LiDAR applications in forest inventory and related statistical issues

Estimations of forest biomass with airborne and satellite optical data: parametric vs. non-parametric approaches

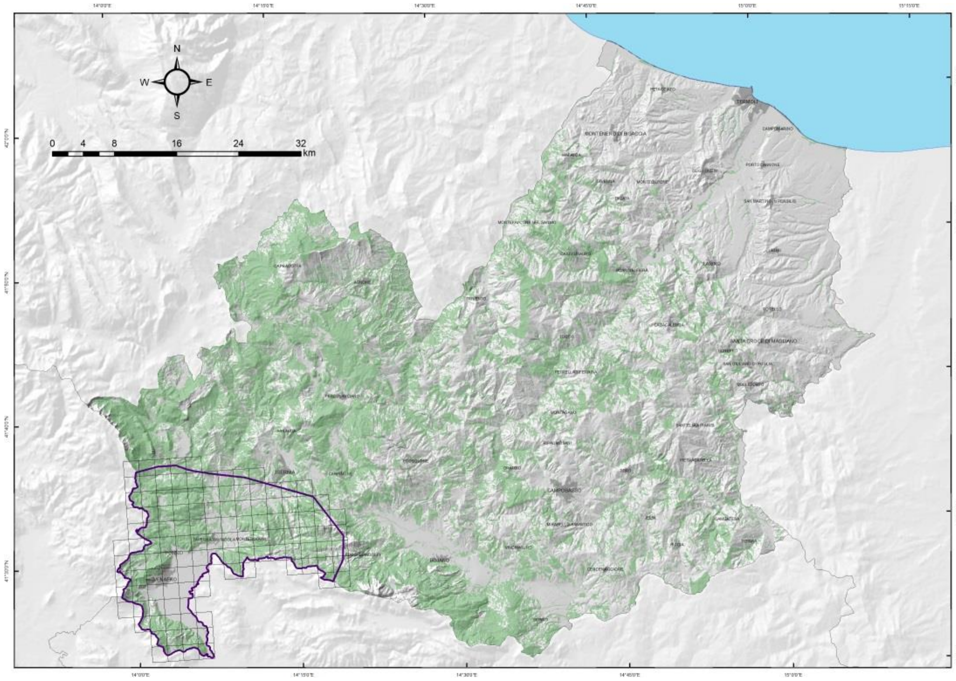
Chirici G.[1], McRoberts M.E.[2], Lopez G.[1], Mura M.[1], Tonti D.[1], Marchetti M.[1]

[1] Università degli Studi del Molise
[2] USDA, Forest Service

AIMS

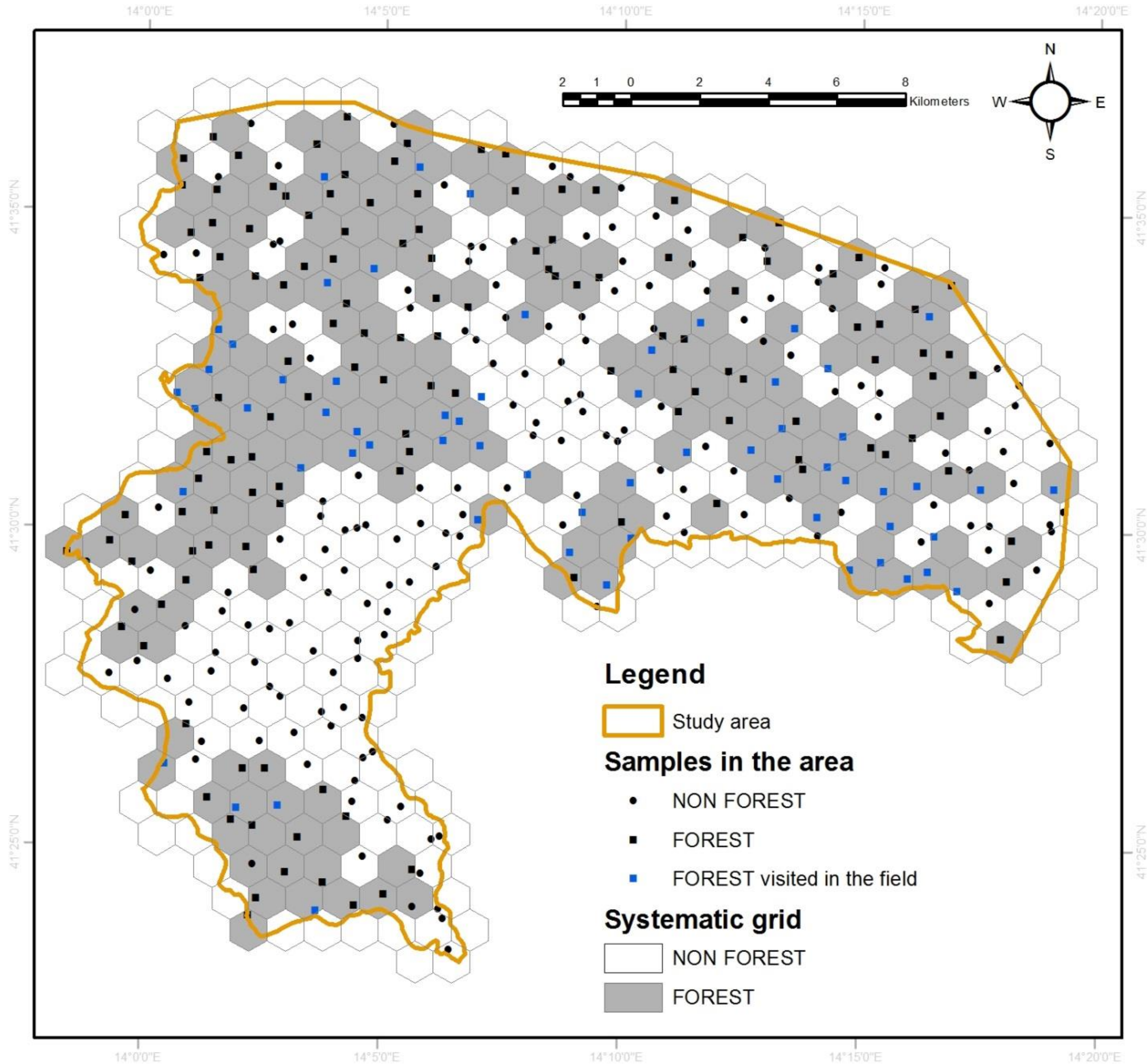
- Estimation of forest total above-ground biomass with a wall-to-wall approach ("mapping" or "spatially explicit" or "geographically continuous")
- Comparing area-based vs. echoes-based approaches
- Comparing parametric vs. non-parametric approaches

