Stratified Estimation and the Choice of Phase I Media

Objective

- Provide a non-technical description of stratified estimation and how stratified estimation influences the selection of a phase I media
- For technical description see
  - General Technical Report SRS-80 '*The Enhanced Forest Inventory and Analysis Program—National Sampling Design and Estimation Procedures*'
    - http://www.treesearch.fs.fed.us/pubs/20371
    - Supplemental material for point in time estimation
    - http://srsfia2.fs.fed.us/publicweb/statistics_band/stat_documents.htm
  - Avery, T.E., Burkhart, H.E. 1994. Forest Measurements 4th edition. McGraw-Hill, New York. 408 p.

Background

- What is FIA?
  - “FIA is a National Program implemented regionally”.
- Example Benefits/Drawbacks
  - Standardized field techniques
  - Standardized processing system
  - Standardized statistical techniques
  - Each region faces the same challenges

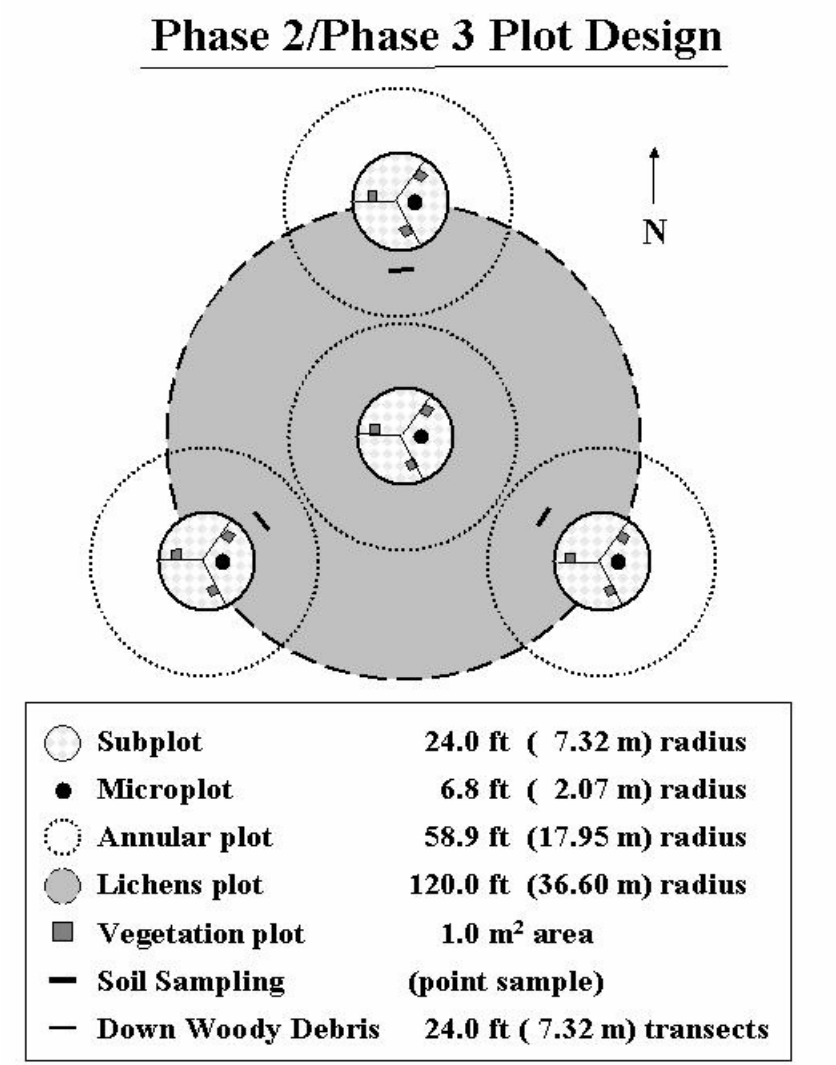
Standardized processing system and statistical techniques

- Processing system: NIMS (National Information Management System).
- Standardized estimation technique: stratified estimation.

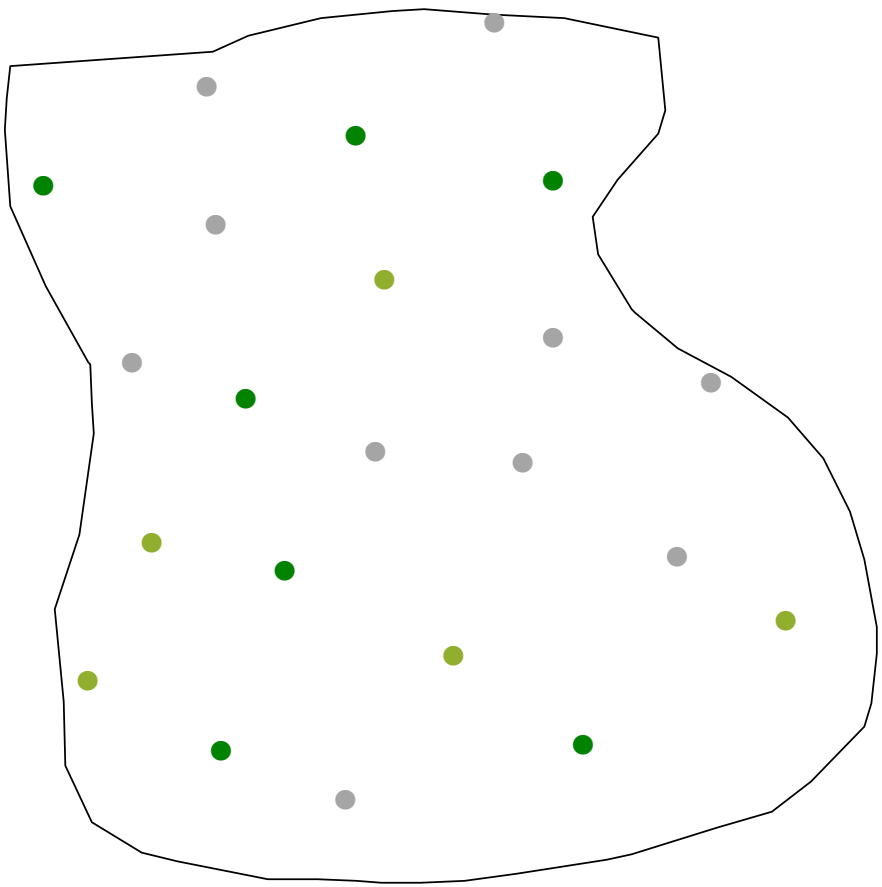
Message

- Given that FIA uses stratified estimation, the media that is used for Phase I has a negligible influence on forest area estimates but can influence certainty of the estimate.

Background: FIA plot design



Background: Simple random sampling



Dark Green Plot = 100% forest
Light Green Plot = 50% forested
Gray Plot = 0% forest

Total Area = 120ha

Average % forest from plots=43.2%

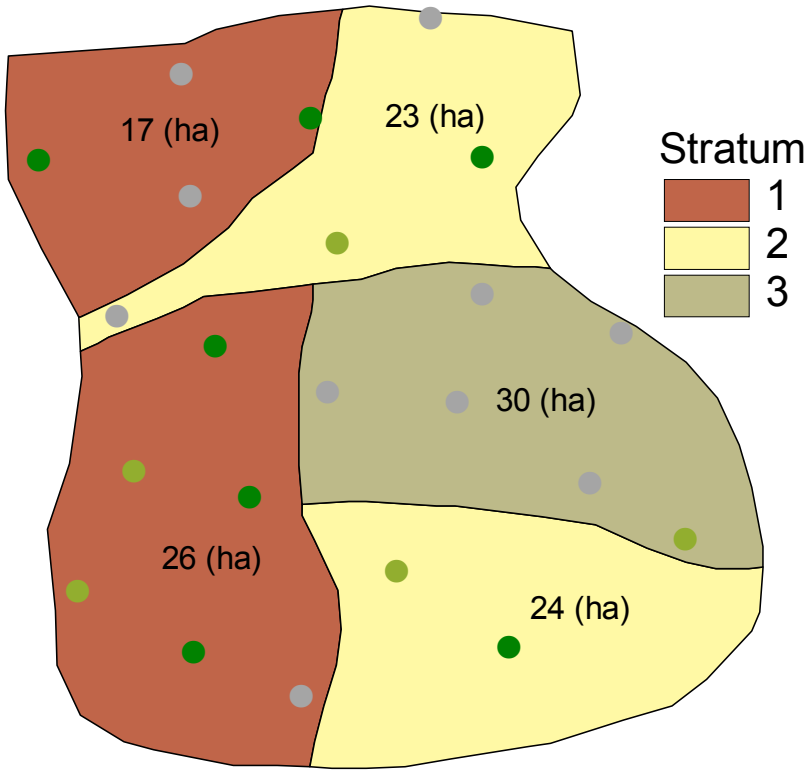
Forest Area=120(ha)*43.2%=51.8 ha

Total=51.8
Sampling error = 21.9%

Stratified Estimation

- Stratification is a statistical approach used to group the population into homogeneous strata.
- The purpose of stratification is variance reduction.
- Example: If we do a good job of grouping forested plots in the same stratum then the variance of our estimate of forest area will be lower.
- No single stratification is optimal for all forest attributes.
- The information used for stratification must be independent of the information collected during p2 plot visits

