Aquatic Resources Monitoring: Survey Design, Analysis and Reporting Using EPA Tools

Garrett Stillings (USEPA-OWOW)

Contributions from:
Karen Blocksom and Michael Dumelle (USEPA-ORD)

2025 Northeast Aquatic Biologists Conference
February 5-7 in Bartlett, New Hampshire

A Quick Note

- The `spsurvey` R package, originally created in 2004, has been used extensively by NARS to carry out survey designs and analyze data
- The survey design and population estimates Shiny apps discussed today are user-friendly, non-R interfaces that actually use `spsurvey` "under the hood"
- `spsurvey` had a major update released in October, 2021 (version 5.0)

Section 1

Monitoring

Clean Water Act

- Monitoring informs progress achieving the Clean Water Act Goals
  - Restore and maintain the chemical, physical and biological integrity of the Nation's waters
    - **Section 101(a)**
  - Water quality provides for the protection and propagation of fish, shellfish and wildlife and recreation in and on the water
    - **Section 101(a)(2)**
- Monitoring supports implementation of CWA programs
  - Report on extent of waters supporting CWA goals (**305b**)
  - Extent water quality is changing over time (**305b, 314**)
  - Leading causes of degradation (**305b, 303d**)
  - List of waters needing TMDLs to achieve goals (**303d**)
  - Level of protection needed to achieve goals (**303c, 303d, 402, 319**)

Common Monitoring Types

1. **Census:** Sample every unit in the population.
2. **Targeted:** Sample a selected area(s) of interest.
3. **Probability:** Randomly sample a subset of population and use the sample to draw inferences regarding the whole population.

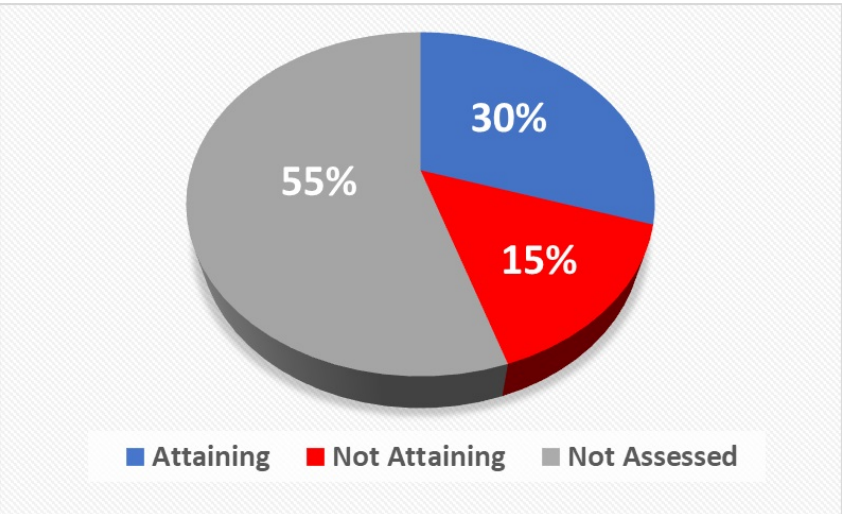


Figure 1: Targeted Monitoring

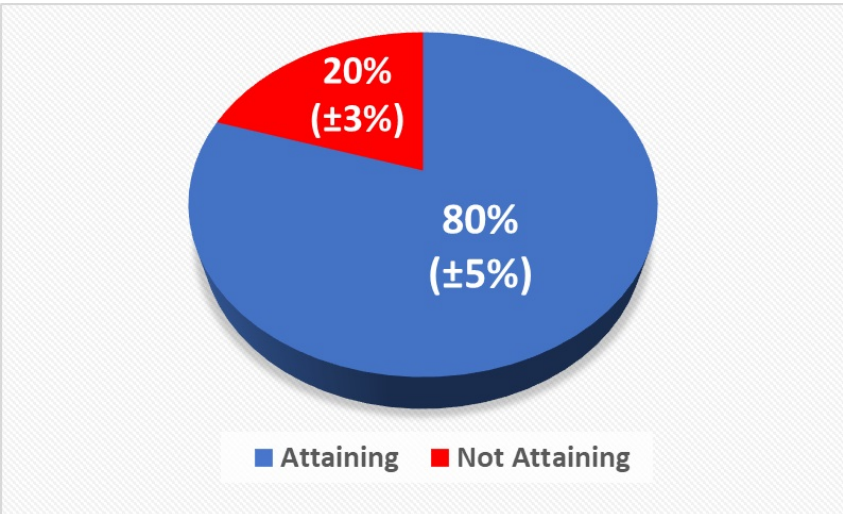


Figure 2: Probability Survey

Generalized Random Tessellation Stratified (GRTS) Design

- Spatial balance disperses the sampling effort across the extent of the resource so that samples achieve a similar spatial distribution as the population (Olsen et al. 2012).
  - More representative of the population.
  - More precise estimates of population parameters.

