Sampling strategies for proportion and rate estimation in a spatially correlated population

Vahid Tadayon[a], Mahmoud Torabi[b,*]

[a]*Department of Statistics, Higher Education Center of Eghlid, Eghlid, Iran*
[b]*Departments of Community Health Sciences & Statistics, University of Manitoba, Winnipeg, Manitoba, Canada*

Abstract

Conventional sampling theory is widely used in environmental and health hazard assessment. However, spatial sampling techniques are among the most efficient methods when sampling units are spatially correlated. Spatial sampling has been introduced and used for population mean estimation. In addition, few works have also been focused for the population proportion estimation. However, in many health-related data applications, we are interested to also know the proportion of non-rare specific health condition (e.g. asthma) or rate of rare specific health condition (e.g. cancer) in each small area rather than overall population to inform public and policy-makers to focus on areas which are most in need. In this paper, we develop design-based and model-based approaches for overall and area-specific proportion and rate estimation. In particular, we expand dependent unit sequential technique method on model-assisted (ranked set sampling) and model-based (small area estimation) approaches which are more efficient than the sampling methods originated from simple random sampling. We evaluate the performance of proposed approaches using stimulation studies and also by a real data application on teen birth rate in Georgia, USA.

Keywords: Dependent Unit Sequential Technique, Proportion/Rate Estimation, Ranked Set Sampling, Small Area Estimation, Spatial Data

1. Introduction

Collecting observations in a two-dimensional framework is the main aim of spatial sampling and has been applied to many disciplines such as soil, mining and health. Two common strategies for spatial sampling are design-based and model-based which differ in

*Mahmoud Torabi
Email address: Mahmoud.Torabi@umanitoba.ca (Mahmoud Torabi)

their elements of randomness [1]. More precisely, in the design-based approach, the only source of randomness comes from selecting locations randomly, where at a given location the value of the variable of interest is assumed to be fixed, but unknown, whereas in the model-based approach each location is not associated with one fixed value, however, with different possible values, each with a defined probability of occurring, thus forming a random variable. Design-based and model-based approaches have been widely used for the population mean estimation of spatially correlated data. In below, we provide more details on the existing approaches for the population mean estimation.

Design-based sampling is generally used for tackling '*how much*' questions and should be used for estimating global properties of the (realized) population of values. Broadly speaking, in the case of a discrete population, the target of inference is a global property such as the *population mean*, say,

$$(1/N)\sum_{i=1}^{N} z\left(s_i\right), \tag{1}$$

where N is the number of members of the population, and s_i is the location of i-th area. If $z\left(s_i\right)$ is quantity value of interest at location i, then (1) is the *population mean* of some specified attribute. Fundamentally, individual $z\left(s_i\right)$ could be the target of inference, however, because design-based estimators disregard most of the information that is available on where the samples are located in the study area, in practice this is either not possible or lead to estimators with poor properties [2]. So, this approach should not be used for the purpose of constructing maps and estimating individual values. All design-based sampling techniques such as systematic sampling, stratified random sampling, cluster sampling and two-step random sampling originate from simple random sampling (SRS) and can only be fully understood with proper knowledge of it [3].

An intuitively simple estimator in the problem of estimating (1) (which we now denote $\bar{Z}$) is given by:

$$(1/n)\sum_{i=1}^{n} z\left(s_i\right) = \bar{z}, \tag{2}$$

where $z\left(s_1\right), \ldots, z\left(s_n\right)$ denote the values in the sample and the error variance of $\bar{z}$ as an estimator of $\bar{Z}$ with finite population correction $f = n/N$ is

$$\frac{1-f}{n}\frac{N}{N-1}\left[\sigma^2 - \frac{2}{N\left(N-1\right)}\sum_{j}\sum_{\substack{i \\ j<i}} Cov\left[z\left(s_j\right), z\left(s_i\right)\right]\right], \tag{3}$$

in which σ^2 is the population variance and the second term inside the square brackets measures the average covariance between all pairs of individuals in the population [2]. The result (3) is based on taking expectations both over the positioning of the randomized sample points and over the distribution of the values $\{Z(s_i)\}$ which have a constant mean (say, μ) and spatial covariance that depends only on distance separation and possibly direction. The estimator (2) can also be used for estimating the proportion of the population when $z(s_i)$ is 0 or 1.

Generally, there is a probabilistic guarantee that each measured observation in a SRS can be considered as a representative of the population. All above mentioned design-based sampling techniques have been developed to provide a good representation of the population. However, another goal in most data collection settings is to minimize the costs associated with obtaining the data. Ranked set sampling (RSS) has been used as a design-based approach for population mean estimation of spatially correlated data. The RSS introduced by [4] in a non-spatial context (and frequently used in spatial analysis), is a relatively recent development that addresses both of these issues. It uses additional information from the population to provide more structure to the data collection process and minimizes the number of measured observations required to achieve the desired precision in making inferences.

To obtain an RSS of n observations from a population, we proceed as follows: first m^2 sample units are drawn at random from the population. These units are then randomly assigned to m sets, each of size m. The m units in each set are ranked from least to greatest on the variable of study without making actual measurement on the units. Let $Z_{[r]}$ $(r = 1, \ldots, m)$ denote the r-th judgment order statistic in a set of size m. The lowest ranked item, $Z_{[1]}$, is selected for quantification from the first set, $Z_{[2]}$ is selected from the second set, and so on until the highest ranked item, $Z_{[m]}$, is selected from the m-th set. In general, in the r-th set the observation having the r-th judgment rank is quantified for $r = 1, \ldots, m$. The entire process is repeated independently $\mathcal{K}$ times (called cycles) to obtain a total ranked set sample of size $m\mathcal{K}$. Thus, to obtain a ranked set sample of size $n = m\mathcal{K}$, a total of $m^2\mathcal{K}$ units must be randomly selected, but only $m\mathcal{K}$ units need to be quantified. The other $m\mathcal{K}(m-1)$ units taken for ranking purposes provide additional information which leads to the improvement of population parameter estimation.

Now let $Z_{[r]k}$ $(r = 1, \ldots, m;\ k = 1, \ldots, \mathcal{K})$ denote the quantified r-th judgment order statistic in the k-th cycle. The RSS estimator of the population mean μ is the average of the RSS observations; that is,

$$\hat{\mu} = \frac{1}{m\mathcal{K}} \sum_{r=1}^{m} \sum_{k=1}^{\mathcal{K}} Z_{[r]k}. \tag{4}$$

On the other hands, model-based approach is of great importance for tackling '*where*'

questions: identifying where in the population threshold values of an attribute are exceeded or where the extreme values are located. This strategy should be used for estimating the parameters of the underlying stochastic model (such as the model mean, say, μ) but not quantities like $\bar{Z}$ (which now represents the mean of just one realization) unless the model is known [2]. In this case, since $\bar{Z}$ is itself a random variable, it is usual to speak of predicting its value which depends on model properties and is optimal with respect to the selected model. Therefore, model-based approach could also be used for predicting values at particular locations and mapping.

Considering model-based approach, two kinds of mean estimation problems have been arisen: the mean of the realized values, $\bar{Z}$, and the mean of the underlying model, μ. The best linear unbiased predictor (BLUP) of $\bar{Z}$ is given by

$$(1/N)\left[\sum_{i=1}^{n} z\left(s_i\right)+\sum_{i=n+1}^{N} \hat{z}\left(s_i\right)\right] \tag{5}$$

where $\hat{z}\left(s_i\right)$ is the BLUP of an unsampled $z\left(s_i\right)$, $i=n+1,\ldots,N$, and $z\left(s_1\right),\ldots,z\left(s_n\right)$ are sample units and the realization of a stochastic model with mean $\mu\left(\cdot\right)$ and variance-covariance matrix Σ. The BLUP of the attribute value at location s_0, which is not included in the sample, is of the form

$$\hat{z}\left(s_0\right)=\mu\left(s_0\right)+\mathbf{c}^{\top}\Sigma^{-1}\left(\mathbf{z}-\boldsymbol{\mu}\right), \tag{6}$$

where $\mathbf{c}=\left(Cov\left[z\left(s_0\right),z\left(s_1\right)\right],\ldots,Cov\left[z\left(s_0\right),z\left(s_n\right)\right]\right)^{\top}$, $\mathbf{z}=\left(z\left(s_1\right),\ldots,z\left(s_n\right)\right)^{\top}$, $\boldsymbol{\mu}=\left(\mu\left(s_1\right),\ldots,\mu\left(s_n\right)\right)^{\top}$ and $\mu\left(s_0\right)$ is the mean evaluated at location s_0. The second term in (6) identifies the simple kriging weights, $\mathbf{c}^{\top}\Sigma^{-1}$, assigned to each datapoint, that yield the BLUP of the unknown attribute value. By contrast, in the design-based strategy $(1/n)\sum_{i=1}^{n} z\left(s_i\right)$ is used as an estimation of the population mean which clearly leads to high variation in particular if the sample size n is small compared to the population size.

