Statistica Sinica Preprint No: SS-2024-0001	
Title	Resampling Method for Generalized One-per-Stratum Sampling Designs
Manuscript ID	SS-2024-0001
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202024.0001
Complete List of Authors	Zhonglei Wang and Zhengyuan Zhu
Corresponding Authors	Zhengyuan Zhu
E-mails	zhuz@iastate.edu
Notice: Accepted version subject to English editing.	

Statistica Sinica

RESAMPLING METHOD FOR GENERALIZED ONE-PER-STRATUM SAMPLING DESIGNS

Zhonglei Wang and Zhengyuan Zhu

Xiamen University and Iowa State University

Abstract:

In areal surveys, one-per-stratum sampling is commonly used since it achieves spatial balance and improves estimation efficiency. The downside of such a design is that it is challenging to have a good variance estimator. In this paper, we propose a generalized one-per-stratum sampling design to generate a spatially balanced sample. The sample is used to get an M-estimator of the parameters in a spatial linear regression model, and the corresponding variance is estimated by a resampling method. Asymptotic properties of the M-estimator are investigated under the proposed one-per-stratum sampling design. Simulation studies show that the proposed one-per-stratum sampling design achieves good spatial balance, and the M-estimator is more efficient compared with existing designs. The resampling method is applied to investigate the relationship between soil erosion and slope in Iowa using a recent sample from the National Resources Inventory survey.

Key words and phrases: Asymptotics, M-estimator, Spatially balanced sampling, Spatial block bootstrap, Survey variance estimation.

1. Introduction

In environmental studies, observations are spatially dependent in the sense that the correlation is a decreasing function of distance, so it is desirable to obtain a spatially balanced sample, which spread over the sampling domain well, to make efficient inference (Cochran, 1946; Stevens and Olsen, 2004; Grafström et al., 2012). For example, stratified sampling is conducted to obtain well-spread samples to study soil erosion by the National Resources Inventory survey (Nusser and Goebel, 1997; Nusser et al., 1998); also see the land surveys by the Bureau of Lang Management and the June Area survey by the National Agricultural Statistics Service. Even though various spatially balanced sampling designs are applied in practice, how to make valid statistical inference is still an open problem (Stevens and Olsen, 2004; Grafström et al., 2012). In this paper, our goal is to propose a general one-per-stratum sampling design and rigorously prove that inference can be made through a resampling method.

Different spatially balanced sampling designs have been proposed. Bartholdi and Platzman (1988) and Lister and Scott (2009) used space-filling curves to obtain spatially balanced samples. Munholland and Borkowski (1996) applied a simple Latin square to draw a spatially balanced sample and demonstrated that estimators under the proposed design are generally more

efficient than those under simple random sampling. Breidt (1995) generalized the one-per-stratum sampling design by introducing dependence in the sampling of neighboring subregions through a Markov structure to achieve better spatial balance. Stevens and Olsen (2004) introduced a generalized random tessellation stratified design to guarantee spatial balance. Grafström et al. (2012) proposed two local pivotal methods both performing better than the generalized random tessellation stratified design. Wang and Zhu (2019) proposed a spatio-temporal balanced sampling design based on the local pivotal method. Tillé et al. (2018) proposed to use renewal chains and multivariate discrete distributions to tune the joint selection probability of neighboring units to achieve better spatial balance. Although spatially balanced samples can be generated by the above designs, unbiased variance estimators are not available.

For spatial analysis, resampling methods are widely used for variance estimation and statistical inference. Nordman and Lahiri (2004) and Nordman et al. (2007) discussed a block-based bootstrap method to estimate the variance when observations are regularly spaced. Politis et al. (1998) proposed a subsampling approach for observations generated by a homogeneous Poisson process. Lahiri and Zhu (2006) discussed both fixed and stochastic sampling designs, and a block resampling method was theoretically inves-

tigated. Also see Lahiri (2018), Hala et al. (2020), Chan et al. (2022) and Zhang et al. (2023). Although valid inference can be made through resampling methods, existing works assumed a *fixed* sampling density function to generate samples, leading to unsatisfactory spatial balance.

We are not aware of any work guaranteeing both spatial balance and valid inference. In this paper, a generalized one-per-stratum sampling design is proposed, and asymptotic properties are investigated under a weak dependent setup (Grenander, 1954; Koul, 1992; Yajima, 1991). The proposed one-per-stratum sampling design has several appealing features. It is flexible and easy to be implemented in practice. For example, an area can be over-sampled by forming more strata. Besides, different sampling density functions can be specified for different strata, and better spatial balance can be achieved using a more concentrated sampling density function within each stratum; see Section 5 for details. Furthermore, a model-based variance estimator can be obtained using a resampling method under the proposed one-per-stratum sampling designs.

The rest of the paper is organized as follows. The one-per-stratum sampling design is proposed in Section 2. An M-estimator under a spatial setting is discussed in Section 3. Section 4 shows the asymptotic properties of the M-estimator under the proposed design. Simulations are conducted

to test the performance of the proposed method in Section 5. Application to the soil erosion analysis for the National Resources Inventory program is presented in Section 6. Section 7 consists of some final discussion.

2. Model setup

2.1 Generalized one-per-stratum sampling design

One-per-stratum sampling is commonly used in area sampling to ensure the spatial balance of the sample, and it partitions a sampling domain into congruent subregions and randomly samples one point from each subregion independently. In this paper, we propose a generalization of the one-per-stratum sampling design, which keeps the independent sampling feature of each subregion, while allows arbitrary spatial sampling distribution within each subregion.

Let $R_0 \subset (-1/2, 1/2]^d$ be a d-dimensional Borel set containing the origin as its interior point. A sampling domain is obtained by $R_n = \lambda_n R_0$, where n is the sample size, $\lambda_n \asymp n^{1/d}$, and $a_n \asymp b_n$ is equivalent to $a_n = O(b_n)$ and $b_n = O(a_n)$. That is, we consider a pure increasing domain asymptotic framework (Cressie, 2015, Section 2.6.3). Denote $\mathcal{A}_n = \{A_i : i = 1, \ldots, n\}$ to be a non-random partition set such that $R_n = \cup_{i=1}^n A_i$ and $A_i \cap A_j = \emptyset$ $(i \neq j)$. We propose to generate $\boldsymbol{s}_i$ from A_i independently for $i = 1, \ldots, n$,

and denote $\mathcal{S}_n = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$ to be a sample of size n. The spatial balance of $\mathcal{S}_n$ is determined by the shape of A_i $(i = 1, \ldots, n)$ as well as the sampling densities to generate $\mathcal{S}_n$; see Section 5 for details.

Remark 1. By selecting a sampling density to be more concentrated in the center of each subregion, we can achieve better spatial balance, with the limiting case being the spatial systematic sampling. The traditional one-per-stratum sampling design corresponds to the special case using a uniform sampling distribution within each subregion. Similar to systematic sampling, we need to introduce a random shift (Stevens and Olsen, 2004) of the subregion boundary to achieve equal sampling rate, and the mathematical details will be presented in a separate paper.