Stratified Estimation: A simple example



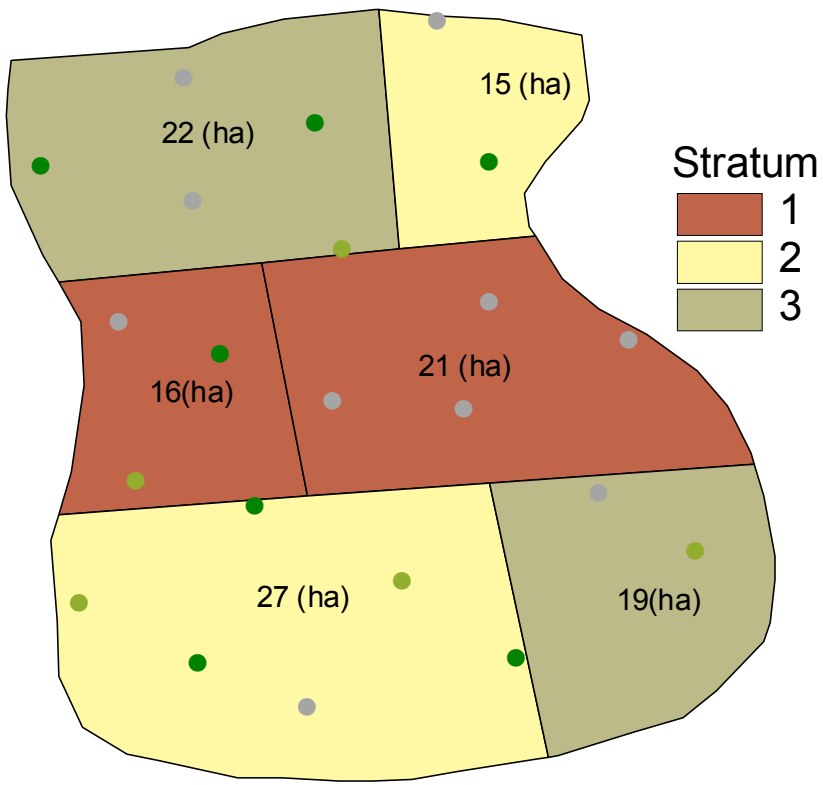
Dark Green Plot = 100% forest
Light Green Plot = 50% forested
Gray Plot = 0% forest

Total Area = 17+26+23+24+30=120ha
Stratum 1 area = 43 ha
Stratum 2 area = 47 ha
Stratum 3 area = 30 ha

Average % forest
Stratum 1=60%
Stratum 2=50%
Stratum 3 =8.3%

Forest Area
Stratum 1 = 0.60*43=25.8
Stratum 2 = 0.50*47=23.5
Stratum 3 = 0.083*30=2.5
Total=51.8
Sampling error = 20.8%

Stratified Estimation: Another simple example



Dark Green Plot = 100% forest
Light Green Plot = 50% forested
Gray Plot = 0% forest

Total Area = 16+21+15+27+22+19=120ha
Stratum 1 area = 37 ha
Stratum 2 area = 42 ha
Stratum 3 area = 41 ha

Average % forest
Stratum 1=21.4%
Stratum 2=62.5%
Stratum 3 =42.9%

Forest Area
Stratum 1 = 0.214*37=7.9
Stratum 2 = 0.625*42=26.3
Stratum 3 = 0.429*41=17.6
Total=51.8
Sampling error = 22.2%

Characteristics of Stratified Estimation

- From the previous example, each stratum contributes to the forest area estimate.
- Information from the plots carries the most weight in the estimate
- The ability to place the plot into the correct stratum is based on spatial overlay is vital. (i.e. good GPS coordinates are important)

Choice of Phase I Media

- Goal of stratification is to group like plots together to reduce the variance of estimates:
  - Examples of different Phase I Media for stratification
    - Photo Interpretation
    - Soil type map
    - Classified satellite imagery
      - NLCD
      - MODIS
    - Classified digital photography
    - Others

What is the National Land Cover Database (NLCD)?

- NLCD is:
  - a land cover classification based on **LANDSAT** (30meter pixels) imagery and ancillary data.
  - Produced by the Multi-Resolution Land Cover (MRLC) consortium.
  - Available free from http://www.mrlc.gov/
- The MRLC is a consortium of the following agencies and programs:

EPA

NOAA

USFS

USGS

LANDFIRE

BLM

NRCS

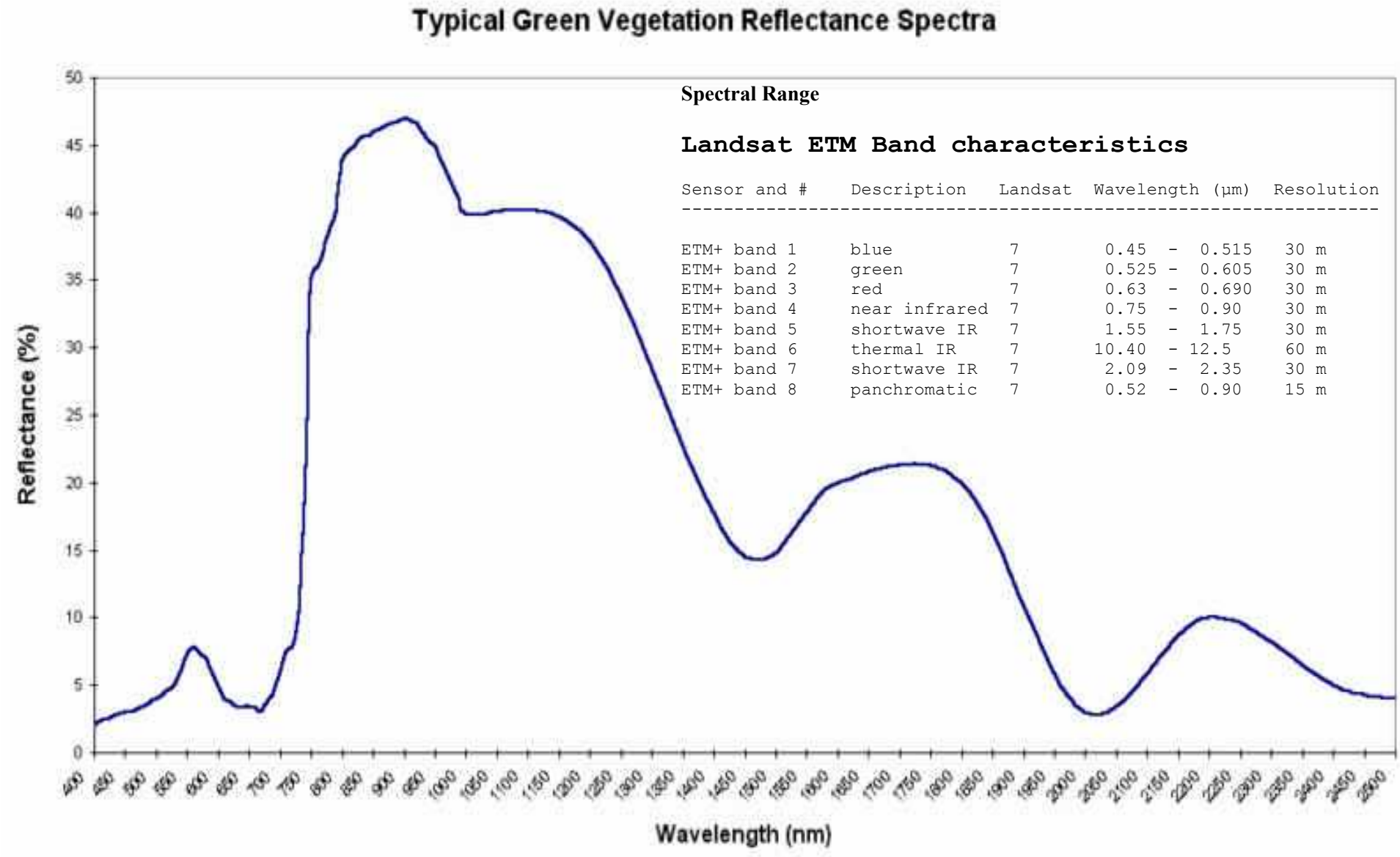
NPS

NASA

USFWS

OSM

What is LANDSAT TM, ETM

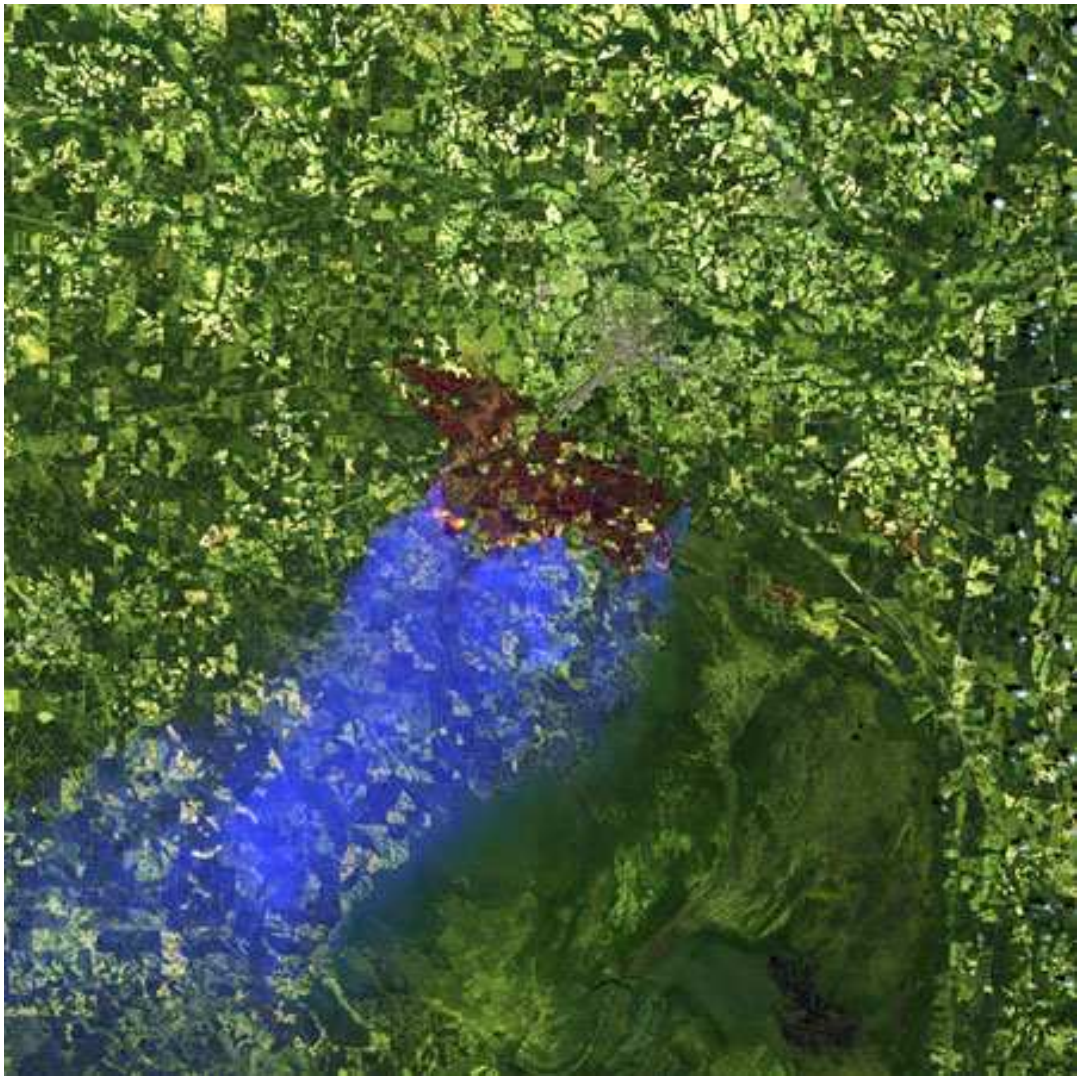


Spectral Range

Landsat ETM Band characteristics

Sensor and #	Description	Landsat	Wavelength (µm)	Resolution
ETM+ band 1	blue	7	0.45 - 0.515	30 m
ETM+ band 2	green	7	0.525 - 0.605	30 m
ETM+ band 3	red	7	0.63 - 0.690	30 m
ETM+ band 4	near infrared	7	0.75 - 0.90	30 m
ETM+ band 5	shortwave IR	7	1.55 - 1.75	30 m
ETM+ band 6	thermal IR	7	10.40 - 12.5	60 m
ETM+ band 7	shortwave IR	7	2.09 - 2.35	30 m
ETM+ band 8	panchromatic	7	0.52 - 0.90	15 m

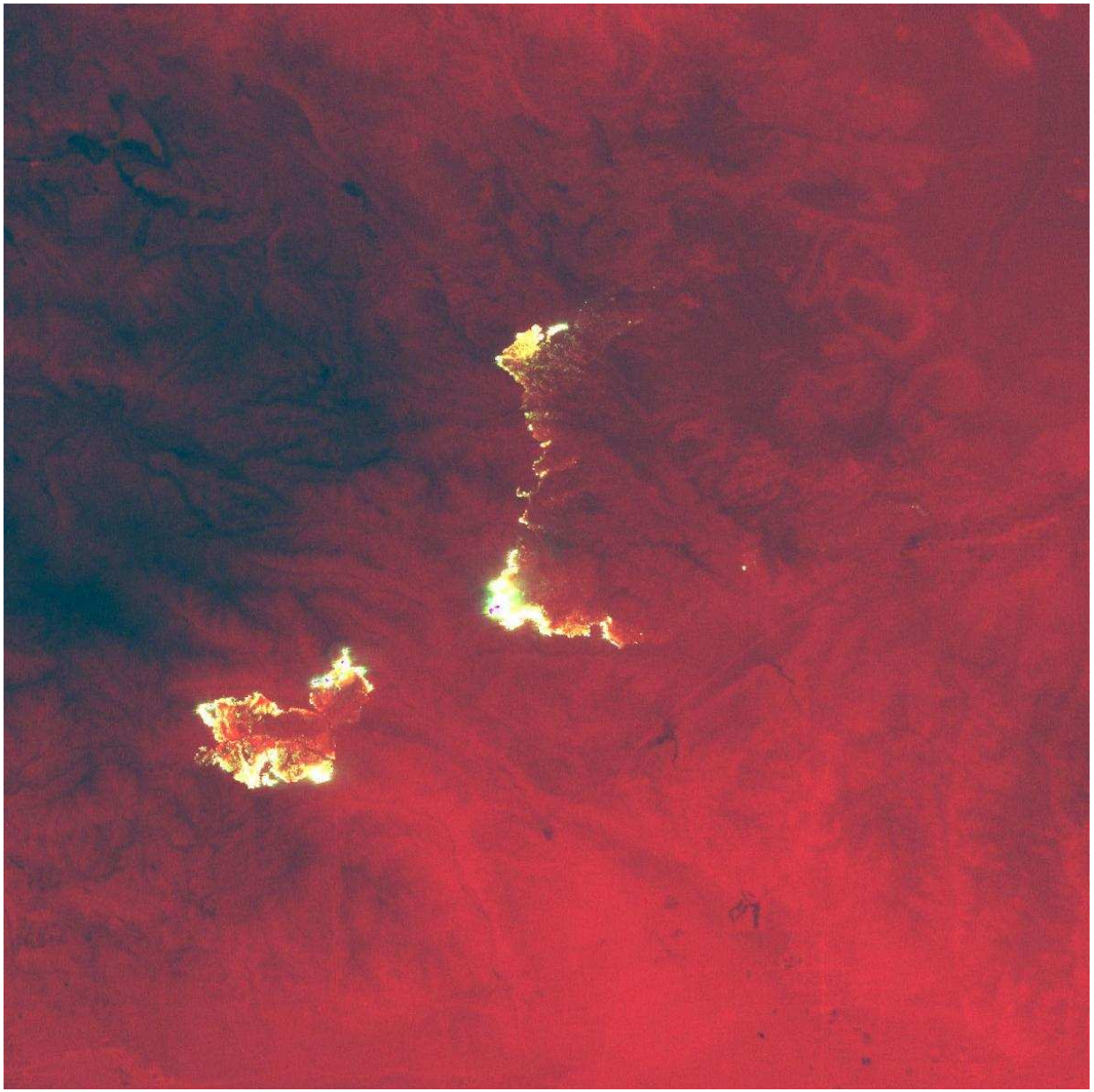
Houston, TX
Bands 4,2,1



Georgia Fires:

April 17, 2007

Bands: 7,5,2



Wildfires in the San Bernardino Mountains at night.

