

eurac research

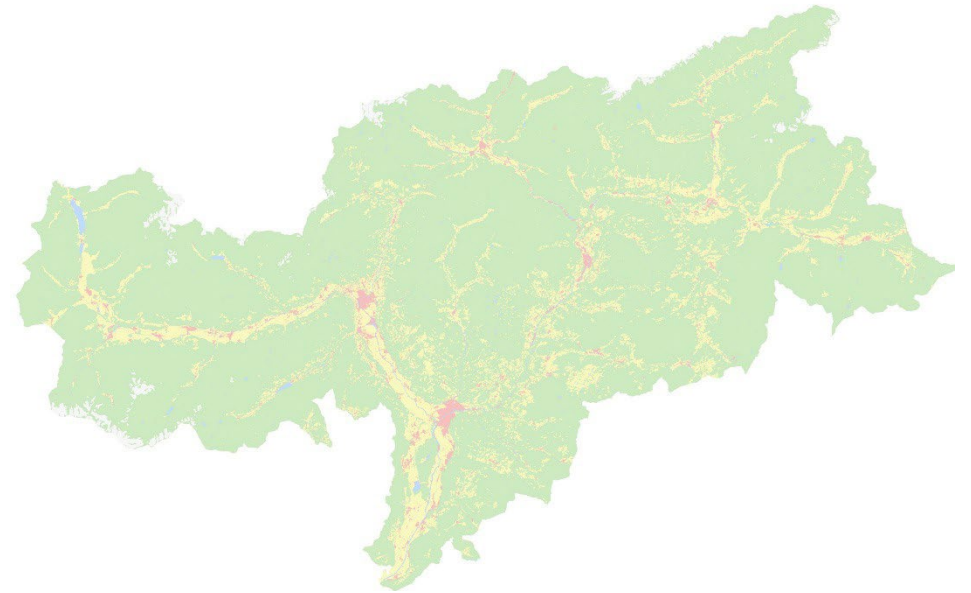
Sampling

Ruth Sonnenschein
Eurac Research



Importance of sampling

- Reference data for training and validation
- Since it is too expensive and difficult to obtain the reference land cover classification for the entire region of interest, statistical sampling becomes a critical component of accuracy assessment.
- A 'sample' is a subset or portion of the region mapped, and a 'sampling design' is a protocol for selecting those locations at which reference data will be collected.



Sample data

Training and validation samples should be:

- statistically independent (not clustered)
- abundant in all classes
- representative for the entire landscape

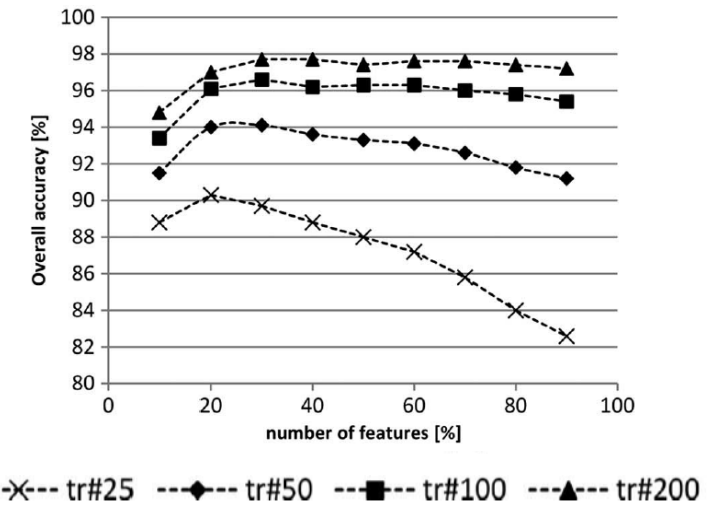
Classification accuracy depends on:

- the quantity of sample data
- the quality of sample data
- the sample selection method used