Figure 1, for example, shows a comparison of the proposed one-per-stratum sampling design with the stochastic design discussed by Lahiri and Zhu (2006), where $R_0 = (-1/2, 1/2)^2$, $\lambda_n = 6$ and $n = 25$. For the proposed one-per-stratum sampling design, we partition the sampling region into $n = 25$ squares with equal size, and a uniform or truncated normal sampling density with mean at its center and $0.25^2\mathcal{I}$ as the covariance matrix is used within each partition. We also consider the same sampling densities for the stochastic design of Lahiri and Zhu (2006); see Section S1 of the Supplementary Material for its brief introduction. From Figure 1, the

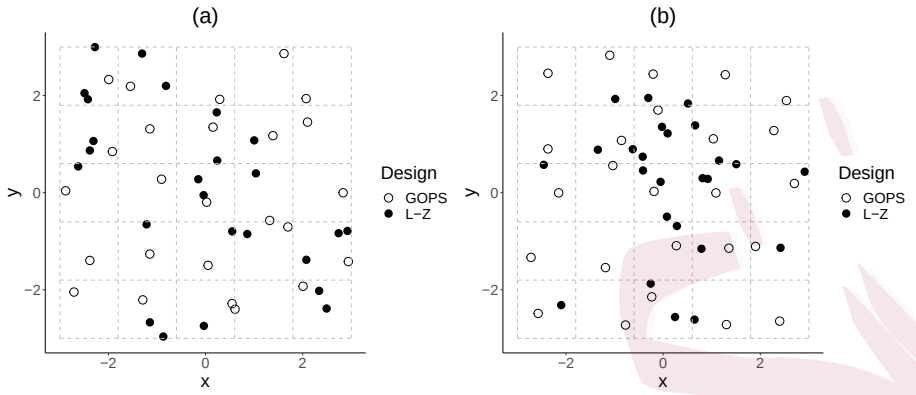

Figure 1: Comparison of the proposed one-per-stratum sampling design (GOPS) with the one of Lahiri and Zhu (2006) (L-Z) under (a) uniform sampling density and (b) truncated bivariate normal sampling density. The subregions are highlighted using dashed lines.

proposed one-per-stratum sampling design generates samples with better spatial balance compared with the stochastic design (Lahiri and Zhu, 2006), especially when a more concentrated sampling density is used.

Remark 2. When sampling trees or housing units, there only exist finite possible locations. Then, we can partition a stratum into Voronoi polygons (Stevens and Olsen, 2004) each centered at one measurable unit, and select one unit within each stratum at random with probability proportional to the integral of the sampling density over the polygon that unit represent. Alternatively, we can pick a random location in each stratum based on the

sampling density and select the unit which is closest to the selected location.

2.2 M-estimator in spatial linear regression models

Consider the following spatial linear regression model (Yajima, 1991; Lahiri and Zhu, 2006):

$$Y(\boldsymbol{s}) = \boldsymbol{x}(\boldsymbol{s})^\top \boldsymbol{\beta}_0 + Z(\boldsymbol{s}) \quad (\boldsymbol{s} \in \mathbb{R}^d), \tag{2.1}$$

where $\boldsymbol{x}(\boldsymbol{s})$ is a known p-dimensional real-valued function of the location $\boldsymbol{s}$, $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ is the regression parameter, $\{Z(\boldsymbol{s}) : \boldsymbol{s} \in \mathbb{R}^d\}$ is a one-dimensional zero-mean stationary random field independent of the proposed one-per-stratum sampling design, and $A^\top$ is the transpose of a matrix A. Under the proposed one-per-stratum sampling design, we observe $\{(\boldsymbol{x}(\boldsymbol{s}_i), Y(\boldsymbol{s}_i)) : \boldsymbol{s}_i \in \mathcal{S}_n\}$, and we are interested in making inference for the regression parameter $\boldsymbol{\beta}_0$. An M-estimator $\hat{\boldsymbol{\beta}}_n$ solves

$$M_n(\boldsymbol{\beta}) = \mathbf{0}, \tag{2.2}$$

where $M_n(\boldsymbol{\beta}) = \sum_{i=1}^n \boldsymbol{x}(\boldsymbol{s}_i)\Psi\{Y(\boldsymbol{s}_i) - \boldsymbol{x}(\boldsymbol{s}_i)^\top \boldsymbol{\beta}\}$, and $\Psi : \mathbb{R} \to \mathbb{R}$ is a known one-dimensional Borel-measurable function satisfying $E[\Psi\{Z(\mathbf{0})\}] = 0$.

Before investigating theoretical properties of $\hat{\boldsymbol{\beta}}_n$, consider the strong mixing condition for the stationary random field $\{Z(\boldsymbol{s}) : \boldsymbol{s} \in \mathbb{R}^d\}$. For $\boldsymbol{s} = (s_1, \ldots, s_d)$, let $\|\boldsymbol{s}\|_1 = \sum_{i=1}^d |s_i|$ and $\|\boldsymbol{s}\|_2 = (s_1^2 + \cdots + s_d^2)^{1/2}$. Denote

vol.(A) and $|A|$ to be the volume and the cardinality of a set $A \subset \mathbb{R}^d$, respectively. For two sets T_1 and T_2 of $\mathbb{R}^d$, denote $d(T_1, T_2) = \inf\{\|\boldsymbol{s}_1 - \boldsymbol{s}_2\|_2 : \boldsymbol{s}_i \in T_i, i = 1, 2\}$. Denote $\mathcal{R}(b) = \{\cup_{i=1}^k D_i : \sum_{i=1}^k \text{vol.}(D_i) \leq b, k \geq 1\}$, where $\{D_i : i = 1, \ldots, k\}$ is a finite set of pairwise disjoint hypercubes in $\mathbb{R}^d$ for $k \geq 1$. The strong-mixing coefficient for the random field $\{Z(\boldsymbol{s}) : \boldsymbol{s} \in \mathbb{R}^d\}$ is defined as

$$\alpha(a; b) = \sup\{\tilde{\alpha}(T_1, T_2); d(T_1, T_2) \geq a, T_1 \in \mathcal{R}(b), T_2 \in \mathcal{R}(b)\}$$

for $a > 0$ and $b > 0$, where $\tilde{\alpha}(T_1, T_2) = \sup\{|P(A \cap B) - P(A)P(B)| : A \in \mathcal{F}_Z(T_1), B \in \mathcal{F}_Z(T_2)\}$, $\mathcal{F}_Z(A) = \sigma\langle Z(\boldsymbol{s}) : \boldsymbol{s} \in A\rangle$ is the sigma-algebra generated by the stationary random field on $A \subset \mathbb{R}^d$. Assume $\alpha(a; b) \leq \alpha_1(a)g_1(b)$, where $\alpha_1(\cdot)$ is a nonincreasing left continuous function satisfying $\lim_{a\to\infty} \alpha_1(a) = 0$, and $g_1(\cdot)$ is a nondecreasing function satisfying $\lim_{b\to\infty} g_1(b) = \infty$. Similar definitions were used by Lahiri (2003), Lahiri and Mukherjee (2004) and Lahiri and Zhu (2006).

Remark 3. Rubin-Bleuer and Schiopu-Kratina (2005) considered a finite-population parameter, and its design-based estimator was investigated by treating the finite population as fixed. Thus, they proposed a product probability space, including both the design space and the model space. However, we do not consider finite populations in this paper, and the parameter of interest is $\boldsymbol{\beta}_0$ in (2.1) above. Besides, theoretical properties are

studied with respect to almost every $\mathcal{S}_n$. Thus, we only concentrate on the design space in the preceding paragraph.

3. Asymptotic properties

We generalize the conditions of Lahiri (2003) and Lahiri and Zhu (2006) to study the asymptotic properties of $\hat{\boldsymbol{\beta}}_n$ under the proposed one-per-stratum sampling design.

1. The prototype R_0 satisfies vol.$(R_0) > 0$ and vol.$(R_0^{\epsilon_n}) \to 0$ as $\epsilon_n \to 0$, where $R_0^{\epsilon} = \{x \in R_0 : (x + \epsilon[-1,1]^d) \cap R_0^{\mathrm{C}} \neq \emptyset\}$, and A^{C} is the complement of a set A.

2. There exists $M_A > 0$ such that vol.$(A_i) \leq M_A$ $(i = 1, 2, \ldots)$.

3. There exists $C_1 > 0$ such that $|\{A_i : A_i \cap B \neq \emptyset, i = 1, \ldots, n\}| \leq C_1 \mathrm{vol.}(B)$ for any ball $B \subset R_n$.

4. There exists $M_f > 0$ such that $f_i(\boldsymbol{s}) \leq M_f$ for $\boldsymbol{s} \in \mathrm{supp}(f_i) \subset A_i$ $(i = 1, 2, \ldots)$, where $\mathrm{supp}(f) = \{x : f(x) > 0\}$ is the support of a function $f(x)$.