July 2006

Bands: 6,5,7

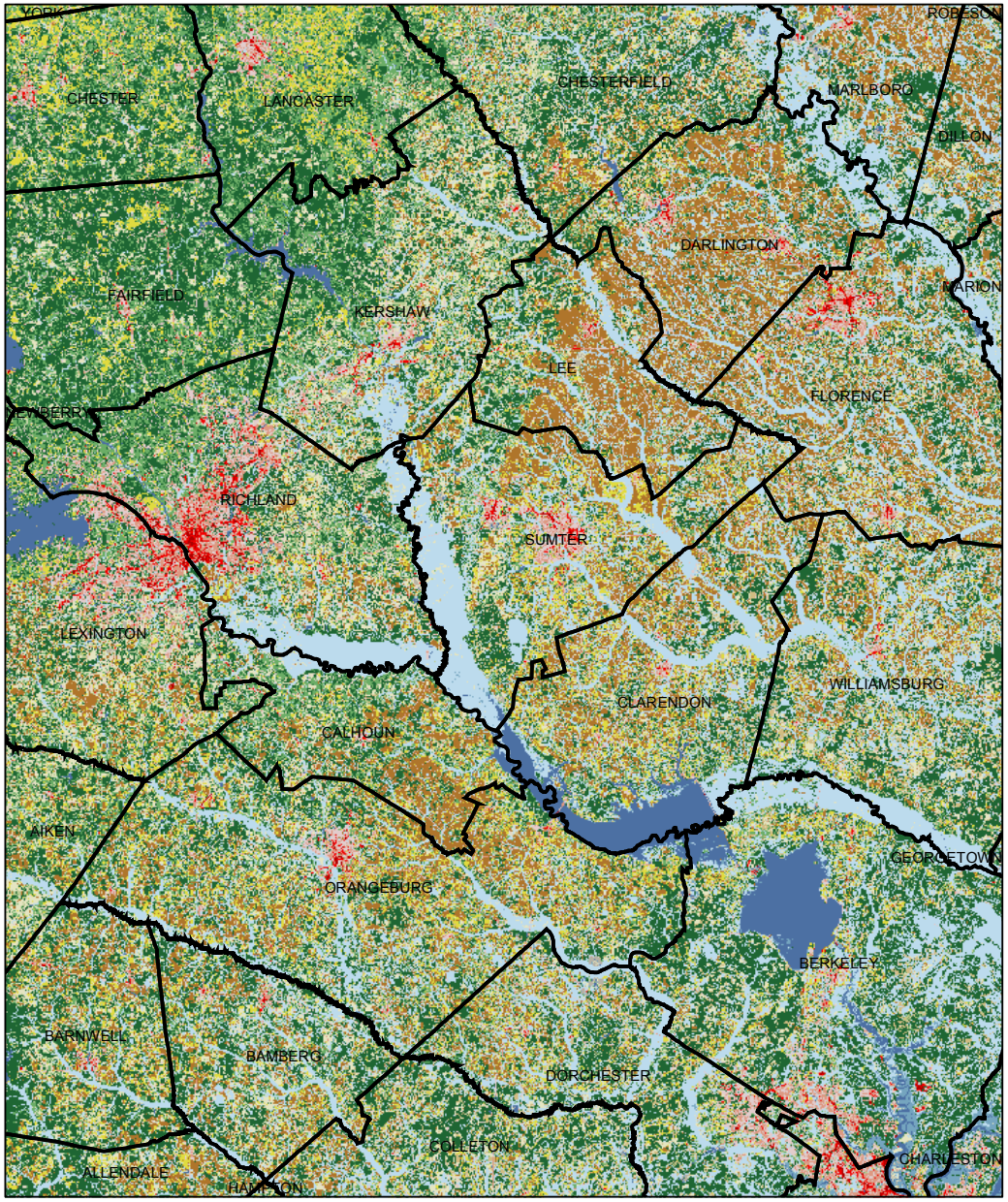
How is the NLCD Created?

- Identification of optimal early, peak and late Landsat TM scene acquisitions for each path/row based on AVHRR seasonal averages. (Yang et al)
- Development of Landsat 7 radiometric and at-sensor reflectance calibrations for imagery to reduce mosaicing difficulty. (Huang et al)
- Development of Tasseled Cap coefficients for Landsat 7 at-sensor reflectance (Huang et al)
- Justification of Tasseled Cap as an optimal way of distilling 6 bands into 3 (Wylie et al)
- Regional mapping using land cover mapping zones (Homer et al)
- Development of imperviousness estimates from TM imagery. (Yang et al)
- Development of tree canopy estimates from TM imagery. (Huang et al)
- Development of image processing rules coupled with ancillary data as needed to classify Landsat data into land cover.
- Comprehensive accuracy assessments of the land cover data and related components.

What Land Cover Classes are Mapped?

2001 NLCD Classes

11. Open Water
12. Perennial Ice/Snow
21. Developed, Open Space
22. Developed, Low Intensity
23. Developed, Medium Intensity
24. Developed, High Intensity
31. Barren Land (Rock/Sand/Clay)
32. Unconsolidated Shore* -
41. Deciduous Forest
42. Evergreen Forest
43. Mixed Forest
51. Dwarf Scrub
52. Shrub/Scrub
71. Grassland/Herbaceous
72. Sedge/Herbaceous
73. Lichens
74. Moss
81. Pasture/Hay
82. Cultivated Crops
90. Woody Wetlands
95. Emergent Herbaceous Wetlands



Land Cover Class

- Open Water
- Developed, Open Space
- Developed, Low Intensity
- Developed, Medium Intensity
- Developed, High Intensity
- Barren Land (Rock/Sand/Clay)
- Deciduous Forest
- Evergreen Forest
- Mixed Forest
- Shrub/Scrub
- Grassland/Herbaceous
- Pasture/Hay
- Cultivated Crops
- Woody Wetlands
- Emergent Herbaceous Wetlands

2001 NLCD:

Part of South Carolina

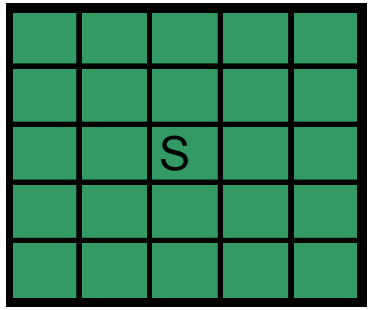
We want to use the NLCD for Stratified Estimation

- There are many ways to use the NLCD but we want a classification that will reduce variance.
- Most of the misclassification of satellite imagery occurs at edges between classes.
- Certain forest characteristics can be associated with edge effects
- We will use this information to construct our stratification
  - Forest (forest pixel surrounded by a 2 pixel buffer of forest pixels)
  - Forest edge (forest pixel within 2 pixels of non-forest
  - Non-forest (non-forest pixel surrounded by a 2 pixel buffer of non-forest pixels.
  - Non-forest edge (non-forest pixel within 2 pixels of forest)

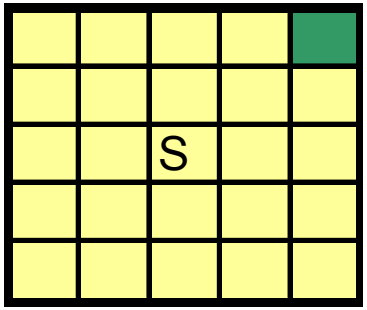
Forest, Non-Forest, Edge

- We use a 5x5 sliding window approach.

