

Design of Soil Monitoring Networks: Some US Perspectives

F. Jay Breidt
Department of Statistics
Colorado State University
Fort Collins, Colorado, USA

European Seminar on
Soil Protection and Sustainable Development
Soria, Spain
May 16, 2002

Some Questions to Consider

- Uses for the data sets?
- Design of the sample?
 - what are the implications of spatial/temporal analysis for design?
- Definition of the sampling units?
 - at what scales should data be collected?
- Combination of measurements from different sources (lab, imagery, etc.)?

Two US Surveys

- National Resources Inventory
 - stratified, two-stage sample of non-federal, rural US lands
 - assess conditions and trends at five-year intervals
 - illustrates monitoring issues at large temporal/spatial scales
- Western Iowa Soil Survey Update
 - multi-phase sample of two Iowa counties
 - improve distributional information in soil reports
 - illustrates monitoring issues at small spatial scales



Uses for the Data Sets

- Tabular estimates of status and change
 - means, proportions, totals, regression coefficients, . . .
- Model development (spatial, temporal?)
- Map construction and small area estimation
- Supervised classification for remotely-sensed images
 - “ground truth”
- Input to biogeophysical process models
 - erosion, carbon storage, etc.
- Frame for further sampling
 - sample points are subsampled for intensive study
- Many other unanticipated uses!
 - choose simple, flexible, extensible design

Design Considerations: Probability Samples

- Define a population of interest
- Construct a frame that identifies population elements
 - GIS coverage
- Determine a design in which all frame elements have a known, positive probability of selection
 - systematic (grid-based) designs
 - random within blocks (geographic stratification)
- Draw a sample and take observations for selected sites
- **Advantages:** unbiased estimators and good measures of their precision
- **Not true for non-probability samples:** (judgement, purposive, or convenience samples)

Design Considerations: Types of Errors in Surveys

- Sampling error: probability sample \neq population
 - but is representative of the population
- Non-sampling error: all other errors
 - coverage error: frame \neq population
 - response error: sampled \neq accessible
 - measurement error: observations \neq true values
 - processing error: data transfer, programming bugs, etc.
 - “timeliness error”: late analysis may not be useful
- Increased sample size, increased use of auxiliary information, more complex design and analyses:
 - decreased sampling error
 - increased non-sampling error

Design Considerations: Auxiliary Information

- Possible sources of auxiliary information
 - sampling frame: latitude, longitude
 - digital elevation models: elevation, aspect, slope
 - administrative records: land ownership, program participation
 - satellite imagery: land cover classifications, vegetation indices
 - aerial photography: land cover and use, plant species
 - geographic information systems: endless variety of indicators

Combining Information: Model-Assisted Estimation

- Consider two nested grid samples:
 - phase 1: fine grid of n_1 sites, coarse information \mathbf{x}
 - (could be wall-to-wall coverage)
 - phase 2: coarse grid of $n_2 < n_1$ sites, fine observation y
- Use \mathbf{x} to predict observation of interest, y
- Model-assisted estimator of $\mu = y$ -average:

$$\hat{\mu} = \sum_{i=1}^{n_1} \frac{(\text{prediction})_i}{n_1} + \sum_{i=1}^{n_2} \frac{(\text{observation})_i - (\text{prediction})_i}{n_2}$$

- nearly unbiased (high accuracy) even if predictions are poor
- small variance (high precision) if predictions are good

Combining Information: Weighting

- Rewrite the model-assisted estimator:

$$\hat{\mu} = \sum_{i=1}^{n_2} (\text{weight})_i (\text{observation})_i$$

- Simple weights ($x \equiv 0$): $(\text{weight})_i = 1/n_2$
- Parametric regression weights (includes ANOVA):

$$(\text{weight})_i = \frac{1}{n_2} + (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' (\mathbf{x}_i \mathbf{x}_i')^{-1} \mathbf{x}_i$$

- Non-parametric weights (kernels, other smooths)

Combining Information: Imputation

- Estimator:

$$\sum_{i=1}^{n_2} (\text{weight})_i (\text{observation})_i + \sum_{i=n_2+1}^{n_1} (\text{weight})_i (\text{imputation})_i$$