5. There exists a sequence of nonsingular matrices $\{\Lambda_n : n \geq 1\}$ such

that

$$\Lambda_n^{-1}\left\{\sum_{i=1}^n \int \boldsymbol{x}(\boldsymbol{s})\boldsymbol{x}(\boldsymbol{s})^\top f_i(\boldsymbol{s})\mathrm{d}\boldsymbol{s}\right\}\Lambda_n^{-1} \to H,$$

$$\Lambda_n^{-1}\left\{\sum_{i=1}^n \sum_{j\neq i} \int \boldsymbol{x}(\boldsymbol{s}+\boldsymbol{h})\boldsymbol{x}(\boldsymbol{s})^\top f_i(\boldsymbol{s}+\boldsymbol{h})f_j(\boldsymbol{s})\mathrm{d}\boldsymbol{s}\right\}\Lambda_n^{-1} \to Q(\boldsymbol{h})$$

as $n \to \infty$, where H is positive definitive, and $Q(\boldsymbol{h})$ is a $p \times p$ matrix-valued function of $\boldsymbol{h} \in \mathbb{R}^d$.

6. $\int Q(\boldsymbol{h})\sigma_\Psi(\boldsymbol{h})\mathrm{d}\boldsymbol{h}$ is positive definite, where $\sigma_\Psi(\boldsymbol{h}) = E[\Psi\{Z(\boldsymbol{0})\}\Psi\{Z(\boldsymbol{h})\}]$.

7. $m_{0n} = \sup\{\|\Lambda_n^{-1}\boldsymbol{x}(\boldsymbol{s})\| : \boldsymbol{s} \in R_n\} = o(n^{-3/8})$. Recall that $\|\boldsymbol{s}\|$ is the l_2 norm defined in Section 2.2.

8. There exists $\delta \in (0, \infty)$ such that $E|\Psi\{Z(\boldsymbol{0})\}|^{2+\delta} < \infty$, $E|\Psi'\{Z(\boldsymbol{0})\}|^{2+\delta} < \infty$ and $\chi_0 = E[\Psi'\{Z(\boldsymbol{0})\}] \neq 0$, where $\Psi'(x)$ is the derivative of $\Psi(x)$. $\alpha_1(a) = O(a^{-\tau})$ and $g_1(b) = o(b^{(\tau-d)/(4d)})$, where $\tau > d(2+\delta)/\delta$.

9. Function $\Psi'(x)$ satisfies a Lipschitz condition of order $\gamma \in (2/3, 1]$ with $C_2 > 0$,

$$|\Psi'(x_1) - \Psi'(x_2)| \leq C_2|x_1 - x_2|^\gamma, \quad (x_1 \in \mathbb{R}, x_2 \in \mathbb{R}).$$

Condition 1 is a mild restriction on the boundary of the prototype R_0, and Lahiri (2003) used a similar condition to avoid pathological boundaries

of R_0. Conditions 2–3 regulate the partition regions, such that a sample, generated by the proposed one-per-stratum sampling design, is spatially balanced. In Condition 2, we only assume that the maximum volume of the partition regions is bounded, but we do not assume that all partition regions have the same volume. By Condition 1 and Condition 3, the number of partition regions on the "boundary" part of R_n is negligible compared with that in its "interior" part, and this result is used to derive the asymptotic properties for the resampling method. Condition 4 is a mild restriction on the sampling density function within each stratum. We do not require $\text{supp}(f_i) = A_i$ $(i = 1, 2, \ldots)$, so a more spatially balanced sample can be generated using a more concentrated sampling density function $f_i(\boldsymbol{s})$. Condition 5 is the Grenader condition for the linear regression model (Grenander, 1954) under the proposed one-per-stratum sampling design. Condition 6 guarantees the existence of the variance matrix for the M-estimator $\hat{\boldsymbol{\beta}}_n$. Condition 7 regulates the covariate $\boldsymbol{x}(\boldsymbol{s})$ and is used to show the convergence of relevant statistics. Condition 8 is needed to show a central limit theorem for the stationary spatial process. Condition 9 is used for the Taylor's expansion when deriving the asymptotic distribution of $M_n(\boldsymbol{\beta}_0)$.

Theorem 1. *Suppose Conditions 2–9 hold. Then, we have*

$$\Lambda_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \to N(0, \chi_0^{-2}\Sigma_{\boldsymbol{\beta}})$$

in distribution $(P_{\cdot|\boldsymbol{S}})$ *almost surely* $(P_{\boldsymbol{S}})$*, where*

$$\Sigma_{\boldsymbol{\beta}} = H^{-1}\sigma_\Psi(\mathbf{0}) + H^{-1}\left\{\int \sigma_\Psi(\boldsymbol{h})Q(\boldsymbol{h})\mathrm{d}\boldsymbol{h}\right\}H^{-1},$$

$P_{\boldsymbol{S}}$ *is the probability measure with respect to the propoased one-per-stratum sampling design, and* $P_{\cdot|\boldsymbol{S}}$ *is the conditional probability measure with respect to (2.1) given the sampled locations.*

The proof of Theorem 1 is given in Section S3 of the Supplementary Material; see Section S2 of the Supplementary Material for the construction of $P_{\boldsymbol{S}}$. Even though Theorem 1 appears similar to Theorem 1 of Lahiri and Zhu (2006), we consider a totally different setups, and the corresponding proof is also different. Specifically, Lahiri and Zhu (2006) assumed that the sampled locations are scaled based on independent and identically distributed random variables generated on the prototype. In this paper, we first scale the prototype to obtain R_n, and generated sampled locations from the partition regions of R_n independently. Futhermore, different sampling densities can be assigned to different partition regions. Theorem 1 shows that the limiting distribution of $\hat{\boldsymbol{\beta}}_n$. For the case that there are more than one solutions to (2.2), Lahiri and Zhu (2006) gave a comprehensive

consideration, which is also applicable under the proposed one-per-stratum sampling design. However, $\Sigma_{\boldsymbol{\beta}}$ is usually intractable due to the lack of information on the sampling density functions as well as the spatial process, so we generalize a resampling method (Lahiri and Zhu, 2006) to make inference in next section.

Before closing this section, we compare the estimation efficiency of the proposed one-per-stratum sampling design with that considered by Lahiri and Zhu (2006). Denote $g(\boldsymbol{s})$ to be a density function on R_0. Assume that $\{\boldsymbol{S}_i^\dagger : i = 1, 2, \ldots\}$ are independently generated from R_0 using $g(\boldsymbol{s})$, and $\{\boldsymbol{S}_i^\dagger : i = 1, 2, \ldots\}$ are independent of the spatial process $\{Z(\boldsymbol{s}) : \boldsymbol{s} \in \mathbb{R}^d\}$. A sample for R_n is obtained by $\boldsymbol{s}_i = \lambda_n \boldsymbol{s}_i^\dagger$, where $\boldsymbol{s}_i^\dagger$ is a realization of $\boldsymbol{S}_i^\dagger$. Such a sampling design is also considered by Shao (2010) and Menezes et al. (2010). Specifically, consider $g(\boldsymbol{s}) = \{\text{vol.}(R_0)\}^{-1}$ $(\boldsymbol{s} \in R_0)$, and denote $\hat{\boldsymbol{\beta}}_{n,iid}$ to be the M-estimator under this design. For the proposed one-per-stratum sampling design, the partition regions satisfy Conditions 2–3 and have the same volume. Since a sample of size n is selected from R_n, let $\lambda_n^d = nc$ with a positive constant c, so $\text{vol.}(A_i) = \lambda_n^d \text{vol.}(R_0)/n = \text{vol.}(R_0)/c$. Let $f_i(\boldsymbol{s}) = n\{\lambda_n^d \text{vol.}(R_0)\}^{-1} \mathbb{1}(\boldsymbol{s} \in A_i)$ $(i = 1, \ldots, n)$ be the sampling density functions, so Condition 4 is also satisfied. Then, we have the following result.

