

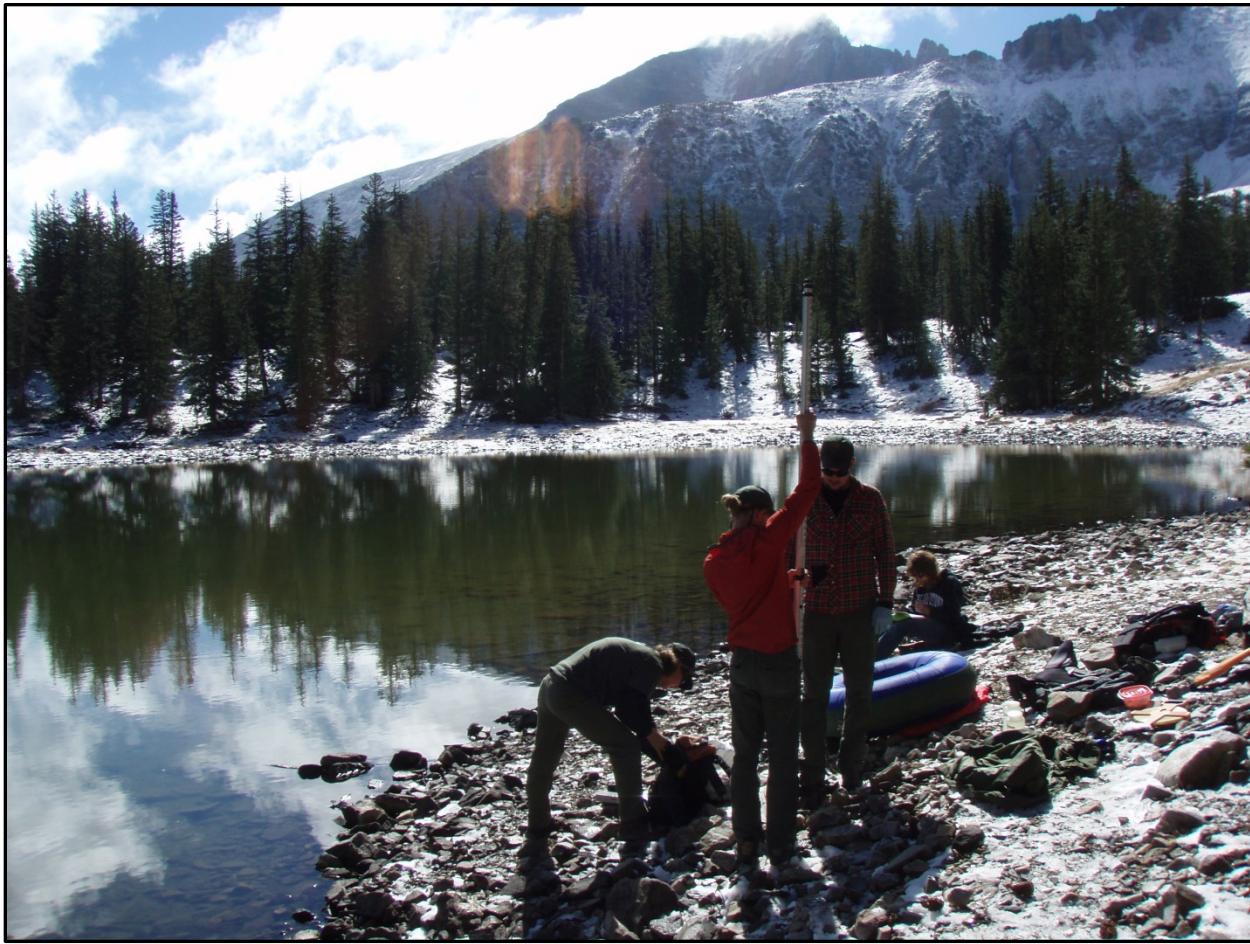


A GRTS User's Manual for the SDrawNPS Package

A graphical user interface for Generalized Random Tessellation Stratified (GRTS) sampling and estimation

Natural Resource Report NPS/PWRO/NRR—2016/1233





ON THIS PAGE

Photograph of observers sampling for the Mojave Desert Network Lake Monitoring protocol
Photograph courtesy of NPS/MOJN I&M

ON THE COVER

Photograph of observers sampling for Invasive Species in the Mojave Desert Network Vegetation Monitoring protocol
Photograph courtesy of NPS/MOJN I&M

A GRTS User's Manual for the SDrawNPS Package

*A graphical user interface for Generalized Random
Tessellation Stratified (GRTS) sampling and estimation*

Natural Resource Report NPS/PWRO/NRR—2016/1233

Leigh Ann Starcevich¹, Guy DiDonato¹, Trent McDonald¹, Jason Mitchell¹

¹Western EcoSystems Technology, Inc.
456 SW Monroe Ave. Suite 106
Corvallis, OR 97330



Prepared for: Nita Tallent² and Stacey Ostermann-Kelm³

²Mojave Desert Network
National Park Service
601 Nevada Way
Boulder City, NV 89005

³Mediterranean Coast Network
National Park Service
401 West Hillcrest Drive
Thousand Oaks, CA 91360

June 2016

U.S. Department of the Interior
National Park Service
Natural Resource Stewardship and Science
Fort Collins, Colorado

The National Park Service, Natural Resource Stewardship and Science office in Fort Collins, Colorado, publishes a range of reports that address natural resource topics. These reports are of interest and applicability to a broad audience in the National Park Service and others in natural resource management, including scientists, conservation and environmental constituencies, and the public.

The Natural Resource Report Series is used to disseminate comprehensive information and analysis about natural resources and related topics concerning lands managed by the National Park Service. The series supports the advancement of science, informed decision-making, and the achievement of the National Park Service mission. The series also provides a forum for presenting more lengthy results that may not be accepted by publications with page limitations.

All manuscripts in the series receive the appropriate level of peer review to ensure that the information is scientifically credible, technically accurate, appropriately written for the intended audience, and designed and published in a professional manner.

This report received formal peer review by subject-matter experts who were not directly involved in the collection, analysis, or reporting of the data, and whose background and expertise put them on par technically and scientifically with the authors of the information.

Views, statements, findings, conclusions, recommendations, and data in this report do not necessarily reflect views and policies of the National Park Service, U.S. Department of the Interior. Mention of trade names or commercial products does not constitute endorsement or recommendation for use by the U.S. Government.

This report is available in digital format from Western EcoSystems Technology, Inc. (WEST) website (<http://west-inc.com>) and the Natural Resource Publications Management website (<http://www.nature.nps.gov/publications/nrpm/>). To receive this report in a format optimized for screen readers, please email irma@nps.gov.

Please cite this publication as:

L. A. Starcevich, G. DiDonato, T. McDonald, and J. Mitchell. 2016. A GRTS user's manual for the SDrawNPS package: A graphical user interface for generalized random tessellation stratified (GRTS) sampling and estimation. Natural Resource Report NPS/PWRO/NRR—2016/1233. National Park Service, Fort Collins, Colorado.

Contents

	Page
Figures.....	v
Photographs and Tables	vii
Executive Summary	ix
Acknowledgments.....	xi
List of Terms.....	xiii
Introduction.....	1
Sampling Design Background	3
General Sampling Design Steps	3
Membership Designs	5
Equal Probability Sampling.....	6
Stratified Sampling.....	6
Unequal Probability Sampling	7
Continuous Probability Sampling.....	7
Two-Stage Sampling	8
Master Sample	8
Nonsampling Error	8
Frame Error	9
Nonresponse Error.....	9
Measurement Error.....	10
Sample Size Approximation.....	11
GRTS Sampling	13
GRTS Algorithm	13
Neighborhood Variance Estimator	14
Flexibility of GRTS Samples	15
Sampling Frames for GRTS Samples.....	15
Case Studies.....	16
MOJN Integrated Upland Monitoring	16
MEDN Stream Monitoring on Santa Rosa Island	16

Contents (continued)

	Page
Installing the SDrawNPS Package.....	17
Drawing a GRTS Sample.....	21
The grts() Function in <i>spsurvey</i>	21
Equal Probability Sampling.....	22
Case Study: Polygon Frame	22
Stratified Sampling.....	27
Case Study: Polygon Frame, User Allocation.....	27
Unequal Probability Sampling	32
Case Study: Finite Frame, Sampling Proportional to a categorical attribute	32
Continuous Probability Sampling.....	37
Case Study: Linear Frame, Sampling Proportional to a continuous attribute	37
Estimation	43
Formatting the GRTS Design File.....	43
Adjusting Design Weights.....	44
Case Study: Data Analysis without Weighting Adjustment.....	46
Case Study: Data Analysis with Weighting Adjustment.....	54
Discussion.....	63
References.....	65
Appendix A: Weight Adjustment for Nonsampling Error.....	1

Figures

	Page
Figure 1. Quadrant-recursive partitioning of a unit square (Stevens and Olsen 2004).....	14
Figure 2. The R command window after the SDrawNPS package is loaded.....	23
Figure 3. The SDrawNPS GUI for an equiprobable GRTS sample of 150 sites drawn from the SRI watersheds polygon sampling frame.....	24
Figure 4. An example equiprobable GRTS sample of 150 sites drawn from the SRI watersheds polygon sampling frame.....	25
Figure 5. The SDrawNPS GUI for a stratified GRTS sample.	27
Figure 6. The SDrawNPS GUI for a stratified GRTS sample from the SRI watersheds shapefile.	28
Figure 7. The SDrawNPS GUI for an stratified GRTS sample from the SRI watersheds shapefile, specifying 120 sites from the Large stratum and 30 sites from the Small stratum.....	29
Figure 8. The SRI watersheds stratified GRTS sample of 120 Large-stratum sites and 30 Small-stratum sites.....	30
Figure 9. Partial screenshot of the SRI watersheds stratified GRTS sample of 120 Large-stratum sites and 30 Small-stratum sites.	32
Figure 10. The SDrawNPS GUI for drawing an unequal probability GRTS sample.	33
Figure 11. The SDrawNPS GUI and necessary information for drawing an unequal probability GRTS sample of vegetated areas in JOTR.	34
Figure 12. The unequal probability GRTS sample across four vegetation associations within JOTR.....	35
Figure 13. Output table showing the sample draw from the unequal probability draw from JOTR vegetation associations.	36
Figure 14. The SDrawNPS GUI for drawing an unequal probability GRTS sample.	38
Figure 15. The SDrawNPS GUI and necessary information for drawing an unequal probability GRTS sample of the streams on Santa Rosa Island using inverse stream length as the weighting variable.....	39
Figure 16. The unequal probability GRTS sample of 30 stream locations on Santa Rosa Island.....	40
Figure 17. Partial table showing the output sample draw from the unequal probability draw of Santa Rosa Island stream segments.	40
Figure 18. Example of a GRTS design file with updated EvalStatus and EvalReason fields and columns added for data collected at each site.	44
Figure 19. The GUI interface used to analyze data collected as part of a GRTS design.	48

Figures (continued)

	Page
Figure 20. The GUI interface showing the browsed data file (WQdata.csv) and the data fields and variable classes read using the “Inspect Sample” button at the bottom left.....	49
Figure 21. The completed analysis GUI to analyze the water quality case study data set.....	50
Figure 22. Screenshot of the files saved following the analysis of the water quality data set.....	51
Figure 23. The cumulative distribution function (CDF) resulting from the analysis of the water quality case study data.....	52
Figure 24. Tabulated results saved in the WQdata_CDF.csv file.....	52
Figure 25. Tabulated results saved in the WQdata_Pct.csv file.....	53
Figure 26. Analysis GUI showing data fields and classes for the vegetation data survey.....	55
Figure 27. The completed GUI to analyze the Mean Canopy Gap Size (m) from the vegetation data set.....	56
Figure 28. Screenshot of the files saved following the analysis of the vegetation data set and output files to the directory containing the data set	57
Figure 29. Accessing the analysis results of the vegetation data set in the R command window.....	57
Figure 30. Partial screenshot of tabulated results saved in the VegAnalysis_Pct.csv file.	58
Figure 31. The first four panels of the cumulative distribution function (CDF) resulting from the analysis of the mean canopy gap size (m) in the vegetation data set	58
Figure 32. Analysis GUI for the analysis of mean canopy gap size with inference to the sampled population.	60
Figure 33. Partial screenshot of tabulated results saved in the Veg_Sampled_Pct.csv file.	60
Figure 34. The completed GUI to analyze both Mean Canopy Gap Size and Mean Basal Gap Size from the vegetation data set.....	61
Figure 35. Analysis GUI for the estimation of frame extent for the target population after accounting for frame error.	62
Figure 36. Partial screenshot of tabulated results saved in the VegFrameExtent.csv file for the analysis of frame extent after accounting for frame error.	62

Photographs and Tables

	Page
Photo 1. MOJN Desert Spring Monitoring.....	ix
Table 1. Post-survey coding for EvalStatus field in GRTS design file.....	43

Executive Summary

Spatially-balanced sampling provides a probabilistic basis for allocating sampling effort across large areas. Benefits of spatially-balanced sampling include avoidance of clustered points that can be obtained through simple random sampling and increased efficiency when the variance estimator can take advantage of the spatial balance.

Many networks of the National Park Service (NPS) have incorporated spatially-balanced sampling into their monitoring programs of Vital Signs at each park. Specifically, many parks have chosen to use generalized random tessellation stratified (GRTS) sampling to obtain spatially-balanced samples of points across these extensive and heterogeneous landscapes. GRTS sample draw and analysis tools available in the R package *spsurvey* have been implemented in a graphical-user interface (GUI) in the *SDrawNPS* package. The tools available in the *SDrawNPS* package include:

- GUI windows for drawing an equiprobable, stratified, or unequal probability GRTS sample
- Options to plot the sampling frame and selected sample points
- GUI windows to analyze data from a GRTS sample
- Options to adjust the design weights for frame error and nonresponse error
- Options to change the scope of inference based on nonresponse error

This manual serves as a companion user's manual for the *SDrawNPS* package and provides statistical background on GRTS sampling, nonsampling error, weighting adjustment, and other details that arise when designing, implementing, and analyzing data from an ecological sample. Example data sets used in the examples are provided with the *SDrawNPS* package.



Photo 1. MOJN Desert Spring Monitoring.

Acknowledgments

The authors would like to thank N. Tallent and S. Ostermann-Kelm for initiating this project which is funded under contract number GS10F-072BA. We would also like to thank D. Gundlach, M. Lehman, and R. Mazor who reviewed the early drafts and supplied the sampling frames and pilot data for examples. We appreciate the thorough reviews of the User's Manual and the SDrawNPS package provided by E. Dinger, B. Hamilton, M. Henk, G. Moret, and T. Phillippi. Thanks also to J. Brackin for providing photos from the Mojave Desert Network I&M program. Finally, we thank the developers of the GRTS sampling design and the *spsurvey* R package, T. Olsen D. Stevens, and T. Kincaid, for many helpful conversations over the years on GRTS sampling and the *spsurvey* package.

List of Terms

Census: A complete survey of a population in which data are completely collected from all sampling units.

Cumulative distribution function (CDF): A function of a random variable that expresses the probability that the outcome of interest takes the given value or less. The CDF ranges from 0 to 1.

Design-based inference: Inference based on the randomization generated by the sampling design rather than a probability distribution. In design-based analysis, the outcome of interest is considered fixed and not random. The random variable is an indicator of sample inclusion selected with a probability called the inclusion probability.

Frame error: The error resulting from the disparity between the target population and the sampled population.

Inclusion probability: The probability that a sampling unit is included in the sample.

Measurement error: Error resulting from the difference in the measurements obtained and the true value of the measure. Sources of measurement error may include instrumentation errors, errors in the detection of units that are difficult to observe, and below-detection-limit data.

Membership design: See Sampling design.

Metric: A measurement representing a parameter of interest; a quantitative assessment of a monitoring characteristic.

Model-based inference: Inference based on a parametric model rather than the sampling design. In model-based analysis, the outcome of interest is treated as a random variable with a known probability distribution.

Multi-density category (mdcaty): A variable used to define design weights for unequal/continuous probability sampling. Note that the multi-density category differs from a stratum because sampling is not conducted independently within levels defined by the multi-density category.

Nonresponse error: Error resulting from the failure to obtain responses for the entire chosen sample. Sources of nonresponse error include site inaccessibility and the loss of samples from laboratory mistakes or the compromise of samples in the field.

Nonsampling error: Error that arises from the imperfect execution of the sampling design, including frame error, nonresponse error, and measurement error.

Overcoverage: A type of frame error that occurs when sampling units that do not represent members of the target population are included in the sampling frame.

Panel: A collection of sampling units always visited during the same sampling occasion.

Parameter: A population-level outcome of interest estimated by a statistic.

Response design: the method by which measurements are collected at a sampling unit. The response design may specify plot size and shape, the types of measurements taken at a site, and the response units from which measurements are made.

Response units: The object on which the measurement is taken. This is the basic unit of observation. In the above example, the observation units would be the individual frogs or possibly clusters of tadpoles.

Revisit design: Pattern for which sampling units are returned to for measurement.

Sample: A subset of a population. This may be chosen through a random, probabilistic mechanism (random sampling) or non-randomly (judgment sampling).

Sampled population: The collection of all possible observation units that might have been chosen from a sample. This is the population from which the sample was actually taken.

Sampling design: Also called membership design. Methods used to obtain a set of sampling units from a sampling frame in order to make inference to a target population.

Sampling frame: The list of sampling units designed to match the target population as closely as possible.

Sampling weight (or design weight): The inverse of the inclusion probability and a measure of the number of population units represented by the sampling unit based on the sampling design.

Sampling unit: The basic unit of replication in the sampling design and the unit selected in the sample from the sampling frame.

Statistic: An estimate of a population parameter.

Stratum: A subpopulation within which independent probabilistic sampling is conducted.

Survey design: Methods used to draw a sample (sampling design), visit the sampling units over time (revisit design), and collect data from the sampling units (response design) to meet quantitative goals.

Target population: Also called the resource population of interest; the complete collection of units to which inference is made.

Undercoverage: A type of frame error that occurs when all sampling units representing the target population are not included in the sampling frame.

Introduction

The National Park Service (NPS) was established by the Congressional “Organic Act” in 1916 “to conserve the scenery and the natural and historic objects and the wild life therein and to provide for the enjoyment of the same in such manner and by such means as will leave them unimpaired for the enjoyment of future generations.” The parks protected by NPS have grown from 35 in 1916 to 405 across the United States and its island territories, and over 270 of these parks are organized into 32 ecoregions called “networks.” Within these networks, signature and highest-priority park resources have been identified as “vital signs” for long-term monitoring. Vital signs include biological, chemical, and physical metrics of resources that represent overall park condition, are expected to respond to stressors, or possess a substantial human value. The NPS Inventory and Monitoring Program (I&M) was initiated in 1998 as a response to the Congressional National Parks Omnibus Management Act charging NPS with the task of obtaining baseline information and monitoring long-term “status” and “trend” of NPS resources (Fancy et al. 2009).

“Status” of a resource is defined as baseline information on the condition of that resource specific to a particular time period, such as a year or season. Estimates of status represent the size of a outcome and include but are not limited to means, medians, modes, variances, percentages, and proportions of the population. “Trend” of a resource describes how the condition of the resource changes over time. The term “trend” can have different definitions, including population-level changes in a resource over time, changes in individual unit’s response over time, averages of the metric of a resource over time, and net change across all individuals in a population (McDonald 2003). In long-term monitoring, trend is often characterized as population-level change in a particular direction (Urquhart and Kincaid 1999). While trend may have non-linear, cyclic, or seasonal components, significant trends in a particular direction will contain a linear component which may be used to quantify the magnitude of change in the resource of interest.

A crucial aspect of long-term monitoring is the selection of sites where measurements are made. An exhaustive census of a population is nearly always impractical because measuring outcomes of interest from all units in a population of interest is physically impossible due to time restraints, accessibility issues, and the inability to identify all units in a population. Sampling involves identifying a subset of units from a population and measuring the outcome of interest for this subset in hopes of gaining information about the entire population (Thompson 2002). Sampling is preferred over a census for three main reasons (Lohr 1999). First, a population sample is less costly than a census of all units in a population. Second, data collection and processing of a sample is less time-consuming than for an entire population, making results more readily available. Finally, estimates based on samples can exhibit higher accuracy than estimates from a census because observers can conduct field operations more carefully and thoroughly, thus reducing measurement errors.

When selecting a sample of sites for long-term monitoring, probability sampling should almost always be used (Fancy et al. 2009). Probability sampling provides a reliable mechanism for obtaining samples that meet the requirements of a wide array of analyses, including design-based and model-based inference. Design-based inference is based on the randomization inherent in the sampling

design and does not depend on a probability distribution (Lohr 1999). The outcome of interest is treated as a fixed, rather than a random, value. The random selection of sampling units selected with a non-zero inclusion probability provides the basis for statistical inference. In model-based inference, the population of the outcome of interest is assumed to be random with a distribution specified by a model. The sample, a single realization of many possible samples from the population, is linked to the population by the assumed joint probability distribution. Inference with a model-based approach does not account for the selection of the sample (Thompson 2002), but model-based inference may be biased if the probability that a sampling unit is included in the sample is a function of the outcome of interest (Schreuder et al. 2001). Design-based analyses have the advantage of few assumptions, and model-based analyses are highly accurate if assumptions are met. Models may be incorporated into design-based analysis approaches to obtain model-assisted methods of inference (Särndal, Swensson, and Wretman 1992). Regardless of the inferential basis, probability sampling provides a critical basis for making inference to the larger population.

NPS networks cover parks with varied resources over large geographic space. Sample sizes for monitoring programs are often small relative to the area of interest due to restricted budgets, narrow sampling windows, and time needed to collect data and travel between sites. The generalized random tessellation stratified (GRTS) sampling design is a spatially-balanced sampling design developed to distribute sampling effort so that selected points are not spatially clustered. When sample units are not spatially clustered, outcomes from those units tend to be less correlated and thus provide more information per unit (Stevens and Olsen 2003, 2004). Spatial balance disperses sampling effort across the extent of the resource so that samples achieve a similar spatial distribution as the population (Olsen et al. 2012). Spatial balance of GRTS samples allows the use of the neighborhood variance estimator (Stevens and Olsen 2003) for more precise estimation of parameters. Tools for sample draws and analysis are available in the *spsurvey* R package (Kincaid and Olsen 2013). GRTS sampling is discussed in greater detail in a later section.

The goal of this document is to serve as a resource for NPS monitoring personnel who design surveys using GRTS sampling designs and subsequently analyze data from those surveys. The document will first review sampling theory terms and sampling designs, including a more detailed discussion of GRTS sampling. Within the context of a spatially-balanced design like GRTS, the potential impact of nonsampling error on design-based estimation will be addressed, as well as methods to mitigate its effects. NPS case studies will be introduced and incorporated into examples of drawing GRTS samples and analyzing data from GRTS samples. Sampling term definitions are provided in Appendix A for quick reference.

Rather than providing R (2014) code for direct analysis, an analysis tool is available in an R package called SDDraw (McDonald et al. 2015). The updated version of this package, SDDrawNPS, incorporates more tools in a menu-driven interface to provide a more user-friendly approach to GRTS sampling and the analysis of data collected from spatially-balanced GRTS samples. The SDDrawNPS package will generate the R code used to draw samples and conduct analyses so that interested users can learn more about using the *spsurvey* package directly.

Sampling Design Background

In this section, the steps for designing a sample survey and issues that arise are discussed. Sampling terminology is defined. A glossary of sampling terms is also provided for reference in Appendix A.

General Sampling Design Steps

The design of any sample survey, including natural resource surveys, begins with a set of clear objectives. Defining the questions of interest is essential to obtaining a sample that meets the needs of the survey. The goals of the survey can often get lost in the myriad decisions necessary for a complex sampling design (Cochran 1977), but defining goals at the outset and referring to these goals during the design process can help maintain focus on monitoring plan objectives. Identifying relevant questions of interest for a monitoring plan may seem to be a straightforward process, but inattention to quantitative aspects of the monitoring program is a primary cause of monitoring program failure or ineffectiveness (Gitzen and Millspaugh 2012). Defining the goals of the study may be an iterative process among a team of researchers or resource managers with diverse, and sometimes divergent, interests. The questions of interest may begin as a set of broad subject-specific questions that require substantial fine-tuning to become a finite set of quantitative hypotheses (Valliant et al. 2013).

The next step in designing a sample survey is to define the population to which inference is made, or the target population. The target population, also called the resource population of interest, defines the extent of units that will be monitored over time. The target population may be further defined by spatial or temporal boundaries that reference a particular question of interest. For example, the target population for an invasive plant monitoring program might include areas within park boundaries containing invasive plants of a particular species. The target population may be further defined as occurring within specific elevational bands and during specific months to eliminate areas and times within which the species are unlikely to be observed. As with monitoring program goals, defining the target population can be a complicated process involving iterations of input from the design team. The definition of the target population should match the goals of the monitoring program, and each design feature may require changes to complement the other.

When the target population is defined, the sampling frame can be constructed. The sampling frame enumerates the set of units, called sampling units, from which the sample is chosen (Lohr 1999, Thompson 2002). Sampling frames may be obtained from inventories of a resource or GIS layers that provide quantitative information on the target population. Sampling frames rarely match the target population perfectly such that every unit in the target population is included only once in the sampling frame (Valliant et al. 2013). For the invasive plant monitoring example, the sampling frame might be areas within the defined elevational bands that fall within park boundaries. The accuracy of the sampling frame might be compromised if the elevation data in the frame contained error, causing areas outside of the elevational range to be included in the frame or areas that fall within the elevational band to be excluded from the sampling frame. Both instances are types of frame error (Lessler and Kalbeek 1992), a form of nonsampling error (error that arises from the imperfect execution of the sampling design) that will be discussed more thoroughly in a subsequent section.

When nonsampling error occurs, the sampled population, the collection of all possible observation units that might have been chosen from a sample, does not match the target population (Lohr 1999) and inference to the target population may be tenuous. If assumptions about the unsampled population can be made or appropriate models applied, then inference may be made to the target population. However, the validity of inference requires that the assumptions are met and the models are robust (Lessler and Kalbeek 1992).

Once a sample frame is constructed, sampling from that frame can commence. At this stage, sampling designs for long-term monitoring programs generally consist of three parts: the membership design, the revisit design, and the response design. The selection of locations in space represents the membership design. This determines which units are members of the sample. Sample locations may be selected in many ways, such as simple random sampling, stratified random sampling, multi-stage sampling, and spatially-balanced sampling. We discuss membership design options in subsequent sections of this manual.

The revisit design dictates the pattern of visits to sites over time (Urquhart and Kincaid 1999). The Vital Signs monitoring program is structured to monitor park resources over time. Sites selected in the membership design are visited in some pattern over time to provide temporal replication needed for trend estimation. Many revisit designs are possible, but temporal replication is necessary to detect trend (McDonald 2003, Urquhart and Kincaid 1999).

A panel is a set of sites that are always visited in the same temporal pattern (Olsen et al. 2012). For a total of P panels, let n_p represent the number of sites in panel p . Then the total sample of unique sites visited in the monitoring plan is $n = \sum_{p=1}^P n_p$. Sites from a GRTS sample are easily assigned to panels by assigning the first n_1 site to Panel 1, the next n_2 sites to Panel 2, and so on. Then each panel is also a spatially-balanced sample of sites because the order of the GRTS sample was followed in assigning sites to panels. Note that there is no rigidity associated with initial assignment of sites to panels. For example, if sampling resources are depleted before the full sample of sites is visited in the first year of monitoring, then unsampled sites originally assigned to Year 1 panels may be moved to the panels sampled in Year 2 and subsequent sites are also reassigned. However, following the order of the GRTS sample for panel assignment is recommended to preserve spatial balance and to ensure the randomization of sites to panels.

The third aspect of sampling design is the response design. The response design outlines the way data will be collected from sampling units. Response units are different from sample units because response units are the objects on which the measurement is taken and represent the basic units of observation, while sample units contain response units and are the basic units of replication in the statistical design. In the invasive plant monitoring example, the response units might be the plant or a section of a plot depending on details of the response design, while the sampling unit is the plot centered on the selected GRTS point. In a survey of Coho salmon redds, the sampling unit might be a stream reach and the response units might be redds within the reach.