What emerges from the literature is that design-based inference may be more robust than model-based inference, but that an appropriate model-based analysis may perform substantially better, provided that the model-based inferences are consistent with the target population [5]. [6] discussed the differences and the choice between the design-based and model-based approaches to sampling. The average suitability of design-based and model-based methods was illustrated in Figure 1 [7]. [8] studied that, in general, design-based method ignores the spatial variation of a population and model-based method ignores the sampling design, while in empirical studies, both of the spatial variation and sampling design influence the inference about the model which may lead to wrong conclusions, if ignored. They proposed to incorporate the population, sample, and inference which constitute a so-called spatial statistic trinity (SST) as a novel approach for integrating spatial autocorrelation and spatial stratified heterogeneity into design-based and model-

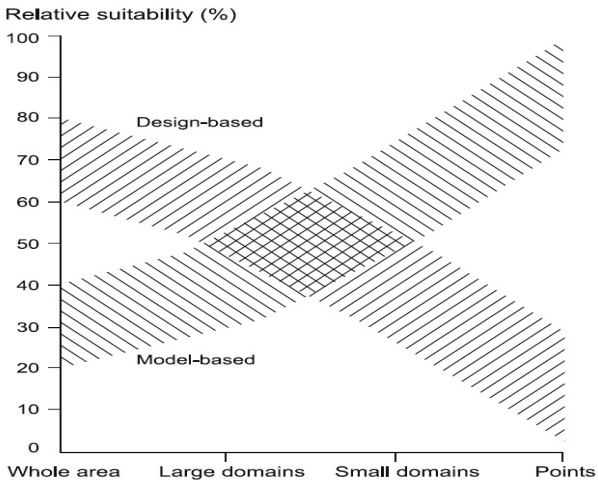

Figure 1: Average relative suitability of the design-based and model-based approaches to sampling, as a function of the spatial resolution at which estimates are required [7].

based framework.

In many health-related data applications, our interest is to estimate population proportion of non-rare specific health condition (e.g. asthma) or estimate population rate of rare specific health condition (e.g. cancer). In particular, [9] used the RSS approach to estimate the population proportion. However, public and policy-makers are also interested to know which (small) areas are more at risk for planning and resource allocation. It also appears that there is no literature on spatial sampling of population rate estimate and also for (small) areas. In this paper, our aim is to address the above issues.

The rest of the paper is organized as follows. In Section 2, we explain design-based and model-assisted approaches, and introduce a model-based approach for the estimation of population proportion and also for small areas. In Section 3, we introduce design-based and model-based approaches for the estimation of population rate and also for small areas. Performance of the proposed approaches is evaluated through stimulation studies (Section 4). We also employ our proposed approach in estimating population rate and also at small areas for the teen birth data in Georgia, USA (Section 5). Finally, concluding remarks are given in Section 6. Computer codes are provided as supplementary materials.

2. Proportion Estimation

When the variable of interest is binary, for instance, diagnosis of asthma, there are only two possible outcomes, present or absent of the asthma where the population mean is equal to the population proportion, say p. In such applications, administrative data initiatives on a health condition may provide unprecedented information about some areas

of interest, e.g., the number of people diagnosed with asthma in small geographic areas. By making use of some auxiliary information, one can help to improve the precision of the proportion estimate [10]. *Spatial dependency* among the adjacent units and *covariates* related to the study binary variable are two important pieces of auxiliary information in making a more efficient statistical inference. In this section, first assuming the proportions of the areal units are known, we estimate the population proportion using the proportions of the areal units in design-based and model-assisted sampling frameworks. Then, in the second part, a more practical problem is developed in a model-based framework where the proportions in the areal units are assumed to be unknown. Subsection 2.1 presents a design-based strategy based on dependent unit sequential technique (DUST) to select more informative samples than the SRS. Subsection 2.2 describes a model-assisted ranking strategy presented by [9] for estimating the proportions of areal units. These proportion estimates are then used to estimate the population proportion. Therefore, for any small area containing respondents to a sample survey, direct estimator for a local area uses sample observations which only come from the sample units in the small area. However, these direct small area estimators typically have low precision due to the fact that the sample sizes in the small areas are disproportionate to the corresponding population. In Subsection 2.3, we develop a model-based approach to estimate the population proportion and also proportion estimate for each small area.

2.1. Design-Based Strategy: Dependent Unit Sequential Technique (DUST)

Spatial dependency among the adjacent areas makes the neighboring areas to be homogeneous. In presence of spatial dependency, near objects are usually more correlated than distant ones, and so, once a particular area is selected in a sample, selection of neighboring areas may not provide considerable information about the underlying population. A spatial sampling method should consider the autocorrelations among the neighbors. This can reduce the costs of surveys based on area sampling whilst maintaining the same level of accuracy.

[10] suggested DUST as a GIS-based sample selection procedure for selection of areal units from spatially correlated population. In this technique, the spatial correlation based on auxiliary character has been used to assign the probability of selection to each area in the population, and so, inclusion probabilities vary at each step. The first area (unit) is selected randomly and the subsequent units are selected sequentially by assigning weights in such a manner that units nearer to earlier selected units in the sample get less probability of selection as compared to the units which are far away from earlier selected unit(s). This sampling procedure is characterized by variable inclusion probabilities at each step. In summary, the DUST works along three steps:

Step I. The spatial autocorrelation γ (which reflects intra-sample correlation, that is, a measure of similarity -correlation- between nearby observations) is estimated at various spatial lags, $\gamma_1, \ldots, \gamma_l$, $l \geq 1$, where γ_l shows the spatial autocorrelation at the lag l.

Table 1: Selection weights for individual sample units.

Sample unit number	Weights
1	1
2	$\left(1-\gamma_1^{d_{12}}\right)$
3	$\left(1-\gamma_1^{d_{13}}\right)\left(1-\gamma_1^{d_{23}}\right)$
$\vdots$	$\vdots$
n	$\prod_{i=1}^{n-1}\left(1-\gamma_1^{d_{in}}\right)$

Let $\mathcal{U}=\{1,2,\ldots,\mathcal{A}\}$ be the set of all areal units in the population. If $\mathcal{P}_i$ and $\mathcal{P}_j$ are respectively the proportions of the binary variable of interest z at i-th and j-th *areal units*, then the spatial autocorrelation γ is given by Moran's I statistic [11] as

$$\gamma=\frac{\sum\sum_{i\neq j=1,\ldots,\mathcal{A}}\omega_{ij}\left(\mathcal{P}_i-\bar{\mathcal{P}}\right)\left(\mathcal{P}_j-\bar{\mathcal{P}}\right)\Big/\sum\sum_{i\neq j=1,\ldots,\mathcal{A}}\omega_{ij}}{\sum_{i=1}^{\mathcal{A}}\left(\mathcal{P}_i-\bar{\mathcal{P}}\right)^2\Big/\mathcal{A}},$$

where $\bar{\mathcal{P}}$ is the population proportion, and ω_{ij} are the weights such that $\omega_{ij}=1$, if i and j are neighbors (e.g., common border) and $\omega_{ij}=0$, otherwise. The spatial autocorrelation between immediate neighbors is shown by γ_1. In order to compute the higher order spatial autocorrelations we assume, for simplicity, $\gamma_l=\gamma^l$ with $l\geq 1$ the lag order.

Step II. Stationarity of the various order correlations is tested usually by the moving sampling techniques [12]. In case not all γ_ls are stationary through space one needs to identify zones where they are at least locally stationary so that each zone can be treated as described in the third step.

Step III. The spatial autocorrelation is employed to assign drawing weights to the individual sample units in the following way: If $\gamma_l=0$, we locate samples randomly as in SRS. If not, the sample is drawn sequentially by assigning a weight varying at each step according to the scheme presented in Table 1, where d_{ij} is the distance between the i-th and the j-th areal unit measured in terms of physical distance between centroids or in terms of the order of neighborhood.

Therefore, any sampling unit has a probability of being drawn which increases as the distance from the areas already sampled increases. It is clear that after a certain distance spatial autocorrelation vanishes, and as a consequence $1-\gamma$ tends to 1 for each unit which means that the criterion for choice is simply randomness. More precisely, in zones displaying a positive spatial autocorrelation, we can save sampling units by scattering

them. In contrast, zones where the spatial autocorrelation is negative need to be sampled more intensely due to the higher irregularity with which data are distributed over space.

2.2. Model-Assisted Strategy: Ranked Set Sampling (RSS)

In addition to spatial dependency, the auxiliary information may also be an easy-to-rank covariate, related to the variable of interest, which helps one ranks a reasonably large set of sample units. Using a logistic regression to aid in the ranking of a binary variable of interest, [9] showed that the estimation of population proportion from a balanced RSS procedure, which features an equal allocation of order statistics in the sample, is at least as efficient as the corresponding SRS estimation.