Theorem 2. *Consider* $g(\boldsymbol{s}) = \{\text{vol.}(R_0)\}^{-1}$ $(\boldsymbol{s} \in R_0)$ *for the stochastic design (Lahiri and Zhu, 2006), and* $f_i(\boldsymbol{s}) = n\{\lambda_n^d \text{vol.}(R_0)\}^{-1}\mathbb{1}(\boldsymbol{s} \in A_i)$ *for* A_i $(i = 1, \ldots, n)$ *of the proposed one-per-stratum sampling design, where the partition regions have the same volume. Suppose Conditions 7–9 hold, and there exists a sequence of nonsingular matrices* $\{\Lambda_{n,iid}\}$ *such that*

$$\Lambda_{n,iid}^{-1}\left[\{\text{vol.}(R_0)\}^{-1}\int_{R_0}\boldsymbol{x}(\lambda_n\boldsymbol{s})\boldsymbol{x}(\lambda_n\boldsymbol{s})^\top \mathrm{d}\boldsymbol{s}\right]\Lambda_{n,iid}^{-1} \to H_{iid}, \tag{3.3}$$

$$\Lambda_{n,iid}^{-1}\left[\{\text{vol.}(R_0)\}^{-2}\int_{R_0}\boldsymbol{x}(\lambda_n\boldsymbol{s}+\boldsymbol{h})\boldsymbol{x}(\lambda_n\boldsymbol{s})^\top \mathrm{d}\boldsymbol{s}\right]\Lambda_{n,iid}^{-1} \to Q_{iid}(\boldsymbol{h}) \tag{3.4}$$

as $n \to \infty$, *where* H_{iid} *is a positive definite matrix,* $Q_{iid}(\boldsymbol{h})$ *is a* $p \times p$ *matrix-valued function such that* $\int Q_{iid}(\boldsymbol{h})\sigma_\Psi(\boldsymbol{h})d\boldsymbol{h}$ *is positive definite, and* $\boldsymbol{x}(\boldsymbol{s}) = 0$ *if* $\boldsymbol{s} \notin R_n$. *Then, we have*

$$\sqrt{n}\Lambda_{n,iid}(\hat{\boldsymbol{\beta}}_{n,iid} - \boldsymbol{\beta}) \to N(0, c\chi_0^{-2}\Sigma_{\boldsymbol{\beta},iid}), \quad \sqrt{n}\Lambda_{n,iid}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \to N(0, \chi_0^{-2}\Sigma_{\boldsymbol{\beta}})$$

in distribution $(P_{\cdot|\boldsymbol{S}})$ *almost surely* $(P_{\boldsymbol{S}})$, *where*

$$\Sigma_{\boldsymbol{\beta},iid} = c^{-1}H_{iid}^{-1}\sigma_\Psi(\boldsymbol{0}) + H_{iid}^{-1}\{\textstyle\int \sigma_\Psi(\boldsymbol{h})Q_{iid}(\boldsymbol{h})\,d\boldsymbol{h}\}H_{iid}^{-1},$$

$$\Sigma_{\boldsymbol{\beta}} = H_{iid}^{-1}\sigma_\Psi(\boldsymbol{0}) + H_{iid}^{-1}\{\textstyle\int \sigma_\Psi(\boldsymbol{h})\{cQ_{iid}(\boldsymbol{h}) - Q_1(\boldsymbol{h})\}\,d\boldsymbol{h}\}H_{iid}^{-1},$$

$$Q_1(h) = \lim_{n\to\infty}\Lambda_{n,iid}^{-1}\left[\{n^2\lambda_n^{-2d}vol.(R_0)^{-2}\}\int_{\cup_{i=1}^n\{A_i\cap(A_i-\boldsymbol{h})\}}\boldsymbol{x}(\boldsymbol{y}+\boldsymbol{h})\boldsymbol{x}(\boldsymbol{y})^\top d\boldsymbol{y}\right]\Lambda_{n,iid}^{-1}.$$

If the limit of $\int \sigma_\Psi(\boldsymbol{h})Q_1(\boldsymbol{h})d\boldsymbol{y}$ *is positive definitive, then* $\hat{\boldsymbol{\beta}}_n$ *is asymptotically more efficient than* $\hat{\boldsymbol{\beta}}_{n,iid}$ *in the sense that* $c\Sigma_{\boldsymbol{\beta},iid} - \Sigma_{\boldsymbol{\beta}}$ *is positive definitive.*

The proof of Theorem 2 is given in Section S4 of the Supplementary Material. Under generality conditions, Theorem 2 shows that $\hat{\boldsymbol{\beta}}_n$ is asymptotically more efficient than $\hat{\boldsymbol{\beta}}_{n,iid}$. The reduction of $\Sigma_{\boldsymbol{\beta}} - c\Sigma_{\boldsymbol{\beta},iid}$ is $H_{iid}^{-1} \int \sigma_\Psi(\boldsymbol{h}) Q_1(\boldsymbol{h}) \mathrm{d}\boldsymbol{y} H_{iid}^{-1}$, and such reduction is made possible due to the fact that the proposed one-per-stratum sampling design prevents two sampled locations from being too close to each other, which reduces the redundancy in the data.

4. Resampling method

We implement a resampling method (Lahiri and Zhu, 2006) to make inference for the M-estimator $\hat{\boldsymbol{\beta}}_n$ under the proposed one-per-stratum sampling design, and its validity is established in Theorem 3. Let $\mathcal{K}_n = \{\boldsymbol{k} \in \mathbb{Z}^d : (\boldsymbol{k}b_n + [0,1)^d b_n) \cap R_n \neq \emptyset\} = \mathcal{K}_{1n} \cup \mathcal{K}_{2n}$, where b_n is the block size satisfying certain conditions, $\mathcal{K}_{1n} = \{\boldsymbol{k} \in \mathbb{Z}^d : (\boldsymbol{k}b_n + [0,1)^d b_n) \subset R_n\}$, and $\mathcal{K}_{2n} = \mathcal{K}_n \cap \mathcal{K}_{1n}^{\mathrm{C}}$. The sampling region R_n can be partitioned by $\{R_n(\boldsymbol{k}) : \boldsymbol{k} \in \mathcal{K}_n\}$, where $R_n(\boldsymbol{k}) = R_n \cap \{\boldsymbol{k}b_n + [0,1)^d b_n\}$ for $\boldsymbol{k} \in \mathcal{K}_n$. That is, each $R_n(\boldsymbol{k})$ is an intersection area of R_n and a certain block. Thus, we have $R_n = \bigcup_{\boldsymbol{k} \in \mathcal{K}_n} R_n(\boldsymbol{k})$. The shape of $R_n(\boldsymbol{k})$ may vary for $\boldsymbol{k} \in \mathcal{K}_{2n}$.

Let $l_n = \{\boldsymbol{l} \in \mathbb{Z}^d : (\boldsymbol{l} + [0,1)^d b_n) \subset R_n\}$ be the index set of hypercubes $(\boldsymbol{l} + [0,1)^d b_n)$ that lie in R_n. Denote $\{I_{\boldsymbol{k}} : \boldsymbol{k} \in \mathcal{K}_n\}$ to be a set of independent

and identically distributed random variables with

$$P_*(I_{\boldsymbol{k}} = \boldsymbol{l}) = \frac{1}{|l_n|} \quad (\boldsymbol{l} \in l_n), \tag{4.5}$$

where P_* is the conditional distribution for the resampling method given $\mathcal{S}_n$ and $Y(\boldsymbol{s}_i)$ $(\boldsymbol{s}_i \in \mathcal{S}_n)$.

