Forest Ecology and Management 489 (2021) 119051

Contents lists available at ScienceDirect

Forest Ecology and Management

journal homepage: www.elsevier.com/locate/foreco

Generalizing systematic adaptive cluster sampling for forest ecosystem inventory

Qing Xu [a], Göran Ståhl [b], Ronald E. McRoberts [c], Bo Li [d], Timo Tokola [e], Zhengyang Hou [f,*]

[a] Key Laboratory of National Forestry and Grassland Administration/Beijing for Bamboo & Rattan Science and Technology, International Centre for Bamboo and Rattan, Beijing 100102, People's Republic of China
[b] Department of Forest Resource Management, Swedish University of Agricultural Sciences, Umeå, Sweden
[c] Department of Forest Resources, University of Minnesota, Saint Paul, MN, United States
[d] Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL, United States
[e] School of Forest Sciences, University of Eastern Finland, Joensuu, Finland
[f] College of Forestry, Beijing Forestry University, Beijing 100083, People's Republic of China

ARTICLE INFO

Keywords:
Design-based inference
Adaptive inventory
Adaptive cluster sampling
Systematic sampling
Estimation
Simulation

ABSTRACT

Reliable statistical inference is central to forest ecology and management, much of which seeks to estimate population parameters for forest attributes and ecological indicators for biodiversity, functions and services in forest ecosystems. Many populations in nature such as plants or animals are characterized by aggregation of tendencies, introducing a big challenge to sampling. Regardless, a biased or imprecise inference would mislead analysis, hence the conclusion and policymaking. Systematic adaptive cluster sampling (SACS) is design-unbiased and particularly efficient for inventorying spatially clustered populations. However, (1) over-sampling is common for nonrare variables, making SACS a difficult choice for inventorying common forest attributes or ecological indicators; (2) a SACS sample is not completely specified until the field campaign is completed, making advance budgeting and logistics difficult; (3) even for rare variables, uncertainty regarding the final sample still persists; and (4) a SACS sample may be variable-specific as its formation can be adapted to a particular attribute or indicator, thus risking imbalance or non-representativeness for other jointly observed variables. Consequently, to solve these challenges, we aim to develop a generalized SACS (GSACS) with respect to the design and estimators, and to illustrate its connections with systematic sampling (SS) as has been widely employed by national forest inventories and ecological observation networks around the world. In addition to theoretical derivations, empirical sampling distributions were validated and compared for GSACS and SS using sampling simulations that incorporated a comprehensive set of forest populations exhibiting different spatial patterns. Five conclusions are relevant: (1) in contrast to SACS, GSACS explicitly supports inventorying forest attributes and ecological indicators that are nonrare, and solved SACS problems of oversampling, uncertain sample form, and sample imbalance for alternative attributes or indicators; (2) we demonstrated that SS is a special case of GSACS; (3) even with fewer sample plots, GSACS gives estimates identical to SS; (4) GSACS outperforms SS with respect to inventorying clustered populations and for making domain-specific estimates; and (5) the precision in design-based inference is negatively correlated with the prevalence of a spatial pattern, the range of spatial autocorrelation, and the sample plot size, in a descending order.

1. Introduction

Reliable statistical inference is central to forest ecology, much of which seeks to estimate population parameters for forest attributes such as forest area, productivity, carbon, or ecological indicators for biodiversity, functions and services in forest ecosystems (Margules and Pressey, 2000; Williams and Brown, 2019). National parameter estimates for these variables are also required by numerous international agreements. For example, the Montréal Process, Forest Europe, and the Convention on Biological Diversity jointly require that member countries assess, monitor, and report on sustainability and biodiversity indicators. The Global Climate Observing System further mandates the

* Corresponding author at: College of Forestry, Beijing Forestry University, Beijing 100083, People's Republic of China.
E-mail address: houzhengyang@bjfu.edu.cn (Z. Hou).

https://doi.org/10.1016/j.foreco.2021.119051
Received 12 December 2020; Received in revised form 27 January 2021; Accepted 8 February 2021
Available online 1 March 2021
0378-1127/© 2021 Elsevier B.V. All rights reserved.

inferential uncertainty for essential variables must not exceed 20% (Sessa and Dolman, 2008), a challenging but necessary requirement, because biased or imprecise procedures from overlooking sampling and inference would mislead analysis, hence conclusion and policy-making (Conn et al., 2017).

There are two inferential frameworks, the model-based inference and the design-based inference (Särndal, 1978), both having important places in forest ecosystem inventory (Kangas et al., 2018). Unbiasedness and small variance are common goals in the respective frameworks. Inferential uncertainty expressed by the mean squared error is decomposable into two independent terms, one representing the variance of a model- or design-based estimator, known as inferential precision, and the other term representing the bias of this estimator, known as inferential bias (Cassel et al., 1977). The population representing a spatial area of interest is tessellated with small units of a given size serving as population units (Cochran, 1977). The two frameworks fork towards different directions by treating the attribute value in a population unit to be random or fixed (Gregoire, 1998). Model-based inference regards the attribute value as a random variable which follows a distribution determined by a superpopulation model. A real-life population is considered a random realization of this superpopulation model, thus making population parameters such as the mean or total random variables as well (Graubard and Korn, 2002). The model-based estimator of a population parameter is generally not considered unbiased, suggesting that inferential uncertainty regarding a realized population may come from both the variance and bias (McRoberts et al., 2018). Because the superpopulation model is unknown in practice, the inferential uncertainty fully relies on a proxy model constructed using a sample and auxiliary data from sources such as remote sensing (McRoberts, 2011; Xu et al., 2018). When the proxy model represents the relationship between the dependent and independent variables without systematic error, the inferential precision is usually greater for the model-based inference than design-based inference (Hou et al., 2018). However, when imbalance-sampled, or when the model mis-specified or mis-used with external model or auxiliary data, a model-based estimator may not only reduce the inferential precision, but increase the inferential bias, and hence inferential uncertainty as a whole (Hou et al., 2017). Therefore, cautions are required, particularly for the use of model-based estimates when reporting to agreements that explicitly advocate elimination of inferential bias (IPCC, 2003).

In contrast, design-based inference regards the attribute value for a population unit as fixed, with the result that the population parameter value is also fixed, rather than random as is the case for model-based inference. Estimators are design-unbiased if they correctly correspond to samples selected using probabilistic designs, meaning the unbiasedness does not depend on assumptions about the population. This assures the inferential bias to be zero, so the inferential uncertainty reduces to the inferential precision which results from the randomness of selecting a sample from the population. Systematic sampling (SS) is design-based and the corresponding estimators are design-unbiased, with a long history of serving official reporting instruments at local, regional, ecosystem and national scales (Kangas and Maltamo 2006). SS is convenient for logistic; specifically, it is often less costly to measure a collection of SS plots than to measure an equal number of plots selected at random (Heikkinen, 2006). Many national forest inventory (NFI) programs including for the Nordic countries, France, the United States of America and China were established using SS (Tomppo et al., 2010; Zeng et al., 2015). An NFI population typically is a country, a state or province from which a sample of regularly spaced population units in the form of sample plots is selected and measured for estimating parameters for a comprehensive set of forest attributes and ecological indicators. The number of sample plots annually measured is phenomenal, typically ranging from 4400 to 15,000 in the Nordic countries, and 1200 to 3500 in individual American states (Hou et al., 2021; Räty et al., 2020). Therefore, cost-efficiency is crucial for field campaigns at this scale, involving a tradeoff between the inferential precision and the number of sample plots. Although SS is expected to improve inferential precision for clustered populations (Cochran 1977), this feature is conditional on a design strategy that can make the within-primary-unit correlation coefficient less than zero ($\rho < 0$); otherwise, the inferential precision for SS would be the same ($\rho = 0$) or even less ($\rho > 0$) than simple random sampling (Thompson, 2012, p. 166).