Let $\mathbf{x}$ denote a vector of explanatory variables. The binomial logistic regression model has the form

$$\text{logit}\left(p_i\right) = \log\left(\frac{p_i}{1-p_i}\right) = \mathbf{x}_i^\top \boldsymbol{\beta}, \tag{7}$$

where p_i is the probability of success at area i, "logit" denotes the logit function and $\boldsymbol{\beta}_{q\times 1}$ is the vector of regression coefficients with corresponding covariates x_i. Let p_r denote the probability of success and $\mathbf{x}_r$ denote the vector of explanatory variables for an area r in a set of size m which selected using SRS (as stated by Chen et al. 9). Accordingly, the estimated probability of success for this area based on a fitted logistic regression model is

$$\hat{p}_r = \frac{\exp\left(\mathbf{x}_\mathbf{r}^\top \hat{\boldsymbol{\beta}}\right)}{1+\exp\left(\mathbf{x}_\mathbf{r}^\top \hat{\boldsymbol{\beta}}\right)}, \qquad r = 1, \ldots, m. \tag{8}$$

Then, the m sample units can be ranked according to their estimated probabilities of success, $\hat{p}_1, \ldots, \hat{p}_m$, and the RSS estimator for the population proportion p is

$$\hat{p}_{RSS} = \frac{1}{m\mathcal{K}} \sum_{r=1}^{m} \sum_{k=1}^{\mathcal{K}} Z_{h[r]k}. \tag{9}$$

However, in the spatially correlated data, it is well-known that if a particular area is selected in the sample, the adjacent areas are not likely to provide any more relevant additional information about the underlying population. An appropriate sampling design for a spatial data is a sampling design which avoids the selection of such neighboring areas. Such a spatial sampling design will be able to reduce the duplication of information which partly contained in areas already sampled. It will also provide more efficient estimation procedure by reducing sample size and survey cost without reducing the reliability of estimators. One may suggest DUST instead of SRS and sample more intensely in the zones where the spatial autocorrelation is negative due to the higher irregularity with

which data are distributed over space. This procedure provides a more efficient estimator for the population proportion in spatially correlated data. We evaluate the performance of this approach in the simulation study.

2.3. Model-Based Strategy: Small Area Estimation (SAE)

In the case that the proportions in the areal units are unknown, direct small area estimators lead to unacceptably large standard errors. This low precision making them statistically unreliable which makes it necessary to find indirect or model based estimators that decrease the standard error for sufficient statistical precision. Small Area Estimation (SAE) models generally provide estimates with adequate precision with borrowing strengths from other resources (e.g., previous surveys and administrative and census data sets) where direct estimation from sample is statistically inadequate [13]. More precisely, this can be done by using the variables of interest from related resources to increase the "effective" sample size. These variables are brought into the estimation process through a model (either implicit or explicit) that provides a link to related areas (domains) through the use of supplementary information related to the variables of interest, such as recent census counts and current administrative records. Therefore, some key areal units are selected through the DUST to form a SAE model. In the context of binomial data, small area methods have been developed using empirical Bayes (EB) and hierarchical Bayes (HB). In this paper, we use the HB approach for the inference since it is straightforward, inferences are "exact" and complex problems using the Markov chain Monte Carlo (MCMC) methods can be handled. Moreover, the HB method is preferred because the EB method may not be feasible for complex models. Proportion for small area i is estimated through the following logit-normal:

- $Z_i | \, \mathcal{P}_i \sim Bin\left(n_i, \mathcal{P}_i\right)$
- $\theta_i = \text{logit}\left(\mathcal{P}_i\right) = \mathbf{x}_i^\top \boldsymbol{\beta} + \nu_i$, with $\nu_i \overset{iid}{\sim} N\left(0, \sigma_\nu^2\right)$
- $\boldsymbol{\beta}$ and σ_ν^2 are mutually independent with $f\left(\boldsymbol{\beta}\right) \propto 1$ and $\sigma_\nu^{-2} \sim Gamma\left(a, b\right)$,

where ν_i is an area-specific random effect and $Z_i = \sum_{j=1}^{n_i} Z_{ij}$, where Z_{ij} is a binary variable for unit j at area i. We note that parameter $\mathcal{P}_i$ is the target of the estimation in this study. One can easily show that:

- $\boldsymbol{\beta} \, | \underline{\mathcal{P}}, \sigma_\nu^2, \mathbf{z} \sim N_q \left[\boldsymbol{\beta}^*, \sigma_\nu^2 \left(\sum_i \mathbf{x}_i \mathbf{x}_i^\top\right)^{-1}\right]; \quad \boldsymbol{\beta}^* = \left(\sum_i \mathbf{x}_i{}^\top \mathbf{x}_i\right)^{-1} \left(\sum_i \mathbf{x}_i^\top \theta_i\right)$
- $\sigma_\nu^2 \, | \boldsymbol{\beta}, \underline{\mathcal{P}}, \mathbf{z} \sim Gamma\left(\frac{A}{2} + a, \frac{1}{2} \sum_i \left(\theta_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2 + b\right)$

- $f\left(\mathcal{P}_i\,|\boldsymbol{\beta},\sigma_\nu^2,\mathbf{z}\right)\propto\mathcal{H}\left(\mathcal{P}_i\,|\boldsymbol{\beta},\sigma_\nu^2\right)\mathcal{K}\left(\mathcal{P}_i\right)$, where $\mathcal{K}\left(\mathcal{P}_i\right)=\mathcal{P}_i^{z_i}(1-\mathcal{P}_i)^{n_i-z_i}$ and

$$\mathcal{H}\left(\mathcal{P}_i\,\middle|\boldsymbol{\beta},\sigma_\nu^2\right)\propto\frac{\partial\theta_i}{\partial\mathcal{P}_i}\exp\left\{-\frac{\left(\theta_i-\mathbf{x}_i^\top\boldsymbol{\beta}\right)^2}{2\sigma_\nu^2}\right\}. \tag{10}$$

The hierarchical Bayes estimate of $\mathcal{P}_i$ and the posterior variance of $\mathcal{P}_i$ are obtained directly from MCMC samples $\left\{\mathcal{P}_1^{(k)},\ldots,\mathcal{P}_m^{(k)},\boldsymbol{\beta}^{(k)},\sigma_\nu^{2(k)}\right\}_{k=d+1}^{d+D}$ generated from the joint posterior $f\left(\mathcal{P}_1,\ldots,\mathcal{P}_m,\boldsymbol{\beta},\sigma_\nu^2\,|\mathbf{Z}\right)$ as

$$\hat{\mathcal{P}}_i^{HB}\approx\frac{1}{D}\sum_{k=d+1}^{d+D}\mathcal{P}_i^{(k)}=\mathcal{P}_i^{(\cdot)},\qquad Var\left(\mathcal{P}_i\,\middle|\hat{\mathcal{P}}_i\right)\approx\frac{1}{D-1}\sum_{k=d+1}^{d+D}\left(\mathcal{P}_i^{(k)}-\mathcal{P}_i^{(\cdot)}\right)^2 \tag{11}$$

[13].

3. Rate Estimation

In many applications, it is common to collect count data observed in spatial locations assuming they have Poisson distribution. For instance, we are interested to know that rate of a specific disease (e.g., lung cancer) at the provincial level and also for small health regions. Our aim in this section is to estimate rate at the population level as well as small area levels in spatially correlated data. In this section, we start assuming small area rates are known. Then, we relax this assumption, similar to the population proportion estimate, and study small area rate for our population study. First, assuming that small areas rates are known, we estimate the population rate using the DUST and SRS approaches and provide their variations through mean squared error (MSE).

To that end, two samples of size ℓ areas (out of $\mathcal{A}$ population areal units) are selected from the population study using DUST and SRS. The set $\lambda_1,\ldots,\lambda_\ell$ is used to show the rates in the areal units selected by DUST and the set $\lambda'_1,\ldots,\lambda'_\ell$ is used to show the rates in the areal units selected by SRS. The population rate estimates based on DUST and SRS are respectively denoted by $\hat{\lambda}=\sum_{h=1}^{\ell}n_h\lambda_h/\sum_{h=1}^{\ell}n_h$ and $\tilde{\lambda}=\sum_{h=1}^{\ell}n_h\lambda'_h/\sum_{h=1}^{\ell}n_h$, where n_h shows the sample size of area h; noting that the areas selected by the DUST method may not be the same areas selected by the SRS method.