Let $B_n(\boldsymbol{l}; \boldsymbol{k}) = R_n(\boldsymbol{k}) - \boldsymbol{k}b_n + \boldsymbol{l}$ $(\boldsymbol{k} \in \mathcal{K}_n, \boldsymbol{l} \in l_n)$, so $B_n(\boldsymbol{l}; \boldsymbol{k})$ is congruent with $R_n(\boldsymbol{k})$. Denote $\mathcal{D}_n(R_n) = \{(\boldsymbol{x}(\boldsymbol{s}_i), Y(\boldsymbol{s}_i)) : \boldsymbol{s}_i \in \mathcal{S}_n\}$ to be the original sample, and a resample is

$$\mathcal{D}_n^*(R_n) = \{(\boldsymbol{x}(\boldsymbol{s}_i^*), Y(\boldsymbol{s}_i^*)) : \boldsymbol{s}_i^* \in \cup_{\boldsymbol{k} \in \mathcal{K}_n} B_n(I_{\boldsymbol{k}}; \boldsymbol{k})\}. \tag{4.6}$$

Let n^* be the sample size of the resampled observations, and it may be different from n by the resampling method. The resampled version of $\hat{\boldsymbol{\beta}}_n$, denoted as $\boldsymbol{\beta}_n^*$, is obtained by solving

$$\sum_{\boldsymbol{k} \in \mathcal{K}_n} \{S_n^*(\boldsymbol{\beta}; \boldsymbol{x}) - \hat{c}_n(\boldsymbol{k})\} = \boldsymbol{0} \tag{4.7}$$

with respect to $\boldsymbol{\beta} \in \mathbb{R}^p$, where

$$S_n^*(\boldsymbol{k}; \boldsymbol{x}) = \textstyle\sum_{i=1}^{n^*} \boldsymbol{x}(\boldsymbol{s}_i^*) \Psi\{Y(\boldsymbol{s}_i^*) - \boldsymbol{x}(\boldsymbol{s}_i^*)^\top \boldsymbol{x}\} \mathbb{1}\{\boldsymbol{s}_i^* \in B_n(I_{\boldsymbol{k}}; \boldsymbol{k})\},$$
$$\hat{c}_n(\boldsymbol{k}) = E_* \left[\textstyle\sum_{i=1}^{n^*} \boldsymbol{x}(\boldsymbol{s}_i^*) \Psi\{\hat{Z}(\boldsymbol{s}_i^*)\} \mathbb{1}(\boldsymbol{s}_i^* \in B_n(I_{\boldsymbol{k}}; \boldsymbol{k})) \right],$$

$\hat{Z}_n(\boldsymbol{s}_i) = Y(\boldsymbol{s}_i) - \boldsymbol{x}(\boldsymbol{s}_i)^\top \hat{\boldsymbol{\beta}}_n$, and E_* is the conditional expectation with respect to the distribution P_* in (4.5). The calibration factor $\hat{c}_n(\boldsymbol{k})$ guarantees

that $\boldsymbol{\beta}_n^*$ is an unbiased estimator of $\hat{\boldsymbol{\beta}}_n$ under the conditional distribution P_*.

Denote $P_{\cdot|\boldsymbol{S}}$ to be the conditional probability given $\mathcal{S}_n = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$ and $E_{\cdot|\boldsymbol{S}}$ and $V_{\cdot|\boldsymbol{S}}$ to be the corresponding conditional mean and variance, respectively. For the resampling method, we have the following result under the proposed one-per-stratum sampling design.

Theorem 3. *Suppose Conditions 1–9 hold and*

$$b_n^{-1} + b_n/\lambda_n = o(1) \quad (n \to \infty). \tag{4.8}$$

Then,

$$\sup_{B \in \mathcal{B}(\mathbb{R}^p)} \left| P_*(T_{1n}^* \in B) - P_{\cdot|\boldsymbol{S}}(T_{1n} \in B) \right| \to 0 \quad \textit{in } P_{\cdot|\boldsymbol{S}}\text{-probability} \tag{4.9}$$

almost surely $(P_{\boldsymbol{S}})$, *where* $\mathcal{B}(\mathbb{R}^p)$ *is the Borel* σ*-algebra on* $\mathbb{R}^p$, $T_{1n}^* = \Lambda_n^{-1}(\boldsymbol{\beta}_n^* - \hat{\boldsymbol{\beta}}_n)$, *and* $T_{1n} = \Lambda_n^{-1}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)$.

The proof of Theorem 3 is given in Section S5 of the Supplementary Material. Theorem 3 shows that $\boldsymbol{\beta}_n^*$ can be used to make inference for $\boldsymbol{\beta}_0$ under the proposed one-per-stratum sampling design.

Remark 4. It is worthy pointing out that block bootstrap is commonly used to make statistical inference for dependent data (Lahiri, 2018; Hala et al., 2020; Chan et al., 2022; Zhang et al., 2023). In this paper, our goal is

not to propose a new resampling method. Instead, we would to show that valid inference can be achieved by a block bootstrap under the proposed sampling design. Other than the resampling method in Lahiri and Zhu (2006), we conjecture that other bootstrap methods (Das and Lahiri, 2019; Kurisu et al., 2023) are also valid for the proposed sampling design, but the associated verification is beyond our scope.

However, the choice of the block size b_n remains an open problem under the proposed one-per-stratum sampling design. We use an empirical method (Hall et al., 1995) to choose the optimal block size. Denote $B_n = \{b_{n,1}, \ldots, b_{n,K}\}$ to be a set of K valid block sizes satisfying (4.8), where $K \geq 1$ is a user-specified number. Let $\{R_n^{(h)} : h = 1, \ldots, H\}$ be a set of pairwise distinct subregions of R_n. For each $b_n \in B_n$, let $b_n^{(h)} = b_n\{\text{vol.}(R_n^{(h)})/\text{vol.}(R_n)\}^{1/d}$ $(h = 1, \ldots, H)$. Based on $R_n^{(h)}$ and $b_n^{(h)}$, obtain the variance estimator of $\hat{\boldsymbol{\beta}}_n^{(h)}$ by the resampling method, say $V_n^{*(h)}$, where $\hat{\boldsymbol{\beta}}_n^{(h)}$ solves (2.2) using the observations in $R_n^{(h)}$. The optimal block size is chosen to be the one that minimizes $\Xi(b_n) = \sum_{i=1}^{p}\sum_{h=1}^{H}(V_{n,i}^{*(h)} - V_{n,i}^{*})^2$, where V_n^* is the estimated variance of $\hat{\boldsymbol{\beta}}_n$ by the resampling method using block size b_n and the original observations, and $V_{n,i}^{*(h)}$ and $V_{n,i}^{*}$ are the i-th diagonal element of $V_n^{*(h)}$ and V_n^*, respectively. Notice that the block size b_n applies as long as it satisfies (4.8). Intuitively, if we scale both the

sampling region and the block size simultaneously, the bootstrap variance estimators should perform similarly, and we can use this intuition to choose an "optimal" block size empirically. That is, we would like to choose an optimal one, which guarantees that the squared difference between the two variance estimators, including $V^*_{n,i}$ and $V^{*(h)}_{n,i}$, are minimized; see Hall et al. (1995) for details.

5. Simulation studies

5.1 Spatial balance test

In this section, the spatial balance of the proposed one-per-stratum sampling design is compared with the generalized random tessellation stratified design (Stevens and Olsen, 2004) and a local pivotal method (Grafström et al., 2012). One finite population consists of 100×100 equally spaced points on the unit square $[0, 1] \times [0, 1]$, and inclusion probability, which is the probability a specific point is sampled, is the same for each point. For the proposed one-per-stratum sampling design, the sampling region is evenly partitioned, and a uniform sampling density function is used within each partition region. Three designs are conducted to generate a sample of size n, and consider $n = 25$, $n = 100$, and $n = 400$.

The Voronoi polygon method (Stevens and Olsen, 2004) is modified to

measure the spatial balance of a given sample. For a sampled location $\boldsymbol{s}_i$, the Voronoi polygon associated with $\boldsymbol{s}_i$, say V_i, is the set of points that are closer to $\boldsymbol{s}_i$ than other sampled elements. If the sample is spatially balanced, we expect that $na_i \approx 1$ $(i = 1, \ldots, n)$, where $a_i = \text{vol.}(V_i)$. Thus, $\zeta = n^{-1} \sum_{i=1}^{n} (na_i - 1)^2$ is a good measure of the spatial balance for a given sample. Denote $\eta_{gops} = \zeta_{gops}/\zeta_{grts}$ and $\eta_{lpm} = \zeta_{lpm}/\zeta_{grts}$, and Grafström et al. (2012) showed that $\eta_{lpm} < 1$, where ζ_{gops}, ζ_{grts} and ζ_{lpm} are associated with the proposed one-per-stratum sampling design, the generalized random tessellation stratified design and the local pivotal method, respectively.