Another aspect of developing a sampling design is determining the appropriate sample size. Guidance for calculating the appropriate sample size for estimating means and proportions is

provided in a later section. Nonsampling error such as nonresponse error or frame error can result in an observed sample size that is substantially smaller than the intended sample size. If nonsampling error is a possibility, then a larger sample should be drawn prior to the survey season so that a spatially-balanced sample of the intended sample size is obtained by the end of the monitoring season. The GRTS tools in the R *spsurvey* package (Kincaid and Olsen 2013) provide an option for an oversample, an additional set of spatially-balanced sites that can be used if needed. These sites are to be used in the GRTS design order only when the main sample has been exhausted. Misuse of the oversample includes replacing a non-target or inaccessible site with a nearby oversample site without following the order of the GRTS design file. Failure to use the oversample properly may result in spatial imbalance.

Note that use of oversample sites does not necessarily rid a sample of the impacts of nonsampling error. If the nonresponding sites differ substantially from the surveyed sites with respect to the outcome of interest, then inference from the obtained sample may be biased. Adding more accessible sites to the sample will not correct bias if the inaccessible sites are significantly different. In this case, model-based or model-assisted methods are needed to correct the biased in status estimates (Oh and Scheuren 1983, Särndal et al. 1992).

Membership Designs

Define a site as a spatial sampling unit delineated in the sampling frame. The membership design specifies how a site is selected as a member of the sample. The simplest form of selection is equal probability sampling. In this case, all members of the sampling frame have equal probability of inclusion. Simple random sampling does not incorporate location as a basis for randomization, so simple random samples in two dimensions result from a simple random sample of both the x-axis and y-axis in a spatial area. Simple random sampling can result in sites that are clustered in space or exhibit areas where no sampling units were selected (McDonald 2012).

Another basis for randomization is spatially-balanced sampling. A spatially-balanced sample ensures that the points within the sample are geographically diverse, which generally means outcomes exhibit less spatial correlation across units. When outcomes are less correlated, outcomes are more spatially independent of one another, thus increasing effective sample sizes. Several types of spatially-balanced samples exist, including two-dimensional systematic (or grid) samples, balanced acceptance sampling (BAS; Robertson et al. 2013), and GRTS (Stevens and Olsen 2003, 2004).

Two-dimensional systematic samples are obtained by imposing a grid over an area of interest and dividing each grid cell into secondary cells (McDonald 2012). A random secondary cell is chosen and the same relative secondary cell is sampled in all grid cells to provide the systematic sample. If the study area is of irregular size, then the sample size may be variable due to the mismatch of the grid to the study area. Replacing missing sites or sites that are out of the sampling frame is more difficult with this sampling design. Spatial balance may be compromised if sampling resources are exhausted before the complete sample is visited.

BAS (Robertson et al. 2013) provides an alternative spatially-balanced sampling design based on a random-start Halton sequence. As with GRTS sampling, BAS selection may originate from sampling

frames of a discrete set of elements or continuous frames for which the pixel size is the sampling unit. Benefits of BAS include computational efficiency and implementation in more than two dimensions, such as across elevational contours or in marine environments.

Many of the NPS networks have selected spatially-balanced GRTS samples for Vital Signs monitoring. GRTS uses a quadrant-recursive process to translate points from a two-dimensional space to a one-dimensional line. Then a systematic sample with a random start is selected and the sample points are translated back to two-dimensional space to obtain a spatially-balanced sample. GRTS samples demonstrate spatial balance for discrete, linear, and continuous frames, and the precision of design-based estimates of metrics from GRTS samples is calculated from the neighborhood variance estimator (Stevens and Olsen 2003) for more precise estimation of parameters. A large suite of analysis tools for GRTS sample data is available in the *spsurvey* R package (Kincaid and Olsen 2013). The purpose of this manual is to demonstrate tools for selecting GRTS samples and for design-based analysis of data from GRTS samples. The following section describes general sampling designs and then the remainder of this document will focus on GRTS samples.

Equal Probability Sampling

As mentioned above, all members of the sampling frame have equal probabilities of inclusion in this scheme. In an equal probability design, a selected randomization scheme is carefully followed to ensure that all members have an equal chance of selection. The assumption of randomness is critical to making proper inferences to the population of interest. Equal probability sampling in R is conducted with a pseudo-random number generator. The randomization is indexed by a “seed” number that can be set using the “set.seed” function prior to a sample draw so that sampling is exactly repeatable and yields the same results. In GRTS sampling, the quadrant-recursive process translates the two-dimensional points to a line. In equiprobable sampling, all locations in the sampling frame are given the same length along the line from which the systematic random sample is taken. Therefore, all locations in the sampling frame have the same probability of inclusion in the GRTS sample.

Stratified Sampling

In stratified sampling, the sampling frame is divided into non-overlapping strata from which independent random samples are drawn. Ideally, strata should be chosen so that members within a stratum are similar to members within the same stratum. Principal reasons to stratify include when estimates of known precision are desired for subpopulations, when stratification provides sampling efficiency, when the sampling design must necessarily differ within subpopulation levels, and when stratification reduces overall sampling error (Cochran 1977). When sampling units within a stratum are similar, within-stratum variance is minimized and the variance of population-level estimates across strata have smaller total variance than if a simple random sample of the same size had been taken.

Typically, the stratification variable is chosen based on physical locations where stratum member are expected to be similar. For long-term monitoring, strata should be defined by characteristics that are not expected to change over time. For example, sites within a vegetation community may be better

stratified by elevation than by specific vegetation types within the community if the vegetation community composition is expected to change over time. When stratum boundaries change relative to the defined strata from the sample, the differences among strata are reduced as are the benefits of increased precision from proper stratification (Johnson 2012). Stratum boundaries cannot be adjusted after the sample is drawn because the sites no longer represent a random selection from the adjusted strata. For clear and valid inference, use features that do not change over time for stratification.

Unequal Probability Sampling

The inclusion probabilities for unequal probability sampling are calculated relative to the distribution of a categorical variable across the population. In GRTS sampling, the one-dimensional line representing two-dimensional locations translated by the quadrant-recursive process are given the lengths proportional to their inclusion probability based on the categorical variable (Olsen et al. 2012). Therefore, sampling units with higher inclusion probabilities correspond to longer portions of the line and are more likely to be selected in the one-dimensional systematic random sample. Note that any values associated with the categories do not affect the inclusion probabilities. For example, assume that a cost-surface model is developed for the area interest that assigns a higher cost to regions that are more difficult or costly to access. The cost surface can be classified into a categorical variable indicating low, medium, and high survey effort. Unequal probability sampling based on the cost category would influence the inclusion probabilities only by the selected sample sizes within each cost class and the distribution of sampling units within each cost class, but the inclusion probabilities would not be influenced by the values of the costs in each category.

In the *spsurvey* package, the sample sizes for each category are specified as the sample size of the categorical variable, or “multi-density category.” Note that the specified sample sizes are not guaranteed within each category as with stratified sampling. The sample sizes for each category specify the inclusion probabilities for each category in unequal probability sampling, but the category-level sample sizes are random variables. The allocation of sample sizes across categories will be close to those specified but not exact. As the sample size increases to the population size, note that the realized sample sizes will converge to the relative population sizes for each category. Therefore, if a large sample (or oversample) is selected to account for nonsampling error or sites that rotate out of the revisit design, the allocation of sites within classes may not closely resemble those specified if the categories are sampled in substantially different proportions to their distribution in the population.

Continuous Probability Sampling

In unequal probability sampling, the probability of selecting a unit is directly proportional to the value of a continuous variable known for every sampling unit in the population. Therefore, the sites with the largest values of the continuous variable have a higher inclusion probability than those with smaller values of the continuous variable. In the cost-surface example, the values of the cost-surface model are directly related to the inclusion probabilities rather than relative sample sizes as with unequal probability sampling. If the goal of the cost-surface model is to reduce the number of high-cost sites selected in the sample, then continuous probability sampling would be conducted for the inverse of the cost value at each site so that low-cost sites would be included more frequently than

high-cost sites. Note that sample sizes are not specified by groups as with unequal probability sampling. The overall sample size is specified for continuous probability sampling and allocated proportionally to the values of continuous covariate in the population.

Two-Stage Sampling

In two-stage sampling, the population is divided into primary sampling units (PSUs) that are composed of secondary sampling units (SSUs). For example, an area of interest might be divided into watersheds (PSUs) that consist of several streams (SSUs). In two-stage sampling, a sample of PSUs is randomly selected within a population (the first stage of sampling), and then SSUs within each selected PSU are sampled (the second stage of sampling). Two-stage cluster sampling is similar to stratified sampling in that sampling is conducted within each PSU but different from stratified sampling in the PSUs are sampled rather than censused. Two-stage sampling is useful when SSUs within each PSU are very similar, so censusing all SSUs within a PSU is an inefficient use of sampling resources (Lohr 1999). Two-stage sampling may also provide survey efficiency when sampling within a PSU is more efficient than equiprobable sampling. For example, travel costs among PSUs may be large relative to costs within PSUs. In this case, sampling a few PSUs and minimizing travel costs within a smaller area may be more cost-effective than spreading a large sample of points across the entire area of interest. GRTS samples may be taken at both the first and second stages to obtain spatial balance among both PSUs and SSUs within selected PSUs.

Master Sample

A master sample is a sample that is used like a sampling frame for multiple surveys (Turner 2003). First, a large initial sample is taken. Then samples are taken from the set of master sampling units for surveys conducted by multiple agencies, monitoring programs, or for different outcomes of interest. The use of a master sample ensures that all samples draw share the same sampling frame so that results from multiple agencies are comparable. Master samples provide a straightforward method for conducting multiple surveys at the same sites (co-location), providing data from the same sites that might be useful at the analysis stage. Subsampling from master samples may focus on specific subpopulations and require a denser set of points for inference.

Several agencies have used dense GRTS samples so that samples at various scales can be used (Larsen et al. 2008, Tiszler *in press*). When specific subpopulations are of interest, sites falling in the desired subpopulations are identified and visited in the order of the GRTS design file to maintain spatial balance within each subpopulation. Selecting a dense sample of sites ensures that adequate sample sizes are achieved within subpopulations of interest. Note that design-based analysis is based on the unconditional probability of sample inclusion calculated as the product of a unit's master sample inclusion probability and the inclusion probability for the sample selected from the master sample conditional on selection in the master sample.

Nonsampling Error

Sampling error is generated by the uncertainty inherent in surveying only a portion of the population. When all members of a population are sampled (resulting in a census of the population), then population parameters are known from the census with no variation. When only a sample is obtained from a population, variation is induced in the estimate of the population parameter because that

quantity cannot be known exactly. Nonsampling error arises from the imperfect execution of a sampling design (Lessler and Kalsbeek 1992). Nonsampling error is avoided only when a survey is designed and executed perfectly, so nearly all surveys are subject to this error source. However, measures can be taken to minimize the effects of nonsampling error both before and after the survey. Careful planning prior to survey implementation and appropriate analysis techniques can mitigate the effects of nonsampling error so that unbiased inference may be obtained. The type of nonsampling error must be determined to identify the appropriate analysis approach. Nonsampling error is comprised of three sources: frame error, nonresponse error, and measurement error (Lessler and Kalsbeek 1992).

Frame Error

Frame error occurs when the sampling frame does not match the target population (Lessler and Kalsbeek, 1992). The associations between the sampling frame and the target population are referred to as *linkages* (Särndal et al. 1992). A one-to-one linkage exists when the units in the target population match a single unit in the sampling frame, and the units in the sampling frame match a single unit in the target population. Frame error arises when a one-to-one linkage does not occur. For instance, when some target population units are not linked to the sampling frame, undercoverage occurs. For example, in a survey design for prairie potholes, data on prairie pothole habitats is unavailable for a few counties. In this case, our sample design does not include all the potential habitats of interest, and our design will show undercoverage. On the other hand, overcoverage occurs when some units in the sampling frame are linked to units that are not members of the target population (i.e. non-target units). In the same example, a series of farm ponds that were mistakenly classified as prairie potholes and included in the sample frame would be overcoverage. Some target population units may be linked to multiple frame units (duplicate listings), and the frame may contain incorrect information that forms the basis for stratification, probability-proportional-to-size selection, or regression estimation. Frames can become obsolete when the population has changed over time or may be of poor resolution to establish linkages with all members of the target population (Särndal et al. 1992). Sampling frames may be subject to one or more of these problems.

Because the sampling frame is constructed to match the target population as closely as possible, frame error may not be identifiable until observers reach the site. The definition of the target population must be clearly defined for an observer team to assess whether a site is a member of the target population. The site is recorded as “Non-target” and no data are collected.

Frame error can impact inference in several ways. Depending on the distribution of the frame error, spatial balance may be impacted by sites omitted as non-target. If attributes of the frame are used as design variables such as strata or multi-density categories, then sample sizes may be adversely impacted for subpopulation-level inference. When frame error is expected, then larger samples (or oversamples) can be selected to provide additional sites if frame error occurs.

Nonresponse Error

Nonresponse error occurs when complete measurements are not obtained for a sampling unit. This type of error may arise from the nonrandom exclusion of a portion of the target population (Cassel, Särndal, and Wretman 1983). For example, site with steep slopes may be excluded from a survey for

safety reasons. If the missing outcomes differ significantly from those collected at the accessible sites, then inference from the set of accessible sites may be biased. Nonresponse error is further classified based on the amount of missing information. If all outcomes for the entire sampling unit are not observed, then the data are subject to unit nonresponse. If at least one outcome of interest is missing but not all outcomes, then the unit is subject to item nonresponse (Little and Rubin, 2002).

Understanding the mechanism that generates the missing data is useful for determining how to address the nonresponse in the analysis phase (Little and Rubin 2002). There are three general classifications for missingness mechanisms of unit nonresponse: missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). The distinctions among the three mechanisms regard the dependence of the missingness on the outcome of interest.

If the missingness mechanism is unrelated to the outcome of interest, any covariates, and any aspect of the sampling design, then the missingness is characterized as MCAR. In this case, the mean of the outcome of interest is assumed to be equal for both the responding and the nonresponding units. Therefore, the reduced sample due to missingness is still regarded as a random subsample of the sampling frame and inference on the obtained sample is valid. An example of MCAR missingness might include site missed because the sampling season ended prematurely due to budget cuts or a water sample was accidentally dropped and destroyed.

When missingness is MAR, the missingness is related to the outcome of interest but can be explained by covariates. For example, if sites that are missing are all on steep slopes, then the slope of each site may be used to account for the missing data. Collecting data related to the missingness at all sites is necessary for MAR nonresponse adjustments such as weighting class adjustments and post-stratification adjustments (Oh and Scheuren 1983).

When data are NMAR, the missingness cannot be explained by covariates and inference based on the partial sample is biased. Adjusting for NMAR missingness requires either a modeling approach that incorporates assumptions on the missing units or a nonrespondent subsample, which is often not possible in ecological sampling. Increasing the sample size does not help because the additional sites represent only the sampled subpopulation rather than the target population.

Nonresponse adjustments are a more complicated aspect of design-based estimation, and each sample requires careful consideration of the missingness mechanism. In this work, we assume MCAR missingness for weighting adjustment. However, the impacts of this assumptions should be reviewed in practice and more advanced adjustment applied if necessary.

Measurement Error

Measurement error, another source of nonsampling error, occurs when an inaccurate response is obtained for a unit (Lessler and Kalsbeek, 1992; Thompson, 1992). Three types of measurement error occur in surveys: the use of an inaccurate instrument, an inaccurate measure provided by a respondent, and an interviewer influencing the response (Särndal et al. 1992). An inaccurate instrument might include an uncalibrated thermometer, poorly constructed questionnaire, or an inadequately-trained observer. Response error, another form of measurement error, is generated

when survey respondents give false information, change their response due to survey involvement, or forget relevant details. The third type of measurement error may result when the interviewer/observer or survey timing affect the outcome. Varying levels of observer experience or trampling effects of observers might result in measurement error in outcomes.

Sample Size Approximation

A consideration fundamental to the success of an environmental monitoring program is sample size, defined here as the number of units necessary to estimate a parameter of interest with a specified precision. How many sampling units should be collected, and how they should be allocated across subpopulations of interest (*e.g.* strata), should be considered early and carefully in developing a monitoring program (Manly 2009). While these general rules of thumb might not apply to every question of sample size allocation, they do provide a means by which to approximate the sample size necessary to efficiently accomplish carefully considered survey objectives.

Sample sizes should be sufficient to estimate important parameters with sufficient precision to be useful (Manly 2009). Utility in environmental monitoring studies often implies the ability to detect existing spatial or temporal differences or changes due to management actions or anthropogenic disturbances. Often, the sample size is dictated not by statistical considerations but instead by logistical or budgetary constraints. In the context of spatially-balanced sampling designs, sample sizes must be large enough to adequately cover the region of interest. Since generating a spatially-balanced sampling design requires the user know how many sampling units to draw, the question of sample size must be considered prior to the GRTS sample draw.

General formulae are provided for guiding the determination of sample sizes (Manly 2009). Represent the error level (or margin of error) for an estimator as δ , where δ is the half-width of the confidence interval on the estimate. Suppose a study is designed to estimate a parameter of interest, and the error level δ is considered acceptable in the context of the study objectives. In a simple random sample, the sample size (n) necessary to estimate a population mean with a $100(1 - \alpha)\%$ -confidence interval is:

$$n = \frac{z_{(1-\alpha/2)}^2 \sigma^2}{\delta^2}$$

where σ is the population standard deviation and $z_{(1-\alpha/2)}$ is the $(1 - \alpha/2)$ -quantile of a standard normal distribution (*e.g.* $z_{(1-\alpha/2)}$ is 1.96 when $\alpha = 0.05$ and 1.645 when $\alpha = 0.10$). Using this formula requires some level of knowledge about the parameter of interest, information often gained from pilot studies, preliminary surveys, or bootstrapping techniques.

To estimate the $100(1 - \alpha)\%$ -confidence interval for a population proportion (for example, the proportion of sites with a pH below a threshold value) from a simple random sample, the approximate sample size is calculated as:

$$n = \frac{z_{(1-\alpha/2)}^2 \pi(1 - \pi)}{\delta^2}$$

where π is the true population proportion. When π equals 0.5, the equation reduces to $n=1/\delta^2$.

If two random samples are taken from populations with the same standard deviation (but potentially different means), the sample size necessary to calculate a $100(1 - \alpha)\%$ -confidence interval for the difference between the population means is:

$$n = \frac{2z_{(1-\alpha/2)}^2 \sigma^2}{\delta^2}$$

where σ is the population standard deviation.

To calculate a $100(1 - \alpha)\%$ -confidence interval for the difference of two population proportions, the necessary sample size is:

$$n = \frac{2z_{(1-\alpha/2)}^2 \pi' (1 - \pi')}{\delta^2}$$

where π' is the average of the two population proportions.

These basic formulas provide guidelines for sample size approximations. For stratified random sampling, the total sample size is then allocated across the strata in one of several ways (Thompson 2002). Samples of the same size may be taken within each stratum (equal allocation). If the strata differ in size, the relative proportions of each stratum in the population may be applied to the total sample size to obtain stratum-level sample sizes (proportional allocation). If one or more strata exhibit high variation, sampling those strata at a rate proportional to their variances may decrease the overall variance in the sample (optimal allocation). When inference is desired for subpopulations not included as strata in the sampling design (called domains), similar considerations must be made to obtain adequate sample sizes within those domains. For more information on sample size approximation and allocation, refer to a sampling text (Cochran 1977, Lohr 1999, Thompson 2002).

For monitoring programs expected to yield both annual estimates of status as well as estimates of trend over time, the allocation of sampling resources across time and space may reflect competing interests. Power analysis (Cohen 1988; Sims et al. 2006) is another useful tool for determining appropriate sampling size. When the monitoring goal involves a statistical hypothesis test (such as a test of a regression coefficient indicating trend), the power of the test can be assessed for a range of sample sizes. Comparing sample size approximation for a single-year estimator of status with the sample size needed to achieve adequate power for trend detection is a useful approach for assessing the trade-offs necessary to attain reasonable monitoring goals.

GRTS Sampling

The necessity for environmental monitoring programs to efficiently and effectively survey habitats, chemical characteristics, or biota, for example, over large survey areas has instigated research to design appropriate survey designs. One particular approach emerging from that research is the generalized random tessellation stratified, or GRTS (Stevens and Olsen 2003), sampling design. A GRTS design generates an ordered list of sample points over a defined population or geographic area that exhibits spatial coverage. One aspect of the sample design, called *reverse hierarchical ordering*, insures that, for a sample of size n , the first n units in the sample will exhibit spatial balance. More generally, any contiguous set of sample units in a reverse hierarchical ordered GRTS sample exhibits spatial balance. This is a powerful aspect of GRTS and other modern spatially balanced designs.

The GRTS and other spatial designs are flexible and adaptable, key considerations in the development and evolution of environmental monitoring programs (Stevens and Olsen 2004). For example, the objective might be to monitor the number of a particular species in an area using a spatially balanced sample, which can be modified to increase sample sizes in certain parts of the species range if necessary (i.e., unequal probability). In addition, monitoring objectives often change through time, and with adequate planning and forethought modern spatially balanced designs provide a means to adjust sample sizes or sampling intensity when necessary without adversely effecting spatial balance of the overall sample (Stevens and Olsen 2004). Design aspects, like stratification, cluster sampling, and variable probability sampling, can easily be incorporated into these designs. The R package, *spsurvey* (Kincaid and Olsen 2013), may be used to draw GRTS samples of various design complexity, the details of which will be detailed in a later section.

GRTS survey designs provide advantages during the analysis phase as well. Natural resources spread over a large region or across a wide range of resource conditions can exhibit wide variation. This variability may obscure real trends in a monitored resource. GRTS sample data analysis incorporates the neighborhood variance estimator, a variance form that computes the components of variation only within neighborhoods of points. This takes advantage of the widely observed fact that nearby sites tend to be more similar to each other than to others from far away. Neighborhood variance estimates were found to be 22% to 58% smaller than variances computed assuming an independent random sample (Stevens and Olsen 2003). The *spsurvey* package (Kincaid and Olsen 2013) provides tools to obtain estimates of resources using data collected through GRTS survey design.

GRTS Algorithm

Environmental resources can come in a variety of forms: a manager might be interested in sampling a resource that looks like points across a region (e.g., prairie potholes), one that looks like a linear network (e.g., stream segments within a watershed), or one that can be characterized by polygons (e.g., clusters of cottonwood trees along river edges). The GRTS design maps two-dimensional space into one-dimensional space using a quadrant-recursive function and creates an ordered spatial address for each resource. The ordering of those points on the line preserves spatial relationships; resources that are close to each other in space tend to be close to each other on this ordered line (Stevens and Olsen 2004). A square grid overlay of a region becomes 4 grid cells within that original

grid cell, and the subdivision continues with grids nesting within grids. When mapped onto a line, the first quarter of the points come from the first quarter of the original grid, the second quarter of points comes from the second quarter of the original grid, and so on (Figure 1). By selecting points from each quarter of the mapped line, a spatially balanced design is guaranteed. An excellent example of this process used to select Oregon lakes for sampling is provided in Olsen et al. (2012).

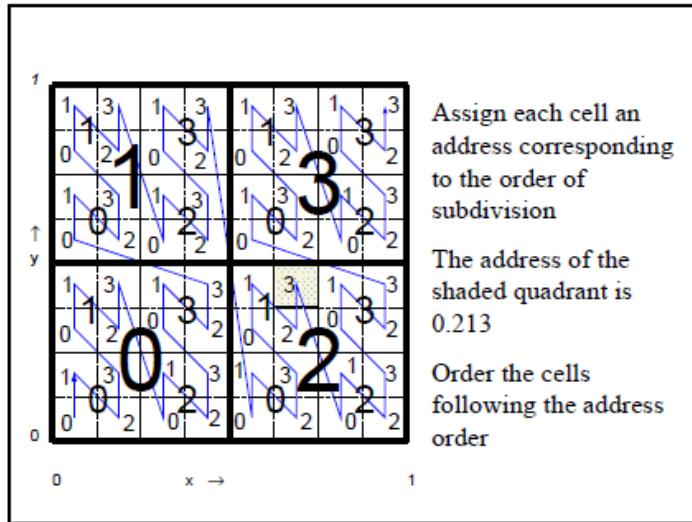


Figure 1. Quadrant-recursive partitioning of a unit square (Stevens and Olsen 2004).

Hierarchical ordering of samples has several benefits. First and foremost, the full set of sample points exhibits spatial balance over the target population. Secondly, the ordered list of points obtained during a GRTS draw allows flexibility when sampling resources fluctuate. For example, assume a GRTS of 100 sites is planned to be surveyed during summer. However, due to unanticipated delays at the start of the field season, only 86 of those points were actually surveyed. If the sequential ordering of those points was maintained resulting in a sample of 86 contiguous points on the list, then those 86 points will achieve spatial balance. As a second example of the utility of this design, suppose those 86 sites were visited again during the following summer. If one or more of the sites revisited were deemed inappropriate for the survey, perhaps because the site was damaged in a spring disturbance, the non-target sites can be replaced with the next sites on the sequential list. The GRTS design maintains the spatial balance in the order of sites, and that spatial balance can be preserved even with slight changes to the survey.

Neighborhood Variance Estimator

Design-based estimates based on data collected from a GRTS sample in a single year are obtained with the Horvitz-Thompson estimator (Horvitz and Thompson 1952). The neighborhood variance estimator (Stevens and Olsen 2003, 2004) is used to account for the spatial balance of the GRTS sample and is calculated as:

$$\hat{V}_{NBH}(\hat{\mu}) = \left(\sum_{s_j \in R} \frac{1}{\pi(s_j)} \right)^2 \sum_{s_j \in R} \sum_{s_i \in D(s_j)} w_{ij} \left(\frac{z(s_j)}{\pi(s_j)} - \sum_{s_k \in D(s_i)} w_{ik} \frac{z(s_k)}{\pi(s_k)} \right)^2,$$

where:

$\hat{\mu}$ is the estimator of the population mean, μ ,

$s_{i,j}$, or k are sample points,

R is the domain of the sampled population defined as a set of points occupied by the population elements,

$D(s_i)$ is the local neighborhood for site i ,

w_{ij} are weights chosen to reflect the behavior of the pairwise inclusion function for GRTS, and are constrained so that $\sum_i w_{ij} = \sum_j w_{ij} = 1$,

$z(s_i)$ is the real-valued outcome measured at the i th point, and

$\pi(s_i)$ is the inclusion density for site i .

Flexibility of GRTS Samples

The GRTS survey design also can accommodate the need for new sites or additional effort. For example, assume a GRTS sample of 100 points and an oversample of 100 additional points are drawn. New sites might be needed if a site “wears out” during the monitoring period. For example, a vegetation survey conducted repeatedly at a site over several years might impact the vegetation present at that site. Trampling damage might affect the monitored outcome, and the site may need to be removed from the sample. The next unsampled site on the ordered GRTS list may be added as the old site is retired. If survey resources become available during the monitoring period, additional sites can be added from the oversample. However, the first priority is to visit all sites in the main sample before starting at the beginning of the oversample list. The oversamples should not serve as alternates for the regular sample list. This is a common misconception of the oversample and using it as such can diminish the spatial balance of the survey. The oversample is to be treated as a large main sample that is used in order as needed.

Sampling Frames for GRTS Samples

Three types of sampling frames may be used to draw GRTS samples: discrete objects, linear features, and areas (Olsen et al. 2012). A sampling frame of discrete objects might include a large list of mountain lakes or a set of timber stands with each lake or stand serving as a sampling unit for selection. Linear features are generally represented by stream networks which have negligible width relative to length. Areal features represent large regions within which points are selected in the sample.

GIS layers or spatial objects in R (from the *sp* package in R) serve as sampling frames for GRTS sampling. The coordinate system for the GIS layer should be an area-preserving projection such as Albers or UTM so that spatial distances are equivalent for all directions (Olsen et al. 2012). The frame should include the complete area corresponding to the target population and exclude areas outside the target population. The corresponding attribute file should contain covariates representing design features such as strata. If areas within the target population are known to be inaccessible prior

to the sample draw, these areas may be omitted before the draw to retain spatial balance after accounting for accessibility issues. However, the scope of inference is reduced to the sampled population (the accessible subpopulation of the target population).

Case Studies

Two case studies from the NPS Vital Signs Monitoring Program are used as examples for this manual. Note that these case studies provide a basis for sample draw examples, but are in most cases not the actual sample draw for the monitoring program.