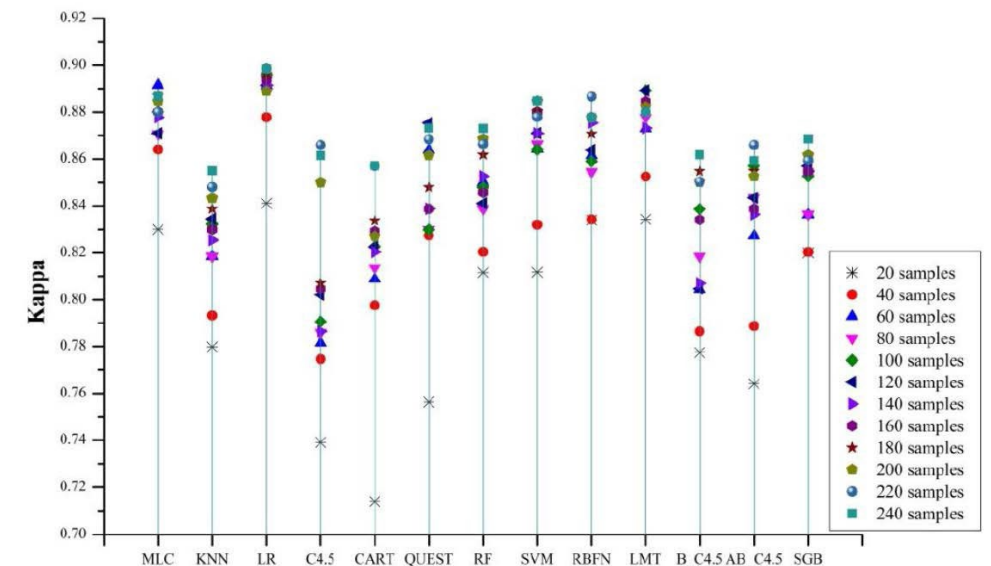
Sample data sets are dependent on the classification method used

Sample size

- High number of samples are beneficial for calibration and validation
- Trade-off between logistical reasons / monetary investment and the number of samples
- Training sample size for each class should not be less than 10-30 times the number of bands
- Sample size is strongly linked to the classifier