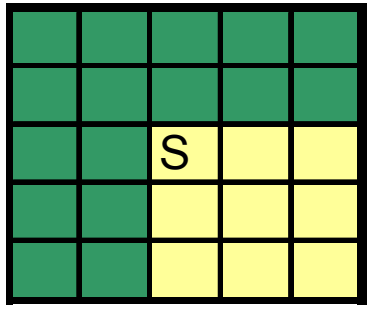
Forest



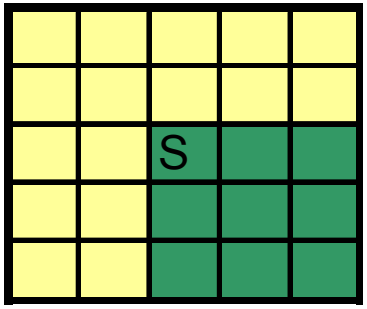
Non-forest



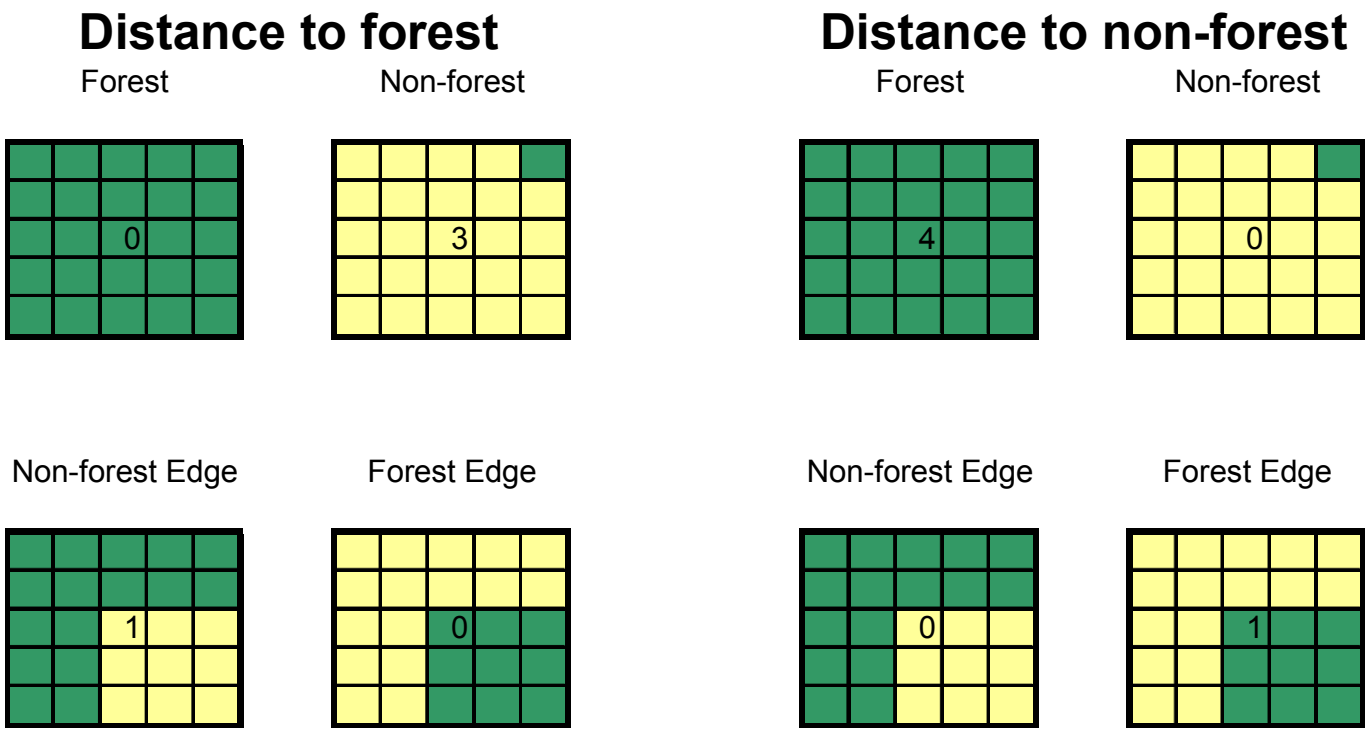
Non-forest Edge



Forest Edge



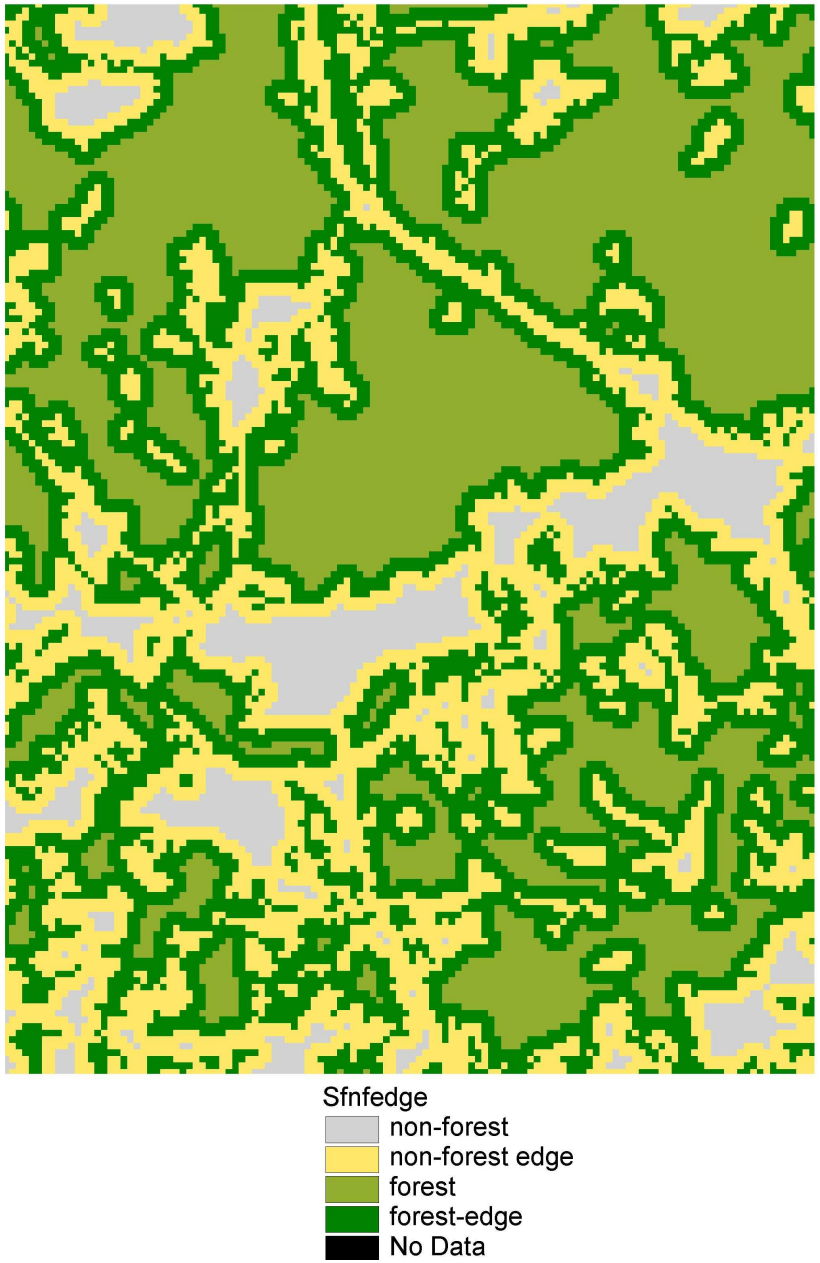
How the Model Works



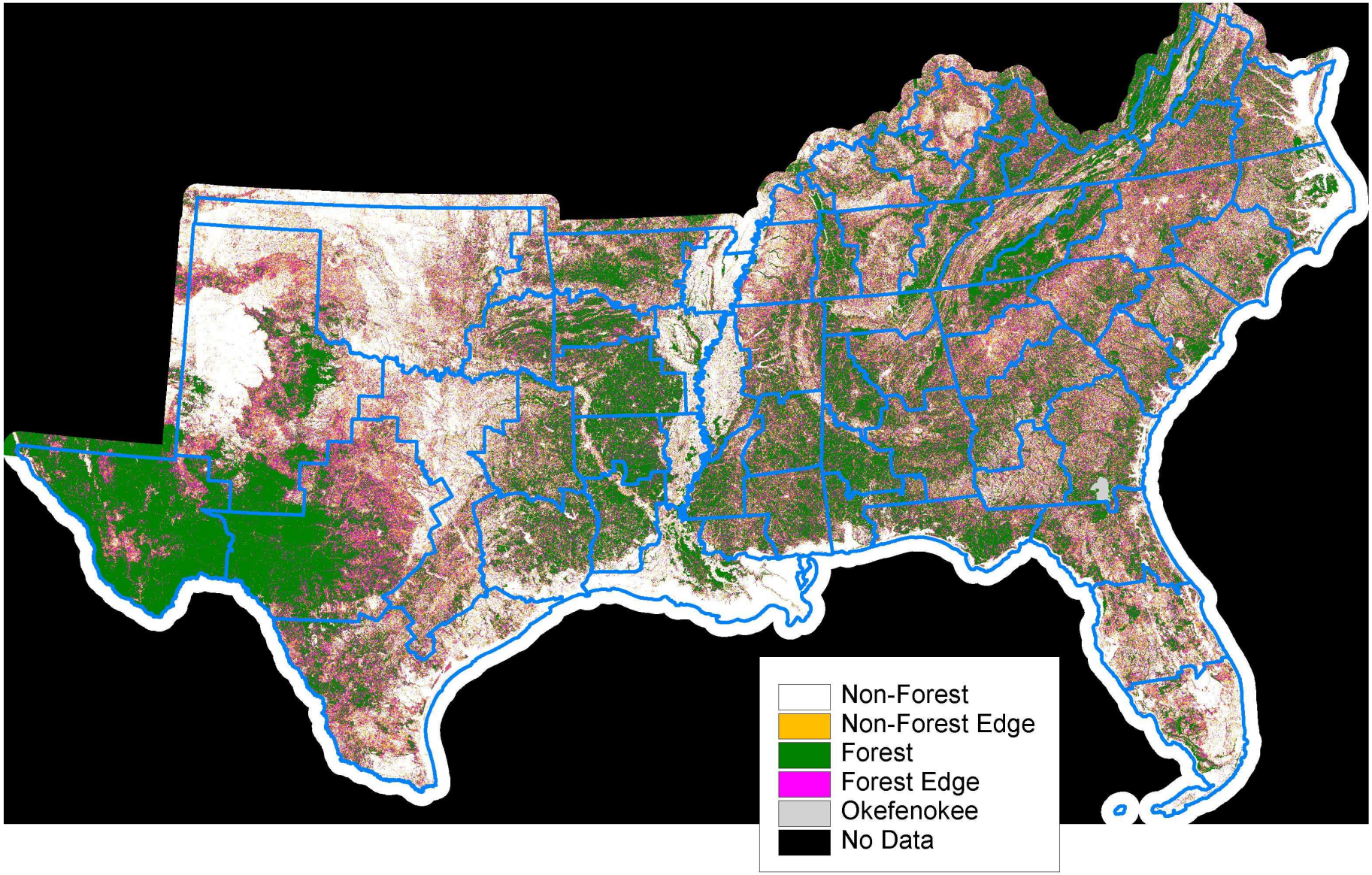
1. If subject pixel is yellow and distance to forest (dtf) >= 3 and Distance from non-forest (dtnf)=0 then output=non-forest
2. If subject pixel is yellow and dtf<3 and dtnf=0 then output=non-forest edge
3. If subject pixel is green and dtf=0 and dtnf>=3 then output=forest
4. If subject pixel is green and dtf=0 and dtnf<3 then output=forest-edge