Much of the attention for SS is focused on optimizing design strategies for which sample selection does not depend in any way on observations made during a campaign, so the entire sample may be selected prior to fieldwork (Magnussen et al., 2020). Occasionally, foresters and ecologists are inclined to improvise during a field campaign, based on what has been observed thus far, as to which plots to, or not to, observe next (Acharyal et al., 2000). For example, in an inventory of a rare, spatially clustered plant population, one may wish to add additional plots to an initial SS sample once a large abundance of this species is encountered. This adaptiveness is sensible, because often the clustering pattern, size and location cannot be predicted beforehand, so that traditional means of increasing precision with SS or stratification are not always efficient (Hou et al., 2015). Excellent discussions about the usefulness of adaptive inventories in ecological studies are available in Brown and Manly (1998).

For clustered populations, systematic adaptive cluster sampling (SACS) provides an option that is robust and more efficient than SS. SACS is also design-based and the corresponding estimators are design-unbiased, as devised by Thompson (1991) for inventorying populations such as plants or animals characterized by aggregation tendencies due to flocking, dispersal patterns, and environmental patchiness. Forests, minerals, fossil fuels, and many other natural resource populations exhibit similar patterns. SACS can increase inferential precision still further, even after SS has been applied (Thompson, 2012, p. 350). However, SACS may cause a complete enumeration, meaning all population units are selected for a SACS sample. This oversampling results from a networking towards neighbor plots that is subject to in-situ measurements during a field campaign, suggesting that (1) oversampling would be particularly common for inventorying nonrare variables whose presence prevails in a population, making SACS a difficult choice for inventorying nonrare or regular forest attributes or ecological indicators; (2) the final SACS sample cannot be certain until the campaign is completed, making advance budgeting and logistics a challenge; (3) even for variables that are rare or scarce, uncertainty regarding the final sample still persists; and (4) a SACS sample may be variable-specific as its growth can be adapted to a particular attribute or indicator, thus risking sample imbalance or non-representativeness for other variables observed (Gattone and Di Battista, 2011; Turk and Borkowski, 2005; Yang et al., 2011, 2016).

Consequently, the objectives of this study are fourfold: (1) to generalize SACS (GSACS) with respect to the design and estimators where GSACS is design-based with design-unbiased estimators, extending the advantages and overcomes the previously noted limitations of SACS; (2) to demonstrate that SS is a special case of GSACS. This allows GSACS, with a reduced number of sample plots, to produce estimates that are identical to those obtained using SS, and thereby introduces a useful feature that retains the advantages of SS while being more cost-effective for clustered populations; (3) to characterize the effects of design strategies and spatial patterns of a forest attribute or ecological indicator on sampling distribution by comparing a comprehensive set of simulation scenarios; and (4) to illustrate that GSACS is 100% compatible with NFI systems constructed with SS, that no modifications to field protocols or designs are required, thereby underscoring the generalizability and prospects of GSACS to large scale inventories.

2. Theory: sampling designs and estimators

2.1. Systematic sampling (SS)

In an inventory, a spatial area of interest is tessellated using small

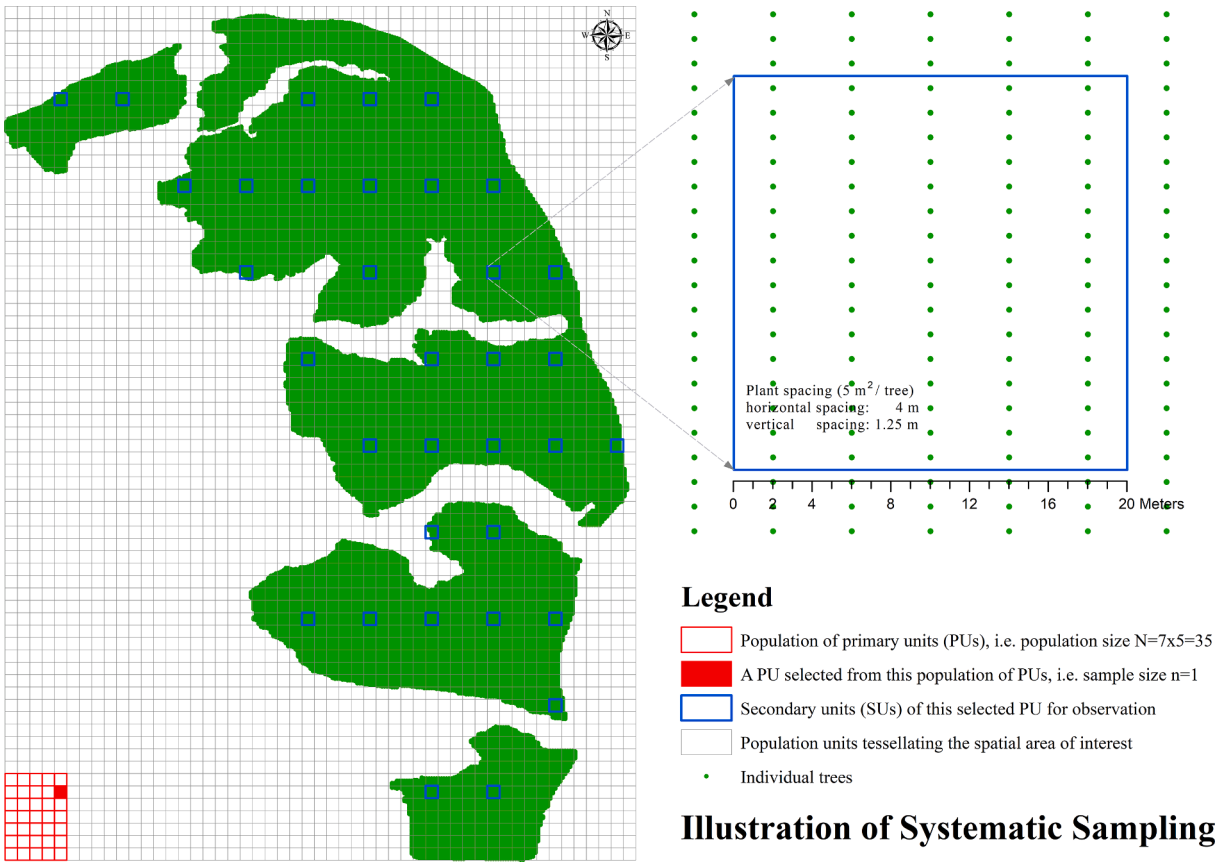

Fig. 1. Illustration of systematic sampling with a partition into primary units and secondary units. The systematic sample results from one randomly selected PU that consists of 34 SUs to be observed.

units of a given size that serve as population units. With systematic sampling, population of these units is partitioned into primary units (PUs) in a way that each PU comprises secondary units (SUs) spaced in a regular pattern over the spatial area. Whenever a PU is selected from the PUs, the values of every SU belonging to this PU are observed. Different choices on the size of population unit or on the partitioning lead to different designs that, through the sample, would affect the inferential precision about population parameters like the mean, μ, or total, τ, of the forest attribute or ecological indicator of interest.