Waske et al., 2010



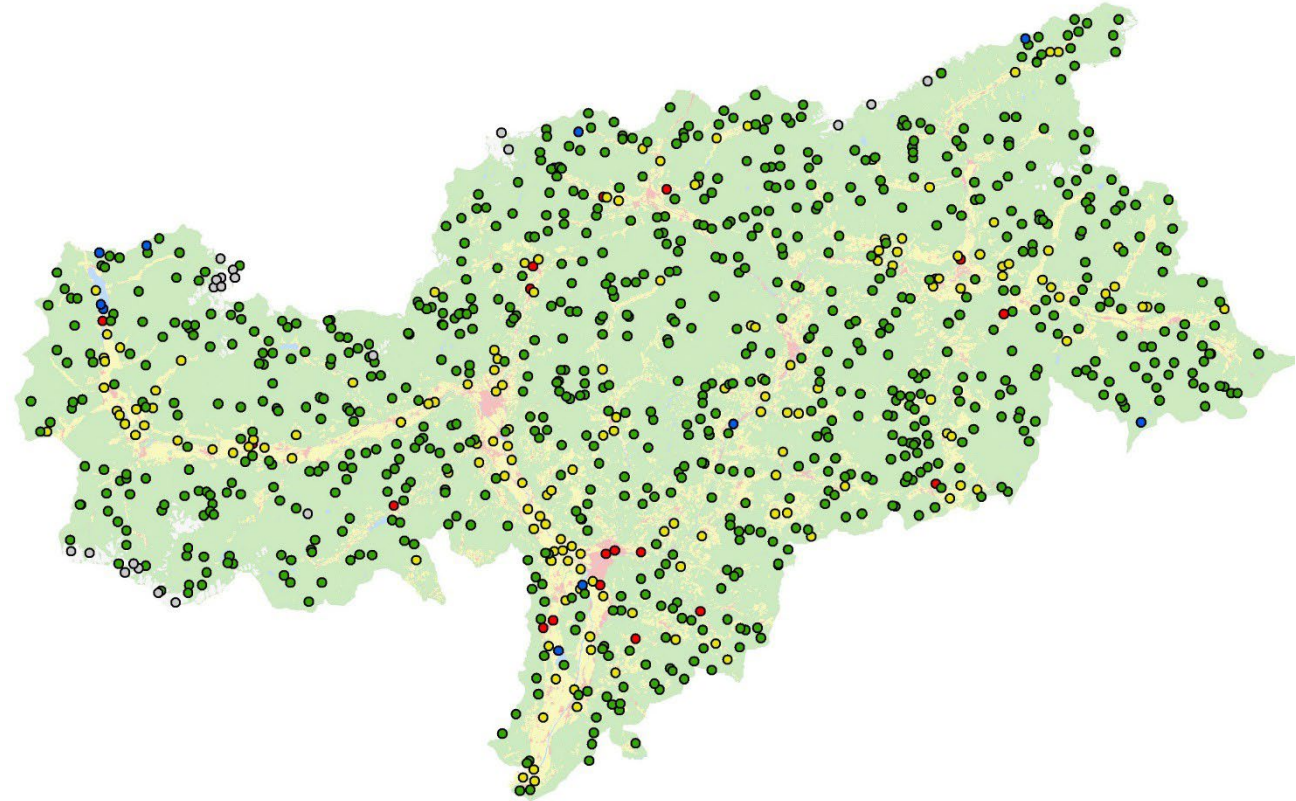
Li et al., 2014

Simple random sampling

- Simple random sampling
- Number of pixels / percentage of overall pixels

Example:

- Overall number of pixel: 1000
- Artificial surfaces: 18
- Agricultural areas: 157
- Forest: 794
- Glacier: 21
- Water: 10