We conduct 1 000 Monte Carlo simulations for each sample size and design, and Table 1 shows the Monte Carlo mean and standard error of statistics η_{gops} and η_{lpm}. Compared with the generalized random tessellation stratified design, the sample from the proposed one-per-stratum sampling design is more spatially balanced when the sample size is larger. Even though the sample generated by the proposed one-per-stratum sampling design is not as spatially balanced as that by the local pivotal method under the simulation setup, we can use a more concentrated sampling density function within each partial region to get a sample with better spatial balance.

Remark 5. As noted by Grafström et al. (2012), the expected computation

Table 1: Monte Carlo mean (outside of the parenthesis) and standard error (inside of the parenthesis) of the spatial balance statistics.

Sample size	η_{gops}	η_{lpm}
$n = 25$	0.896 (0.069)	0.887 (0.078)
$n = 100$	0.748 (0.017)	0.701 (0.016)
$n = 400$	0.716 (0.005)	0.645 (0.004)

complexity for the local pivotal method is $O(N^2)$, where N is size of the finite population. The algorithm for generating a sample by the generalized random tessellation stratified design is even slower than the local pivotal method. Once the partition regions are given, the computation complexity for the proposed one-per-stratum sampling design is $O(n)$, which is much smaller than its two competitors. Furthermore, since the sample is independently obtained within each partition region, the proposed sampling design can be further accelerated by trivial parallelization, while the other two cannot. Another advantage of the proposed one-per-stratum sampling design is that it can generate sample from an *infinite* population, which is the case in many spatial area sampling problems.

5.2 Spatial linear regression model

In this section, the resampling method is tested under the proposed one-per-stratum sampling design. The prototype area is $R_0 = (-1/2, 1/2] \times (-1/2, 1/2]$. The spatial linear regression model is

$$Y(\boldsymbol{s}) = \beta_0 + \beta_1 \log(1 + |s_1|) + Z(\boldsymbol{s}) \quad (\boldsymbol{s} \in R_n), \tag{5.10}$$

where $(\beta_0, \beta_1) = (1, 1)$, and $Z(\boldsymbol{s})$ is a zero-mean stationary process with spherical semivariogram that has unit sill and range r; see (Cressie, 2015, Section 2.3.1) for details. Consider $r \in \{1, 2\}$, $n \in \{400, 900\}$, and set the sampling rate as $n/\lambda_n^2 = 25/36$. For the proposed one-per-stratum sampling design, squares with 1.2 on a side are used to partition the sampling region, and the sampling density function within each partition region is uniform or a truncated bivariate normal distribution with mean at the center and $0.15^2\mathcal{I}$ as the covariance matrix, where $\mathcal{I}$ is the 2×2 identity matrix. Thus, the sampling design with truncated bivariate normal sampling density functions can generate a sample with better spatial balance compared with the one using uniform sampling density functions.

First, the relative efficiency of the proposed one-per-stratum sampling design is compared with a stochastic design (Lahiri and Zhu, 2006), and a uniform distribution is used for the latter. For $i = 0, 1$, denote $eff(\beta_i) =$

$V_{gops}(\hat{\beta}_{n,i})/V_{iid}(\hat{\beta}_{n,i})$, where $(\hat{\beta}_{n,0}, \hat{\beta}_{n,1})$ solves (2.2), and $V_{gops}(\hat{\beta}_{n,i})$ and $V_{iid}(\hat{\beta}_{n,i})$ are the variances of $\hat{\beta}_{n,i}$ under the proposed one-per-stratum sampling design and the stochastic design (Lahiri and Zhu, 2006), respectively. We conduct 1 000 Monte Carlo simulations to estimate $eff(\beta_0)$ and $eff(\beta_1)$, and Table 2 shows the comparison results. Values of the relative efficiency are less than one for all scenarios, indicating that the proposed one-per-stratum sampling design has a better estimation efficiency. Specifically, under uniform sampling, since the sampling rate $n\lambda_n^{-2}$ does not change for $n = 400$ and $n = 900$, $eff(\beta_i)$ is almost the same regardless of the sample size when the spatial dependence is fixed, where $i = 0, 1$. Besides, as the spatial dependence becomes stronger, the proposed one-per-stratum sampling design has better estimation efficiency compared with the stochastic design (Lahiri and Zhu, 2006) since $\sigma_\Psi(\boldsymbol{h})$ decays to 0 slower. Refer to Theorem 2 for the theoretical justification of the two finding under uniform sampling. When the truncated bivariate normal sampling density is used, the estimation efficiency of the proposed one-per-stratum sampling design becomes better as the sample size increases. In addition, as the spatial dependence becomes stronger, the efficiency gain of the proposed one-per-stratum sampling design is compromised since the "effective sample size" decreases.

Table 2: Summary statistic for the relative efficiency of the M-estimator by the sample generated by the proposed one-per-stratum sampling design and the stochastic design (Lahiri and Zhu, 2006).

Density	Dependence	$n = 400$		$n = 900$	
		$eff(\beta_0)$	$eff(\beta_1)$	$eff(\beta_0)$	$eff(\beta_1)$
Uniform	$r = 1$	0.87	0.86	0.83	0.82
	$r = 2$	0.81	0.79	0.83	0.83
Normal	$r = 1$	0.70	0.73	0.74	0.78
	$r = 2$	0.73	0.74	0.76	0.78

Uniform: uniform sampling design function; Normal: bivariate normal sampling density.

Next, we generate 1 000 Monte Carlo samples to obtain the M-estimator of $\boldsymbol{\beta}$, and 1 000 resamplings are conducted for each sample. The set of valid block sizes is $B_{n,1} = \{2, 3\}$ when $\lambda_n = 24$, and $B_{n,2} = \{3, 4\}$ when $\lambda_n = 36$. The subregions are chosen to be the four conjugate halves of the original sampling region for choosing an optimal block size. To test the performance of the resampling method, we consider the square root of mean square error, the relative bias for the variance estimator and the coverage rate of the 90% confidence interval constructed by the resampling method under the proposed one-per-stratum sampling design. Table 3 summarizes the

estimation results. As the sample size increases, the square root of mean square error and the absolute value of relative bias for the variance estimator decrease for both sampling designs. For a fixed sample size and block size, the square root of mean square error and the absolute value of relative bias increase as the spatial dependence becomes stronger since the number of effective sampling size decreases; see (Cressie, 2015, Section 4.6.2) for details. Besides, the coverage rate gets closer to 90% as the sample increases. Please notice that even when $n = 900$, the coverage rates are far less than the nominal truth 0.9 for the case $r = 2$. There are two main reasons for this unfavorable result. First, the effective sample size is small even when $n = 900$. More importantly, even when the block size b_n equals to 4, the spatial correlation cannot be captured well. In an additional simulation study, we have increased the sample size to $n = 6\,400$ and the block size to $b_n = 10$ for the case $r = 2$. When uniform sampling density functions are used, the coverage rates for β_0 and β_1 are 0.91 and 0.91, respectively. They are 0.88 and 0.89 when truncated normal sampling density functions are applied. Thus, as the sample size and block size diverge, we can get better coverage rates, and this observation is exactly what we have required in Theorem 3. From Table 3, we can conclude that the selected optimal block sizes are reasonable, and a design with bivariate normal density functions

performs better since samples with better spatial balance are generated.

For comparison, we also consider a naive method using simple linear regression to make inference for the regression parameters. When the spatial dependence is strong, the relative bias of the variance estimator is less than -0.5 and the corresponding coverage rate of the 90% confidence interval is only around 0.7; see Section S6 of the Supplementary Material for details. Thus, simple linear regression can lead to erroneous statistical inference for the regression parameters if the spatial dependence is ignored.

