Survey errors in land-based surveys

Sarah Nusser

Department of Statistics

Center for Survey Statistics and Methodology

Iowa State University

Partially supported by USDA NRCS CESU agreement 68-7482-11-534 and USDA NASS cooperative agreement 58-3AEU-1-0012

Land-based surveys

- Environmental and agricultural surveys with area samples
- Why this topic?
  - Underdeveloped understanding and assessment of survey quality for some surveys
  - Geospatial foundation plays a larger role
  - TSE concepts are similar, but can play out differently

Topics

- The nature of land-based populations and surveys based on area samples
- Some more prominent features that affect survey error
- Observations on TSE components in land-based surveys

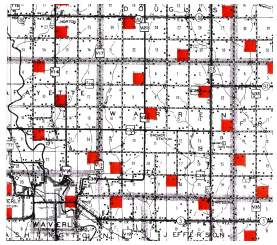
- Due to time limitations, will focus mainly on natural resource and environmental surveys

Natural resource and environmental surveys

- Measure extent and quality of resources
  - Lakes and streams, productive land (forest, grazing, crop), soil, wetlands, parks
  - Often involve multiple resources or resource types (e.g., forest types, crop types, or combinations such as all non-federal land)
- Sampled as land areas, points, features
- Data collected via field visits &/or remote sensing
- Often specifically designed for monitoring change

National Resources Inventory (NRI)

- Monitor status and trends of natural resource conditions on non-federal land
  - Land cover/use dynamics, soil erosion, wetland loss/ gain, prime farmland loss to development, grazingland health, etc.
  - Model response data to predict ag run-off in streams



- Stratified 2-stage area sample (70K)
- Most data collection via remote sensing
  - High res aerial photos, local administrative data
  - Smaller surveys with field observations on points

Differences from household (HH) and establishment surveys

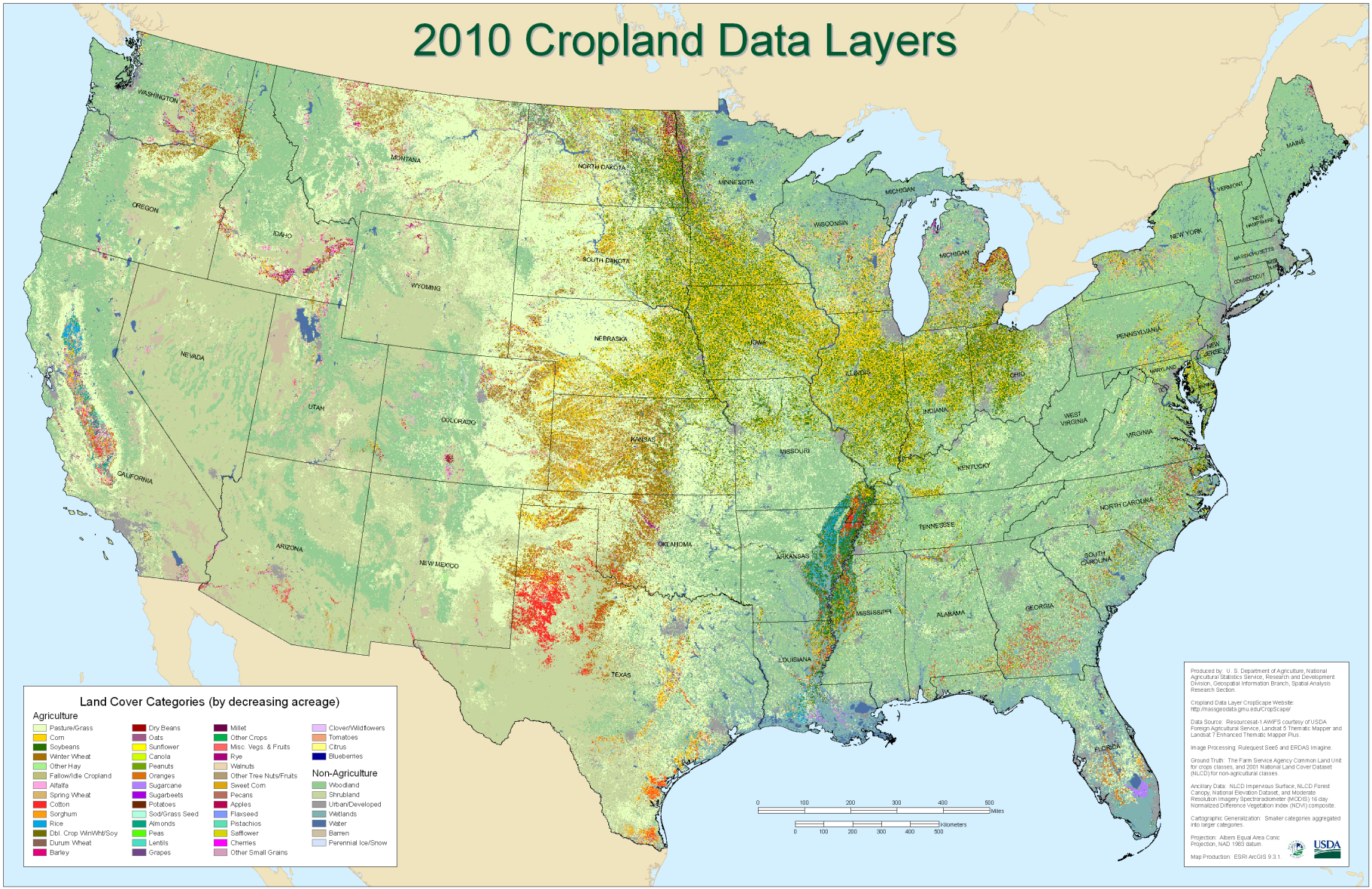
Many differences derive from geographic nature of populations and analysis domains