Study area located in Regione Molise in Italy 36380 ha wide



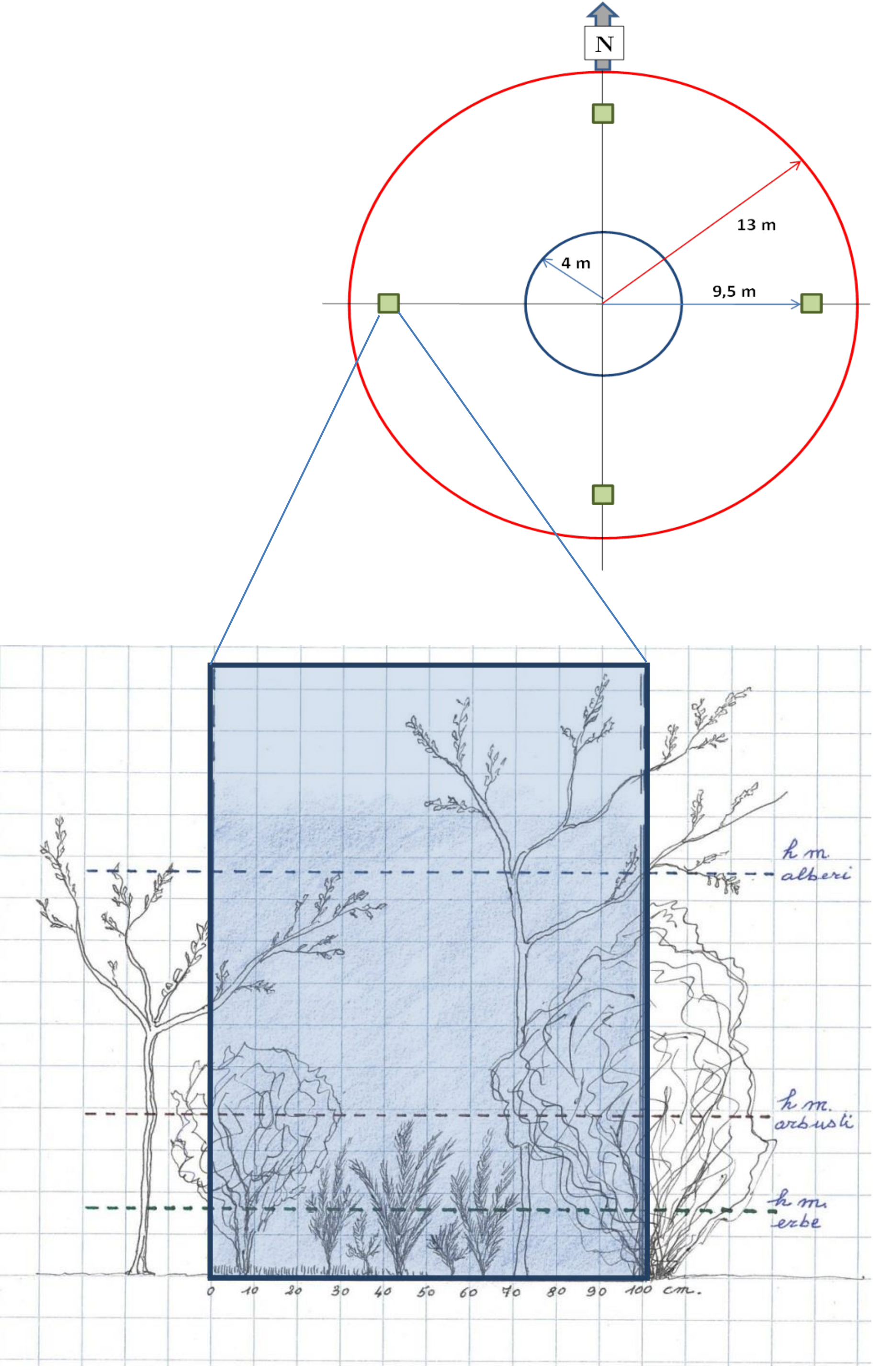
SAMPLING DESIGN

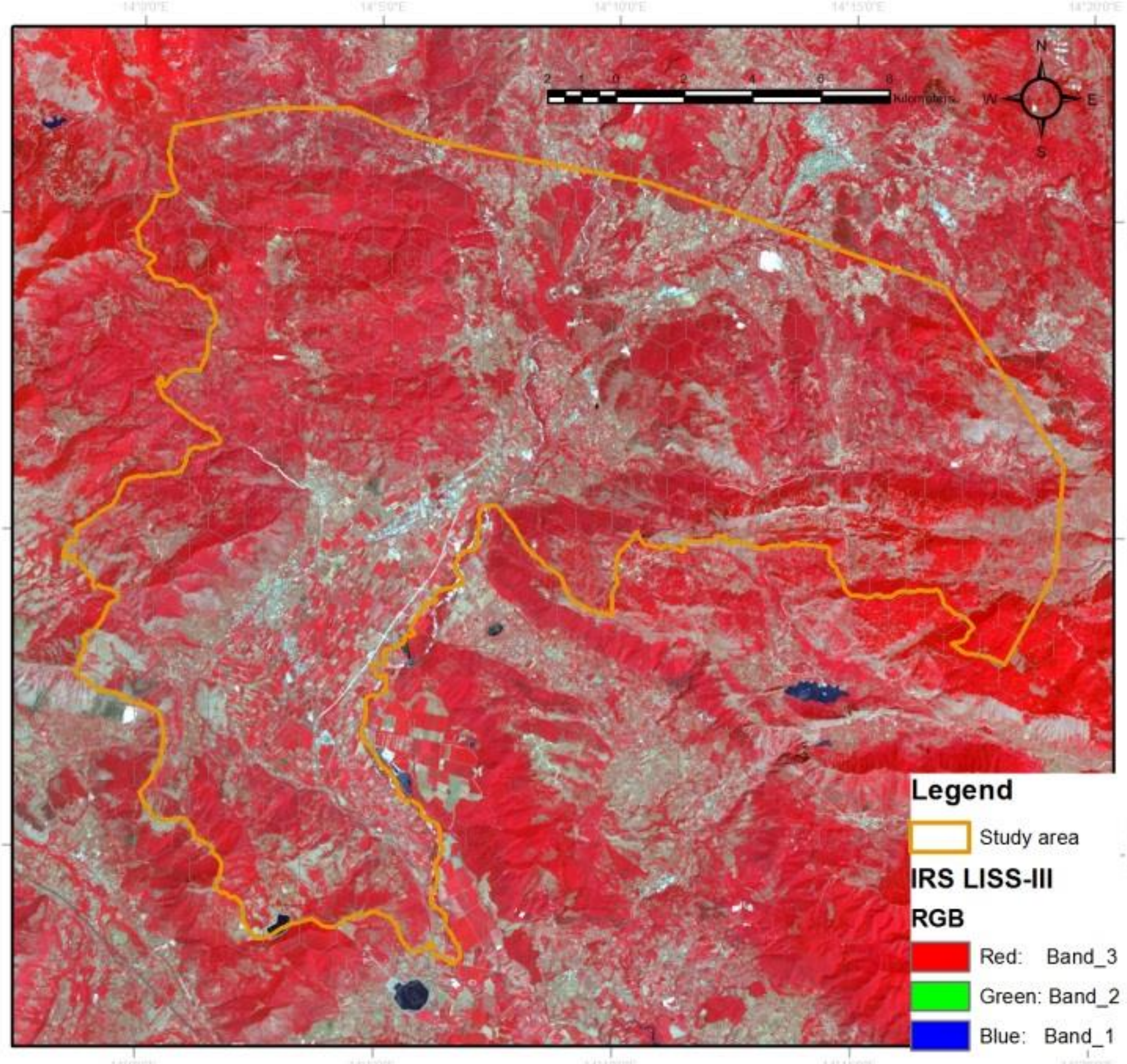
- Systematic sampling design random origin with hexagons of 1 km^2 each
- One random sampling unit in each hexagon = 368 units located inside the study area
- The sampling units were classified as "forest" and "non forest" on the basis of an aerial high resolution photography
- Resulting in 171 "non forest" units and 197 "forest" units.