In SS, a sample is selected from the PUs, even though the actual measurements are made on SUs. Because drawing a sample is with respect to the PUs, the population size, N, refers to the number of PUs, and the sample size, n, the number of selected PUs. The number of SUs, M, refers to the number of SUs in a PU, so the total number of population units is MN. When a PU is selected, its SUs will be taken as sample plots and then measured for that forest attribute. What is important in a systematic arrangement is that whenever a SU of a PU, denoted by u_{ij}, is included in the sample, the remaining SUs of this PU must be included as well. Fig. 1 illustrates a SS design deployed to a simulated forest compartment where the visualizations of respective notions are given.

In this study, we focus on μ per population unit since $\tau = MN\mu$. When a sample is selected from PUs by simple random sampling without replacement, the unbiased expansion estimator (Thompson, 2012, p. 343–344) is

$$\widehat{\mu}_1 = \frac{1}{Mn}\sum_{i=1}^{n}\sum_{j=1}^{M} y_{ij} = \frac{1}{Mn}\sum_{i=1}^{n} Y_i \tag{1}$$

where y_{ij} is the attribute or indicator value in the j^{th} SU of the i^{th} PU, u_{ij}; and $Y_i = \sum_{j=1}^{M} y_{ij}$ is the total of the y-values in the i^{th} PU.

The variance of $\widehat{\mu}_1$ is

$$Var\left(\widehat{\mu}_1\right) = \frac{N-n}{M^2Nn}\cdot\frac{\sum_{i=1}^{N}(Y_i - M\mu)^2}{N-1} \tag{2}$$

An unbiased estimator of this variance is therefore

$$\widehat{Var}\left(\widehat{\mu}_1\right) = \frac{N-n}{M^2Nn}\cdot\frac{\sum_{i=1}^{n}\left(Y_i - M\widehat{\mu}_1\right)^2}{n-1} \tag{3}$$

2.2. Systematic adaptive cluster sampling (SACS)

With systematic adaptive cluster sampling (SACS), the systematic sample of selected PUs in Section 2.1 was used as an initial sample to which additional population units are subsequently added from a clearly defined neighborhood and for a clearly defined condition of interest, C. In an inventory, the role of this C is to specify a domain for making inference; the role of the neighborhood is to transform the population of units to a population of networks (Thompson, 1991, 2012).

SACS pertains to sampling the population of networks rather than the population of units. A SACS sample takes form in a recursive manner such that each unit in the initial SS sample is a seed to be evaluated against C. When the C criterion is met, neighbors of this seed, typically those contiguous to the left, right, top and bottom, are then evaluated respectively. A valid neighbor will be enrolled with that seed in a network, and taken as a new seed for another iteration of this process with the aim of growing this network. This recursive process exhaustively searches and grows towards additional population units permissible under the neighborhood constraint, and automatically stops when all PU-intersected networks are included in a SACS sample. Fig. 2 illustrates a SACS sample under the typical neighborhood definition when applied to inventorying stem volume (m^3/ha), a nonrare or regular forest attribute. Thorough SACS design descriptions are available in

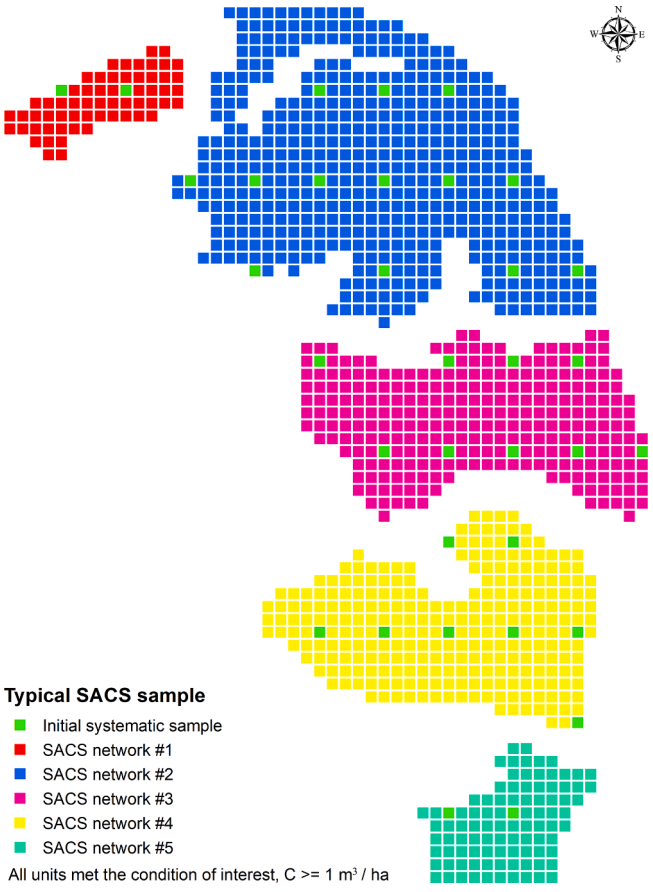

Fig. 2. Inventorying a nonrare forest attribute, e.g. stem volume (m^3/ha), with SACS caused a complete enumeration, a situation disastrous for forest ecosystem inventory.

Thompson (1991, 2012).

A SACS sample consists of networks instead of individual population units as is the case for SS. This distinction highlights the differences between estimators derived for SACS and SS, with the former taking network-level observations into account, and the latter unit-level observations into account. For SACS, an unbiased Hanse-Hurwitz estimator based on partial selection probability (Thompson, 2012, p. 345) is

$$\widehat{\mu}_2 = \frac{1}{Mn}\sum_{i=1}^{n}\sum_{k=1}^{K}\frac{y_k I_{ik}}{x_k} = \frac{1}{n}\sum_{i=1}^{n}\omega_i \tag{4}$$

where the notation M, N, n and PU are consistent with SS in Section 2.1; K is the number of networks; y_k is the total of the y-values in the k^{th} network; I_{ik} is an indicator variable, with $I_{ik} = 1$ if the i^{th} PU intersects the k^{th} network or $I_{ik} = 0$, otherwise; $x_k = \sum_{i=1}^{N} I_{ik}$ is the number of PUs in N intersecting the k^{th} network; and $\omega_i = \frac{1}{M}\sum_{k=1}^{K}\frac{y_k I_{ik}}{x_k}$ is a PU-specific estimate for μ.

The variance of $\widehat{\mu}_2$ is

$$Var\left(\widehat{\mu}_2\right) = \frac{N-n}{Nn}\cdot\frac{\sum_{i=1}^{N}(\omega_i-\mu)^2}{N-1} \tag{5}$$

An unbiased estimator of this variance is therefore

$$\widehat{Var}\left(\widehat{\mu}_2\right) = \frac{N-n}{Nn}\cdot\frac{\sum_{i=1}^{n}\left(\omega_i-\widehat{\mu}_2\right)^2}{n-1} \tag{6}$$

2.3. Generalized systematic adaptive cluster sampling (GSACS)

As Fig. 2 indicates, SACS may cause a complete enumeration, meaning all population units are network members of a SACS sample. This oversampling results from the recursive process that is dynamic relative to in-situ or on-the-spot measurements during a field campaign, leading to a series of complex effects noted in the Introduction. To resolve these challenges, we generalized SACS (GSACS) with respect to the design and estimators, prior to illustrating its connections with SS.