6. Soil erosion analysis

We investigate the relationship between the soil erosion and slope for the cropland of Iowa based on the most recent sample from the National Resources Inventory program. The National Resources Inventory was initially mandated by the Rural Development Act of 1972 , and it is a longitudinal survey to provide scientific information about natural resources, such as soil, water and other related resources, on the Nation's non-federal land; see `https://www.nrcs.usda.gov` for details.

In order to guarantee spatial balance, a two-stage stratified area sample is generated from the 49 continental States (exclude Alaska), Puerto Rico, and the Virgin Islands. Small political areas or geographic units are

used to stratify the sampling region. In Iowa, for example, the size of each stratum is about 2×6 miles. The primary sampling units are segments of land of size 0.5×0.5 miles. The original sampling design selects 1–4 primary sampling units within each stratum, and 1–3 points are selected within each primary sampling units. In order to achieve better spatial balance, restricted randomization is used to select primary sampling units and points; see Nusser and Goebel (1997) and Nusser et al. (1998) for details. In this study, we treat the primary sampling units as sample elements by averaging information of the points within each of them. The most recent survey was conducted in 2017, and there are 5981 primary sampling unites selected in Iowa. The left panel of Figure 2 shows the locations of the corresponding primary sampling units. The sample is spatially balanced, and the sampling rates differ for different counties. For example, Lyon, the county on the northwest, was over-sampled.

The soil erosion of the cropland is a critical issue when we consider effectiveness of soil and water conservation practices, and we are interested in investigating the relationship between the soil erosion and slope of cropland in Iowa. The right panel of Figure 2 shows the relationship between the log soil erosion and log slope, so it is reasonable to consider the spatial linear regression model in log scale. Berg and Chandra (2014) considered

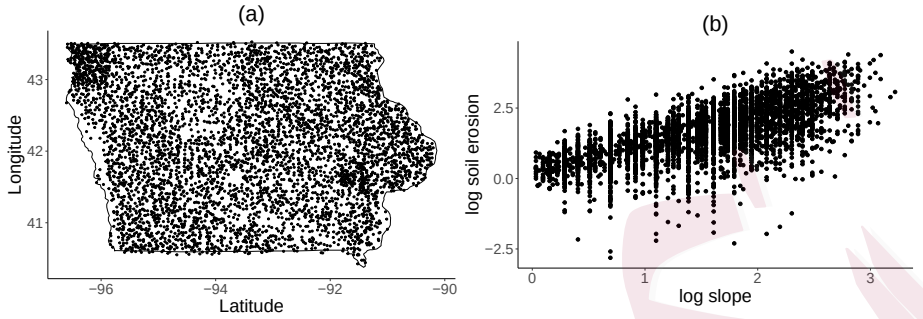

Figure 2: The left panel (a) shows locations of the primary sampling units, and the right panel (b) shows the relationship between the log soil erosion and log slope. Due to the confidential restriction, the real locations are randomly shifted.

the following model:

$$\log\, y_i = \beta_0 + \beta_1 \log\, x_i + Z_i, \tag{6.11}$$

where (x_i, y_i) represents the slope and soil erosion of the ith primary sampling units, and Z_i corresponds to a zero mean spatial process. Notice that the slope is bounded by a constant for each location, so the conditions in Section 3 are satisfied. We adopt the same model, and our goal is to provide a valid confidence interval for the regression parameters in (6.11) using the resampling method developed in this paper. For comparison, a simple linear regression analysis is conducted by assuming that $\{Z_i : i = 1, \ldots, n\}$ are

independently generated, and such an assumption is often used in survey sampling; see (Breidt and Fuller, 1999, Section 2.3) for details.

Hall's method (Hall et al., 1995) is used to select an optimal block size, and let $B_n = \{0.15, \ldots, 0.9\}$ be a set of block sizes with step size 0.05; see the last paragraph of Section 4 for details about Hall's method. Figure 3 shows values of $\Xi(b_n)$ ($b_n \in B_n$), and we conclude that the target function $\Xi(b_n)$ is minimized with $b_n = 0.35$. Thus, $b_n = 0.35$ is used as the optimal block size for the following analysis.

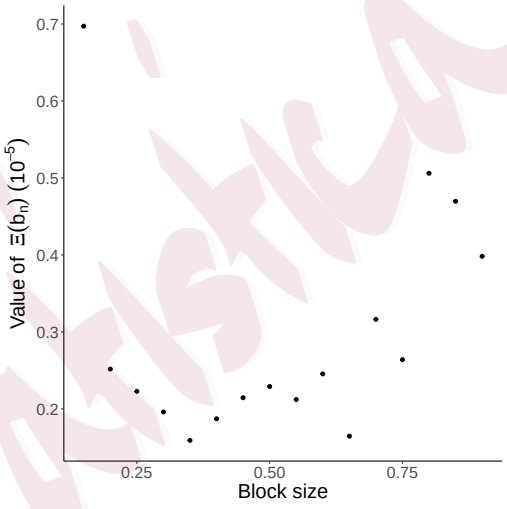

Figure 3: Values of Ξ_b for different valid block sizes.

We have repeated the resampling method 1 000 times to obtain the 90% confidence interval for the estimated regression parameters in (6.11), and Table 4 shows the estimation results. Compared with the simple linear

regression, we get larger variance estimators and wider confidence intervals since spatial dependence reduces the amount of information. The slope has a significant positive effect in estimating the soil erosion. Thus, a cropland with larger slope suffers more from the soil erosion, so policy makers should pay more attention on such croplands. The estimated value for β_1 is 0.97, which is close to the theoretical value in the universal soil loss equation (Wischmeier and Smith, 1978).

7. Discussion

The proposed one-per-stratum sampling design is a special case of the one discussed by Krewski and Rao (1981). The difference between these two is that we are interested in making inference for the parameters in the super-population model (2.1), while Krewski and Rao (1981) focused on the finite population. Since the spatial balance of the sample is emphasized, only one element is selected from each stratum. Generally, more than one elements can be selected from each stratum, and the theoretical properties of the resampling method still apply under certain conditions; also see Zhang and Fuller (2019).

In this paper, one basic assumption is that the sampling design is ignorable in the sense that the sampling mechanism only depends on the

covariate $\boldsymbol{x}(\boldsymbol{s})$ (Pfeffermann, 1993). It is an interesting topic to investigate the theoretical properties of the resampling method under non-ignorable one-per-stratum sampling designs in the future.

Supplementary Materials

The online supplementary material contains a brief description of the Stochastic sampling design of Lahiri and Zhu (2006) (S1), the proofs of Theorem 1 (S3), Theorem 2 (S4) and Theorem 3 (S5), and an additional simulation result for a simple linear regression (S6).

Acknowledgments

We are grateful for Dr. Xiaotong Shen, an anonymous associated editor and two referees for the constructive comments to improve the accessibility of this paper. Wang is partially supported by National Key R&D Program of China 2022YFA1003800, NSFC (No.: 72033002,12231011, 71988101), Humanities and Social Sciences Foundation of the Ministry of Education of China Grant (23YJA910005) and NSSFC (No. 23CMZ005). Zhu is partially supported by NSF SES 1952007 and NSF AST 2232461.

References

Bartholdi, J. J. and L. K. Platzman (1988). Heuristics based on spacefilling curves for combinatorial problems in Euclidean space. *Manag. Sci. 34*(3), 291–305.

Berg, E. and H. Chandra (2014). Small area prediction for a unit-level lognormal model. *Comput. Stat. Data. Anal. 78*, 159–175.

Breidt, F. J. (1995). Markov chain designs for one-per-stratum spatial sampling. In *Proceedings of the Section on Survey Research Methods, American Statistical Association, Washington, DC*, pp. 356–361.

Breidt, F. J. and W. A. Fuller (1999). Design of supplemented panel surveys with application to the national resources inventory. *Journal of Agricultural, Biological, and Environmental Statistics 4*(4), pp. 391–403.

Chan, N. H., R. Zhang, and C. Y. Yau (2022). Inference for structural breaks in spatial models. *Statistica Sinica 32*(4).