In the following, a more practical problem is examined while the rates in the areal units are assumed to be *unknown*. To that end, we consider three approaches: A SRS design is used to estimate the population rate in the areas selected using the SRS (called SRS-SRS) and the DUST (called DUST-SRS). The third approach is DUST-SAE which adopts a SAE approach to estimate the areas selected using the DUST. To do the first approach, a sample of size ℓ areal units is selected using the SRS. The rates of the selected areal units,

denoted by λ_h, $h = 1, \ldots, \ell$, are estimated using SRS which is given by $\tilde{\lambda}_h = z_h/n_h$ where z_h denotes the observed number of cases and n_h is the number of people at risk in the h-th area. The population rate is estimated as $\tilde{\lambda} = \sum_{h=1}^{\ell} n_h \tilde{\lambda}_h / \sum_{h=1}^{\ell} n_h$.

The second approach is similar to the first approach but we only select the areal units using the DUST rather than the SRS. In the third approach, a sample of size ℓ areal units is first selected using the DUST. Then, a conventional approach in SAE is used to estimate the rate at each small area selected by the DUST. In particular, we estimate λ_i through the following HB approach:

- $Z_i \,|\lambda_i \overset{ind}{\sim} P\left(n_i \lambda_i\right)$
- $\xi_i = \log\left(\lambda_i\right) | \, \boldsymbol{\beta}, \sigma^2 \overset{iid}{\sim} N\left(\mathbf{x}_i^\top \boldsymbol{\beta}, \sigma^2\right)$
- $\boldsymbol{\beta}$ and σ^2 are independent with $f\left(\boldsymbol{\beta}\right) \propto 1$ and $\sigma^{-2} \sim Gamma\left(a, b\right)$,

and,

- $f\left(\lambda_i \,|\boldsymbol{\beta}, \sigma^2, \mathbf{z}\right) \propto \lambda_i^{z_i - 1} \exp\left\{-n_i \lambda_i - \frac{1}{2\sigma^2}\left(\xi_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2\right\}$
- $\boldsymbol{\beta} \,|\boldsymbol{\lambda}, \sigma^2, \mathbf{z} \sim N_q\left[\boldsymbol{\beta}^*, \sigma^2\left(\sum_i \mathbf{x}_i \mathbf{x}_i^\top\right)^{-1}\right]$
- $\sigma^2 \,|\boldsymbol{\lambda}, \boldsymbol{\beta}, \mathbf{z} \sim Gamma\left(\frac{\mathcal{A}}{2} + a, \frac{1}{2}\sum_{i=1}^{\mathcal{A}}\left(\xi_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2 + b\right)$,

[14].

4. Simulation Study

4.1. Proportion estimation

In this subsection we illustrate the proposed approaches related to proportion estimation numerically through a simulation study and compare the results with their SRS counterparts. For demonstration purpose, we need to generate spatially correlated binary data. Several authors have proposed different methods for generating correlated binary data. [15] developed a copula method to generate spatial correlated binary data. Based on their proposed method, spatially correlated binary variables $Z\left(s_i\right)$ with $E\left[Z\left(s_i\right)\right] = p\left(s_i\right)$ and $\rho\left[Z\left(s_i\right), Z\left(s_j\right)\right] = \rho_{ij}$ can be generated through the following algorithm:

We assume that $Z\left(s_i\right)$ are spatially correlated based on the spatial correlation of $V\left(s_i\right)$. Let $(0, 10) \times (0, 10)$ is the regular grid under study with $\mathcal{A} = 100$ areas and a random set of locations of size $N = 3500$ has been chosen to generate spatially correlated binary variables $Z\left(s_i\right)$ with $E\left[Z\left(s_i\right)\right] = \mathcal{P}$, for $\mathcal{P} = 0.35$ and 0.65. These locations have been presented in Figure 2. The exponential correlation type $\rho\left(\|h\|\right) = \exp\left\{-\|h\| / \varphi\right\}$ with $\varphi = 3$ was

Algorithm 1 Generate spatially correlated binary data through a copula method

I. Generate normally distributed and spatially correlated $V(s_i)$ with the cumulative distribution function $F(v(s_i))$ and $\rho[V(s_i), V(s_j)] = \rho_{ij}$,
II. Determine $U(s_i) = F(V(s_i))$,
III. Define $Z(s_i) = I\{U(s_i) < p(s_i)\}$.

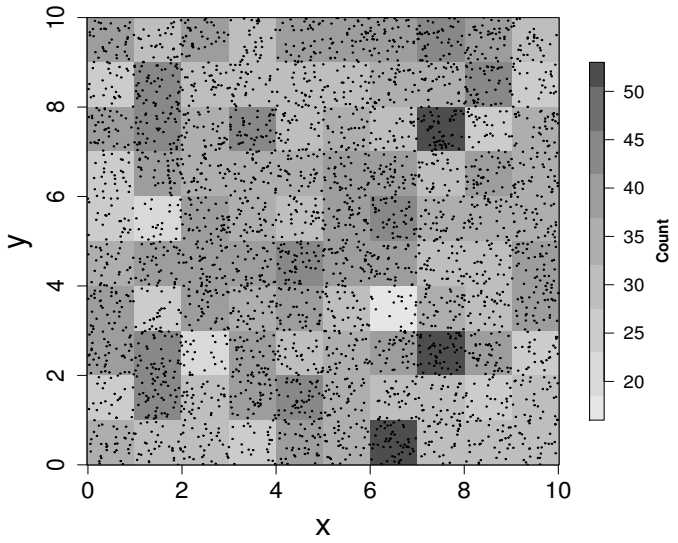

Figure 2: Sampling locations in each area under population study.

considered for $V(s_i)$, which allows spatial correlation to vary freely from 0.01 to 0.99. $V(s_i)$ is generated on the selected locations with the mentioned correlation and mean $\beta_0 + \beta_1 x_i$, where $\beta_0 = 1$, $\beta_1 = 0.5$ and $x_i \sim N(0, 1)$. $\mathcal{R} = 5000$ independent datasets (spatial binary variable $Z(s_i)$) are generated by the simple transformation, $Z(s_i) = I\{F(V(s_i)) < \mathcal{P}\}$.

In each simulation run, $\ell = 20$ areal units are selected through the DUST method and the estimation of population proportion is calculated. We also estimate the population proportion using a SRS and with the same sample size as in the DUST ($\ell = 20$). We observed that the spatial autocorrelation vanishes after the third lag as $\gamma_3 \leq 0.001$ which means that beyond the third-order neighborhood any areal unit has the same probability as in SRS. We refer $\hat{\mathcal{P}}$ and $\tilde{\mathcal{P}}$ as an estimation of the proportion parameter $\mathcal{P}$ with the DUST and the SRS designs, respectively. Table 2 shows that the estimate based on the DUST is more efficient than the SRS design. The ratio of mean squared error (MSE) of the DUST over the MSE of SRS (called RM) shows a reduction in MSE of the estimate based on the DUST compared to its SRS counterpart. Another essential concept of geographical

Table 2: Proportion estimation (and standard error) using SRS and DUST methods based on $\mathcal{R} = 5000$ simulations for different values of $\mathcal{P}$ assuming the small areas proportion are known; RM is the ratio of MSE of DUST over the MSE of SRS.

$\mathcal{P}$	SRS	DUST	RM
0.35	0.358 (0.062)	0.350 (0.059)	0.8763
0.65	0.620 (0.061)	0.649 (0.054)	0.8508
Moran's I	0.21	0.69	
q-statistic	0.18	0.78	

phenomena besides spatial autocorrelation (as a property of either the population or the sample [8]) is spatial stratified heterogeneity (SSH). The SSH refers to the phenomena that within strata are more similar than between strata which results in the within strata variance is less than the between strata variance. [16] proposed a q-statistic $\in [0, 1]$ for measuring the degree of spatial stratified heterogeneity as

$$q = 1 - \frac{\sum_{i=1}^{\mathcal{A}} N_i \sigma_i^2}{N\sigma^2},$$

where N stands for the size of the population and σ^2 stands for variance of the attribute. In particular, $q = 0$ shows no spatial stratified heterogeneity and $q = 1$ indicates a fully spatial stratified population. Therefore, q discloses the percent of the variance of an attribute explained by the stratification. Table 2 also reports the measurements of both Moran's I and q values. The results confirm better performance of DUST method than SRS since the q value indicates that 78% of the spatial variation in population has been explained by the sample was taken by DUST. Moreover, Figure 3 shows the empirical semi-variogram of the sampled data by DUST (left panel) and SRS (right panel) which also shows the sampled data by DUST are more spatially correlated than the SRS sampled data.

To evaluate the performance of the model-assisted RSS and compare with its SRS counterpart, we generate two covariates $x_1 \sim Ber\,(0.5)$ (e.g. gender; with 1 for male) and x_2 which is a random number from 1 to 4 (e.g. age-group: child, young, middle-aged and elderly). Then two model-assisted ranking settings are considered: DUST-RSS and SRS-RSS. To obtain a ranked set sample of size $n = 360$, the logistic regression model in (7) is considered with two values 3 and 6 as the set size, m, which requires $\mathcal{K} = 120$ and $\mathcal{K} = 60$ cycles, respectively. The simulation results in Table 3 show that RSS design outperforms its SRS counterpart in terms of bias and MSE for different value of proportion parameter $\mathcal{P}$ and also different set size m.