1. Populations have high degree of heterogeneity and uneven spatial distributions
2. Land is continuous, but land cover transitions are messy
   - Classification systems are difficult to define
   - Boundaries are difficult to discern, change over time
3. Auxiliary and collected data often in geospatial formats

1. Pop/spatial heterogeneity

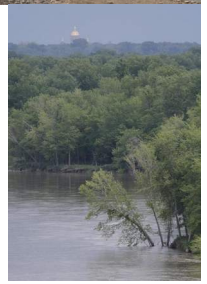
- Numerous and varied land and farm types
  - Multi-resource surveys involve lots of land types
    - Crop, pasture, wetland, rangeland, forest, developed, water types, barren and other rural land such as gravel pits, mines, etc.
  - Ag surveys cover many farm types
  - Complex survey instruments
- Strongly uneven spatial distribution of land or farm types
  - Few land or farm types are ubiquitous or evenly distributed nationally
  - Variation in spatial distribution across states (or equiv.)

2010 Cropland Data Layers

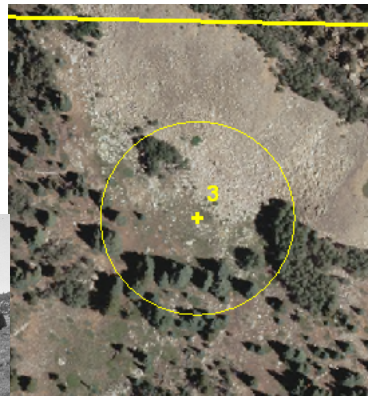


2. Defining land categories

- Requires a practical definition that separates land categories
- What is an agricultural field?
  - Only the productive land? The associated conservation practices? (where do these end?)

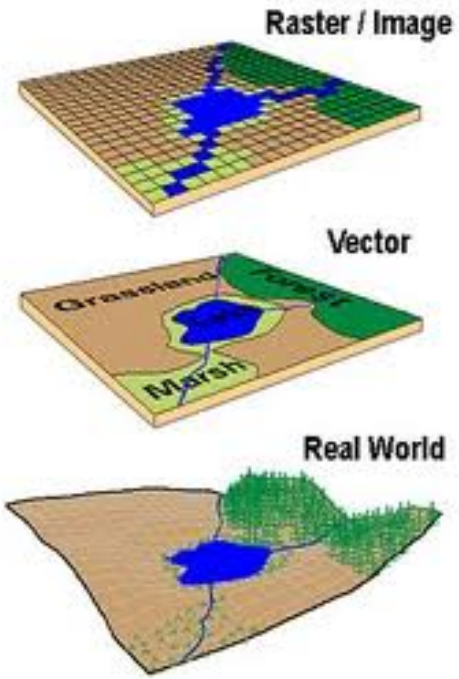


- What is the area of a lake? A stream?
  - Water levels change within and across years
- Where is the border between rangeland and forest land in Western US?



