Forest Science, 2023, **XX**, 1–8
https://doi.org/10.1093/forsci/fxad022
Advance access publication 11 April 2023
Research Article

Biometrics

Risk-Averse Importance Sampling of Tree Attributes in High-Risk Forested Areas

Francis A. Roesch,[1*] Todd A. Schroeder,[2] and Joseph M. McCollum[2]

[1]USDA Forest Service, Southern Research Station, Forest Inventory and Analysis, 200 W.T. Weaver Blvd., Asheville, NC 28804, USA (Francis.A.Roesch@usda.gov).
[2]USDA Forest Service, Southern Research Station, Forest Inventory and Analysis, 4700 Old Kingston Pike, Knoxville, TN 37919-5206, USA (Todd.Schroeder@usda.gov); (Joseph.McCollum@usda.gov)

[*]Corresponding author email: Francis.A.Roesch@usda.gov

Abstract

This study develops the theory of risk-averse importance sampling and explains its potential application to forest inventory estimation through the use of a heuristic simulation. When the risk-producing elements of the landscape are known, a risk-averse sampling strategy can be created that results in fewer samples in high-risk areas. Our simulation shows that for certain high-risk populations, risk-averse importance sampling can be highly effective at reducing both risk to field crew members (requiring only 10% of the plot visits in the riskiest category) and sample variance relative to simple random sampling. The method is shown to be especially helpful when a population of values of interest decreases with increasing risk, with a reduction in mean square error (MSE) of 84% to 99% in these cases. The simulation also showed the opposite effect on MSE can be expected when values of interest increase with increasing risk. By increasing field crew safety, risk-averse importance sampling should also improve the frequency and accuracy of field observations, potentially leading to even bigger gains in estimate precision. We recommend risk-averse importance sampling any time hazardous conditions can result in a high number of missing observations and reasonably accurate characterizations of landscape risks can be developed.

Study Implications: During the collection of forest inventory data, the safety of field personnel is always of primary importance, but never has the safety of personnel been a component of the sample design. This study develops a risk-averse importance sampling strategy that provides a low-risk probability sample of field observations in high-risk areas for interested practitioners. The low-risk probability sample can be created when the risk-producing elements of the forested area are known and can be described in functional form. The method is shown to be especially helpful when a population of values of interest decrease with increasing risk. It is hypothesized that if a probability sample is defined that will reduce the sample in the riskiest areas and increase safety in the field, both response rates and the accuracy of field measurements will increase, and, in turn, will lead to a reduction in the variance of the final estimates. Risk-averse importance sampling is recommended any time it is likely to lead to a higher level of safety and observational success.

Keywords: forestry, sampling, non-response, risk, estimation, modelling

The land area of interest in extensive forest inventories often include sub-areas within which obtaining field samples by the usual protocols can have a higher risk than normal. Usually, these risks are known and obvious; however, they are rarely accounted for in a probability sample design and sometimes result in a field crew decision to not measure the sample plot. Of course, missing observations are highly problematic in a probability sample, which has led some investigators to the conclusion that a non-probability sampling method known as purposive sampling is more useful (Monga et al. 2022; Stringer et al. 2015). Purposive sampling is exactly as it sounds: population elements are selected for a sample "on purpose" because they meet some set of desirable qualities. In the case of natural resource inventories, those qualities would sometimes include a component or set of components that would indicate that a short travel distance or a low risk would be incurred while making the observation (Trettin et al. 2021).

Purposive sampling is a way to gain insight for a population when a probability sample cannot be obtained but its drawbacks are serious enough to advise against using it when a probability sample can be obtained. Naturally, when one is purposely selecting population elements that meet certain eligibility criteria, one is not sampling the portion of the population that does not meet the criteria. Therefore, unlike a probability sample, estimators based on a purposive sample cannot, in general, be shown to be unbiased for the population of interest, although they may, in some cases, provide unbiased estimates of identifiable subpopulations within the population of interest. Also, as in Tang et al. (2017) and Trettin et al. (2021), the use of ancillary data to inform a stratified sample, in conjunction with a purposive sample, can lead to reduced bias in the estimators. Given that purposive sampling does not provide a mechanism for ensuring the primary statistical goal of being able to generalize the

Received: August 17, 2022. Accepted: March 15, 2023.

results, we offer a probability sample design for high-risk areas.

