}; j \in s_i\}] - \mathbb{E}[\mathbb{E}\{\varepsilon_{ia} \mid \varepsilon_{ij}; j \in s_i\}^2] \\ &= \mathbb{E}[(\nu_i + \xi_{ia})(\nu_i + \xi_{ib})] - \mathbb{E}[\mathbb{E}\{\varepsilon_{ia} \mid \varepsilon_{ij}; j \in s_i\}^2] \\ &= \sigma^2\rho - \mathbb{E}\{\mathbb{E}(\varepsilon_{ia}|\varepsilon_{ij}; j \in s_i)^2\}, \end{aligned} \tag{5}$$

where $\mathbb{E}[(\nu_i + \xi_{ia})(\nu_i + \xi_{ib})] = \sigma_\nu^2 = \sigma^2\rho$ is used for $a, b \in r_i$ with $a \neq b$.

One can use the best linear prediction of the small area means along with its MSPE using a linear mixed model with known parameters (Prasad & Rao , 1990), however, the linear form only holds under the special case of the Gaussian copula. Hence, the BUP of small area means yields a more general predictor. For details, see Rivest, Verret & Baillargeon (2016).

3. PROPOSED METHODS

Since the model parameters are often unknown, one needs to use a fitted model to formulate the empirical BUP along with its MSPE. Given sample data $\{(Y_{ij}, x_{ij}); i = 1, \ldots, m, j = 1, \ldots, n_i\}$, one can fit the model (1) through a two-stage estimation procedure, by first estimating the marginal error distribution and then the copula parameter.

3.1. Estimation of the Marginal Error Distribution

To estimate the marginal error distribution F_ε, first the regression vector β is estimated using the ordinary least squares (OLS) under the linear regression model and the residuals $e_{ij} = Y_{ij} - x_{ij}^\top \hat{\beta}$ are obtained. The residuals e_{ij} are then used to estimate the marginal error distribution.

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

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

In the absence of a suitable parametric model, one can estimate the marginal error distribution using the empirical cumulative distribution function (cdf) of the residuals, scaled by $n/(n+1)$, i.e.,

$$\tilde{F}_\varepsilon(e) = \frac{1}{n+1} \sum_{i=1}^{m} \sum_{j \in s_i} 1\{e_{ij} \leq e\},$$

where $n = \sum_{i=1}^{m} n_i$. The denominator $1/(n+1)$ is preferred over $1/n$ to avoid computational issues in the evaluation of the copula density at the boundary. This approach is often referred to as rank transformation and is commonly used in copula estimation (Genest, Ghoudi & Rivest , 1995). The pseudo-observations for each small area are then defined as

$$(\tilde{u}_{i,1}, \ldots, \tilde{u}_{i,n_i}) = (\tilde{F}_\varepsilon(e_{i,1}), \ldots, \tilde{F}_\varepsilon(e_{i,n_i})).$$

Note that the OLS estimator $\hat{\beta}$ is consistent for β, and even though it is less efficient than the generalized least squares estimator, it can achieve asymptotic efficiency under some conditions (Watson , 1967; Kruskal , 1968). The OLS estimator is suggested here primarily due to its simplicity; more sophisticated approaches include estimating β by the maximum likelihood method under a mixed-effects model with normality (Rivest, Verret & Baillargeon , 2016). The consistency of the parametric estimator $\hat{F}_\varepsilon(e_{ij}; \hat{\delta})$ for $F_\varepsilon(\varepsilon_{ij}; \delta)$ follows from the consistency of $\hat{\delta}$ for δ and the consistency of e_{ij} for ε_{ij}. Similarly, the consistency of the nonparametric estimator $\tilde{F}_\varepsilon(e_{ij})$ can be derived following the proof of Rivest, Verret & Baillargeon (2015), with a difference that no centring is performed for the residuals in our case.

3.2. Estimation of the Copula Parameter

After obtaining the pseudo-observations, one can estimate the copula parameter α of the multivariate exchangeable copula C_α using the maximum pseudo-copula log-likelihood. This approach differs from the one in Rivest, Verret & Baillargeon (2016) in that the latter uses the empirical Kendall's tau of the residuals to estimate α which does not fully utilize the underlying copula model. On the other hand, our proposed approach relies on the pseudo-copula log-likelihood.

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

$$\ell(\alpha; \{\hat{u}_{i1}, \ldots, \hat{u}_{in_i}\}_{i=1}^m) = \sum_{i=1}^m \log[c_{1:n_i}(\hat{u}_{i1}, \ldots, \hat{u}_{in_i}; \alpha)]. \tag{6}$$

The copula parameter estimate $\hat{\alpha}$ is obtained by maximizing (6).

For the two-stage semi-parametric approach, where the margins are estimated using the rank transformation, the copula parameter estimate $\tilde{\alpha}$ is obtained by maximizing

$$\ell(\alpha; \{\tilde{u}_{i1}, \ldots, \tilde{u}_{in_i}\}_{i=1}^m) = \sum_{i=1}^m \log[c_{1:n_i}(\tilde{u}_{i1}, \ldots, \tilde{u}_{in_i}; \alpha)].$$

The asymptotic distributions of $\hat{\alpha}$ and $\tilde{\alpha}$ are obtained by generalizing the results in, respectively, Joe (2005) and Genest, Ghoudi & Rivest (1995) to clustered data with potentially unequal cluster sizes.

Proposition 1 *Let α_0 be the true copula parameter. Then, under standard regularity conditions, as $m \to \infty$, we have the following results:*

a) $\sqrt{m}(\hat{\alpha} - \alpha_0)$ has an asymptotic normal distribution with mean zero and variance $\nu_{\hat{\alpha}}$.
b) $\sqrt{m}(\tilde{\alpha} - \alpha_0)$ has an asymptotic normal distribution with mean zero and variance $\nu_{\tilde{\alpha}}$.

The proof of Proposition 1 is provided in Appendix.

3.3. Empirical Best Unbiased Prediction (EBUP) of Small Area Means

The EBUP of small area means is constructed from the BUP in (2) with model parameters replaced by their estimates. For brevity, here we focus on the two-stage parametric estimation. We first approximate the conditional expectation in (3) of an un-sampled error term in the i^{th} small area given the residual data using

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

where the weight function is defined as in (4), but with the estimated marginal distribution and copula parameter.

Then, plugging-in the corresponding estimates in (2), we obtain the EBUP of γ_i as

$$\hat{\gamma}_i^{EBUP} = \bar{X}_i^\top \hat{\beta} + \frac{n_i}{N_i}\bar{e}_{is} + \frac{N_i - n_i}{N_i}\hat{e}_i. \tag{7}$$

For the two-stage semi-parametric estimation, the EBUP of small area means, $\tilde{\gamma}_i^{EBUP}$, is obtained with replacing $\hat{e}_i$ and $\hat{F}_\epsilon(e_{i,j})$ in (7) by $\tilde{e}_i$ and $\tilde{F}_\epsilon(e_{i,j})$.

3.4. MSPE of the EBUP of Small Area Means

In the case where the two-stage parametric estimation is employed, the prediction error of $\hat{\gamma}_i^{EBUP}$ can be written by

$$\hat{\gamma}_i^{EBUP} - \gamma_i = (\hat{\gamma}_i^{EBUP} - \hat{\gamma}_i^{BUP}) + (\hat{\gamma}_i^{BUP} - \gamma_i),$$

which yields the following decomposition of the MSPE of $\hat{\gamma}_i^{EBUP}$ as

$$\begin{aligned}
\mathrm{MSPE}(\hat{\gamma}_i^{EBUP}) &= \mathbb{E}\{(\hat{\gamma}_i^{EBUP} - \gamma_i)^2\} \\
&= \mathbb{E}\{(\hat{\gamma}_i^{BUP} - \gamma_i)^2\} + \mathbb{E}\{(\hat{\gamma}_i^{EBUP} - \hat{\gamma}_i^{BUP})^2\} \\
&\qquad\qquad\qquad + 2\,\mathbb{E}\{(\hat{\gamma}_i^{BUP} - \gamma_i)(\hat{\gamma}_i^{EBUP} - \hat{\gamma}_i^{BUP})\} \\
&\equiv M_{1i} \quad + \quad M_{2i} \quad + \quad M_{3i}.
\end{aligned} \tag{8}$$

Here M_{1i} gives $\mathrm{MSPE}(\hat{\gamma}_i^{BUP})$ in (5). In practical settings, it becomes necessary to estimate $\mathrm{MSPE}(\hat{\gamma}_i^{EBUP})$ as it is a function of model parameters. While Rivest, Verret & Baillargeon (2016) employed the jackknife method to estimate $\mathrm{MSPE}(\hat{\gamma}_i^{EBUP})$, this approach can capture only the first two terms in (8) and ignores the cross-product term M_{3i}. Hence, the jackknife method can lead to significant bias in the estimation of $\mathrm{MSPE}(\hat{\gamma}_i^{EBUP})$. In light of this, it is necessary to account for M_{3i} in the estimation procedure of MSPE of the EBUP of small area means.