- randomly “impute” observations from probability distribution f
- $(\text{imputation})_i \sim f(y; \mathbf{x}_i)$
- imputed points may replace missing values
- imputed points may be used for convenience

Design Considerations: Spatial Structure

- “First Law of Geography”: nearby units tend to be similar
 - positively correlated implies increased sampling error
 - resources wasted by essentially measuring the same thing twice
- Implication: sample should be well-dispersed spatially
 - systematic sampling, stratified within small geographic strata, Markov chain designs
- Adaptive sampling designs:
 - possible to use spatial analysis to augment/reduce existing network
 - usually not worthwhile for multi-purpose surveys
 - may be useful for some special-purpose surveys (e.g., rare, highly clustered populations)

Design Considerations: Small Area Estimation

- Rule of 100: to estimate a proportion to within $\pm 10\%$ with 95% confidence, need 100 observations in the area (regardless of its size!)

$$\pm 1.96 \sqrt{\frac{(1/2)(1/2)}{100}} \simeq \pm 0.10 = 10\%$$

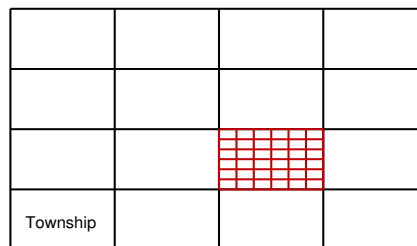
- Often interested in smaller areas than those supported by design
- Requires good auxiliary information and good model
 - hard problem!
- Avoid as much as possible via fine stratification
 - guarantee some observations in most areas
 - augment as necessary with additional samples for known subpopulations of interest

Spatial Design of the National Resources Inventory

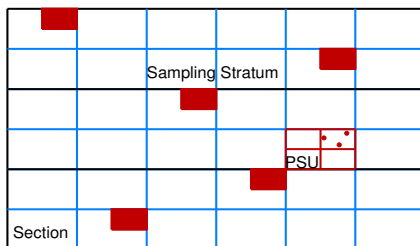
- 300,000 primary sampling units, three points each



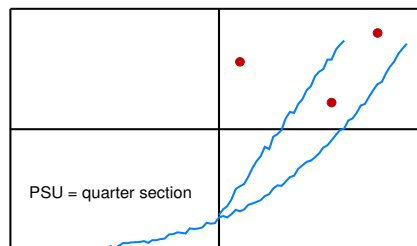
COUNTY (24 x 24 miles)



TOWNSHIP (6 x 6 miles)



SECTION (one square mile)



Choice of Sampling Unit in the NRI

- Ultimate sampling unit is a 0-dimensional point
 - collect ownership, land cover and use, soil properties, habitat composition, overland water flow, irrigation, soil erosion, wetlands, conservation practices, and salinity
 - convenient for data collection
 - poor for capturing 1-dimensional features like streams, roads, windbreaks, buffer strips
- Primary sampling unit is 2-dimensional square (0.25 sq mi)
 - too labor-intensive to collect all information at PSU level
 - collect additional surface area data on urban land, farmsteads, streams, and water bodies
 - used to detect changes too rare to be captured in the point observations
 - good for capturing 1-dimensional features

Temporal Design of the NRI

- Traditional NRI design: 300,000 PSUs every five years
 - 1982, 1987, 1992, 1997
- Problems:
 - data are not timely
 - data collectors require re-hiring, re-training
 - workload is very heavy in the inventory years
- Redesigned NRI: supplemented panel design
 - core sample observed in every year (hang on to this core if funding is cut!)
 - additional units observed on rotating basis
 - units may rotate out and back in to sample
 - gives data collectors a more continuous workload
 - easier to keep them trained and ready