Cochran, W. G. (1946). Relative accuracy of systematic and stratified random samples for a certain class of populations. *Ann. Math. Stat. 17*(2), 164–177.

Cressie, N. A. C. (2015). *Statistics for Spatial Data* (Revised ed.). New York: John Wiley.

Das, D. and S. Lahiri (2019). Second order correctness of perturbation bootstrap M-estimator of multiple linear regression parameter. *Bernoulli 25*(1), 654 – 682.

Grafström, A., N. L. P. Lundström, and L. Schelin (2012). Spatially balanced sampling through

the pivotal method. *Biometrics 68*(2), 514–520.

Grenander, U. (1954). On the estimation of regression coefficients in the case of an autocorrelated disturbance. *Ann. Math. Stat. 25*(2), 252–272.

Hala, M. V., S. Bandyopadhyay, S. N. Lahiri, and D. J. Nordman (2020). A general frequency domain method for assessing spatial covariance structures. *Bernoulli 26*(4), 2463 – 2487.

Hall, P., J. L. Horowitz, and B.-Y. Jing (1995). On blocking rules for the bootstrap with dependent data. *Biometrika 82*(3), 561–574.

Koul, H. L. (1992). M-estimators in linear models with long range dependent errors. *Statist. Probab. Lett. 14*(2), 153–164.

Krewski, D. and J. N. K. Rao (1981). Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods. *Ann. Statist.*, 1010–1019.

Kurisu, D., K. Kato, and X. Shao (2023). Gaussian approximation and spatially dependent wild bootstrap for high-dimensional spatial data. *Journal of the American Statistical Association* ((Accepted))), 1–21.

Lahiri, S. N. (2003). Central limit theorems for weighted sums of a spatial process under a class of stochastic and fixed designs. *Sankhya (2003–2007) 65*(2), 356–388.

Lahiri, S. N. (2018). Uncertainty quantification in robust inference for irregularly spaced spatial data using block bootstrap. *Sankhya A 80*, 173–221.

REFERENCES

Lahiri, S. N. and K. Mukherjee (2004). Asymptotic distributions of M-estimators in a spatial regression model under some fixed and stochastic spatial sampling designs. *Ann. Inst. Statist. Math. 56*(2), 225–250.

Lahiri, S. N. and J. Zhu (2006). Resampling methods for spatial regression models under a class of stochastic designs. *Ann. Statist. 34*(4), 1774–1813.

Lister, A. J. and C. T. Scott (2009). Use of space-filling curves to select sample locations in natural resource monitoring studies. *Environ. Monit. Assess. 149*(1), 71–80.

Menezes, R., C. Ferreira, and P. García-Soidán (2010). Nonparametric spatial prediction under stochastic sampling design. *J. Nonparametr. Stat. 22*(3), 363–377.

Munholland, P. L. and J. J. Borkowski (1996). Simple latin square sampling+ 1: A spatial design using quadrats. *Biometrics 52*(1), 125–136.

Nordman, D. J. and S. N. Lahiri (2004). On optimal spatial subsample size for variance estimation. *Ann. Statist. 32*(5), 1981–2027.

Nordman, D. J., S. N. Lahiri, and B. L. Fridley (2007). Optimal block size for variance estimation by a spatial block bootstrap method. *Sankhya (2003–2007) 69*(3), 468–493.

Nusser, S., F. Breidt, and W. Fuller (1998). Design and estimation for investigating the dynamics of natural resources. *Ecological Applications 8*(2), 234–245.

Nusser, S. and J. Goebel (1997). The national resources inventory: a long-term multi-resource monitoring programme. *Environmental and Ecological Statistics 4*(3), 181–204.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *Int. Stat. Rev. 61*(2), 317–337.

Politis, D. N., E. Paparoditis, and J. P. Romano (1998). Large sample inference for irregularly spaced dependent observations based on subsampling. *Sankhya A 60*(2), 274–292.

Rubin-Bleuer, S. and I. Schiopu-Kratina (2005, 12). On the two-phase framework for joint model and design-based inference. *Ann. Statist. 33*(6), 2789–2810.

Shao, X. (2010). The dependent wild bootstrap. *J. Amer. Statist. Assoc. 105*(489), 218–235.

Stevens, D. L. and A. R. Olsen (2004). Spatially balanced sampling of natural resources. *J. Amer. Statist. Assoc. 99*(465), 262–278.

Tillé, Y., L. Qualité, and M. Wilhelm (2018). Sampling designs on finite populations with spreading control parameters. *Statist. Sinica*, 471–504.

Wang, Z. and Z. Zhu (2019). Spatiotemporal balanced sampling design for longitudinal area surveys. *J. Agric. Biol. Environ. Stat. 24*(2), 245–263.

Wischmeier, W. H. and D. D. Smith (1978). *Predicting Rainfall Erosion Losses: A Guide to Conservation Planning*. Number 537. Department of Agriculture, Science and Education Administration.

Yajima, Y. (1991). Asymptotic properties of the LSE in a regression model with long-memory stationary errors. *Ann. Statist. 19*(1), 158–177.

Zhang, R., N. H. Chan, and C. Chi (2023). Nonparametric testing for the specification of spatial

trend functions. *Journal of Multivariate Analysis 196*, 105180.

Zhang, X. and W. A. Fuller (2019, 12). A Sampling Design for Ordered Populations. *J. Surv. Stat. Methodol. 9*(1), 121–140.

Wang Yanan Institute for Studies in Economics and School of Economics, Xiamen University

E-mail: wangzl@xmu.edu.cn

Department of Statistics, Iowa State University

E-mail: zhuz@iastate.edu

Table 3: Summary statistics for the resampling method under the proposed one-per-stratum sampling design for different scenarios using different sampling densities.

Stat	r	Uniform sampling density						Truncated normal sampling density					
		$n = 400$			$n = 900$			$n = 400$			$n = 900$		
		b_n	β_0	β_1	b_n	β_0	β_1	b_n	β_0	β_1	b_n	β_0	β_1
RMSE	1	2	0.63†	0.14†	3	0.25†	0.04†	2	0.52†	0.12†	3	0.21†	0.04†
		3	0.73	0.16	4	0.30	0.05	3	0.65	0.14	4	0.26	0.05
	2	2	1.79†	0.44†	3	0.68†	0.11†	2	1.88†	0.44†	3	0.65†	0.12†
		3	1.67	0.38	4	0.69	0.11	3	1.76	0.37	4	0.66	0.12
RB	1	2	-0.14†	-0.10†	3	-0.09†	-0.03†	2	-0.12†	-0.06†	3	-0.04†	-0.02†
		3	-0.15	-0.07	4	-0.10	-0.02	3	-0.14	-0.04	4	-0.05	-0.02
	2	2	-0.34†	-0.31†	3	-0.24†	-0.18†	2	-0.37†	-0.32†	3	-0.24†	-0.22†
		3	-0.28	-0.22	4	-0.22	-0.13	3	-0.31	-0.22	4	-0.21	-0.17
CR	1	2	0.86†	0.88†	3	0.89†	0.89†	2	0.88†	0.88†	3	0.89†	0.90†
		3	0.87	0.88	4	0.88	0.90	3	0.87	0.88	4	0.89	0.90
	2	2	0.81†	0.83†	3	0.85†	0.86†	2	0.82†	0.82†	3	0.85†	0.85†
		3	0.83	0.86	4	0.85	0.88	3	0.83	0.84	4	0.84	0.86

RMSE: square root of the mean square error; RB: relative bias; CR: coverage rate; †: optimal block size.

Table 4: Estimation results for the regression parameters in (6.11) based on the proposed method (Proposed) and simple linear regression (SLR).

Par	Est	Proposed		SLR	
		SE	90%CI	SE	90%CI
β_0	0.172	0.025	(0.129, 0.214)	0.021	(0.138, 0.207)
β_0	0.966	0.017	(0.936, 1.000)	0.013	(0.944, 0.988)

Est: estimated parameter; SE: estimated standard error; 90% CI: 90% confidence interval.