3.5. Estimation of MSPE of $\hat{\gamma}_i^{EBUP}$

For complex small area models, a closed analytical expression of the MSPE of small area mean predictors is often not available. In such situations, the estimation of the MSPE of the EBUP of small area means becomes cumbersome. One way to tackle this issue is to use re-sampling methods. In order to properly capture all components of the MSPE of the EBUP of small area means, we propose a bootstrap approach. While bootstrap methods have been previously employed in SAE (Hall & Maiti , 2006a,b; González-Manteiga et al., 2008; Torabi , 2012), these methods are not directly applicable to the exchangeable copula model (1). Furthermore, a careful treatment is needed when the two-stage semi-parametric estimation is employed. Below, we first outline the parametric bootstrap method in Algorithm 1 for the case where the two-stage parametric estimation is employed, and then present the semi-parametric bootstrap approach in Algorithm 2 for the case where the two-stage semi-parametric estimation is used.

Both algorithms start with the generation of copula data using the fitted copula $C(\cdot;\ \hat{\alpha})$ to define the 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 and using linear interpolation on the residuals to obtain the error terms. That is, if $\tilde{u}_{ij}^{(b)}$ falls outside of the range of the empirical cdf values, the corresponding bootstrap error $\varepsilon_{ij}^{*(b)}$ takes the minimum or maximum of the residuals in the sample, and if $\tilde{u}_{ij}^{(b)}$ falls between the (ordered) empirical cdf values $\tilde{u}_{i(k)}$ and $\tilde{u}_{i(k+1)}$, the corresponding bootstrap error $\varepsilon_{ij}^{*(b)}$ is obtained via linear interpolation on $e_{i(k)}$ and $e_{i(k+1)}$. 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 obtaining simple random samples from the bootstrap populations, performing estimation in the same way as in the original sample and getting the EBUP of the small area means. Following González-Manteiga et al. (2008), one can show that the parametric bootstrap MSPE estimator is consistent.

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

```
Given the parameter estimates $(\hat\beta, \hat\delta, \hat\alpha)$ from the dataset $\{(Y_{ij}, x_{ij});\ i = 1, \dots, m;\ j = 1, \dots, n_i\}$, perform the following for $b = 1, \dots, B$:

**Bootstrap population:**
  Generate $(\hat{u}_{i1}^{(b)}, \dots, \hat{u}_{iN_i}^{(b)})$ from $C_{1:N_i}(\cdot; \hat\alpha)$, $(i = 1, \dots, m)$.
  Use the inverse cdf method to obtain the bootstrap error terms $\varepsilon_{ij}^{(b)} = F_\varepsilon^{-1}(\hat{u}_{ij}^{(b)}; \hat\delta)$, $(i = 1, \dots, m;\ j = 1, \dots, N_i)$.
  Obtain the response data for the bootstrap population $Y_{ij}^{(b)} = x_{ij}^\top \hat\beta + \varepsilon_{ij}^{(b)}$, $(i = 1, \dots, m;\ j = 1, \dots, N_i)$.
  Compute the mean of the bootstrap population as $\gamma_i^{(b)} = N_i^{-1} \sum_{j=1}^{N_i} Y_{ij}^{(b)}$, $(i = 1, \dots, m)$.

**Bootstrap sample:**
  Get a bootstrap sample $\{(Y_{ij}^{(b)}, x_{ij}); i = 1, ..., m; j = 1, ..., n_i\}$ from the bootstrap population $\{(Y_{ij}^{(b)}, x_{ij}); i = 1, ..., m; j = 1, ..., N_i\}$ using simple random sampling without replacement.
  Perform the two-stage parametric estimation using the bootstrap sample to get the bootstrap estimates $(\hat\beta^{(b)}, \hat\delta^{(b)}, \hat\alpha^{(b)})$.
  Calculate the bootstrap EBUP of small area means
      $$\hat\gamma_i^{EBUP(b)} = \bar{X}_i^\top \hat\beta^{(b)} + \frac{n_i}{N_i} \bar{e}_{is}^{(b)} + \frac{N_i - n_i}{N_i} \hat{e_i}^{(b)}.$$

**MSPE estimation:**
  Using $B$ bootstrap results, calculate the parametric bootstrap MSPE estimate
      $$\mathrm{mspe}_{\mathrm{boot}}(\hat\gamma_i^{EBUP}) = \frac{1}{B} \sum_{b=1}^{B} (\hat\gamma_i^{EBUP(b)} - \gamma_i^{(b)})^2.$$
```

4. SIMULATION STUDY

This section contains the results from simulation studies that evaluate the performance of the proposed approach in comparison to the RVB method (Rivest, Verret & Baillargeon , 2016). The latter was performed using the CopulaSA package available from the authors. Following Rivest, Verret & Baillargeon (2016), we consider a population with $m = 20$ or 40 small areas, where each area consists of $N_i = 200$ units.

The population data are generated as follows: we first generate $R = 500$ independent sets of copula data $\{u_{ij}^{(r)}; i = 1, \dots, m; j = 1, \dots, N_i; r = 1, \dots, R\}$ using the exchangeable copula model (1). For the copula family, we consider the Gaussian, Clayton, Frank, and Gumbel copulas with copula parameter values that correspond to ICC of $\rho = 0.5$. For the marginal error distribution, we consider standard normal distribution or skewed normal distribution with

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

```
Given the parameter estimates $(\tilde{\beta}, \tilde{\delta}, \tilde{\alpha})$ from the dataset $\{(Y_{ij}, x_{ij});\ i = 1, \ldots, m;\ j = 1, \ldots, n_i\}$, perform the following for $b = 1, \ldots, B$:

Bootstrap population:
    1: Generate $(\tilde{u}_{i1}^{(b)}, \ldots, \tilde{u}_{iN_i}^{(b)})$ from $C_{1:N_i}(\cdot; \tilde{\alpha})$, $(i = 1, \ldots, m)$.
    2: Get the bootstrap error terms $\varepsilon_{ij}^{*(b)}$ $(i = 1, \ldots, m;\ j = 1, \ldots, N_i)$ by mapping the quantiles of $\tilde{u}_{ij}^{(b)}$ to those of the empirical cdf values $\tilde{u}_{ij}$ from the sample and using linear interpolation on the residuals.
    3: Obtain the response data for the bootstrap population $Y_{ij}^{*(b)} = x_{ij}^\top \tilde{\beta} + \varepsilon_{ij}^{*(b)}$, $(i = 1, \ldots, m;\ j = 1, \ldots, N_i)$.
    4: Compute the mean of the bootstrap population as $\gamma_i^{*(b)} = N_i^{-1} \sum_{j=1}^{N_i} Y_{ij}^{*(b)}$, $(i = 1, \ldots, m)$.

Bootstrap sample:
    5: Get a bootstrap sample $\{(Y_{ij}^{*(b)}, x_{ij});\ i = 1, ..., m;\ j = 1, ..., n_i\}$ from the bootstrap population $\{(Y_{ij}^{*(b)}, x_{ij});\ i = 1, ..., m;\ j = 1, ..., N_i\}$ using simple random sampling without replacement.
    6: Perform the two-stage semi-parametric estimation using the bootstrap sample to get the bootstrap estimates $(\tilde{\beta}^{(b)}, \tilde{\delta}^{(b)}, \tilde{\alpha}^{(b)})$.
    7: Calculate the bootstrap EBUP of small area means
           $$\tilde{\gamma}_i^{EBUP(b)} = \bar{X}_i^\top \tilde{\beta}^{(b)} + \frac{n_i}{N_i} \bar{e}_{is}^{*(b)} + \frac{N_i - n_i}{N_i} \tilde{e_i}^{(b)}.$$

MSPE estimation:
    Using $B$ bootstrap results, calculate the semi-parametric bootstrap MSPE estimate
           $$\text{mspe}_{\text{boot}}(\tilde{\gamma}_i^{EBUP}) = \frac{1}{B} \sum_{b=1}^{B} (\tilde{\gamma}_i^{EBUP(b)} - \gamma_i^{*(b)})^2.$$
```

mean $\mu_\varepsilon = 0$, variance $\sigma_\varepsilon^2 = 1$, and the skewness parameter $\phi = 0$ or 10, respectively. Here the skewed normal distribution, $SKN(\zeta, \tau, \phi)$ is defined in terms of the location (ζ), scale (τ), and skewness (ϕ) parameters, with conversions $\mu_\varepsilon = \zeta + \tau\kappa\sqrt{2/\pi}$ and $\sigma_\varepsilon^2 = \tau^2(1 - 2\kappa^2/\pi)$, where $\kappa = \frac{\phi}{\sqrt{1+\phi^2}}$. For $\mu_\varepsilon = 0$, $\sigma_\varepsilon^2 = 1$, and $\phi = 10$, we get $\zeta = -1.31$ and $\tau = 1.65$. For each of $R = 500$ generated independent sets of copula data, we obtain the corresponding error terms $\{e_{ij}^{(r)}; i = 1, \ldots, m; j = 1, \ldots, N_i; r = 1, \ldots, R\}$ using the inverse cdf method for each marginal distribution. The auxiliary variable x_{ij} is generated from $N(1, 1)$ independently and kept fixed across the Monte-Carlo replicates. Consequently, $R = 500$ population datasets