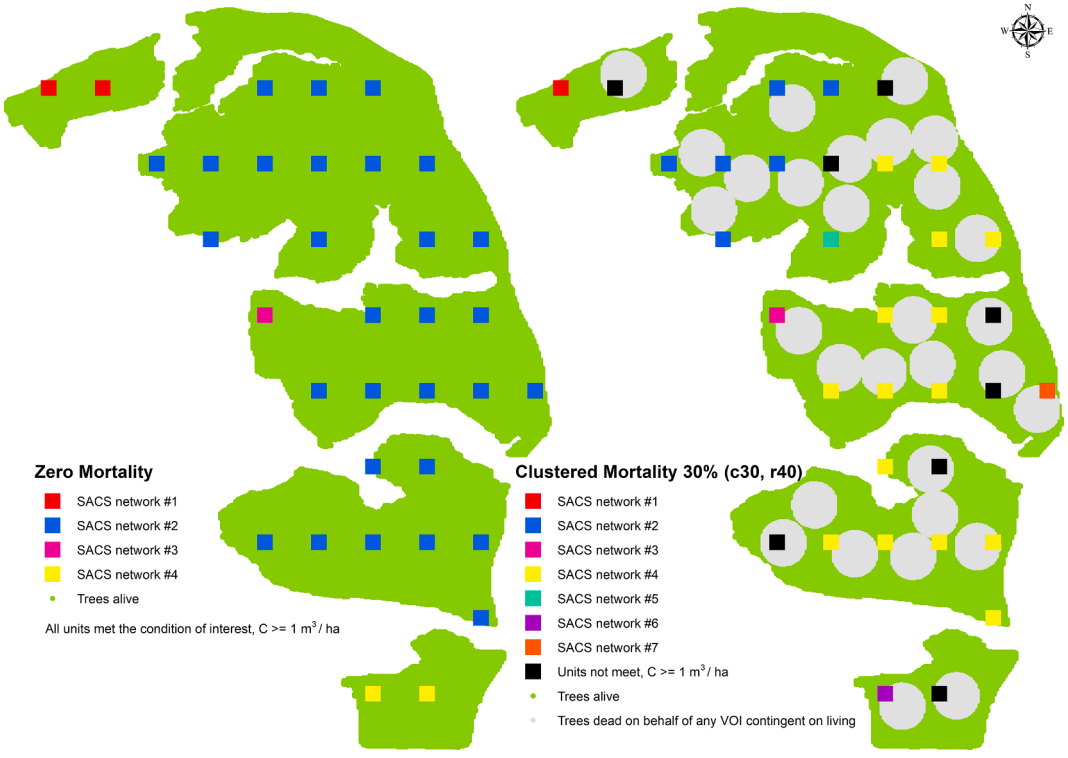

Fig. 3. the final form of GSACS samples in two simulated forest compartments.

Table 1
Summary of artificial populations generated for sampling simulations.

Population alias	μ(stem volume, m^3/ha)	Mortality (%)	No. of clusters	Cluster radius (m)	No. of Trees	No. of dead trees
Zero Mort	244.64	0	n.a.	n.a.	101,166	0
Random.10	220.07	10	n.a.	n.a.	101,166	10,006
Random.20	195.32	20	n.a.	n.a.	101,166	20,122
Random.30	171.44	30	n.a.	n.a.	101,166	30,307
Cluster.10.c161.r10	220.10	10	161	10	101,166	10,182
Cluster.10.c40.r20	220.30	10	40	20	101,166	10,061
Cluster.10.c18.r30	220.22	10	18	30	101,166	10,189
Cluster.10.c10.r40	220.12	10	10	40	101,166	10,062
Cluster.20.c322.r10	195.14	20	322	10	101,166	20,323
Cluster.20.c81.r20	194.96	20	81	20	101,166	20,397
Cluster.20.c36.r30	195.69	20	36	30	101,166	20,360
Cluster.20.c20.r40	195.72	20	20	40	101,166	20,129
Cluster.30.c483.r10	170.51	30	483	10	101,166	30,406
Cluster.30.c121.r20	170.89	30	121	20	101,166	30,457
Cluster.30.c54.r30	170.64	30	54	30	101,166	30,544
Cluster.30.c30.r40	171.50	30	30	40	101,166	30,194

In GSACS, the neighbor of a seed is spatially noncontiguous, constrained to be population units to the left, right, top and bottom following the spacing pattern identical to the initial SS sample. Spatial contiguity is not a requirement (Thompson, 2012, p. 341). Thereby, a network is always specific to a PU, because networks can only grow within the initial sample, achieving a transformation from the unit-level observations to the network-level observations. As a result, the final form of a GSACS sample becomes tethered and predictable and is similar or even identical to the initial SS sample. Fig. 3 illustrates the final form of GSACS samples and the formation of GSACS networks for two forest compartments.

While the SACS estimators in Section 2.2 remain valid, GSACS has a reduced form that is easier to use and analytically comparable with SS,

$$\widehat{\mu}_3 = \frac{1}{Mn}\sum_{i=1}^{n}\sum_{k=1}^{K} y_k I_{ik} = \frac{1}{n}\sum_{i=1}^{n}\varphi_i \tag{7}$$

where $\varphi_i = \frac{1}{M}\sum_{k=1}^{K} y_k I_{ik}$ is a PU-specific estimate for μ evaluated base on the y-values of respective networks intersected by the i^{th} PU. Note $x_k = 1$ in GSACS, because only one PU can intersect the k^{th} network, which is illustrated in Fig. 3.

The variance of $\widehat{\mu}_3$ is

$$Var\left(\widehat{\mu}_3\right) = \frac{N-n}{Nn}\cdot\frac{\sum_{i=1}^{N}(\varphi_i-\mu)^2}{N-1} \tag{8}$$

An unbiased estimator of this variance is therefore

$$\widehat{Var}\left(\widehat{\mu}_3\right) = \frac{N-n}{Nn}\cdot\frac{\sum_{i=1}^{n}\left(\varphi_i-\widehat{\mu}_3\right)^2}{n-1} \tag{9}$$

Interestingly, the SS estimators in Section 2.1 can be regarded as a special case of the GSACS estimators, i.e. $\widehat{\mu}_3 = \widehat{\mu}_1$, $Var\left(\widehat{\mu}_3\right) = Var\left(\widehat{\mu}_1\right)$, and $\widehat{Var}\left(\widehat{\mu}_3\right) = \widehat{Var}\left(\widehat{\mu}_1\right)$, demonstrated as follows:

$$\widehat{\mu}_1 = \frac{1}{Mn}\sum_{i=1}^{n} Y_i$$

$$= \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{M}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\varphi_i$$

$$= \widehat{\mu}_3; \text{ and}$$

$$Var\left(\widehat{\mu}_1\right) = \frac{N-n}{M^2Nn}\cdot\frac{\sum_{i=1}^{N}(Y_i-M\mu)^2}{N-1}$$

$$= \frac{N-n}{M^2Nn}\cdot\frac{\sum_{i=1}^{N}\left(M\left(\frac{Y_i}{M}-\mu\right)\right)^2}{N-1}$$

$$= \frac{N-n}{Nn}\cdot\frac{\sum_{i=1}^{N}\left(\frac{Y_i}{M}-\mu\right)^2}{N-1}$$

$$= \frac{N-n}{Nn}\cdot\frac{\sum_{i=1}^{N}(\varphi_i-\mu)^2}{N-1}$$

$$= Var\left(\widehat{\mu}_3\right)$$

Likewise, $\widehat{Var}\left(\widehat{\mu}_1\right) = \widehat{Var}\left(\widehat{\mu}_3\right)$. Note $\varphi_i = \frac{Y_i}{M}$ in the same domain of interest, which is illustrated in Fig. 3; both $\widehat{Var}\left(\widehat{\mu}_1\right)$ for SS (Thompson, 2012, p. 162) and $\widehat{Var}\left(\widehat{\mu}_3\right)$ for GSACS are not applicable when $n = 1$, namely when the sample size is only one. However, as the sample size increases, the variance estimators for GSACS and SS converge equally fast at the rate of $\frac{N-n}{Nn}$.