In this article, we recognize that all forest inventory efforts contain some component of risk due to the physical demands of traversing to plot locations and the potential for encountering inaccessible terrain features and wild animals. We consider this a base level of risk not requiring a special sample design. We also note that some risk factors beyond the base level might occur along a definable gradient. Recognizing those gradients could assist us in defining a low-risk probability sample that will allow us to obtain as close to a full set of observations in high-risk areas as possible. Here, we describe an alternative or auxiliary sample design for field plot locations that has a risk-averse component in which the sample locations are selected with a probability proportional to the size of a risk-aversion function ($\ddot{r}$), which we will give an example of below. The method has its roots in probability proportional to size (*pps*) sampling (Cutkosky 1951; Hansen and Hurwitz 1949), and importance (*ppi*) sampling (Gregoire et al. 1986; Kahn 1954). It was Gregoire et al. (1986) that brought importance sampling to the attention of the forest sampling community. Subsequent work by those authors and a few colleagues helped to reinforce the method and its application. For the interested reader, an understanding of the usual justification for importance sampling as a variance-reduction technique and an inclination as to its wider applicability can be gained by a reading of Gregoire et al. (1986), on which the theory of this article is based. The major difference between previous work and this study is that here we simply seek an arguably less biased estimator in the presence of nonresponse while assuming that variance reduction will also be achieved through an increase in observational success. Our argument here is that by reducing the sample in the riskiest areas, we will increase field safety, which in turn will increase response rates and the accuracy of field measurements. The overall effect will likely be a reduction in the variance of the final estimates. We present our theory of risk aversion sampling and provide a simple heuristic simulation to explore the benefits and costs of application both with and without the presence of nonresponse.

Theory of Risk-Aversion Sampling

Without loss of generality, we assume that we can construct a well-defined map of probable higher-risk areas based on identifiable ecosystem attributes, such as certain forest types, proximity to coastal and riparian areas, and presence of unnavigable terrain features, etc. By well-defined, we mean that the map is segmented in its entirety and the areal segments have boundaries that are absolute rather than fuzzy and are therefore mutually exclusive. Usually, we would create an "access map" constituting a network of access areas, line segments and points, such as roads, navigable waterways, and aircraft landing or helicopter drop sites. The access map should exceed the boundaries of the areas of interest to the extent that the closest access to all area within the inventory area can be determined. We also assume that a risk function (r) can be developed based on the joint distribution of the attributes of the mapped areas and access map. We then define a probability of point selection that is proportional to a function ($\ddot{r}$) that is averse to that risk function. Although not evaluated here, the risk aversion function, $\ddot{r}$, could also be influenced by a combination of other factors, such as known crocodile nesting sites, sudden elevation changes, forest blowdowns, mining operations, ocean tides, etc. The key is to ensure that all area within the mapped population (or subpopulation) has an assigned $\ddot{r} > 0$, a positive probability of inclusion. We then select sample points with probability proportional to $\ddot{r}$.

We need a probability density function based on a variate d, $f(d)$, such that

$$\int_{D_{\min}}^{D_{\max}} f(d)\,\partial d = 1$$
$$f(d) > 0 \text{ for } D_{\min} \le d \le D_{\max}, \text{ and 0 otherwise, where } D_{\min} \text{ and } D_{\max} \text{ are the } \textit{minimum and maximum possible values for } d, \text{ respectively} \tag{1}$$

A value (V) of interest can be integrated with respect to the range of d:

$$V = \int_{D_{\min}}^{D_{\max}} v(d)\ \partial d \tag{2}$$

$$V = \int_{D_{\min}}^{D_{\max}} \frac{v(d)}{f(d)} f(d)\ \partial d. \tag{3}$$

The sample design is risk-averse when $f(d) = \ddot{r}$.

To facilitate the development, we also include the integration over the area of interest:

$$V = \int_{y_{\min}}^{y_{\max}} \int_{x_{\min}}^{x_{\max}} \int_{D_{\min}}^{D_{\max}} v(d)\ \partial d \partial x \partial y \tag{4}$$

To select a sample point, we consider two stages:

(1) Select the distance variate d ($d = d_s$) with probability density $f(d_s)$ to identify possibly many areal regions in the set S with attribute d_s.
(2) Define a bounding box for set S, with the limits of x-axis set at the minimum and maximum x values within S (x_{min} and x_{max}, respectively), and the limits of the y axis set at the minimum and maximum y values within S, (y_{min} and y_{max}, respectively). Then select a random point within the bounding box and accept it if the point is also within S, otherwise reject it. To select the random point within the bounding box:

   a. Select the x coordinate with uniform probability density $f(x_s) = 1/L_x$ where L_x equals x_{max}-x_{min}.
   b. Select the y coordinate with uniform probability density $f(y_s) = 1/L_y$ where L_y equals y_{max}-y_{min}.