Forest, Non-Forest, Edge

- Do for each of the $2x10^9$ pixels cover the South



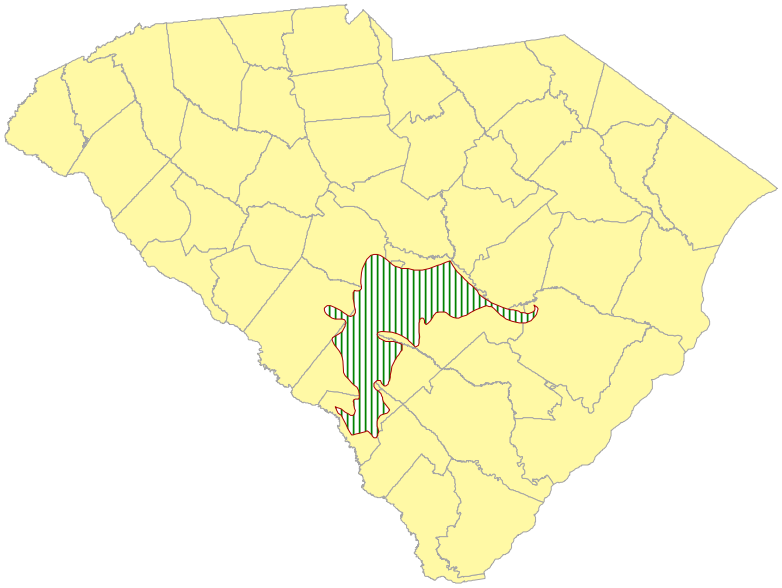
NLCD Phase I: South



Phase I Information For Stratified Estimation: Plot Stratum Assignments and Estimation Unit Area

- NIMS requirements
  - Each phase 2 plot must be assigned to a stratum.
  - For each estimation unit we need to know the proportion of each stratum in the estimation unit.
- Phase 1 alternatives
  - PI
    - Photo interpret 27x grid
    - Photo interpret P2 plots
    - To avoid bias, all interpretation (both Phase I and II) should be done by the same person, using the same imagery, on the same computer.
  - NLCD
    - Sample the Phase I map with the plot locations
    - Summarize the estimation units
    - These steps are automated and generally take about 30 minutes per state to complete.

Forest Area Estimation: Real examples



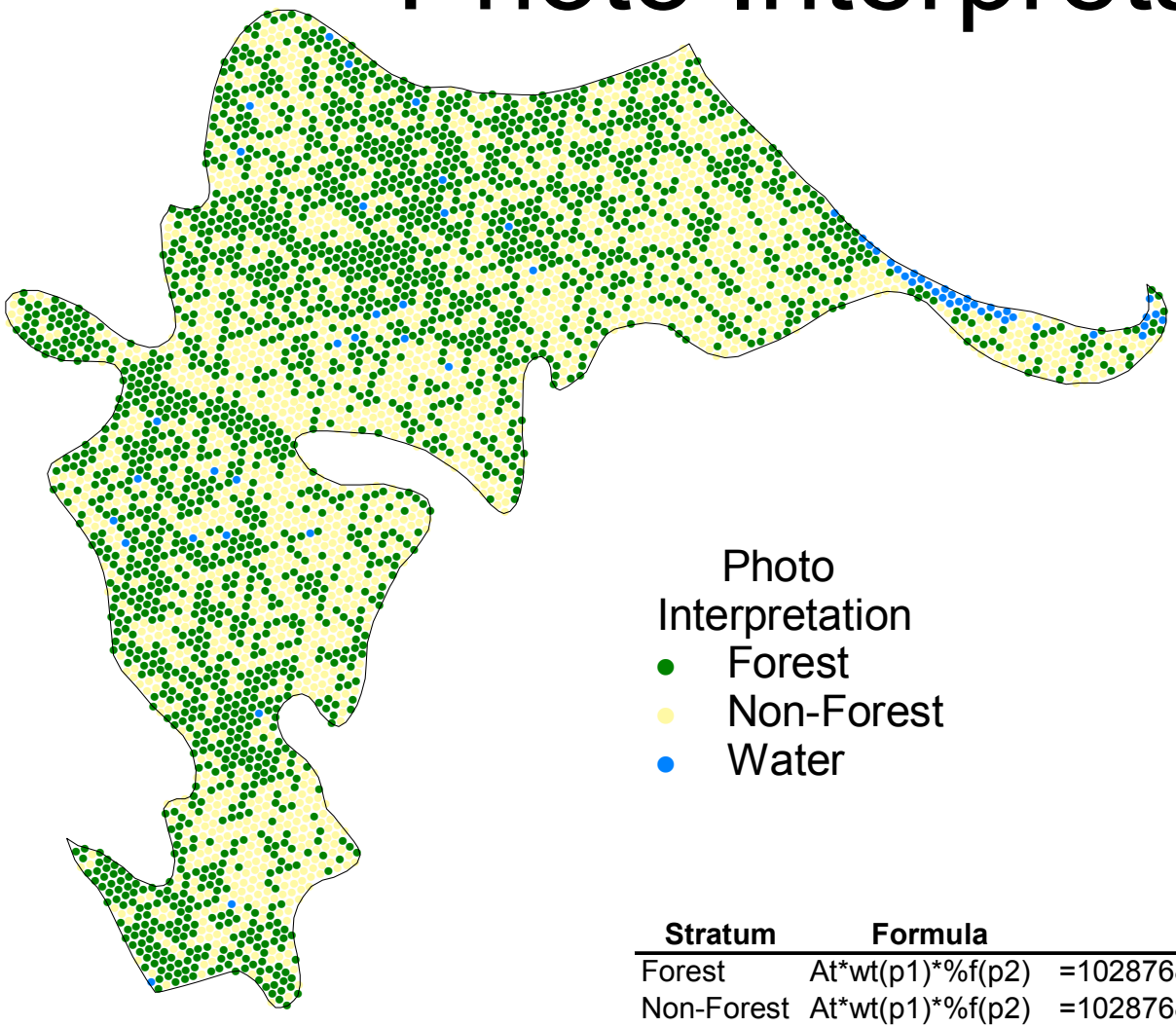
Goal: Provide a estimate of forest area
Population: Loam Hill ecoregion subsection
Total Area: 1,028,768 acres
Method1: Stratified estimation
Phase I:

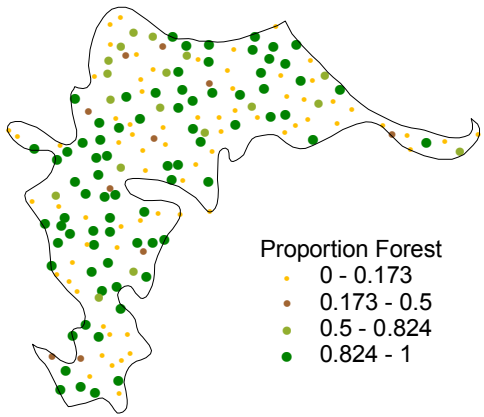
- Photo Interpretation
- NLCD
- Elevation Class

Method2: Double sampling
Phase I:

- Photo Interpretation

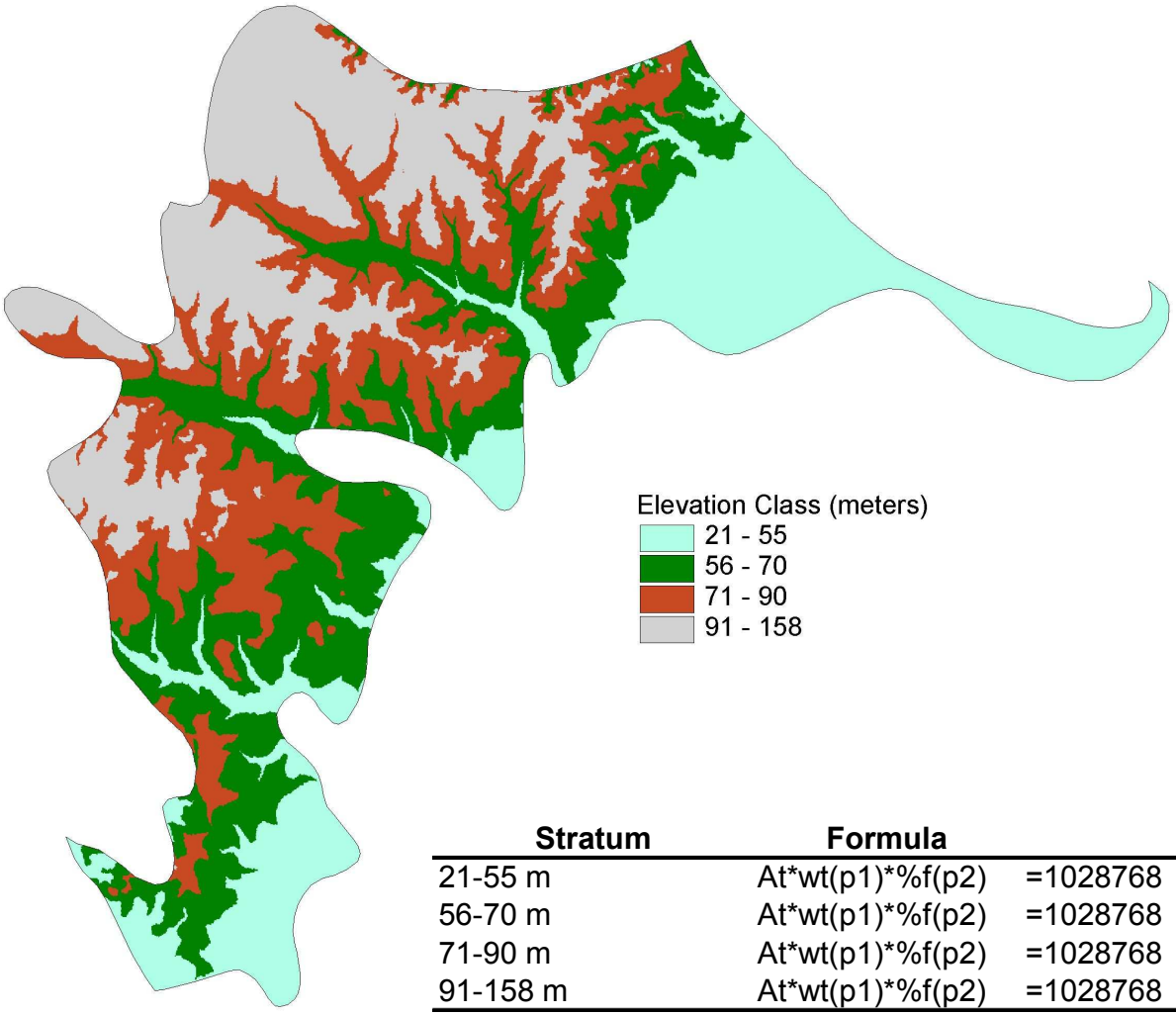
Photo Interpretation

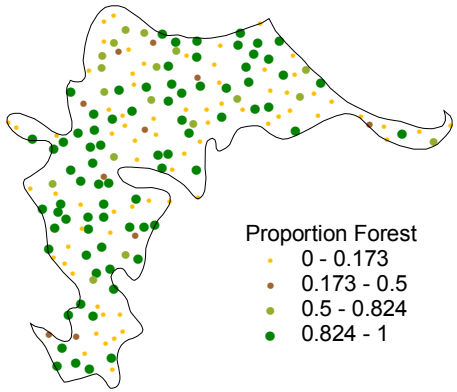




Stratum	Formula						Forest Area (ac)
Forest	At*wt(p1)*%f(p2)	=1028768	*	0.513	*	0.888	469,251
Non-Forest	At*wt(p1)*%f(p2)	=1028768	*	0.474	*	0.17	83,003
Water	At*wt(p1)*%f(p2)	=1028768	*	0.013	*	0	0
						total forest area (ac)	**552,254**
						percent sampling error	4.1

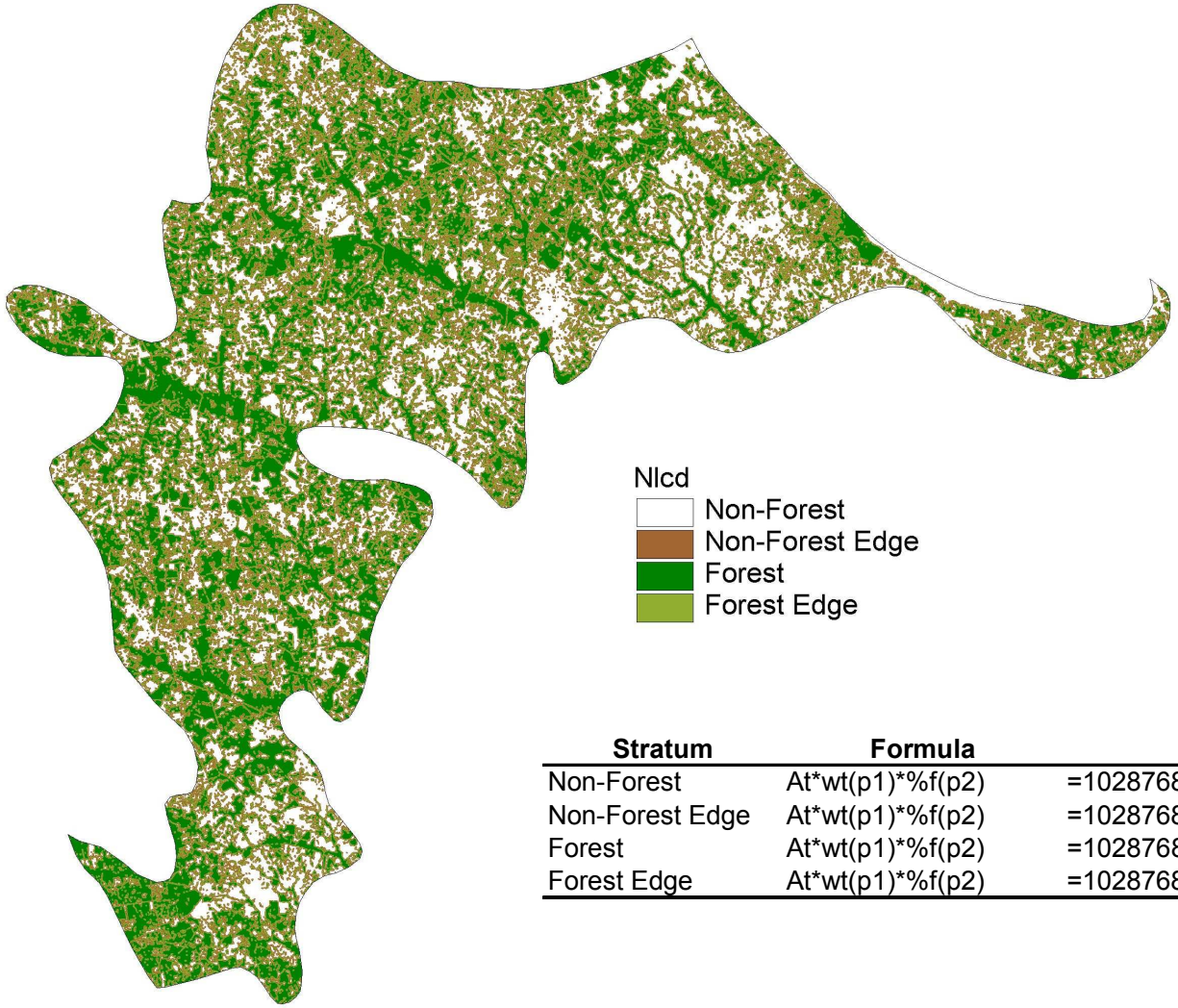
Elevation Class Stratification

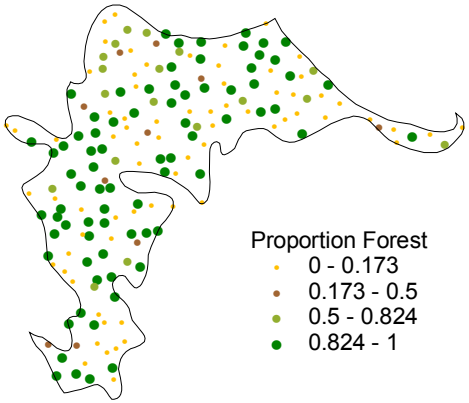




Stratum	Formula					Forest Area (ac)
21-55 m	At*wt(p1)*%f(p2)	=1028768	*	0.2609	* 0.461	123,874
56-70 m	At*wt(p1)*%f(p2)	=1028768	*	0.2494	* 0.577	147,927
71-90 m	At*wt(p1)*%f(p2)	=1028768	*	0.2474	* 0.686	174741
91-158 m	At*wt(p1)*%f(p2)	=1028768	*	0.2422	* 0.434	108064
					total forest area (ac)	**554,606**
					percent sampling error	6.46

NLCD Stratification





Stratum	Formula		Forest Area (ac)
Non-Forest	At*wt(p1)*%f(p2)	=1028768 * 0.2462 * 0.238	6,032
Non-Forest Edge	At*wt(p1)*%f(p2)	=1028768 * 0.2419 * 0.397	98,693
Forest	At*wt(p1)*%f(p2)	=1028768 * 0.2639 * 0.995	270108
Forest Edge	At*wt(p1)*%f(p2)	=1028768 * 0.2479 * 0.799	203800
		total forest area (ac)	**578,633**
		percent sampling error	3.76