3. Simulation: validation and comparison

3.1. Sampling distribution

Because the procedures for selecting a sample and for producing estimates are clear, the validation, effectiveness, and behavior of GSACS versus SS were analyzed using sampling simulations. Although much of the attention was devoted to the simulation of the populations and sampling strategies specified in Sections 3.2 and 3.3, the basic idea of sampling simulation involves three steps. First, construct an artificial population that mimics as much as possible a real one. Because the parameters of a simulated population are readily known, estimators and sampling strategies can be assessed. Therefore, second, carry out a sampling strategy iteratively on this simulated population. Third, for each iteration, draw a sample following a prescribed design and then use it for estimating the population parameter.

The distribution of these estimates over respective iterations is the sampling distribution. The sampling distribution depends on the sampling design and estimation procedure but is Gaussian as per the central limit theorem and follows $N(E(\widehat{\mu}), Var(\widehat{\mu}))$. Incidentally, the difference

Fig. 4. Artificial populations generated for sampling simulations; living trees in green, dead trees in red; dead trees are surrogate for any forest attributes or ecological indicators relying on living status; mortality patterns represent the spatial distribution of any surrogated attributes or indicators. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

between a standard error and a standard deviation is that the former refers to the standard deviation of a sampling distribution, $SE = \sqrt{Var(\widehat{\mu})}$, and thus $\sqrt{\widehat{Var}(\widehat{\mu})}$ is an estimate for SE. Because GSACS, SACS and SS are design-unbiased, meaning $E(\widehat{\mu}) = \mu$ and hence $N(\mu, Var(\widehat{\mu}))$, the inferential uncertainty is explicitly associated with the inferential precision. For this study, this precision is evaluated using the coefficient of variation, $CV\% = \frac{SE}{\mu} \times 100$, for comparisons between estimators and among sampling strategies. The $CV\%$ is also termed as sampling error.

3.2. Populations

A total of 16 artificial populations was generated using real data collected from eucalyptus compartments located in Hainan province, China. These populations were generated at tree-level following Hou et al. (2015) in a way that the spatial pattern of a forest attribute or ecological indicator is either random or clustered. Tree locations were generated for these compartments with a density of 5 m^2 per tree (4 by 1.25 m), forming a systematic lattice of known coordinates. Summary statistics of these populations are detailed in Table 1 with visuals in Figs. 1 and 4.

In these populations, tree mortality was used as surrogate for any

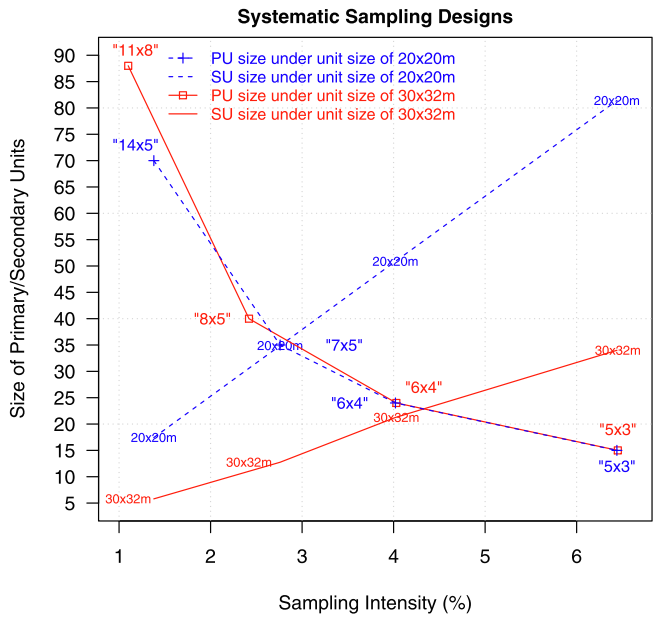

Fig. 5. Strategy for partitioning population units into primary and secondary units.

forest attributes or ecological indicators relying on living status, because an associated variable would be deprived as its carrier is deceased. For simplicity, variable values were zeros for dead trees. Spatial patterns were simulated for dead trees using the DNR Sampling Tools (DNR, 2020). The prevalence of a pattern is expressed using mortality (%), and the range of spatial autocorrelation using cluster diameter (meters). In a cluster, all trees are dead. In this context, investigating the spatial effects of a variable is equivalent to investigating the spatial effects of mortality.

The associated variable we chose was the stem volume (m^3/ha) because related attributes such as biomass, carbon and their changes that are important in forestry and ecology can be approximated using stem volume and expansion factors (Petersson et al., 2012). Stem volume was assigned for every tree using an allometric model based on class frequencies of compartment-wise diameter distribution (Hou et al., 2015). A dead tree was, of course, deprived of stem volume. With this comprehensive set of populations, we gain clear insight into the behavior of GSACS when used in practice.

3.3. Sampling designs

A sample was selected following SS and then networked following GSACS. Nevertheless, in practice, GSACS works independently and does not have to require a SS sample observed in the first place. There were two plot sizes, or equivalently, two population unit sizes considered and compared, the large unit was 30 by 32 m, and the small unit was 20 by 20 m, both rectangular. The number of plots in a sample is determined by the partitioning into PU as explained in Section 2.1, which was represented by sampling intensity. The sampling intensity is expressed as the ratio of the sampled area and the entire area. The sampled area refers to the product of the number of plots and the plot size. The partitioning employed in this study is graphed in Fig. 5 which stresses the comparability of the two plot sizes for similar sampling intensities, with other conditions being equal.

4. Results and discussion

4.1. Behavior of GSACS

For the respective sampling strategies, sampling distributions of GSACS and SS are summarized in Fig. 6. The sampling distribution is presented in the form of coefficient of variation, and the sampling strategy in the form of sampling intensity.

As expected, SS turns out to be a special case of GSACS, and the estimates for GSACS and SS are identical, i.e. $\widehat{\mu}_3 = \widehat{\mu}_1$, $Var\left(\widehat{\mu}_3\right) = Var\left(\widehat{\mu}_1\right)$ and $\widehat{Var}\left(\widehat{\mu}_3\right) = \widehat{Var}\left(\widehat{\mu}_1\right)$, consistent with the demonstration in Section 2.3. This identity is spatially or temporally invariant and is not conditional on population characteristics intrinsic to the spatial distribution of forest attributes or design characteristics driven by sampling strategies. In Fig. 6, the mortality is on behalf of any variable of forest attributes or ecological indicators relying on the living status. When a variable is randomly distributed, or precisely, follows a spatial Poisson point process (Daley and Vere-Jones, 2007), GSACS behaves exactly the same as SS as illustrated in Fig. 6 (A) by the overlaps between the dotted lines for GSACS and the solid lines for SS.

When a variable is spatially clustered, following a Matérn cluster point process (Matérn 1986), GSACS outperforms SS in terms of cost-efficiency, because GSACS produces estimates identical to SS while requiring fewer sample plots, as shown in Fig. 6 (B to D) in the form of a blank space between the dotted lines for GSACS and the solid lines for SS. The reduced plots are saved from those not meeting C. As the prevalence of clustering increases from 10% to 30%, or as the range of spatial autocorrelation increases from 10 m to 40 m in radius, this advantage becomes increasingly prominent for GSACS, suggesting that GSACS is more efficient than SS for sampling clustered populations Fig. 6 (B to D). In practice, although the spatial pattern may be unknown, GSACS applies where SS applies, and typically is more effective given the aggregation tendency common in nature.