FIELD PROTOCOL

- 13 m radius circular plot
- trees callipered with Diameter at Breast Height (DBH) limit of 2.5 in the 4 m subplot
- for callipered trees, height, species, and tree location registered
- in 4 subplot of 1x1 for each plot biomass removed (small trees, regeneration, bushes, herbs)
- use of biomass equations developed in the framework of the last Italian National Forest Inventory (Tabacchi et al., 2011) for tree biomass
- dry weight for subplot biomass
- total aboveground biomass aggregated at plot level

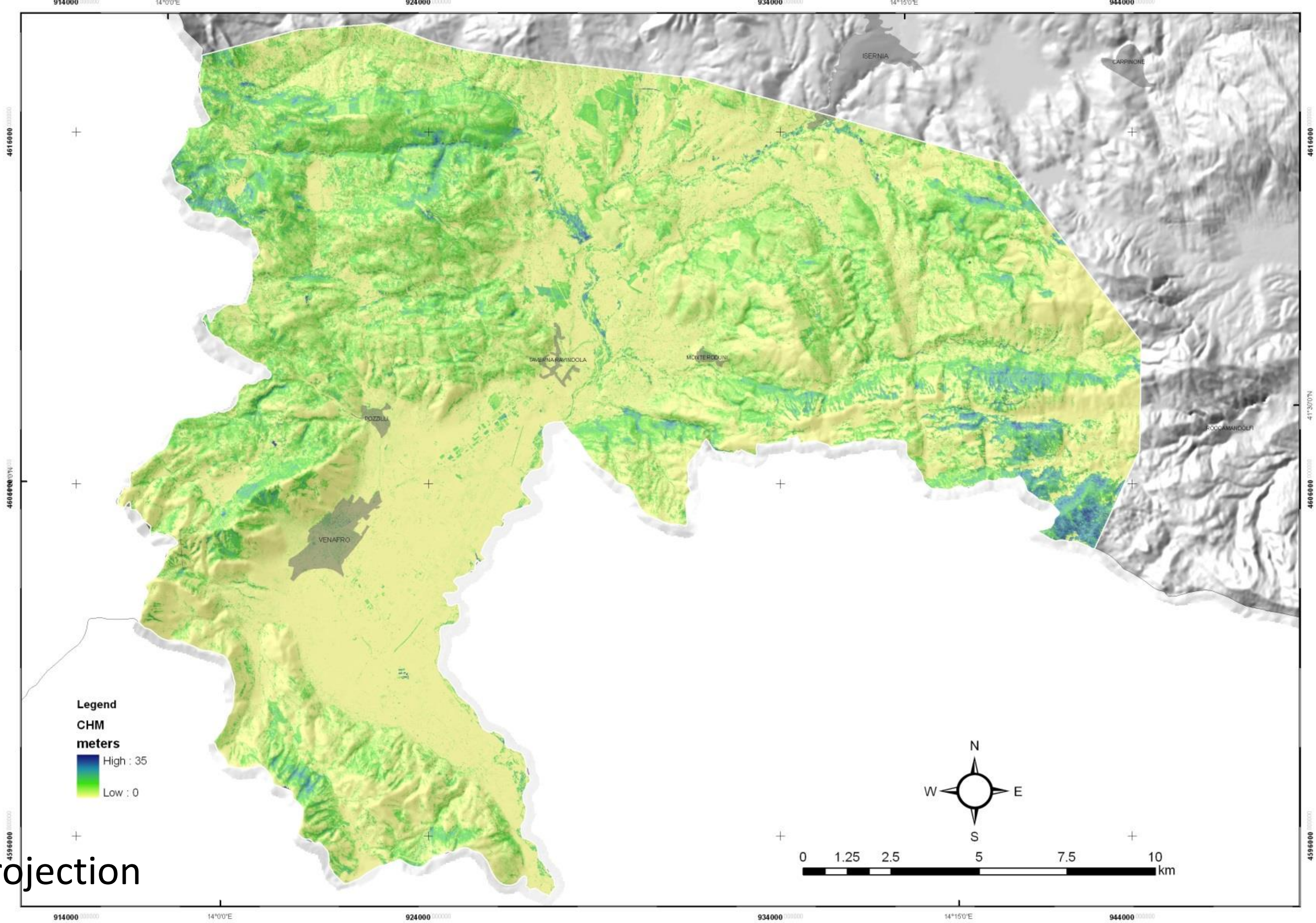




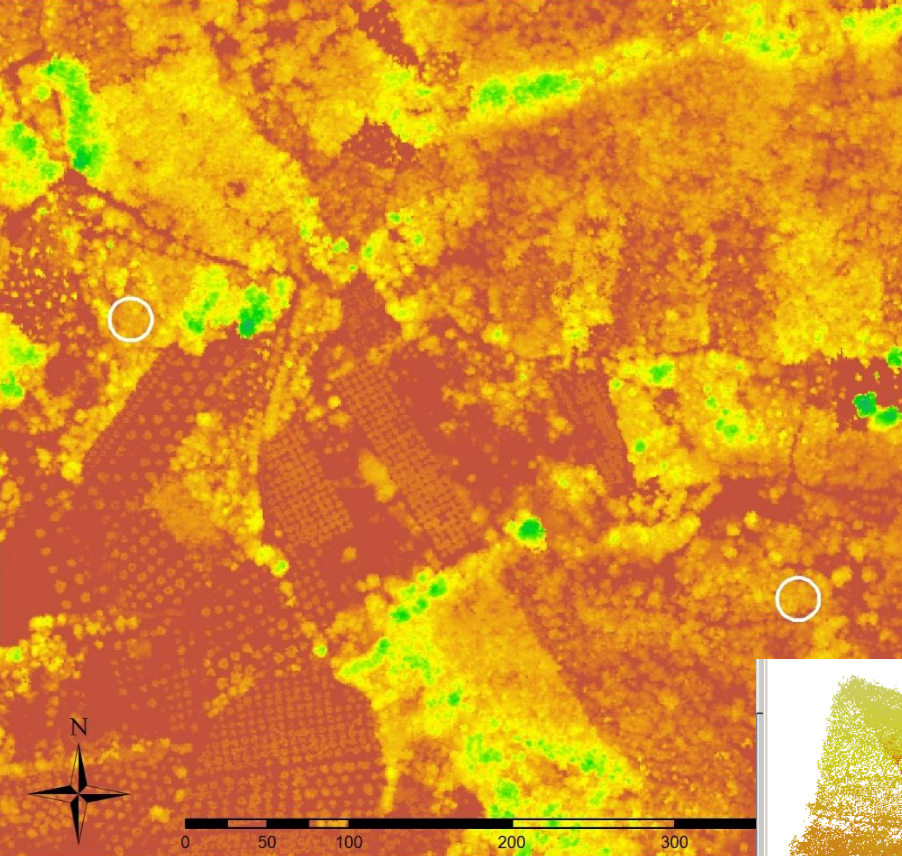
IRS LISS-III

- 4 spectral bands at 0.52-0.59, 0.62-0.68, 0.77-0.86, 1.55-1.70 µm covering green, red, near Infra Red (IR) and medium IR channels
- Geometric resolution 20 m