If i ranges from 1 to the number of random sample points (m) and s_i has attribute d_i between D_{min} and D_{max}, then an unbiased estimator of V is

$$\hat{V} = \sum_{i=1}^{m} \frac{v(d_i)}{\ddot{r}_i} \tag{5}$$

$\hat{V}$ is a linear combination of random variables and therefore an estimator for the variance of $\hat{V}$ is

$$\hat{var}\left(\hat{V}\right) = \sum_{i=1}^{m} \hat{var}\left(\frac{v(d_i)}{\ddot{r}_i}\right) + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left[\hat{cov}\left(\left(\frac{v(d_i)}{\ddot{r}_i}\right), \left(\frac{v(d_j)}{\ddot{r}_j}\right)\right)\right] \tag{6}$$

Sampling High-Risk Coastal Mangrove Forests: a Motivating Example

The inventory of mangrove forests, which grow in small patches along the coastline of the southern United States and Caribbean islands, has associated with it a special set of well-defined risks, so we use them as starting point for our discussion of how to practically implement a risk-aversion importance sample. We give an example of a risk-aversion function in the context of estimation of attributes within mangrove forest types. To start, we assume that risk is related to mangrove community type and access distance. Within each mangrove type (or community type), there is a risk assumed to be a constant, in other words, saltwater mangroves have a higher probability of the presence of crocodiles, which are more aggressive than the alligators found more commonly in freshwater mangroves. We further assume that this additional risk will be addressed by additional precautions and possibly by adjustments to sample size and allowable sampling error. This simplification allows us to temporarily ignore this community-specific risk by sampling within each community type. Assume that we have merged a mangrove community type map with a map of access points including roads, boat landings, and navigable waterways. For each point (s_i) within the areas of interest, there is a distance $d(s_i)$ in meters to the nearest point of access as determined by an overlay of these two maps. The maximum distance considered in community type C is defined as

$$D_C = \max\left(d\left(s_i\right)\right),\ s_i \supset C \tag{7}$$

We would define the proportional risk in mangrove community type C at point s_i as

$$r_C\left(s_i\right) = d\left(s_i\right)/D_{\ C} \tag{8}$$

We define a risk aversion function that is equal to one minus the proportional risk; therefore, the value of this function at point i is

$$\ddot{r}_C\left(s_i\right) = 1 - r_C\left(s_i\right) \tag{9}$$

Although the population is one of an infinite set of points, we could use a finite approximation of the population to determine the sample locations by using a dense grid approximation. For simplicity, we discuss this discrete approach here, in which we overlay the merged map with a dense systematic grid using a random starting point for the grid. By dense grid, we mean one in which the number of candidate points is much larger than the desired sample size (say on the order of 1,000 times larger). One might use the pixel centers of a raster map for this purpose. For each grid point (s_i) within the areas of interest, we calculate the distance $d(s_i)$ in meters to the nearest point of access as determined by the access map.

We would then use equations 6 through 8 to obtain $\ddot{r}_C\left(s_i\right)$ for each point, resulting in empirical distributions for $d(s_i)$ and $\ddot{r}_C\left(s_i\right)$ for each community type. Risk aversion requires that we randomly select, with replacement, specific points with probability proportional to $\ddot{r}_C\left(s_i\right)$. This strategy ensures that the sample plot locations would be weighted toward the relative safety of the access network. In this example, from a practical perspective, field crew professionals would know of the rare instances when a sample location is a significant distance from an access point and would have the opportunity (and support) to take additional precautions while measuring the riskiest plots.

A Heuristic Simulation

We devised a simple heuristic simulation to demonstrate the advantages and potential pitfalls of enacting a risk-averse sampling design. We first constructed twelve populations that are intended to demonstrate the extremes of potential results from implementing a risk-averse design.

Populations

To construct the twelve populations, we first defined three value sets with ten thousand units each.

The sum of values in each value set is five thousand, whereas the sets differ in the structural diversity of the values. Value set 1 has ten thousand unique values. Value set 2 has two hundred copies of each of fifty unique values, grouped into five subsets with ten unique values per subset. Value set 3 also has two hundred copies of fifty unique values grouped into two subsets with twenty-five unique values per subset. A more detailed description of the value sets is given in the appendix. The sum of values in each value set is five thousand. The means and variances for the three value sets are given in Table 1. These statistics are constant for all four populations created from each value set.

We then created four populations from each value set. One population from each value set is the value set itself, with no association to a distance pattern, and three additional populations which are created by joining each value set with each of three distance patterns. Distance pattern 1 is randomly dispersed. Distance pattern 2 is increasing whereas distance pattern 3 is decreasing. This resulted in three unordered populations and nine distance-dependent populations. Three risk aversion distributions (**RA**1, **RA**2, and **RA**3) were derived, one from each distance pattern by subtracting each distance from 1. The specific details of the distance patterns and their complementary risk aversion distributions are given in the appendix.