What makes GSACS advantageous relative to SS is the difference in the inferential basis. GSACS pertains to selecting a sample of networks from a population of networks, whereas SS pertains to selecting a sample of units from a population of units. Units not meeting C are irrelevant to GSACS and are thus excluded from estimators (Thompson, 2012). Nevertheless, these units are always required by SS. This explains why using GSACS instead of SS would be more efficient for field campaigns, and attractive for construction of ecological observation networks where expensive mensuration equipment must be deployed and maintained at appointed locations (FIA, 2020).

Sample data may have a shelf-life for a timely reason, but GSACS also works for mining historic data through domain-specific or small-domain estimation. Domain estimation is realized through the networking procedure that transforms population units to networks within a sample. Because this networking is subject to C, the domain specifier, domain-specific estimates are thus obtainable by adjusting C. This convenient feature is somewhat analogous to quantile regression that exploits extra information from the same sample (Xu et al., 2019). However, GSACS is design-unbiased, whereas quantile regression is model-unbiased. Although SS would achieve a similar effect on estimation by manually choosing qualified units from the specified domain, GSACS is automated, computationally fast, and thus particularly useful for large scale analysis with national forest inventory databases as an example (FIA, 2020).

Although it may appear that a comparison between SACS and GSACS or even SACS and SS was omitted, in effect, this comparison was already made. This is because GSACS is the SACS under a particular neighborhood specification. When the neighborhood is spatially noncontiguous, constrained to be population units to the left, right, top and bottom following the spacing pattern identical to the initial SS sample, SACS exactly is GSACS with an identical performance, and thus the comparison between GSACS and SS is equivalent to the comparison between SACS and SS. Alternatively, when the neighborhood is spatially contiguous, constrained to be population units to the left, right, top and bottom, SACS is not GSACS, and SACS would introduce a complete enumeration as in Fig. 2, which leads straight to μ as in Table 1. Upon the

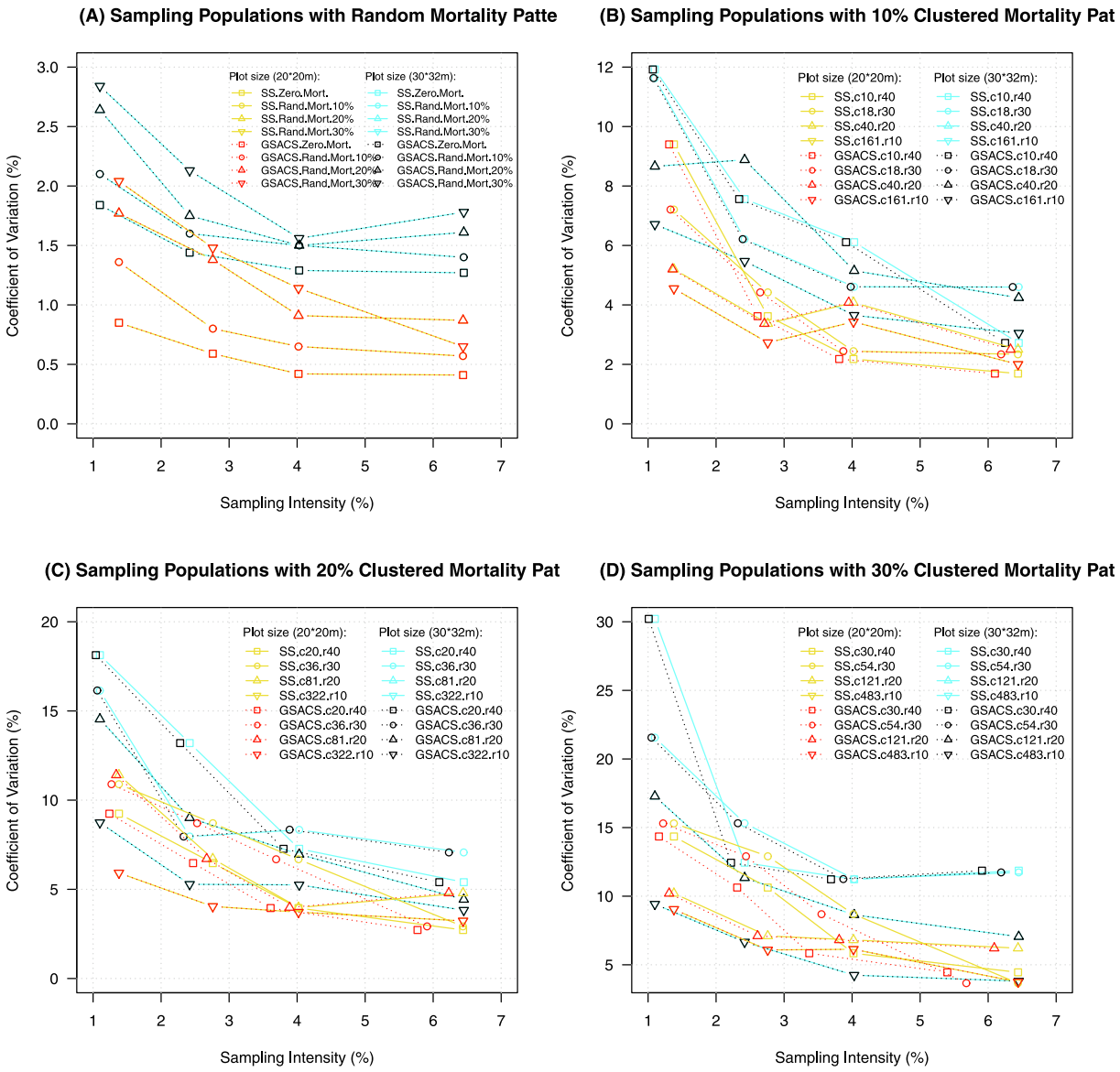

Fig. 6. Identical estimates for GSACS and SS, regardless of plot size or spatial patterns; the more prevalent a pattern or the greater the spatial autocorrelation, the fewer observations required for GSACS, making it more cost-effective for GSACS than SS.

complete enumeration or oversampling, SACS certainly outperforms GSACS, but would merit little consideration in practice, if any.

4.2. *Effects of plot size*

Reducing mean squared error, $MSE = Variance + Bias^2$, is a target when deriving estimators for sampling designs (Thompson, 2012). GSACS and SS estimators are design-unbiased, meaning $Bias = 0$ assured by their respective probabilistic designs. However, this variance term expressed by *CV*% turns out to be strongly affected by the population unit size, specifically the sample plot size, highlighting the effects of plot size on the gain or loss in inferential precision.

When a forest attribute or ecological indicator is spatially random (Fig. 6, A), *CV*% is consistently less for the small plots than the large plots. Regardless, the range of *CV*% for respective sampling intensities rooted in different sampling strategies is slightly wider for the small plots. This widening could be associated with the within-PU observations being slightly more variable for small-sized units because small plots outnumber large plots at a sampling intensity, contributing to the increase of within-PU variance or the decrease of within-PU correlation. However, as per Eq. (2) or (3) and Eq. (8) or (9), the variance estimators account for the between-PU variability rather than the within-PU variability (Thompson, 2012, p. 163).

With spatial clustering (Fig. 6, B-D), however, the *CV*% is consistently less, again, for the small plots than for the large plots. The range of *CV*% at respective sampling intensities becomes increasingly narrower for the small plots as the prevalence of clustering increases from 10% to 30%, and as the range of spatial correlation increases from 10 m to 40 m, suggesting that large plots are more vulnerable to the clustering effect and the between-PU variances are smaller for small plots.