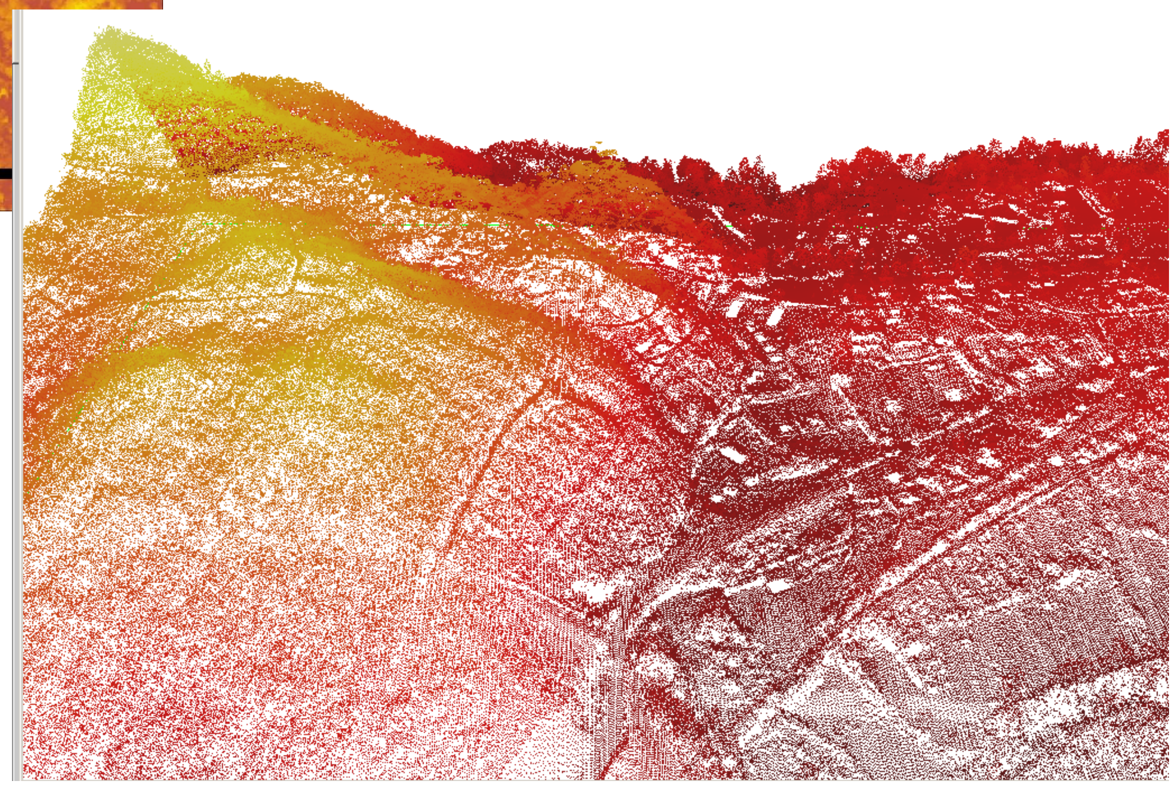
LiDAR pre-elaborations



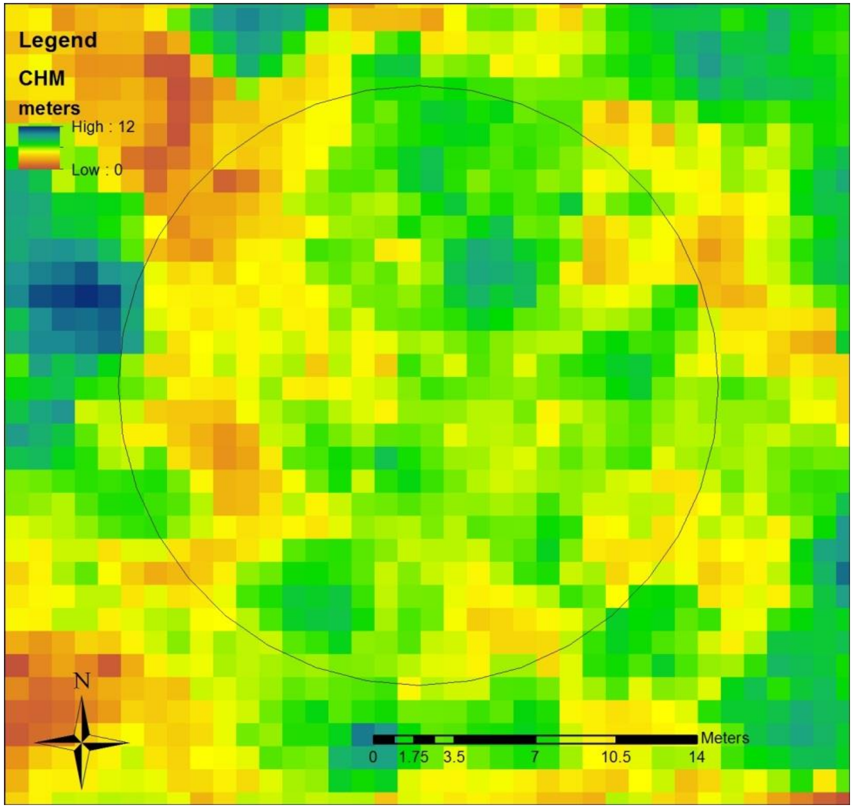
UTM 33N WGS 84 projection
Echoes filtering
Tiling (2 km x 2 km)
Ground / non-ground classification
DTM generation
Relative heights above ground calculation
DSM generation
DCM generation



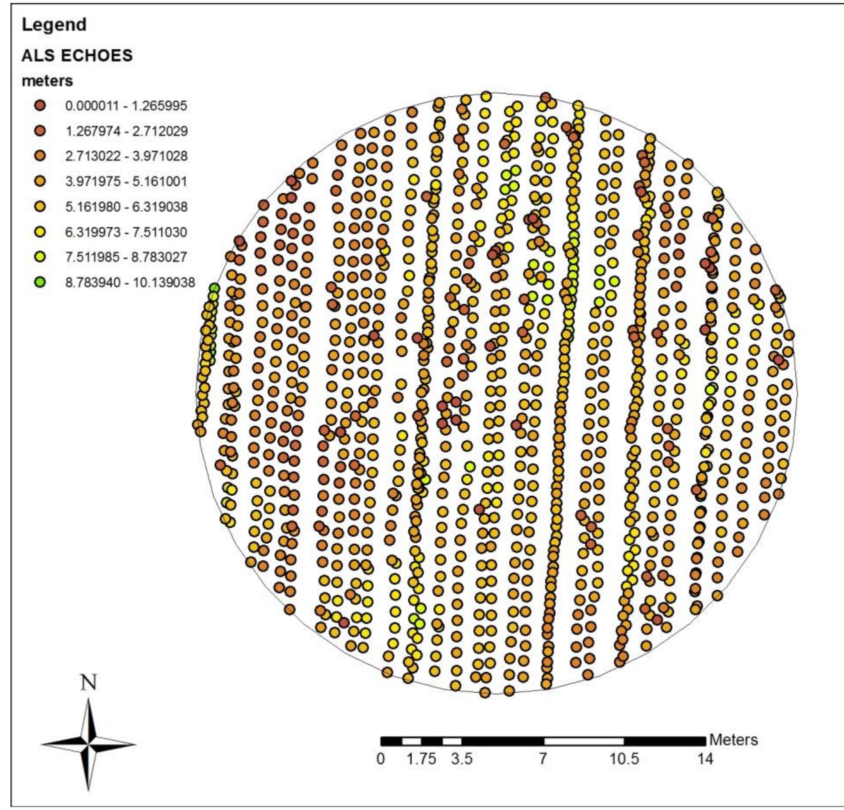
APPROACH 1
CHM for «area based» approach
1 m resolution