To evaluate the performance of the model-based DUST-SAE and compare with its DUST-SRS counterpart, we pursue two goals. The first goal is to estimate the population

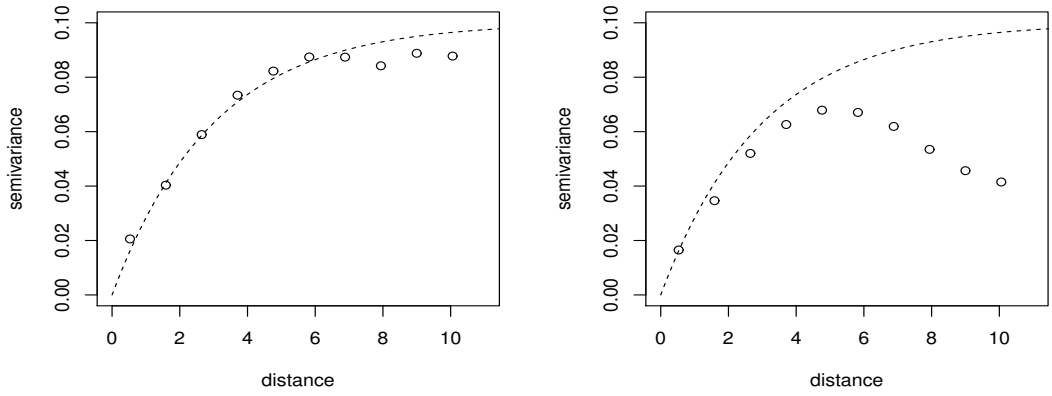

Figure 3: The empirical semi-variogram of the sampled data by DUST (left panel) and SRS (right panel).

Table 3: Proportion estimation (and standard error) using SRS-RSS and DUST-RSS methods based on $\mathcal{R} = 5000$ simulations for different values of $\mathcal{P}$ and m.

	$m = 3$		$m = 6$	
$\mathcal{P}$	SRS-RSS	DUST-RSS	SRS-RSS	DUST-RSS
0.35	0.355 (0.056)	0.350 (0.048)	0.354 (0.056)	0.350 (0.047)
0.65	0.671 (0.058)	0.651 (0.052)	0.668 (0.058)	0.650 (0.052)

proportion for the overall study using the DUST and SAE technique (DUST-SAE) and compare the results with the direct estimate using the SRS (DUST-SRS). The second goal is to assess the performance of DUST-SAE and DUST-SRS in estimating each areal proportion.

To do the first goal, we assume that small areas proportions are unknown. We consider a regular grid $(0, 5) \times (0, 5)$ with $\mathcal{A} = 25$ small areas, and $N = 20,000$ binary variables are spatially generated all over the grid $(0, 5) \times (0, 5)$ which will fall into 25 small areas. We also generate two covariates for the all population individuals: $x_1 \sim Ber\,(0.5)$ and x_2 as a random number from 1 to 4. Then, in each small area, we draw a sample of size $ceiling\,(0.05 \times n_i) = \lceil 0.05 \times n_i \rceil$ from the locations and assume that the values of $Z\,(s_i)$ are available only for these n_i individuals, however, we have values of $\mathbf{x}_i^\top = (\mathbf{x}_{i1}, \mathbf{x}_{i2})$ for $i = 1, \ldots, 25$, as average values of x_1 and x_2 in the i-th small area and collect them in matrix $X_{25\times 2} = (\mathbf{x}_1, \mathbf{x}_2)$. Values of $Z\,(s_i)$ are generated using algorithm 1 with $E\,[V\,(s_i)] = 0.5x_1 + 0.8x_2$ and $Var\,[V\,(s_i)] = 0.1$. The true population proportion for 20,000 generated data is $\mathcal{P} = 0.3001$.

$\ell = 6$ areas are selected using the DUST. Here, $\mathcal{R} = 5000$ simulations are carried out. First, we use the SRS for each selected area to estimate the proportion, and then use those areas proportion estimates to calculate the population proportion (DUST-SRS). Second, we use the SAE method to estimate the proportion for each selected area by the DUST. In particular, based on the independence of the regression coefficients vector $\boldsymbol{\beta} = (\beta_1, \beta_2)^\top$ and σ_ν^2, we assume $f(\boldsymbol{\beta}) \propto 1$ and $\sigma_\nu^{-2} \sim Gamma(a, b)$ where a and b are chosen to reflect vague prior information. In particular, we proceed the following steps for each selected area with letting $d = 0$:

I. Draw $\mathcal{P}_i^*$ from a uniform distribution on $(0, 1)$,

II. Generate $\theta_i^{(d)} \sim N_{25}\left(x_i^\top \boldsymbol{\beta}, \sigma_\nu^2\right)$, and calculate

$$\mathcal{P}_i^{(d)} = \text{logit}^{-1}\left(\theta_i^{(d)}\right) = \frac{e^{\theta_i^{(d)}}}{1 + e^{\theta_i^{(d)}}},$$

III. Calculate

$$r\left(\mathcal{P}_i^{(d)}, \mathcal{P}_i^*\right) = \min\left\{\frac{\mathcal{K}\left(\mathcal{P}_i^*\right)}{\mathcal{K}\left(\mathcal{P}_i^{(d)}\right)}, 1\right\},$$

IV. Select u from a uniform distribution on $(0, 1)$ and let $\mathcal{P}_i^{(d+1)} = \mathcal{P}_i^*$ if $u \leq r\left(\mathcal{P}_i^{(d)}, \mathcal{P}_i^*\right)$,

The above algorithm is repeated until $D, (d = 1, ..., D)$, samples are obtained. A simple estimate of the i-th areal proportion and its posterior variance can be calculated using (11). The model was fitted using JAGS software interfaced to R through *rjags* package. We then estimate the population proportion using $\ell = 6$ selected areas proportion estimate (DUST-SAE). The results of the DUST-SRS and DUST-SAE methods for $\mathcal{R} = 5000$ simulations are shown in Table 4. As it shows, the population proportion estimate using the SAE method is unbiased with smaller standard error unlike the SRS method.

As the second goal for this particular simulation study, we now use the SAE and SRS to estimate each area proportion. In the case of SAE technique, we follow the same procedure as above (for selected areas) to estimate each area proportion. Similarly, we can estimate

Table 4: Model parameter estimates using the DUST-SAE method, and population proportion estimate (and standard error) using the DUST-SAE and DUST-SRS methods.

	DUST-SAE				DUST-SRS
Parameter	β_1	β_2	σ_ν^2	$\mathcal{P}$	$\mathcal{P}$
Estimate	0.43	0.71	0.19	0.30	0.24
SE	0.02	0.02	0.04	0.02	0.11

each area proportion using the SRS method. The results are shown in Figures 4 and 5. In particular, the boxplots of areal proportion estimates through the $\mathcal{R}$ simulations are shown in Figure 4 for the both SRS and SAE methods. As shown in Figure 4, the variations of areal proportions estimate for the SAE method are smaller than the SRS method. In Figure 5, we present the boxplots of population proportion estimate for the SAE and SRS and also true methods. As it is clear from this Figure, the population proportion estimates using the SAE method resemble the true method unlike the SRS method. We also provide the standard errors (SEs) of areal proportions using the SRS and SAE methods (Figure 5). It shows that the SEs of areal proportion estimates using the SAE method are consistently smaller than the SRS method. We also provide the empirical mean squared prediction error (EMSPE) as $(1/\mathcal{R})\sum_{i=1}^{\mathcal{R}}\left(\hat{\mathcal{P}}_i-\mathcal{P}_i\right)^2$ for the SAE method and similarly $(1/\mathcal{R})\sum_{i=1}^{\mathcal{R}}\left(\tilde{\mathcal{P}}_i-\mathcal{P}_i\right)^2$ for the SRS (direct estimate). Figure 6 shows boxplots of EMSPE for the both SAE and SRS methods and corresponding variances of areal proportion estimates. Again, we see superiority of SAE method compared to the SRS method in terms of MSPE of areal proportion estimates.