$\{Y_{ij}^{(r)}; i = 1, \dots, m; j = 1, \dots, N_i; r = 1, \dots, R\}$ are obtained using the model

$$Y_{ij}^{(r)} = \beta_0 + \beta_1 x_{ij} + e_{ij}^{(r)},$$

with $\beta_0 = \beta_1 = 1$. The population mean of the i^{th} area in the r^{th} simulation run is calculated as

$$\gamma_i^{(r)} = N_i^{-1} \sum_{j=1}^{N_i} Y_{ij}^{(r)}.$$

The sample data $\{(Y_{ij}^{(r)}, x_{ij}); i = 1, \dots, m; j = 1, \dots, n_i\}$ are drawn from the r^{th} population ($r = 1, \dots, R$), using simple random sampling without replacement in each area. For the samples, we consider areas of equal sample size ($n_i = 4$) as well as with varying sample sizes (ranging from 1 to 4). Using the sample data $\{(Y_{ij}^{(r)}, x_{ij}); i = 1, \dots, m; j = 1, \dots, n_i\}$, the model parameters $\beta = (\beta_0, \beta_1)$, δ, and α are estimated as described in Section 3, and the EBUP of small area means $\hat{\gamma}_i^{EBUP(r)}$ is obtained using the parametric approach. Similarly, we obtain the EBUP of small area means $\tilde{\gamma}_i^{EBUP(r)}$ using the semi-parametric approach. For the estimation of the MSPE of these predictors, we employ Algorithm 1 and Algorithm 2 with $B = 100$ bootstrap samples.

Performance of the parametric, semi-parametric, and RVB methods is assessed using the relative bias (RB) and the empirical MSPE (EMSPE) of the EBUP of small area means. For the parametric approach, we calculated

$$\text{RB}(\hat{\gamma}_i^{EBUP}) = \frac{1}{R} \sum_{r=1}^{R} \{\hat{\gamma}_i^{EBUP(r)} - \gamma_i^{(r)}\} / \sum_{r=1}^{R} \gamma_i^{(r)},$$

and

$$\begin{aligned}
\text{EMSPE}(\hat{\gamma}_i^{EBUP}) &= \frac{1}{R} \sum_{r=1}^{R} \{\hat{\gamma}_i^{EBUP(r)} - \gamma_i^{(r)}\}^2 \\
&= \frac{1}{R} \sum_{r=1}^{R} \{\hat{\gamma}_i^{BUP(r)} - \gamma_i^{(r)}\}^2 + \frac{1}{R} \sum_{r=1}^{R} \{\hat{\gamma}_i^{EBUP(r)} - \hat{\gamma}_i^{BUP(r)}\}^2 \\
&\quad + \frac{2}{R} \sum_{r=1}^{R} \{\hat{\gamma}_i^{BUP(r)} - \gamma_i^{(r)}\}\{\hat{\gamma}_i^{EBUP(r)} - \hat{\gamma}_i^{BUP(r)}\} \\
&= \widetilde{M}_{1i} + \widetilde{M}_{2i} + \widetilde{M}_{3i}, \qquad (9)
\end{aligned}$$

where $\hat{\gamma}_i^{BUP(r)}$ is the BUP of the i^{th} small area mean in the r^{th} sample. To evaluate performance of the MSPE estimation of the EBUP of small area means under the three approaches, we define the RB of each MSPE estimator (mspe) as

$$\text{RB}(\text{mspe}_{\text{boot}}(\hat{\gamma}_i^{EBUP})) = \frac{\sum_{r=1}^{R} \text{mspe}_{\text{boot}}(\hat{\gamma}_i^{EBUP(r)})/R}{\text{EMSPE}(\hat{\gamma}_i^{EBUP})} - 1$$

Note that mspe is calculated using bootstrap for the parametric and semi-parametric methods, and jackknife for the RVB method.

We investigate four scenarios in our simulation study: (i) the copula family and the marginal error distribution are correctly specified, (ii) misspecified copula where the Clayton, Frank and Gumbel copula families are misspecified as the Gaussian copula in the estimation, (iii) misspecified margins where the skewed normal distribution is misspecified as a normal distribution in the parametric method, and (iv) different sample sizes in each small area where the sample sizes range from 1 to 4. The results are summarized in Sections 4.1 to 4.4.

4.1. Results Under Correctly Specified Joint Model

We first investigate the estimation performance of the proposed methods in comparison to the RVB method under the scenario where the marginal error distribution and the copula family are correctly specified. Tables 1 and 2 summarize the results for different copula families (Clayton, Gaussian, Frank, or Gumbel), number of small areas ($m = 20$ or 40), and marginal error distributions as normal or skewed normal.

TABLE 1: Decomposition of the EMSPE for the three small area mean predictors (parametric, semi-parametric, and RVB) under the Gaussian, Clayton, Frank, and Gumbel copulas with standard normal ($\phi = 0$) and skewed normal ($\phi = 10$) marginal error distributions for $m = 20$ and 40 small areas, each with sample size $n_i = 4$.

			Parametric				Semi-parametric				RVB			
Copula	m	ϕ	$\widetilde{M}_1$	$\widetilde{M}_2$	$\widetilde{M}_3$	EMSPE	$\widetilde{M}_1$	$\widetilde{M}_2$	$\widetilde{M}_3$	EMSPE	$\widetilde{M}_1$	$\widetilde{M}_2$	$\widetilde{M}_3$	EMSPE
Gaussian	20	0	0.117	0.011	-0.020	0.109	0.117	0.020	-0.028	0.110	0.117	0.019	-0.027	0.109
		10	0.114	0.012	-0.021	0.105	0.114	0.020	-0.030	0.105	0.114	0.020	-0.030	0.104
	40	0	0.108	0.005	-0.011	0.103	0.108	0.010	-0.017	0.102	0.108	0.010	-0.017	0.101
		10	0.109	0.006	-0.010	0.106	0.109	0.012	-0.015	0.106	0.109	0.011	-0.015	0.105
Clayton	20	0	0.083	0.010	-0.011	0.082	0.083	0.021	-0.019	0.085	0.083	0.022	-0.016	0.089
		10	0.083	0.018	-0.010	0.091	0.083	0.019	-0.012	0.089	0.083	0.017	-0.011	0.089
	40	0	0.077	0.005	-0.005	0.077	0.077	0.011	-0.010	0.079	0.077	0.012	-0.009	0.080
		10	0.079	0.009	-0.006	0.082	0.079	0.010	-0.008	0.082	0.079	0.009	-0.006	0.082
Frank	20	0	0.090	0.013	-0.012	0.091	0.090	0.016	-0.014	0.092	0.090	0.015	-0.013	0.091
		10	0.087	0.013	-0.012	0.088	0.087	0.016	-0.015	0.088	0.087	0.015	-0.015	0.087
	40	0	0.081	0.007	-0.007	0.081	0.081	0.009	-0.008	0.082	0.081	0.008	-0.008	0.081
		10	0.077	0.007	-0.007	0.077	0.077	0.009	-0.008	0.077	0.077	0.008	-0.008	0.077
Gumbel	20	0	0.092	0.012	-0.016	0.088	0.092	0.018	-0.021	0.088	0.092	0.016	-0.020	0.088
		10	0.079	0.013	-0.012	0.079	0.079	0.018	-0.018	0.078	0.079	0.016	-0.019	0.075
	40	0	0.083	0.005	-0.005	0.082	0.083	0.009	-0.009	0.082	0.083	0.008	-0.008	0.082
		10	0.071	0.006	-0.005	0.071	0.071	0.008	-0.008	0.072	0.071	0.007	-0.008	0.069

In Table 1, $\widetilde{M}_1$ is obtained by averaging $\widetilde{M}_{1i}$ over all small areas, where $\widetilde{M}_{1i}$ is calculated assuming the model parameters are known (see equation (9)). Hence, $\widetilde{M}_1$ remains the same under the three methods. Similarly, $\widetilde{M}_2$ and $\widetilde{M}_3$ are obtained by averaging $\widetilde{M}_{2i}$ and $\widetilde{M}_{3i}$ over all small areas, noting that $\widetilde{M}_2$ captures the variation due to the estimation of model parameters. It is worth mentioning that the jackknife method used by RVB captures the variations of $\widetilde{M}_1$ and $\widetilde{M}_2$, while the proposed bootstrap methods capture all variations including the cross-product term $\widetilde{M}_3$. As seen in Table 1, $\widetilde{M}_3$ is not ignorable, especially when $m = 20$. The magnitude of

TABLE 2: Percent RB of EBUP, MSPE estimate (mspe), and percent RB of the MSPE estimate of the small area mean predictors of the three methods (parametric, semi-parametric, and RVB) under the Gaussian, Clayton, Frank, and Gumbel copulas with standard normal ($\phi = 0$) and skewed normal ($\phi = 10$) marginal error distributions for $m = 20$ and 40 small areas, each with sample size $n_i = 4$.