APPROACH 2
Raw echo pulses



area-based approach



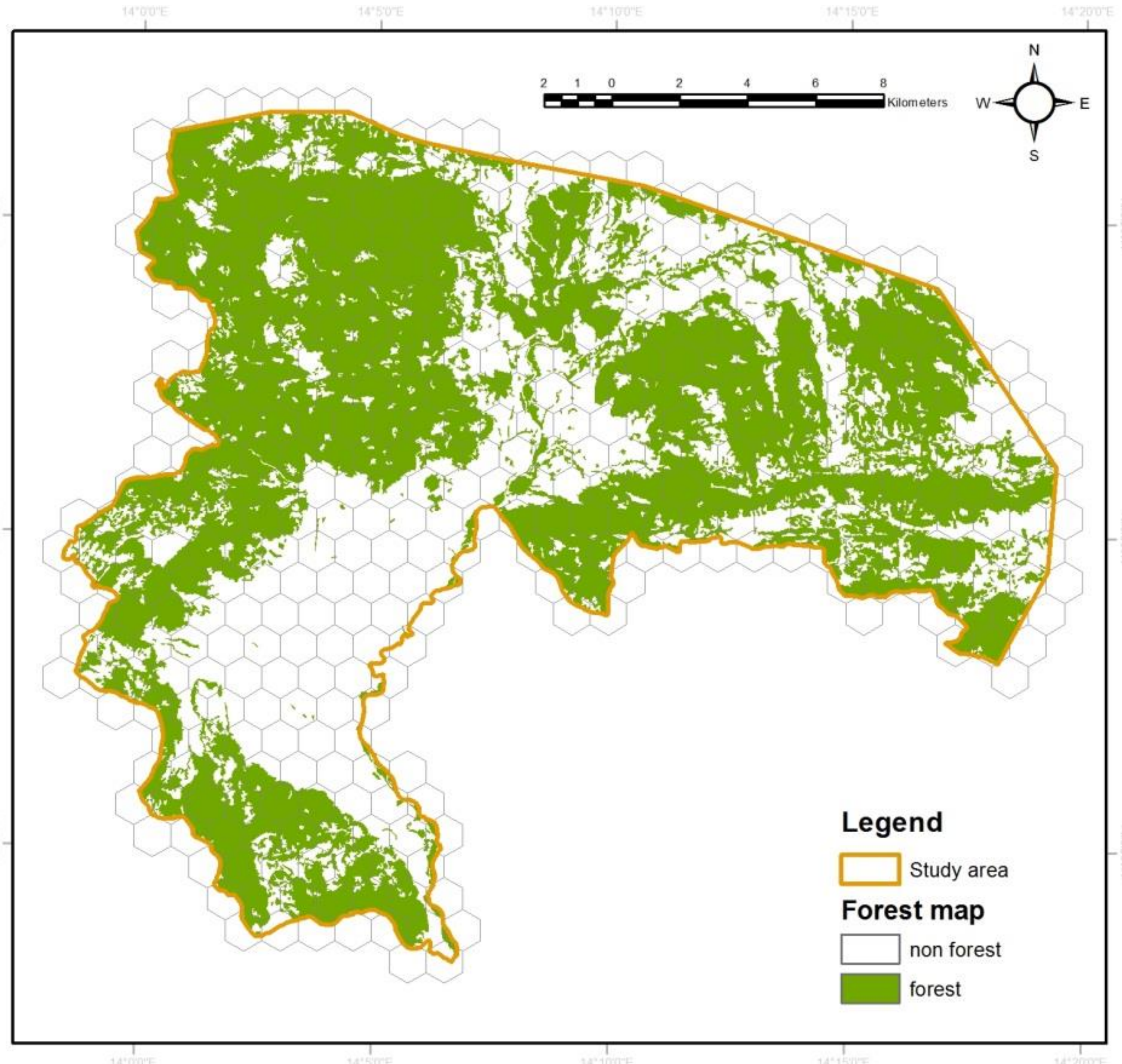
echoes-based approach

For each one of the 62 field plots metrics from CHM and from raw echo pulses were calculated

HEIGHT Minimum	LIDAR heights	minimum
HEIGHT Maximum	LIDAR heights	maximum
HEIGHT Total	LIDAR heights	sum
HEIGHT Average	LIDAR heights	average
HEIGHT Range	LIDAR heights	range
HEIGHT STDDEV	LIDAR heights	standard deviation
HEIGHT above 2m Minimum	LIDAR heights above 2 meters	minimum
HEIGHT above 2m Maximum	LIDAR heights above 2 meters	maximum
HEIGHT above 2m Total	LIDAR heights above 2 meters	sum
HEIGHT above 2m Average	LIDAR heights above 2 meters	average
HEIGHT above 2m Range	LIDAR heights above 2 meters	range
HEIGHT above 2m STDDEV	LIDAR heights above 2 meters	standard deviation
IRS B1 Minimum	Digital number from IRS LISS-III B1	minimum
IRS B1 Maximum	Digital number from IRS LISS-III B1	maximum
IRS B1 Total	Digital number from IRS LISS-III B1	sum
IRS B1 Average	Digital number from IRS LISS-III B1	average
IRS B1 Range	Digital number from IRS LISS-III B1	range
IRS B1 STDDEV	Digital number from IRS LISS-III B1	standard deviation
IRS B2 Minimum	Digital number from IRS LISS-III B2	minimum
IRS B2 Maximum	Digital number from IRS LISS-III B2	maximum
IRS B2 Total	Digital number from IRS LISS-III B2	sum
IRS B2 Average	Digital number from IRS LISS-III B2	average
IRS B2 Range	Digital number from IRS LISS-III B2	range
IRS B2 STDDEV	Digital number from IRS LISS-III B2	standard deviation
IRS B3 Minimum	Digital number from IRS LISS-III B3	minimum
IRS B3 Maximum	Digital number from IRS LISS-III B3	maximum
IRS B3 Total	Digital number from IRS LISS-III B3	sum
IRS B3 Average	Digital number from IRS LISS-III B3	average
IRS B3 Range	Digital number from IRS LISS-III B3	range
IRS B3 STDDEV	Digital number from IRS LISS-III B3	standard deviation
IRS B4 Minimum	Digital number from IRS LISS-III B4	minimum
IRS B4 Maximum	Digital number from IRS LISS-III B4	maximum
IRS B4 Total	Digital number from IRS LISS-III B4	sum
IRS B4 Average	Digital number from IRS LISS-III B4	average
IRS B4 Range	Digital number from IRS LISS-III B4	range
IRS B4 STDDEV	Digital number from IRS LISS-III B4	standard deviation
NUMB HITS Minimum	Number of LIDAR hits	minimum
NUMB HITS Maximum	Number of LIDAR hits	maximum
NUMB HITS Total	Number of LIDAR hits	sum
NUMB HITS Average	Number of LIDAR hits	average
NUMB HITS Range	Number of LIDAR hits	range
NUMB HITS STDDEV	Number of LIDAR hits	standard deviation
INTENSITY Minimum	LIDAR intensity	minimum
INTENSITY Maximum	LIDAR intensity	maximum
INTENSITY Total	LIDAR intensity	sum
INTENSITY Average	LIDAR intensity	average
INTENSITY Range	LIDAR intensity	range
INTENSITY STDDEV	LIDAR intensity	standard deviation