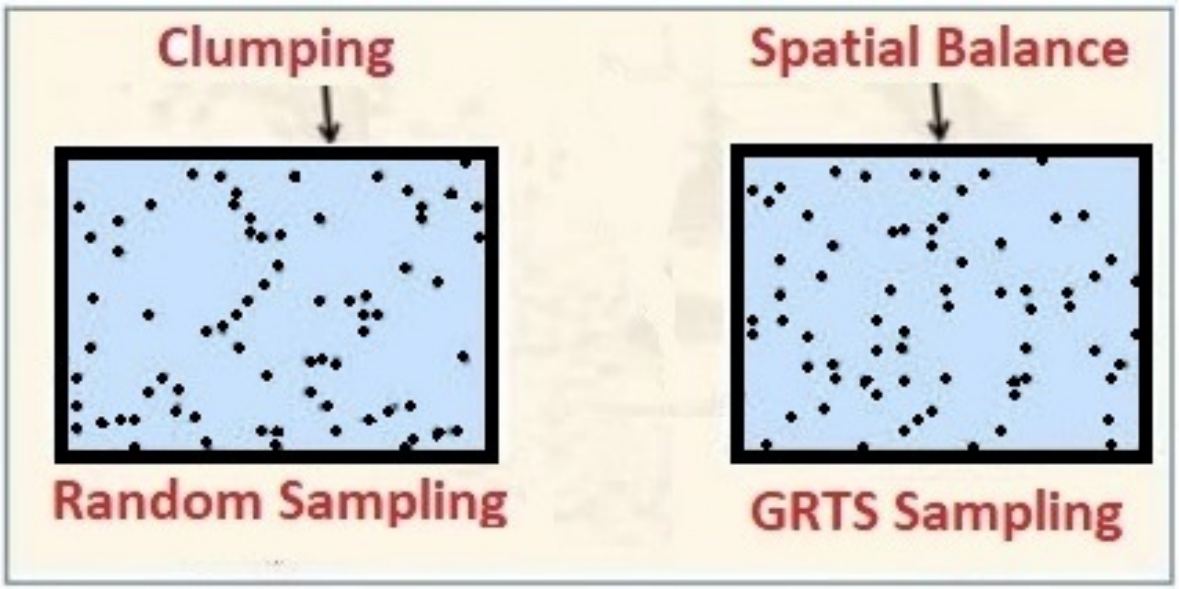


Figure 3: Spatial Balance

Section 2

Survey Design

Survey Design Tool

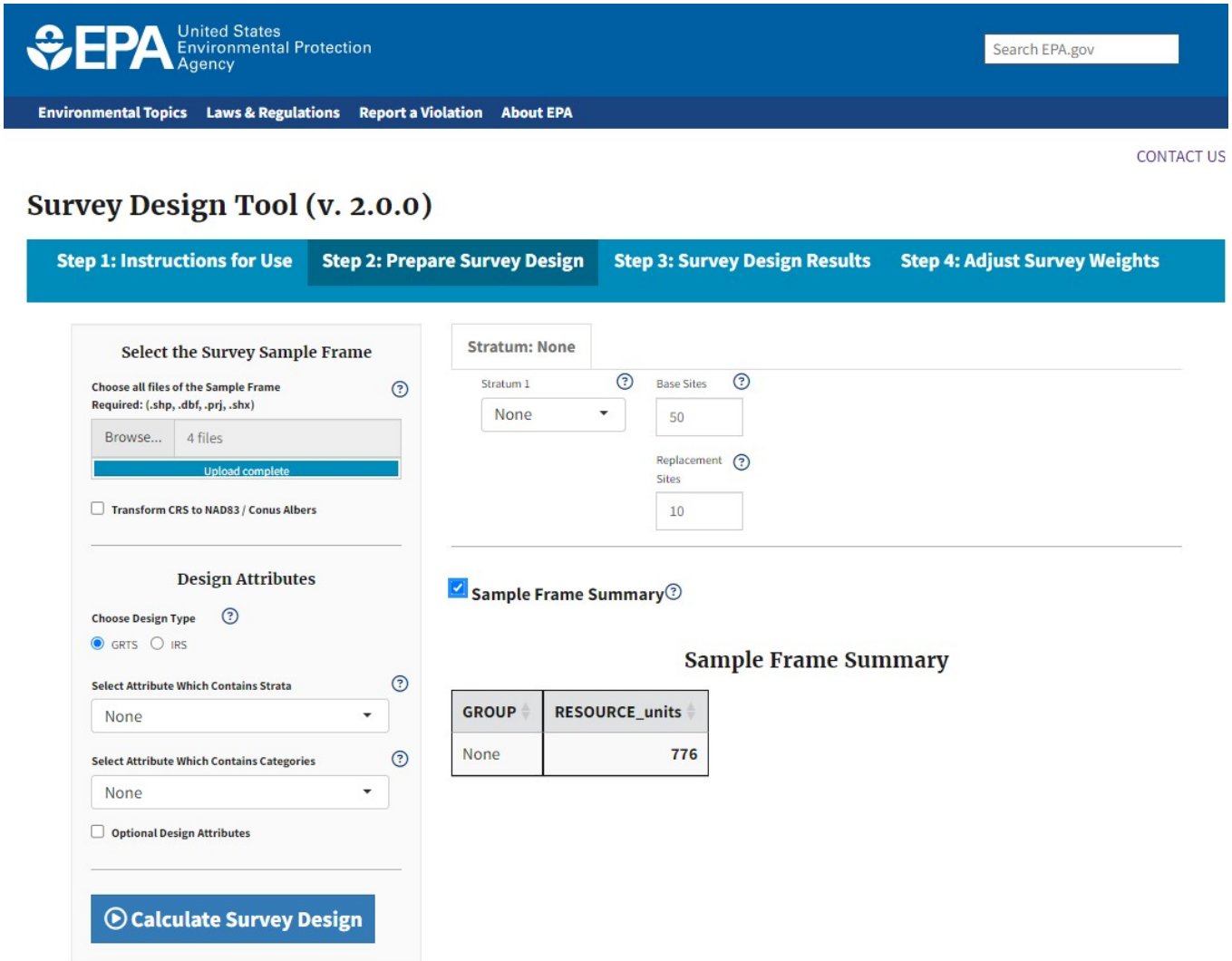


Figure 4: https://rconnect-public.epa.gov/survey-design-tool/

Define Survey Objectives

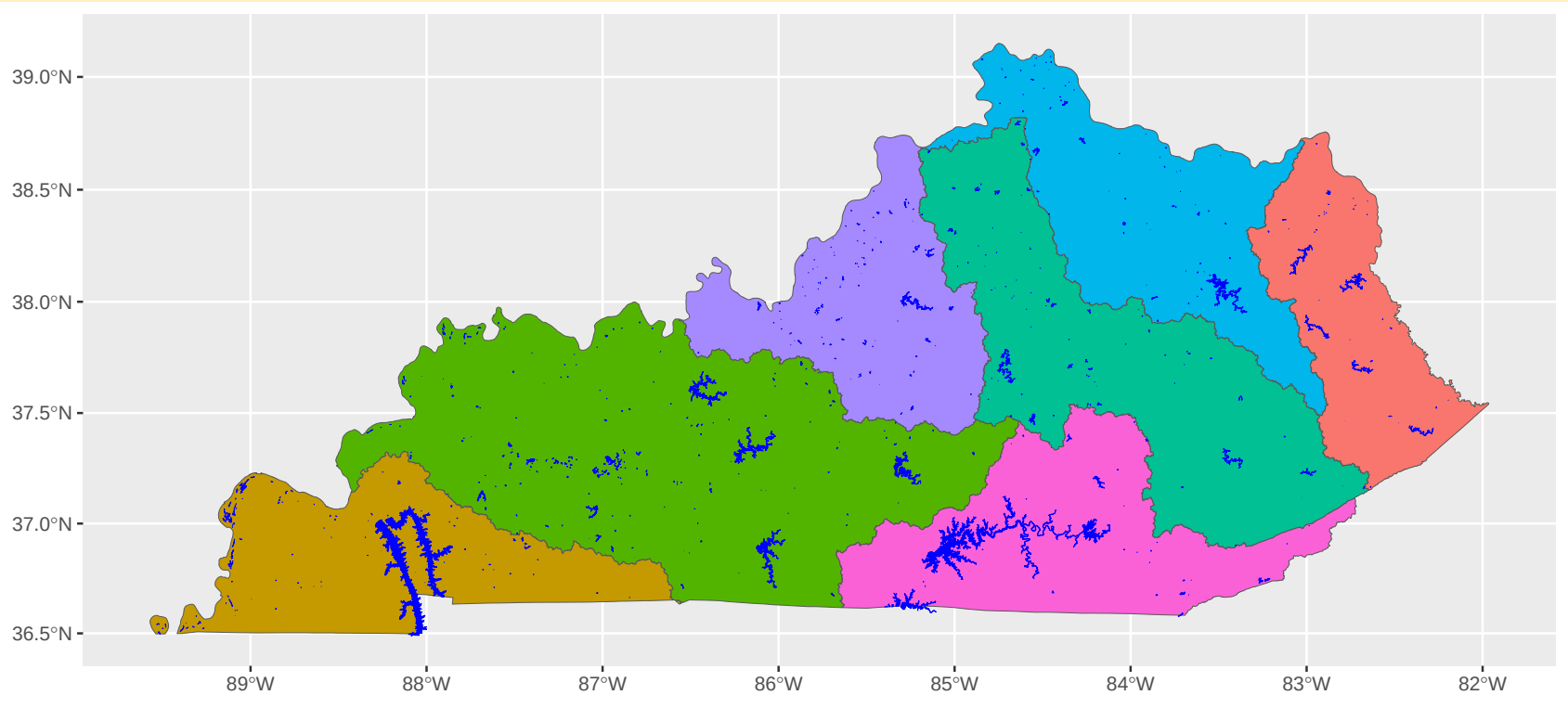
- Objectives should be linked to management decisions and reporting requirements.
- Ensure the level of effort will sufficiently answer the questions.
- Define the **target population** of interest.