Temporal Design of the NRI: Supplemented Panel

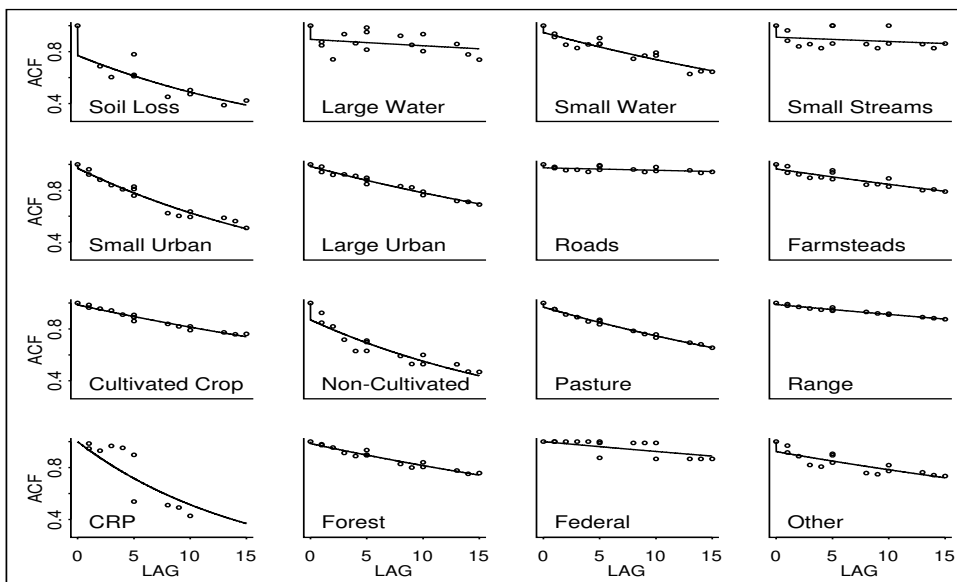
- Sample of n units each year:

	Year: 1	2	3	...	T
θn matched units	×	×	×	...	×
	×				
$(1 - \theta)n$ unmatched units		×			
			×	...	
					×

- Choices for θ :
 - $\theta = 1$ is a **pure panel**
 - $\theta = 0$ is a **new sample** at each time point
 - $\theta = 1/2$ is suggested for continuous NRI

Optimal Choice of θ

- Best fraction revisited every year depends on question:
 - for estimation of **change**, choose $\theta = 1$
 - for estimation of **level**, choose $0 \leq \theta \leq 1$ (temporal dependence structure determines best choice)



- for **multiple objectives**, no general conclusions, but $\theta \geq 1/2$ might be good compromise

Combination of Information in the NRI

- Data from different kinds of units can be combined into single database
- NRI uses imputation:
 - fills in missing data
 - some PSU-level data not captured in the point data
 - create **pseudo-points** that reflect **PSU** data
- NRI uses regression weighting:
 - incorporates population-level data (known acreages from GIS coverages)
 - creates one data set with information consistent across time
 - (e.g., 1997 dataset can be used to reproduce 1982 estimates)
- Rather complicated estimation procedures

Western Iowa Soil Survey Update: Background

- National Cooperative Soil Survey:
 - US Dept of Agriculture + state agency
 - county-level soil maps and soil series descriptions
 - intensive field work; rarely updated
 - used by farmers, contractors, scientists
- Map traditionally constructed from:
 - conceptual model of soil formation
 - subjective evaluation of topography, vegetation, parent material, etc.
 - occasional observation of soil properties at purposively-located sites
- Poor distributional information

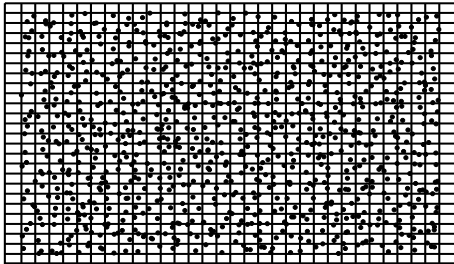
Western Iowa Soil Survey Update, Continued

- Heavy use of technological innovations
- Geographic Information System
 - detailed stratification using digitized soil maps
- Global Positioning System
 - precise navigation to randomly-selected points
- Personal Digital Assistants
 - simple, consistent field data collection, with edits

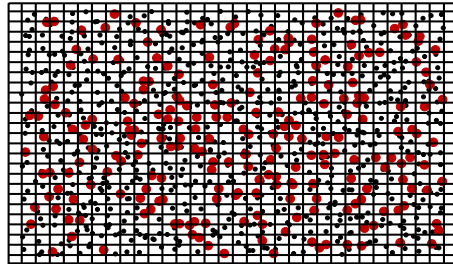
Spatial Design of Western Iowa Soil Survey