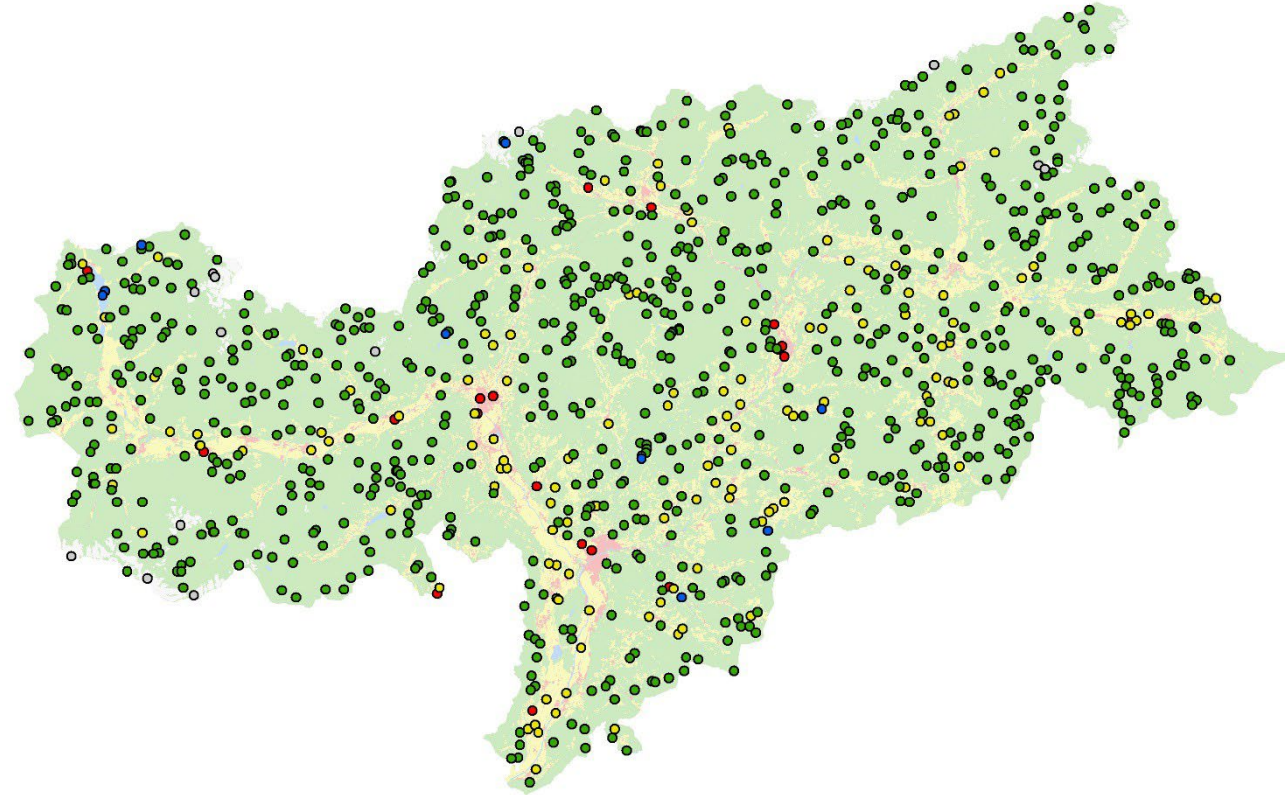


Stratified random sampling

- Number of samples in each stratum is proportional to the area of the class

Example:

- Overall sample size: $0.135\% = 1000$
- Artificial surfaces: 20
- Agricultural areas: 134
- Forest: 824
- Glacier: 13
- Water: 9