MOJN Integrated Upland Monitoring

The MOJN integrated upland monitoring program is designed to monitor both status and trend of indicators of the function, composition, and structure of MOJN ecosystems. Species richness, vegetative cover measurements, frequency, and density are indicators of this Vital Sign that will be monitored over time. The MOJN integrated upland polygon sampling frame is an areal sampling frame of polygons ranging in size from 0.03 to 153 ha and ranging from 1000 to 1400 m in elevation (Pan and Tallent 2012). Four plant associations of interest were identified: creosote bush (LT), Mojave Yucca/big galleta (PR), Joshua tree/blackbrush (CR), and desert almond/California juniper/Nevada Mormon tea (JCPF). This sampling frame is used to demonstrate sampling from points in a continuous area.

Because the macroplots used in the MOJN integrated upland monitoring program are 1-hectare in area, positioning the macroplots in randomly selected points proved to be problematic in small polygons during the first year of sampling. A finite sampling frame consisting of the centroid of each polygon is created from the areal polygon. In this scenario, polygons serve as the sampling unit and macroplots are centered on the centroid of selected polygons to demonstrate sampling from a finite sampling frame.

MEDN Stream Monitoring on Santa Rosa Island

The linear sampling frame example comes from a monitoring program of streams on Santa Rosa Island (SRI) in Channel Islands National Park. SRI has been devastated by browsing, and removal of ungulates is expected to improve riparian condition. The monitoring program will sample SRI streams and measure indicators of riparian condition, geomorphology, benthic macroinvertebrate populations, and water quality over time. A mapping survey was recently completed and will form the basis for the sampling frame, but some undercoverage of SRI streams is expected. Because a minimum requirement for stream discharge defines the target population, a high rate of frame error due to overcoverage is also expected. Within discrete reaches, GRTS points will be selected. The subset of the selected reach that meets the definition of the target population will be identified for sampling. MEDN expects to sample about 10 points each year and may incorporate panel structure in a temporal revisit design. Outcomes such as stream depth and temperature are collected.

Installing the SDrawNPS Package

Installing the SDrawNPS package either from a local zipped file or from a binary source file is a straightforward process. The installation files should be available on the user's hard drive in a known directory. These files may be downloaded from the following web site:

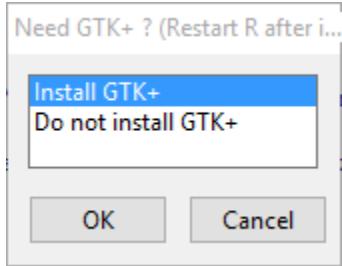
<https://irma.nps.gov/DataStore/Reference/Profile/2229965>.

To begin installation of SDrawNPS, open the R program, and, in the command window at the prompt (indicated by a greater-than symbol: >), type the following and hit Enter to install the dependent packages:

```
install.packages( c("RGtk2", "spsurvey", "utils", "rgdal", "rgeos", "sp"), repos="http://cran.r-project.org")
```

This will download and install the packages required by SDrawNPS. Note that you must have an internet connection to do this, as the packages will be downloaded from a CRAN mirror site. Furthermore, if the version of R is recent, a message may appear that “utils” is included in the base package and should not be updated. This is fine. Older versions of R, though, may require “utils” to be downloaded and installed as a separate package.

Once these packages are installed, type require(RGtk2) at the prompt. In some cases, loading the RGtk2 package will identify an issue with a missing component (GTK+). If this is the case, a pop-up window will prompt an additional installation, stating: “Need GTK+? (Restart R after installation)”.



Click OK to install GTK+, and the necessary programs will download and install. Error messages may occur at this point. Close and reopen the R window.

After re-opening R, check to see if the GTK+ component installed correctly. To do this, type require(RGtk2). If an error message occurs, attempt to reinstall GTK+ following the above instructions. If no error message occurs, install the SDrawNPS package. We recommend using a recent version of R (3.2.X suggested, the most recent version is 3.2.5) for sample draws and analyses; however, slightly older versions of R (3.1.3 for example) have been used to successfully execute draws and analyses. Versions older than 3.1.3 should be avoided.

This program has been installed on many different PCs with different operating systems (Windows 7 and Windows 10) using different versions of R. Installation usually proceeds most smoothly using

the 32-bit version of R (e.g. ‘R 386i 3.1.3’). If the user’s computer is a 64-bit computer, SDrawNPS should be installed using 64-bit R. Once installed, either the 32- or 64-bit version of R can be used to run the SDrawNPS package.

To install from the zipped installation file (SDrawNPS_1.0.zip), open R and select the drop-down menu item Packages>Install package(s) from local zip files... then use the browser window to locate and select the zipped installation file. If any dependent packages are missing, an error messages will print indicating which packages are needed. Install these missing packages and the SDrawNPS package. When the SDrawNPS package is successfully installed, the message “package ‘SDrawNPS’ successfully unpacked and MD5 sums checked” will print to the console.

The remaining instructions will focus on installing directly from a binary source program. Download the SDrawNPS installation file `SDrawNPS_1.0.tar.gz` and store it in an appropriate directory on your computer. Let “`..PATH..`” represent the directory on the user’s computer where the file `SDrawNPS_1.0.tar.gz` is stored. Install SDrawNPS by typing the following line and hitting enter:

```
install.packages("..PATH../SDrawNPS_1.0.tar.gz", type="source", repos=NULL)
```

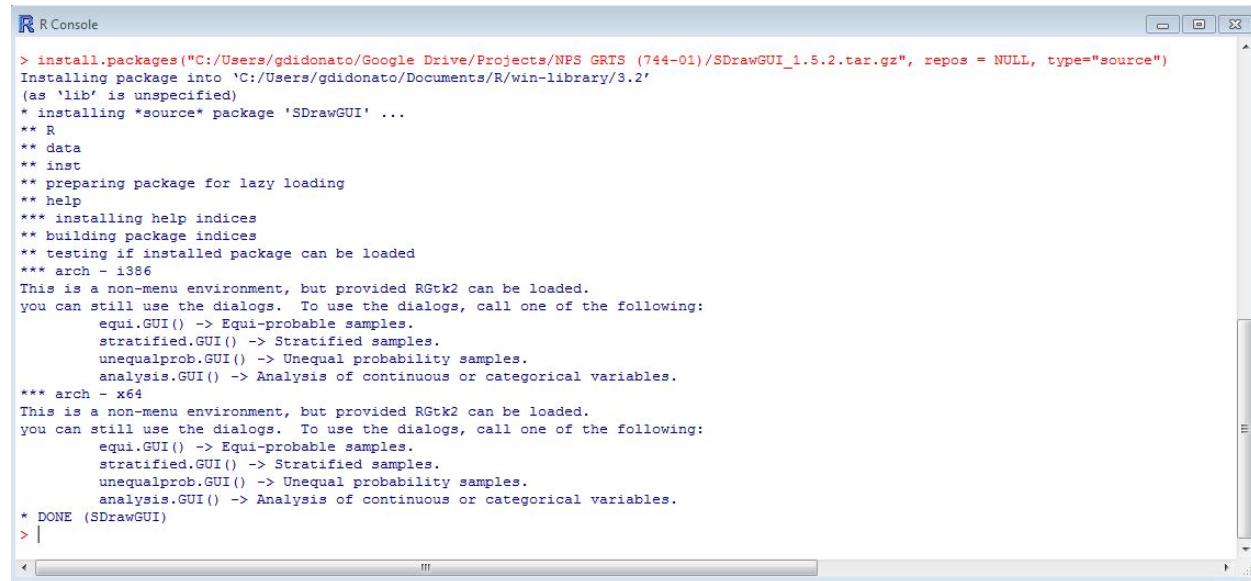
For example, if the installation file is downloaded to the TEMP directory on the C drive, then the path is “`C:/TEMP`” and the following command would be used to install SDrawNPS:

```
install.packages("C:/TEMP/SDrawNPS_1.0.tar.gz", type="source", repos=NULL)
```

If you have already changed your working directory to the location of the installation file, then enter:

```
install.packages("SDrawNPS_1.0.tar.gz", type="source", repos=NULL)
```

Next, the following text (or similar) will appear in the R window, indicating a successful installation:



The screenshot shows the R Console window with the following text output:

```
R Console
> install.packages("C:/Users/gdidonato/Google Drive/Projects/NPS GRTS (744-01)/SDrawGUI_1.5.2.tar.gz", repos = NULL, type="source")
Installing package into 'C:/Users/gdidonato/Documents/R/win-library/3.2'
(as 'lib' is unspecified)
* installing *source* package 'SDrawGUI' ...
** R
** data
** inst
** preparing package for lazy loading
** help
*** installing help indices
** building package indices
** testing if installed package can be loaded
*** arch - i386
This is a non-menu environment, but provided RGtk2 can be loaded.
you can still use the dialogs. To use the dialogs, call one of the following:
  equi.GUI() -> Equi-probable samples.
  stratified.GUI() -> Stratified samples.
  unequalprob.GUI() -> Unequal probability samples.
  analysis.GUI() -> Analysis of continuous or categorical variables.
*** arch - x64
This is a non-menu environment, but provided RGtk2 can be loaded.
you can still use the dialogs. To use the dialogs, call one of the following:
  equi.GUI() -> Equi-probable samples.
  stratified.GUI() -> Stratified samples.
  unequalprob.GUI() -> Unequal probability samples.
  analysis.GUI() -> Analysis of continuous or categorical variables.
* DONE (SDrawGUI)
> |
```

Once the package is installed, the user can load the SDrawNPS package into the R environment by either typing at the command line:

```
require(SDrawNPS)
```

or by navigating on the menu to Packages/Load package... and selecting SDrawNPS from the list of installed packages.

As the package is loaded, the following text will appear in the R window:

“Loading required package: sp”

“SDrawNPS – Sample Draws (vers 1.0)”

If the user’s version of R is earlier than 3.2.3, an additional message will appear in the R window:

“Warning message:”

“package ‘SDrawNPS’ was built under R version 3.2.3”

Installing SDrawNPS on a more recent version of R will eliminate this warning message. However, the package should function properly despite the warning.

SDrawNPS will also appear in the menu bar along the top of the R window. Sample draws and sample analyses can now be done. When the SDrawNPS package is successfully loaded, the following text will appear in the command window:

To use the dialogs, call one of the following:

equi.GUI() -> Equi-probable samples.

stratified.GUI() -> Stratified samples.

unequalprob.GUI() -> Unequal probability samples.

analysis.GUI() -> Analysis of continuous or categorical variables.

These commands are used as an alternative to the drop-down list menu items or when the drop-down menu is not available, such as when using RStudio as an interface for R.

The SDrawNPS package is the package developed as a companion tool to this User’s Manual. Note that the SDrawNPS package is a forked software product developed from an existing package called SDraw. Similar packages will be updated in the future. Subsequent SDraw packages may offer tools not described in this document, so the authors encourage users to explore future SDraw packages for new sampling and analysis tools.

Drawing a GRTS Sample

The grts() Function in *spsurvey*

When drawing a GRTS sample in SDrawNPS, the *spsurvey* package (Kincaid and Olsen 2013) for drawing GRTS samples is called. The *grts* function is the basic function for drawing GRTS samples, and the function requires specific formatting for the input objects. The SDrawNPS package converts the GUI inputs into the proper formatting for easier use, and we review the commonly-used inputs here.

The “design” input contains the relevant sampling design features such as strata definitions, sample size allocation, random selection type (equal, unequal, or continuous), and oversample size, if specified. This information is formatted as an R list, which can be complicated to organize properly. By specifying this information in the SDrawNPS GUI, the inputs can be clearly organized. Note that the SDrawNPS GUI does not provide panel specification. This input is omitted due to user confusion on the process of assigning sites to panels. Panels are used in temporal revisit designs, and sites are allocated to panels in sections of designated size. When sample sizes differ from what was originally planned, confusion can occur in panel assignment when panels are assigned with the GRTS sample. We omit this option here and encourage resource managers to assign sites to panels each year according to the revisit design and available sampling resources.

The “*type.frame*” option specifies the sampling frame type (finite, linear, or area).

The “*src.frame*” option takes one of three values: “shapefile” if a shapefile is used for the sampling frame, “sp.object” if the sampling frame is an object from the *sp* package, or “att.frame” if the sampling frame is represented by the *att.frame* field (below) and sampling is from a finite frame.

The “*in.shape*” option is the name of the sampling frame shapefile when *src.frame* = “shapefile”.

The “*sp.object*” option is the name of the sampling frame *sp* object when *src.frame* = “sp.object”.

The “*att.frame*” option provides the name of the attribute data frame to match the sampling frame. This data frame contains the multi-density category or stratification data if specified. This is also the name of the sampling frame for finite sampling when *src.frame* = “att.frame”.

The “*xcoord*” and “*ycoord*” inputs provide the x- and y-coordinate data needed for the draw. Note that these coordinates should reflect an area-preserving projection such as Albers or UTM.

The “*stratum*” input provides the name of the stratification variable in the *att.frame* data frame.

The “*mdcaty*” field provides the multi-density category for unequal or continuous selection sampling.

The “*out.shape*” field allows the user to specify the name of the GRTS sample file that is created.

In the R command window, enter `?grts` for more information on using the *grts* function in the *spsurvey* package. Note that the *grts* function call in the SDrawNPS package is written to a text file (*.log) when the sample is drawn so that inputs to the function are clearly archived. The output file is

called the GRTS design file. This file provides the appropriate order of the sites needed to obtain spatial balance. The original GRTS design file and the sample draw commands from the log file should be archived and not changed over time so that the sampling design used may be referenced without error in the future. Copies of the GRTS design file may be used to add the survey evaluation information (EvalStatus field, see more in the Analysis section of this manual), the outcomes of interest by site, and field notes. Then the modified design file will include the design weights, spatial coordinates, site evaluation information (especially important for dealing with nonsampling error), and the data and can be used for design-based estimation using the analysis tools available in SDrawNPS. Given the ease with which data files can be corrupted or mismanaged, archiving the original GRTS design file and the sample draw commands is highly recommended.

Equal Probability Sampling

In equiprobable sampling, all sampling units in the population have the same probability of selection. An example of equi-probable GRTS sampling incorporates a monitoring design from Channel Islands National Park (CHIS).

Case Study: Polygon Frame

One hundred ninety-nine watersheds were identified and mapped on Santa Rosa Island (SRI), one of five islands in CHIS. Suppose an invasive plant had been detected on SRI, and monitoring personnel are interested in collecting presence/absence information about this plant across the island to determine its distribution across the island and its watersheds. An equiprobable GRTS design would provide a spatially-balanced survey design that would allow a comprehensive assessment of the occupancy rate of this particular plant on this island.

To draw an equiprobable GRTS sample, first the SDrawNPS user would select the SDrawNPS dropdown menu list from the R window. Then, select “Sample Draws/Equi-Probable . . .” to obtain an SDrawNPS interface through which the user can type in or select the necessary inputs to create the desired survey design (Figure 2).

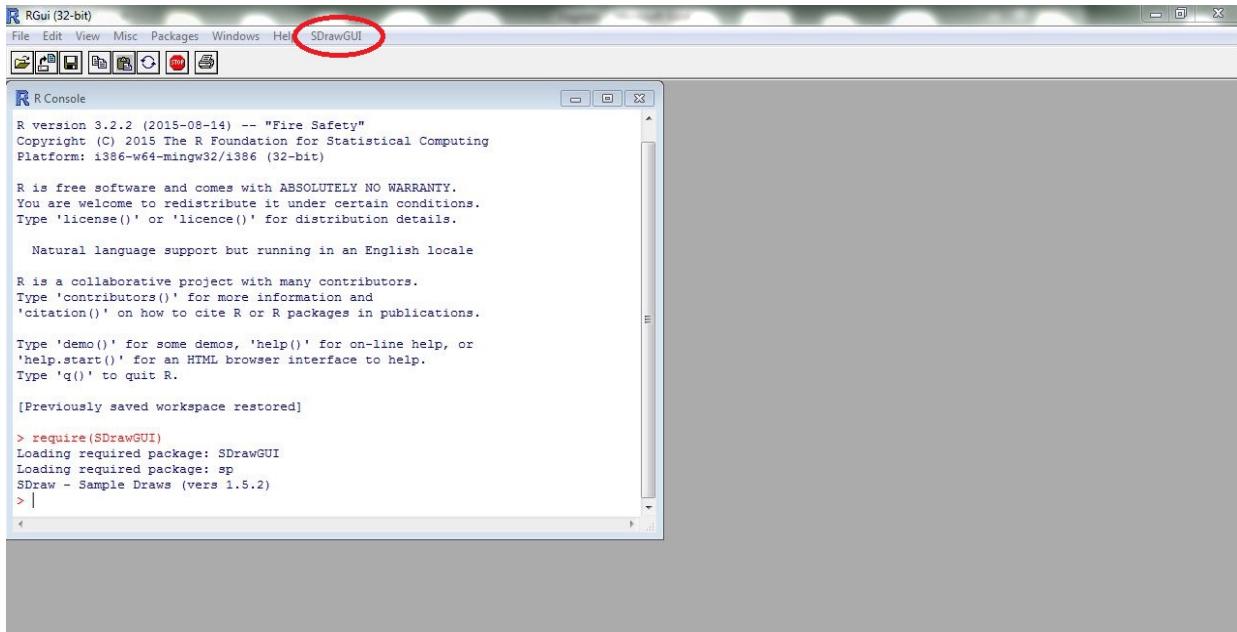


Figure 2. The R command window after the SDrawNPS package is loaded.

In the Sample Type box, select the drop-down menu option for “GRTS – Generalized Random Tessellation Stratified”. Assume that SRI resource managers have determined that a sample of 150 sites across this 53,000 acre island is sufficient for the survey design. Then, type in the desired sample size of 150. The next required input is the shapefile that contains the polygon boundaries. In this case, the shapefile is called “SRI_Watersheds_shapefile.shp” and was provided by NPS.

The analyst must also provide a name for the GRTS output and type it in the “Sample’s R name” box. A more informative name, such as “SRI_Polygon_Sample”, can be entered and the GRTS design file for the sample will be saved with this name. Lastly, the random number seed should be selected so that the sample can be redrawn if the GRTS design files are lost or corrupted. Selecting a random seed allows the user to obtain an identical draw when the same random seed is used. Archiving the random seed may be useful the sample must be redrawn because the original design files were lost or if more sites are needed. If the random seed box is left empty, a random seed will be automatically generated and archived with GRTS sample draw information. A random seed can be generated with the following R code in the R command window:

```

sample(100000000,1) # run to get initial random seed
[1] 26550867
set.seed(26550867)

```

An image of the inputs for this draw is shown in Figure 3.

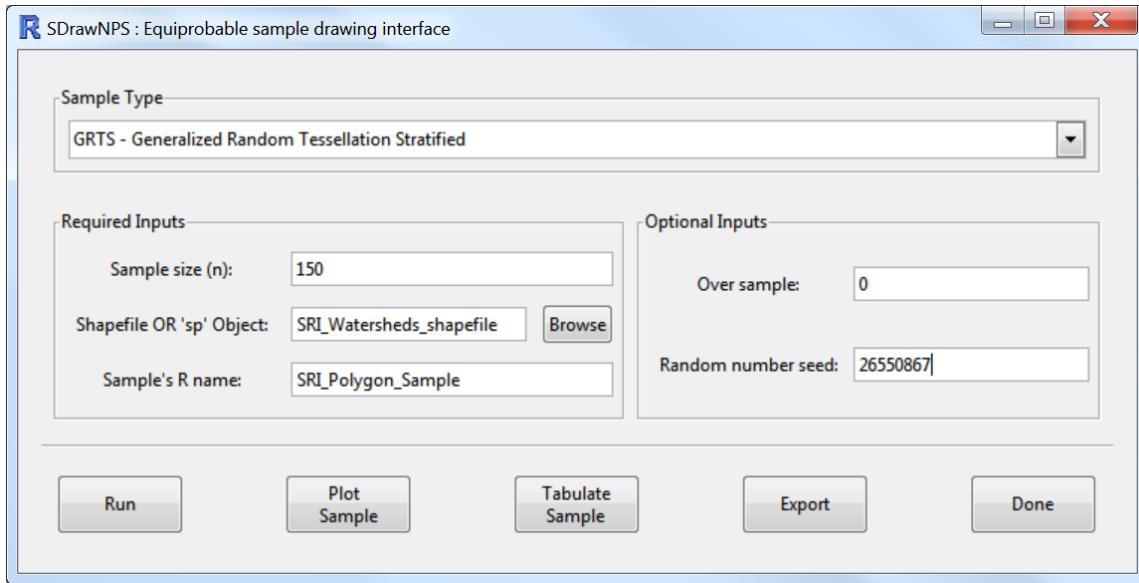


Figure 3. The SDrawNPS GUI for an equiprobable GRTS sample of 150 sites drawn from the SRI watersheds polygon sampling frame.

Now that all the required inputs and desired optional inputs have been specified, the SDrawNPS user can click the Run button to draw the equiprobable GRTS sample of 150 sites from the SRI watersheds polygon sampling frame. Assuming no errors occurred in the sample draw, the output in the R window will show the first 10 sample coordinates. Clicking the Plot Sample button will bring up a map showing the sample locations across the identified watersheds of Santa Rosa Island (Figure 4). The sample draw can be viewed in tabular form using the View Sample button, and the file can be exported as a spatial data file (e.g., ESRI shapefile, Google Earth KML, or Garmin GPX) or as a .csv using the Export button.

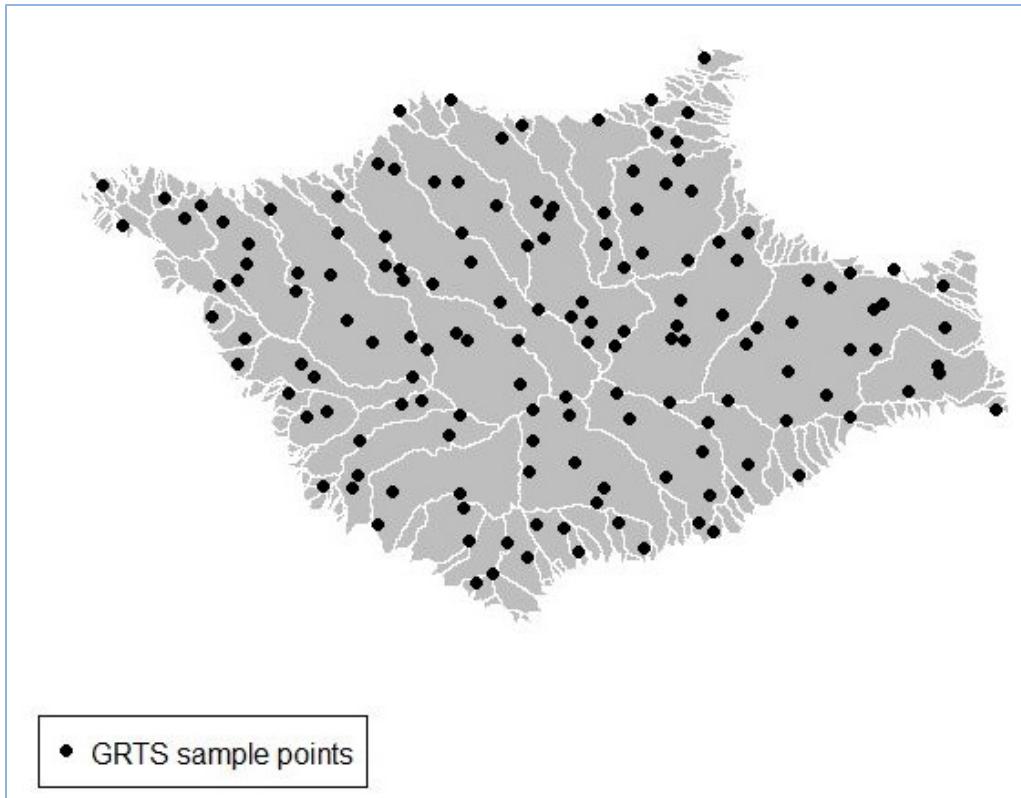


Figure 4. An example equiprobable GRTS sample of 150 sites drawn from the SRI watersheds polygon sampling frame.

Figure 4 shows that the equiprobable GRTS survey points cover the island and appear to exhibit some level of spatial balance. Many of the larger watersheds located in the central island region have multiple sample locations, while some of the smaller ones along the coast do not have any survey points. Because an equiprobable GRTS sample across a set of polygons is equivalent to a probability-proportional-to-size sample relative to watershed area, this result is expected given the small sizes of the coastal watersheds.

For comparison, the R code needed to draw the same sample directly from *spsurvey* is given below for two sampling frames, a shapefile named **SRI_Watersheds_shapefile** and an *sp* object named **SR.watershed**.

```
require(spsurvey)
require(SDrawNPS)

# Define aspects of the sampling design:
# sample size of 150 sites, equal selection, "None" is single stratum name
SRIWatershed_DesignSpecs_Equi <- list("None"=list(panel=c(Main=150),
seltype="Equal"))
```

```

set.seed(26550867)           # matches the seed in the SDrawNPS sample

# Obtain attribute information

att_SRIWatershed<- read.dbf('SRI_Watersheds_shapefile')

# For use with shapefile = SRI_Watersheds_shapefile

SRIWatershed_Equi_Sample_shp <- grts(design=SRIWatershed_DesignSpecs_Equi,
                                       DesignID='Site',
                                       type.frame='area',
                                       src.frame='shapefile',
                                       in.shape='SRI_Watersheds_shapefile',
                                       att.frame=att_SRIWatershed,
                                       out.shape='SRIWatershed_Equi_Sample')

# For use with sp.object = SR.watershed

# Read shapefile to sp.object

require(rgdal)

SR.watershed<-readOGR(".", "SRI_Watersheds_shapefile")

SRIWatershed_Equi_Sample_sp <- grts(design=SRIWatershed_DesignSpecs_Equi,
                                       DesignID='Site',
                                       type.frame='area',
                                       src.frame='sp.object',
                                       sp.object=SR.watershed,
                                       att.frame=SR.watershed@data,
                                       out.shape='SRIWatershed_Equi_Sample')

# Compare these samples with the one drawn in SDrawNPS:

SRIWatershed_Equi_Sample_sp[1:5,]

SRIWatershed_Equi_Sample_shp[1:5,]

SRI_Polygon_Sample[1:5,]

```

Stratified Sampling

In stratified sampling, independent sampling is conducted within levels of the stratification variable. The stratification variable is a categorical variable that may be of interest for inference, may provide more convenient sampling, or exhibits different variability within its levels. Stratified GRTS sampling is conducted for the monitoring design from Channel Islands National Park (CHIS).

Case Study: Polygon Frame, User Allocation

Suppose, as in the previous case study, the 199 watersheds on Santa Rosa Island were to be sampled but sampling effort was stratified according to watershed size. Specifically, there are 19 watersheds on Santa Rosa Island that are greater than 250 hectares (Large) and 180 watersheds that are smaller than 250 hectares (Small). A survey design that allocates sampling effort across these strata is a stratified design. The user can create a design with 150 sample sites selected, with 120 sites in the large watersheds and the remaining 30 sites in the small watersheds.

To create this sample design, the resource manager would first select the SDrawNPS dropdown menu list from his R window and choose “SDrawNPS/Sample Draws/Stratified...” The pop-up user interface window, shown below, looks a little different than the one that drew an equiprobable sample (Figure 5). The differences reflect the additional information required to draw a stratified sample using SDraw. Specifically, the user must supply the “Name of Stratifying Variable”, that is, the levels across which the sample will be allocated, as well as information necessary to allocate the sampling effort.