Target Population

- The target population defines the complete collection of units that will be monitored.
- Must define what elements make up the target population (e.g., Lakes >1m depth).
- The target population should align with your organization's assessment goals and monitoring strategy.

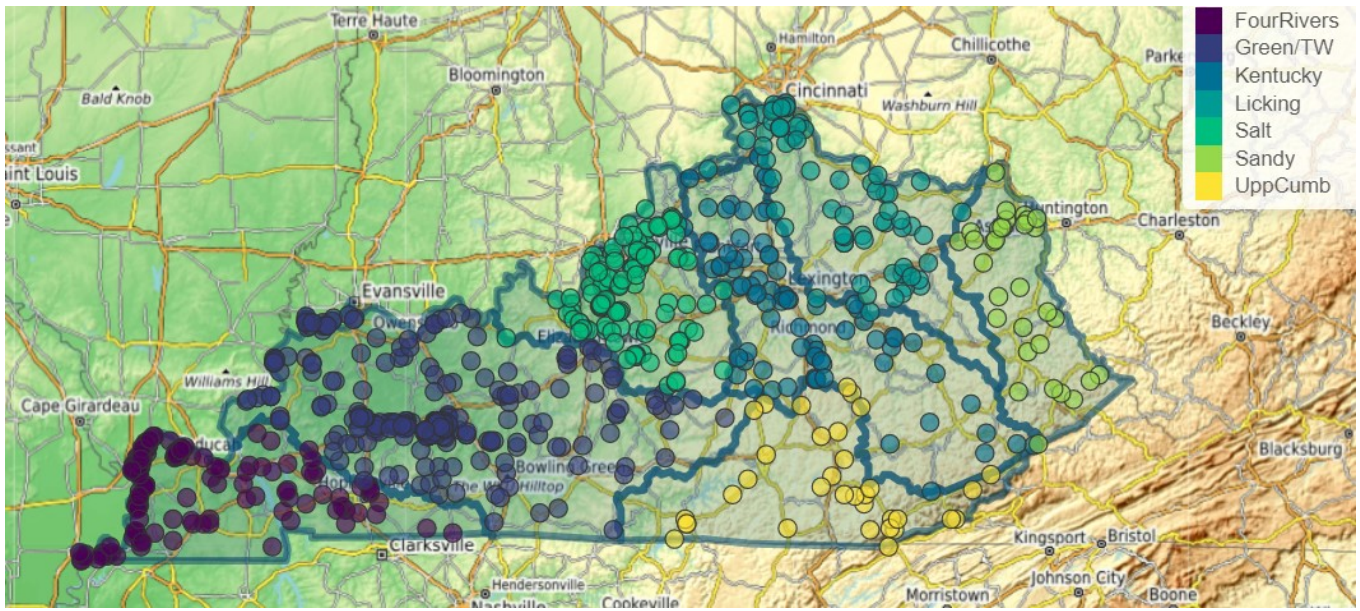
Target Population

Example - In this example, we chose to target all lakes in Kentucky which are publicly accessible. This consists of 776 lakes across seven river basins.



Sampling Frame

- A **sampling frame** is a list that contains all sites available to be sampled and intended to be in the target population.
- Sampling frames are often stored as GIS representations (e.g., ESRI shapefile) and contain additional information about the site that can be used in the design process:
  - Stratification variables (Stratified sampling)
  - Categorical/Continuous variables (Unequal or Proportional sampling)



Sampling Frame Types

1. **Discrete Objects:** Represented by **points** which serve as sampling units for selection (e.g. ponds and lakes treated as a whole).
2. **Linear Features:** Represented by **polylines** in which a sampling unit is selected along the networks length (e.g. river and stream length).
3. **Areas:** Represented by **polygons** where sampling units are selected within these regions (e.g. lake, wetland and coastline area).

Section 3

Spatially Balanced Sampling

Inclusion Probability Methods

1. **Equal Probability:** Selection where all members of the sample frame have equal probabilities of being selected.
2. **Unequal Probability:** Selection where the chance of being included is calculated relative to the proportion of a **Categorical Variable** (e.g., size class) across the population (user samples sizes are not guaranteed). Can give smaller populations a greater chance of being selected.

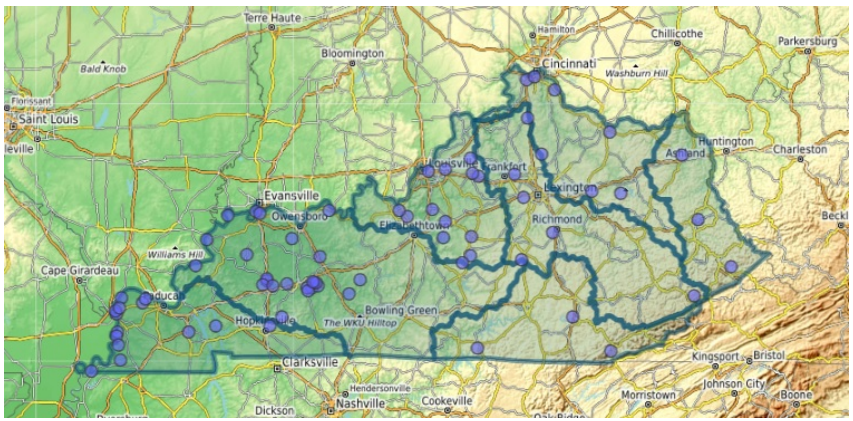


Figure 6: Equal Probability.