			Parametric			Semi-parametric			RVB		
Copula	m	ϕ	RB(%) of EBUP	mspe	RB(%) of mspe	RB(%) of EBUP	mspe	RB(%) of mspe	RB(%) of EBUP	mspe	RB(%) of mspe
Gaussian	20	0	0.026	0.106	-2.3	0.025	0.108	-1.1	0.009	0.114	4.3
		10	0.584	0.108	4.2	0.202	0.110	5.7	0.180	0.116	12.3
	40	0	0.245	0.103	2.0	0.255	0.104	4.3	0.253	0.107	7.7
		10	-0.166	0.105	0.7	-0.323	0.106	2.6	-0.309	0.109	5.4
Clayton	20	0	- 0.777	0.082	0.7	-0.295	0.089	6.8	-1.000	0.100	14.6
		10	0.103	0.088	-2.9	-0.102	0.089	1.2	-0.798	0.098	10.5
	40	0	-0.466	0.076	0.3	-0.096	0.081	4.4	-0.538	0.084	6.4
		10	0.037	0.079	-2.1	0.030	0.081	0.3	-0.270	0.084	4.2
Frank	20	0	0.000	0.088	-1.9	0.157	0.094	3.0	-0.398	0.097	7.1
		10	0.005	0.084	-3.4	0.077	0.089	2.8	-0.512	0.091	5.5
	40	0	-0.396	0.081	1.4	-0.352	0.084	4.1	-0.527	0.086	6.3
		10	-0.193	0.077	0.7	-0.146	0.080	3.8	-0.514	0.081	5.2
Gumbel	20	0	0.103	0.081	-6.5	0.463	0.086	-1.5	0.091	0.089	2.6
		10	-0.048	0.072	-7.4	0.122	0.077	-0.5	-0.294	0.080	7.1
	40	0	-0.185	0.077	-3.9	0.072	0.080	-0.2	-0.135	0.082	1.9
		10	-0.012	0.065	-6.9	0.042	0.067	-4.4	-0.179	0.071	5.2

$\widetilde{M}_3$ is the same as $\widetilde{M}_2$ in most cases and has a negative sign in all cases. As a result of ignoring the cross-product term, the MSPE estimator in the RVB method overestimates the true EMSPE, which leads RB to be positive and larger than the corresponding parametric and semi-parametric bootstrap approaches as shown in Table 2. We also observe that $\widetilde{M}_2$ is uniformly smaller under the parametric approach with the correctly specified marginal distribution than the two semi-parametric methods, both of which are based on ranks. This result is expected since there is a loss in efficiency when using ranks instead of the parametrically estimated marginal distribution (Joe , 2005).

Note that the relative bias of the small area mean predictors behaves differently for the three methods which depend on the form of the copula family. However, in terms of MSPE estimation, for instance, under the Clayton copula with skewed normal error margins, the absolute difference between the RB of the parametric and RVB methods is 7.6 units and between the RB of the semi-parametric and RVB methods is 9.3 units in the case of $m = 20$ small areas.

We also provide boxplots of the absolute RB of the three methods under the four copula families with normal error margins in the case of $m = 20$ which suggest that the variability in the absolute RB is relatively small for the parametric and semi-parametric methods in comparison with the RVB method (Figure 1). We also compare the EMSPE of EBUP for $m = 20$ small areas using the three methods in the case of Gaussian copula with normal error margins (Figure 2). As shown in Figure 2, in most small areas, the EMSPE of EBUP for both parametric and semi-parametric methods is smaller than the corresponding value from the RVB method.

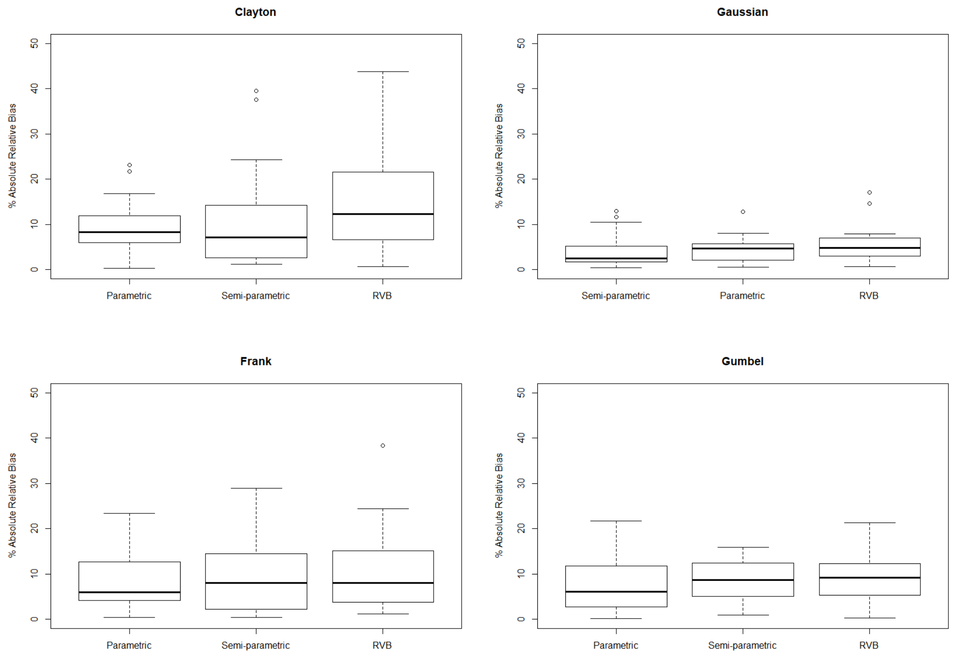

FIGURE 1: Boxplots of absolute RB (%) of different copula families for each method when $m = 20$ and model is correctly specified.

To assess the impact of the cross-product term in the MSPE estimation for a smaller m, we also considered $m = 10$ with standard normal marginal error distribution (see Table A1 in Appendix). For instance, the percent RB of the MSPE estimators (averaged over all small areas) under the Clayton copula are -5.6, 2.1, and 30.5 for the parametric, semi-parametric, and RVB methods, respectively. This suggests that the RVB method has an even worse performance in terms of RB when the population consists of small number of small areas.

We also compare our proposed EBUP methods with the EBLUP where error margins are from the standard normal distribution for different copula families. In Appendix, we provide the result for $m = 20$. As shown in Table A2, the EBLUP performs similar to our EBUP methods in terms of EMSPE in the case of Gaussian copula, however, the efficiency of the EBLUP will be lost for the other copula families, as expected, compared to our proposed EBUP methods.

4.2. Results Under Copula Family Misspecification

We investigate the impact of misspecifying the copula family when fitting the small area model (1) on the EBUP of the small area means and on its MSPE. We focus on the case where the marginal error distribution is standard normal and the true copula families are Clayton, Frank and Gumbel. Table 3 summarizes the EMSPE decomposition for the three methods when the Clayton, Frank, and Gumbel copulas are misspecified as the Gaussian copula. To facilitate comparisons, we also report the results under the correct copula model.

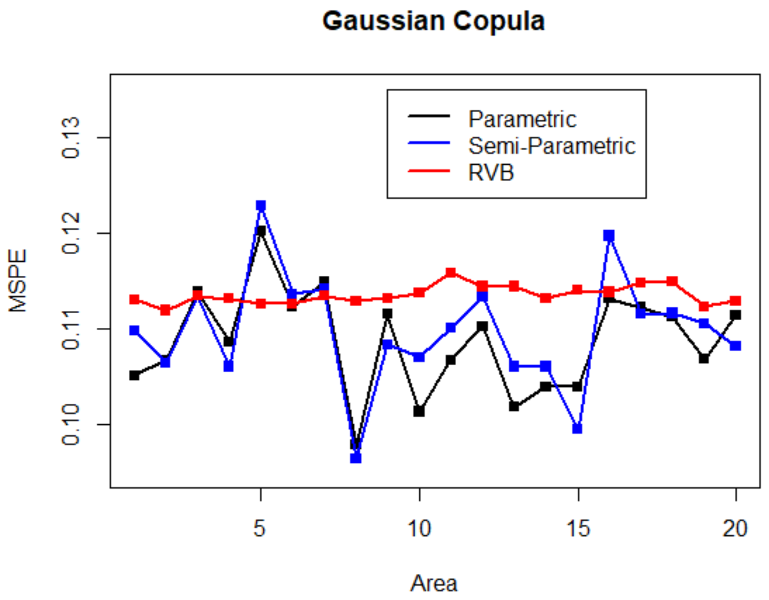

FIGURE 2: MSPE of the EBUP for $m = 20$ small areas in the case of Gaussian copula with normal error margins when model is correctly specified under each small area.