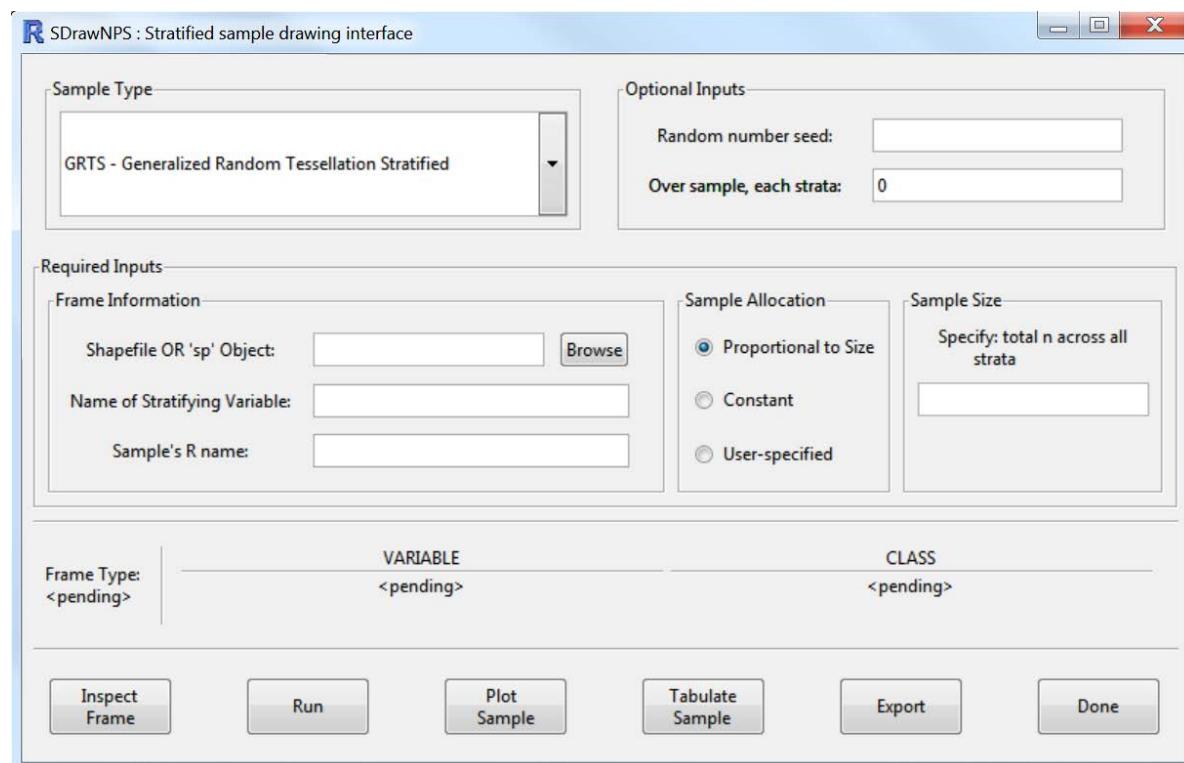


Figure 5. The SDrawNPS GUI for a stratified GRTS sample.

The first step is to supply the Required Inputs (the middle section of the GUI). To draw a stratified draw across Santa Rosa Island Large and Small watersheds, the watershed shapefile with the appropriate stratifying information should be selected. Once selected, one can preview the contents of that shapefile by pushing the “Inspect Frame” button (Figure 6). SDrawNPS reads the shapefile and returns information about the fields.

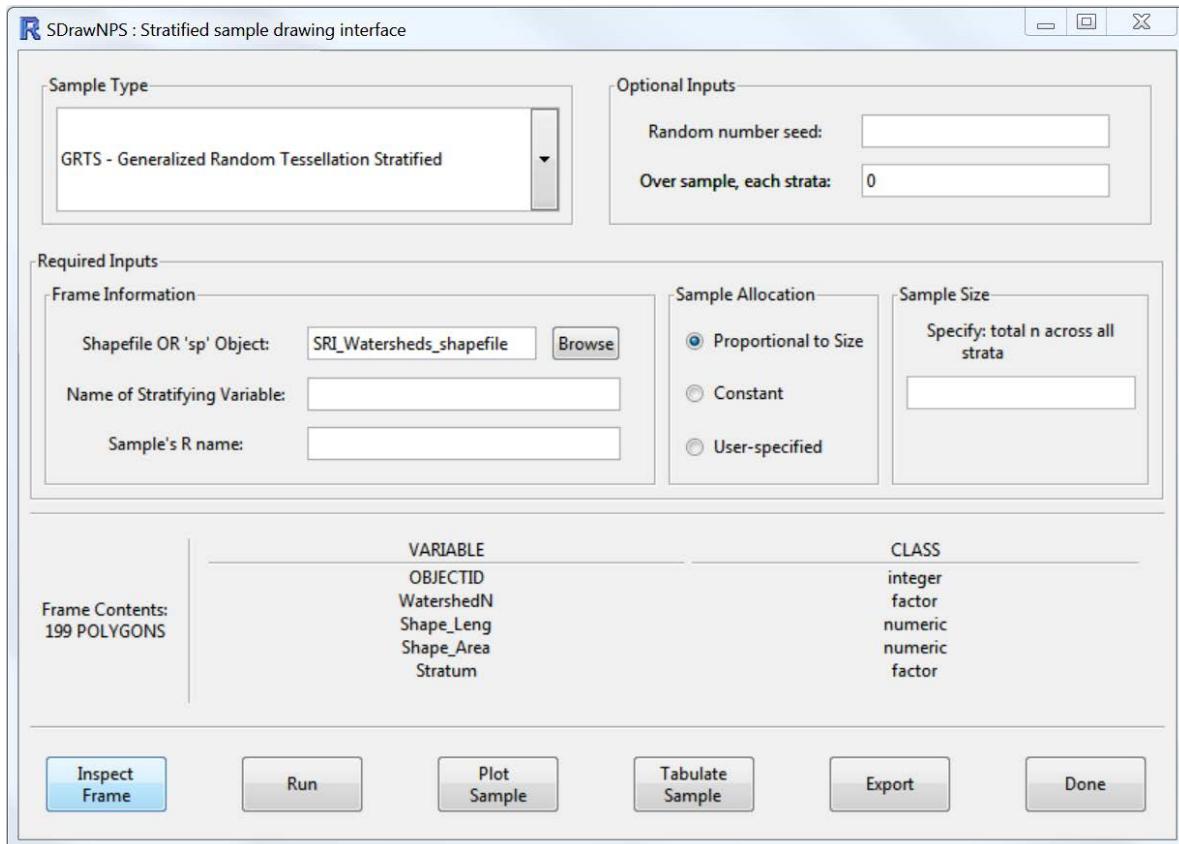


Figure 6. The SDrawNPS GUI for a stratified GRTS sample from the SRI watersheds shapefile.

The shapefile contains 199 polygons, and the variable names include: “OBJECTID”, “WatershedN”, “Shape_Leng”, “Shape_Area”, and “Stratum.” For this example, the stratifying variable is Stratum, so that should be entered in the Name of Stratifying Variable text box. Additionally, an appropriate name for the output file should be typed in the Sample’s R name text box.

The last required input relates to the sample size. In the previous example, there were 150 sample points cast across the entire island. In this example, we want to specify the sample size across the two strata: 120 in the Large stratum, 30 in the Small stratum. To do this, the User-specified radio button should be selected and this information must be entered in the Sample Size box. These numbers must be separated by a comma and listed in strata (alphabetical) order (Figure 7). When all the information is provided, push the Run button to complete the sample draw. The selected sample locations from this draw are shown in the next screenshot (Figure 8).

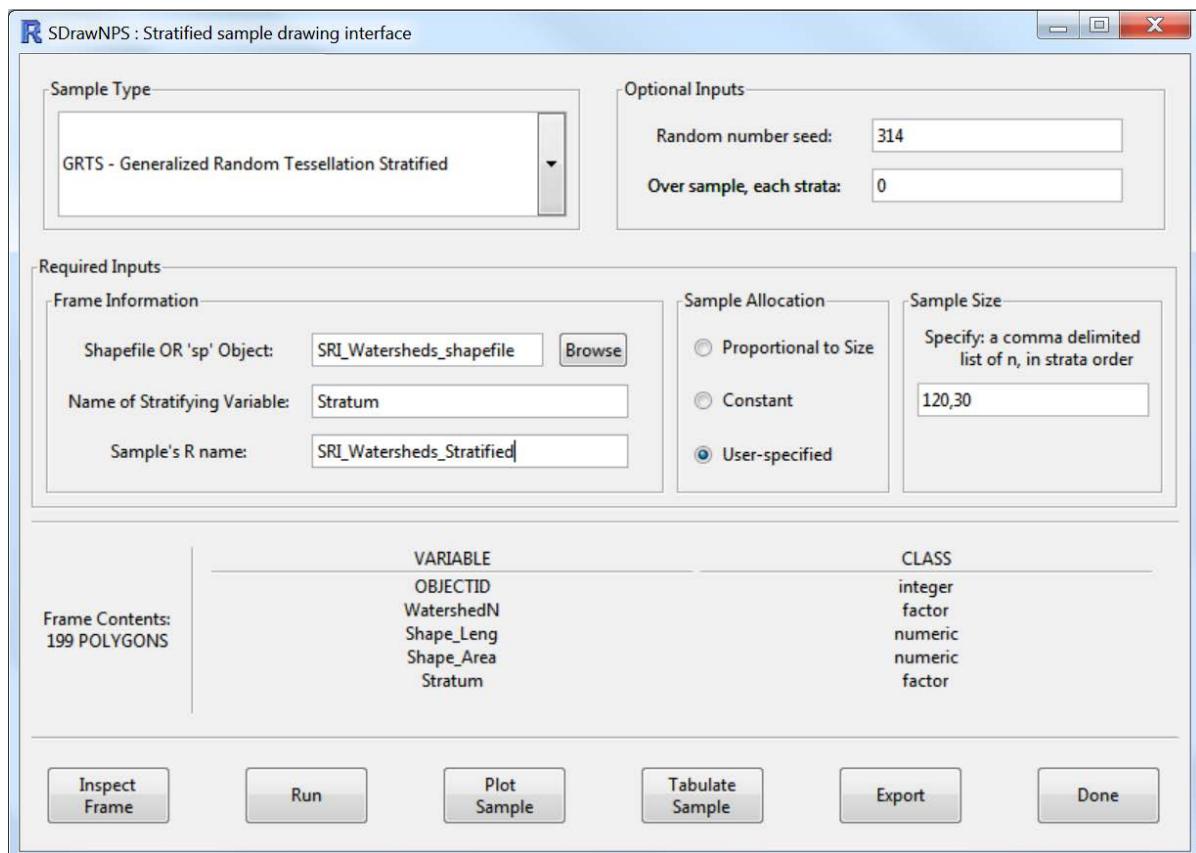


Figure 7. The SDrawNPS GUI for an stratified GRTS sample from the SRI watersheds shapefile, specifying 120 sites from the Large stratum and 30 sites from the Small stratum.

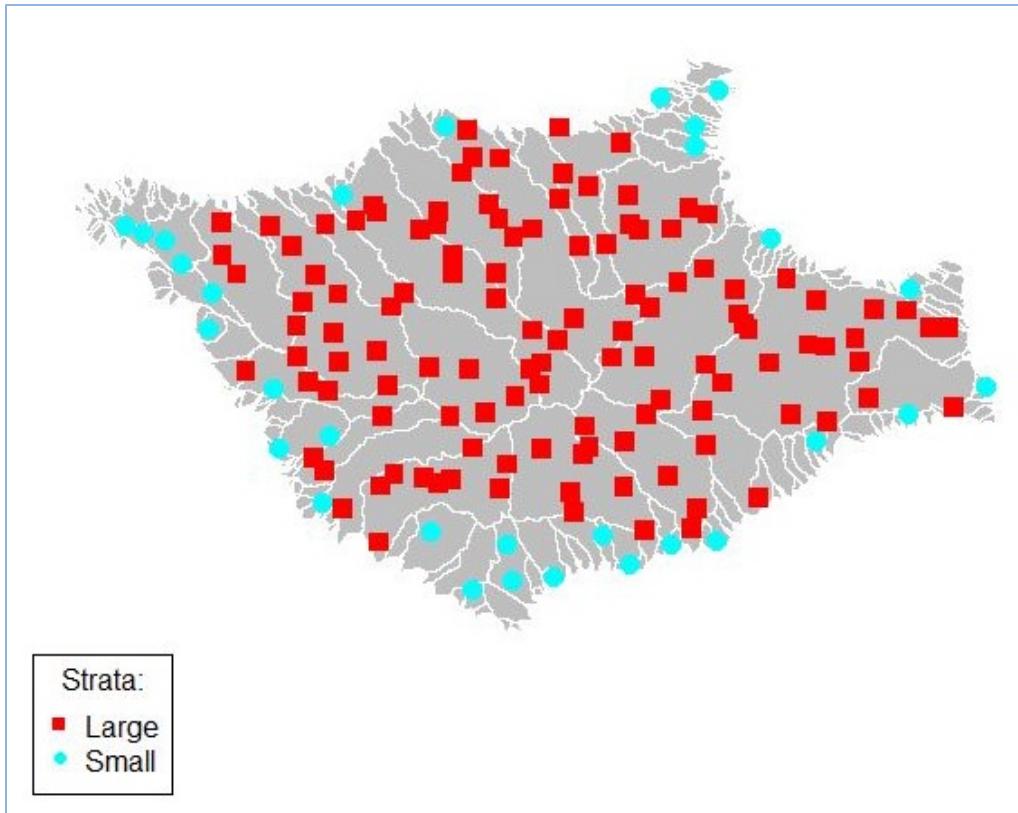


Figure 8. The SRI watersheds stratified GRTS sample of 120 Large-stratum sites and 30 Small-stratum sites.

Here, sample sites within the large watersheds are indicated by a red box, while the sample sites located within the small watersheds are indicated by blue circles.

For comparison, the R code needed to draw the same stratified GRTS sample directly from *spsurvey* is given below for two sampling frames, a shapefile named *SRI_Watersheds_shapefile* and an *sp* object named *SR.watershed*.

```
# Set sampling design
SRIWatershed_DesignSpecs_Strata<- list(
  "Large"=list(panel=c(Main=120), seltype="Equal"),
  "Small"=list(panel=c(Main=30), seltype="Equal"))

# For use with Shapefile
set.seed(314)
SRIWatershed_Strata_Sample_shp <-
  grts(design=SRIWatershed_DesignSpecs_Strata,
```

```

DesignID='Site',
type.frame='area',
src.frame='shapefile',
in.shape='SRI_Watersheds_shapefile',
stratum='Stratum',
att.frame=att_SRIWatershed,
out.shape='SRIWatershed_Equi_Sample_shp')

# For use with sp.object

set.seed(314) # need to set before every draw

SRIWatershed_Strata_Sample_sp <-
grts(design=SRIWatershed_DesignSpecs_Strata,
DesignID='Site',
type.frame='area',
src.frame='sp.object',
sp.object=SR.watershed,
stratum='Stratum',
att.frame=SR.watershed@data,
out.shape='SRIWatershed_Equi_Sample_sp')

# Compare draws

SRIWatershed_Strata_Sample_shp[1:5,]

SRIWatershed_Strata_Sample_sp[1:5,]

sdraw.2015.04.16.113011[1:5,]      # Note that this name will change when testing

```

The entire sample draw can be viewed using the View Sample button, which displays a table of all the points selected, with the XY coordinate information, the Stratum, and the WatershedN (name) (Figure 9). The information displayed on the left side of the table is specific to the sample draw, while the information on the right side comes from the attribute data of the shapefile.

	siteID	xcoord	ycoord	mdcaty	wgt	stratum	panel	EvalStatus	EvalReason	OBJECTID	WatershedN	Shape_Leng	Shape_Area
1	STRAT-001	771986.9	3764156	Equal	1263789	Large	PanelOne	NotEval		122	Water	25766	12196253
2	STRAT-002	766311.1	3764052	Equal	1263789	Large	PanelOne	NotEval		97	Dry	20558	6862212
3	STRAT-003	772051.6	3761548	Equal	1263789	Large	PanelOne	NotEval		139	Quemada	32718	18637558
4	STRAT-004	765695.4	3759276	Equal	1263789	Large	PanelOne	NotEval		169	Acapluco	19288	4746658
5	STRAT-005	771276	3763792	Equal	1263789	Large	PanelOne	NotEval		122	Water	25766	12196253
6	STRAT-006	763787.9	3763476	Equal	1263789	Large	PanelOne	NotEval		137	Arlington	32676	11816050
7	STRAT-007	770989.8	3758526	Equal	1263789	Large	PanelOne	NotEval		171	Wreck	19668	7418128
8	STRAT-008	767274	3761415	Equal	1263789	Large	PanelOne	NotEval		116	Soledad	33640	12258313
9	STRAT-009	769713.2	3767585	Equal	1263789	Large	PanelOne	NotEval		95	Lobo	20930	4703413
10	STRAT-010	775305.3	3762035	Equal	1263789	Large	PanelOne	NotEval		139	Quemada	32718	18637558

Figure 9. Partial screenshot of the SRI watersheds stratified GRTS sample of 120 Large-stratum sites and 30 Small-stratum sites.

Lastly, the data from the draw can be exported in a variety of formats using the Export button. Once pressed, the SDrawNPS package prompts the user for the file's name, destination, and the specific format. These data can be exported as an ESRI shapefile, a Google Earth .KML file, a Garmin .GPX file, or a comma delimited format (.CSV).

If the user doesn't have a specific sample allocation in mind, s/he can also select a constant or proportional sample size. For constant allocation, the user inputs the number of sites to be drawn from each stratum. The proportional allocation option, if selected, prompts for the total sample size across all strata (in our example, 150) and then allocates the sampling effort across strata based on the proportional area. Had the shapefile been stream segments, samples would have been allocated proportional to total length in each stratum. If they had been finite points, the sample size would have been allocated proportional to total number of points in each stratum. Note that, if the sample size is so small that sampling points are not allocated to a relatively-small stratum, then an error message (e.g. "The design list does not contain any valid values of the panel argument for stratum CR") indicating the unsampled stratum (in this example, "CR") is written to the console. In this case, a user-specified allocation of sample size is recommended. Note that at least 4 sites are needed to compute the neighborhood variance estimate at the stratum level (Stevens and Olsen 2003).

Unequal Probability Sampling

Oftentimes, a survey design will not require sample points allocated evenly across an entire resource, nor will it have any obvious stratification. Instead, a resource manager will want to allocate the sampling effort unequally across the resource, perhaps with sample sites allocated proportional to resource size or cost of sampling. In this case, an unequal probability GRTS design will likely provide an optimal design for sampling. Two case studies of the unequal probability design follow.

Case Study: Finite Frame, Sampling Proportional to a categorical attribute

An example of an unequal survey design based on a categorical attribute can be demonstrated using vegetated areas of Joshua Tree National Park (JOTR). JOTR, located in southern California, contains two distinct desert ecosystems, the Colorado and the Mojave deserts. The National Park Service has mapped within JOTR 4 distinct plant associations: *Coleogyne ramosissima*, *Juniperus californica/Prunus fasciculata*, *Larrea tridentata*, and the *Pleuraphis rigida* associations. These four

associations were provided for this example in an ESRI shapefile. In that shapefile, an attribute called ASSOC_CO (for association code) was created, and the following codes used: CR, JCPF, LT, and PR, respectively.

To draw a GRTS sample of points from the JOTR sampling frame using an unequal probability design, the user must first select the “SDraw/Sample Draws/Unequal...” menu option, bringing up this interface (Figure 10). This GUI appears very similar to the ones for the equiprobable and stratified draws, with fields to specify the necessary spatial data, sample allocation and size, and then to view the draw output.

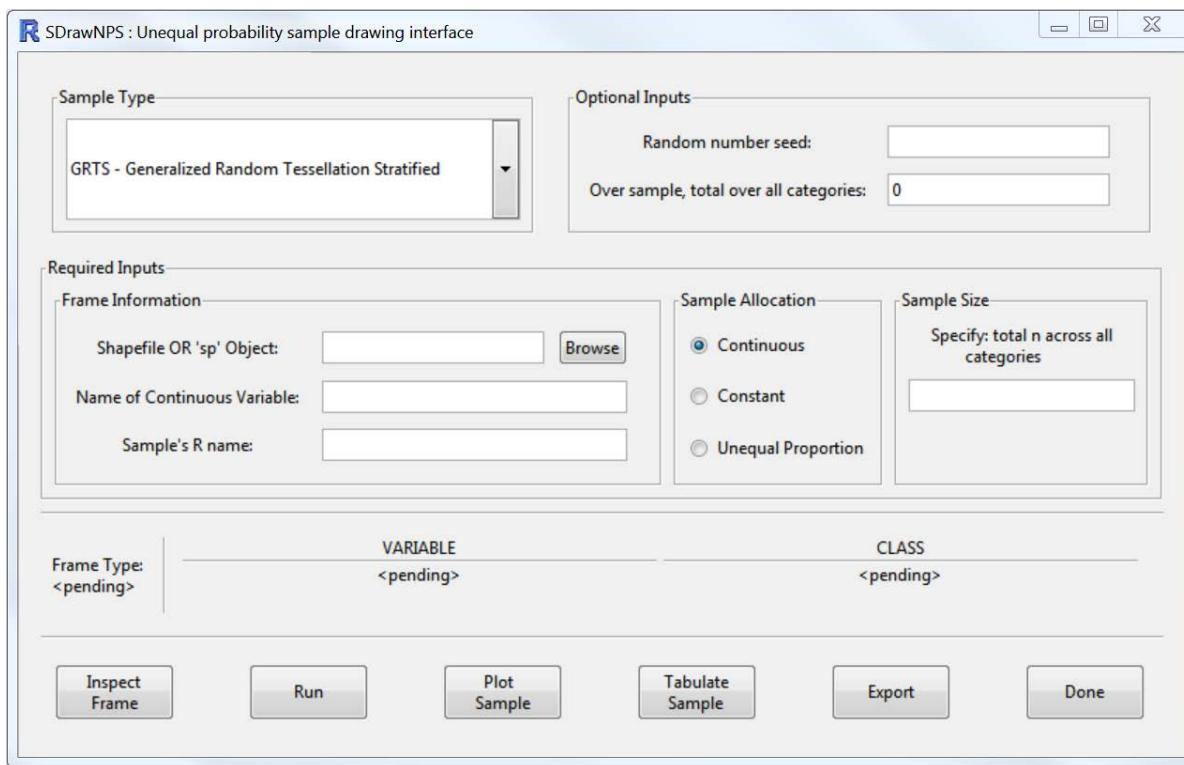


Figure 10. The SDrawNPS GUI for drawing an unequal probability GRTS sample.

In the Sample Type box, select the drop-down menu option for “GRTS – Generalized Random Tessellation Stratified”. The next step is to supply the Required Inputs (the middle section of the GUI), such as the sampling frame, variable for unequal probability weighting, and the name of the sample draw output file. The appropriate shapefile, in this case, is the shapefile is called JOTR_INUP_YUBR_Final.” Use the “Browse” button to navigate to the directory that contains that shapefile, which will be listed in the “Shape file OR ‘sp’ Object” box after selection. Then the field names of the shapefile attribute table may be viewed by clicking the “Inspect Frame” button (Figure 10). From the resulting list of variable names, the appropriate variable to be used for unequal probability weighting is entered in the “Name of Categorical Variable” box. For this example, unequal probability sampling will be conducted relative to the categorical variable indicating the plant association, “ASSOC_CO”. Finally, an appropriate sample draw output name should be created

and entered in the “Sample’s R name” box. For this case study, the output file will be called “JOTR_unequal_proportional” to indicate that the unequal probability design with a categorical attribute was used.

Next, the Sample Allocation and Sample Size information must be entered. For this example, the sample size is allocated relative to a categorical variable (ASSOC_CO), so the “Unequal Proportion” radio button should be selected. The Sample Size box includes text asking the user to specify a comma delimited list of n , in alphabetical categorical order. Since the four levels in this case study are CR, JCPF, LT, and PR, we can draw 5 sites from the CR category and 10 from each of the other three categories by typing “5, 10, 10, 10” into the Sample Size box. The last bit of information to supply (if necessary) is the “Random number seed.” This must be provided to recreate the same draw at another time. The GUI interface with the necessary information for an unequal probability draw using a categorical variable to allocate sampling effort is shown in Figure 11. Clicking the Run button will begin the sample draw.

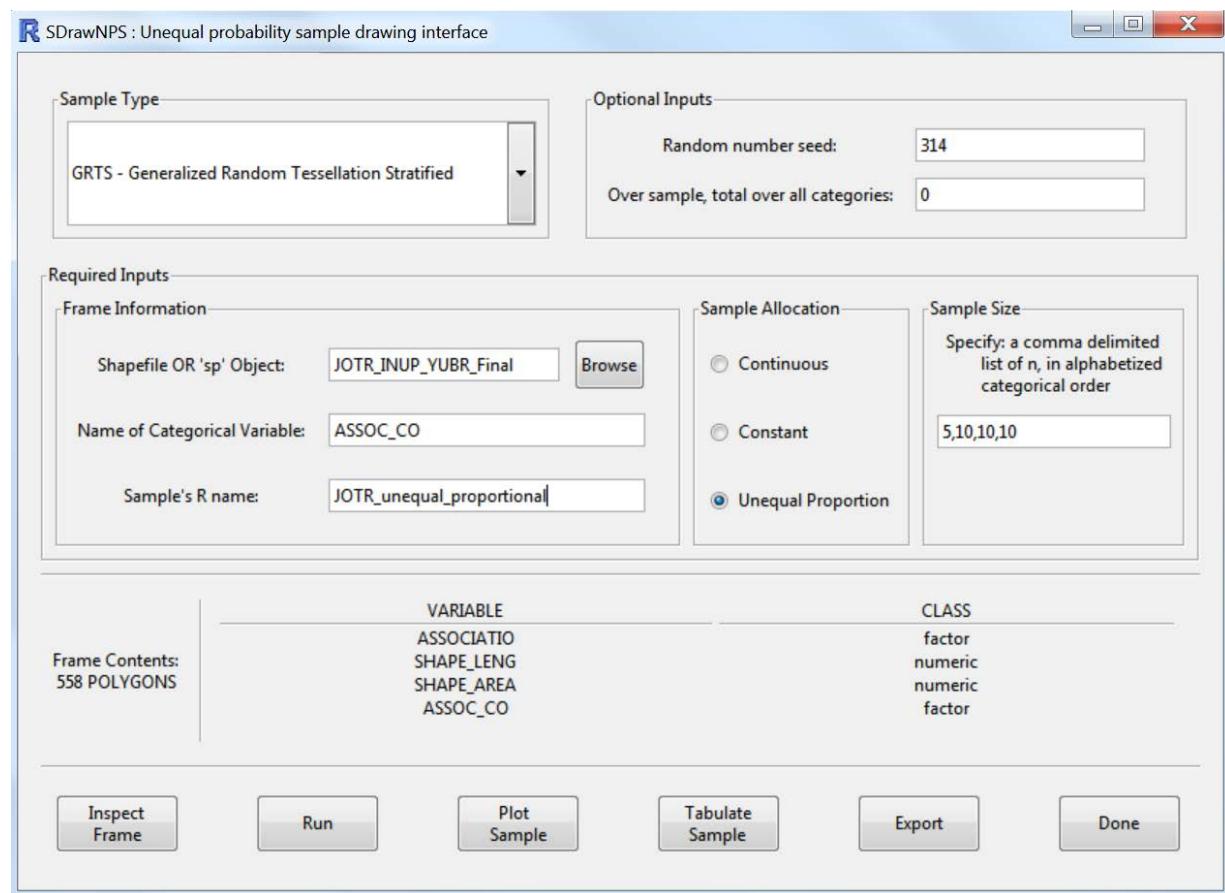


Figure 11. The SDrawNPS GUI and necessary information for drawing an unequal probability GRTS sample of vegetated areas in JOTR.

When the draw is finished, the complete sample draw can be visualized in the R window by pressing the Plot Sample button (Figure 12). Figure 12 shows in the legend the symbols used for vegetation

classes. Using the “Tabulate Sample” button will bring up the sample information in a table (Figure 13).

Lastly, the sample draw information, called the GRTS design file, can be exported in a variety of formats using the Export button. Once pressed, the SDrawNPS package prompts the user for the file name, destination, and the specific format. These data can be exported as an ESRI shapefile, a Google Earth .KML file, a Garmin .GPX file, or a comma delimited format (.CSV).

