E-forest platform: Approaches to data analysis

SC3 mid-term Meeting, Vienna, Austria, 12-13 Jnuary 2010

Sampling frames and Strata

Simplified:

- the sampling frame is the land area in which the sampling plots are randomly generated by a unique sampling scheme
- usually, the sampling frame covers the whole country
- if not, the sampling frame should be implemented in the e-forest platform (terra incognita, like Belgium, or Lichtenstein)
- for some countries, more then one sampling frame is needed, because sampling schemes differ by region
- the surface area of sampling frames must be known
- an other term for sampling frame is sampling stratum

Sample

- a sample is a finite set of randomly selected sampling units (points) in the sampling frame
- the sample must be a representative sample for the infinite population of points in the sampling frame
- this means: the selection (or sample inclusion) probability of each sampling unit (point) has to be known (but selection probabilities may be unequal), and each point in the sampling frame must have a chance to be selected as a sampling unit
- we are only interested in samples with field data collected (terrestrial grids)

Target variables and domains

- for SC3, we have only one target variable: basal area
- for each target variable, we need to declare the domain on which the target variable is defined and assessed in the NFI
- a domain indicator variable is needed for each target variable in the plot file (discussion of yesterday)
- the intended domain for the basal area is forest land
- a domain indicator may have three values:
  - 0: plot is in a domain on which the target variable is not defined and assessed
  - 1: plot is in the domain on which the target variable is defined and has been collected
  - -1: plot is in the domain on which the target variable is defined but it has not been collected (is missing)
- another target variable, we may implement now: forest area
- domain is usually the whole sample (only 1 and -1), the target variable would have to be delivered (1 or 0)

Plot configuration

- differs in size and form between inventories (countries, sampling strata within countries)
  - complications in variance estimation when plots from different inventories are not separated $\rightarrow$ stratification needed
  - occurence of species, for instance, depends on form and size of sampling units (occurence of any *event* depends on form and size of sampling units)
  - many variables can be standardised to local densities (hectare values) $\rightarrow$ unbiased point estimates

Shared Plots (Plots at the forest boundary)

- forest definition; at the moment, reference definition as good as possible
- definition of forest and non-forest plot is not clear for shared plots (plots at the forest edge)
  - plot centre decides within SC3
  - center inside forest: simplified procedure acceptable for SC3: expand the not-corrected density with factor 1/p (p being the proportion of the plot laying within forest)
  - centre outside forest: simply drop these plots (domain 0)

Plot locations

- all plots of a cluster should be shifted by the same vector (maintaining cluster geometry)
- correction of plots located sharply outside the country (after shifting and compared with the e-forest GIS layers) is not mandatory (at least not needed to fulfill SC3)

Unequal sample inclusion probabilities of points (plots)

- must be taken into account in the (design-based) data analysis, because the mean of plot values is not an unbiased estimate of the population parameter
- with regular grids, even under cluster sampling, selection probabilities are usually equal (within a stratum)
- known cases are:
  - Italy
  - France for some areas (could drop some plot to simplify our life)
  - Norway mountainous regions
  - ...
  - aggregation units (50x50) overlapping country (strata) boundaries

Design-based estimation

- we need the relative weights a_j for all plots from all countries (strata)
- is very simple for countries with equal selection probabilities: $a_j = 1$ for all plots in the sample
- others need to think... (most likely already know what to do)
- with y_j being the local densities per hectare of the target variable, the weights must be such, that

$$\hat{Y}_s = \frac{\sum_{j=1}^{n_s} \frac{a_j}{\bar{a}_s} y_j}{n_s}$$

  is a design-unbiased estimator of the true density Y_s in stratum s
- n_s is the number of plots and $\bar{a}_s$ is the mean (average) weight in s

Overlapping aggregation units

- simple case (for design-based estimation) would be to make separate estimates $\hat{Y}_s$ for the strata and combine them with strata weights; known relative strata sizes p_s within aggregation unit (50x50 km cell)

$$\hat{Y}_d = \sum_{s=1}^{S} p_s \hat{Y}_s$$

- problem: p_s may not be available or implemented in the system; ok for country boundaries? But strata boundaries within countries may not be available?
- Alternative:

$$\hat{Y}_d = \frac{\sum_{j=1}^{n_d} \frac{b_j}{\bar{b}_d} y_j}{n_d}$$

where the $b_j = (a_j/\bar{a}_s) \times (\lambda_s/n_s)$ are plot weights standardised with the represented area of the plots (λ_s is the size of stratum s).

Cluster sampling

- no problem for equal selection probabilities of clusters
- weights to compensate for unequal selection probabilities have to be given for the cluster (the clusters are selected with unequal probabilities, not the individual plots), so that

$$\hat{Y}_s = \frac{\sum_{j=1}^{n_c} \frac{b_j}{\bar{b}_s} g_j \bar{y}_j}{n_c}$$

  is an unbiased estimator for the population parameter Y_s. b_j are the cluster weights, g_j is the number of plots in the cluster, $\bar{y}_j$ is the cluster mean of the target variable, and the sum is over the number of clusters (n_c).
- design-based approximately unbiased estimates for aggregation units overlapping strata boundaries and with different cluster designs seem possible (to be checked)

Variance estimation

- without major problems for aggregation units completely within a single stratum
- $\rightarrow$ no problem when GIS layers of strata are available, formulae are worked out
- problems for aggregation units overlapping strata boundaries without GIS layer
  - plot configurations are different (plot sizes, mixture of angle count sampling and fixed radius plot sampling)
  - mixture of cluster and single plot sampling
- not needed for SC3
- some good ideas may come up

Aggregation techniques

- Design-based estimation procedures (limited to larger area), in SC3 also used as a tool for "predictions" within aggregation units (sound, but large errors)
- Model-dependent estimation
  - different basis: design is not important (sample is fixed, forest is random)
  - spatial auto-correlation models (variogram): local models needed (plot configuration are different), implies a lot of modelling work, may be (or may be not) better for relatively small aggregation units
- Model-based estimation (prediction) (kNN and derivates): needs high resolution auxiliary information (satellite images and other geo-data), big effort to adapt models regionally, high potential for error reduction and for small area estimation
- Interpolation (smoothing): immediately available tools for visualising plot values, quality of generated values and means are not known, smoothing over the whole land base