Again, combining each value set with each of the distance patterns creates nine distance-dependent populations. Because the order of units in each value set is nondescending, a combination of any value set with distance pattern 1 creates a random population in one-dimensional space, the same dimension as distance. A combination of any value set with distance pattern 2 creates a population with values that are either increasing or nondecreasing with increasing distance, and a combination with distance pattern 3 creates a population that is either decreasing or nonincreasing with increasing distance. To help the reader visualize these populations, each of these nine populations is categorized into value and distance categories, formed using the algorithm: $x_{cat} = ((round\ up(x * 10))/10)$. The percent of the population in each combination of value and distance category is given in Table 2.

Table 1. The means and variances for the three value sets of 10,000 units each.

	Value set 1	Value set 2	Value set 3
Mean	0.5000	0.5000	0.5000
Variance	0.08334	0.10684	0.14584

Table 2. The percent of population elements by value and distance categories for the nine populations created by combining value sets 1 through 3 with distance patterns 1 through 3.

Value set		Distance pattern 1										Distance pattern 2										Distance pattern 3									
	Value	Distance category										Distance category										Distance category									
1	Category	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
	0.1	1.2	0.9	1.1	0.9	1.0	1.1	1.1	1.0	0.8	1.0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10
	0.2	1.0	0.9	1.0	1.0	1.1	1.0	1.0	0.8	1.1	1.1	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0
	0.3	1.1	0.9	1.0	0.9	1.0	1.0	1.1	1.0	1.0	1.0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0
	0.4	1.1	1.0	1.0	1.0	1.1	1.0	1.2	0.9	0.9	1.0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0
	0.5	1.0	0.9	1.0	0.9	1.0	1.0	1.2	1.1	0.9	0.9	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0
	0.6	1.0	1.0	0.8	1.0	1.2	1.0	1.1	1.0	1.0	0.9	0	0	0	0	0	10	0	0	0	0	0	0	0	0	10	0	0	0	0	0
	0.7	1.1	1.0	0.8	1.0	0.9	0.9	1.0	1.2	1.1	1.0	0	0	0	0	0	0	10	0	0	0	0	0	0	10	0	0	0	0	0	0
	0.8	1.1	0.9	0.9	1.1	1.0	1.0	1.1	0.9	1.0	1.1	0	0	0	0	0	0	0	10	0	0	0	0	10	0	0	0	0	0	0	0
	0.9	1.0	1.0	0.9	1.1	1.1	0.9	1.1	1.0	0.9	1.0	0	0	0	0	0	0	0	0	10	0	0	10	0	0	0	0	0	0	0	0
	1	1.1	1.0	1.2	1.0	1.0	0.9	1.1	1.0	0.9	0.8	0	0	0	0	0	0	0	0	0	10	10	0	0	0	0	0	0	0	0	0
2	Value	Distance category										Distance Category										Distance Category									
	category	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
	0.1	2.2	1.8	2.0	1.9	2.1	2.1	2.1	1.8	1.9	2.1	10	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	10
	0.3	2.2	1.9	2.0	1.9	2.0	1.9	2.3	1.9	1.8	2.0	0	0	10	10	0	0	0	0	0	0	0	0	0	0	0	0	10	10	0	0
	0.5	1.0	0.9	1.0	0.9	1.0	1.0	1.2	1.1	0.9	0.9	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0
	0.6	1.0	1.0	0.8	1.0	1.2	1.0	1.1	1.0	1.0	0.9	0	0	0	0	0	10	0	0	0	0	0	0	0	0	10	0	0	0	0	0
	0.8	2.2	1.9	1.6	2.0	1.9	1.9	2.0	2.1	2.2	2.1	0	0	0	0	0	0	10	10	0	0	0	0	10	10	0	0	0	0	0	0
	1	2.1	2.0	2.1	2.1	2.1	1.8	2.2	2.0	1.8	1.8	0	0	0	0	0	0	0	0	10	10	10	10	0	0	0	0	0	0	0	0
3	Value	Distance category										Distance Category										Distance Category									
	category	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
	0.1	2.2	1.8	2.0	1.9	2.1	2.1	2.1	1.8	1.9	2.1	10	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	10
	0.2	2.2	1.9	2.0	1.9	2.0	1.9	2.3	1.9	1.8	2.0	0	0	10	10	0	0	0	0	0	0	0	0	0	0	0	0	10	10	0	0
	0.3	1.0	0.9	1.0	0.9	1.0	1.0	1.2	1.1	0.9	0.9	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0
	0.8	1.0	1.0	0.8	1.0	1.2	1.0	1.1	1.0	1.0	0.9	0	0	0	0	0	10	0	0	0	0	0	0	0	0	10	0	0	0	0	0
	0.9	2.2	1.9	1.6	2.0	1.9	1.9	2.0	2.1	2.2	2.1	0	0	0	0	0	0	10	10	0	0	0	0	10	10	0	0	0	0	0	0
	1	2.1	2.0	2.1	2.1	2.1	1.8	2.2	2.0	1.8	1.8	0	0	0	0	0	0	0	0	10	10	10	10	0	0	0	0	0	0	0	0

Table 3. The sample selection outcomes following the first one thousand iterations of samples of size three hundred, for the nine populations created by combining value sets 1 through 3 with distance patterns 1 through 3. The values are in percent of importance sampling size by value and distance categories.