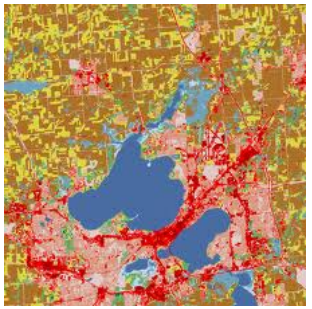
3. Geospatial data formats

- The form of Yi for geospatial data
  - Index (i) = coordinate
    - Typically 2D (latitude, longitude)
    - Increasingly 3D (with elevation)
  - Response or auxiliary data (Y) = attribut
  - Error arises in index and response
- Two main formats
  - "Picture" or raster = array of pixels
  - "Objects" or vector = points, lines, polygons

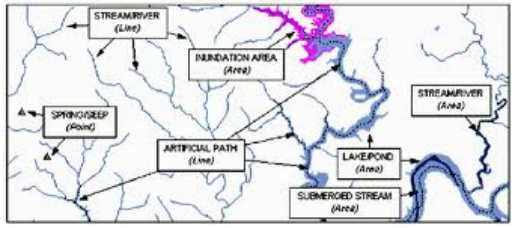


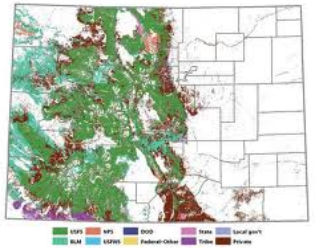
Geospatial data examples

- Raster: pixel data derived from imagery signals or classifying imagery into land cover categories
  - Cropland Data Layer (CDL)
  - National Land Cover Data (NLCD)

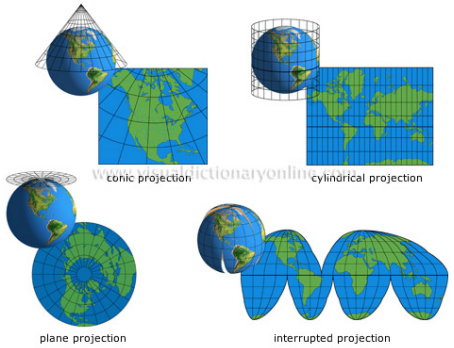


- Vector: feature boundaries denoted by photointerpretation, field records, cadastral administrative data
  - National Hydrography Data (NHD)
  - Agency land ownership databases
  - Delineated features from data collection

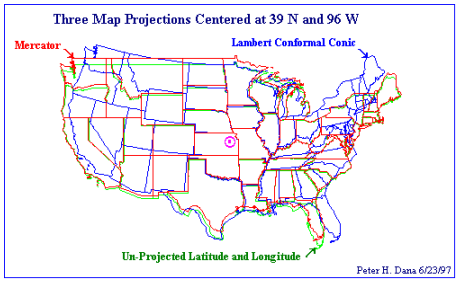




Errors in geospatial da



- Location error
  - Earth surface projected to one of many 2D coordinate systems
  - Resolution of coordinate system matters
  - Overlays of data layers can have considerable linkage error



- Attribute error
  - Classification error from imagery (raster data)
  - Collected features (vector data) have boundary and land cover (attribute, photointerp) error

Geospatial elements in longitudinal surveys

- Temporal monitoring of land more feasible
  - Unlike households, ground doesn't move, but …
- Borders of population, domains and sometimes reporting units are dynamic
  - Federal-private land swaps, farm dynamics
  - Anticipate possible future boundaries in design
- Location error in revisiting sample points leads to observation of a different location
  - Ripples through to measurement (attribute)