The advantage of SS, which also applies to GSACS, is that it is usually more economic to observe a collection of SS units than to observe an equal number of units selected at random from a population. The effectiveness of SS and GSACS depends on the variance resulting from the PUs of a given size and shape as well as the cost of observing them. Theoretically, SS is efficient if the between-PU variance is small relative to the overall population variance (Thompson, 2012, p. 164); and once so, GSACS will be more efficient than SS (Fig. 6).

The ideal size and shape for the PUs can be determined by a variance function or a cost function, which are not necessarily simple in practice.

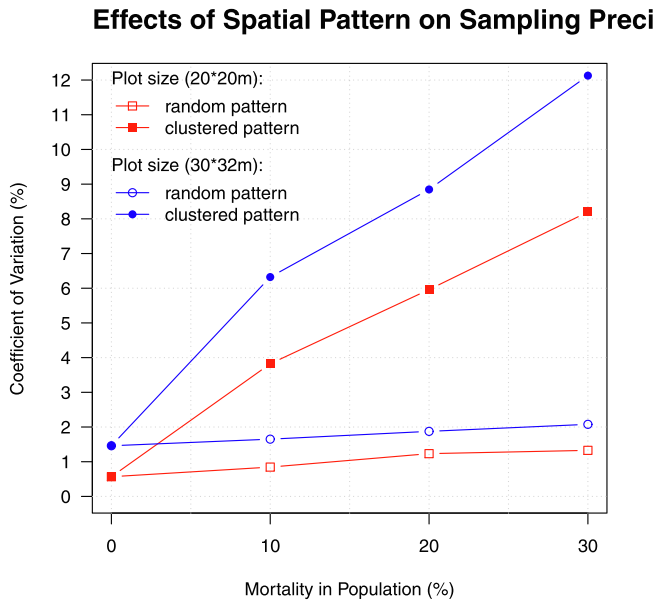

Fig. 7. Effects of spatial patterns on inferential precision.

Examples of these functions are available in Cochran (1977) and Jessen (1978). However, since the decision regarding plot size (or population unit size) is prior to the optimization for PU, it now becomes clear that small plots (or population units) outperform large ones in terms of inducing a smaller variance, thereby becoming a straightforward way for enhancing the effect of optimization. Hence, a sensible choice would be using small sample plots or population units for design-based inference, particularly for GSACS and SS.

4.3. *Effects of spatial patterns*

GSACS and SS estimators are design-unbiased, but not necessarily ideal for variance reduction. In addition to the plot size, the spatial distribution of a forest attribute or ecological indicator in the form of random or clustered patterns was found to affect the inferential precision more substantially than the plot size (Fig. 7).

For spatially random patterns, the effect on the variance in the form of *CV*% is relatively slight for both plot sizes (Fig. 7). This suggests that for homogeneous populations, measures such stratification could be ineffective for increasing inferential precision. For such a population, effective measures include, but are not limited to, increasing the sampling intensity by increasing the sample size or intensifying the sample, or for a permissible sampling intensity using smaller population units or sample plots. However, attention must be paid to a saturation point after which the decrease in*CV*% through increasing the sampling intensity becomes slow, which was 4% for this study (Fig. 6, A).

For spatially clustered patterns, the effect on *CV*% becomes increasingly stronger as the prevalence and autocorrelation increases (Fig. 7). The more prevalent a pattern, the larger the variance, and the more spatially autocorrelated, the larger the variance, with prevalence affecting variance more than autocorrelation (Fig. 6, B-D). Similar findings were also reported in Lessard et al. (2002). Measures such as stratification, increasing sample size or reducing plot size would be necessary for SS, but not necessarily for GSACS, because GSACS is relatively more effective for clustered populations as discussed in Section 4.1.

The SS variance can alternatively be examined in terms of the within-PU correlation coefficient, ρ (Thompson, 2012, p. 166). When $\rho = 0$, the inferential precision is about the same for SS and simple random sampling with an equal number of sample plots; when $\rho > 0$, the simple random sample produces greater inferential precision; and conversely when $\rho < 0$. With many natural populations, population units close to each other tend to be similar. With SS, the SUs of each PU are spaced relatively far apart, so $\rho < 0$ may well be the case. For this reason, SS is inherently efficient with many real populations; and once so, GSACS will be more efficient than SS (Fig. 6).

Last but not the least, the estimators for SS and GSACS can further be improved with a procedure derived from the Rao-Blackwell theorem (Blackwell, 1947). It will take a conditional expectation for the sampling distribution of an estimator given the minimal sufficient statistic observed (Cassel et al. 1977). Today, procedures of this kind are categorized as data assimilation (Hou et al., 2019, 2021).

5. Conclusions

Five conclusions are relevant. First, we generalized systematic adaptive cluster sampling (GSACS) and demonstrated that systematic sampling (SS) is a special case of GSACS. Second, in contrast to SACS, GSACS explicitly supports inventorying forest attributes and ecological indicators that are nonrare or common and resolves SACS problems with oversampling, uncertain sample form, and sample imbalance for alternative attributes or indicators. Third, empowered by networking, even with fewer sample plots, GSACS produces estimates identical to those for SS. Fourth, GSACS outperforms SS with respect to inventorying clustered populations and for making domain-specific estimates. Fifth, the precision for design-based inference is negatively correlated with the prevalence of a spatial pattern, the strength of spatial autocorrelation, and the size of sample plot, in descending order.

The equivalence between GSACS and SS would benefit forest ecology inventory for three aspects. First, with reduced observations, GSACS produces point and variance estimates identical to those for SS. The reduced observations are those spared from measuring population units not satisfying *C*, which must be measured in SS, anyway. This saving suggests reduced sample plots, reduced cost, reduced labor and increased cost-efficiency, factors making a difference for field campaigns. Second, like SS, GSACS also supports inventorying multiple attributes from a common sample. This is unlike SACS for which a sample is adapted to one specific attribute, thereby inducing sample imbalance for other attributes which, again, reiterates the generalizability of GSACS. Third, there is a wide potential for replacing SS with GSACS considering the popularity of SS, particularly for national forest inventories and international ecological observation networks. A distinguishing feature of GSACS resides in its compatibility with original inventory systems by not imposing any modifications to the field protocols, and in its efficiency for analyzing historic data for making domain-specific estimates that are crucial for change detection and monitoring.

6. Authors' contributions

QX and ZH conceived the study, derived sampling estimators, designed methodology and coded simulators; TT collected data; QX, GS, BL and ZH analyzed the data; QX, REM and ZH led the writing of the manuscript. All authors contributed critically to the article and gave final approval for publication.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 32001252) and the International Center for Bamboo and Rattan (Grant No. 1632020029). Dr. Todd S. Fredericksen and

anonymous reviewers are sincerely acknowledged for their excellent services and helpful comments.

References

Acharyal, B., Bhattarai, G., de Gier, A., Stein, A., 2000. Systematic adaptive cluster sampling for the assessment of rare tree species in Nepal. Forest Ecol. Manag. 137, 65–73.

Blackwell, D., 1947. Conditional expectation and unbiased sequential estimation. Ann. Math. Stat. 18, 105–110.

Brown, J.A., Manly, B.F.J., 1998. Restricted adaptive cluster sampling. Environ. Ecol. Stat. 5, 47–62.

Cassel, C.M., Särndal, C.E., Wretman, J.H., 1977. Foundations of Inference in Survey Sampling. Wiley, New York.