Value set		Distance pattern 1										Distance pattern 2										Distance pattern 3									
1	Value	Distance category										Distance category										Distance category									
	category	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
	0.1	2.0	1.6	1.6	1.3	1.0	0.9	0.7	0.5	0.3	0.1	19.0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.0
	0.2	1.6	1.9	1.3	1.4	1.0	1.0	0.7	0.4	0.4	0.1	0	16.9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3.0	0
	0.3	2.1	1.3	1.5	1.3	1.2	0.9	0.7	0.6	0.3	0.1	0	0	15.0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.0	0	0
	0.4	1.9	1.7	1.5	1.4	1.2	0.9	0.7	0.6	0.3	0.1	0	0	0	12.9	0	0	0	0	0	0	0	0	0	0	0	0	7.1	0	0	0
	0.5	1.6	1.5	1.5	1.3	1.1	1.1	0.7	0.5	0.4	0.1	0	0	0	0	11.1	0	0	0	0	0	0	0	0	0	0	9.0	0	0	0	0
	0.6	1.8	1.8	1.4	1.5	1.2	0.8	0.8	0.5	0.3	0.1	0	0	0	0	0	9.0	0	0	0	0	0	0	0	0	11.0	0	0	0	0	0
	0.7	1.6	2.1	1.7	1.2	1.1	0.8	0.8	0.4	0.3	0.1	0	0	0	0	0	0	7.1	0	0	0	0	0	0	13.0	0	0	0	0	0	0
	0.8	1.9	1.7	1.4	1.3	1.0	1.0	0.6	0.6	0.3	0.1	0	0	0	0	0	0	0	5.0	0	0	0	0	15.0	0	0	0	0	0	0	0
	0.9	1.9	1.6	1.6	1.3	1.1	0.8	0.8	0.5	0.4	0.1	0	0	0	0	0	0	0	0	3.0	0	0	17.1	0	0	0	0	0	0	0	0
	1	1.8	1.7	1.4	1.3	1.0	0.9	0.8	0.5	0.3	0.1	0	0	0	0	0	0	0	0	0	1.0	18.9	0	0	0	0	0	0	0	0	0
2	Value	Distance category										Distance category										Distance category									
	Category	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
	0.1	3.6	3.5	2.9	2.7	2.0	1.9	1.4	1.0	0.7	0.2	19.0	16.9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3.0	1.0
	0.3	4.0	3.0	3.0	2.7	2.3	1.7	1.5	1.2	0.6	0.2	0	0	15.0	12.9	0	0	0	0	0	0	0	0	0	0	0	0	7.1	5.0	0	0
	0.5	1.6	1.5	1.5	1.3	1.1	1.1	0.7	0.5	0.4	0.1	0	0	0	0	11.1	0	0	0	0	0	0	0	0	0	0	9.0	0	0	0	0
	0.6	1.8	1.8	1.4	1.5	1.2	0.8	0.8	0.5	0.3	0.1	0	0	0	0	0	9.0	0	0	0	0	0	0	0	0	11.0	0	0	0	0	0
	0.8	3.5	3.9	3.0	2.5	2.1	1.9	1.4	1.0	0.6	0.2	0	0	0	0	0	0	7.1	5.0	0	0	0	0	15.0	13.0	0	0	0	0	0	0
	1	3.7	3.3	2.9	2.6	2.2	1.7	1.5	1.0	0.7	0.2	0	0	0	0	0	0	0	0	3.0	1.0	18.9	17.1	0	0	0	0	0	0	0	0
3	Value	Distance category										Distance category										Distance category									
	Category	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
	0.1	3.6	3.5	2.9	2.7	2.0	1.9	1.4	1.0	0.7	0.2	19.0	16.9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3.0	1.0
	0.2	4.0	3.0	3.0	2.7	2.3	1.7	1.5	1.2	0.6	0.2	0	0	15.0	12.9	0	0	0	0	0	0	0	0	0	0	0	0	7.1	5.0	0	0
	0.3	1.6	1.5	1.5	1.3	1.1	1.1	0.7	0.5	0.4	0.1	0	0	0	0	11.1	0	0	0	0	0	0	0	0	0	0	9.0	0	0	0	0
	0.8	1.8	1.8	1.4	1.5	1.2	0.8	0.8	0.5	0.3	0.1	0	0	0	0	0	9.0	0	0	0	0	0	0	0	0	11.0	0	0	0	0	0
	0.9	3.5	3.9	3.0	2.5	2.1	1.9	1.4	1.0	0.6	0.2	0	0	0	0	0	0	7.1	5.0	0	0	0	0	15.0	13.0	0	0	0	0	0	0
	1	3.7	3.3	2.9	2.6	2.2	1.7	1.5	1.0	0.7	0.2	0	0	0	0	0	0	0	0	3.0	1.0	18.9	17.1	0	0	0	0	0	0	0	0