Total survey error framework

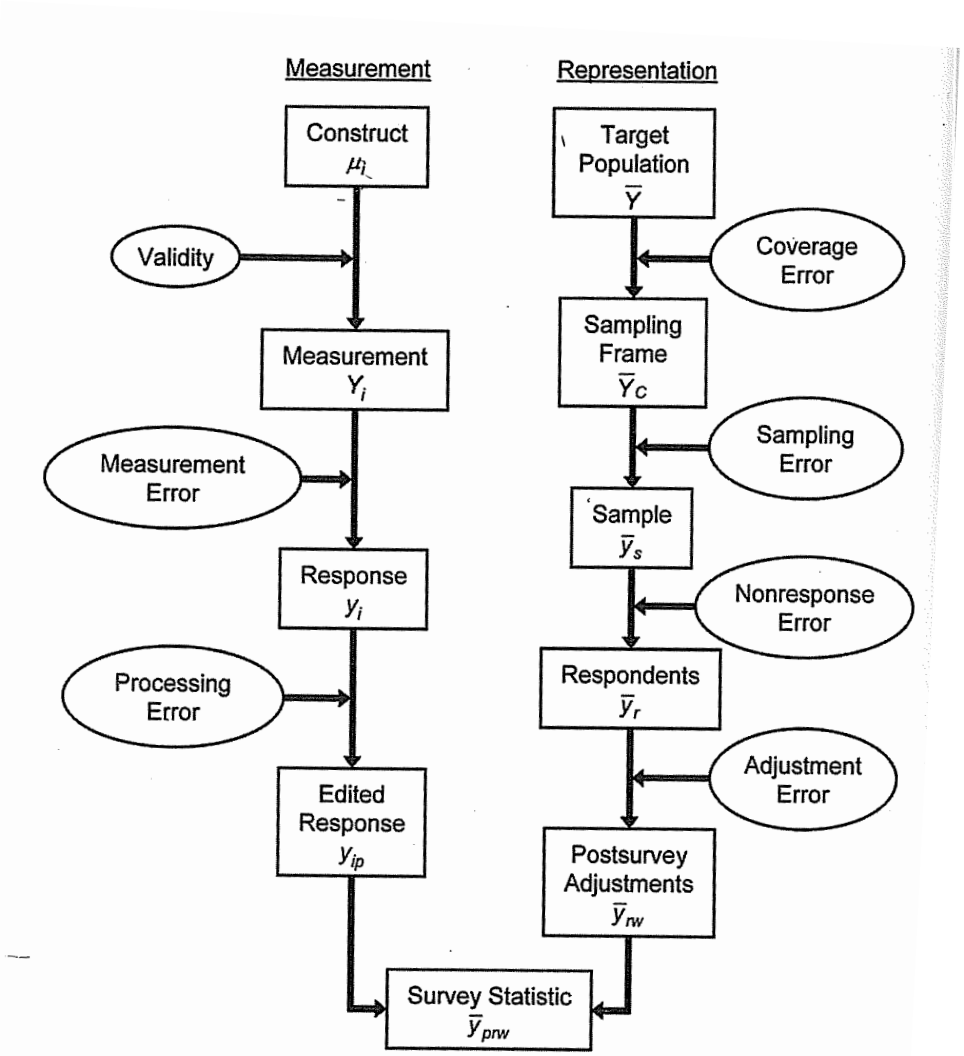


Figure 2.5 Survey lifecycle from a quality perspective.

Administrative and other data linkage, imputation, disclosure limitation, etc.

Population representation: coverage and sampling errrors

- Fairly similar to other types of surveys
  - Area samples tend to have low coverage error
  - Complex sample designs used to address heterogeneity while controlling sampling error
- Problems with determining eligibility
  - Error in constructing frame (linkage error, attribute error, lack of information)
  - Difficulty applying definitions in data collection