Cochran, W.G., 1977. Sampling techniques, third ed. John Wiley, New York.

Conn, P.B., Thorson, J.T., Johnson, D.S., 2017. Confronting preferential sampling and analyzing population distributions: diagnosis and model-based triage. Methods Ecol. Evol. 8, 1535–1546.

Daley, D.J., Vere-Jones, D., 2007. An Introduction to the Theory of Point Processes: Volume II: General Theory and Structure. Springer Science & Business Media.

DNR, 2020. DNR Sampling Tool. Minnesota Depart. of Natural Resources. Available online at http://files.dnr.state.mn.us/aboutdnr/bureaus/mis/gis/tools/arcview/extensions/sampling/dnr_sampling_tool.pdf.

FIA, 2020. Forest inventory and analysis: we are the nation's forest census. Available online at https://www.fia.fs.fed.us.

Gattone, S.A., Di Battista, T., 2011. Adaptive cluster sampling with a data driven stopping rule. Stat. Methods Appl. 20, 1–21.

Graubard, B.I., Korn, E.L., 2002. Inference for superpopulation parameters using sample surveys. Stat. Sci. 17, 73–96.

Gregoire, T.G., 1998. Design-based and model-based inference in survey sampling: appreciating the difference. Can. J. Forest Res. 28, 1429–1447.

Heikkinen, J., 2006. Assessment of uncertainty in spatially systematic sampling. Springer, Dordrecht, The Netherlands, pp. 155–176.

Hou, Z., Domke, G., Russell, M., Coulston, J., Nelson, M., Xu, Q., McRoberts, R.E., 2021. Updating annual state- and county-level forest inventory estimates with data assimilation and FIA data. Forest Ecol. Manag. 483, 118777.

Hou, Z., McRoberts, R.E., Ståhl, G., Packalen, P., Greenberg, J., Xu, Q., 2018. How much can natural resource inventory benefit from a finer resolution auxiliary data? Remote Sens. Environ. 209, 31–40.

Hou, Z., Mehtätalo, L., McRoberts, R.E., Ståhl, G., Tokola, T., Rana, P., Siipilehto, J., Xu, Q., 2019. Remote sensing-assisted data assimilation and simultaneous inference for forest inventory. Remote Sens. Environ. 234, 111431.

Hou, Z., Xu, Q., Hartikainen, S., Antilla, P., Packalen, T., Maltamo, M., Tokola, T., 2015. Impact of Plot Size and Spatial Pattern of Forest Attributes on Sampling Efficacy. Forest Sci. 61, 847–860.

Hou, Z., Xu, Q., McRoberts, R.E., Greenberg, J.A., Liu, J., Heiskanen, J., Pitkänen, S., Packalen, P., 2017. Effects of temporally external auxiliary data on model-based inference. Remote Sens. Environ. 198, 150–159.

IPCC, 2003. Forest lands. Intergovernmental Panel on Climate Change Guidelines for National Greenhouse Gas Inventories, Volume 4, Chapter 4. Institute for Global Environmental Strategies (IGES), Hayama, Japan, 2006. p. 4.48.

Jessen, R.J., 1978. Statistical Survey Techniques. Wiley, New York.

Kangas, A., Astrup, R., Breidenbach, J., Fridman, J., Gobakken, T., Korhonen, K.T., Maltamo, M., Nilsson, M., Nord-Larsen, T., Næsset, E., et al., 2018. Remote sensing and forest inventories in Nordic countries–roadmap for the future. Scand. J. For. Res. 33, 397–412.

Kangas, A., Maltamo, M., 2006. Forest Inventory: Methodology and Applications. Springer, the Netherlands.

Lessard, V.C., Drummer, T.D., Reed, D.D., 2002. Precision of density estimates from fixed-radius plots compared to N-tree distance sampling. Forest Sci. 48, 1–6.

Matérn, B., 1986. Spatial Variation, second ed. Springer-Verlag, Berlin.

Magnussen, S., McRoberts, R.E., Breidenbach, J., Nord-Larsen, T., Ståhl, G., Fehrmann, L., Schnell, S., 2020. Comparison of estimators of variance for forest inventories with systematic sampling – results from artificial populations. Forest Ecosyst. 7, 17.

Margules, C.R., Pressey, R.L., 2000. Systematic conservation planning. Nature 405, 243–253.

McRoberts, R.E., 2011. Satellite image-based maps: scientific inference or pretty pictures? Remote Sens. Environ. 115, 715–724.

McRoberts, R.E., Næsset, E., Gobakken, T., Chirici, G., Condés, S., Hou, Z., Saarela, S., Chen, Q., Ståhl, G., Walters, B.F., 2018. Assessing components of the model-based mean square error estimator for remote sensing assisted forest applications. Can. J. Forest Res. 48, 642–649.

Petersson, H., Holm, S., Stahl, G., Alger, D., Fridman, J., Lehtonen, A., Lundström, A., Mäkipää, R., 2012. Individual tree biomass equations or biomass expansion factors for assessment of carbon stock changes in living biomass—A comparative study. Forest Ecol. Manag. 270, 78–84.

Räty, M., Kuronen, M., Myllymäki, M., Kangas, A., Mäkisara, K., Heikkinen, J., 2020. Comparison of the local pivotal method and systematic sampling for national forest inventories. Forest Ecosyst. 7, 54.

Särndal, C.E., 1978. Design-based and model-based inference in survey sampling. Scand. J. Stat. 5, 27–52.

Sessa, R., Dolman, H., 2008. Terrestrial essential climate variables for climate change assessment, mitigation and adaption. Food and Agricultural Organization of the United Nations. Edition: GTOS 52.

Thompson, S.K., 1991. Adaptive cluster sampling: Designs with primary and secondary units. Biometrics 47, 1103–1115.

Thompson, S.K., 2012. Sampling, third ed. Wiley, New Jersey.

Tomppo, E., Gschwantner, T., Lawrence, M., McRoberts, R.E., 2010. National Forest Inventories: Pathways for Common Reporting. Springer, Dordrecht, The Netherlands.

Turk, P., Borkowski, J.J., 2005. A review of adaptive cluster sampling: 1990–2003. Environ. Ecol. Stat. 12, 55–94.

Williams, B.K., Brown, E.D., 2019. Sampling and analysis frameworks for inference in ecology. Methods Ecol. Evol. 10, 1832–1842.

Xu, Q., Li, B., Maltamo, M., Tokola, T., Hou, Z., 2019. Predicting tree diameter using allometry described by non-parametric locally estimated copulas from tree dimensions derived from airborne laser scanning. Forest Ecol. Manag. 434, 205–212.

Xu, Q., Man, A., Fredrickson, M., Hou, Z., Pitkänen, J., Wing, B., Ramirez, C., Li, B., Greenberg, J.A., 2018. Quantification of uncertainty in aboveground biomass estimates derived from small-footprint airborne LiDAR. Remote Sens. Environ. 216, 514–528.

Yang, H., Kleinn, C., Fehrmann, L., Tang, S., Magnussen, S., 2011. A new design for sampling with adaptive sample plots. Environ. Ecol. Stat. 18, 223–237.

Yang, H., Magnussen, S., Fehrmann, L., Mundhenk, P., Kleinn, C., 2016. Two neighborhood-free plot designs for adaptive sampling of forests. Environ. Ecol. Stat. 23, 279–299.

Zeng, W., Tomppo, E., Healey, S.P., Gadow, K.V., 2015. The national forest inventory in China: history – results – international context. Forest Ecosyst. 2, 23.