As shown in Table 3, copula family misspecification affects $\widetilde{M}_2$ more adversely than the other two terms; it does not alter $\widetilde{M}_1$ and has only negligible effect on $\widetilde{M}_3$. This result is expected as $\widetilde{M}_2$ captures the variation due to the estimation of the model parameters, which suffers from bias under a misspecified model. The magnitude of $\widetilde{M}_2$ is at least tripled under a misspecified copula in comparison to the case when the copula family is correctly specified, irrespective of the method used for inference in most cases. The results suggest that copula family misspecification can lead to a significant increase in the MSPE of the small area mean predictors. Hence, in practice, the choice of copula family requires care (see Section 6 for further discussion).

4.3. Results Under Misspecification of the Marginal Error Distribution

We next assess the impact of misspecifying the marginal error distribution when fitting the small area model (1) on the EBUP of the small area means and on its MSPE. Since such distributional assumptions are only made in the two-stage parametric estimation, this investigation concerns specifically performance of the parametric approach. We consider the case where the marginal error distribution is skewed normal with skewness parameter $\phi = 10$, but misspecified as normal distribution. The results from the three methods with both true and misspecified marginal error distributions are reported in Table 4 under different copula families (Gaussian, Clayton, Frank, or Gumbel) and for different number of small areas ($m = 20$ or 40).

As seen in Table 4, misspecification of the marginal error distribution affects the results of the EMSPE only in the parametric method, and mainly in terms of $\tilde{M}_2$. The EMSPE of small area means remains the same in the semi-parametric and RVB methods. In practice, depending on the available information, one would move forward with either a parametric or a semi-parametric approach in order to achieve good performance in terms of the EMSPE of small area mean

TABLE 3: Decomposition of the EMSPE for the three small area mean predictors (parametric, semi-parametric, and RVB) when the true Clayton, Frank and Gumbel copula is misspecified as Gaussian copula in the case of standard normal marginal errors for $m = 20$ and 40 small areas, each with sample size $n_i = 4$.

			Parametric				Semi-parametric				RVB			
		m	$\tilde{M}_1$	$\tilde{M}_2$	$\tilde{M}_3$	EMSPE	$\tilde{M}_1$	$\tilde{M}_2$	$\tilde{M}_3$	EMSPE	$\tilde{M}_1$	$\tilde{M}_2$	$\tilde{M}_3$	EMSPE
True Copula	Fitted Copula													
Clayton	Clayton	20	0.083	0.010	-0.011	0.082	0.083	0.021	-0.019	0.085	0.083	0.022	-0.016	0.089
	Clayton	40	0.077	0.005	-0.005	0.077	0.077	0.011	-0.010	0.079	0.077	0.012	-0.009	0.080
Clayton	Gaussian	20	0.084	0.039	-0.016	0.107	0.084	0.048	-0.025	0.108	0.084	0.048	-0.024	0.108
	Gaussian	40	0.079	0.034	-0.008	0.105	0.079	0.040	-0.013	0.106	0.079	0.040	-0.013	0.107
Frank	Frank	20	0.090	0.013	-0.012	0.091	0.090	0.016	-0.014	0.092	0.090	0.015	-0.013	0.091
	Frank	40	0.081	0.007	-0.007	0.081	0.081	0.009	-0.008	0.082	0.081	0.008	-0.008	0.081
Frank	Gaussian	20	0.091	0.043	-0.021	0.113	0.091	0.043	-0.023	0.111	0.091	0.041	-0.023	0.108
	Gaussian	40	0.084	0.035	-0.007	0.112	0.084	0.033	-0.008	0.108	0.084	0.032	-0.008	0.107
Gumbel	Gumbel	20	0.092	0.012	-0.016	0.088	0.092	0.018	-0.021	0.088	0.092	0.016	-0.020	0.088
	Gumbel	40	0.083	0.005	-0.005	0.082	0.083	0.009	-0.009	0.082	0.083	0.008	-0.008	0.082
Gumbel	Gaussian	20	0.089	0.035	-0.012	0.112	0.089	0.039	-0.017	0.111	0.089	0.038	-0.016	0.111
	Gaussian	40	0.081	0.029	-0.006	0.104	0.081	0.033	-0.009	0.105	0.081	0.033	-0.009	0.105

predictions. In the absence of a suitable parametric model for the marginal error distribution, we recommend employing the semi-parametric approach.

The investigations so far concern the choices made by a user when fitting the small area model (1) to obtain the EBUP of small area means. For simplicity, we assumed that the sample sizes are the same in each small area. However, the study design, particularly the choice of sample size may also affect the results.

4.4. Different Sample Sizes

This section evaluates performance of the all three methods when the sample size is different in small areas. We consider the case of $m = 20$ small areas with sample sizes ranging from 1 to 4, focusing on the Frank copula with the standard normal marginal error distribution. In the study design, areas $1 - 5$ have sample size 1, areas $6 - 10$ have sample size 2, areas $11 - 15$ have sample size 3, and areas $16 - 20$ have sample size 4.

The results based on this experiment are summarized in Tables 5 and 6, along with the MSPE estimation. Turning to the evaluation of the three approaches, the EMSPE of small area mean predictors decreases with an increasing sample size. Also, the EMSPE of small area mean predictors in this scenario (sample sizes vary from 1 to 4) is consistently larger than the corresponding values in Table 2, where the sample sizes are 4 in all small areas.

While, in general, the parametric approach shows a smaller RB with increasing sample sizes (see Table 6), no obvious pattern is noticed in the semi-parametric and the RVB methods. These results suggest that the parametric method successfully captures the variations of true EMSPE with varying sample sizes. On the other hand, considering the empirical versions of the estimation procedure, the semi-parametric approach performs better than the RVB method in terms of the RB of the MSPE estimation of the small area mean predictions.

TABLE 4: Decomposition of the EMSPE for the three small area mean predictors (parametric, semi-parametric, and RVB) under the Gaussian, Clayton, Frank, and Gumbel copulas when the true (denoted by *) skewed normal ($\phi = 10$) marginal error distribution is misspecified as normal distribution ($\phi = 0$) for $m = 20$ and 40 small areas, each with sample size $n_i = 4$.

			Parametric				Semi-parametric				RVB			
Copula	ϕ	m	$\tilde{M}_1$	$\tilde{M}_2$	$\tilde{M}_3$	EMSPE	$\tilde{M}_1$	$\tilde{M}_2$	$\tilde{M}_3$	EMSPE	$\tilde{M}_1$	$\tilde{M}_2$	$\tilde{M}_3$	EMSPE
Gaussian		True margin												
	10^*	20	0.114	0.012	-0.021	0.105	0.114	0.020	-0.030	0.105	0.114	0.020	-0.030	0.104
		40	0.109	0.006	-0.010	0.106	0.109	0.012	-0.015	0.106	0.109	0.011	-0.015	0.105
		Misspecified margin												
	0	20	0.115	0.034	-0.018	0.131	0.115	0.021	-0.031	0.106	0.115	0.021	-0.031	0.105
		40	0.109	0.030	-0.006	0.133	0.109	0.012	-0.015	0.106	0.109	0.011	-0.015	0.105
Clayton		True margin												
	10^*	20	0.083	0.018	-0.010	0.091	0.083	0.019	-0.012	0.089	0.083	0.017	-0.011	0.089
		40	0.079	0.009	-0.006	0.082	0.079	0.010	-0.008	0.082	0.079	0.009	-0.006	0.082
		Misspecified margin												
	0	20	0.082	0.037	-0.011	0.109	0.082	0.017	-0.013	0.086	0.082	0.016	-0.011	0.087
		40	0.079	0.035	-0.005	0.109	0.079	0.011	-0.007	0.083	0.079	0.010	-0.006	0.083
Frank		True margin												
	10^*	20	0.087	0.013	-0.012	0.088	0.087	0.016	-0.015	0.088	0.087	0.015	-0.015	0.087
		40	0.077	0.007	-0.007	0.077	0.077	0.009	-0.008	0.077	0.077	0.008	-0.008	0.077
		Misspecified margin												
	0	20	0.083	0.022	-0.011	0.094	0.083	0.016	-0.015	0.084	0.083	0.015	-0.015	0.083
		40	0.076	0.017	-0.004	0.089	0.076	0.008	-0.007	0.077	0.076	0.008	-0.007	0.077
Gumbel		True margin												
	10^*	20	0.079	0.013	-0.012	0.079	0.079	0.018	-0.018	0.078	0.079	0.016	-0.019	0.075
		40	0.071	0.006	-0.005	0.071	0.071	0.008	-0.008	0.072	0.071	0.007	-0.008	0.069
		Misspecified margin												
	0	20	0.077	0.027	-0.015	0.088	0.077	0.018	-0.018	0.077	0.077	0.016	-0.019	0.074
		40	0.068	0.020	-0.010	0.078	0.068	0.008	-0.009	0.067	0.068	0.008	-0.009	0.067

TABLE 5: Decomposition of the EMSPE for the three small area mean predictors (parametric, semi-parametric, and RVB) under the Frank copula with standard normal marginal errors for $m = 20$ small areas, with different sample sizes.