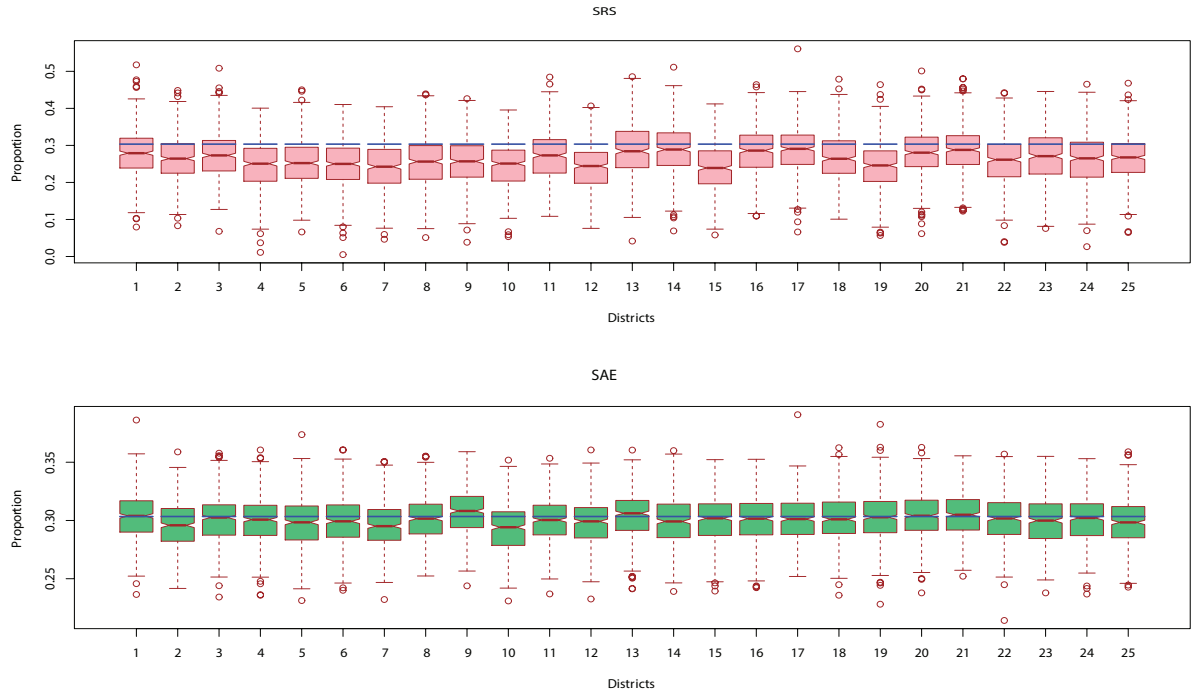

Figure 4: Boxplots of areal proportion estimates using SRS and SAE methods. The horizontal line in each area shows the true proportion.

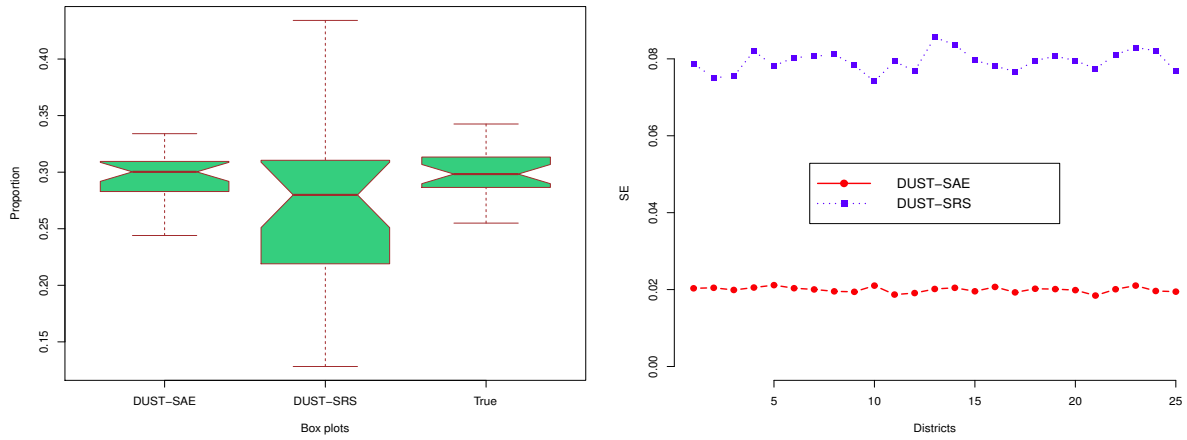

Figure 5: Left panel: Boxplots for the estimates of population proportions based on two approaches (SRS and SAE) vs. true values. Right panel: Standard errors of areal proportion estimates based on SRS and SAE methods.

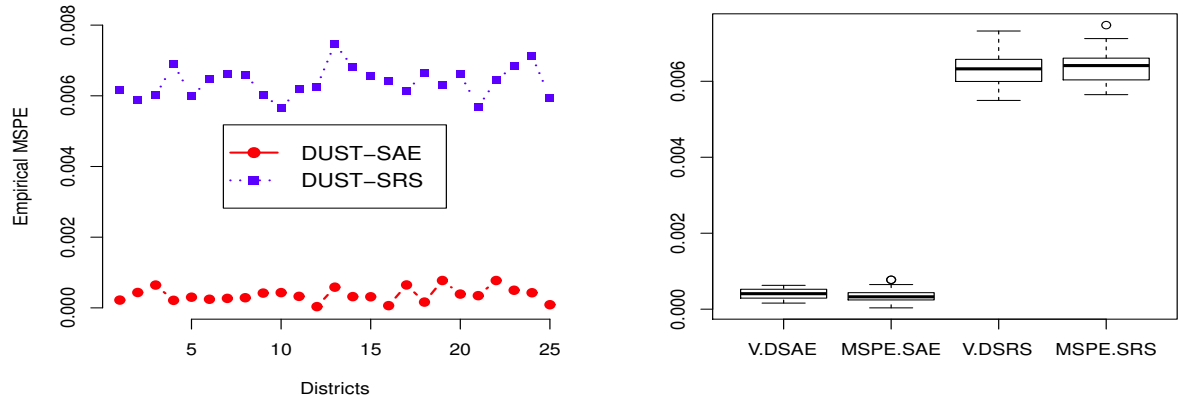

Figure 6: Left panel: Empirical mean squared prediction error of areal proportion estimates. Right panel: Boxplots of EMSPE for SAE (MSPE.SAE) and SRS (MSPE.SRS) and corresponding variances of the areal proportion estimates (V.DSAE and V.DSRS).

4.2. Rate Estimation

This subsection demonstrates the objectives of Section 3. A simple way to simulate spatial correlated counts has already proposed by [17] through the algorithm 2 below.

For the simulation study on rate estimation, we consider a grid $(0, 10) \times (0, 10)$ with $\mathcal{A} = 100$ areas and the random set of locations of size $N = 500$ are considered (Figure

Algorithm 2 Generate spatially correlated count data

I. Spatially correlated normals $V(s_i)$ are generated (see the Algorithm 1),
II. $V(s_1), \ldots, V(s_n)$ are transformed to lognormals $\exp\{V(s_i)\}$ for $i = 1, 2, \ldots, n$,
III. Conditionally independent $Z_i \sim Poisson(\exp\{V(s_i)\})$ are generated.

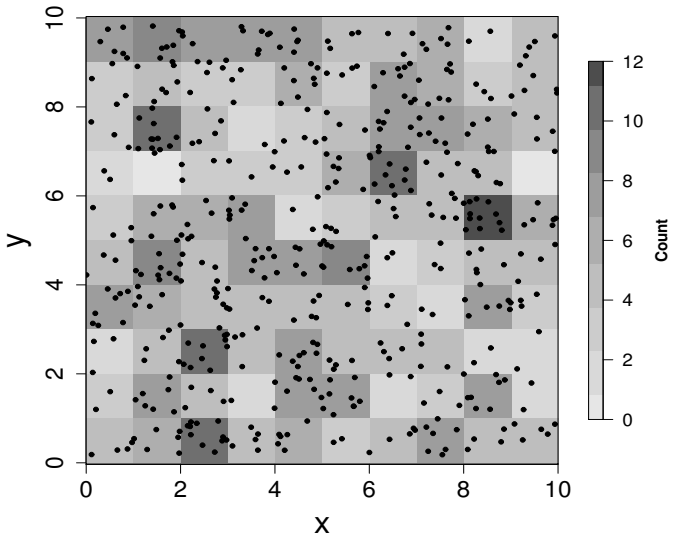

Figure 7: Sampling locations of spatial count data.

7). To evaluate the performance of the DUST and SRS methods on population rate estimate, we first assume that the areal rates are known. We follow the algorithm 2 to generate spatially correlated count data Z_i for area $i\,(= 1, \ldots, \mathcal{A})$. We carry out $\mathcal{R} = 4000$ simulations and consider two different values for population rate (0.019 and 0.022). The results of population rate estimate using the SRS and DUST (based on $\ell = 15$ selected areas) methods are shown in Table 5. As you can see from Table 5, the DUST rate estimate has less bias and MSE compared to the SRS rate estimate. Moreover, the measurements of both Moran's I and q values indicate better performance of DUST method than SRS since the q value indicates that 70% of the spatial variation in population has been explained by the sample was taken by DUST. Figure 8 shows the empirical semi-variogram of the sampled data by DUST (left panel) and SRS (right panel) which also shows that the sampled data by DUST are more spatially correlated than the SRS sampled data.