Table 4. RRMSE ((*sqrt(MSE)*)/*T*), and RBIAS (*Bias/T*) in parentheses, for random sampling of each unordered value set in the first column and importance sampling of the distance-dependent populations in the remaining columns, for estimators of the total ($T = 5000$), after 10,000 iterations of samples of size 300.

Value set	Random sample	Distance pattern 1	Distance pattern 2	Distance pattern 3
1	0.0334 (-0.0001)	0.1151 (-0.0020)	0.6984 (0.0056)	0.0000 (0.0000)
2	0.0378 (-0.0001)	0.1207 (-0.0020)	0.6963 (0.0056)	0.0078 (0.0000)
3	0.0445 (-0.0002)	0.1306 (-0.0020)	0.6971 (0.0057)	0.0178 (-0.0004)

Table 5. RRMSE ((*sqrt(MSE))/T*), and RBIAS (*Bias/T)* in parentheses, for the SRS estimators of the total ($T = 5000$) for each of the nine populations under the distance-dependent (threshold of .7) missing-value scenario after ten thousand iterations.

Value set	Distance pattern 1	Distance pattern 2	Distance pattern 3
1	0.0400 (0.0001)	0.3011 (-0.2298)	0.3011 (0.2298)
2	0.0454 (0.0001)	0.3370 (-0.3355)	0.3372 (0.3356)
3	0.0532 (0.0017)	0.3667 (-0.3641)	0.3666 (0.3641)

Table 6. RRMSE ((*sqrt(MSE))/T*), and RBIAS (*Bias/T*) in parentheses, for the risk-averse sample estimators of the total ($T = 5000$) for each of the twelve missing-value scenarios after 10,000 iterations. MAR, missing at random.

Value set	30% MAR	Distance pattern 1	Distance pattern 2	Distance pattern 3
1	0.0399 (-0.0002)	0.2412 (-0.2390)	0.4628 (-0.4617)	0.0000 (0.0000)
2	0.0452 (-0.0002)	0.2417 (-0.2390)	0.4904 (-0.4891)	0.0078 (0.0274)
3	0.0530 (-0.0003)	0.2412 (-0.2377)	0.5129 (-0.5109)	0.0178 (0.0491)

Sampling Simulation

We drew ten thousand random samples of size three hundred for each of the twelve populations, one random sample for each value set population and one risk-aversion importance sample as described above for each of the nine value set-distance pattern populations. Table 3 gives the relative importance sampling sizes in percent by value category and distance for each population, after the first one thousand samples. Finally, we show the results obtained first in the absence of nonresponse, and then in the presence of risk-influenced nonresponse.

Results

We calculated the simulation MSEs and empirical bias for the estimated total (T) for each scenario after the ten thousand sampling iterations. Because ten thousand estimates are obtained in each scenario (S) for each population total of $T = 5000$,

$$MSE_S = \frac{\sum_{i=1}^{10000}\left(\hat{T}_{S,i} - 5000\right)^2}{10000}, \text{ and} \tag{10}$$

$$Bias_S = \frac{\sum_{i=1}^{10000}\left(\hat{T}_{S,i} - 5000\right)}{10000} \tag{11}$$

The relative root *MSE* (RRMSE = (*sqrt(MSE))/T*) and relative bias (RBIAS = *Bias/T*) for each scenario, after the ten thousand sampling iterations, assuming that there are no missing values, are given in Table 4. To put the statistics on the same scale without losing the direction of bias and to make the tables directly comparable, we report RRMSE and RBIAS.

Next, we simulated the effects of missing observations for the nine distance-dependent populations assuming that any unit with an assigned distance (or risk) greater than 0.70 was not observed and simple random sampling (SRS) and the corresponding SRS estimators were used. RRMSE and RBIAS for each of these missing value scenarios are given in Tables 5.

To demonstrate the effects of nonrandom missing values on the importance sampling estimator, we present the results of one final simulation in Tables 6. In the first column of the table, we show the results for the three random samples of value sets 1 through 3, after a random reduction of 30% of the observations, that is ninety of the three hundred observations were missing at random. For the remaining nine populations, we assumed that any unit with an assigned distance (or risk) greater than 0.70 was not observed in the importance sampling simulation described above. The RRMSE and RBIAS for each of these missing value scenarios are given in Table 6.