METRICS

Here we present results from the «area-based» approach only

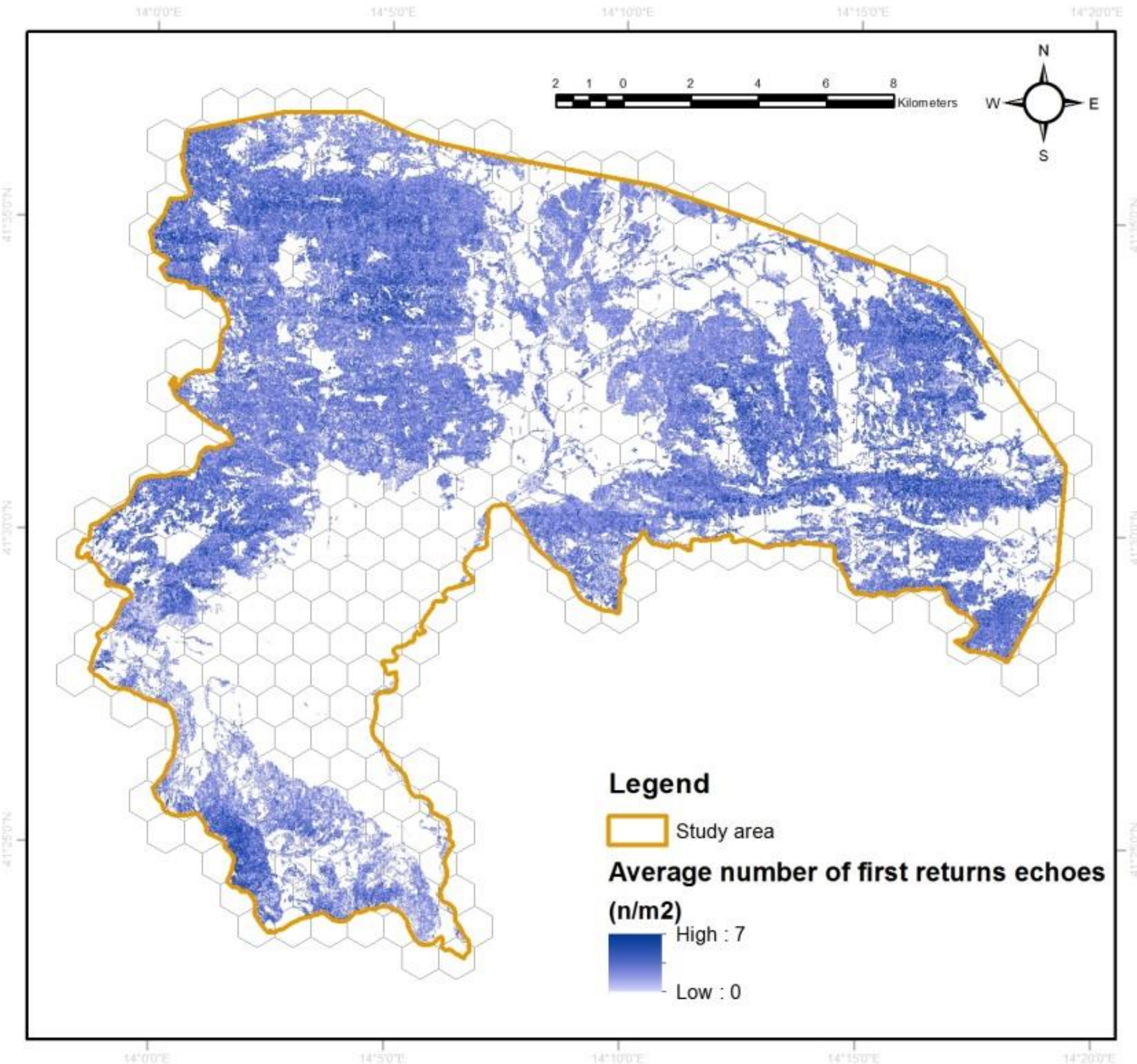


Generation of a 23 x 23 m systematic raster grid

Classification of each pixel as forest and non-forest on the basis of a local forest type map (scale 1:10.000, MMU 0.5 ha)

Total of forest pixels:

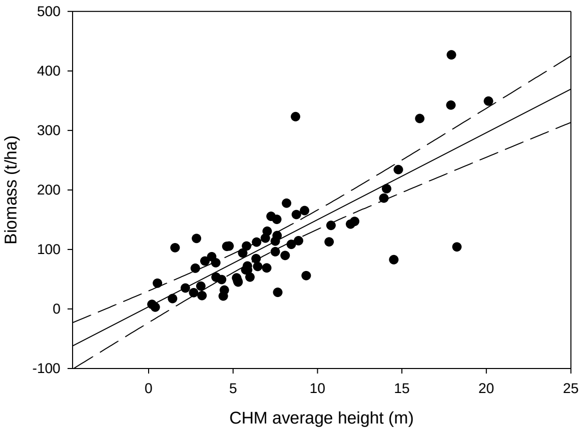
N_F = 405233



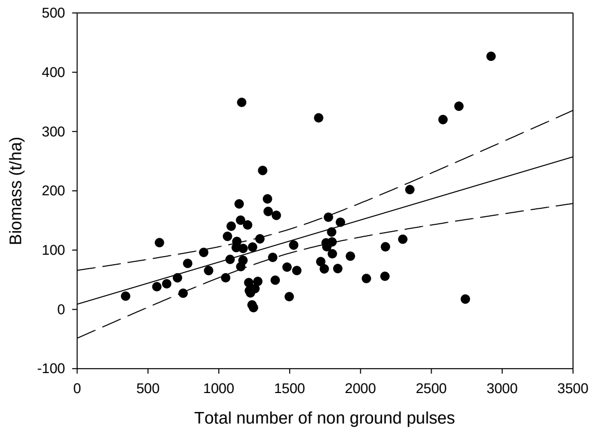
Number of echoes standardized on the number of flight lines to reduce the effect strip overlaps

Calculation of LiDAR and IRS metrics for each one of the 405233 forest pixels

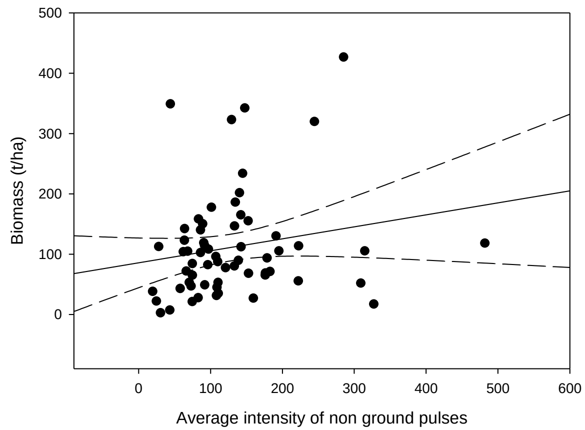
Study of the univariate relationship between biomass (dependent variable) and LiDAR and IRS metrics (independent variables)