		Parametric				Semi-parametric				RVB			
Area	n_i	$\widetilde{M}_1$	$\widetilde{M}_2$	$\widetilde{M}_3$	EMSPE	$\widetilde{M}_1$	$\widetilde{M}_2$	$\widetilde{M}_3$	EMSPE	$\widetilde{M}_1$	$\widetilde{M}_2$	$\widetilde{M}_3$	EMSPE
1-5	1	0.267	0.027	-0.011	0.282	0.267	0.031	-0.014	0.284	0.267	0.035	-0.013	0.289
6-10	2	0.169	0.026	-0.019	0.176	0.169	0.033	-0.023	0.179	0.169	0.034	-0.022	0.180
11-15	3	0.117	0.024	-0.020	0.121	0.117	0.030	-0.023	0.124	0.116	0.032	-0.024	0.124
16-20	4	0.096	0.021	-0.018	0.099	0.096	0.026	-0.022	0.100	0.096	0.026	-0.022	0.100
1-20		0.162	0.024	-0.017	0.169	0.162	0.030	-0.020	0.172	0.162	0.032	-0.020	0.174

5. APPLICATION: LANDSAT DATA

In this section, we analyze the Landsat dataset (Battese, Harter & Fuller , 1988), considering 12 counties in north central Iowa for predicting the areas planted with soybeans and corn. The areas

TABLE 6: Percent RB of EBUP, MSPE estimate (mspe), and percent RB of the MSPE estimate of the small area mean predictors of the three methods (parametric, semi-parametric, and RVB) under the Frank copula with standard normal marginal errors for $m = 20$ small areas, with different sample sizes.

		Parametric			Semi-parametric			RVB		
Area	n_i	RB(%) of EBUP	mspe	RB(%) of mspe	RB(%) of EBUP	mspe	RB(%) of mspe	RB(%) of EBUP	mspe	RB(%) of mspe
1-5	1	1.396	0.253	-10.3	1.339	0.270	-4.9	1.394	0.312	7.9
6-10	2	0.578	0.165	-6.2	0.533	0.177	-1.1	0.022	0.207	15.0
11-15	3	-0.088	0.121	0.8	0.034	0.123	6.5	-0.564	0.147	18.5
16-20	4	-0.094	0.099	-1.0	0.101	0.100	5.0	-0.438	0.110	10.0
1-20		0.447	0.159	-5.9	0.502	0.167	-2.9	0.103	0.194	12.7

of soybeans and corn are determined by interviewing farm operators, which featured 36 segments in 12 counties. Approximately 250 hectares (each pixel is about 0.45 hectare) are represented by each segment. Further, each segment is estimated from satellite images by counting the number of individual pixels. The aim of this study is to predict the mean hectares of corn in each county (small area). We fit the following model to these data:

$$y_{ij} = \beta_0 + x_{ij1}\beta_1 + x_{ij2}\beta_2 + \varepsilon_{ij},\ i = 1, \ldots, 12;\ j = 1, \ldots, N_i,$$

where x_{ij1} is the number of pixels classified as corn, and x_{ij2} is the number of pixels classified as soybeans in the jth segment of the ith county. We report the small area mean predictors in Table 7, along with the root mean square prediction error (Rmspe) estimation where $B = 1000$ bootstrap samples are used in the bootstrap approach. The residuals display an asymmetric distribution in small areas. Therefore, we consider a skewed normal distribution in the parametric approach. The estimated model parameters of fixed effects are $\hat{\beta}_0 = 44.40, \hat{\beta}_1 = 0.33, \hat{\beta}_2 = -0.10,$ and the estimated parameters of skewed normal are 15.15, 22.00, and -1.66 for location, scale, and skewness parameters, respectively, indicating that the residuals are not normal. For the choice of copula family, the maximum log-likelihood is obtained under the Frank copula, and thus, estimation is carried forward by comparing the EBUP of small area means in each method under the Frank copula family. The log-likelihood values are -4.21 and -3.75, average bootstrap residuals are 145.3 and 160.1, and the Kendall's tau values are 0.27 and 0.28, respectively, for the parametric and semi-parametric approaches.

As shown in Table 7, in the cases of parametric and semi-parametric methods, the Rmspe values significantly decrease with an increase in the number of sample segments (sizes). Furthermore, the improvement in Rmspe is modest when the sample size is greater than 3 or more. However, the 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. In the MSPE estimation of the RVB method, the Rmspe values do not decrease with increasing sample sizes unlike the parametric and semi-parametric approaches. A possible explanation for the higher magnitude of Rmspe for counties 4 and 7 in the RVB method is the use of jackknife, which does not capture the all variations of the EBUP of small area means, as also shown in the simulation study.

TABLE 7: EBUP of small area means and corresponding Rmspe for Landsat data under Frank copula.

		Parametric		Semi-parametric		RVB	
County	n_i	EBUP	Rmspe	EBUP	Rmspe	EBUP	Rmspe
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

6. CONCLUSION

Following Rivest, Verret & Baillargeon (2016), this paper has considered a general linear model framework in small area estimation where the joint error distribution belongs to the family of multivariate exchangeable copulas. Under the same setting as in Rivest, Verret & Baillargeon (2016), unit-level regression model, the two-stage parametric and the two-stage semi-parametric methods for predicting small area means have been proposed, where the copula parameter was estimated via maximizing pseudo-copula log-likelihood. The paper has contributed a bootstrap approach to obtain a nearly-unbiased estimate of the MSPE of the EBUP of small area means.

Extensive simulation studies have been provided to evaluate performance of the proposed methods in comparison to the RVB method of Rivest, Verret & Baillargeon (2016) under various scenarios addressing the impact of misspecifying the copula family, misspecifying the marginal error distribution, and having a study design with different sample sizes in each small area. Overall, the parametric approach has shown relatively better performance in the EBUP of the small area means and the MSPE estimation of the EBUP when the model is correctly specified. While the semi-parametric and the RVB methods have given similar results in most of the scenarios, the semi-parametric method has shown an improved performance in terms of RB of the MSPE estimation of the EBUP of small area means.

An important finding in the paper was that, unlike the traditional small area model based on multivariate normal distribution, the cross-product term involved in the MSPE of the EBUP of small area means was not negligible in the multivariate exchangeable copula model as empirically shown in Section 4. Our results have indicated that the proposed bootstrap method was able to capture all variations in the MSPE of the EBUP of small area means including the cross-product term, hence outperformed the jackknife method used in the RVB method. We could also use the double bootstrap method (Hall & Maiti , 2006a,b) to further improve the MSPE estimation of the EBUP of small area means. In consideration with the computational cost, we recommend the single phase bootstrap method for the estimation of the MSPE of the EBUP of small area means as we got reasonably good results in terms of RB for our unit-level regression model.

Our investigations on the impact of the copula family misspecification have suggested that the choice of copula family is critical for the MSPE estimation in practical settings. A particular challenge in the current framework is that, unlike many copula applications, copula family selection based on visual inspections is not straightforward due to the small and possibly different number of observations across small areas. While one can exploit the exchangeability and invariance properties of the copula and inspect, for instance, all pairwise observations within an area, or choose a copula family based on the maximum (pseudo) likelihood criterion or cross-validated MSPE, a more elaborate investigation is needed. The estimation of regression coefficients can be improved using an iterative reweighted least squares instead of the ordinary least squares method. This, however, comes at an increased computational cost and may not guarantee an overall improvement in efficiency in the estimation of other model parameters. The choice of the marginal error distribution is relevant only for the parametric approach. Hence, one can employ the complementary semi-parametric approach in the absence of a suitable parametric distribution.

In conclusion, a major advantage of proposed likelihood framework is that it provides a more flexible approach for small area estimation allowing for both parametric and semi-parametric estimation methods. The proposed approach can be extended to settings involving discrete outcomes. This is a subject of future study.

ACKNOWLEDGEMENTS

Constructive comments and suggestions of the editor, an associate editor, and three referees, which led to an improved version of this article, are greatly appreciated. This research is a part of MSc thesis work of Kanika Grover. This work was supported by research grants from the Natural Sciences and Engineering Research Council of Canada (NSERC) to Elif Acar and Mahmoud Torabi.

SUPPLEMENTARY MATERIALS

The supplementary materials contain R codes and corresponding “readme” files for the simulations and application conducted in this paper.

APPENDIX

Proof of Proposition 1.