Discussion

A comparison of Tables 2 and 3 demonstrates what can be expected in terms of sample outcome from the application of a risk-averse importance sample. Table 3 shows, for the nine distance-dependent populations after the first one thousand iterations, the empirical distributions (in percent) of the sample by value and distance categories. In all nine sub-tables in Table 3, the sample is heavily weighted toward shorter distances (or greater risk aversion), as opposed to the population distributions shown in Table 2, as designed. In Table 3, the upper-left sub-table shows the relatively even assignment of value category to each distance category, as we would expect from a random allocation of values through the range of distances. The top row, center sub-table in Table

3 shows a value category increasing with distance category (or decreasing with increasing risk aversion.) Finally, the upper-right sub-table shows a value category decreasing with increasing distance category (or decreasing with decreasing risk-aversion). The next two rows of sub-tables show the same effects, for the distance-dependent populations created from value sets 2 and 3, respectively. As we mentioned earlier, these populations were contrived for discussion purposes, not as representative of any sample of a known population of values. We would expect to see reasonable results from the samples shown in the left-hand column of sub-tables, extremely poor results from the samples shown in the center column of sub-tables, and extremely good results from the samples shown in the right-hand column of sub-tables. Additionally, if nonresponse increases with risk (or distance according to our model), we would expect an effective sample size–induced reduction of variance for the population estimates represented by the three sub-tables in the left-hand column, a large bias in the estimates from samples of populations represented by the center column of sub-tables, as high values are disproportionately lost from the sample, and very little bias from the samples of populations represented by the right-hand column. These expectations are borne out by the results presented in Tables 4 through 6.

The advantage of a high correlation between the importance sampling variate and a value of interest to the importance sampling estimator, when no values are missing, is evidenced by comparing the first column in Table 4 with the corresponding cells in the final column of that table. It bears some mention that a result of zero for either RBIAS or RRMSE is highly unlikely in practice but does occur in these tables due to rounding to four decimal places.

The disadvantage of a negative correlation is seen in Table 4 when comparing the results for distance pattern 2 with the results for the random sample and the other two distance patterns. Clearly, this column shows that risk-averse importance sampling is ill-advised for variables with values that increase with risk.

Table 5 shows the danger in making a missing at random assumption for simple random sampling, when values are not missing at random, in the results for distance patterns 2 and 3 compared with the results for distance pattern 1. For distance pattern 1, the RBIAS in Table 5 is extremely small. Additionally, the comparison of column 1 in Table 5 with the corresponding cells in column 1 of Table 4 clearly shows the effect of missing values when missing values occur at random, because distance is randomized in distance pattern 1. That effect being mostly an increase in variance, due to fewer observations being made.

Comparing the final column of Table 4 with the corresponding cells in Table 5 shows a large advantage to the importance sampling approach when the value of interest is highly correlated with risk aversion (or inversely correlated with distance). The opposite is true for distance pattern 2.

Again, in the final column of Table 6, as in the final column of Table 4, we see the advantage of a high correlation between the importance sampling variate and a value of interest to the importance sampling estimator as evidenced by comparing that column with the corresponding cells in the first column. The disadvantage of a negative correlation is also seen in Table 6 when comparing the results for distance pattern 2 with the results in the first column for the missing at random sample.

Conclusions

We have developed the theory of risk-averse importance sampling and explained its application through the use of a heuristic simulation. The simulation has shown that risk-averse importance sampling can be highly effective and variance reducing when judiciously used in high-risk areas. The method is especially helpful when a population of values of interest decrease with increasing risk. It has also shown that it can be variance increasing when used for populations of values that increase with risk. These properties of importance sampling are known to be true for any proxy function used as a sample selection mechanism in importance sampling.

The results in this article are based on comparison of known levels of sampling success. The major justification for this work lies in the argument that if we define a probability sample that will reduce the sample in the riskiest areas and increase safety in the field, we will increase response rates and the accuracy of the field measurements, which, in turn will lead to a reduction in the variance of the final estimates. Although risk-averse importance sampling will always reduce the risk of sample observation and is recommended any time it is likely to lead to a higher level of observational success, only application in specific situations will determine whether it actually does lead to higher response rates. Further research will be needed to determine whether risk-averse importance sampling can be used to improve estimation of characteristics in specific high-risk forested environments, such as on steep slopes, in storm-damaged areas, and in coastal mangroves.

Acknowledgments

An anonymous reviewer, an Associate Editor, and the Editor provided valuable comments that led to the improvement of this article.

Funding

This research was funded by the USDA Forest Service, Southern Research Station, Forest Inventory and Analysis Program.

Conflict of Interest

None declared.