Double Sampling for area: Photo Interpretation

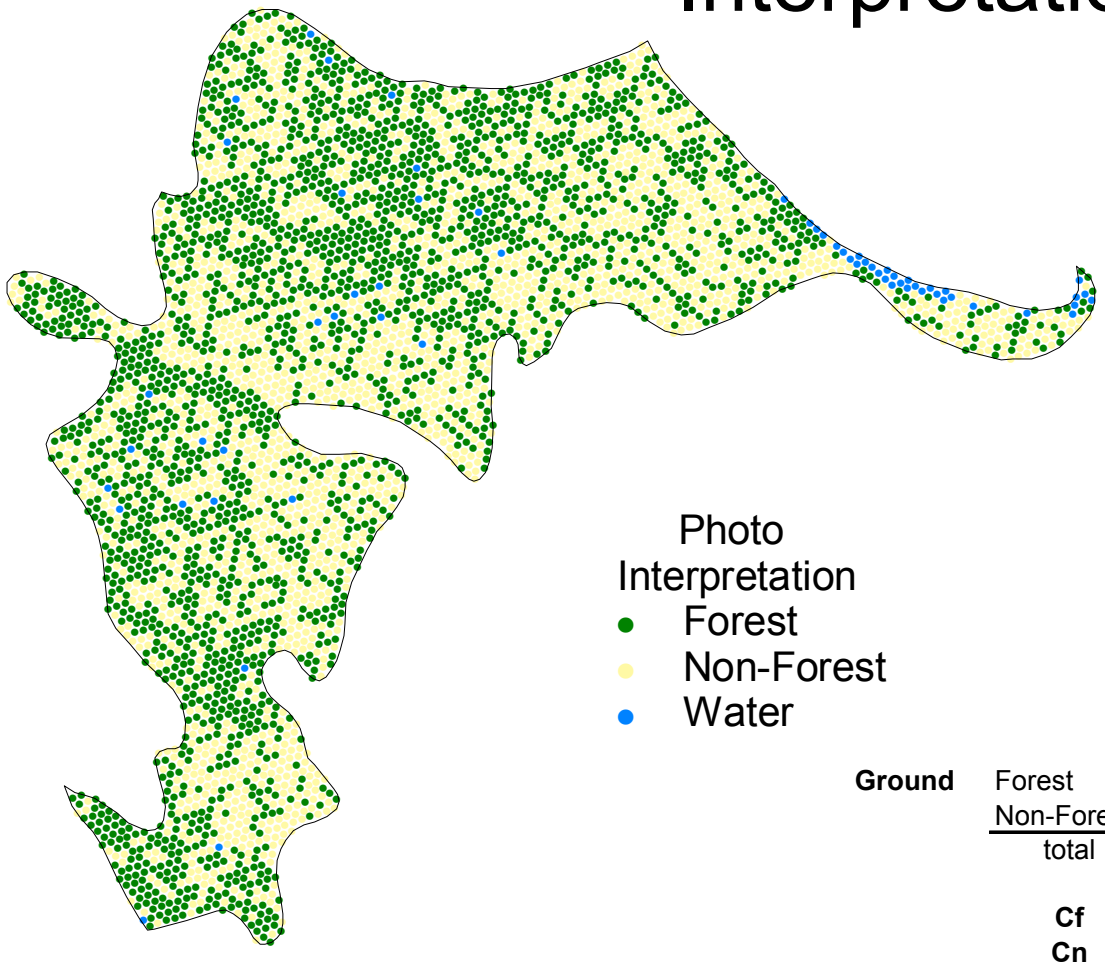
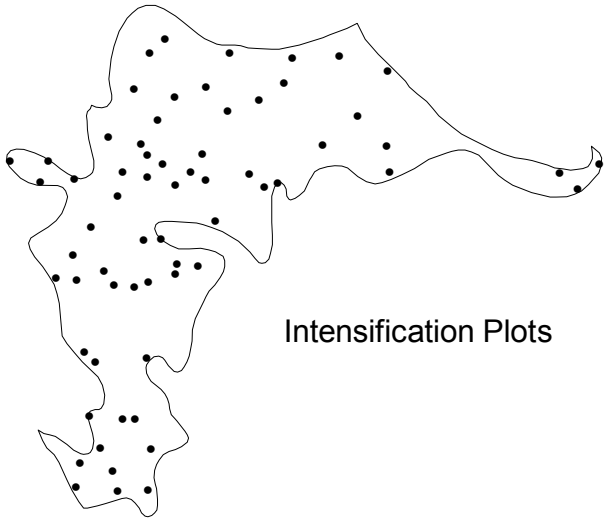


Photo Interpretation
- Forest
- Non-Forest
- Water



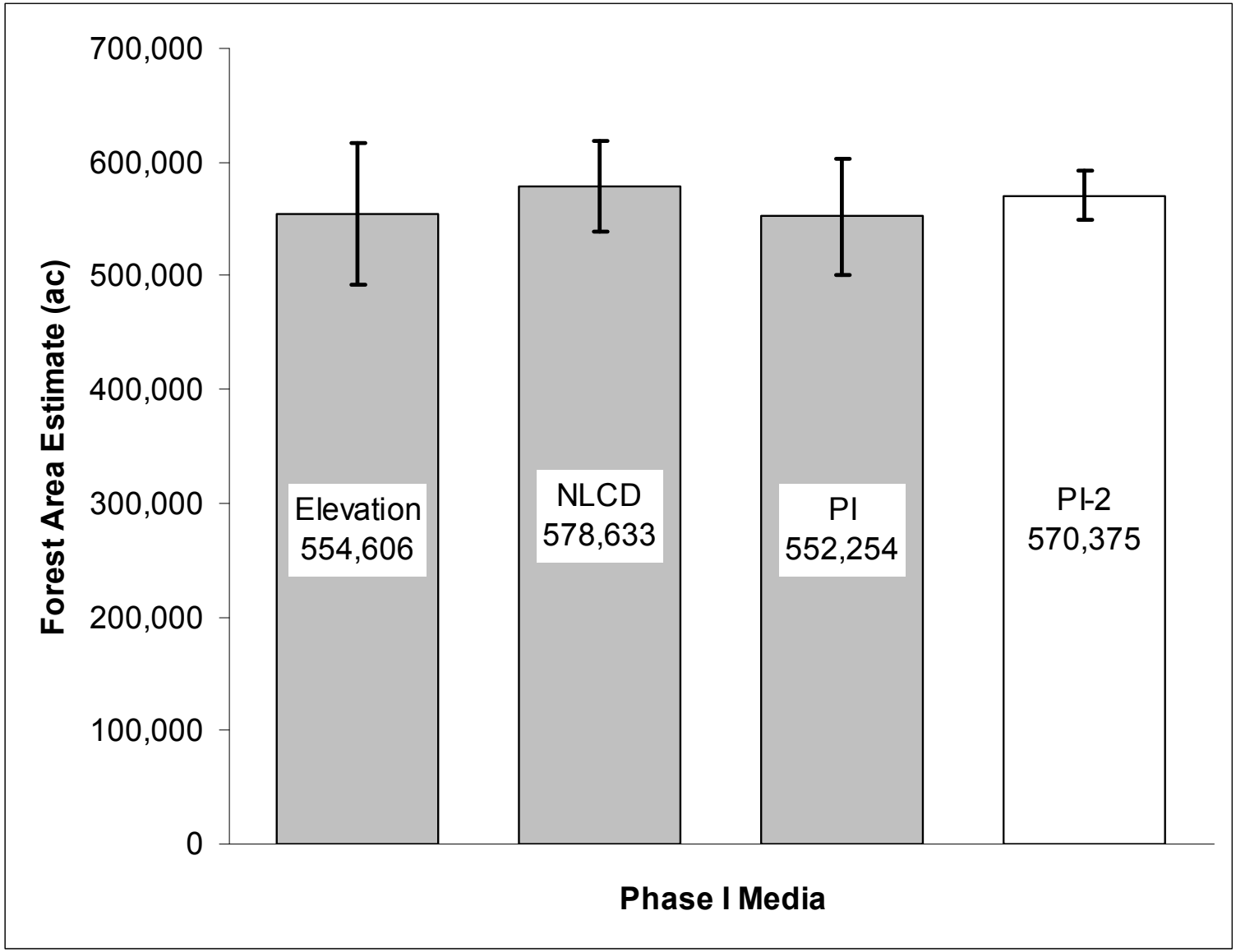
Intensification Plots

		Photo Interpretation		
		Forest	Non-forest	total
Ground	Forest	112	17	129
	Non-Forest	6	106	112
	total	118	123	241

Cf	0.949
Cn	0.138

Proprtion Forest =((2402*0.949) + (2278*0.138)) / (2402+2278) = 0.554
Area Forest = 0.554 *1028768 = 570,374
Percent Sampling error = 1.9

All Estimates Have a Level of Certainty: Forest Area Estimates and 95% Confidence Intervals



Conclusions from the example

- All estimates are statistically the same
- Sampling errors: stratified estimation
  - Lowest: NLCD
  - Middle: Photo Interpretation
  - Highest: Elevation
- Double sampling for area:
  - Estimate had half the sampling error.
  - Intensification plots cost approximately $6000 per million ac.
  - NIMS does not produce currently produce double sampling for area estimates
- Therefore: the optimal phase I media should minimize cost and sampling error

Costs: Manual PI

Photo Interpretation

23 class Example KY

Total P1 plots	117605		
Employee	hrs	salary+fringe	cost
XXX	162	XXX	XXX
XXX	78	XXX	XXX
XXX	91	XXX	XXX
XXX	12	XXX	XXX
		Total cost ($)	9204.38
		Cost ($/plot)	0.078

4 class Example SC

Total P1 plots	93190		
Employee	hrs	salary+fringe	cost
XXX	168	XXX	XXX
XXX	14	XXX	XXX
XXX	10	XXX	XXX
		Total cost ($)	4796.82
		Cost ($/plot)	0.0515

Estimated cost for TX

Total P1 plots	781785
Total Cost--23 class ($)	61186.57
Total Cost--4 class ($)	40241.25

Costs: Automated Phase 1

Automated Phase I

Activity	Employee	salary+fringe	hrs
Implement code to process NLCD (Source: Andy Lister NRS)	XXa	XXX	10
Error check and product comparison	XXb	XXX	10
Write code to create EUS and PSA tables (This code was given to NRS to help with their processing)	XXa	XXX	30
Technology transfer with IM	XXa	XXX	15
		Total Cost ($)	3075.7625

Estimated cost for TX	hrs
IM Phase I processing	0.5

Conclusions

- With stratified estimation the choice of phase I media will have little influence on estimates.
- The choice of phase I media will influence the variance of population estimates.
- There are several phase I media options
- No single stratification will be optimal for all variables.
- We can realize significant cost saving by using automated approaches to phase I stratification and the introduction of bias is minimized.