The proof of Proposition 1 follows similar arguments as in Andersen (2005); Prenen, Braekers & Duchateau (2017). Without loss of generality, denote by δ the parameter vector of the marginal distribution and by $\ell(\delta, \alpha)$ the log-likelihood function of marginal and dependence parameters. Let $(\bar{\delta}, \bar{\alpha})$ be the maximum likelihood estimator obtained simultaneously from the score equations $s_\delta(\delta, \alpha) = \partial\ell(\delta, \alpha)/\partial\delta = 0$ and $s_\alpha(\delta, \alpha) = \partial\ell(\delta, \alpha)/\partial\alpha = 0$. Under standard regularity conditions, $\sqrt{m}\,(\bar{\delta} - \delta_0, \bar{\alpha} - \alpha_0)$ converges to a normal distribution with mean zero and variance-covariance matrix $\mathbf{I}^{-1}$, where $\mathbf{I}$ is the Fisher information matrix with components

$$\mathbf{I} = \begin{pmatrix} \mathbf{I}_{\delta\delta} & \mathbf{I}_{\delta\alpha} \\ \mathbf{I}_{\alpha\delta} & \mathrm{I}_{\alpha\alpha} \end{pmatrix}.$$

Hence, for the maximum likelihood estimator $\bar{\alpha}$, the variance is given by

$$\mathrm{Var}(\bar{\alpha}) = \frac{1}{\mathrm{I}_{\alpha\alpha}} + \frac{\mathbf{I}_{\alpha\delta}{\mathbf{I}_{\delta\delta}}^{-1}\mathbf{I}_{\delta\alpha}}{{\mathrm{I}_{\alpha\alpha}}^{2}}.$$

Now, consider the two-stage parametric estimator, where $\hat{\delta}$ is obtained from the score equation $s_\delta(\delta) = \partial\ell(\delta)/\partial\delta = 0$. By Taylor expansion of the score function around the true parameter δ_0, we have

$$s_\delta(\hat{\delta}) = 0 = s_\delta(\delta_0) + \left.\frac{\partial s_\delta(\delta)}{\partial\delta}\right|_{\delta=\delta_0}(\hat{\delta}-\delta_0) + o_p(\sqrt{m}).$$

Using the contributions of small areas to the score function, we obtain

$$-\frac{1}{m}\left.\frac{\partial s_\delta(\delta)}{\partial\delta}\right|_{\delta=\delta_0} = -\frac{1}{m}\sum_{i=1}^{m}\frac{\partial}{\partial\delta}s_{i;\delta}(\delta_0)$$

which converges in probability to $\mathbf{I}^* = \mathbb{E}\left[\frac{\partial}{\partial\delta}\mathrm{s}_{1;\delta}(\delta_0)\right]$ by the law of large numbers.

The two-stage parametric estimator $\hat{\alpha}$ is obtained from the score equation $s_\alpha(\hat{\delta},\alpha) = \partial\ell(\hat{\delta},\alpha)/\partial\alpha = 0$. Expanding this function via Taylor approximation around the true parameters δ_0 and α_0, we get

$$\begin{aligned} s_\alpha(\hat{\delta},\hat{\alpha}) = 0 = s_\alpha(\delta_0,\alpha_0) + & \left.\frac{\partial s_\alpha(\delta,\alpha)}{\partial\delta}\right|_{(\delta,\alpha)=(\delta_0,\alpha_0)}(\hat{\delta}-\delta_0) \\ + & \left.\frac{\partial s_\alpha(\delta,\alpha)}{\partial\alpha}\right|_{(\delta,\alpha)=(\delta_0,\alpha_0)}(\hat{\alpha}-\alpha_0) + o_p(\sqrt{m}). \end{aligned}$$

By the law of large numbers, we obtain the following convergence in probability results,

$$-\frac{1}{m}\left.\frac{\partial s_\alpha(\delta,\alpha)}{\partial\delta}\right|_{(\delta,\alpha)=(\delta_0,\alpha_0)} = -\frac{1}{m}\sum_{i=1}^{m}\frac{\partial}{\partial\delta}s_{i;\alpha}(\delta_0,\alpha_0) \xrightarrow{\mathrm{P}} \mathbf{I}_{\alpha\delta},$$

and

$$-\frac{1}{m}\left.\frac{\partial s_\alpha(\delta,\alpha)}{\partial\alpha}\right|_{(\delta,\alpha)=(\delta_0,\alpha_0)} = -\frac{1}{m}\sum_{i=1}^{m}\frac{\partial}{\partial\alpha}s_{i;\alpha}(\delta_0,\alpha_0) \xrightarrow{\mathrm{P}} \mathrm{I}_{\alpha\alpha}.$$

Combining these results, the following terms are asymptotically equivalent in terms of limiting distributions:

$$\frac{1}{\sqrt{m}}\begin{pmatrix} s_\delta(\delta_0) \\ s_\alpha(\delta_0,\alpha_0) \end{pmatrix} \overset{\mathrm{d}}{\equiv} \sqrt{m}\begin{pmatrix} \mathbf{I}^* & 0 \\ \mathbf{I}_{\alpha\delta} & \mathrm{I}_{\alpha\alpha} \end{pmatrix}\begin{pmatrix} \hat{\delta}-\delta_0 \\ \hat{\alpha}-\alpha_0 \end{pmatrix}.$$

Also, by the central limit theorem, we have

$$\frac{1}{\sqrt{m}}\begin{pmatrix} s_\delta(\delta_0) \\ s_\alpha(\delta_0,\alpha_0) \end{pmatrix} \xrightarrow{\mathrm{d}} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \mathbf{V} & 0 \\ 0 & \mathrm{I}_{\alpha\alpha} \end{pmatrix}\right),$$

where $\mathbf{V} = \mathrm{Var}(\mathrm{s}_{1,\delta}(\delta_0))$ with $s_{1,\delta}$ the contribution of the first small area to the score function in the first stage estimation. Hence, we obtain

$$\sqrt{m}\begin{pmatrix} \hat{\delta}-\delta_0 \\ \hat{\alpha}-\alpha_0 \end{pmatrix} \xrightarrow{\mathrm{d}} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \mathbf{I}^* & 0 \\ \mathbf{I}_{\alpha\delta} & \mathrm{I}_{\alpha\alpha} \end{pmatrix}^{-1}\begin{pmatrix} \mathbf{V} & 0 \\ 0 & \mathrm{I}_{\alpha\alpha} \end{pmatrix}\begin{pmatrix} \mathbf{I}^* & 0 \\ \mathbf{I}_{\alpha\delta} & \mathrm{I}_{\alpha\alpha} \end{pmatrix}^{-1^{T}}\right),$$

which yields the desired result with the variance term

$$\nu_{\hat{\alpha}} = \frac{1}{\mathrm{I}_{\alpha\alpha}} + \frac{\mathbf{I}_{\alpha\delta}(\mathbf{I}^{*})^{-1}\mathbf{V}(\mathbf{I}^{*})^{-1}\mathbf{I}_{\delta\alpha}}{\mathrm{I}^{2}_{\alpha\alpha}}.$$

The proof for part (b) can be obtained mimicking the proof in Genest, Ghoudi & Rivest (1995).

■

Results for Correctly Specified Joint Model in the Case of m=10.

We also examined the case when number of small areas decreases to 10 from 20 under the same setting as defined in the sub-section 4.1, where error margins are from standard normal distribution. The results are summarized in Table A1. As expected, the cross-product term and corresponding RB are relatively higher in all cases and in particular for the RVB method in the case of $m = 10$ compared to $m = 20$. 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.

TABLE A1: Decomposition of EMSPE, percent RB of EBUP, MSPE estimate (mspe), and percent RB of MSPE estimate of small area mean predictors for three different methods (parametric, semi-parametric, and RVB) when copula family (Clayton, Gaussian, Frank, or Gumbel) and margins (normal) are correctly specified for $m = 10$.

	Parametric				Semi-parametric				RVB			
Copula	$\tilde{M}_1$	$\tilde{M}_2$	$\tilde{M}_3$	EMSPE	$\tilde{M}_1$	$\tilde{M}_2$	$\tilde{M}_3$	EMSPE	$\tilde{M}_1$	$\tilde{M}_2$	$\tilde{M}_3$	EMSPE
Gaussian	0.132	0.023	-0.039	0.116	0.132	0.038	-0.056	0.114	0.132	0.033	-0.051	0.114
Clayton	0.098	0.024	-0.025	0.096	0.098	0.039	-0.038	0.100	0.098	0.036	-0.034	0.100
Frank	0.111	0.025	-0.021	0.115	0.111	0.031	-0.027	0.116	0.111	0.027	-0.024	0.114
Gumbel	0.112	0.032	-0.036	0.108	0.112	0.044	-0.049	0.107	0.111	0.037	-0.046	0.102
	RB(%) of EBUP	mspe	RB(%) of mspe		RB(%) of EBUP	mspe	RB(%) of mspe		RB(%) of EBUP	mspe	RB(%) of mspe	
Gaussian	0.343	0.111	-3.5		0.335	0.116	2.4		0.381	0.132	16.5	
Clayton	-0.649	0.090	-5.6		0.048	0.101	2.1		-1.231	0.129	30.5	
Frank	0.666	0.100	-11.9		0.975	0.109	-4.1		-0.109	0.116	4.4	
Gumbel	-0.202	0.090	-16.0		0.385	0.100	-5.1		-0.287	0.111	10.6	

■

Results of EBUP in comparison to empirical best linear unbiased predictor (EBLUP).