Population representation: nonresponse error

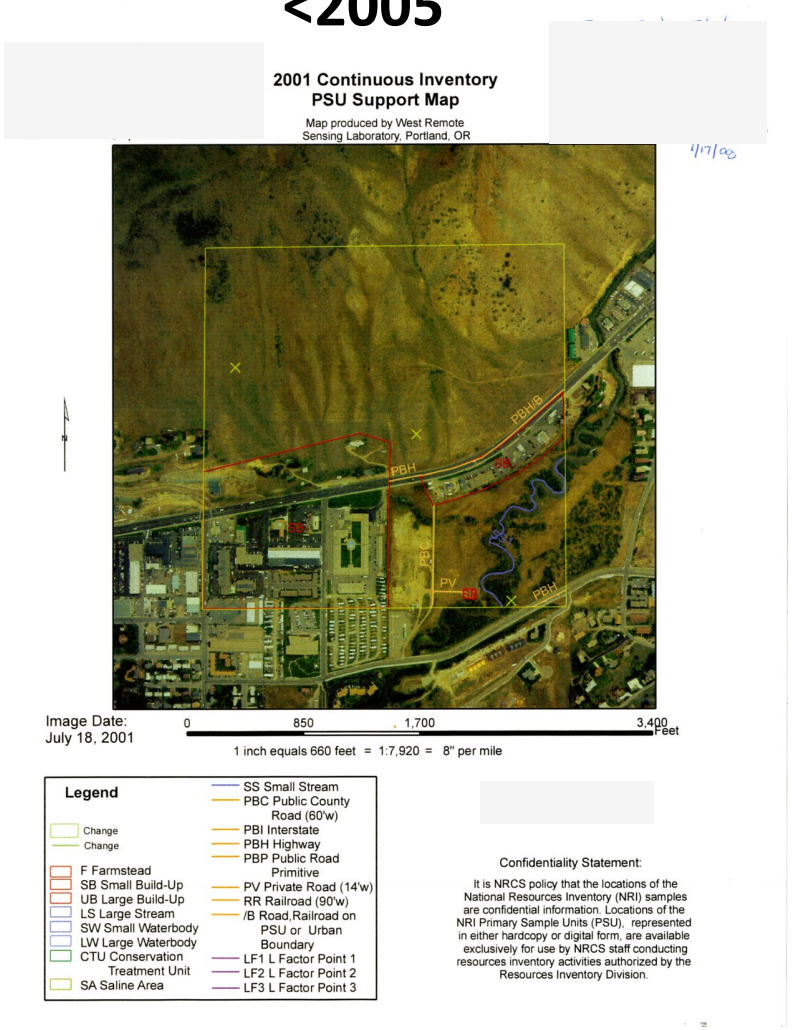
- Remote sensing has low nonresponse rates
  - No-fly zones in US, cloud cover, other
- Field visits require access and permission
  - Lack of permission, refusal of land manager to provide information – akin to HH nonresponse
  - Unit is physically or politically inaccessible
    - Range survey: protected lands, steep slopes
    - Stream survey: non-navigable waters
  - Can be strongly related to response variables

Population representation: location error in longitudinal surveys

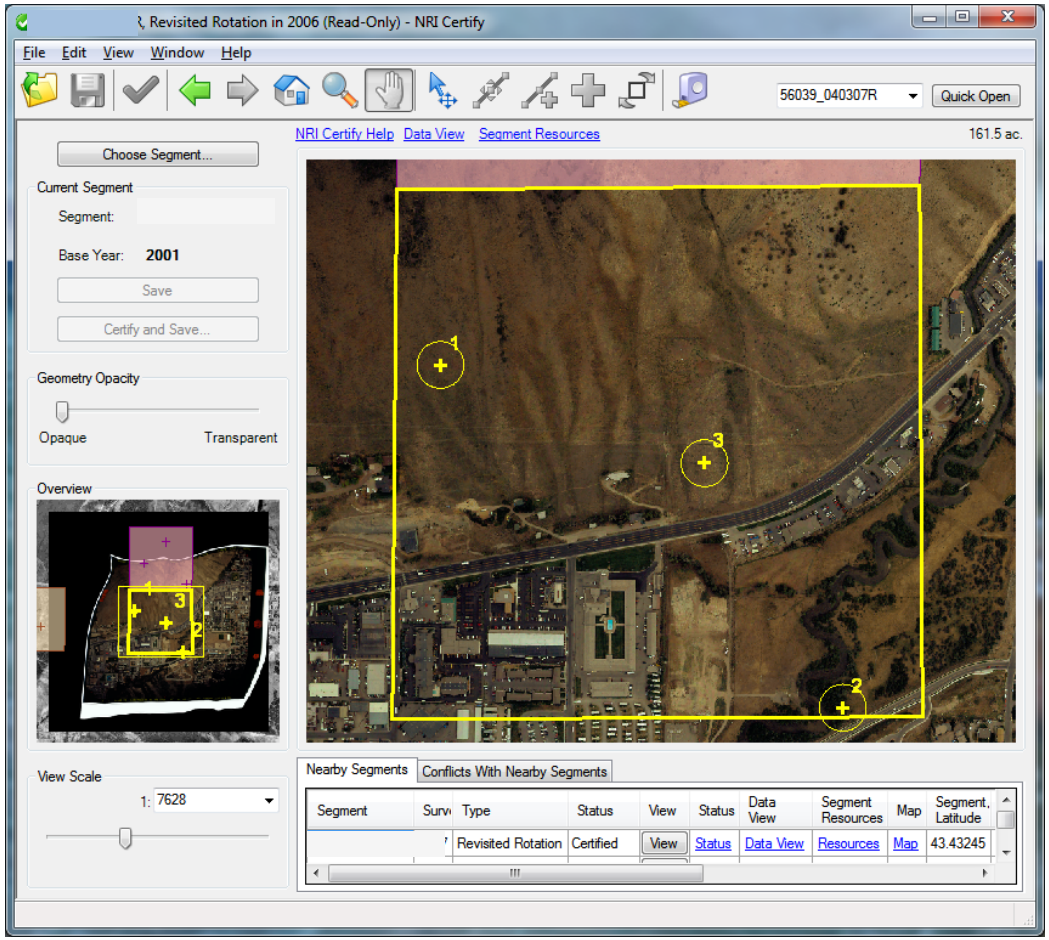
- Finding appropriate unit on ground
  - NRI data collectors to return to the location on the ground where data were previously collected
  - Return to same location on image (not same coordinate) when photos that have slightly different coordinate system each year
  - GPS has its own error and without a field marker, difficult to identify location of revisited point
- Ripples through to measurement

