Probability Sampling for Spatial Domains Workshop; August 2012
Dr. Don L. Stevens, Jr.

Workshop description:

Statistical science has multiple paradigms that provide prescriptions for collecting and analyzing data. Most introductory texts to statistics describe the frequentist and Bayesian paradigms. These are both model-based in that they begin with an assumed distributional form and investigate how inferences should be drawn to unspecified parameters in that distributional form. A third statistical paradigm that is useful in environmental studies is survey sampling. In contrast to model-based paradigms, survey sampling explicitly creates the probability distribution that is used for inference. This workshop will explore the application of the survey sampling paradigm to spatial domains, especially environmental studies, and as time permits, contrast that with model-based paradigms.

Prof. Stevens has worked on large-scale natural resource surveys for over 30 years, primarily with the US EPA to develop the statistical theory and practice of designing and analyzing surveys of lakes, streams, wetlands, and estuaries. Much of the lecture material will be motivated by that aquatic context, but will also apply to other types of resources, e.g., forests. The workshop will begin with a discussion of the survey sampling inference paradigm, provide the minimal required sampling theory, and cover topics such as using ancillary information in design & analysis; structuring a sample through space & time; plot design and spatial support of metrics; components of variance and allocation of fixed resources; panel designs; re-weighting to account for missing data; imputation of missing data; and variance estimation for complex survey designs; Master Sample concept & practice; merging legacy sites with a probability sample.

Morning Session 1, Day 1: Context for environmental sampling

> This session explores some basic sampling concepts in the context of sampling a spatial environmental. The notions of target population, sampling objectives, sampling strategy, sampling design, response design, and analysis strategy are introduced and discussed.

Break

Morning Session 2, Day 1: Relevant sampling theory

> This session introduces the general theoretical structure of inference from a probability sample through the Horvitz-Thompson theorem. The HT theorem is used to investigate the relationship between design and precision. That relationship is exploited to derive efficient design strategies.

Lunch Break

Afternoon Session 1, Day 1: Design and measurement of sampling units

> This session investigates the relationship between the high level sampling objective and the data that is actually collected in the field. The concepts of metric, indicators and indices are explored. The importance of response design is stressed, and the impact of plot shape, size, and orientation is explored. Also, the potential bias resulting from boundary or edge effect is evaluated, and some strategies for addressing edge-effect bais are discussed.

Break

Afternoon Session 2, Day 1: Spatial control: maps, projections, grids, tessellations, and spatial stratification

> Almost all spatial populations have spatial pattern, and this has long been recognized in strategies design to sample spatial populations. In this session, we explore methods for using spatial pattern to achieve more efficient samples.

Morning Session 1, Day 2: Using ancillary information in design and analysis

> We almost always know something about the spatial population we want to sample, either from previous samples, or from landscape level information available from GIS or satellite sources. That knowledge can be used to obtain more precise estimates, either at the design stage or at the analysis stage. In this session, we explore some ways using prior knowledge to design a survey or to use covariate information in the analysis.

Break

Morning Session 2, Day 2: Allocation of resources through space and time, Master Sample concept & practice

> The spatial control session looked at some ways of spreading a sample through space. In this session, we look at techniques for designing a sample to both estimate current status and trend or change in that status. The notion of rotating panel studies is introduced, and techniques for creating spatially interpenetrating panels are provided.
>
> A Master Sample is simply a very large (in a spatial context, a spatially dense) sample that can be used by various organizations. The concept is that a particular study can select a subset of the Master Sample tailored to the needs of that study. The structure can be used to coordinate studies of the same resource by different organizations, and, over time, can be used to compile a "site history" that can be used as ancillary information to inform future sampling.

Lunch Break

Afternoon sessions for Day 2 will be devoted to hands-on exercises using R software. We will illustrate drawing a probability sample from populations embedded in a spatial domain using the R package *spsurvey*. We will illustrate sample selection using a finite population, stream network, and a two-dimensional continuous domain. I will be running the software on my laptop; if you want to duplicate on your own laptop, I recommend that prior to the workshop, you download and install R software and several R packages. (I will also have these available on USB drives at the workshop). R and the packages can be downloaded from the R Project website (www.**r-project**.org/) . You will need to select an R mirror website for the download from the list available on the R Project website.

You will need the following packages: *spsurvey, sp, deldir, rgeos, stringr, and plyr*. You might also want the package *shapefiles*.

R has a built-in source editor, but it's not very convenient. I will be using TINN-R as an Integrated Development Environment (IDE). Others are available. TINN-R is a free download from http://sourceforge.net/projects/tinn-r/. I suggest version 1.19.4.7. Later versions are available, but have had some stability problems. I'll be using 1.19.4.7.