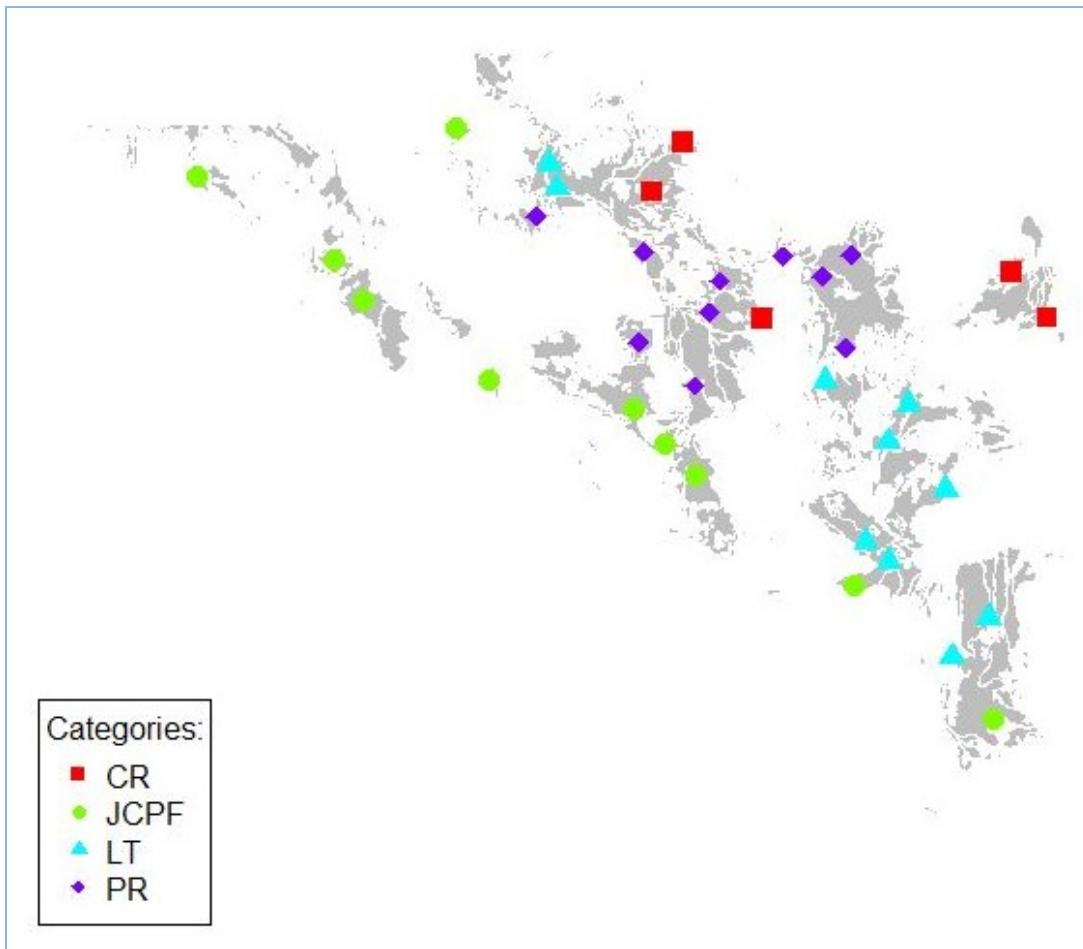


Figure 12. The unequal probability GRTS sample across four vegetation associations within JOTR.

R JOTR_unequal_proportional

siteID	xcoord	ycoord	mdcaty	wgt	stratum	panel	EvalStatus	EvalReason	ASSOCIATION
1 Site-01	556545.2	3768648	JCPF	4968588	None	PanelOne	NotEval		YUBR-JUCA/EPNE Association YUBR/PRFA Association
2 Site-02	570241.9	3767029	PR	5194779	None	PanelOne	NotEval		YUBR/PLRI Association
3 Site-03	576644	3756618	JCPF	4968588	None	PanelOne	NotEval		YUBR-JUCA/EPNE Association YUBR/PRFA Association
4 Site-04	581798.5	3764608	PR	5194779	None	PanelOne	NotEval		YUBR/PLRI Association
5 Site-05	587010.6	3749280	LT	5621028	None	PanelOne	NotEval		YUBR/LATR-YUSC/PLRI Association
6 Site-06	580190.3	3765421	PR	5194779	None	PanelOne	NotEval		YUBR/PLRI Association
7 Site-07	576650.6	3760204	PR	5194779	None	PanelOne	NotEval		YUBR/PLRI Association
8 Site-08	586764	3756044	LT	5621028	None	PanelOne	NotEval		YUBR/LATR-YUSC/PLRI Association
9 Site-09	583061.7	3752183	JCPF	4968588	None	PanelOne	NotEval		YUBR-JUCA/EPNE Association YUBR/PRFA Association
10 Site-10	574908.6	3768010	CR	1308072	None	PanelOne	NotEval		YUBR/CORA Association
11 Site-11	574368.8	3761948	PR	5194779	None	PanelOne	NotEval		YUBR/PLRI Association
12 Site-12	584457.3	3753117	LT	5621028	None	PanelOne	NotEval		YUBR/LATR-YUSC/PLRI Association
13 Site-13	562099.3	3765278	JCPF	4968588	None	PanelOne	NotEval		YUBR-JUCA/EPNE Association YUBR/PRFA Association
14 Site-14	570755.5	3769127	LT	5621028	None	PanelOne	NotEval		YUBR/LATR-YUSC/PLRI Association
15 Site-15	589402.1	3764782	CR	1308072	None	PanelOne	NotEval		YUBR/CORA Association
16 Site-16	582732.7	3761734	PR	5194779	None	PanelOne	NotEval		YUBR/PLRI Association
17 Site-17	567017.1	3770574	JCPF	4968588	None	PanelOne	NotEval		YUBR-JUCA/EPNE Association YUBR/PRFA Association
18 Site-18	577238.6	3763162	PR	5194779	None	PanelOne	NotEval		YUBR/PLRI Association
19 Site-19	577652.2	3764427	PR	5194779	None	PanelOne	NotEval		YUBR/PLRI Association

Figure 13. Output table showing the sample draw from the unequal probability draw from JOTR vegetation associations. The ‘mdcaty’ field shows the four levels of the categorical attribute (CR, JCPF, LT, PR), and the ‘wgt’ field shows the resulting weight of each level.

In this example, the CR habitat is small compared to the other three associations. One advantage gained from using the unequal probability design over an equiprobable design is that sites in the CR association can be specified to have a higher inclusion probability in the sample draw than if an equiprobable sample was drawn. Specifying a target sample size does not guarantee that sample size within each category as with stratification, but category-level sample sizes allow the analyst to directly specify the inclusion probability of sites in the CR habitat. For comparison, the R code needed to draw the same unequal probability GRTS sample directly from *spsurvey* is given below for two sampling frames, a shapefile named JOTR_INUP_YUBR_Final and an *sp* object named JOTR_INUP_YUBR.

```
# Set sampling design
JOTR_IntUp_DesignSpecs<- list(None=list(panel=c(Main=35),
                                         seltype="Unequal",
                                         caty.n=c('CR'=5, 'JCPF'=10, 'LT'=10, 'PR'=10), over=0))
```

```
att <- read.dbf("JOTR_INUP_YUBR_Final")
```

```
# For use with Shapefile
set.seed(314)
JOTR_IntUp_GRTS <- grts(design=JOTR_IntUp_DesignSpecs,
                         DesignID='MOJN JOTR YUBR',
                         type.frame='area',
```

```

src.frame='shapefile',
in.shape='JOTR_INUP_YUBR_Final',
att.frame=att,
mdcaty="ASSOC_CO",
prjfilename='JOTR_INUP_YUBR_Final',
out.shape='JOTR_YUBR_IntUp_GRTS_sample')

# For use with sp.object

set.seed(314) # need to set before every draw

JOTR_IntUp_GRTS <- grts(design=JOTR_IntUp_DesignSpecs,
                          DesignID='MOJN JOTR YUBR',
                          type.frame='area',
                          src.frame='sp.object',
                          sp.object='JOTR_INUP_YUBR',
                          att.frame=JOTR_INUP_YUBR@data,
                          mdcaty="ASSOC_CO",
                          out.shape='JOTR_YUBR_IntUp_GRTS_sample')

```

Continuous Probability Sampling

Case Study: Linear Frame, Sampling Proportional to a continuous attribute

A resource management objective may require a GRTS sample for a linear resource (for example, stream channels), and often the samples must be allocated unequally across those linear features. Using the tools in the spsurvey package, drawing a GRTS sample with the SDrawNPS package may be designed to target specific features of interest while giving lower probability to features that are of lesser interest, are less variable, or are not expected to change greatly over time. The streams of Santa Rosa Island provide a useful case study for this type of draw.

NPS provided a GIS shapefile for mapped streams on Santa Rosa Island. The shapefile contains 434 linear features ranging in size from less than 1m to 4,100 meters in length. In this example, the survey design requires that smaller streams on the island are sampled with greater intensity. This can easily be done using an unequal probability survey design and letting the sampling weights vary inversely with stream length; that is, shorter streams would have a greater chance of inclusion in our survey design than longer ones.

To draw this sample, the resource manager would first select the SDrawNPS dropdown menu list from his R window and choose “SDraw/Sample Draws/Unequal...” The user interface that appears is seen in Figure 14.

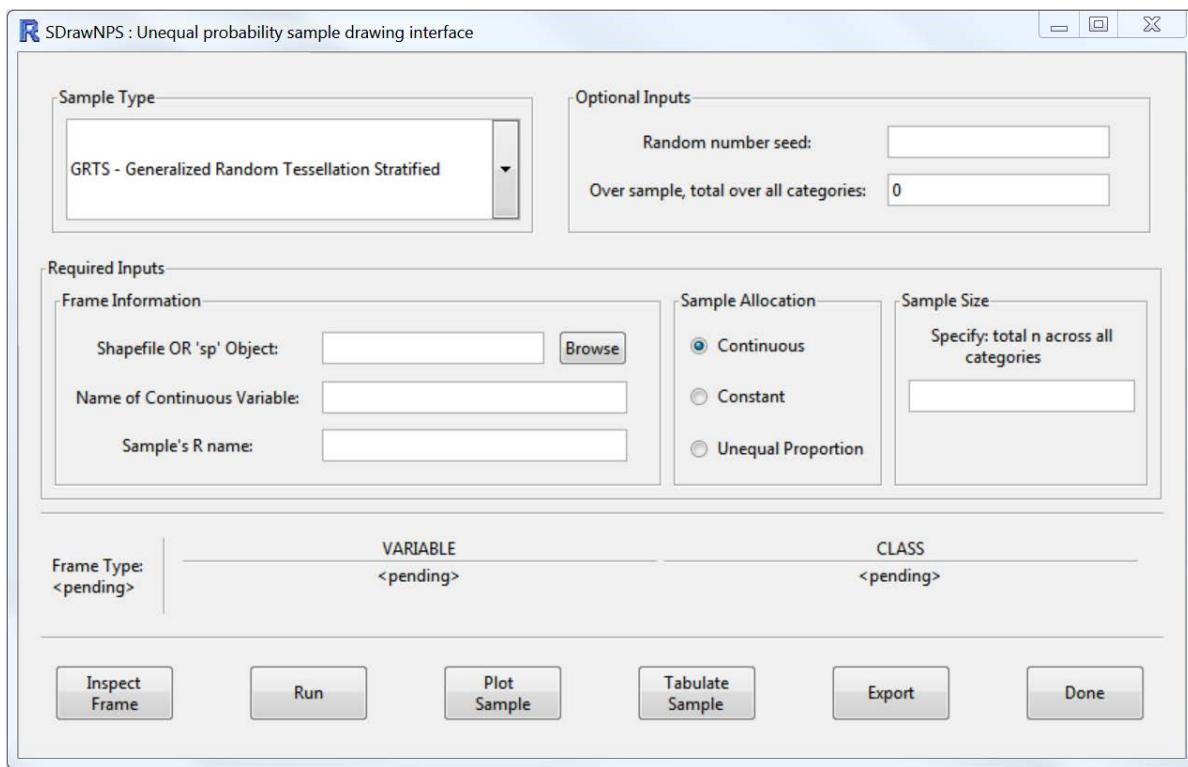


Figure 14. The SDrawNPS GUI for drawing an unequal probability GRTS sample.

In the Sample Type box, select the drop-down menu option for “GRTS – Generalized Random Tessellation Stratified”. The next step is to supply the Required Inputs (the middle section of the GUI), such as the sampling frame, variable for unequal probability weighting, and the name of the sample draw output. To draw an unequal probability sample across Santa Rosa Island mapped streams, the stream shapefile must be specified. In this case, the appropriate shapefile is called “SRI_Stream_Derived_Lengths_shapefile.” Use the “Browse” button to navigate to the directory that contains that shapefile and select it, which will be listed in the “Shape file OR ‘sp’ Object” box. The field names within the shapefile can be previewed by pushing the “Inspect Frame” button (Figure 15). SDrawNPS reads the shapefile and returns information about the fields. In this example, “SHAPE_Leng” is the length (in m) for each of the 434 mapped segments, and another variable (“Inv_Leng”) is calculated as 1/SHAPE_Leng. Inv_Leng can be used to sample shorter segments in this example, so that variable should be entered into the “Name of Continuous Variable” box. Finally, an appropriate output name should be created and entered in the “Sample’s R name” box. For this case study, the output file will be called “SRI_streams_unequal_continuous” to indicate that the unequal probability design using with continuous attribute was used.

Next, the Sample Allocation and Sample Size information must be entered. For this example, the samples are allocated based on a continuous variable (inverse stream length), so the Continuous radio button should be selected. The Continuous button is the default when the GUI first opens, but the user should ensure that it is selected. The Sample Size box includes text asking the user to specify the total n across all categories. For this example, a sample size of 30 is desired. The last bit of information to supply (if necessary) is the Random number seed which must be provided and used to recreate the same draw at another time. The GUI interface with the necessary information for an unequal probability draw using a continuous variable to allocate sample size is shown in Figure 15. Clicking the Run button will begin the sample draw.

The complete sample draw can be visualized in the R window by pressing the Plot Sample button, and the following representation appears (Figure 16). Figure 16 shows in the legend the symbols used for stream segments of different inverse lengths, and the distribution of samples demonstrates the sample weighting in the draw.

The screenshot shows the SDrawNPS GUI window titled "SDrawNPS : Unequal probability sample drawing interface".

- Sample Type:** GRTS - Generalized Random Tessellation Stratified
- Optional Inputs:**
 - Random number seed: 314
 - Over sample, total over all categories: 0
- Required Inputs:**
 - Frame Information:**
 - Shapefile OR 'sp' Object: SRI_Stream_Derived_Lengths, Browse
 - Name of Continuous Variable: Inv_Leng
 - Sample's R name: SRI_streams_unequal_continuous
 - Sample Allocation:** Continuous (radio button selected)
 - Sample Size:** Specify: total n across continuous variable range, 30
- Frame Contents:** 434 LINES

VARIABLE	CLASS
OBJECTID	integer
Notes	factor
SHAPE_Leng	numeric
created_us	factor
created_da	factor
last_edite	factor
last_edi_1	factor
Stratum	factor
Inv_Leng	numeric
- Buttons:** Inspect Frame, Run, Plot Sample, Tabulate Sample, Export, Done

Figure 15. The SDrawNPS GUI and necessary information for drawing an unequal probability GRTS sample of the streams on Santa Rosa Island using inverse stream length as the weighting variable.

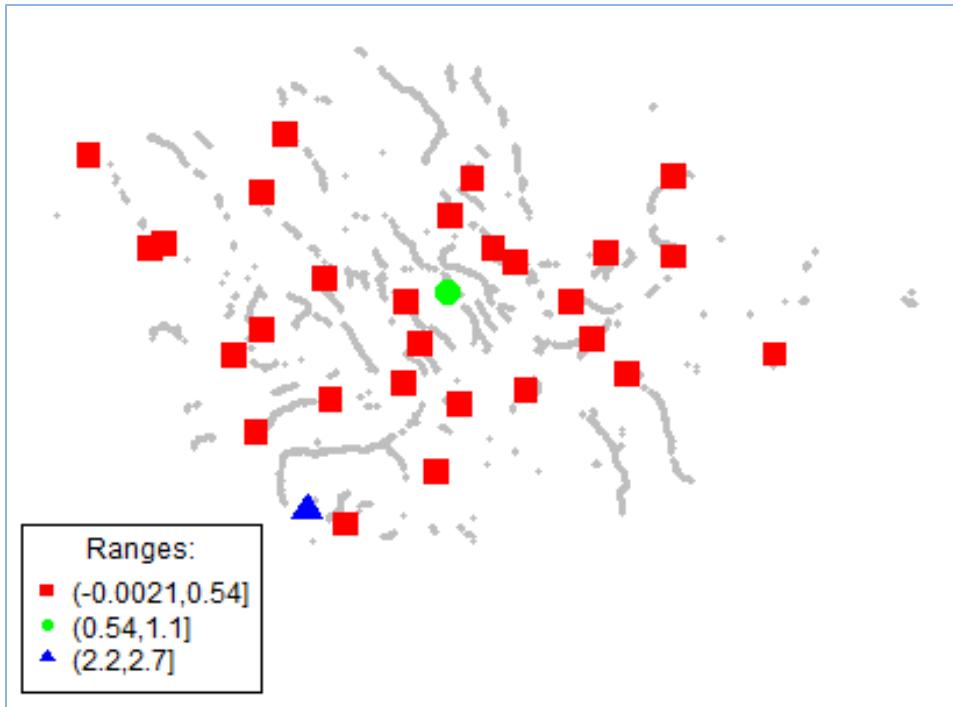


Figure 16. The unequal probability GRTS sample of 30 stream locations on Santa Rosa Island.

The sample draw can also be examined using the Tabulate Sample button (output shown in Figure 17). The ‘mdcaty’ field is the stream length (in m) from the shapefile, and the ‘wgt’ field indicates the *spsurvey*-calculated weight of that segment used in the sample draw.

	siteID	xcoord	ycoord	mdcaty	wgt	stratum	panel	EvalStatus	EvalReason
1	Site-01	764558.7	3756323	0.02334693	619.639	None	PanelOne	NotEval	
2	Site-02	762431.1	3761306	0.0006260728	23107	None	PanelOne	NotEval	
3	Site-03	767834.7	3765219	0.01536876	941.3032	None	PanelOne	NotEval	
4	Site-04	771797.1	3760180	0.04108651	352.1026	None	PanelOne	NotEval	
5	Site-05	763570.9	3756686	2.705421	5.347288	None	PanelOne	NotEval	
6	Site-06	766503.8	3760960	0.1008399	143.4617	None	PanelOne	NotEval	
7	Site-07	771257.6	3763295	0.08680045	166.6658	None	PanelOne	NotEval	
8	Site-08	768366.2	3763431	0.00116782	12387.75	None	PanelOne	NotEval	
9	Site-09	767506.9	3759425	0.1499794	96.45766	None	PanelOne	NotEval	
10	Site-10	767157	3762288	0.7926545	18.25091	None	PanelOne	NotEval	
11	Site-11	769213.2	3759767	0.1162266	124.4695	None	PanelOne	NotEval	
12	Site-12	757962.3	3765801	0.05858657	246.928	None	PanelOne	NotEval	
13	Site-13	766076.3	3759962	0.07180573	201.4695	None	PanelOne	NotEval	
14	Site-14	762408	3764861	0.01277391	1132.517	None	PanelOne	NotEval	
15	Site-15	770919.7	3761060	0.007888559	1833.879	None	PanelOne	NotEval	
16	Site-16	761717.6	3760649	0.08772372	164.9117	None	PanelOne	NotEval	
17	Site-17	775613.4	3760689	0.01350298	1071.369	None	PanelOne	NotEval	
18	Site-18	773026.8	3763209	0.2024388	71.46194	None	PanelOne	NotEval	
19	Site-19	766929.2	3757662	0.02613428	553.5514	None	PanelOne	NotEval	

Figure 17. Partial table showing the output sample draw from the unequal probability draw of Santa Rosa Island stream segments.

For comparison, the R code needed to draw the same unequal probability GRTS sample directly from *spsurvey* is given below for a shapefile named SRI_Stream_Derived_Lengths and an *sp* object named SRI_Streams.

```
# Set sampling design

SRI_Streams_DesignSpecs<- list(None=list(panel=c(PanelOne=30),
                                         seltype="Continuous", over=0))

att <- read.dbf("SRI_Stream_Derived_Lengths")

# For use with shapefile

set.seed(314)

SRI_IntUp_GRTS <- grts(design=SRI_Streams_DesignSpecs,
                        DesignID='SITE',
                        type.frame='linear',
                        src.frame='shapefile',
                        in.shape='SRI_Streams_Derived_Lengths',
                        att.frame=att,
                        mdcaty="Leng_Inv",
                        shapefile=FALSE) # will not automatically output shapefile

# For use with sp.object

set.seed(314) # need to set before every draw

SRI_IntUp_GRTS <- grts(design=SRI_Streams_DesignSpecs,
                        DesignID='SITE',
                        type.frame='linear',
                        src.frame='sp.object',
                        sp.object='SRI_Streams',
                        att.frame=SRI_Streams@data,
                        mdcaty="Leng_Inv",
                        out.shape='SRI_Streams_sample')
```


Estimation

After the data have been collected and reviewed for accuracy, status estimates are obtained with design-based tools available in the spsurvey package and accessed through the SDrawNPS package. Typically, the tools available in the spsurvey package are used to estimate the status of an ecological resource within a single year, although multiple years of data may contribute to a status estimate if assumptions on the behavior of the outcome over time are made. In this manual, we assume that a single year or season of survey data contribute to an estimate of resource status which may be measured as a mean, median, percentile, proportion falling above or below a given threshold, or cumulative distribution function (CDF). For GRTS sampling, the neighborhood variance estimator (Stevens and Olsen 2003, 2004) provides the basis for confidence interval construction. Status estimates may be obtained for levels of design strata and domains of interest not incorporated *a priori* in the sampling design.

Formatting the GRTS Design File

The GRTS design file is created when the sample is drawn and contains the list of sampled sites in spatially-balanced order, the initial design weights, site coordinates and other fields when the GRTS draw is initially conducted. A blank field called “EvalStatus” is used to store the survey evaluation status for each site after the survey is conducted. This information is used to adjust the design weights for nonsampling error such as frame error and nonresponse error.

To use the weighting adjustment option in the SDrawNPS analysis tool, the EvalStatus field must be formatted specifically as outlined below. The EvalStatus field takes on the values listed in Table 1.

Table 1. Post-survey coding for EvalStatus field in GRTS design file.

EvalStatus	Description
"Target - Surveyed"	The site was assessed and found to be a member of the target population. Data were successfully collected.
"Target - Not Surveyed"	The site was assessed and found to be a member of the target population. Data were not collected due to an issue such as inaccessibility or landowner denial of access.
"Non-Target"	The site was not a member of the target population.
"Not Evaluated"	The site was not evaluated or sampled. This could occur when the sampling season is ending and there is not enough time to visit all sites. This would also be the appropriate survey evaluation for a site in the sample but not surveyed in the given year.

Appropriate coding of the EvalStatus field is required for accurate weighting adjustment in the SDrawNPS package (Figure 18). More detail on why a site was found to be non-target or was non-responding may be entered in the EvalReason field. This additional information may be useful for assessing frame error or minimizing future nonresponse. Note that the survey evaluation status for each site will often differ from year to year. For example, if a site is in a panel that is not visited annually, that site will be classified as “Not Evaluated” in years that it is not surveyed. Nonsampling error may vary from year to year. For example, a water-quality site that is found to be a “Target –

Surveyed” site in one year may be classified as “Non-Target” during a dry year when no water is present at the site.

The importance of archiving the original sampling frame, the R code used to draw the GRTS sample, and the resulting GRTS design file containing spatial coordinates, sampling weights, and original siteID names cannot be overemphasized. These three pieces may be needed in the future for updating the sampling frame, revising the sampling design, changing the sample size, and making comparisons to other samples. Sites may become inaccessible over time or no longer be a part of the target population. Having the original GRTS sampling design products provides a basis from which new sites may be added and spatial balance maintained. Furthermore, an archive of all sampling design discussions provides a record for the motivations behind each decision and may be valuable to future resource managers.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	siteID	xcoord	ycoord	mdcaty	wgt	panel	EvalStatus	EvalReason	Mean Peri	Mean Can	Mean Bas	Mean Can	Mean Bas	Gap	Percent		
22	Site-021	584528	3753115	LT	5621028	Main	Target - Surveyed		31.3	7.3	38.9	84.5	99.9				
23	Site-022	589569	3764521	CR	1308072	Main	Target - Surveyed		28	2.4	7.6	70.9	97.7				
24	Site-023	575352	3760216	JCPF	4968588	Main	Non-Target	Vegetation mismatch; No associations present within transect area.									
25	Site-024	578788	3765782	PR	5194779	Main	Target - Not Sampled	Cultural resources									
26	Site-025	588656	3751664	LT	5621028	Main	Target - Surveyed		32.7	2.3	5.7	82	96.4				
27	Site-026	582096	3765019	PR	5194779	Main	Target - Surveyed		22	4.2	9.8	92.6	99.2				
28	Site-027	577184	3755786	JCPF	4968588	Main	Target - Surveyed		60	2	3.8	57.3	94.9				
29	Site-028	570105	3761923	JCPF	4968588	Main	Non-Target	Vegetation mismatch; None of the target associated species are present; likely burned and not tollerant to fire.									
30	Site-029	590330	3765523	LT	5621028	Main	Target - Not Sampled	Multi-terrain Types; Presence of roads									
31	Site-030	582551	3762148	PR	5194779	Main	Target - Surveyed		34	5.8	10.9	91.2	99				
32	Site-031	573307	3769256	PR	5194779	Main	Target - Surveyed		26.7	6	9.4	96.6	99.6				
33	Site-032	561771	3767284	CR	1308072	Main	Non-Target	Vegetation mismatch; Target community <80% of macroplot									
34	Site-033	588271	3748371	LT	5621028	Main	Target - Surveyed		48	1.6	4.8	63.5	98.5				
35	Site-034	585726	3758331	LT	5621028	Main	Target - Surveyed		40	2.1	3.6	75.5	94.2				
36	Site-035	578300	3764743	PR	5194779	Main	Target - Surveyed		26.7	3.7	6.6	88.1	98.3				
37	Site-036	576571	3767890	PR	5194779	OverSamp	Target - Not Sampled	Inaccesible; Part of plot blocked by fences									
38	Site-037	585879	3751516	CR	1308072	OverSamp	Target - Surveyed		32.7	1.3	3.9	69.6	97.5				
39	Site-038	590624	3764770	CR	1308072	OverSamp	Target - Surveyed		40.7	3.5	7.6	87.2	95.8				
40	Site-039	573989	3761055	JCPF	4968588	OverSamp	Target - Surveyed		32.7	3.5	5.7	88.3	97.5				
41	Site-040	561942	3765029	JCPF	4968588	OverSamp	Target - Not Sampled	Too close to trail									
42	Site-041	588074	3750368	LT	5621028	OverSamp	Target - Surveyed		25.3	2.6	5.7	80.8	98.5				
43	Site-042	583265	3763626	PR	5194779	OverSamp	Target - Surveyed		41.3	3.1	3.8	88.6	92.5				

Figure 18. Example of a GRTS design file with updated EvalStatus and EvalReason fields and columns added for data collected at each site.

Adjusting Design Weights

When the sample size for a given year is not equal to the sample size in the original GRTS sample draw, the design weights must be adjusted for analysis. This includes samples that are smaller or larger than planned, instances where an oversample is used, samples impacted by frame error or nonresponse error, and large samples that are allocated across panels for visits in different years. In these cases, the design weight provided in the GRTS design file must be adjusted to account for the obtained sample size.

For example, assume that the size of the resource is 20,000 units. These units may represent stream miles, area, or pixels in a sampling frame. An equiprobable GRTS sample of 100 sites is drawn and allocated to 4 panels for a sample of 25 sites within any given year. The original design weights for

all units i in the equiprobable sample, calculated as $w_i = \frac{N}{100} = 200$, would be adjusted as follows:

$$w_i^* = \left(\frac{n_{ALL}}{n} \right) w_i = \left(\frac{100}{25} \right) \left(\frac{N}{100} \right) = \left(\frac{100}{25} \right) 200 = \frac{20,000}{25} = 800 ,$$

where w_i^* is the adjusted weight for site i , n_{ALL} is the sample size of the original GRTS draw (which may contain sites for multiple years), n is the sample size of the current sample (e.g. for a single year), w_i is the original GRTS design weight for site i , and N is the population size (here, 20,000 units). The adjusted weights reflect that each of the sampled sites now represent a larger portion of the population because only a quarter of the sites in the original draw are surveyed in a given year.