We now assume that the areal rates are unknown, and evaluate the performance of three approaches: SRS-SRS, DUST-SRS, and DUST-SAE. Table 6 presents the results of the SRS-SRS and DUST-SRS where $\ell = 15$ areas are selected based on the DUST

Table 5: Population rate estimates based on DUST and SRS methods.

			MSE	
λ	SRS	DUST	SRS	DUST
0.019	0.017	0.018	0.048	0.032
0.022	0.025	0.023	0.051	0.031
Moran's I	0.17	0.62		
q-statistic	0.15	0.70		

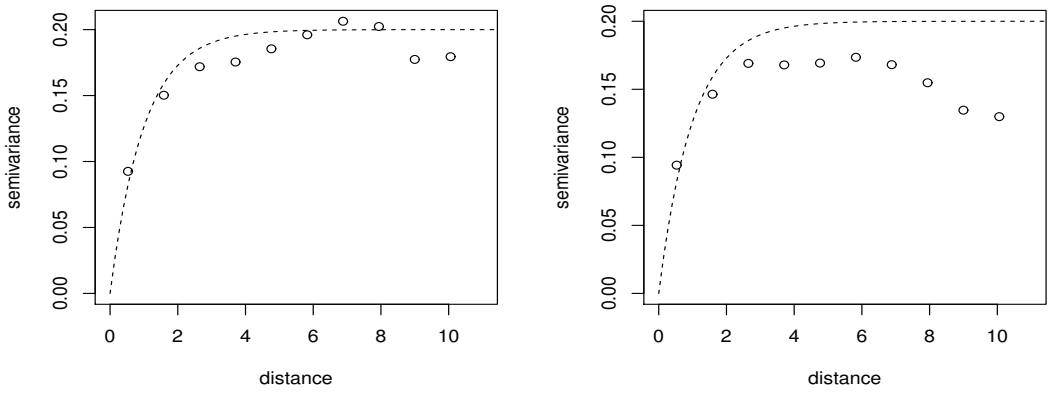

Figure 8: The empirical semi-variogram of the sampled data by DUST (left panel) and SRS (right panel).

approach for two different values of true population rate.We also observe that, in the case of areal rates are unknown, the population rate estimate using the DUST-SRS outperforms the SRS-SRS in terms of bias and MSE.

We now evaluate the performance of the DUST-SRS and DUST-SAE methods in the case of areal rates are unknown. In particular, for the SAE method, we consider the same two covariates (binary, and a random number from 1 to 4 for the second covariate) as in the proportion estimation simulation study (subsection 4.1). The results are shown in Table 7 and Figure 9. In Table 7, we have population rates estimates using the DUST-SRS and DUST-SAE methods with also model parameter estimates. As expected, the DUST-SAE outperforms the DUST-SRS in terms of bias and MSE, and also model parameters are estimated well using the SAE technique. In Figure 9, we have boxplots of population rate estimates using the DUST-SRS, DUST-SAE, and true methods which show the DUST-SAE method resembles the true population rates very well unlike the DUST-SRS method. In Figure 9, we also see that the standard errors of population rate estimates using the

Table 6: Population rate estimates using the SRS-SRS and DUST-SRS approaches.

			MSE	
λ	SRS-SRS	DUST-SRS	SRS-SRS	DUST-SRS
0.020	0.023	0.022	0.062	0.045
0.032	0.026	0.036	0.067	0.050

DUST-SAE are consistently smaller than the corresponding values from the DUST-SRS method. Figure 10 shows boxplots of EMSPE of areal rate estimates using the DUST-SRS and DUSRT-SAE methods. It is clear from this Figure that the MSPE of areal rate estimates for the DUST-SAE is consistently smaller than the DUST-SRS method. We also provide boxplots of EMSPE and variances of areal rate estimates using the DUST-SRS and DUST-SAE methods (Figure 10). It is also clear from the boxplots that the DUST-SAE method is doing better than the DUST-SRS method in terms of MSPE of areal rate estimate and corresponding variances.

5. Application: Teen Birth Rate Data

The teen birth rate (TBR) in the USA hit a high of 61.8 percent in 1991 and over the past two decades USA teens have been far more likely to give birth than in any other industrialized country in the world. For instance, USA teens are two and a half times as likely to give births as compared to teens in Canada, around four times as likely as teens in Germany or Norway, and almost 10 times as likely as teens in Switzerland [18]. In 2016, TBR in the USA was 20.3 births for every 1000 females aged 15-19 and is substantially higher than other western countries, although disparities exist between racial/ethnic groups and geographic regions. However, limited information is available about teenage pregnancy and childbearing specially in rural areas. Determining rate estimates of teen pregnancy is important to design and implement prevention programs to reduce the TBR. Georgia County Health Rankings data in 2018 contains TBR and county-level socio-demographic

Table 7: Population rate estimates using the DUST-SAE and DUST-SRS with corresponding standard errors (SEs), and also model parameters estimate using the SAE method.

	DUST-SAE				DUST-SRS
Parameter	β_1	β_2	σ^2	λ	λ
True	3	4	5	0.016	0.016
Estimate	3.26	3.91	5.27	0.017	0.025
SE	0.82	0.91	1.19	0.032	0.073

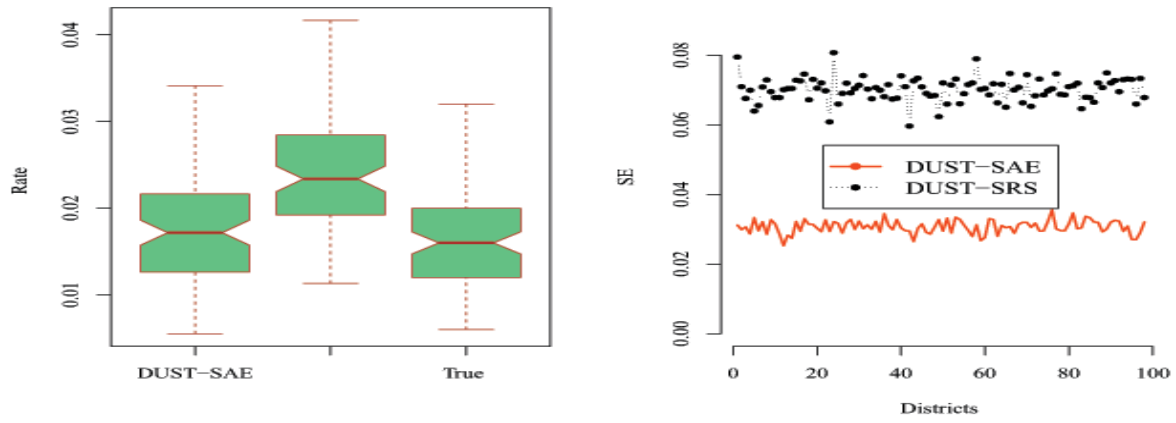

Figure 9: Left panel: Boxplots of the average of areal rate estimates based on DUST-SAE, DUST-SRS , and true values. Right panel: Standard errors of areal rate estimates based on DUST-SRS and DUST-SAE methods.

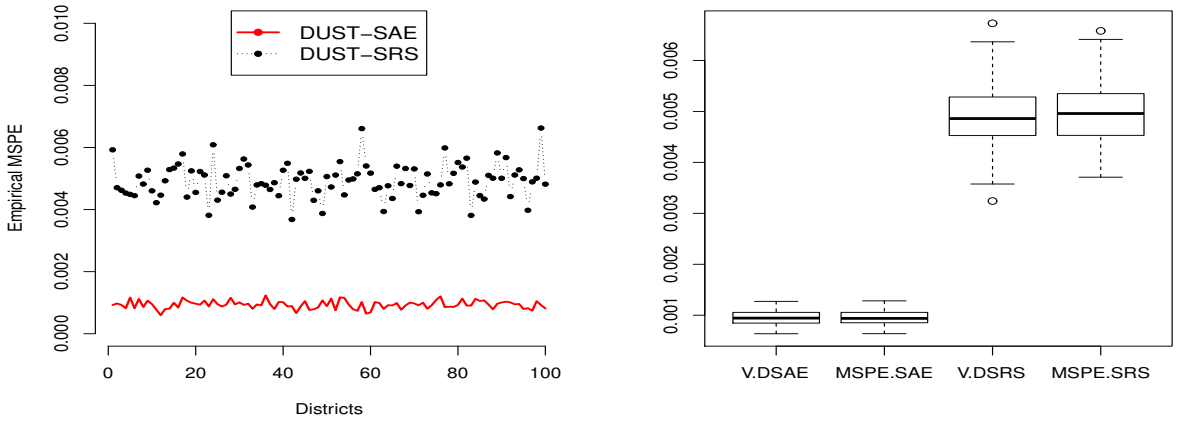

Figure 10: Left panel: Empirical mean squared prediction error (MSPE) of areal rate estimates using the DUST-SRS and DUST-SAE methods. Right panel: Boxplots of EMSPE of areal rate estimates using the DUST-SAE (MSPE.SAE) and DUST-SRS (MSPE.SRS) methods, and corresponding variances of the areal rate estimates (V.DSAE and V.DSRS).

information for 159 counties. Georgia was chosen because it has one of the higher TBR in the USA and had no missing data for key variables. Various individual, family and community characteristics have been previously linked to teen pregnancy, e.g., race/ethnicity, community-level socio-economic status (SES), urban vs rural and single-parent household.