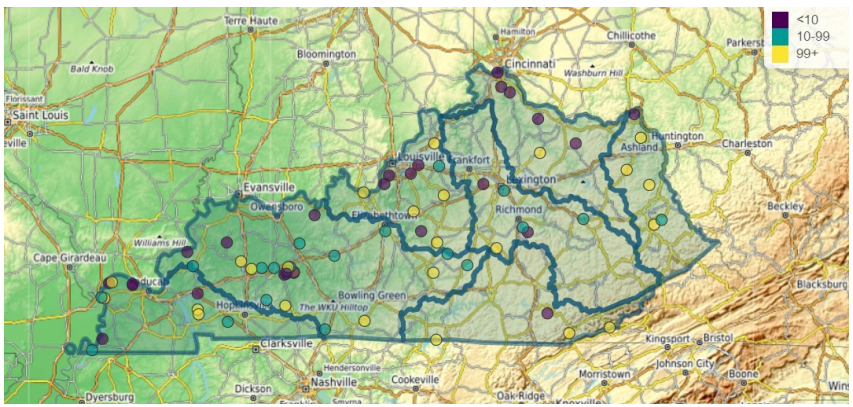


Figure 7: Unequal Probability.

Inclusion Probability Methods

3. **Proportional Probability Sampling:** Selection where the chance of being included is proportional to the values of a positive **Auxiliary Variable** (e.g., numeric size). Larger values of the auxiliary variable result in higher inclusion probabilities.

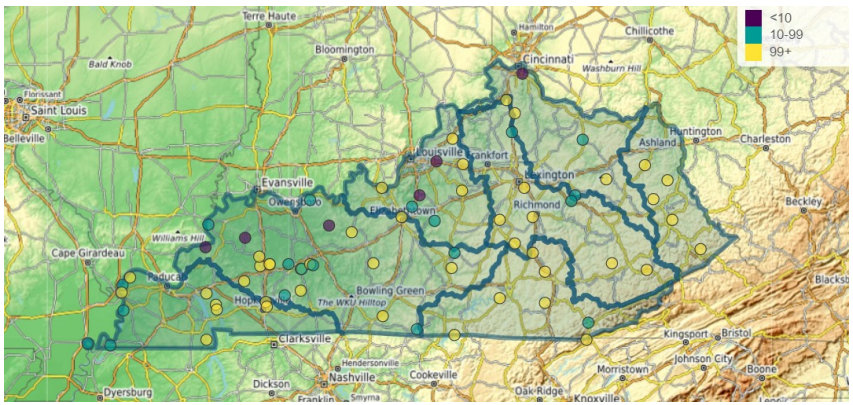


Figure 8: Proportional Sampling.

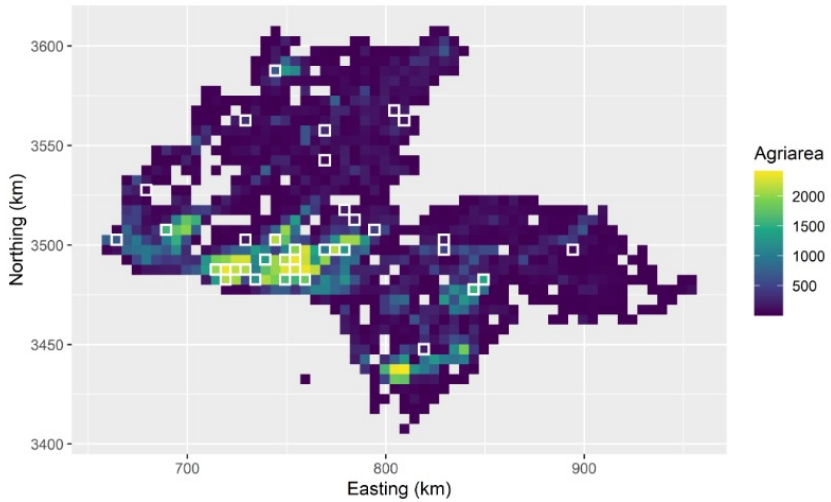


Figure 9: Image from: Dick Brus (2023). Spatial sampling with R.

Sample Selection

a. **Unstratified Sampling:** Random samples are drawn from the sample frame as a whole.
b. **Stratified Sampling:** The sample frame is divided into non-overlapping **Strata** (e.g., river basin) from which independent random samples are drawn.

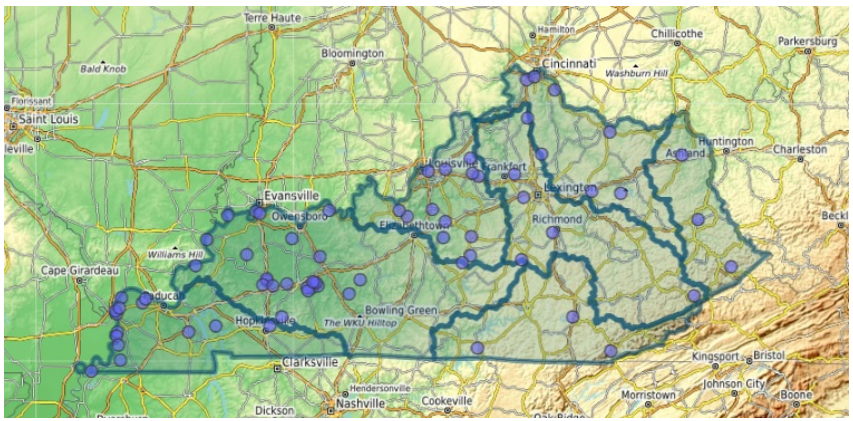


Figure 10: Unstratified Sampling (Equal Probability).