Similarly, nonsampling error requires design weight adjustment prior to data analysis. Frame error affects the size of the population. In ecological sampling, frame overcoverage (inclusion of sites that do not meet the definition of the target population) is more conveniently assessed than undercoverage (omission of target sites from the sampling frame) because sites defined as outside the target population are not included in the sample. In the case of frame overcoverage, weighting adjustment decreases the size of the population, N , by the proportion of sampled sites that are verified to be outside the target population. In this case, the weights are reduced to account for the smaller frame extent. Nonresponse error, often caused by lack of access to sites, results in a smaller than expected sample, so the sampling weight must be adjusted to give each sampled site more weight to account for the smaller sample size. Weighting adjustment in the SDrawNPS package assumes that the sites missed due to nonresponse are missing completely at random (MCAR). This assumption implies that missingness is unrelated to the sampling design, the outcome of interest, or any related covariates. If the MCAR assumption is valid, the sample obtained after nonresponse may still be treated as a random sample and unbiased inference may be made.

Inference in the SDrawNPS tool may be made to one of two populations: the target population or the sampled population. The target population represents all sites in the sampling frame minus any frame error. The sampled population represents the subpopulation of sites that would be successfully surveyed if sampled, *i.e.* would not be subject to nonresponse error. Note that the sampled population is a subset of the target population, so estimates of the population total for the sampled population would be less than estimates of the population total for the target population when nonresponse occurs.

Extending the earlier example, assume that, of the 25 sites identified as the first year sample, only 20 were visited and evaluated before the sampling season ended. Assume that 4 sites were found to be non-target sites and 4 sites were inaccessible for a total of 12 sites that were successfully surveyed. Assuming inference to the sampled population, the adjusted design weights for a single year would be obtained as follows:

$$w_i^{*S} = \left(\frac{n}{n_E} \right) w_i^* = \left(\frac{25}{20} \right) w_i^* = 1.25 * 800 = 1000 ,$$

where n_E is the number of evaluated sites where frame membership is assessed and a survey is attempted. We see that each sampling unit represents 1000 population units after accounting for the smaller sample size due to nonresponse. Note that the sum of the adjusted design weights for all evaluated sampling units is $12*1000 = 12,000$ which is the adjusted extent of the population after accounting for 20% frame error (4/20 non-target sites) and 25% nonresponse (4/16 nonresponding sites). In other words, the extent of the subpopulation of sites that meet the target population definition and are accessible is estimated as $20,000*(1 - 0.20)*(1 - 0.25) = 12,000$ units.

Assuming inference to the target population, the adjusted design weights for a single year would be obtained as follows:

$$w_i^{*T} = \left(\frac{nn_T}{n_E n_R} \right) w_i^* = \left(\frac{25*16}{20*12} \right) w_i^* = 1.667 * 800 = 1333,$$

where n_T is the number of evaluated sites that meet the definition of the target population, and n_R is the number of responding sites, which are the evaluated sites that were successfully sampled and for which data were collected. Now each successfully-surveyed sampling unit carries 67% more weight to account for the nonresponse in the sample. Note that the sum of the weights for the responding units sums to $12*1333=16,000$ units which represents the adjusted extent of the target population after accounting for the 20% frame error exhibited in the sample but not adjusting for nonresponse (i.e. $20,000$ units * $(1 - 0.2)$ = $16,000$ units). After accounting for frame error, the extent of the target population is larger than the extent of the sampled population because the target population includes sites that would be found to be inaccessible if selected. Note that estimates of the mean are the same for the two scopes of inference but estimates of the total will be larger for inference to the target population due to the larger extent. However, if the mean of the outcome of interest at the missing sites is significantly different than the mean from accessible sites, inference from the obtained sample may be biased.

The weighting adjustment tool provided in the SDrawNPS package adjusts weights at the sample level. Weighting adjustments within strata or post-strata must be computed outside of the SDrawNPS GUI prior to analysis. For more details on the weighting adjustment, inferential populations, and missing data assumptions, please refer to Appendix A.

Case Study: Data Analysis without Weighting Adjustment

The first analysis case study comes from a program to monitor benthic macroinvertebrates in California. The California Stream Condition Index (CSCI) combines an index of taxonomic completeness with an index of ecological structure to assess stream condition across a wide gradient of California landscapes (Mazor et al. 2016). The sampling frame was based on the National Hydrography Dataset for perennial wadeable streams, divided into six regions and ten subregions, and stratified by ecoregional and hydrologic features (Peck et al. 2006). The example data were collected from a GRTS sample of southern California rivers and streams during a low water year which caused considerable frame and nonresponse error. Post-stratification (Oh and Scheuren 1983) by stream Strahler order and zone was used to adjust the original GRTS design weights to account

for varying nonresponse rates and CSCI values among post-strata. Because a more complicated weighting adjustment was used, we apply the adjusted weights provided from the post-stratification and demonstrate status estimation without adjusting the given weights further. This approach is analogous to estimation for a sample with nonsampling error for which weighting adjustment is also unnecessary. The example data are tabulated in a comma-separated file included in the “data” directory of the SDrawNPS package called “WQdata.csv.” This data set contains CSCI measurements from 65 sites at perennial wadeable streams in southern California. Note that these data were selected to demonstrate the SDrawNPS package tools and are not appropriate for inference.

After loading the SDrawNPS package and selecting the analysis pathway where the GRTS design file containing information such as the design weights and outcomes of interest, the analysis GUI window will appear (Figure 19). Like the sample draw GUIs, this GUI has fields at the top of the window for required inputs and includes a few places for optional input towards the bottom. In addition, at the bottom of the GUI window are four buttons that can be used to inspect the data, run the analysis, tabulate the sample, and complete the analysis.

To begin the data analysis process, the analyst should locate the data file. Using the “Browse” button at the top of the GUI (Figure 19), find the data stored in a comma-separated file (.csv) or as a data.frame object in the current R environment. In this example the data is saved as a .csv file in the SDrawNPS R library directory (*e.g.* C:\...\Documents\R\win-library\3.1\SDrawNPS\data). After the data file is located, the analyst should click the “Inspect Sample” button at the bottom of the screen to read the data file. A list containing the variables in the data set and their data classes will appear on the right side of the GUI. This list is useful for populating the remaining required fields in the GUI. The GUI window with the selected data file and its fields is shown in Figure 20.

The remaining fields are populated before running the analysis. Only the “Stratum/Subpopulation...” fields are not mandatory if stratification was not a sampling design element. The output file name specifies the name of the variable containing analysis output. Additionally, the “Site-ID Variable” field should contain the name of the ID field; here it is “siteID.” The “Evaluation Status Variable” identifies the survey status of each data record (See: Estimation: Formatting the GRTS Design File for more details). In our example data, the field that contains this information is called EvalStatus (Figure 20). All data in this example set were collected at Target sites, so all records have “Target – Surveyed” in the EvalStatus field. The Weight Variable is “AdjWgt_km”, as shown in the Data Info box on the right side of the GUI. The “X-Coordinate” and “Y-Coordinate” boxes are populated with “xcoord” and ycoord” fields, respectively. Lastly, the analyst includes in the “Outcome(s)” text box the name of the variable (or variables) that contain the response(s). In this example, the measured response is contained in the “CSCI” field.

In this example, there are no stratifying variables, so the “Stratum” text boxes are left blank. Since the weights in the data file had been adjusted prior to analysis, the radio button in the “Adjust Weights?” box should be set to the No default. The Population Inference radio button does not need to be selected if weighting adjustment is not used. In this example, nonsampling error is addressed in the post-stratification weighting and the inferential population (sampled vs. target) is defined by the

weighting adjustment (*i.e.* the sum of the adjusted weights in the sample provide the estimate of the inferential population size). The completed GUI interface for this case study is shown in Figure 21.

The screenshot shows the R SDrawNPS: Continuous variable analysis interface. The window title is "R SDrawNPS : Continuous variable analysis interface".

Required Inputs - Sample Information:

- CSV file OR data.frame object: [Input Field]
- Output File: [Input Field]
- Site-ID Variable: siteID
- Evaluation Status Variable: EvalStatus
- Inclusion Identifier: Target - Surveyed
- Stratum / Subpopulation 1: [Input Field]
- Stratum / Subpopulation 2: [Input Field]
- Weight Variable: wgt
- X-Coordinate: xcoord
- Y-Coordinate: ycoord
- Outcome(s): variable1, variable2, variable3, ...
- Confidence Level: 95

Data Info:

VARIABLE	CLASS
<pending>	<pending>

Optional Weighting Inputs:

Adjust Weights? Yes No Population Inference Target Sampled

Figure 19. The GUI interface used to analyze data collected as part of a GRTS design.

R SDrawNPS : Continuous variable analysis interface

Required Inputs		Data Info	
- Sample Information		Sample Contents: 6 VARIABLES	
CSV file OR data.frame object:	WQdata.csv	Browse	
Output File:			
Site-ID Variable:	siteID		
Evaluation Status Variable:	EvalStatus		
Inclusion Identifier:	Target - Surveyed		
Stratum / Subpopulation 1:			
Stratum / Subpopulation 2:			
Weight Variable:	wgt		
X-Coordinate:	xcoord		
Y-Coordinate:	ycoord		
Outcome(s):	variable1, variable2, variable3, ...		
Confidence Level:	95		
Optional Weighting Inputs - Adjust Weights? <input type="radio"/> Yes <input checked="" type="radio"/> No			
Population Inference <input checked="" type="radio"/> Target <input type="radio"/> Sampled			
<input type="button" value="Inspect Sample"/>	<input type="button" value="Run"/>	<input type="button" value="Tabulate Sample"/>	<input type="button" value="Done"/>

Figure 20. The GUI interface showing the browsed data file (WQdata.csv) and the data fields and variable classes read using the “Inspect Sample” button at the bottom left.

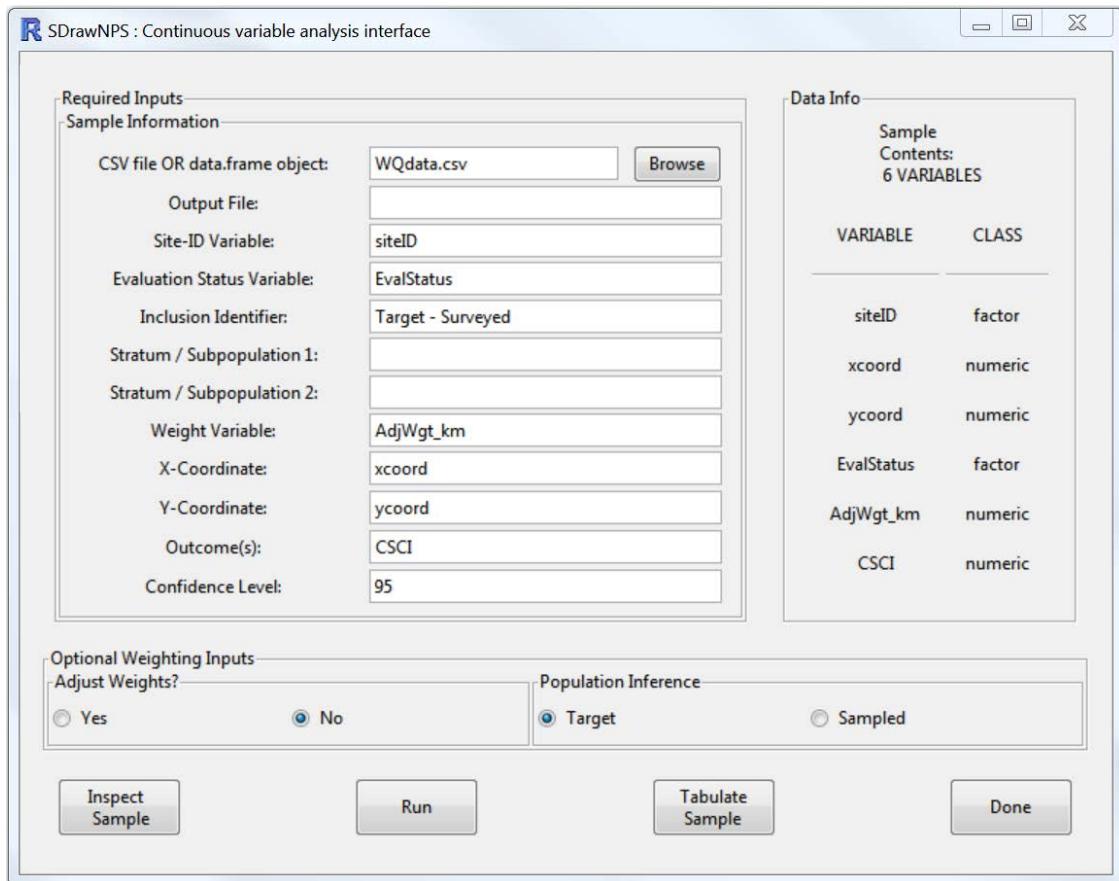


Figure 21. The completed analysis GUI to analyze the water quality case study data set.

After pressing the “Run” button, a window will pop up with the message: “You have selected to not weight your sample.” This message reminds the analyst that weighting adjustment will not be used in case this option was selected in error.

After running the analysis, two files are written to the R working directory. The first is a *.log text file showing the R window commands used to run the analysis. This file is named by appending the text string “Analysis of” to the data file name (*e.g.* “Analysis of WQdata”). The second file is a cumulative distribution function (CDF) plot saved as an Adobe Acrobat file (*.pdf). The CDF plot file is saved as the name as the data file appended to “ – CDF Plots” (*e.g.* “WQdata – CDF Plots”).

The analysis results files, LOG file of the R code used in the spsurvey package, and the CDF plots PDF file are written to the location of the data set (Figure 22). These two analysis results files contain CDF estimates and estimates of percentiles, means, totals, and variances (named “_CDF” and “_Pct”, respectively). The analysis results can also be viewed in the R window by using the Tabulate Sample button.

Name	Date modified	Type	Size
Analysis of WQdata	12/16/2015 11:55 ...	Text Document	2 KB
WQdata - CDF Plots	12/16/2015 11:55 ...	Adobe Acrobat D...	6 KB
WQdata	12/4/2015 5:38 PM	Microsoft Excel C...	5 KB
WQdata_CDF	12/16/2015 11:56 ...	Microsoft Excel C...	12 KB
WQdata_Pct	12/16/2015 11:56 ...	Microsoft Excel C...	2 KB

Figure 22. Screenshot of the files saved following the analysis of the water quality data set.

The “_CDF” file contains the CDF data that forms the basis for the CDF plot which shows the sample data expressed as the proportion (percent) of values at or below a given response. This CDF contains the summary of the output from this analysis and is shown in Figure 23. The CDF plot indicates that the proportion of CSCI values at or below a CSCI of 0.5 is 15% or so and the median (50th percentile) CSCI value is about 0.7.

In the CDF file, the output shows not only the information associated with the sample design but then also provides the results from the proportional estimates (that is, the CDF as percentage values). This percentage and associated estimates of variation are labeled with a “.P” in the field name (Figure 24). Since the analysis expresses the CSCI measurement as proportional estimates, the analyst can read the proportional estimate for any of the measured values. For example, the value of CSCI at the 51.96% was 0.686241. The standard error (7.84) associated with this proportional estimate, as well as a 95% lower (36.58) and upper (67.33) confidence interval bounds, are obtained with the neighborhood variance estimator (Stevens and Olsen 2003). The estimates in the CDF file labeled with a “.U” in the field name represent the cumulative size of the population units represented by each percentile. Note that the 100th percentile represents the size of the entire inferential population estimate, which is also obtained by summing the design weights for the sample.

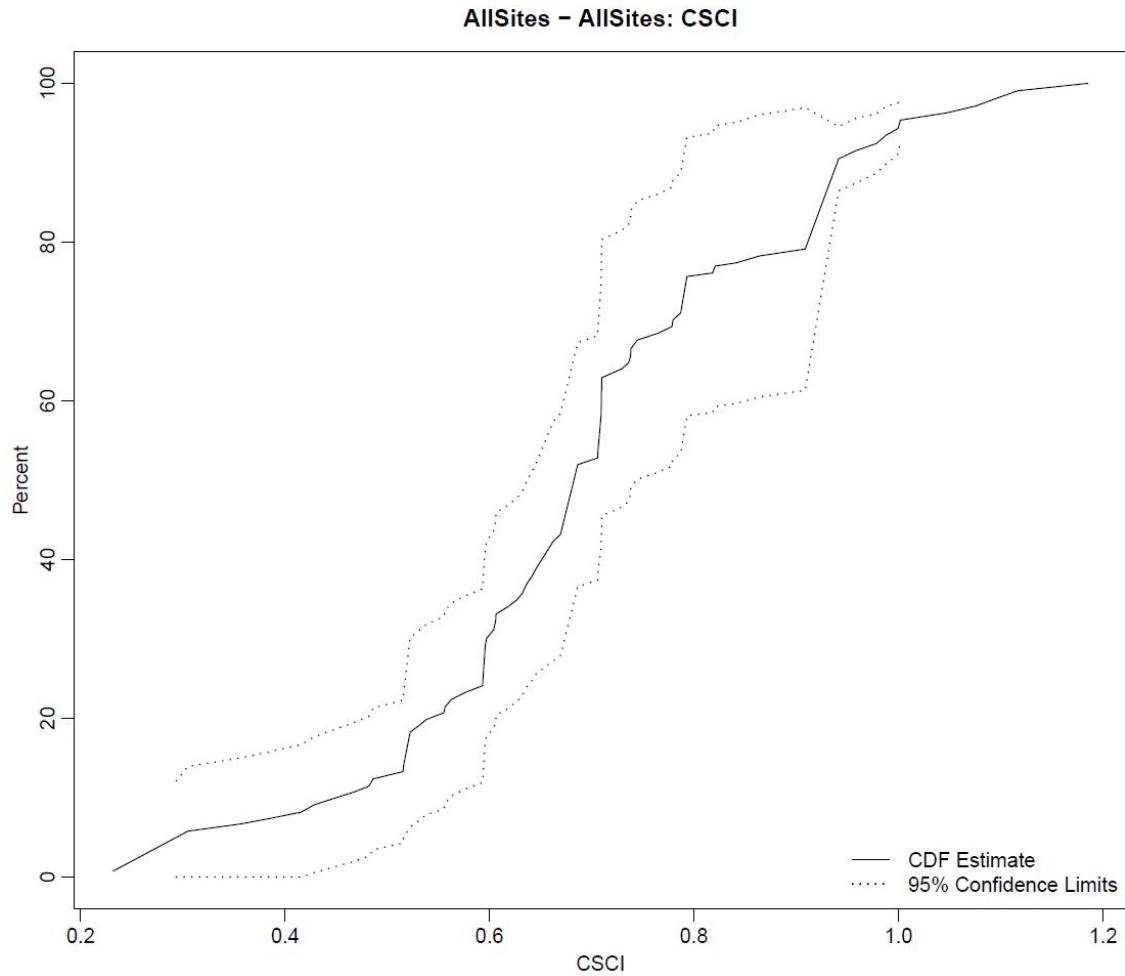


Figure 23. The cumulative distribution function (CDF) resulting from the analysis of the water quality case study data.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
Type	Subpopulation	Indicator	Value	NResp	Estimate.P	StdError.P	LCB95Pct.P	UCB95Pct.P	Estimate.U	StdError.U	LCB95Pct.U	UCB95Pct.U		
1	AllSites	CSCI	0.23185	1	0.738204675	0.64385099	0	2.000129426	0.527528716	0.457513956	0	1.424239592		
2	AllSites	CSCI	0.304925	2	5.768952935	4.164019445	0	13.93028108	4.122553594	2.984299619	0	9.971673366		
3	AllSites	CSCI	0.357049	3	6.691903551	4.245541944	0	15.01301285	4.782103675	3.040068314	0	10.74052808		
4	AllSites	CSCI	0.387436	4	7.430108225	4.295137386	0	15.84842281	5.309632391	3.071750781	0	11.33015329		
5	AllSites	CSCI	0.415688	5	8.1683129	4.338917763	0	16.67243545	5.837161107	3.099296412	0	11.91167045		
6	AllSites	CSCI	0.428265	6	9.082996281	4.362102337	0.533432805	17.63255976	6.490803337	3.111595113	0.39218898	12.58941769		
7	AllSites	CSCI	0.446492	7	9.848164536	4.411874791	1.201048842	18.49528023	7.037600507	3.141252689	0.88085837	13.19434264		
8	AllSites	CSCI	0.467261	8	10.69698255	4.457269738	1.960894395	19.43307071	7.644174663	3.167318639	1.436344203	13.85200512		
9	AllSites	CSCI	0.482034	9	11.43518723	4.49404292	2.627024957	20.24334949	8.171703379	3.187436344	1.924442942	14.41896382		
10	AllSites	CSCI	0.486168	10	12.35813784	4.559697432	3.421295094	21.29498059	8.83125346	3.226557548	2.507316873	15.1519005		
11	AllSites	CSCI	0.515439	11	13.28108846	4.586487832	4.29173749	22.27043942	9.490803541	3.236609438	3.14716561	15.83444147		
12	AllSites	CSCI	0.515924	12	14.01929313	4.61012956	4.983605229	23.05498103	10.01833226	3.245064423	3.658122861	16.37854165		
13	AllSites	CSCI	0.522382	13	18.24789387	6.09079842	6.310148333	30.18563941	13.0401342	4.114824068	4.975227222	21.10504118		

Figure 24. Tabulated results saved in the WQdata_CDF.csv file.

The other exported file generated by the analysis summarizes the percentile values, as well as the population total, the mean, the variance, and the standard deviation (Figure 25). The percentile values are obtained by inverting the CDF estimate and an associated error is not reported. This file provides a convenient way to report percentiles of interest (the first 10% of CSCI is at or below 0.450207), or

the mean value with its associated spread. Standard errors are calculated using nearest neighborhood variance estimator.

	A	B	C	D	E	F	G	H	I	J	K
1	Type	Subpopul.	Indicator	Statistic	NResp	Estimate	StdError	LCB95Pct	UCB95Pct		
2	1 AllSites	AllSites	Cont1	5Pct		1	0.293755	0.271736	0.347106		
3	2 AllSites	AllSites	Cont1	10Pct		7	0.450207	0.240735	0.526589		
4	3 AllSites	AllSites	Cont1	25Pct		20	0.593617	0.498073	0.638199		
5	4 AllSites	AllSites	Cont1	50Pct		34	0.6825	0.619039	0.738102		
6	5 AllSites	AllSites	Cont1	75Pct		48	0.79227	0.708364	0.98352		
7	6 AllSites	AllSites	Cont1	90Pct		54	0.940089	0.787076	1.185298		
8	7 AllSites	AllSites	Cont1	95Pct		59	1.0011	0.964831	1.097518		
9	8 AllSites	AllSites	Cont1	Total		65	50.00048	5.741078	38.74818	61.25279	
10	9 AllSites	AllSites	Cont1	Mean		65	0.699689	0.031892	0.637181	0.762196	
11	10 AllSites	AllSites	Cont1	Variance		65	0.038461	0.007746	0.02328	0.053642	
12	11 AllSites	AllSites	Cont1	Std. Devia		65	0.196115	0.019747	0.157411	0.234819	
13											

Figure 25. Tabulated results saved in the WQdata_Pct.csv file.

The following code can be used in the R window to do the same analysis. This code is obtained from the LOG text file generated when the analysis is conducted in the SDrawNPS package.

```
# Read in the CSV file of interest for which analysis is required. Note that you must specify the appropriate file path on your computer.
```

```
df <- read.csv("C:/..PATH ..//WQdata.csv")
```

```
# Identify your chosen weighting scheme.
```

```
df$wgt <- df$AdjWgt_km
```

```
# Prepare the analysis for use in the continuous analysis function.
```

```
the.sites <- data.frame(siteID=df$siteID, df$EvalStatus=="Target - Surveyed")
the.subpop <- data.frame(siteID=df$siteID, Popn1=rep('AllSites',nrow(df)))
the.design <- data.frame(siteID=df$siteID, wgt=df$AdjWgt_km, xcoord=df$xcoord,
ycoord=df$ycoord)
the.data.cont <- data.frame(siteID=df$siteID, Cont1=df$CSCI)
```

```
# Analyze the sample via the cont.analysis function in package spsurvey and write the results to an R object named ans.cont.
```

```
ans.cont <- cont.analysis(sites=the.sites,
subpop=the.subpop,
design=the.design,
data.cont=the.data.cont,
```

```
total=TRUE,  
conf=95)
```

```
# Plot the resulting empirical distribution functions and write the results to an Adobe Acrobat file  
named WQdata - CDF Plots.pdf.
```

```
cont.cdfplot("C:/..PATH../WQdata - CDF Plots.pdf", ans.cont$CDF,cdf.plot=1)
```

Case Study: Data Analysis with Weighting Adjustment

The second case study for analysis is obtained from a subset of vegetative cover measurements collected for the MOJN integrated upland monitoring program. These data should be used for demonstration of the SDrawNPS package and not for inference. The vegetation data are tabulated in a comma-separated file included in the “data” directory of the SDrawNPS package called “Vegdata.csv.” This data set contains 35 sites from the sample draw and an additional 301 oversample records. The data file indicates that all 35 of the main sample locations and 7 of the oversample locations were visited during the survey period. Unlike the previous case study, though, not all the sample sites were successfully surveyed. Four (4) of the main sample sites were deemed to be Non-Target sites because the vegetation alliance at that site was not represented by the four alliances that define the target population. Note that a site that is misclassified as in a different vegetation alliance but one that falls within the target population is still considered a member of the target population. Furthermore, four of the target locations were judged to be part of the sampling frame but were not sampled for other reasons, such as falling too close to a trail, inaccessibility due to a fence, and the presence of cultural resources.

Estimation of vegetation parameters is conducted similarly to the water quality data analysis shown above. One exception is that in this example the weights will be adjusted with tools available in the SDrawNPS package to account for frame and nonresponse error. After selecting the subdirectory containing the GRTS design file, the analyst should navigate using the Browse button to the location of the survey data. After selecting the data file, pressing the “Inspect Sample” button will read the file and show the data fields and classes for the survey data (Figure 26).

R SDrawNPS : Continuous variable analysis interface

Required Inputs

Sample Information

CSV file OR data.frame object:	Vegdata.csv	Browse
Output File:		
Site-ID Variable:	siteID	
Evaluation Status Variable:	EvalStatus	
Inclusion Identifier:	Target - Surveyed	
Stratum / Subpopulation 1:		
Stratum / Subpopulation 2:		
Weight Variable:	wgt	
X-Coordinate:	xcoord	
Y-Coordinate:	ycoord	
Outcome(s):	variable1, variable2, variable3, ...	
Confidence Level:	95	

Optional Weighting Inputs

Adjust Weights? Yes No

Population Inference Target Sampled

Data Info

Sample Contents: 13 VARIABLES

VARIABLE	CLASS
siteID	factor
xcoord	integer
ycoord	integer
mdcaty	factor
wgt	numeric
panel	factor
EvalStatus	factor
EvalReason	factor
Mean.Percent.Cover	numeric
Mean.Canopy.Gap.Size.m	numeric
Mean.Basal.Gap.Size.m	numeric
Mean.Canopy.Gap.Percent	numeric
Mean.Basal.Gap.Percent	numeric

Figure 26. Analysis GUI showing data fields and classes for the vegetation data survey.

To complete the analysis, the analyst should then use the field information from the right side of the interface box to populate the required fields. Many of these fields autofill, so the analyst should verify the field name before proceeding. Like in the previous case study, the Evaluation Status Variable is contained in the field EvalStatus, and the Inclusion Identified is “Target – Surveyed.” Unlike the previous example, this data set does have sites that were classified as either “Non – Target” or “Target – Not Surveyed.” Only the records with “Target – Surveyed” in the EvalStatus field will be used in the analysis. In this example, a four-level vegetation classification was used to define unequal probability sampling, and this variable is contained in the “mdcaty” field. To obtain estimates by each classification, enter “mdcaty” in the “Stratum / Subpopulation 1” text box. The Weight Variable is “wgt”, and the X- and Y-Coordinate variables are xcoord and ycoord, respectively. Finally, the analyst should enter an outcome name or a set of outcomes names separated by commas in the Outcome(s) field. In this example, the status estimates for the Mean.Canopy.Gap.Size.m. outcomes are obtained.