- Multi-phase design

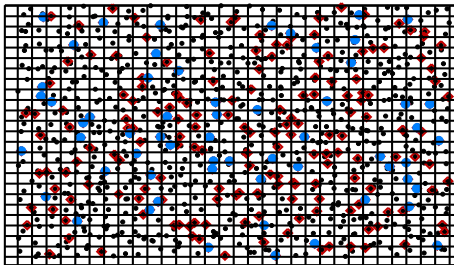
Phase One: Surface Horizons



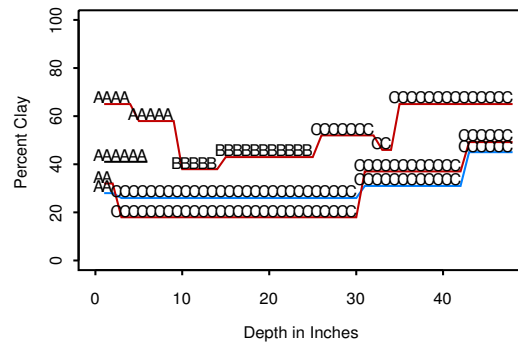
Phase Two: Full Profiles



Phase Three: Lab Data



Point Profiles

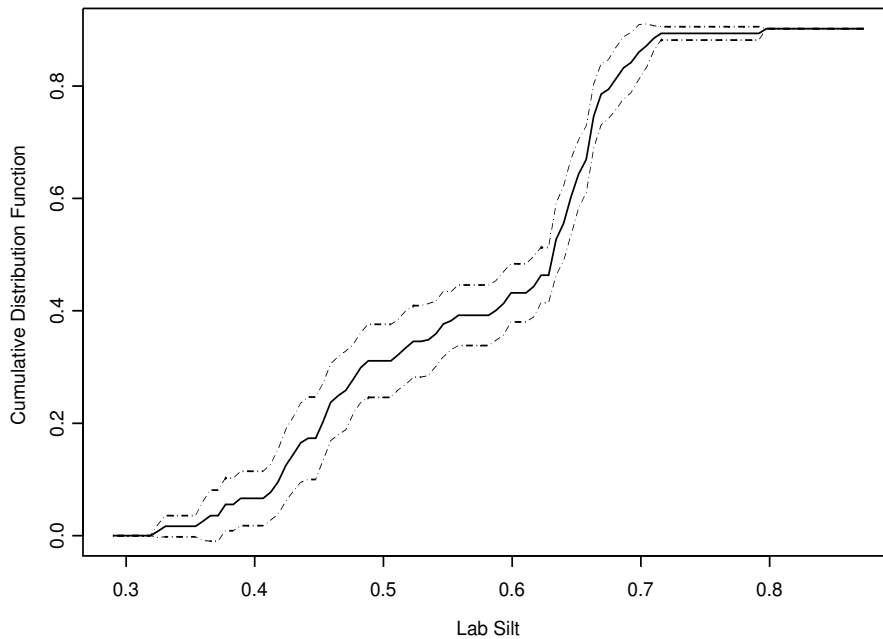


Choice of Sampling Unit in WISSU

Data Element	Phase 1	Phase 2	Phase 3
Map unit symbol	X	X	X
Landscape data (slope, shape, etc.)	X	X	X
Surface horizon data (depth, color, texture, effervescence, roots, pores, etc.)	X	X	X
Beyond surface horizon (depth, color, texture, effervescence, roots, pores, etc.)		X	X
Lab analysis (particle size, organic carbon, pH, texture, etc.)			X

Combination of Information in WISSU

- Combination of information across different phases
- Population \supset sample \supset subsample
- Combine information across phases using weighting



Summary

- **Data Uses:** many, varied
- **Sample Design:** keep it simple, flexible
 - select a probability sample via a unified design
 - use geographic stratification to avoid small area problems
 - consider a supplemented panel design
 - avoid adaptive designs
- **Sampling Units:**
 - collect data at the most appropriate scale
 - unit needs to be precisely defined (e.g., no moving site for convenience)
- **Combination of Measurements:**
 - many good sources of auxiliary information
 - exploit via weighting and/or imputation