Location Certification

<2005

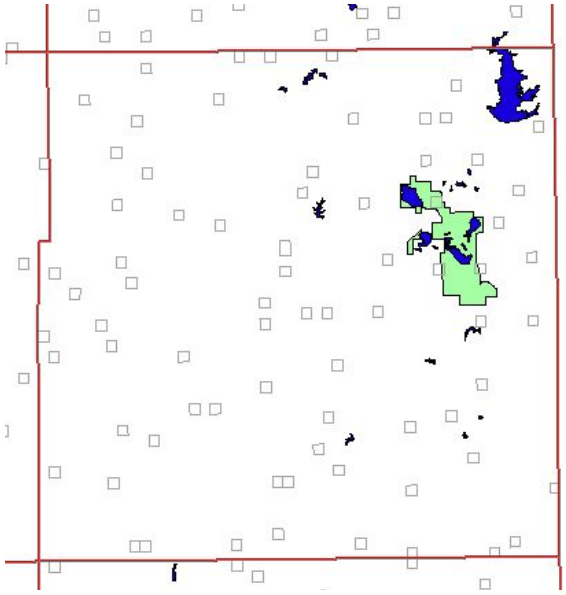


≥2005



Population representation: adjustment error

- Geospatial data can be used to set population totals for area weights
  - Area of states, counties in US
- NRI sample is of entire US area, but inference population is non-federal land
- NRI develops layers to quantify ineligible Federal and large water/river areas within the US



Response measurement: validity

- Definitions may not match target concept
- Natural resource "health" as a concept is underdeveloped in ecological sciences relative to human health
  - Rangeland scientists define health in relation to ecological site potential, an unobservable condition with "ideal" plant communities
  - Protocols developed tend to be subjective

Response measurement: measurement error

- Tends to be the largest component
  - Especially with remotely sensed area sample surveys
- Many difficulties of classifying land
  - Subjective aspects of protocols, incomplete information in imagery
  - Supporting auxiliary data layers have their own error
- Location error also causes measurement error
  - Delineated boundaries of collected data
  - Revisited sample units
  - Linked information captured for database

Response measurement: processing error

- Vector data need to be processed into a number for estimation
  - Most of this is straightforward
  - NRI turns housing unit marks into developed land area

- Error from further manipulation collected data
  - Simulation and model error
  - NRI uses complex erosion calculations, imputation models
  - NRI analysts use simulations of cropping outcomes for policy analysis

Summary

- Lots of similarities to HH and establishment surveys
  - Complex sample designs to address heterogeneity in population
  - Interview-based data collection has similar measurement error and nonresponse
- Some design options yield more favorable error characteristics
  - Area sampling: good coverage
  - Remote sensing: lower nonresponse rates
  - Measurement and processing error become dominant nonsampling errors

Summary

- Many issues revolve around making usable definitions and aligning them across information sources and the survey process
  - Target population
  - Frame
  - Eligibility
  - Response
  - Post-survey adjustment inputs
- Geographic location introduces new twists that propagate into measurement or processing error

Thank you!

Sarah Nusser

nusser@iastate.edu

Fuller (1999), Gallego and Delince (2010), Groves, et al (2004),
Karl, et al (2012), Marker and Stevens (2009),
Nusser and House (2009), Verma et al (2010)

Partially supported by USDA NRCS CESU agreement 68-7482-11-534 and USDA NASS cooperative agreement 58-3AEU-1-0012