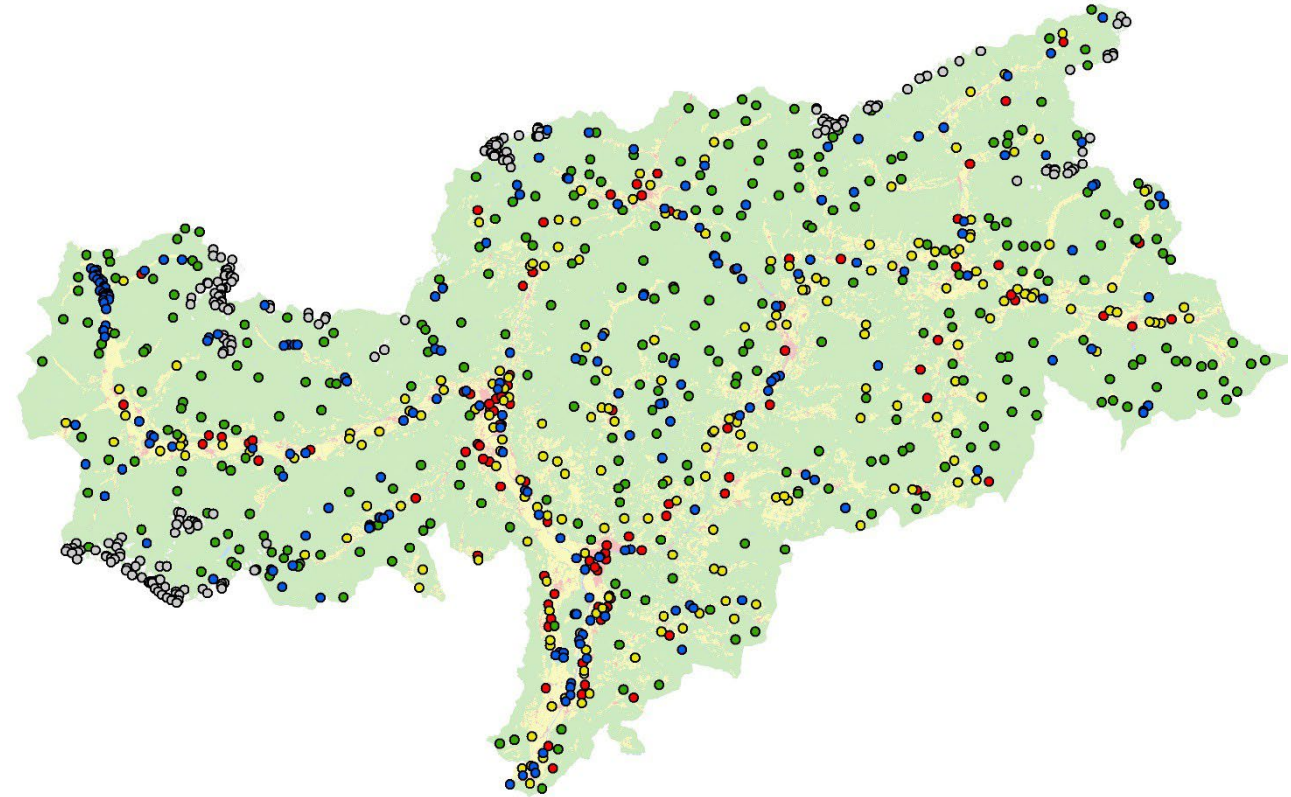


Disproportional stratified random sampling

- The size of the classes are specified and set to intermediate values between the proportions of the areas of the classes and a simple equalized approach

Example:

- Overall sample size: 1000
- Artificial surfaces: 100
- Agricultural areas: 200
- Forest: 300
- Glacier: 200
- Water: 200

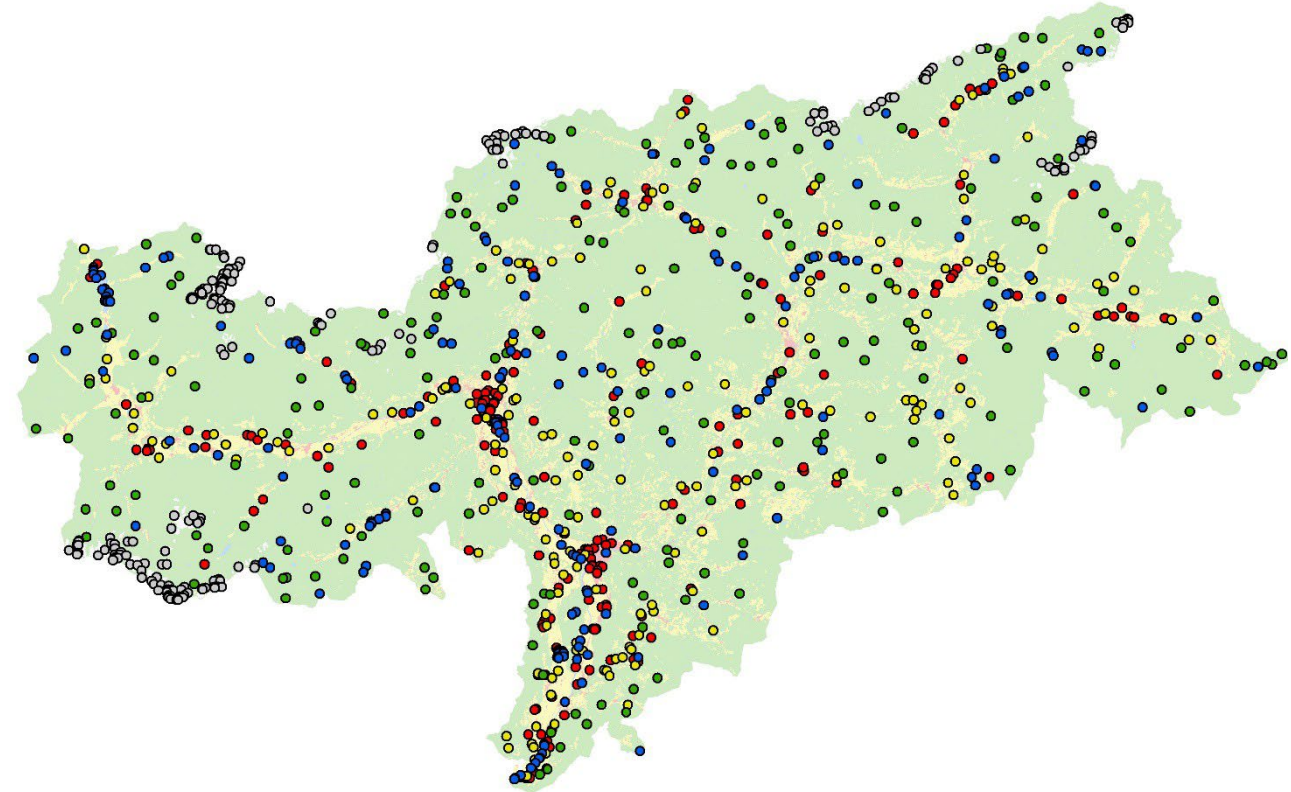


Equalized stratified random sampling

- Number of samples in each stratum is the same irrespective of the area on the map