Table 8: Overall TBR estimate and corresponding MSE based on SRS-SRS, DUST-SRS, and DUST-SAE methods.

				MSE		
Observed	SRS-SRS	DUST-SRS	DUST-SAE	SRS-SRS	DUST-SRS	DUST-SAE
0.0306	0.048	0.038	0.031	0.12	0.08	0.03

The SES index (consists of percentages of unemployed and single parent households, high school graduation rate, and median household income), the percentage of African-American and the percentage of uninsured teens (with coefficients β_1, β_2 and β_3, in order) are included in the model. Since the locations of the events were not known, the centroid of each district is considered as the location in demand for modeling. Table 8 demonstrates the overall TBR estimate using the SRS-SRS, DUST-SRS, and DUST-SAE methods, and corresponding MSE, note that $\ell = 15$ areas are selected based on the DUST method. It shows that the DUST-SAE method outperforms other methods in terms of point estimate and corresponding MSE. Figure 11 provides the boxplots of district-level TBR estimate for the SRS-SRS, DUST-SRS, and DUST-SAE methods which show that the DUST-SAE method has smaller variation unlike the other methods. We also provide the standard error of district-level TBR estimate where the DUST-SAE method has consistently smaller standard errors compared to the SRS-SRS and DUST-SRS methods. Model parameters estimate based on the SAE method is provided in Table 9. The results are also supported by the literature where higher SES results in lower TBR, and higher African-American and uninsured teens result in higher TBR, and these contributions are statistically significant.

Table 9: Model parameters estimate and corresponding standard error for the TBR data based on the SAE method.

Parameter	β_0	β_1	β_2	β_3	σ^2
Estimate	1.02	-0.35	1.9	1.6	1.05
SE	0.15	0.03	0.04	0.07	0.08

6. Conclusion

In this paper, we have developed efficient estimations for the population proportion and rate when the variable of study is spatially correlated. In particular, in the context of spatial statistic trinity (SST) [8], our "population" is a spatially correlated population, our

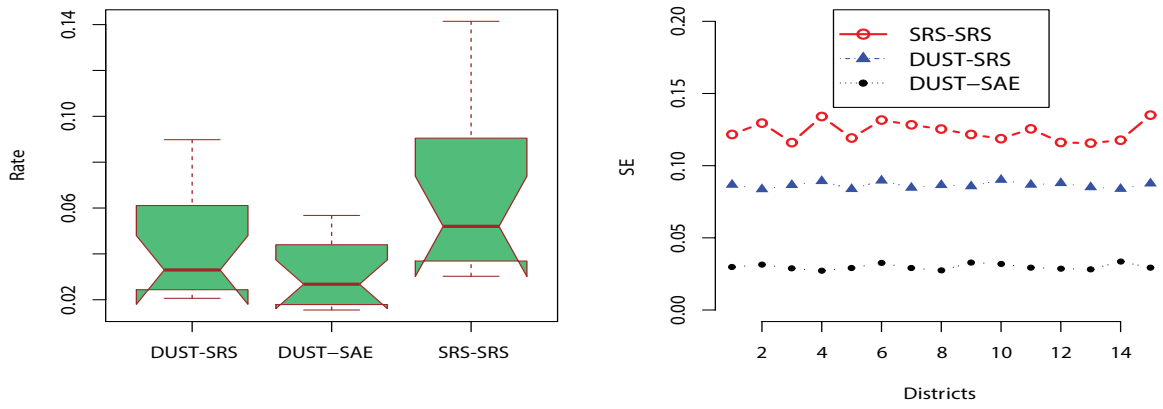

Figure 11: Left panel: Boxplots of the average of district-level TBR estimate based on SRS-SRS, DUST-SRS, and DUST-SAE methods. Right panel: Standard errors of district-level TBR estimate based on SRS-SRS, DUST-SRS, and DUST-SAE methods.

"sampling" is a DUST approach, and our "estimator/predictor" is small area estimation technique. In above set-up, our approach outperforms its counterparts.

The literature for proportion estimate on spatially correlated binary data was mainly focused on the population level assuming area-level proportion is known. However, sometimes public and policy-makers need to also know proportion estimate of certain conditions (e.g., asthma) at area-level to identify areas which are more at risk for resource allocation. To that end, we have developed two approaches (DUST-RSS and DUST-SAE) which were more efficient than the SRS-SRS method. We also showed that the DUST-SAE method is the best one for population proportion estimate when area-level proportions are not known.

To the best of our knowledge, we could not find any literature on rate estimation (overall or area-level) for spatially correlated count data. This design is very useful when our interest is to know rate estimate of rare conditions (e.g., cancer) at area-level to help policy-makers for resource allocation and possible interventions and prevention. To that end, we have developed novel techniques for the overall rate estimate and also for area-level rate estimate. We have shown that the proposed DUST-SRS and DUST-SAE methods are more efficient than the SRS-SRS method, and in particular the DUST-SAE method is the best one for population rate estimate when area-level rates are not known.

The proposed approaches for the proportion and rate estimates have been evaluated using simulation studies. We have also employed our rate estimation approach to the teen birth rate dataset in Georgia, USA.

Supplementary materials

Supplementary materials contain R codes and corresponding "readme" files for the simulations and real data application conducted in this paper.

Acknowledgments

Constructive comments and suggestions of two reviewers, which led to an improved version of this article, are greatly appreciated. This research was partially supported by Discovery Grants from the Natural Sciences and Engineering Research Council of Canada (NSERC) to M. Torabi.

References

[1] D. J. Brus, J. J. DeGruijter, Design-based versus model-based estimates of spatial means: Theory and application in environmental soil science, Environmetrics 4 (1993) 123–152.

[2] R. P. Haining, R. Haining, Spatial data analysis: theory and practice, Cambridge University Press, 2003.

[3] J. F. Wang, A. Stein, B. B. Gao, Y. Ge, A review of spatial sampling, Spatial Statistics 2 (2012) 1–14.

[4] G. McIntyre, A method for unbiased selective sampling, using ranked sets, Australian Journal of Agricultural Research 3 (1952) 385–390.

[5] W. Overton, Probability sampling and population inference in monitoring programs, Environmental modeling with GIS (1993) 470–480.

[6] J. De Gruijter, Spatial sampling schemes for remote sensing, in: Spatial statistics for remote sensing, Springer, 1999, pp. 211–242.

[7] J. De Gruijter, D. J. Brus, M. F. Bierkens, M. Knotters, Sampling for natural resource monitoring, Springer Science & Business Media, 2006.

[8] J. Wang, B. Gao, A. Stein, The spatial statistic trinity: A generic framework for spatial sampling and inference, Environmental Modelling & Software 134 (2020) 104835.

[9] H. Chen, E. A. Stasny, D. A. Wolfe, Ranked set sampling for efficient estimation of a population proportion, Statistics in medicine 24 (2005) 3319–3329.

[10] G. Arbia, The use of GIS in spatial statistical surveys, International Statistical Review/Revue Internationale de Statistique (1993) 339–359.

[11] P. Moran, Notes on continuous stochastic phenomena, Biometrika 37 (1950) 17–23.

[12] G. Arbia, On second-order non-stationarity in two dimensional lattice processes, Computational Statistics & Data Analysis 9 (1990) 147–160.

[13] J. Rao, I. Molina, Small area estimation, Second Ed., Hoboken, New (2015).

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

[15] R. Jin, S. Wang, F. Yan, J. Zhu, Generating spatial correlated binary data through a copulas method, Science Research 3 (2015) 206–212.

[16] J.-F. Wang, T.-L. Zhang, B.-J. Fu, A measure of spatial stratified heterogeneity, Ecological Indicators 67 (2016) 250–256.

[17] L. Madsen, D. Dalthorp, Simulating correlated count data, Environmental and Ecological Statistics 14 (2007) 129–148.

[18] M. S. Kearney, P. B. Levine, Why is the teen birth rate in the united states so high and why does it matter?, Journal of Economic Perspectives 26 (2012) 141–63.