Appendix: Constructing the Populations

To construct the twelve populations, we first defined three sets of values with ten thousand units in each value set. In each of the three value sets, the units are assigned values as described below. The sum of values in each value set is 5000, while the sets differ in the structural diversity of the values. The value sets are defined as:

Value set 1: Ten thousand unique values ranging from (and uniformly dispersed over) 0.00005 to 0.99995. The first value is.00005 and each successive value is incremented by.0001.

Value set 2: Five subsets in clusters of two hundred for each of fifty unique values, with ten unique values per subset. The subsets are defined as:

Subset 2a: The first of ten unique values is 0.005 and each successive unique value is incremented by 0.01.
Subset 2b: The next of ten unique values is 0.205 and each successive unique value is incremented by.01.
Subset 2c: The next of ten unique values is 0.455 and each successive unique value is incremented by 0.01.
Subset 2d: The next of ten unique values is 0.705 and each successive unique value is incremented by 0.01.
Subset 2e: The next of ten unique values is 0.905 and each successive unique value is incremented by 0.01.

Value set 3: Two subsets in clusters of two hundred for each of fifty unique values, with twenty-five unique values per subset. The subsets are defined as:

Subset 3a: The first of the twenty-five unique values is 0.005 and each successive unique value is incremented by 0.01.
Subset 3b: The first of the next twenty-five unique values is 0.755 and each successive unique value is incremented by 0.01.

Next, we defined three distance patterns, each consisting of a vector of length 10,000. distance patterns 1 through 3 were defined as:

Distance pattern 1 (**D**1): The distance assigned to each unit is a random draw from a uniform (0, 0.9999) distribution, creating a randomly dispersed population in the one-dimensional space of distance.
Distance pattern 2 (**D**2): Ten thousand unique distances ranging from 0.00005 to 0.99995. The first distance is 0.00005 and each successive distance is incremented by 0.0001.
Distance pattern 3 (**D**3): Ten thousand unique distances ranging from 0.99995 to 0.00005. The first distance is.99995 and each successive distance is decremented by 0.0001.

We then created four populations from each value set. One population from each value set is the value set itself, with no association to a distance pattern, and three additional populations are created by joining the value set with each of three distance patterns. That is, we applied each of these three distance patterns to each of the three value sets to create nine distance-dependent populations.

Three risk aversion patterns are derived, one from each distance pattern (or vector). Each distance vector of length 10,000 is subtracted from a vector (**1**) of 10,000 ones, such that

RA1 = **1**-**D**1, **RA**2 = **1**-**D**2, and **RA**3 = **1**-**D**3.

Literature Cited

Cutkosky, R.E. 1951. “A Monte Carlo Method for Solving a Class 6f Integral Equations.” *Journal of Research of the National Bureau of Standards* 47(2), August 1951. Research Paper 2234. 113.

Gregoire, T.G., H.T. Valentine, and G.M. Furnival. 1986. “Estimation of Bole Volume by Importance Sampling.” *Canadian Journal of Forest Research* 16 (3): 554–557.

Hansen, M.H., and W.N. Hurwitz. 1949. “On the Determination of Optimum Probabilities in Sampling.” *The Annals of Mathematical Statistics* 20 (3): 426–432.

Kahn, H. 1954. “*Applications of Monte Carlo*.” No. AECU-3259; RM-1237-AEC. Santa Monica, CA: RAND Corp.

Monga, E., M.M. Mangora, and C.C. Trettin. 2022. “Impact of Mangrove Planting on Forest Biomass Carbon and Other Structural Attributes in the Rufiji Delta, Tanzania.” *Global Ecology and Conservation* 35: e02100. doi: 10.1016/j.gecco.2022.e02100.

Stringer, C.E., C.C. Trettin, S.J. Zarnoch, and W. Tang. 2015. “Carbon Stocks of Mangroves within the Zambezi River Delta.” *Forest Ecology and Management* 354 (2015): 139–148. doi: 10.1016/j.foreco.2015.06.027.

Tang, W., W. Feng, M. Jia, J. Shi, H. Zuo, C.E. Stringer, and C.C. Trettin. 2017. “A Cyber-enabled Spatial Decision Support System to Inventory Mangroves in Mozambique: Coupling Scientific Workflows and Cloud Computing.” *International Journal of Geographical Information Science* 31 (5): 907–938. doi:10.1080/13658816.2016.1245419.

Trettin, C.C., Z. Dai, W. Tang, D. Lagomasino, N. Thomas, S.K. Lee, M. Simard, *et al*. 2021. “Mangrove Carbon Stocks in Pongara National Park, Gabon.” *Estuarine, Coastal and Shelf Science* 259 (2021): 107432. doi:10.1016/j.ecss.2021.107432.