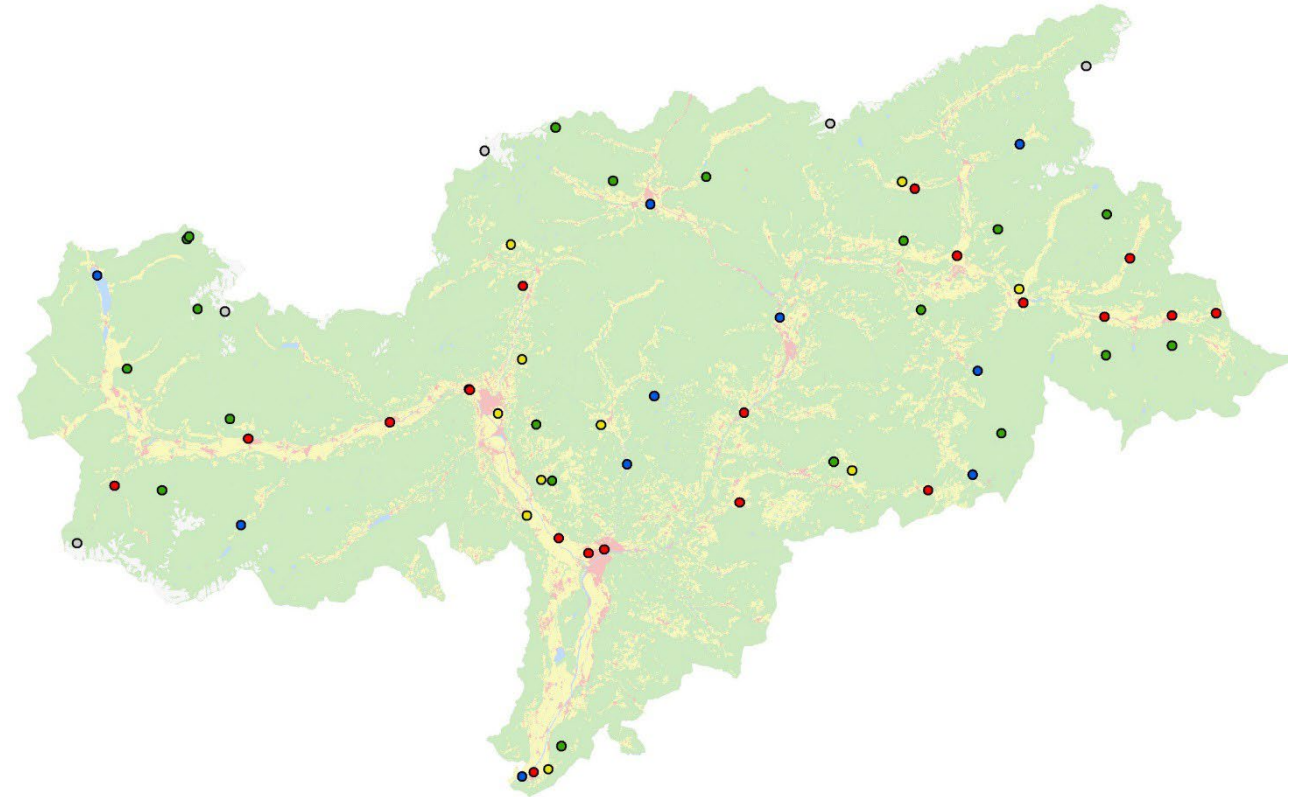
Example:

- Minimum sample size: 200
- Artificial surfaces: 200
- Agricultural areas: 200
- Forest: 200
- Glacier: 200
- Water: 200