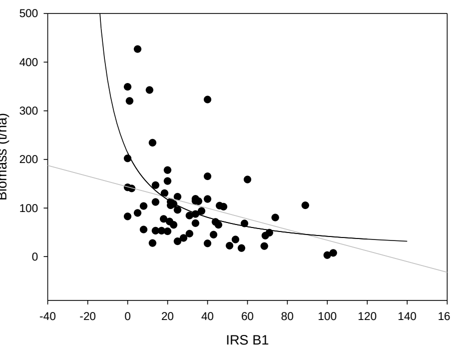
R = 0.7771
SE= 55.67



R = 0.4443
SE= 79.23



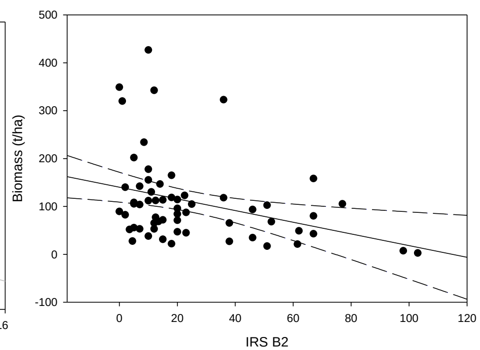
R = 0.1892
SE= 76.85



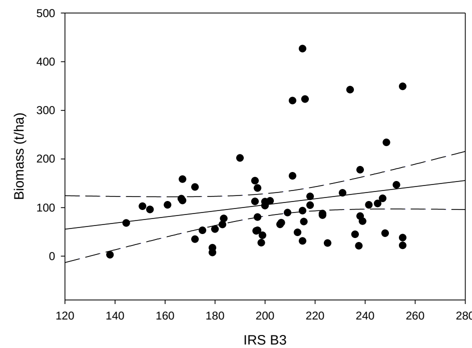
Linear: R = 0.4534, SE= 79.56

Non linear: R = 0.5141, SE= 76.56

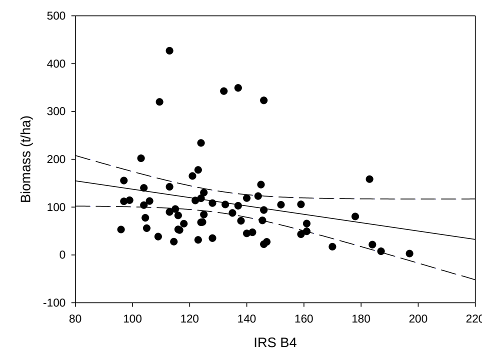
f = (a*b)/(b+x)



R = 0.3374
SE= 84.03



R = 0.2134
SE= 87.20



R = 0.2413
SE= 86.62

Parametric estimation

First a forward stepwise regression selected only two variables.
The resulting linear model was:

	Coef.	Std. Coeff.	Std. Error	F-to-Remove	P
Constant	-44.412		19.986		
HEIGHT Average	13.336	0.702	1.542	74.795	<0.001
NUMB HITS Total	0.0394	0.248	0.0129	9.321	0.003

With R = 0.811, Rsqr = 0.657 , Adj Rsqr = 0.645, Standard Error of Estimate = 52.70

To accommodate both linearity and the nonlinearity, we selected the following model:

$$y_i = \beta_0 \cdot x_{i1}^{\beta_1} \cdots x_{ip}^{\beta_p} + \varepsilon \qquad \text{(Eq: 1)}$$

The model can be linearized as:

$$\ln(y_i) = \ln(\beta_0) + \beta_1 \cdot \ln(x_{i1}) + \cdots + \beta_p \ln(x_{ip}) + \varepsilon \qquad \text{(Eq: 2)}$$

We fit the model of Eq. (2) using all possible combinations of all numbers of the independent variables. Then we fit the model of Eq. (1) using the best combinations from fitting the model of Eq. (2).
For all models, the best combinations were very stable: average ALS height and total number of ALS hits. Adding any more variables did not statistically significantly improve the quality of fit of the model to the data. The model selected was,

$$\hat{y}_i = 9.6389 \cdot x_{i1}^{0.9251} \cdot x_{i2}^{0.5664}$$

Where:
x_{i1}=average ALS height for the i^{th} plot
x_{i2}=total number of hits for the i^{th} plot.
Pseudo-R^2=0.70

Model assisted estimator

.

The initial estimator of mean biomass per unit area is:

$$\hat{\mu}_{initial} = \frac{1}{N_F}\sum_{j=1}^{N}\hat{y}_j$$

Where N_F is the population size of the forest pixels and $\hat{y}_j$ is the model prediction for the *j*th pixel.

However, this estimator may be biased as a result of systematic model prediction error.

The bias can be estimated as:

$$B\hat{i}as(\hat{\mu}_{initial}) = \frac{1}{n}\sum_{U_i \varepsilon S}(\hat{y}_j - y_j)$$

where *Ui* is a grid cell in the field sample, S, and n is the field sample size. The design-based, model-assisted regression estimator is:

$$\hat{\mu}_{MA} = \hat{\mu}_{initial} - B\hat{i}as(\hat{\mu}_{initial}) = \frac{1}{N}\sum_{j=1}^{N}\hat{y}_j - \frac{1}{n}\sum_{U_j \varepsilon S}(\hat{y}_j - y_j)$$

The variance is estimated as:

$$V\hat{a}r(\hat{\mu}_{MA}) = \frac{1}{n(n-1)}\sum_{U_j \varepsilon S}(\hat{y}_j - y_j)^2$$

Non-parametric estimation: *k*-NN

Several approaches under test: *k*-Nearest Neighbours, CART, Random Forests, Stochastic Gradient Boosting

k-NN approach following a local optimization (Chirici et al., 2008):

1 HEIGHT Minimum
2 HEIGHT Maximum
3 HEIGHT Total
4 HEIGHT Average
5 HEIGHT Range
6 HEIGHT STDDEV
7 IRS B1 Average
8 IRS B2 Average
9 IRS B3 Average
10 IRS B4 Average
11 NUMB HITS Minimum
12 NUMB HITS Maximum
13 NUMB HITS Total
14 NUMB HITS Average
15 NUMB HITS Range
16 NUMB HITS STDDEV
17 INTENSITY Minimum
18 INTENSITY Maximum
19 INTENSITY Total
20 INTENSITY Average
21 INTENSITY Range
22 INTENSITY STDDEV

Number of variables	Number of combinations	Combination of variables with smallest RMSE	k	RMSE	PseudoR ^{2}
1	22	4	5	59.73	0.537
2	231	4 12	5	58.15	0.561
3	1,540	4 6 12	4	57.26	0.574
4	**7,315**	**4 6 12 14**	**3**	**56.10**	**0.591**
5	26,634	4 6 11 12 14	3	56.10	0.591
6	74,613	4 6 11 12 14 17	2	56.13	0.591
7	170,544	1 4 5 6 11 12 15	2	56.66	0.583
8	319,770	1 4 5 6 11 12 15 16	4	56.66	0.583
9	497,420	1 4 5 6 11 12 15 16 17	2	56.66	0.583
10	646,646	1 2 3 4 5 6 7 8 9 13	4	57.94	0.564