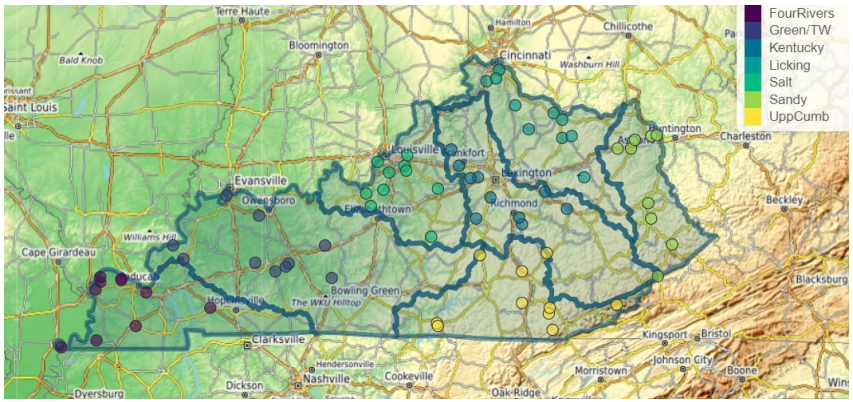


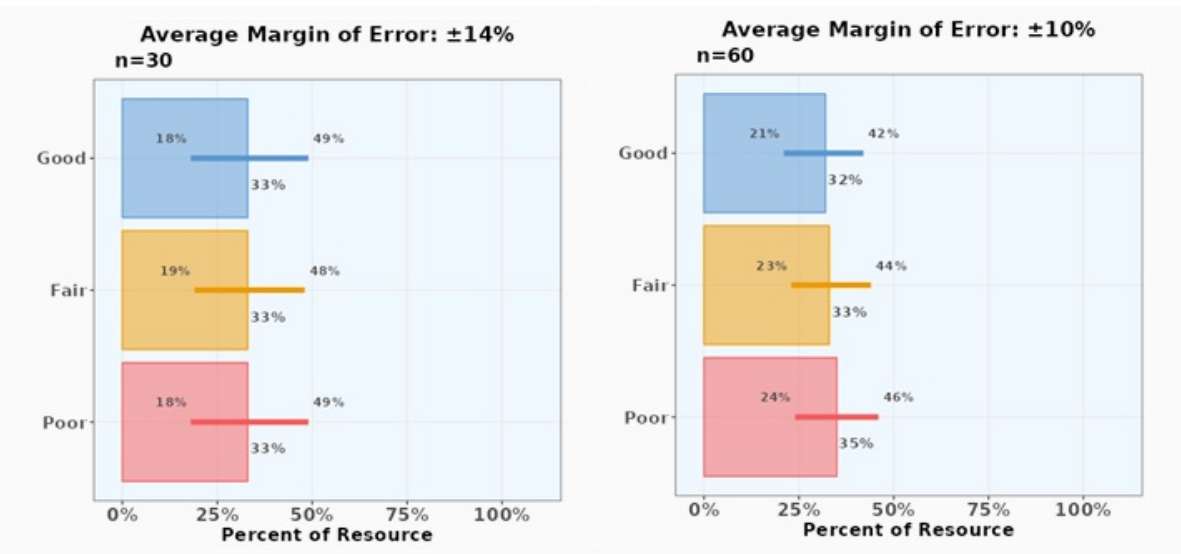
Figure 11: Stratified Sampling (Equal Probability).

Additional Sampling Methods

1. **Replacement Sampling:** Provides additional sites to replace sites in the base sample for which data cannot be collected. **Sites should be replaced with the next site in Site ID order within the same stratum (if applicable)**.
   - **Reverse Hierarchical Ordering (Default):** Sites are first selected using the GRTS algorithm then determined as base sites or replacement sites in a way that preserves as much spatial balance as possible.
   - **Nearest Neighbor:** Closest site (measured by Euclidean distance) to the base site (within a stratum).
2. **Legacy Sampling:** Sites that were selected in a previous random design and should be included in the current sample.
3. **Minimum Distance Sampling:** Enforces a minimum distance between sites.

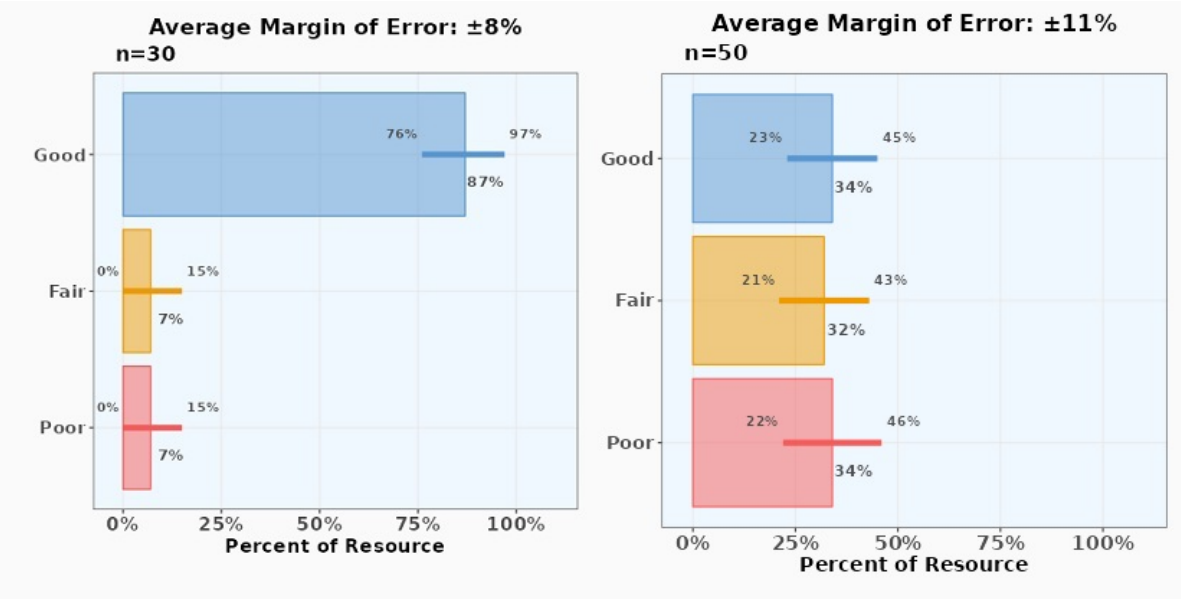
Determining the Sample Size

- Setting an appropriate sample size and considering how they should be allocated across a sample frame is a fundamental step in designing a successful survey.
- Many surveys are limited by budgetary and logistical constraints. The designer must determine a sample size which can overcome these constraints while ensuring the survey estimates the parameter(s) of interest with a low margin of error.



Sample Size Considerations

- Compare the **Spatial Balance** of surveys as estimates from spatially balanced surveys are more precise.
- Set a reasonable sample size of replacement sites for increased spatial balance and lower variance estimates.
- Consider if the parameter(s) of interest will result in low variation across the survey. A smaller sample size can yield an estimate with a low margin of error. Use the Survey Design tools **Population Estimate Simulation** to test.



Design Weights

- Design weights (i.e., sampling weights) are assigned to each site and **quantify the number of units** in the target population this site represents.
- Combined with the observed data to construct population estimates that represent the target population.