We also examined the case when the data was estimated using the EBLUP (Rao & Molina , 2015) for the set-up when the number of small areas is 20 under the same setting as defined in the sub-section 4.1, where error margins are from standard normal distribution. The estimated mean squared prediction error (mspe) was calculated using parametric bootstrap (González-Manteiga et al., 2008). The results are summarized in Table A2. As shown in Table A2, the EMSPE of EBLUP in the case of Gaussian copula is close to the other EBUP methods, however, in the case of other copula families, the EMSPE of EBLUP provides worse results compared to the other EBUP methods. In the case of RB of small area mean predictors and mspe, the EBLUP method behaves similar to the proposed EBUP methods.

■

TABLE A2: Decomposition of EMSPE of EBLUP and EBUP (parametric, semi-parametric, and RVB), percent RB of EBLUP and EBUP, MSPE estimate (mspe), and percent RB of MSPE estimate of small area mean predictors when copula family (Clayton, Gaussian, Frank, or Gumbel) and error margins (normal) are correctly specified for $m = 20$.

	EBLUP				Parametric				Semi-Parametric				RVB			
Copula	$\tilde{M}_1$	$\tilde{M}_2$	$\tilde{M}_3$	EMSPE	$\tilde{M}_1$	$\tilde{M}_2$	$\tilde{M}_3$	EMSPE	$\tilde{M}_1$	$\tilde{M}_2$	$\tilde{M}_3$	EMSPE	$\tilde{M}_1$	$\tilde{M}_2$	$\tilde{M}_3$	EMSPE
Gaussian	0.099	0.007	0.000	0.105	0.117	0.011	-0.020	0.109	0.117	0.020	-0.028	0.110	0.117	0.019	-0.027	0.109
Clayton	0.099	0.006	-0.002	0.103	0.083	0.010	-0.011	0.082	0.083	0.021	-0.019	0.085	0.083	0.022	-0.016	0.089
Frank	0.103	0.007	0.001	0.111	0.090	0.013	-0.012	0.091	0.090	0.016	-0.014	0.092	0.090	0.015	-0.013	0.091
Gumbel	0.099	-0.008	-0.002	0.105	0.092	0.012	-0.016	0.088	0.092	0.018	-0.021	0.088	0.092	0.016	-0.020	0.088
	RB(%) of EBLUP	mspe	RB(%) of mspe		RB(%) of EBUP	mspe	RB(%) of mspe		RB(%) of EBUP	mspe	RB(%) of mspe		RB(%) of EBUP	mspe	RB(%) of mspe	
Gaussian	0.005	0.100	-4.3		0.026	0.106	-2.3		0.025	0.108	-1.1		0.009	0.114	4.3	
Clayton	-0.338	0.099	-3.1		- 0.777	0.082	0.7		-0.295	0.089	6.8		-1.000	0.100	14.6	
Frank	0.034	0.105	-4.8		0.000	0.088	-1.9		0.157	0.094	3.0		-0.398	0.097	7.1	
Gumbel	0.236	0.098	-6.0		0.103	0.081	-6.5		0.463	0.086	-1.5		0.091	0.089	2.6	

BIBLIOGRAPHY

Andersen, E.W. (2005). Two-stage estimation in copula models used in family studies. *Lifetime Data Analysis*, 11, 333-350.

Battese, G. E., Harter, R. M., & 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, 28–36.

Chambers, R., Chandra, H., Salvati, N., & Tzavidis, N. (2014), Outlier robust small area estimation. *Journal of the Royal Statistical Society, Series B*, 76, 47–69.

Chambers, R. & Tzavidis, N. (2006). M-quantile models for small area estimation. *Biometrika*, 93, 255–268.

Chandra, H. & Chambers, R. (2011). Small area estimation under transformation to linearity. *Survey Methodology*, 37, 39–51.

Clayton, D. & Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, 43, 671–681.

Datta, G.S., Ghosh, M., & Waller, L.A. (2000). Hierarchical and empirical Bayes methods for environmental risk assessment, in P.K. Sen and C.R. Rao (Eds.), *Handbook of Statistics*, Volume 18, Amsterdam: Elsevier Science B.V., pp. 223–245.

Dean, C.B. & MacNab, Y.C. (2001). Modeling of rates over a hierarchical health administrative structure. *Canadian Journal of Statistics*, 29, 405–419.

Diallo, M.S. & Rao, J.N.K. (2018). Small area estimation of complex parameters under unit-level models with skew-normal errors. *Scandinavian Journal of Statistics*, 45, 1092–1116.

Fay, I.R.E. & 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, 269-277.

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

Ghosh, M., Natarajan, K., Stroud, T.W.F., & Carlin, B.P. (1998). Generalized linear models for small area estimation. *Journal of American Statistical Association*, 93, 273–282.

Ghosh, M., Natarajan, K., Waller, L.A., & Kim, D. (1999), Hierarchical Bayes GLMs for the analysis of spatial data: an application to disease mapping. *Journal of Statistical Planing and Inference*, 75, 305–318.

González-Manteiga, W., Lombardía, M.J., Molina, I., Morales, D., & Santamaría, L. (2008). Bootstrap mean squared error of a small-area EBLUP. *Journal of Statistical Computation and Simulation*, 78, 443–462.

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

Hall, P. & Maiti, T. (2006 b). Nonparametric estimation of mean-squared prediction error in nested-error regression models. *The Annals of Statistics*, 34(4), 1733-1750.

Jiang, J. & Nguyen, T. (2012). Small area estimation via heteroscedastic nested-error regression. *Canadian Journal of Statistics*, 40, 588–603.

Jiongo, V.D., Haziza, D., & Duchesne, P. (2013). Controlling the bias of robust small-area estimation. *Biometrika*, 100, 843–858.

Joe, H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, 94, 401-419.

Kruskal, W. (1968). When are Gauss–Markov and least squares estimators identical? A coordinate-free approach. *Annals of Mathematical Statistics*, 39(1), 70-75.

Langford, I.H., Leyland, A.H., Rasbash, J., & Goldstein, H. (1999). Multilevel modelling of the geographical distribution of diseases. *Applied Statistics*, 48, 253–268.

López-Vizcaíno, E., Lombardía, M.J., & Morales, D. (2013). Multinomial-based small area estimation of labour force indicators. *Statistical Modelling*, 13, 153–178.

López-Vizcaíno, E., Lombardía, M.J., & Morales, D. (2015), Small area estimation of labour force indicators under a multinomial model with correlated time and area effects. *Journal of the Royal Statistical Society, Series A*, 178, 535–565.

MacGibbon, B. & Tomberlin, T.J. (1989). Small area estimation of proportions via empirical Bayes techniques. *Survey Methodology*, 15, 237–252.

Malec, D., Sedransk, J., Moriarity, C.L., & LeClere, F.B. (1997). Small area inference for binary variables in national health interview survey. *Journal of the American Statistical Association*, 92, 815–826.

Molina, I., Saei, A., & Lombardía, M.J. (2007). Small area estimates of labour force participation under a multinomial logit mixed model. *Journal of the Royal Statistical Society, Series A*, 170, 975–1000.

Nandram, B., Sedransk, J., & Pickle, L. (1999). Bayesian analysis of mortality rates for U.S. health service areas. *Sankhya, Series B*, 61, 145–165.

Prasad, N.G.N. & Rao, J.N.K. (1990). The estimation of mean squared error of small-area estimators. *Journal of the American Statistical Association*, 85, 163-171.

Prenen, L., Braekers, R., & Duchateau, L. (2017). Extending the Archimedean copula methodology to model multivariate survival data grouped in clusters of variable size. *Journal of Royal Statistical Society Series B*, 79(2), 483-505.

Rao, J. N.K. & Molina, I. (2015). *Small Area Estimation, 2nd Edition*. Wiley Online Library.

Rivest, L. P., Verret, F., & Baillargeon, S. (2015). Estimation of the parameters in copula models for small areas. *Statistical Society of Canada, Proceedings of the Survey Methods Section 2015*

Rivest, L. P., Verret, F., & Baillargeon, S. (2016). Unit level small area estimation with copulas. *The Canadian Journal of Statistics*, 44(4), 397-415.

Sinha, S.K. & Rao, J.N.K. (2009). Robust small area estimation. *Canadian Journal of Statistics*, 37, 381–399.

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.

Torabi, M. (2019). Spatial generalized linear mixed models in small area estimation. *The Canadian Journal of Statistics*, 47, 426–437.

Torabi, M. & Rao, J. N.K. (2010). Mean squared error estimators of small area means using survey weights. *The Canadian Journal of Statistics*, 38(4), 598-608.

Torabi, M. & Shokoohi, F. (2015). Non-parametric generalized linear mixed models in small area estimation. *The Canadian Journal of Statistics*, 43, 82–96.

Tsutakawa, R.K., Shoop, G.L., & Marienfield, C.J. (1985). Empirical Bayes estimation of cancer mortality rates. *Statistics in Medicine*, 4, 201–212.

Watson, G.S. (1967). Linear least squares regression. *Annals of Mathematical Statistics*, 38(6), 1679-1699.

Received 17 September 2018
Accepted 11 February 2020