Since the sample included some non-target sites and some inaccessible sites (nonresponse), weighting adjustment is required for unbiased inference under the assumption of MCAR missingness. Inference may be made to the target population (minus the misclassified sites that are outside the target population) or the sampled population (the subpopulation of sites that would be successfully surveyed if sampled). In this example, the population status is estimated with inference

to the target population. Choose the “Target” button in the “Population Inference” box and click the “Yes” button in the “Adjust Weights?” box. The GUI for this analysis is depicted in Figure 27.

Pressing the Run button will bring up a window reminding the analyst that the weighting adjustment option has been selected. If the analysis is successful, a window with a message indicating that the analysis is complete will pop up. As before, two types of analysis results are provided: the CDF estimates and the percentiles, mean, and total estimates. Standard errors are based on the neighborhood variance estimator. The “_CDF” and “_Pct” summary files, CDF plots, and a LOG text file containing the generated R commands are written to the folder where the data were located.

Figure 28 shows a screenshot of the files in the directory from running this analysis. The estimates may also be viewed by examining a variable in the R workspace with the variable name specified in the analysis GUI with the suffix “.cont”. For example, if the output file name is specified in the GUI as “Veg_Target”, then the estimates of the percentiles, mean, and total can be obtained in the R command window with the command Veg_Target.cont\$Pct and the CDF estimates can be obtained with the command Veg_Target.cont\$CDF (Figure 29).

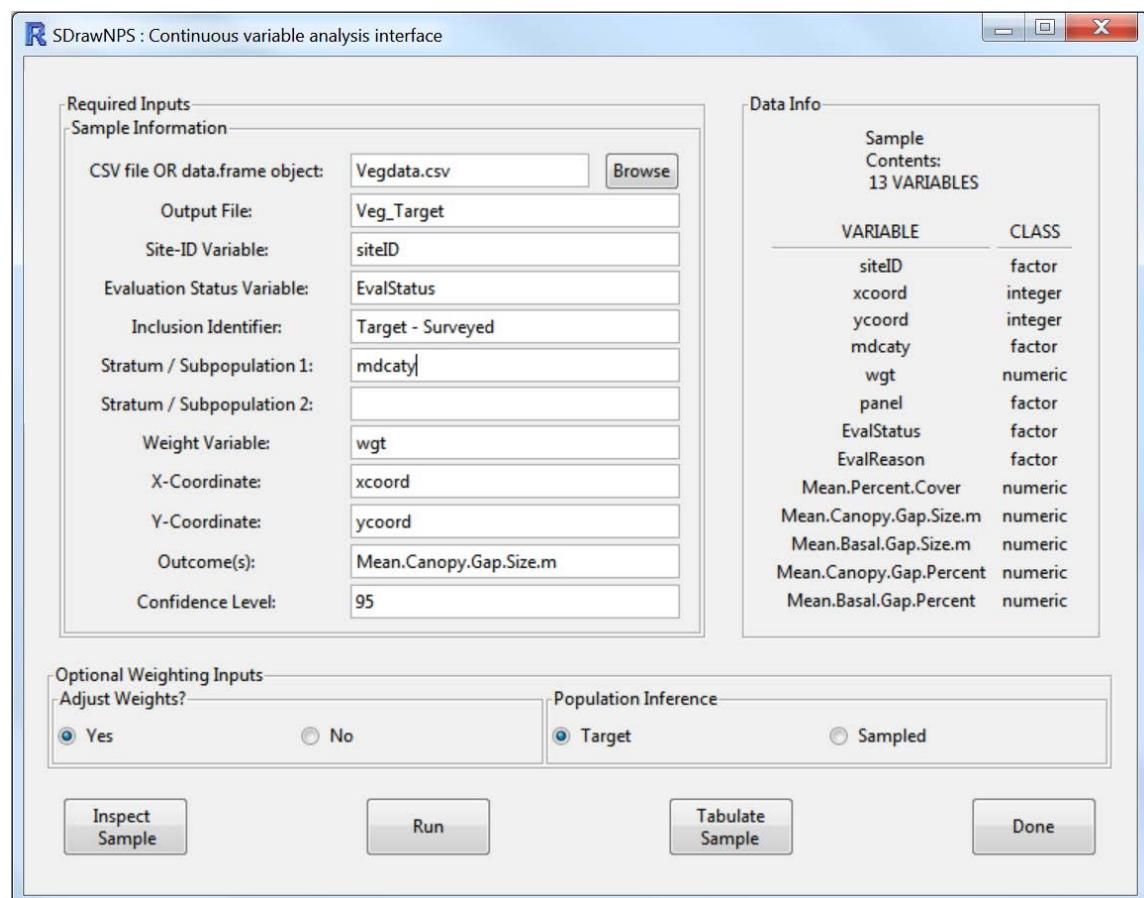


Figure 27. The completed GUI to analyze the Mean Canopy Gap Size (m) from the vegetation data set.

Name	Date modified	Type	Size
Analysis of Vegdata - Veg_Target	12/28/2015 11:54 ...	Text Document	2 KB
Veg_Target_CDF	12/28/2015 11:56 ...	Microsoft Excel C...	19 KB
Veg_Target_Pct	12/28/2015 11:56 ...	Microsoft Excel C...	6 KB
Vegdata - Veg_Target - CDF Plots	12/28/2015 11:54 ...	Adobe Acrobat D...	8 KB
Vegdata	12/4/2015 5:38 PM	Microsoft Excel C...	23 KB

Figure 28. Screenshot of the files saved following the analysis of the vegetation data set and output files to the directory containing the data set.

```

R Console

> Veg_Target.cont$Pct[1:11,]
   Type Subpopulation Indicator Statistic NResp Estimate StdError LCB95Pct UCB95Pct
1 AllSites AllSites Mean.Canopy.Gap.Size.m 5Pct 2 1.527665e+00 1.360000e+00 1.752478e+00
2 AllSites AllSites Mean.Canopy.Gap.Size.m 10Pct 4 1.743722e+00 1.363325e+00 1.909811e+00
3 AllSites AllSites Mean.Canopy.Gap.Size.m 25Pct 9 2.040596e+00 1.797376e+00 2.312377e+00
4 AllSites AllSites Mean.Canopy.Gap.Size.m 50Pct 18 2.633480e+00 2.295197e+00 2.903864e+00
5 AllSites AllSites Mean.Canopy.Gap.Size.m 75Pct 25 3.575023e+00 2.892978e+00 4.132873e+00
6 AllSites AllSites Mean.Canopy.Gap.Size.m 90Pct 31 4.306939e+00 3.653366e+00 7.256705e+00
7 AllSites AllSites Mean.Canopy.Gap.Size.m 95Pct 32 5.914889e+00 4.110346e+00 7.300000e+00
8 AllSites AllSites Mean.Canopy.Gap.Size.m Total 34 4.459197e+08 39094547.1308132 3.692958e+08 5.225436e+08
9 AllSites AllSites Mean.Canopy.Gap.Size.m Mean 34 3.056663e+00 0.208313302657489 2.648377e+00 3.464950e+00
10 AllSites AllSites Mean.Canopy.Gap.Size.m Variance 34 1.841246e+00 0.610138282046932 6.453971e-01 3.037095e+00
11 AllSites AllSites Mean.Canopy.Gap.Size.m Std. Deviation 34 1.356925e+00 0.224823838206736 9.162786e-01 1.797572e+00
> Veg_Target.cont$CDF
   Type Subpopulation Indicator Value NResp Estimate.P StdError.P LCB95Pct.P UCB95Pct.P Estimate.U
1 AllSites AllSites Mean.Canopy.Gap.Size.m 1.3 1 0.8351147 0.7811576 0.0000000 2.366156 1218303
2 AllSites AllSites Mean.Canopy.Gap.Size.m 1.5 2 4.0072188 2.7004390 0.0000000 9.299982 5845909
3 AllSites AllSites Mean.Canopy.Gap.Size.m 1.6 3 7.5958610 4.0461074 0.0000000 15.526086 11081180
4 AllSites AllSites Mean.Canopy.Gap.Size.m 1.7 4 8.4309758 4.1465051 0.3039752 16.555976 12299483
5 AllSites AllSites Mean.Canopy.Gap.Size.m 1.8 5 12.0196180 5.0377579 2.1457939 21.893442 17534754
6 AllSites AllSites Mean.Canopy.Gap.Size.m 2.0 9 23.2041213 6.6574495 10.1557600 36.252483 33851205
7 AllSites AllSites Mean.Canopy.Gap.Size.m 2.1 11 27.6278783 7.0673691 13.7760893 41.479667 40304779
8 AllSites AllSites Mean.Canopy.Gap.Size.m 2.2 12 30.7999824 7.1833056 16.7209622 44.879003 44932386
9 AllSites AllSites Mean.Canopy.Gap.Size.m 2.3 14 37.5607287 6.8027101 24.2276618 50.893796 54795263
10 AllSites AllSites Mean.Canopy.Gap.Size.m 2.4 16 41.7123546 6.8833124 28.2213103 55.203399 60851840
11 AllSites AllSites Mean.Canopy.Gap.Size.m 2.5 17 45.3009968 7.0518336 31.4796569 59.122337 66087111

```

Figure 29. Accessing the analysis results of the vegetation data set in the R command window.

The analysis output is summarized in the Veg_Target_Pct.csv file in the folder, and a partial screenshot of those output are provided in Figure 30. The results include estimates across and by vegetation type since the “mdcaty” field was entered in the “Stratum / Subpopulation 1” text box. The mean estimate is found in the results row where Statistic is “Mean.” Across all vegetation types, the mean canopy gap size (m) was 3.06 m (95%-CI: 2.65m – 3.46m); for the vegetation class “CR”, the mean canopy gap size was 2.17 (1.65 – 2.68).

A	B	C	D	E	F	G	H	I	J	K
1	Type	Subpopulation	Indicator	Statistic	NResp	Estimate	StdError	LCB95Pct	UCB95Pct	
2	1 AllSites	AllSites	Mean.Canopy.Gap.Size.m	5Pct		2 1.527664535		1.3	1.752478	
3	2 AllSites	AllSites	Mean.Canopy.Gap.Size.m	10Pct	4	1.743721946		1.363325	1.909811	
4	3 AllSites	AllSites	Mean.Canopy.Gap.Size.m	25Pct	9	2.040596232		1.797376	2.312377	
5	4 AllSites	AllSites	Mean.Canopy.Gap.Size.m	50Pct	18	2.633479789		2.295197	2.903864	
6	5 AllSites	AllSites	Mean.Canopy.Gap.Size.m	75Pct	25	3.575023482		2.892978	4.132873	
7	6 AllSites	AllSites	Mean.Canopy.Gap.Size.m	90Pct	31	4.306938652		3.653368	7.256705	
8	7 AllSites	AllSites	Mean.Canopy.Gap.Size.m	95Pct	32	5.914889009		4.110346	7.3	
9	8 AllSites	AllSites	Mean.Canopy.Gap.Size.m	Total	34	445919671.7	39094547.13	3.69E+08	5.23E+08	
10	9 AllSites	AllSites	Mean.Canopy.Gap.Size.m	Mean	34	3.056663442	0.208313303	2.648377	3.46495	
11	10 AllSites	AllSites	Mean.Canopy.Gap.Size.m	Variance	34	1.841246112	0.610138282	0.645397	0.307095	
12	11 AllSites	AllSites	Mean.Canopy.Gap.Size.m	Std. Deviation	34	1.356925242	0.224823838	0.916279	1.797572	
13	12 mdcaty	CR	Mean.Canopy.Gap.Size.m	5Pct	1	1.3		1.3	1.690664	
14	13 mdcaty	CR	Mean.Canopy.Gap.Size.m	10Pct	1	1.3		1.3	1.841328	
15	14 mdcaty	CR	Mean.Canopy.Gap.Size.m	25Pct	1	1.65		1.3	2.010664	
16	15 mdcaty	CR	Mean.Canopy.Gap.Size.m	50Pct	3	2		1.3	3.219261	
17	16 mdcaty	CR	Mean.Canopy.Gap.Size.m	75Pct	4	2.35		1.864904	3.5	
18	17 mdcaty	CR	Mean.Canopy.Gap.Size.m	90Pct	5	3.26		2.06878	3.5	

Figure 30. Partial screenshot of tabulated results saved in the VegAnalysis_Pct.csv file.

The resulting CDF plot is divided up into panels representing the different vegetation classes as classified in the “mdcaty” field. The first panel (in the upper left) is for all classes combined, and then the remaining 4 panels are for each of the unique vegetation classes (Figure 31).

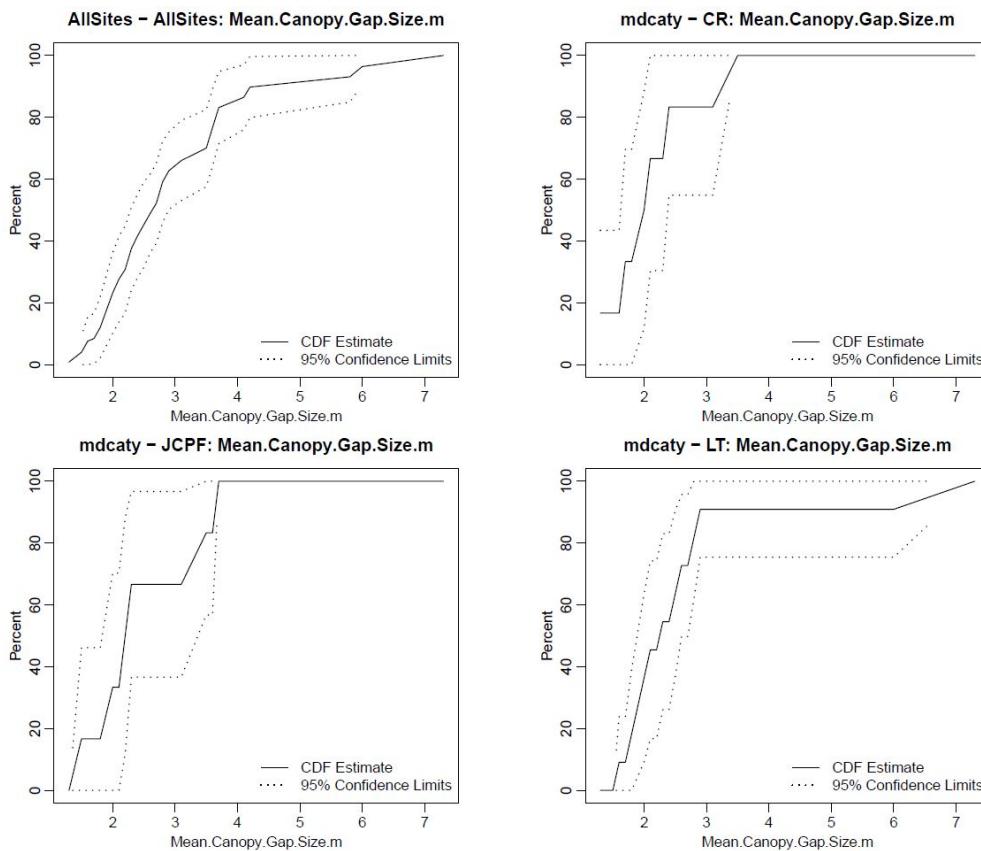


Figure 31. The first four panels of the cumulative distribution function (CDF) resulting from the analysis of the mean canopy gap size (m) in the vegetation data set. The upper left panel shows the results across all vegetation classes, while the other panels are for unique vegetation classes. The last panel is not shown.

The following code can be used in the R window to recreate the analysis with inference to the target population.

```
# Read in the csv file of interest for which analysis is required.  
df <- read.csv("C:/...PATH.../Vegdata.csv")  
  
# Identify the pop'n of inference. Keep in mind you have chosen to focus on the entire pop'n.  
popn <- "Target"  
  
# Identify your chosen weighting scheme. Keep in mind SDrawNPS is adjusting your weights.  
df$oldwgt <- df$wgt  
df$wgt <- Adjwgt_FrameNR(dat=df, popn=popn, evalstatus="EvalStatus", wgt="oldwgt")  
  
# Prepare the analysis for use in the continuous analysis function.  
  
the.sites <- data.frame(siteID=df$siteID, df$EvalStatus=="Target - Surveyed")  
the.subpop <- data.frame(siteID=df$siteID, Popn1=rep('AllSites',nrow(df)), Popn2=df$mdcaty)  
names(the.subpop) <- c('siteID','AllSites','mdcaty')  
the.design <- data.frame(siteID=df$siteID, wgt=df$wgt, xcoord=df$xcoord, ycoord=df$ycoord)  
the.data.cont <- data.frame(siteID=df$siteID,  
Mean.Canopy.Gap.Size.m=df$Mean.Canopy.Gap.Size.m)  
  
# Analyze the sample via the cont.analysis function in package spsurvey.  
  
ans.cont <- cont.analysis(sites=the.sites,  
    subpop=the.subpop,  
    design=the.design,  
    data.cont=the.data.cont,  
    total=TRUE,  
    conf=95)  
  
# Plot the resulting empirical distribution functions.  
cont.cdfplot("C:/...PATH.../Vegdata - Veg_Target - CDF Plots.pdf",ans.cont$CDF,cdf.plot=4)
```

In this example, the analyst can also specify that the inference be made only to the sampled population. Using the Population Inference box in the GUI as shown in Figure 32, the results of this analysis can be written to a new set of output files with the name “Veg_Sampled”. This analysis is similar to the analysis for inference to the target population, but the estimate of the total (not interpretable for this example of canopy gap) is smaller for inference to the sampled population because inference is not made to the proportion of sites for which nonresponse occurs (Figure 33).

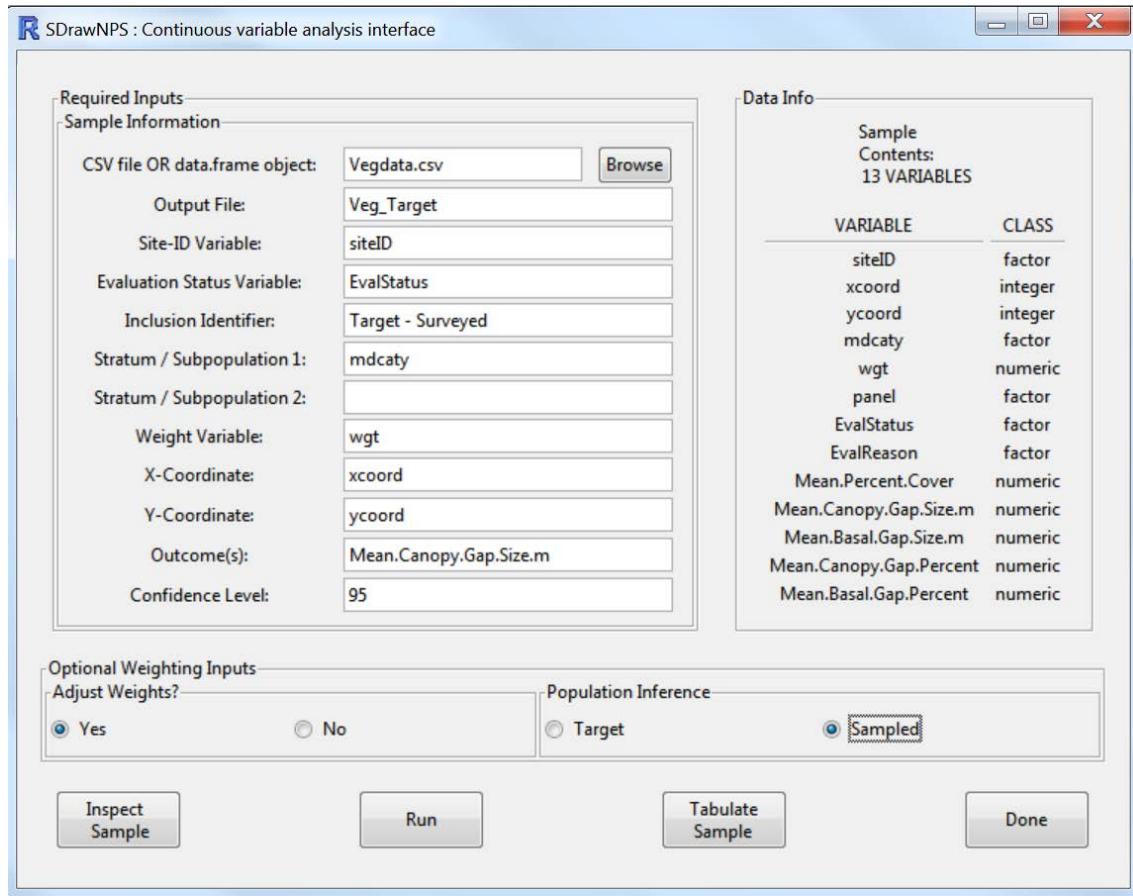


Figure 32. Analysis GUI for the analysis of mean canopy gap size with inference to the sampled population.

A	B	C	D	E	F	G	H	I	J	K
Type	Subpopulation	Indicator	Statistic	NResp	Estimate	StdError	LCB95Pct	UCB95Pct		
1	AllSites	AllSites	Mean.Canopy.Gap.Size.m	5Pct	2	1.527664535		1.3	1.752478336	
2	AllSites	AllSites	Mean.Canopy.Gap.Size.m	10Pct	4	1.743721946	1.363324874	1.909810898		
3	AllSites	AllSites	Mean.Canopy.Gap.Size.m	25Pct	9	2.040596232	1.797375518	2.312376755		
4	AllSites	AllSites	Mean.Canopy.Gap.Size.m	50Pct	18	2.633479789	2.295196544	2.903863995		
5	AllSites	AllSites	Mean.Canopy.Gap.Size.m	75Pct	25	3.575023482	2.892977656	4.132872997		
6	AllSites	AllSites	Mean.Canopy.Gap.Size.m	90Pct	31	4.306938652	3.653367593	7.256704923		
7	AllSites	AllSites	Mean.Canopy.Gap.Size.m	95Pct	32	5.914889009	4.11034568	7.3		
8	AllSites	AllSites	Mean.Canopy.Gap.Size.m	Total	34	398980758.9	34979331.64	330422528.7	467538989.1	
9	AllSites	AllSites	Mean.Canopy.Gap.Size.m	Mean	34	3.056663442	0.208313303	2.648376871	3.464950012	
10	AllSites	AllSites	Mean.Canopy.Gap.Size.m	Variance	34	1.841246112	0.610138283	0.645397052	3.037095171	
11	AllSites	AllSites	Mean.Canopy.Gap.Size.m	Std. Deviation	34	1.356925242	0.224823838	0.916278616	1.797571868	
12	mdcaty	CR	Mean.Canopy.Gap.Size.m	5Pct	1	1.3		1.3	1.690664143	
13	mdcaty	CR	Mean.Canopy.Gap.Size.m	10Pct	1	1.3		1.3	1.841328286	
14	mdcaty	CR	Mean.Canopy.Gap.Size.m	25Pct	1	1.65		1.3	2.010664143	
15	mdcaty	CR	Mean.Canopy.Gap.Size.m	50Pct	3	2		1.3	3.219260618	
16	mdcaty	CR	Mean.Canopy.Gap.Size.m	75Pct	4	2.35	1.864904209	3.5		

Figure 33. Partial screenshot of tabulated results saved in the Veg_Sampled_Pct.csv file.

While the water quality data set contained only one response variable (CSCI), the vegetation data set has five (5) responses: Mean Percent Cover, Mean Canopy Gap Size (m), Mean Basal Gap Size (m), Mean Canopy Gap (%), and Mean Basal Gap (%). An analysis for these five variables can occur simultaneously if the analyst enters multiple variables in the Outcome(s) text box. These variables

need to be separated by a comma. The analysis GUI (Figure 34) shows the input fields for the analysis of Mean Canopy Gap Size and Mean Basal Gap Size.

VARIABLE	CLASS
siteID	factor
xcoord	integer
ycoord	integer
mdcaty	factor
wgt	numeric
panel	factor
EvalStatus	factor
EvalReason	factor
Mean.Percent.Cover	numeric
Mean.Canopy.Gap.Size.m	numeric
Mean.Basal.Gap.Size.m	numeric
Mean.Canopy.Gap.Percent	numeric
Mean.Basal.Gap.Percent	numeric

Figure 34. The completed GUI to analyze both Mean Canopy Gap Size and Mean Basal Gap Size from the vegetation data set. Note that the full variable name is not visible in the text box. Variable names must be separated by commas.

Given that this sample was subject to frame error, an estimate of the frame extent may be obtained from an analysis of the indicator of frame inclusion. In the R console, read the sample data into a dataframe called ‘Vegdata’ and define the frame indicator, FrameInd, to be 1 if the sampled site was in the target population and 0 otherwise, as follows:

```
Vegdata = read.csv("Vegdata.csv", header=TRUE)
Vegdata$FrameInd <- as.numeric(!Vegdata$EvalStatus %in% c("Non-Target", "NotEval"))
```

Then analyze the data in the Continuous variable analysis interface GUI as in Figure 35 using the R data.frame object “Vegdata” rather than the CSV file Vegdata.csv in this analysis so that the FrameInd field can be used. Note that inference in this example is made to the Target population. Click Tabulate Sample to see the results (Figure 36). The total size of the frame after accounting for frame error is 145,884,452 m² for the target population.

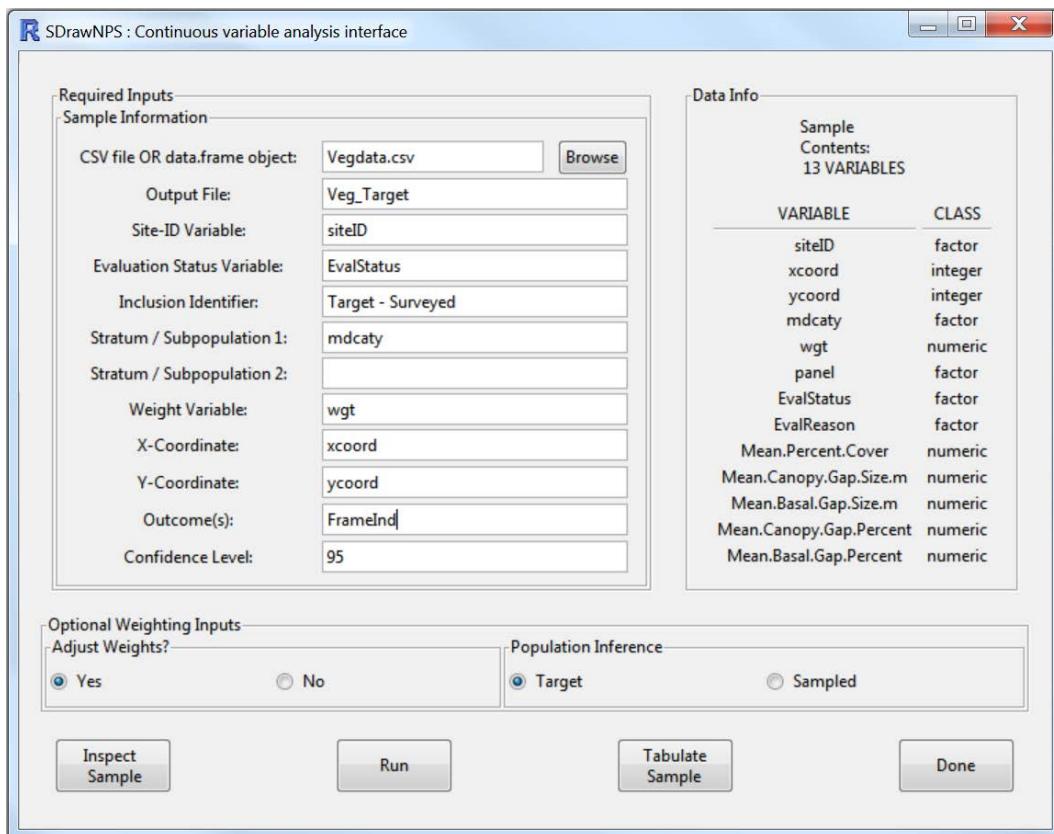


Figure 35. Analysis GUI for the estimation of frame extent for the target population after accounting for frame error.