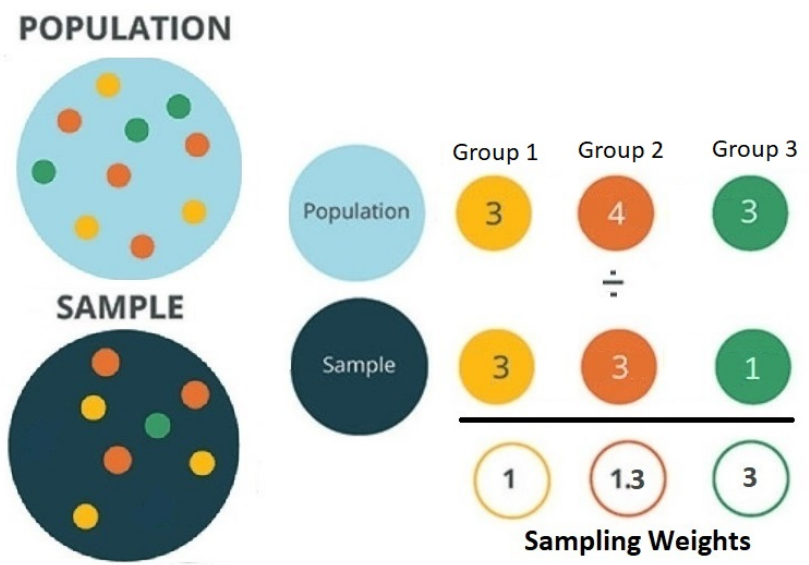


Figure 14: Design Weights.

Section 4

Adjusting Weights

Site Evaluation

- Often the design does not go exactly as planned and the original design weights must be adjusted so that they continue to accurately represent the target population.
- A sampling frame has **overcoverage** if it contains some sampling units that are missing from the target population and/or **undercoverage** if it is missing some sampling units present in the target population.

Site Evaluation Designations

1. **Target Site**: A site which was in the target population and was sampled.
2. **Non-Target Site**: A site which was **not** part of the target population and was **not** sampled.
3. **Non-Response Site**: A site which was in the target population and was **not** sampled (inaccessible site, landowner denial)

Design Weight Adjustments

- **Weight Category**: A weight adjustment category represents if a Stratum and/or a category was used in the design. Separate weight adjustments are applied to each initial weight so the final weights sum to a desired frame size.
- **Frame Size**: Size of the target population of the sampling frame. **Note** The Survey Design Tools Sample Frame Summary can assist.

Survey Design Tool Weight Adjustment Outputs

- WGT_TP_EXT: Weights based on all target and non-target sites (design as implemented).
- WGT_TP_CORE: Weights based on the target population. **Use WGT_TP_CORE for population estimates**

Section 5

Population Estimates

Population Estimate Tool

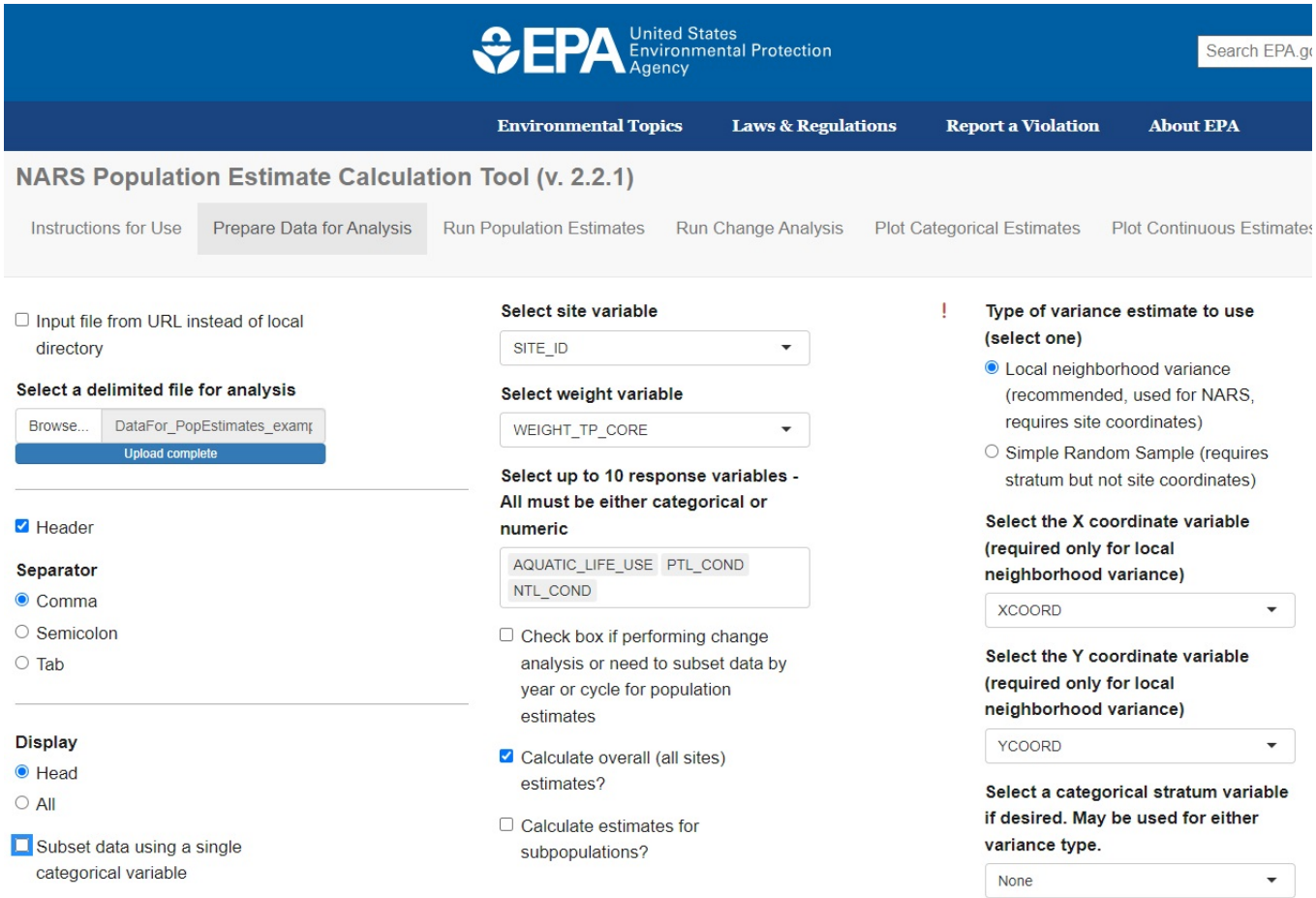


Figure 15: https://rconnect-public.epa.gov/NARS_Population_Estimates_Tool/

Analyses available in app

1. Categorical variable analysis (Proportions and Totals)
2. Continuous variable analysis (CDFs, Percentiles)
3. Change analysis (Categorical and Continuous variables)

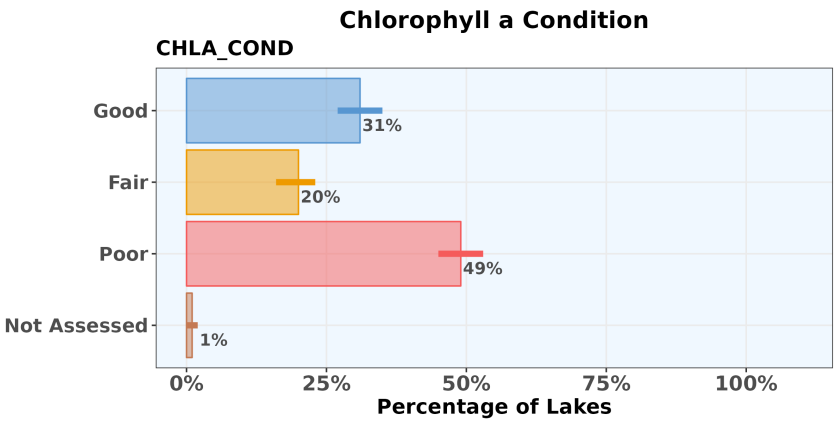


Figure 16: Proportion Estimates.