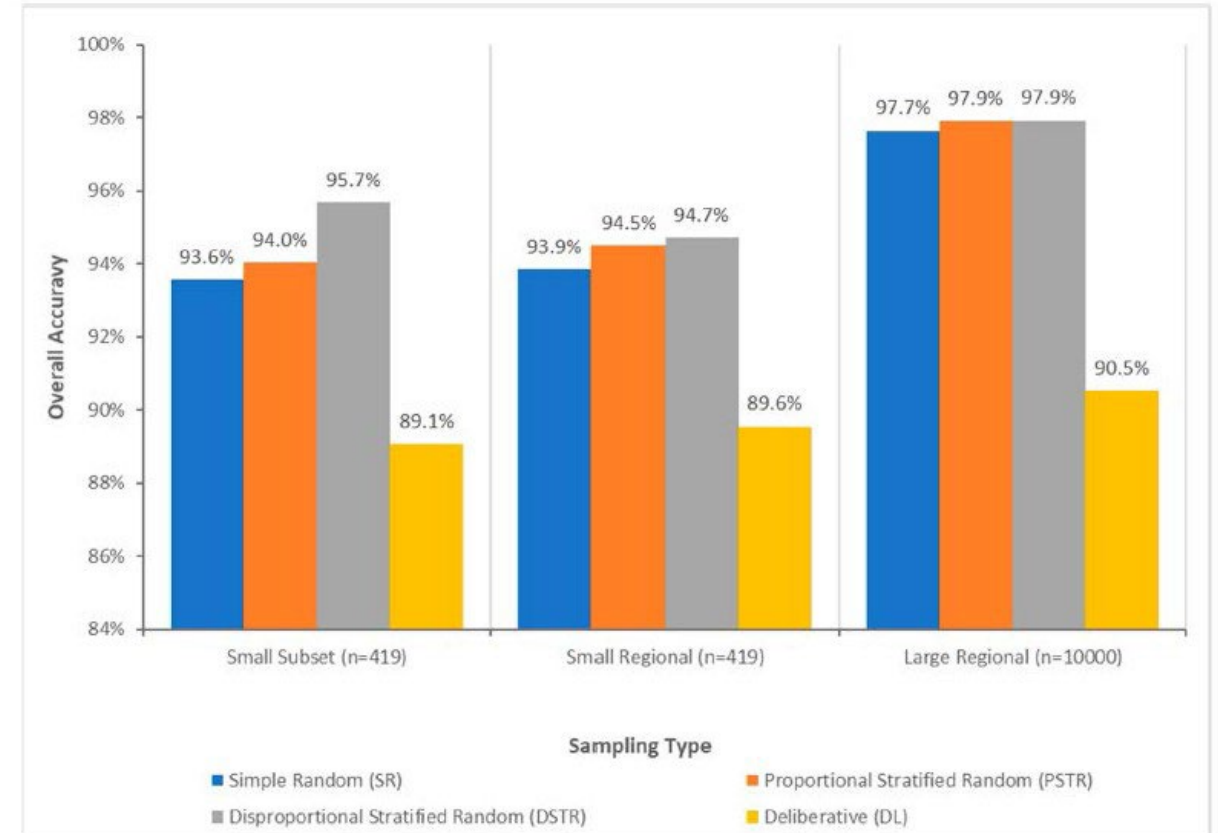
Deliberate sampling

- Selection based on non-random method
- Necessary if data access limitations or other issues constrain the sampling
- Allows for incorporation of expert knowledge into the sampling process



Spatial distribution of samples

- If in situ observations are required for sample characterization and the cost of traveling between sites is high, the spatial distribution of samples becomes a central focus.
- Samples should be well distributed across the entire study region



Ramezan et al., 2019

Summary

General recommendations:

- High number of samples are beneficial for calibration and validation
- Well distributed across space
- Avoid imbalanced training data sets
- Autocorrelation - minimum distance between pixels
- Mixed pixels - include or avoid depending on the classifier