Multidimensional distance metrics: euclidean, mahalanobis, **fuzzy** (Chirici et al., 2008)

k-NN predictions: a design based perspective

The forested study area was partitioned into a population F of N_F pixels (N_F=405233).
Denote by y_j the value of the forest attribute Y (biomass) at pixel j.
An estimate of $\bar{Y}_F$ for any $j \in F$ is needed, together with an estimate of their mean over the whole study area:

$$\tilde{\bar{Y}}_F = \frac{1}{N_F}\sum_{j\in F}\tilde{y}_j + \frac{N}{N_F}\frac{n_F}{n}\bar{e}_F$$

With estimated variance:

$$\tilde{V}\left(\tilde{\bar{Y}}\right) = \left(\frac{N}{N_F}\frac{n_F}{n}\right)^2 \frac{1}{n_F(n_F-1)}\sum_{j\in S_F}(e_j-\bar{e}_F)^2$$

Respect to Baffetta et al. 2009 in Molise one more phase in sampling was introduced. Estimators under development!!
For the moment even, if potentially biased, the average of the k-NN predictions was considered the estimator for the whole area:

$$\tilde{\bar{Y}}_{k\text{-NN}} = \frac{1}{N}\sum_{j\in U}\tilde{y}_j$$

Baffetta, F., Fattorini, L., Franceschi, S., & Corona, P. (2009). Design-based approach to k-nearest neighbours technique for coupling field and remotely sensed data in forest surveys. Remote Sensing of Environment, 113, 463-475

Forest biomass design based estimation

N number of first phase hexagons sampling units (368)
A total area of the N hexagons (36800 ha)
a plot area (530.929 m^2 or 0.0530929 ha)
first phase sampling units in forest (204)
n second phase sampling units (62)
S second phase sample
B_j biomass measured in the *j-th* plot

Total estimated biomass on the basis of plot *j:* $$\hat{T}_j = \frac{A}{a} B_j$$

Second phase estimation of the total biomass: $$\hat{\hat{T}}_2 = \frac{N_f}{N} \frac{1}{n} \sum_{j \in S} \hat{T}_j$$

With variance estimated as:

$$V_{\hat{\hat{T}}_2} = \frac{N_f - n}{N} \frac{1}{n(n-1)} \sum_{j \in S} (\hat{T}_j - \hat{\hat{T}}_2)^2 + \frac{N_f}{N^2 n} \sum_{j \in S} \hat{T}_j^2 - \frac{2}{N^2 (N-1)} \frac{N_f (N_f - 1)}{n(n-1)} \sum_{h > j \in S} \hat{T}_j \hat{T}_h$$

…. still under calculation

Thanks to Prof. Fattorini!

Results: whole area estimations

Design based estimate

2,277,061 tons
111.62 t/ha
Variance under calculation

Model based estimations

Parametric approach

2,147,030 tons
100.16 t/ha with SE 6.27

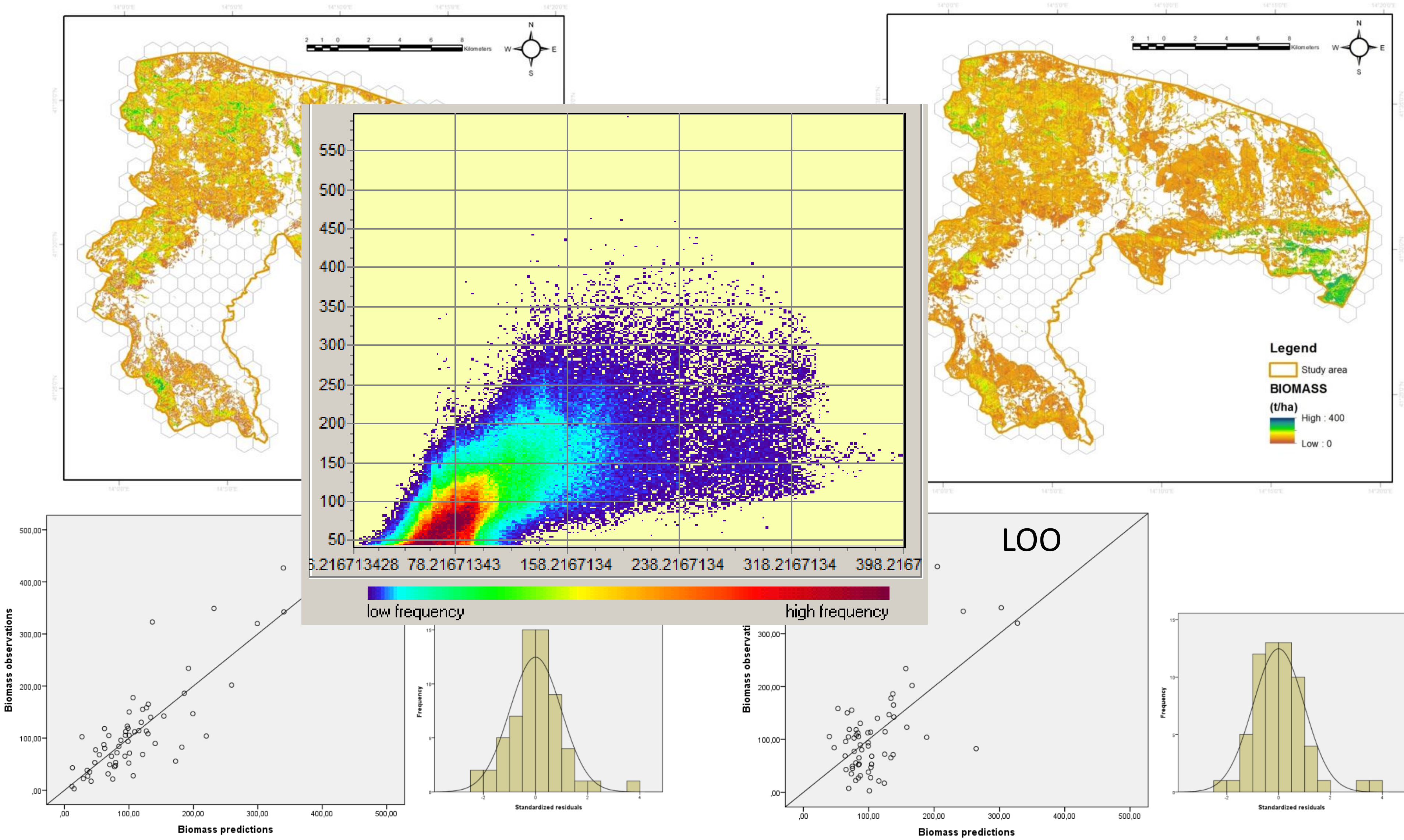
Non-Parametric approach k-NN

2,187,287 tons
102.04 t/ha with SE 8.07

Results: pixel level performances

Parametric

Non-parametric



RMSE= 37.75 t ha^{-1}

RMSE= 35.44 t ha^{-1}

Conclusions

- Both parametric and non-parametric (k-NN) approaches are able to estimate total above area biomass
- The multispectral information from IRS do not improve the performances based on ALS metrics only
- The use of non a linear model in the parametric approach increased very little the parametric model performance
- The selection of predictors from parametric and non-parametric approaches independently lead to the selection of the same variables

- To be done:
  - design based estimators
  - comparison with echoes metrics
  - use of other non-parametric approaches (Salford Systems)

A special thanks to Lorenzo Fattorini for design-based estimators and Lorenzo Bottai for LiDAR processing