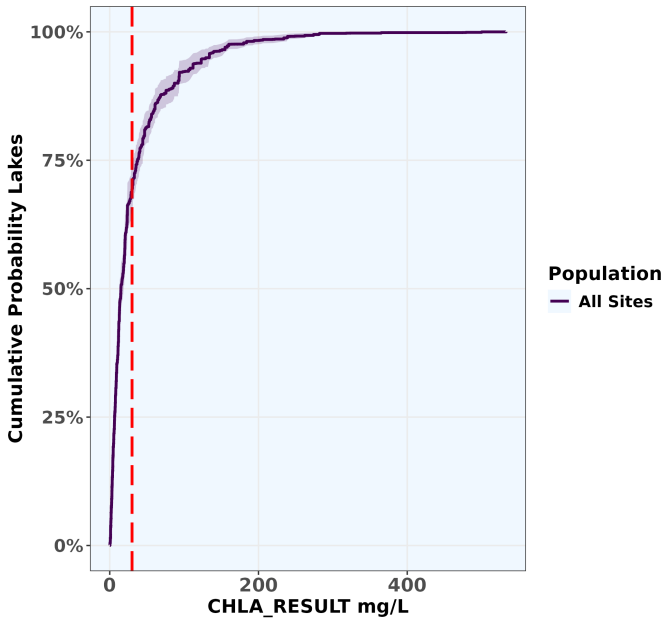


Figure 17: Cumulative Distribution Function.

Analyses available using R code

4. Risk analysis
   - Relative risk, attributable risk, risk difference
5. Trend analysis
   - Categorical and continuous variables
   - e.g., indicator trend over 3+ survey cycles

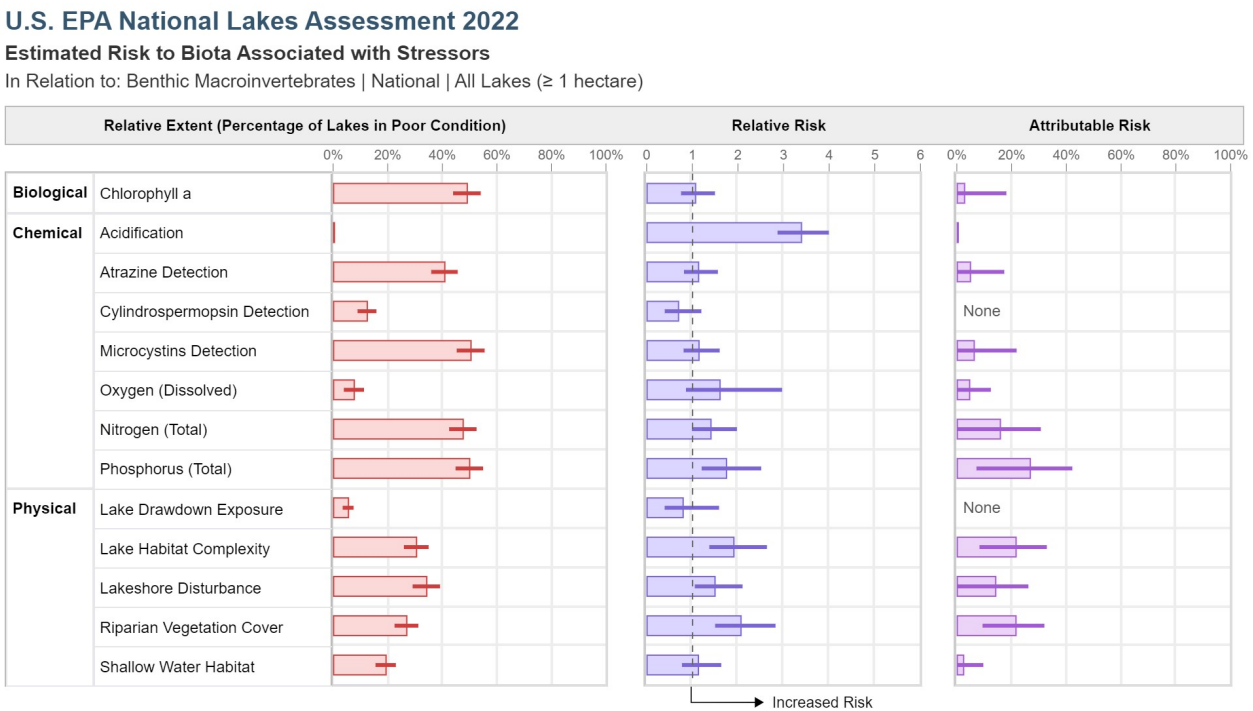


Figure 18: Risk Analysis

Requirements for Analysis

1. All data must be in a single file
   - One row per site and year/cycle. No Visit 2's.
   - For Change Analysis, if legacy sites are included, site variable needs to be the same name.
2. Weight Column
3. Response Variable Column
   - Categorical or Continuous variable(s)
4. Subpopulation(s) Column (optional)
   - Variable for different subsets of the population which estimates will be calculated for.
5. Variance Estimator
   - Local neighborhood variance is preferable and decreases variance estimates
6. X/Y Coordinates Columns
   - Should be an area-preserving projection CRS instead of geographic (XCOORD/YCOORD).