Type	Subpopulation	Indicator	Statistic	NResp	Estimate	StdError	LCB95Pct
1	AllSites	AllSites	FrameInd	5Pct	34	1	1
2	AllSites	AllSites	FrameInd	10Pct	34	1	1
3	AllSites	AllSites	FrameInd	25Pct	34	1	1
4	AllSites	AllSites	FrameInd	50Pct	34	1	1
5	AllSites	AllSites	FrameInd	75Pct	34	1	1
6	AllSites	AllSites	FrameInd	90Pct	34	1	1
7	AllSites	AllSites	FrameInd	95Pct	34	1	1
8	AllSites	AllSites	FrameInd	Total	34	145884452	6730657.68068165 132692606
9	AllSites	AllSites	FrameInd	Mean	34	1	0
10	AllSites	AllSites	FrameInd	Variance	34	0	0
11	AllSites	AllSites	FrameInd	Std. Deviation	34	0	0

Figure 36. Partial screenshot of tabulated results saved in the VegFrameExtent.csv file for the analysis of frame extent after accounting for frame error.

Discussion

The SDrawNPS package provides a simple user interface for ecologists who want to draw and analyze data from spatially-balanced samples. This tool represents a range of complex designs but certainly does not attempt to encompass all possible design complexity. The R code archived in the log file for a sampling design provided in the SDrawNPS package may be used as a template for a more complex design not yet available in the SDrawNPS package. For a two-stage design, an initial GRTS sample of polygons treated as finite and then serve as the sampling frame for a second-stage sample of points within each first-stage polygon by treating each first-stage polygon as a stratum. In this way, the tools provided in the SDrawNPS package may be used to facilitate additional design complexity.

The information provided in this manual covers a number of issues that natural resource managers face when designing and implementing a monitoring plan and analyzing data. The complex nature of sampling design in ecological systems prohibits the anticipation of all possible issues that could arise in the design, implementation, and analysis phases. Consultation with a statistician in the design and analysis stages may circumvent design issues or analytical errors not discussed in this manual.

References

- Cassel, C. M., C-E. Särndal, and J. H. Wretman. 1983. Some uses of statistical models in connection with the nonresponse problem. Pages 143-160 in W. G. Madow, I. Olkin, and D. B. Rubin, editors. Incomplete data and in sample surveys. (Vol. 3). Academic Press, New York.
- Cochran, W. G. 1977. Sampling techniques, 3rd edition. John Wiley and Sons, New York.
- Cohen, J. 1988. Statistical power analysis for the behavioral sciences, 2nd edition. Lawrence Erlbaum Associates, Hillsdale, New Jersey.
- Fancy, S. G., J. E. Gross, and S. L. Carter. 2009. Monitoring the condition of natural resources in US national parks. *Environmental Monitoring and Assessment* 151:161-174.
- Horvitz, D. G., and D. T. Thompson. 1952. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 47:663-685.
- Johnson, D. H. 2012. Monitoring that matters. In R. A. Gitzen, J. J. Millspaugh, A. B. Cooper, and D. S. Licht, editors. Design and analysis of long-term ecological monitoring studies. Cambridge University Press, England, U.K.
- Kincaid, T. M., and A. R. Olsen. 2013. *spsurvey: Spatial Survey Design and Analysis*. R package version 2.6. Available online: <http://www.epa.gov/nheerl/arm/> (accessed December 2015).
- Larsen, D. P., A. R. Olsen, and D. L. Stevens. 2008. Using a master sample to integrate stream monitoring programs. *Journal of Agricultural, Biological, and Environmental Statistics* 13:243-254.
- Lessler, J. T., and W. D. Kalsbeek. 1992. Nonsampling errors in surveys. John Wiley and Sons, New York.
- Little, J. A. R., and D. B. Rubin. 2002. Statistical analysis with missing data, 2nd edition. John Wiley and Sons, Inc., New Jersey.
- Lohr, S. 1999. Sampling: design and analysis. Duxbury Press, Pacific Grove, California.
- Manly, B. F. J. 2009. Statistics for environmental science and management, 2nd edition. Chapman & Hall/CRC, Boca Raton, Florida.
- Mazor, R. D., A. C. Rehn, P. R. Ode, M. Engeln, K. C. Schiff, E. D. Stein, D. Gillett, and C. P. Hawkins. 2016. Bioassessment in complex environments: designing an index for consistent meaning in different settings. Freshwater Science, 0-0. Retrieved from <http://www.jstor.org/stable/10.1086/684130>.
- McDonald, T. 2003. Review of environmental monitoring methods: Survey design. *Environmental Monitoring and Assessment* 85:277–292.

- McDonald, T. 2012. Spatial sampling designs for long-term ecological monitoring. In R. A. Gitzen, J. J. Millspaugh, A. B. Cooper, and D. S. Licht, editors. *Design and analysis of long-term ecological monitoring studies*. Cambridge University Press, England, U.K.
- McDonald, T. 2015. SDraw: Sample Draw. R package version 1.5.1 (accessed December 2015).
- Oh, H. L., and F. J. Scheuren. 1983. Weighting adjustment for unit nonresponse. Pages 143-184 in W. G. Madow, I. Olkin, and D. B. Rubin, editors. *Incomplete data and in sample surveys*. (Vol. 2). Academic Press, New York.
- Olsen, A. R., T. M. Kincaid, and Q. Payton. 2012. Spatially balanced survey designs for natural resources. In R. A. Gitzen, J. J. Millspaugh, A. B. Cooper, and D. S. Licht, editors. *Design and analysis of long-term ecological monitoring studies*. Cambridge University Press, England, U.K.
- Pan, J., and N. Tallent. 2012. Mojave Desert Network Integrated Upland Monitoring: 2012 pilot study for Joshua Tree National Park. National Park Service, Boulder City Nevada. Natural Resource Technical Report NPS/MOJN/NRTR—2014/850.
- R Core Team. 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available online: <http://www.R-project.org/> (accessed December 2015).
- Reynolds, J. H. 2012. An overview of statistical considerations. In R. A. Gitzen, J. J. Millspaugh, A. B. Cooper, and D. S. Licht, editors. *Design and analysis of long-term ecological monitoring studies*. Cambridge University Press, England, U.K.
- Robertson, B. L., J. A. Brown, T. McDonald, and P. Jaksons. 2013. BAS: Balanced acceptance sampling of natural resources. *Biometrics* 69:776-784.
- Särndal, C-E., B. Swensson, and J. Wretman. 1992. *Model assisted survey sampling*. Springer, New York.
- Schreuder, H. T., T. G. Gregoire, and J. P. Weyer. 2001. For what applications can probability and non-probability sampling be used? *Environmental Monitoring and Assessment* 66:281-291.
- Sims, M., S. Wanless, M. P. Harris, P. I. Mitchell, and D. A. Elston. 2006. Evaluating the power of monitoring plot designs for detecting long-term trends in the numbers of common guillemots. *Journal of Applied Ecology* 43:537-546.
- Stevens, D. L., and A. R. Olsen. 2003. Variance estimation for spatially balanced samples of environmental resources. *Environmetrics* 14:594-610.
- Stevens, D. L., and A. R. Olsen. 2004. Spatially balanced sampling of natural resources. *Journal of the American Statistical Association* 99(465):262-278.
- Thompson, S. K. 2002. *Sampling*, 2nd edition. John Wiley and Sons, New York.

Tiszler, J., D. Rodriguez, K. Lombardo, T. Sagar, L. Lee, T. Handley, K. McEachern, L. Starcevich, and T. Philippi. In press. Terrestrial vegetation monitoring in the Mediterranean Coast Network: Cabrillo National Monument, Channel Islands National Park, and Santa Monica Mountains National Recreation Area. Natural Resource Report NPS/XXXX/NRR—20XX/XXX. National Park Service, Fort Collins, Colorado.

Turner, A. G. 2003. Sampling frames and master samples. Expert group meeting to review the draft handbook on designing of household sample surveys. United Nations Secretariat. Statistics Division.

Urquhart, N. S., and T. M. Kincaid. 1999. Designs for detecting trend from repeated surveys of ecological resources. *Journal of Agricultural, Biological, and Environmental Statistics* 4(4):404-414.

Valliant, R., J. A. Dever, and F. Kreuter. 2013. Practical tools for designing and weighting survey samples. Springer, New York.

Appendix A: Weight Adjustment for Nonsampling Error

Introduction

The SDrawNPS package provides tools for analyzing data from a GRTS sample including an option to adjust sampling weights when samples are subject to nonsampling error. Three main sources of nonsampling error are nonresponse error, frame coverage error, and measurement error (Lessler and Kalsbeek 1992). Design weights are impacted by nonresponse error and frame error, so measurement error will not be addressed in this setting. Nonresponse error occurs when a unit in the sample does not provide a complete outcome of interest, such as when a site is inaccessible to observer crews. However, when the original sample is allocated to panels and not completely surveyed in a single year, the design weights from the original sample must be adjusted to account for the sites not surveyed in the current year. Frame error occurs when the sampling frame does not match the target population. Non-target sites that were erroneously included in the sample must be addressed in the analysis phase for unbiased design-based inference. In this appendix, the methods used to adjust the sampling weights are outlined, and considerations on nonsampling error and inferential scope are discussed.

A primary consideration in any discussion of missing observations is the nature of how they were generated. When data are missing completely at random (MCAR), the missingness mechanism is by definition unrelated to the outcome of interest, the sampling design, or any covariates (Little and Rubin 2002). In this case, the responding portion of the sample may be treated as a random sample from the population. When observations are missing at random given covariates (MAR), the missingness can be explained by related covariates. Techniques such as weighting class adjustment and post-stratification adjustment may be used to obtain unbiased model-assisted estimates if the covariates are available for units in the sample and if sample size for each level of the covariate(s) is sufficient (Oh and Scheuren 1983). When observations are not missing at random (NMAR), the missing observations are significantly different from the non-missing observations and inference based only on the non-missing portion of the sample may provide biased inference. In this case, a separate non-respondent sample or model is needed to provide additional information. Nonresponse adjustments are an added complication. Data from ecological surveys are often insufficient for weighting adjustments because at least five respondents and five non-respondents are needed within each weighting class. Modeling NMAR nonresponse is beyond the scope of this project. In the SDrawNPS package, weights are adjusted assuming MCAR only. More advanced weighting adjustments can be performed outside of SDrawNPS and used directly in analysis.

The nature of frame errors must also be considered for unbiased design-based estimation. Overcoverage occurs when the sampling frame contains sampling units that are not members of the targeted population. Undercoverage occurs when members of the target population are omitted from the sampling frame. Frame undercoverage is generally not assessed in monitoring programs because survey effort is not expended outside of the sampling frame. However, frame overcoverage can be assessed when observers visit a site. If the site does not meet the criteria for inclusion in the target population, the site is recorded as “Non-target.” The proportion of “Non-target” sites in the sample can then be used to reduce the extent of the target population for inference.

A final consideration is the scope of inference. The analysis must assume that outcomes from the missing units are not significantly different from those from the observed units in order to make inferences to the target population. If this assumption is not valid, inference is limited to the subset of the target population that would respond if sampled, or the “sampled population.” Note that the sampled population is a subset of the target population and represents the subset of the target population for which an outcome would be observed if included in the sample. If missingness is not MCAR but data are insufficient for more advanced weighting adjustments, the safest approach is to limit inference to the sampled population. However, it is not always possible to identify the specific units that define this subpopulation of the target population.

Weighting Adjustment

To describe the weighting adjustments made by SDrawNPS under the MCAR assumption, let:

- n be the initial number of units (here, units are sites) from the main sample
- n_E be the number of units evaluated in the sample ($n_E \leq n$).
- n_T be the number of units within the set of evaluated sites that meet the definition of the target population ($n_T \leq n_E$).
- n_R denote the number of units that were determined to meet the definition of the target population and from which an outcome (e.g., habitat metric) was obtained ($n_R \leq n_T$)

Let the probability of including site i in the initial sample (of size n) be π_i , and the inclusion weight (or design weight) as its inverse, $w_i = 1/\pi_i$. These weights are produced by SDrawNPS when the sample is drawn. The notation may be extended as needed with subscripting for strata, domains of interest, or replication over time.

The extent of the sampling frame is measured as the number of individual units in a finite sampling frame, the total length of features in a linear resource, or the total area of the features in the GIS coverage of an area sampling frame. Let the size of the sampling frame, N , represent the number of features in the sampling frame before consideration of nonsampling error. Let the extent of each feature in the sampling frame be represented by $|R_i|$, such that the total extent of the sampling frame is given by $|R| = \sum_{i=1}^N |R_i|$. For a finite sampling frame, each member of the population is represented

by a point with no dimension so the extent is assumed to be 1. Therefore, the extent of a finite sampling frame is $|R| = \sum_{i=1}^N 1 = N$. For a linear sampling frame, such as a stream network, the extent of the sampling frame is the total linear stream length across all stream features in the frame. Similarly, the extent of an areal sampling frame is the sum of the area of each polygon in the sampling frame.

When frame overcoverage occurs, the size and extent of the frame are reduced to reflect the proportion of evaluated sites that met the definition of the target population, $\frac{n_T}{n_E}$. Let $|R|^* = \frac{n_T}{n_E}|R|$ represent the reduced frame extent after this adjustment for frame error. N^* represents an estimate of the number of units in the target population, but the specific geographic locations of the (approximate) N^* units in the target population cannot be known unless missingness is related to known covariates.

For inference to the target population, define $w_i^T = \frac{w_i}{\left(\frac{n_E}{n}\right)\left(\frac{n_R}{n_T}\right)}$, and for inference to the sampled

population, define $w_i^S = \frac{w_i}{\left(\frac{n_E}{n}\right)}$. Details of this adjustment for each of the general sampling designs

follow.

Equal probability sampling

The unadjusted design weight from the equiprobable GRTS sample draw is:

$$w_i^E = \frac{|R|}{n}.$$

When equal probability sampling is used and inference may be made to the target population, the adjusted design weight accounts for the number of evaluated sites and is defined as:

$$w_i^{ET} = \frac{w_i^E}{\left(\frac{n_E}{n}\right)\left(\frac{n_R}{n_T}\right)} = \frac{|R|}{\frac{n_E n_R}{n_T}} = \frac{|R|^*}{n_R}.$$

Weights should fulfill the requirement that $\sum_{i=1}^{n_R} w_i = |R|$. When nonresponse occurs, the summation occurs over the n_R responding sites, and under equal probability sampling,

$$\sum_{i=1}^{n_R} w_i^{ET} = n_R \frac{|R|}{\frac{n_E n_R}{n_T}} = \frac{n_T}{n_E} |R| = |R|^*,$$

which is a necessary condition and represents the estimated extent of the target population after adjustment for frame error.

When equal probability sampling is used and inference is made only to the sampled population, the adjusted design weight is:

$$w_i^{ES} = \frac{w_i^E}{\left(\frac{n_E}{n} \right)} = \frac{|R|}{n_E}.$$

Under equal probability sampling and assuming inference to the sampled population, we have that:

$$\sum_{i=1}^{n_R} w_i^{ES} = \sum_{i=1}^{n_R} \frac{|R|}{n_E} = \frac{n_R |R|}{n_E} = \frac{n_R |R|^*}{n_T}.$$

This value is the extent of the sampled population (the responding portion of the target population), which is less than the adjusted target population extent by a factor equal to the response rate $\frac{n_R}{n_T}$.

Note that we do not explicitly adjust for overcoverage error in the weighting adjustment. The adjustment in each case reduces the sampling frame appropriately for each scope of inference.

Unequal-probability sampling

When GRTS unequal probability sampling is used, the unadjusted design weights output by SDrawNPS for unit i in class c is,

$$w_i^U = w_c^U = \frac{|R_c|}{n_c},$$

where $|R_c|$ is the extent of the target population within class c and n_c is the initial sample size specified for class c . Note that the actual number of sites selected in class c , n_c^* , is a random variable, and the sample size specified for class c defines the inclusion weight for all units in that class. The extent of the target population is calculated as:

$$\sum_{i=1}^n w_c^U = \sum_{i=1}^n \frac{|R_c|}{n_c} = \sum_{c=1}^C \frac{n_c^* |R_c|}{n_c}.$$

Notice that, unlike equal probability sampling, the sum of the weights does not give an exact calculation of the extent of the target population within each category due to the random sample sizes within each category. We assume that, under repeated sampling, and for samples that are a small fraction of the total population within each class c , the expectation of the random sample size within each category is equal to the designated sample size, or:

$$E(n_c^*) = n_c.$$

Note that, as the sample sizes within each class increase, the relative proportions of the sample within each class converge to the relative proportions of the extent in the target population. Regardless of how the relative sample sizes are designated among the C categories, the sample size cannot exceed the resolution (number of pixels in the sampling frame) of the frame extent within a given category.

Applying the methods used for equal-probability sampling for inference to the target population, the adjusted design weight for all evaluated sites is defined as:

$$w_i^{UT} = w_c^{UT} = \frac{w_c^U}{\left(\frac{n_E}{n} \right) \left(\frac{n_R}{n_T} \right)} = \frac{|R_c|}{\frac{n_c n_E n_R}{nn_T}} = \frac{|R_c|^*}{\hat{n}_{cR}},$$

where $|R_c|^* = \frac{n_T}{n_E} |R_c|$ is the extent of the target population within class c after adjusting for

overcoverage error for the entire sample and $\hat{n}_{cR} = \left(\frac{n_E}{n} \right) \left(\frac{n_T}{n_E} \right) \left(\frac{n_R}{n_T} \right) n_c = \left(\frac{n_R}{n} \right) n_c$ is the expected

number of respondents in the sample within category c . Using this weight to obtain the extent of the population in category c , we have:

$$\sum_{i=1}^{n_R} w_i^{UT} = \sum_{i=1}^{n_R} \frac{|R_c|}{\frac{n_c n_E n_R}{\hat{n}_{cR}}} = \sum_{i=1}^{n_R} \frac{|R_c|^*}{\hat{n}_{cR}} = \sum_{c=1}^C \frac{n_{cR}^* |R_c|^*}{\hat{n}_{cR}},$$

where n_{cR}^* is the number of responding units in class c . The asterisk notation is to emphasize that the number of units in category c of the sample is n_c^* which is not necessarily equal to n_c in unequal-probability sampling. Therefore, if the response rate is equal among all categories, we have $n_{cR}^* \approx \hat{n}_{cR}$ and $\sum_{i=1}^{n_R} w_i^{UT} \approx |R|^*$. Therefore, the validity of this weighting approach depends on the assumption that the response rate is independent of and roughly equal among the C categories.

Similarly for the case when inference is made only to the sampled population, the adjusted design weight is:

$$w_i^{US} = w_c^{US} = \frac{w_c^U}{\left(\frac{n_E}{n} \right)} = \frac{n}{n_E} \frac{|R_c|}{n_c}.$$

Therefore the weights are increased to account for the proportion of the original sample that was actually evaluated. The extent of the sampled population within class c after adjusting for overcoverage error is calculated as:

$$\sum_{i=1}^{n_R} w_i^{US} = \sum_{i=1}^{n_R} \frac{n}{n_E} \frac{|R_c|}{n_c} = \frac{n}{n_E} \sum_{i=1}^{n_R} \frac{|R_c|}{n_c} = \frac{n}{n_E} \sum_{c=1}^C \frac{n_{Rc}^* |R_c|}{n_c} = \frac{n}{n_T} \sum_{c=1}^C \frac{n_{Rc}^* |R_c|^*}{n_c}.$$

Continuous-probability sampling

When continuous-probability sampling is used and inference is made to the target population, the unadjusted design weight from the GRTS sample draw is:

$$w_i^C = \frac{\sum_{i=1}^N |R_i| X_i}{nX_i}.$$

This method of sampling is appropriate when the sample size, n , is small relative to the population size, N . Otherwise, the weights, w_i^C , can take values less than one. The estimate of the extent of the resource for this sampling approach is:

$$\sum_{i=1}^n w_i^C = \sum_{i=1}^n \frac{\sum_{i=1}^N |R_i| X_i}{nX_i} = \left(\sum_{i=1}^N |R_i| X_i \right) \left(\frac{1}{n} \sum_{i=1}^n \frac{1}{X_i} \right).$$

Notice that, unlike equal probability sampling and similar to unequal probability sampling, the sum of the design weights does not give an exact calculation of the extent of the target population due to the variability in the continuous variable used to define the inclusion probabilities.

The adjusted design weight for inference to the target population accounts for the number of evaluated sites and is defined as:

$$w_i^{CT} = \frac{\frac{\sum_{i=1}^N |R_i| X_i}{nX_i}}{\left(\frac{n_E}{n} \right) \left(\frac{n_R}{n_T} \right)} = \frac{\frac{n_T}{n_R} \sum_{i=1}^N |R_i| X_i}{n_E X_i}.$$

Recall that $\sum_{i=1}^{n_R} w_i^{CT}$ must equal the extent of the target population after adjusting for overcoverage error. For continuous-probability sampling with nonresponse and frame overcoverage, we have:

$$\sum_{i=1}^{n_R} w_i^{CT} = \sum_{i=1}^{n_R} \frac{\frac{n_T}{n_E} \sum_{i=1}^N |R_i| X_i}{n_E X_i} = \left(\frac{n_T}{n_E} \sum_{i=1}^N |R_i| X_i \right) \left(\frac{1}{n_E} \sum_{i=1}^{n_R} \frac{1}{X_i} \right).$$

Note that the sum of the continuous variable for the population is now scaled down by the proportion of target sites among the total number evaluated. Also the mean of the inverse of the continuous variable for the sample is taken for only the responding sites.

The adjusted design weight for inference to the sampled population is defined as:

$$w_i^{CS} = \frac{w_i^C}{\left(\frac{n_E}{n}\right)} = \frac{\sum_{i=1}^N |R_i| X_i}{\left(\frac{n_E}{n}\right)} = \frac{\sum_{i=1}^N |R_i| X_i}{n_E X_i}.$$

Recall that $\sum_{i=1}^{n_R} w_i^{CS}$ must equal the extent of the target population after adjusting for overcoverage error. For continuous-probability sampling with nonresponse and frame overcoverage, we have:

$$\sum_{i=1}^{n_R} w_i^{CS} = \sum_{i=1}^{n_R} \frac{\sum_{i=1}^N |R_i| X_i}{n_E X_i} = \left(\frac{n_R}{n_E} \sum_{i=1}^N |R_i| X_i \right) \left(\frac{1}{n_R} \sum_{i=1}^{n_R} \frac{1}{X_i} \right).$$

The sum of the continuous variable for the population is scaled further in this case by the proportion of responding sites of the total number evaluated. As with w_i^T , the mean of the inverse of the continuous variable for the sample is taken for only the responding sites. Note that the extent of the sampled population is smaller than the extent of the target population by a factor equivalent to the response rate, $\frac{n_R}{n_T}$, or:

$$\frac{\sum_{i=1}^{n_R} w_i^{CS}}{\sum_{i=1}^{n_R} w_i^{CT}} = \frac{\left(\frac{n_R}{n_E} \sum_{i=1}^N |R_i| X_i \right) \left(\frac{1}{n_R} \sum_{i=1}^{n_R} \frac{1}{X_i} \right)}{\left(\frac{n_T}{n_E} \sum_{i=1}^N |R_i| X_i \right) \left(\frac{1}{n_R} \sum_{i=1}^{n_R} \frac{1}{X_i} \right)} = \frac{n_R}{n_T}$$

Design-based inference

In the *spsurvey* package, the *cat.analysis* function may be used to analyze an indicator of frame membership to estimate the target population extent, $|R|^*$. Estimates of the target population extent by each of the C categories may also be obtained with the *cat.analysis* function by estimating the total for an indicator of target population inclusion (1 if the unit is a member of the target population and 0 otherwise) and specifying estimation by the subpopulation of interest.

For status estimation of both means and totals, weight is assigned only to the responding sites. Estimates of the mean and its standard error are the same for inference to both the target population and the sampled population. However, the estimated total for the target population is larger than that of the sampled population because the extent of the target population is larger than the extent of the sampled population when nonresponse occurs. Therefore, inference to the target population includes sites that are nonresponsive. Inference to the sampled population represents a more conservative approach that might serve as a lower bound for inference on abundance. Therefore, reporting both estimates may provide useful information to managers.

For the equiprobable sample without nonsampling error, the population mean, μ , is estimated as:

$$\hat{\mu} = \frac{\sum_{i=1}^{n_E} y_i w_i}{\sum_{i=1}^{n_E} w_i} = \frac{\sum_{i=1}^{n_E} y_i \frac{|R|}{n_E}}{\sum_{i=1}^{n_E} \frac{|R|}{n_E}} = \frac{\frac{|R|}{n_E} \sum_{i=1}^{n_E} y_i}{\frac{|R|}{n_E} n_E} = \frac{\sum_{i=1}^{n_E} y_i}{n_E} = \bar{y},$$

where \bar{y} is the mean of the n_E evaluated sampling units.

For inference to the target population when nonsampling error occurs, the mean is estimated as:

$$\hat{\mu}^T = \frac{\sum_{i=1}^{n_R} y_i w_i^{ET}}{\sum_{i=1}^{n_R} w_i^{ET}} = \frac{\sum_{i=1}^{n_R} y_i \frac{|R|}{n_E n_R}}{\sum_{i=1}^{n_R} \frac{|R|}{n_E n_R}} = \frac{\frac{|R|}{n_E n_R} n_T \sum_{i=1}^{n_R} y_i}{\frac{|R|}{n_E n_R} n_R} = \frac{\sum_{i=1}^{n_R} y_i}{n_R} = \bar{y},$$

where \bar{y} is the mean of the outcome of interest for the n_R responding units.

For inference to the sampled population when nonsampling error occurs, the mean is estimated as:

$$\hat{\mu}^S = \frac{\sum_{i=1}^{n_R} y_i w_i^{ES}}{\sum_{i=1}^{n_R} w_i^{ES}} = \frac{\sum_{i=1}^{n_R} y_i \frac{|R|}{n_E}}{\sum_{i=1}^{n_R} \frac{|R|}{n_E}} = \frac{\frac{|R|}{n_E} \sum_{i=1}^{n_R} y_i}{\frac{|R|}{n_E} n_R} = \frac{\sum_{i=1}^{n_R} y_i}{n_R} = \bar{y}.$$

Therefore, for estimation of the mean, the weighting does not impact inference. However, appropriate weighting is very important for estimates of the population total. For inference to the target population when no nonsampling error occurs for an equiprobable sample, the population total, τ , is estimated as:

$$\hat{\tau} = \sum_{i=1}^{n_E} y_i w_i = \sum_{i=1}^{n_E} y_i \frac{|R|}{n_E} = |R| \bar{y}.$$

For inference to the target population when nonsampling error occurs for an equiprobable sample, the population total, is estimated as:

$$\hat{\tau}^T = \sum_{i=1}^{n_R} y_i w_i^{ET} = \sum_{i=1}^{n_R} y_i \frac{|R|}{n_E n_R} = \frac{|R| n_T}{n_E n_R} \sum_{i=1}^{n_R} y_i = \frac{n_T}{n_E} |R| \bar{y} = |R|^* \bar{y}.$$

For inference to the sampled population when nonsampling error occurs for an equiprobable sample, the population total is estimated as:

$$\hat{\tau}^S = \sum_{i=1}^{n_R} y_i w_i^{ES} = \sum_{i=1}^{n_R} y_i \frac{|R|}{n_E} = \frac{n_R}{n_E} |R| \bar{y} = \frac{n_R}{n_E} \frac{n_E}{n_T} |R|^* \bar{y} = \frac{n_R}{n_T} |R|^* \bar{y}.$$

Therefore, estimates of the population total for inference to the sampled population is smaller than the population total for inference to the target population by a factor equivalent to the response rate,

$\frac{n_R}{n_T}$, or:

$$\frac{\hat{\tau}^S}{\hat{\tau}^T} = \frac{\frac{n_R}{n_T} |R|^* \bar{y}}{|R|^* \bar{y}} = \frac{n_R}{n_T}.$$

These results generalize to the unequal- and continuous probability cases as well when the appropriate population extents are applied.

The Department of the Interior protects and manages the nation's natural resources and cultural heritage; provides scientific and other information about those resources; and honors its special responsibilities to American Indians, Alaska Natives, and affiliated Island Communities.

NPS 963/133314, June 2016



Natural Resource Stewardship and Science

1201 Oakridge Drive, Suite 150
Fort Collins, CO 80525

www.nature.nps.gov

2016

National Park Service
CENTENNIAL

EXPERIENCE YOUR AMERICA™