Plotting Options

- Data can come from within the app or can upload a file of analysis results
- For categorical data, plot all categories for single subpopulation **OR** all subpopulations for single category
- For continuous data, only CDFs can be plotted
- For continuous data, confidence bounds and reference lines can be added
- Plots can be saved as .png files

Section 6

Communication

ATTAINS-Surveys Module

- Information about the population surveyed
  - Resource Type (Streams/Rivers, lakes, wetlands, estuaries)
  - Population (statewide, basins, individual waterbodies)
  - Extent and number of sites sampled
- Condition estimates for designated uses, conditions and stressors
  - Margin of error
  - Confidence level

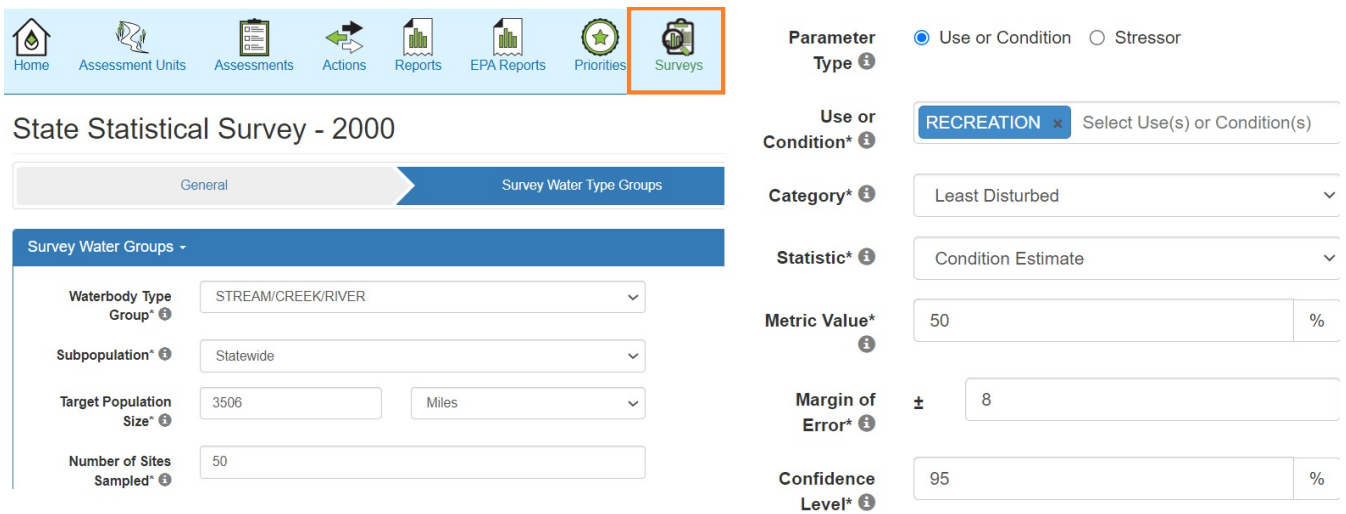


Figure 19: ATTAINS Surveys Module

ARMADA

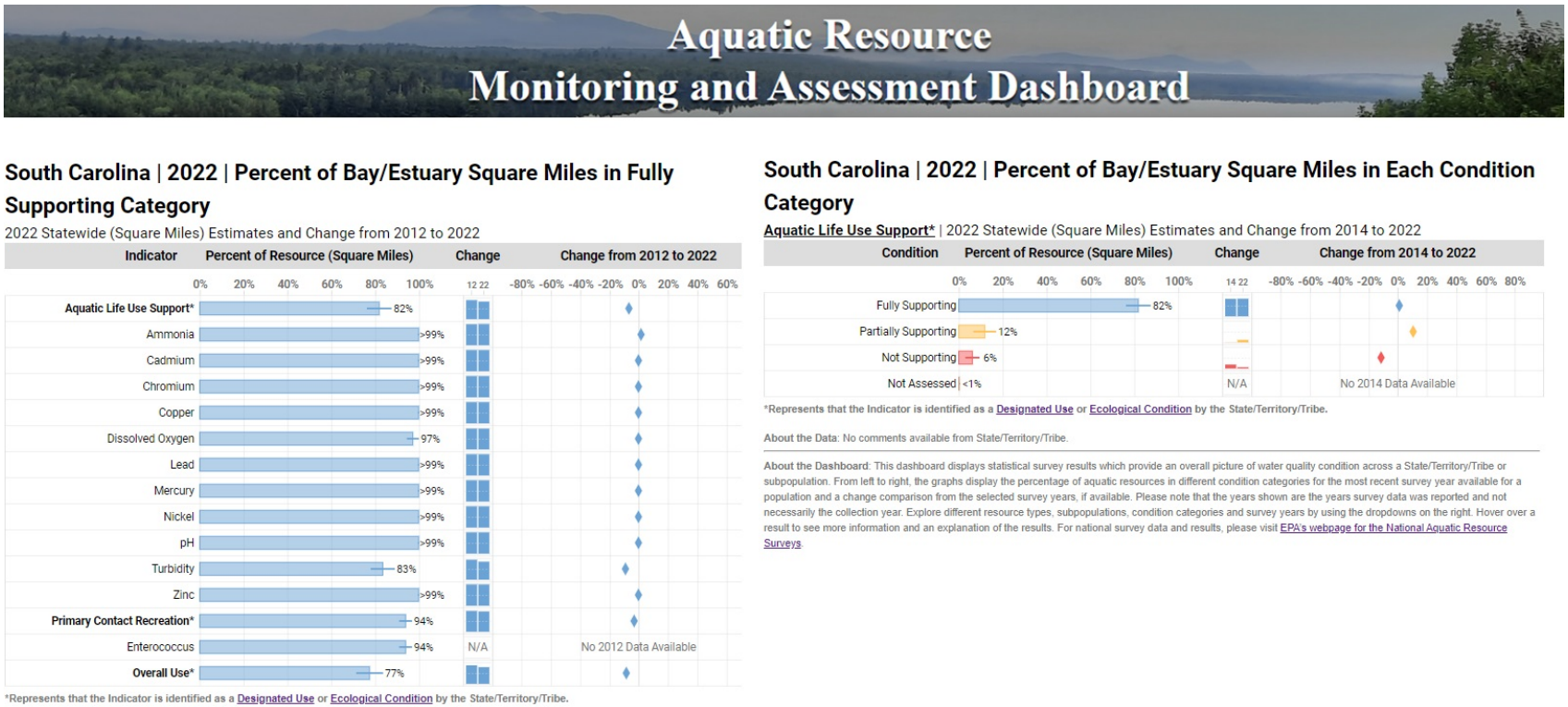


Figure 20: https://rconnect-public.epa.gov/armada/

Section 7

Resources

Applications and Websites

- *spsurvey Website*
- *Survey Design Tool Link*
- *Survey Design Tool GitHub*
- *Population Estimate Shiny Tool Link*
- *Population Estimate Shiny Tool GitHub*
- *ARMADA*

Additional Resources

1. *spsurvey: Spatial Sampling Design and Analysis in R.*
2. *Spatially Balanced Sampling of Natural Resources*
3. *A GRTS User's Manual for the SDrawNPS Package*
4. *NARS Website*
5. Olsen, A., Kincaid, T., & Payton, Q. (2012). Spatially balanced survey designs for natural resources.